Nandini Ramakrishnan

Readings

Moving on from Weiser’s Vision of Calm Computing:

Engaging UbiComp Experiences

Yvonne Rogers

School of Informatics, Indiana University, 901 East 10th Street,

Bloomington, IN47408, USA

yrogers@indiana.edu

Abstract. A motivation behind much UbiComp research has been to make our

lives convenient, comfortable and informed, following in the footsteps of

Weiser’s calm computing vision. Three themes that have dominated are context

awareness, ambient intelligence and monitoring/tracking. While these avenues

of research have been fruitful their accomplishments do not match up to any-

thing like Weiser’s world. This paper discusses why this is so and argues that is

time for a change of direction in the field. An alternative agenda is outlined that

focuses on engaging rather than calming people. Humans are very resourceful

at exploiting their environments and extending their capabilities using existing

strategies and tools. I describe how pervasive technologies can be added to the

mix, outlining three areas of practice where there is much potential for profes-

sionals and laypeople alike to combine, adapt and use them in creative and con-

structive ways.

Keywords: calm computing, Weiser, user experiences, engaged living, Ubi-

Comp history, pervasive technologies, proactive computing.

1 Introduction

Mark Weiser’s vision of ubiquitous computing has had an enormous impact on the

directions that the nascent field of UbiComp has taken. A central thesis was that while

“computers for personal use have focused on the excitement of interaction...the most

potentially interesting, challenging and profound change implied by the ubiquitous

computing era is a focus on calm.” [46]. Given the likelihood that computers will be

everywhere, in our environments and even embedded in our bodies, he argued that

they better “stay out of the way” and not overburden us in our everyday lives. In con-

trast, his picture of calm technology portrayed a world of serenity, comfort and aware-

ness, where we are kept perpetually informed of what is happening around us, what is

going to happen and what has just happened. Information would appear in the centre

of our attention when needed and effortlessly disappear into the periphery of our at-

tention when not.

Now regarded as the forefather of UbiComp, Weiser has inspired governments, re-

searchers and developers across the globe. Most prominent was the European Com-

munity’s Disappearing Computer initiative in the late 90s and early 2000s, that

funded a large number of research projects to investigate how information technology

could be diffused into everyday objects and settings and to see how this could lead to

Moving on from Weiser’s Vision of Calm Computing 405

new ways of supporting and enhancing people’s lives that went above and beyond

what was possible using desktop machines. Other ambitious and far-reaching projects

included MIT’s Oxygen, HP’s CoolTown, IBM’s BlueEyes, Philips Vision of the

Future and attempts by various telecom companies and academia to create the ulti-

mate ‘smart home’, e.g., Orange-at-Home and Aware Home. A central aspiration

running through these early efforts was that the environment, the home, and our pos-

sessions would be aware, adapt and respond to our varying comfort needs, individual

moods and information requirements. We would only have to walk into a room, make

a gesture or speak aloud and the environment would bend to our will and respond or

react as deemed appropriate for that point in time.

Considerable effort has gone into realizing Weiser’s vision in terms of developing

frameworks, technologies and infrastructures. Proactive computing was put forward

as an approach to determine how to program computers to take the initiative to act on

people’s behalf [43]. The environment has been augmented with various computa-

tional resources to provide information and services, when and where desired, with

the implicit goal of “assisting everyday life and not overwhelming it” [1]. An assort-

ment of sensors have been experimented with in our homes, hospitals, public build-

ings, physical environments and even our bodies to detect trends and anomalies, pro-

viding a dizzying array of data about our health, movements, changes in the environ-

ment and so on. Algorithms have been developed to analyze the data in order for

inferences to be made about what actions to take for people. In addition, sensed data

is increasingly being used to automate mundane operations and actions that we would

have done in our everyday worlds using conventional knobs, buttons and other physi-

cal controls. For example, our favorite kind of music or TV show that we like to exer-

cise to will automatically play as we enter a gym. Sensed data is also being used to

remind us of things we often forget to do at salient times, such as detecting the ab-

sence of milk in the fridge and messaging us to buy a carton when passing the grocery

store.

But, as advanced and impressive as these endeavors have been they still do not

match up to anything like a world of calm computing. There is an enormous gap be-

tween the dream of comfortable, informed and effortless living and the accomplish-

ments of UbiComp research. As pointed out by Greenfield [20] “we simply don’t do

‘smart’ very well yet” because it involves solving very hard artificial intelligence

problems that in many ways are more challenging than creating an artificial human

[26]. A fundamental stumbling block has been harnessing the huge variability in what

people do, their motives for doing it, when they do it and how they do it. Ethno-

graphic studies of how people manage their lives – ranging from those suffering from

Alzheimer’s Disease to high-powered professionals – have revealed that the specifics

of the context surrounding people’s day-to-day living are much more subtle, fluid and

idiosyncratic than theories of context have led us to believe [40]. This makes it diffi-

cult, if not impossible, to try to implement context in any practical sense and from

which to make sensible predictions about what someone is feeling, wanting or need-

ing at a given moment. Hence, while it has been possible to develop a range of simple

UbiComp systems that can offer relevant information at opportune moments (e.g.,

reminding and recommending to us things that are considered useful and important) it

is proving to be much more difficult to build truly smart systems that can understand

or accurately model people’s behaviors, moods and intentions.

406 Y. Rogers

The very idea of calm computing has also raised a number of ethical and social

concerns. Even if it was possible for Weiser’s dream to be fulfilled would we want to

live in such a world? In particular, is it desirable to depend on computers to take on

our day-to-day decision-making and planning activities? Will our abilities to learn,

remember and think for ourselves suffer if we begin to rely increasingly on the envi-

ronment to do them for us? Furthermore, how do designers decide which activities

should be left for humans to control and which are acceptable and valuable for the

environment to take over responsibility for?

In this paper I argue that progress in UbiComp research has been hampered by in-

tractable computational and ethical problems and that we need to begin taking stock

of both the dream and developments in the field. In particular, we need to rethink the

value and role of calm and proactive computing as main driving forces. It is without

question that Weiser’s enormous legacy will (and should) continue to have an impact

on UbiComp developments. However, sufficient time has passed since his untimely

death and it should be possible now for researchers to take a critical stance. As part of

this exercise, I propose that the field needs to broaden its scope, setting and address-

ing other goals that are more attainable and down-to-earth. New agendas need also to

be outlined that can guide, stimulate and challenge UbiComp (and other) researchers

and developers, building upon the growing body of research in the field.

To this end, I propose one such alternative agenda which focuses on designing

UbiComp technologies for engaging user experiences. It argues for a significant shift

from proactive computing to proactive people; where UbiComp technologies are

designed not to do things for people but to engage them more actively in what they

currently do. Rather than calm living it promotes engaged living, where technology is

designed to enable people to do what they want, need or never even considered before

by acting in and upon the environment. Instead of embedding pervasive computing

everywhere in the environment it considers how UbiComp technologies can be cre-

ated as ensembles or ecologies of resources, that can be mobile and/or fixed, to serve

specific purposes and be situated in particular places. Furthermore, it argues that peo-

ple rather than computers should take the initiative to be constructive, creative and,

ultimately, in control of their interactions with the world – in novel and extensive

ways.

While this agenda might appear to be a regressive step and even an anathema to

some ardent followers of Weiser’s vision, I argue that it (and other agendas) will turn

out to be more beneficial for society than persisting with following an unrealistic

goal. Current technological developments together with emerging findings from user

studies, showing how human activities have been positively extended by ‘bounded’

(as opposed to pervasive) technologies, suggest that much can be gained from re-

conceptualizing UbiComp in terms of designing user experiences that creatively,

excitedly, and constructively extend what people currently do. This does not mean

that the main tenet of Weiser’s vision be discarded (i.e., computers appearing when

needed and disappearing when not) but rather we begin to entertain other possibilities

– besides calmness – for steering UbiComp research. Examples include extending and

supporting personal, cognitive and social processes such as habit-changing, problem-

solving, creating, analyzing, learning or performing a skill. Ultimately, research and

development should be driven by a better understanding of human activity rather than

Moving on from Weiser’s Vision of Calm Computing 407

what has tended to happen, namely, “daring to intervene, clumsily, in situations that

already work reasonably well” [20, p231].

In the remainder of this paper I offer a constructive critique of Weiser’s vision and

the subsequent research that has followed in its footsteps. I then outline an alternative

agenda for UbiComp, highlighting pertinent questions, concerns and illustrative ex-

amples of how it can be achieved.

2 Weiser’s Vision Revisited and Early Research

To illustrate how his early vision of ubiquitous computing could work, Weiser [47]

presented a detailed scenario about a day in the life of Sal, an executive single mother.

The scenario describes what Sal gets up to, as she moves from her domestic world to

her work place, during which she is perpetually informed of the goings on of her

family, neighbors, fellow citizens and work colleagues. With this knowledge she is

able to keep up-to-date, avoid obstacles, make the most of her time and conduct her

work – all in smooth and effective ways. The scenario emphasizes coziness, comfort

and effortlessness:

“Sal awakens: she smells coffee. A few minutes ago her alarm clock, alerted by her

restless rolling before waking, had quietly asked “coffee?”, and she had mumbled

“yes.” “Yes” and “no” are the only words it knows.

Sal looks out her windows at her neighborhood. Sunlight and a fence are visible

through one, but through others she sees electronic trails that have been kept for her

of neighbors’ coming and going during the early morning. Privacy conventions and

practical data rates prevent displaying video footage, but time markers electronic

tracks on the neighborhood map let Sal feel cozy in her street.”

In this small excerpt we see how the world evolves around Sal’s assumed needs,

where computers, cameras and sensors are embedded into her world to make her life

super efficient, smooth and calm. It is as if she glides through life, where everything

is done or laid out for her and whenever there is potential for frustration, such as a

traffic jam or parking problem, the invisible computers come to her rescue and gently

inform her of what to do and where to go. It is worth drawing an analogy here with

the world of the landed aristocracy in Victorian England who’s day-to-day live was

supported by a raft of servants that were deemed to be invisible to them. This scenario

also highlights the ethical issues that such an informed world needs to address,

namely the importance of establishing appropriate levels of privacy that are consid-

ered acceptable by a community (e.g., having abstract digital trails rather than video

footage to ensure anonymity).

The core topics raised in Weiser’s seminal papers have motivated much subsequent

UbiComp research. Most prominent themes are context-aware computing, ambi-

ent/ubiquitous intelligence and recording/tracking and monitoring. (N.B. It should be

noted that these are not mutually exclusive but overlap in the aims and methods used.)

2.1 Context-Aware Computing

Context-aware computing focuses on detecting, identifying and locating people’s

movements, routines or actions with a view to using this information to provide

408 Y. Rogers

relevant information that may augment or assist a person or persons. Many projects

have been conducted under this heading to the extent that it has been noted that ubiq-

uitous computing is sometimes called context-aware computing [12]. In a nutshell,

context is viewed as something that can be sensed and measured using location, time,

person, activity type and other dimensions. An example of an early context-sensitive

application was comMotion that used location information and a speech output sys-

tem to inform people when they were driving or cycling past a store to buy the grocer-

ies they needed [30].

A motivation behind much context-aware computing is to find ways of compensat-

ing for limitations in human cognition, e.g., attention, memory, learning, comprehen-

sion, and decision-making, through the use of sensor-based and computational tools.

For example, augmented cognition – originating in military research – seeks to de-

velop methods “to open bottlenecks and address the biases and deficits in human

cognition” by continually sensing the ongoing context and inferring what strategies to

employ to help people in their tasks [5].

Key questions in context-aware computing concern what to sense, what form and

what kind of information to represent to augment ongoing activities. A number of

location and tagging technologies have been developed, such as RFID, satellite, GPS

and ultrasonics, to enable certain categories of information to be tracked and detected.

Many of these, however, have been beset with detection and precision limitations,

sometimes resulting in unreliable and inaccurate data. Recent advances in cognitive

radio technology that is software defined (SDR), promises to be more powerful; wire-

less systems will be able to locate and link to locally unused radio frequency, based

on the ability to sense and remember various factors, such as human behavior, making

them more dependable and more aware of their surroundings [4]. The advocates of

this new technology portray its potential for highly complex settings, such as combat

war zones to help commanders from different friendly forces stay appraised of the

latest situation, through voice, data and video links, thereby reducing collateral dam-

age [4].

While newer technological developments may enable more accurate data to be de-

tected and collected it is questionable as to how effectively it can be used. It still in-

volves Herculean efforts to understand, interpret and act upon in real-time and in

meaningful ways. Context-aware systems that attempt to guide a person through cer-

tain activities require models of human behavior and intentionality that are based on

rationality and predictability [40]. However, as already mentioned, people often be-

have in unpredictable and subtle ways in their day-to-day contexts. Therefore, it is

likely that context-aware systems will only ever be successful in highly constrained

settings.

2.2 Ambient and Ubiquitous Intelligence

Another dominant theme that has emerged in the field of UbiComp is ubiquitous or

ambient intelligence, i.e., computational intelligence that is part of both the physical

and the digital worlds. This approach follows on from work in artificial intelligence.

The phrase ‘right place/right time/right means’ has been sloganized with visions of

smart worlds and smart things, embedded with intelligence, that will predict people’s

needs and react accordingly [25]. Instead of reaching for the remote to change the TV

Moving on from Weiser’s Vision of Calm Computing 409

channel the smart entertainment system will do it for us, instead of browsing the web

the smart internet will find the information we need and so on. Just as it is becoming

increasingly common place for supermarkets to automatically open their doors as we

walk towards them, toilets to flush when we stand up and taps to release water as we

wave our hands under them it is envisioned that information will appear on our TVs,

watches, walls, and other displays as and when needed (e.g., children will be alerted

of dangers and tourists will be informed of points of interest when walking through an

unfamiliar city).

However, similar to context-aware computing, ambient intelligence is proving to

be a hard nut to crack. While there have been significant advances in computer vision,

speech recognition and gesture-based detection, the reality of multimodal interfaces –

that can predict and deliver with accuracy and sensitivity what is assumed people

want or need – is a long way off. One of the most well known attempts at implement-

ing ambient intelligence was IBM’s BlueEyes project, that sought to develop com-

puters that could “see” and “feel” like humans. Sensing technology was used to iden-

tify a person’s actions and to extract key information that was then analyzed to deter-

mine the person’s physical, emotional, or informational state. This was intended to be

used to help make people “more productive by performing expected actions or by

providing expected information.” The success of the BlueEyes project, however, was

limited; an example of an achievement that is posted on its website is of a television

that would turn itself on when a person in the room made eye contact with it. To turn

it off, the person could ‘tell’ it to switch off.

Such meager accomplishments in both context-aware computing and ambient intel-

ligence reflect just how difficult it can be to get a machine to behave like a human.

But it is essential that such systems be accurate for them to be accepted by humans in

their everyday context. Reading, interpreting and acting upon people’s moods, inten-

tions, desires, etc, at any given moment in an appropriate way is a highly developed

human skill that when humans get it wrong can lead to misunderstanding. When a

ubiquitous computing system gets it wrong – which is likely to be considerably more

frequent – it is likely to be more frustrating and we are likely to be less forgiving. For

example, when the system decides to switch on the TV because we happen momen-

tarily to stare into space while reading a book, it is likely to be unnerving and ex-

tremely annoying, especially if ‘it’ persistently gets it wrong.

2.3 Recording, Tracking and Monitoring

The push towards developing assistive applications through sensing and alerting has

been most marked for vulnerable people; a number of UbiComp systems have been

built to constantly check up on the elderly, the physically and mentally disabled [34].

The movements, habits, health and mishaps of such people are recorded, tracked and

presented via remote monitors to the families, carers and other people responsible for

them, who can then use the information to make decisions about whether to intervene

or administer alternative forms of medical care or help. In particular, there has been a

move towards developing ubiquitous computing systems to aid elderly people, who

need to be cared for, by helping them take their medicines regularly, checking up on

their physical health, monitoring their whereabouts and detecting when they have

fallen over [e.g., 13].

410 Y. Rogers

A number of assisted living applications and services has also been developed to

help people with loss of vision or deteriorating memory to be more independent in

their lives. For example, Cyber Crumbs was designed to help people with progressive

vision loss find their way around a building using a reader badge system that reads out

directions and warns of obstacles, such as fire hydrants [39]. Cook’s Collage was

developed as an aid for people with memory loss. It replays a series of digital still

images in a comic strip reel format depicting people’s cooking actions in situ, in-

tended to help them remember if they have forgotten a step (e.g., adding a particular

ingredient) after being distracted [45].

A reason for there being so much interest in helping the less able in UbiComp is

that explicit needs and benefits can be readily identified for these user groups. More-

over, there is an assumption that pervasive technologies offer more flexibility and

scope for providing solutions compared with other computing technologies since they

can sense, monitor and detect people’s movements, bodily functions, etc., in ways not

possible before. There is a danger, however, that such techniques may probe too far

into the lives of less able people resulting in – albeit unintentionally – ‘extreme’

forms of recording, tracking and monitoring that these people may have no control

over. For example, consider the extent to which a group of researchers went to in

order to help with the care of old people in a residential care home [6]. A variety of

monitoring devices were installed in the home, including badges on the patients and

the caregivers and switches on the room doors that detected when they were open or

closed. Load sensors were also used to measure and monitor weight changes of peo-

ple while in their beds; the primary aim was to track trends in weight gain or loss over

time. But the sensors could also be used to infer how well someone was sleeping. If

significant movement was detected during the night this could enable a caregiver to

see whether the person was having trouble sleeping (and if there was a huge increase

in weight this could be inferred as someone else getting in or on the bed).

Such panopticon developments elicit a knee-jerk reaction of horror in us. While the

motives behind such projects are altruistic they can also be naïve, overlooking how

vulnerable people’s privacy and self-respect may be being violated. Not surprisingly,

there has been enormous concern by the media and other social scientists about the

social implications of recording, tracking and re-representing people’s movements,

conversations, actions and transactions. Inevitably, a focus has been on the negative

aspects, namely a person’s right to privacy being breached. Is it right to be videoing

and sensing people when sleeping, eating, etc., especially when they are not at their

best [2]? Is it right to be providing information to other family members about their

granny’s sleeping habits, especially if it can be inferred from the sensed data that she

might have got into bed with another patient, which none of the vested parties might

want to share or let the others know about.

While most projects are sensitive to the privacy and ethical problems surrounding

the monitoring of people, they are not easy to solve and have ended up overwhelming

UbiComp research. Indeed, much of the discussion about the human aspects in the

field has been primarily about the trade-offs between security and privacy, conven-

ience and privacy, and informedness and privacy. This focus has often been at the

expense of other human concerns receiving less airing, such as how recording, track-

ing and re-representing movements and other information can be used to facilitate

social and cognitive processes.

Moving on from Weiser’s Vision of Calm Computing 411

My intention here is not to diminish the importance of awareness, ambience and

monitoring to detect and inform people in their everyday lives, together with the ethi-

cal and social issues they raise. Rather, my overview of the projects in these areas has

revealed how difficult it is to build calm computing systems and yet the attempts have

largely dominated the field of UbiComp. Those that have tried have fallen short, re-

sulting in prototype systems that can sometimes appear to be trivial or demeaning.

Conversely, there has been less focus on other areas of research that could prove to be

easier to achieve and potentially of more benefit to society. The time is ripe for other

directions to take center stage in UbiComp. One such avenue promoted here is to

consider how humankind’s evolved practices of science, learning, health, work and

play can be enhanced. This involves thinking about UbiComp not in terms of embed-

ding the environment with all manner of pervasive technologies but instead as

bounded ensembles of entities (e.g., tools, surfaces and lenses) that can be mobile,

collaborative or remote, through which information, other people and the environment

are viewed and interacted with when needed. Importantly, it argues for rethinking the

nature of our relationship with the computer.

3 A New Agenda for UbiComp: Engaging User Experiences

I suggest here that it is highly profitable to recast UbiComp research in the context of

a central motivation that computers were originally designed for, namely, as tools,

devices and systems that can extend and engage people in their activities and pursuits.

My reason for proposing this is based on the success of researchers who have started

to take this approach. In particular, a number of user studies, exploring how UbiComp

technologies are being appropriated, are revealing how the ‘excitement of interaction’

can be brought back in innovative ways; that is not frustrating and which is quite

different from that experienced with desktop applications. For example, various

mixed reality, physical-digital spaces and sensor-rich physical environments have

been developed to enable people to engage and use multiple dynamic representations

in novel ways: in scientific and working practices and in collaborative learning and

experimental games. More extensive inquiries and decisions have been enabled in

situ, e.g., determining the effects of deforestation in different continents and working

out when is the best time to spray or pick grapes in a vineyard.

Recently, world famous computer scientist John Seely Brown put forward his up-

dated vision of UbiComp 1 in a keynote, outlining ‘a common sense’ model that em-

phasizes how UbiComp can help to catalyze creativity [41]. He proposed that creating

and learning be seen as integral to our work and leisure that are formed through re-

creation and appropriation activities. In a similar vein, I argue that it is timely to

switch from a reactive view of people towards a more proactive one. Instead of aug-

menting the environment to reduce the need for humans to think for themselves about

what to do, what to select, etc., and doing it for them, we should consider how Ubi-

Comp technologies can be designed to augment the human intellect so that people can

perform ever greater feats, extending their ability to learn, make decisions, reason,

create, solve complex problems and generate innovative ideas. Weiser’s idea that

1 John Seely Brown was a co-author of the paper written by Weiser on calm technology.

412 Y. Rogers

technologies be designed to be ‘so embedded, so fitting and so natural’ that we use

them without thinking about them needs to be counter-balanced; we should also be

designing them to be exciting, stimulating and even provocative – causing us to re-

flect upon and think about our interactions with them. While Weiser promoted the

advantages of calm computing I advocate the benefits of engaging UbiComp experi-

ences that provoke us to learn, understand and reflect more upon our interactions with

technologies and each other.

A central concern of the engaging UbiComp experiences agenda is to fathom out

how best to represent and present information that is accessible via different surfaces,

devices and tools for the activity at hand. This requires determining how to make

intelligible, usable and useful, the recordings of science, medicine, etc., that are

streaming from an increasing array of sensors placed throughout the world. It also

entails figuring out how to integrate and replay, in meaningful and powerful ways, the

masses of digital recordings that are begin gathered and archived such that profes-

sionals and researchers can perform new forms of computation and problem-solving,

leading to novel insights. In addition, it involves experimenting more with creative

and constructive uses of UbiComp technologies and archived digital material that will

excite and even make people feel uncomfortable.

In terms of who should benefit, it is useful to think of how UbiComp technologies

can be developed not for the Sal’s of the world, but for particular domains that can be

set up and customized by an individual firm or organization, such as for agriculture

production, environmental restoration or retailing. At a smaller scale, it is important to

consider how suitable combinations of sensors, mobile devices, shared displays, and

computational devices can be assembled by non-UbiComp experts (such as scientists,

teachers, doctors) that they can learn, customize and ‘mash’ (i.e., combine together

different components to create a new use). Such toolkits should not need an army of

computer scientists to set up and maintain, rather the inhabitants of ubiquitous worlds

should be able to take an active part in controlling their set up, evolution and destruc-

tion. Their benefits should be clear: enabling quite different forms of information flow

(i.e., ways and means of accessing information) and information management (i.e.,

ways of storing, recording, and re-using information) from older technologies, making

it possible for non-UbiCompers to begin to see how to and subsequently develop their

own systems that can make a difference to their worlds. In so doing, there should be

an emphasis on providing the means by which to augment and extend existing prac-

tices of working, learning and science.

As quoted by Bruner [10] “to assist the development of the powers of the mind is

to provide amplification systems to which human beings, equipped with appropriate

skills, can link themselves” (p.53). To enable this to happen requires a better under-

standing of existing human practices, be it learning, working, communicating, etc.

Part of this reconceptualization should be to examine the interplay between technolo-

gies and their settings in terms of practice and appropriation [15]. “Practices develop

around technologies, and technologies are adapted and incorporated into practices.”

(Dourish, 2001, p. 204). More studies are needed that examine what people do with

their current tools and devices in their surrounding environments. In addition, more

studies are needed of UbiComp technologies being used in situ or the wild – to help

illuminate how people can construct, appropriate and use them [e.g., 16, 22, 23, 29].

Moving on from Weiser’s Vision of Calm Computing 413

With respect to interaction design issues, we need to consider how to represent and

present data and information that will enable people to more extensively compute,

analyze, integrate, inquire and make decisions; how to design appropriate kinds of

interfaces and interaction styles for combinations of devices, displays and tools; and

how to provide transparent systems that people can understand sufficiently to know

how to control and interact with them. We also need to find ways of enabling profes-

sionals and laypeople alike to build, adapt and leverage UbiComp technologies in

ways that extend and map onto their activities and identified needs.

A more engaging and bounded approach to UbiComp is beginning to happen but in

a scattered way. Three of the most promising areas are described below: (i) playful

and learning practices, (ii) scientific practices and (iii) persuasive practices. They

show how UbiComp technologies can be developed to extend or change human ac-

tivities together with the pertinent issues that need to be addressed. Quite different

practices are covered, reflecting how the scope of UbiComp can be broad but at the

same time targeted at specific users and uses.

3.1 Playful and Learning Practices

One promising approach is to develop small-scale toolkits and sandboxes, comprising

interlinked tools, digital representations and physical artifacts that offer the means by

which to facilitate creative authoring, designing, learning, thinking and playing. By a

sandbox it is not meant the various senses it has been used in computing but more

literally as a physical-digital place, kitted out with objects and tangibles to play and

interact with. Importantly, these should allow different groups of people to participate

in novel activities that will provoke and extend existing repertoires of technology-

augmented learning, playing, improvising and creating. An example of a promising

UbiComp technology toolkit is PicoCrickets, developed at MIT Media Lab, arising

from the work of Mitch Resnick and his colleagues. The toolkit comprises sensors,

motors, lights, microcomputers, and other physical and electrical devices that can be

easily programmed and assembled to make them react, interact and communicate,

enabling “musical sculptures, interactive jewelry, dancing creatures and other playful

inventions” to be created by children and adults alike. An advantage of such light-

weight, off-the-shelf tangible toolkits is that they offer many opportunities for differ-

ent user groups (e.g., educators, consultants) to assemble and appropriate in a range of

settings, such as schools, waiting rooms, playgrounds, national parks, and museums.

A nagging question, however, is how do the benefits of such UbiComp toolkits and

sand boxes compare with those offered by more conventional ones – that are much

cheaper and more practical to make? Is it not the case that children can be highly

creative and imaginative when given simply a cardboard box to play with? If so, why

go to such lengths to provide them with new tools? The debate is redolent of whether

it is better for children to read a book or watch a 3D Imax movie. One is not necessar-

ily better than the other: the two provide quite different experiences, triggering differ-

ent forms of imagination, enjoyment and reflection. Likewise, UbiComp and physical

toys can both provoke and stimulate, but promote different kinds of learning and

collaboration among children. However, a benefit of UbiComp toolkits over physical

artifacts is that they offer new opportunities to combine physical interaction, through

manipulation of objects or tools or through physical body postural movement and

414 Y. Rogers

location, with new ways of interacting, through digital technology. In particular, they

provide different ways of thinking about the world than interacting solely with digital

representations or solely with the physical world. In turn, this can encourage or even

enhance further exploration, discovery, reflection and collaboration [35].

Examples of projects that have pioneered the design of novel physical-digital

spaces to facilitate creativity and reflection include the Hunting of the Snark [32],

Ambient Wood [36], RoomQuake [33] Savannah [17], Environmental Detectives

[27], Drift Table [19] and Feeding Yoshi [7]. Each of these have experimented with

the use of mobile, sensor and fixed technologies in combination with wireless infra-

structures to encourage exploration, invention, and out of the box thinking.

The Hunting of the Snark adventure game provoked young children into observing,

wondering, understanding, and integrating their fragmented experiences of novel

physical-digital spaces that subsequently they reflected upon and shared as a narrative

with each other. A combination of sensor-based, tangible, handheld and wireless

technologies was used to create the physical-digital spaces, where an imaginary vir-

tual creature was purported to be roaming around in. The children had to work out

how to entice the creature to appear in them and then gather evidence about its per-

sonality, moods, etc, by walking with it, feeding it and flying with it. Similarly, Sa-

vannah was designed as a physical-digital game to encourage the development of

children’s conceptual understanding of animal behavior and interactions in an imagi-

nary virtual world. The project used GPS and handheld computers to digitally overlay

a school playing field with a virtual plain. Children took on the roles of lions, had to

hunt animals in the virtual savannah and capture them to maintain energy levels. After

the game, the children reflected on their experiences by interacting with a visualiza-

tion on a large interactive whiteboard, that showed the trails they made in the Savan-

nah and the sounds and images that they encountered at specific place.

The Ambient Wood project used an assortment of UbiComp technologies to en-

courage more self-initiation in inquiry and reflective learning. Various wireless and

sensor technologies, devices and representational media were combined, designed and

choreographed to appear and be used in an ‘ambient’ woodland. Several handcrafted

listening, recording and viewing devices were created to present certain kinds of digi-

tal augmentations, such as sounds of biological processes, images of organisms, and

video clips of life cycles. Some of these were triggered by the children’s exploratory

movements, others were collected by the children, while still others were aggregated

and represented as composite information visualizations of their exploratory behavior.

RoomQuake was designed to encourage children to practice scientific investigatory

practices: an earthquake was simulated in a classroom using a combination of inter-

connected ambient media, string and physical styrofoam balls. The ambient media

provided dynamic readings of the simulated earthquakes, which students then re-

represented as physical models using the physical artifacts. The combination of com-

puter-based simulations and physical-based artifacts enabled the whole class to take

part in the measuring, modeling, interpreting, sparking much debate and reflection

among the children about the seismic events.

As part of the Equator collaboration, a number of innovative ‘seamful games’ have

been developed. The inherent limitations of ubiquitous technologies have been delib-

erately exploited to provoke the players into thinking about and acting upon their

significance to the ongoing activity. Two examples are Treasure in which players had

Moving on from Weiser’s Vision of Calm Computing 415

to move in and out of a wireless network connectivity to collect and then deposit gold

tokens and Feeding Yoshi where the players were required to feed virtual creatures

scattered around a city with virtual fruits that popped up on their displays as a result

of their location and activity therein.

Evaluations of this emerging genre of physical-digital spaces for learning and play-

ing have been positive, highlighting enhanced understanding and an immense sense of

engagement. Children and adults have been able to step back and think about what

they are doing when taking part in the game or learning experience, examining the

rationale behind their choices when acting out and interacting with the UbiComp-

based technologies in the space. However, many of the pioneering projects were tech-

nology, resource and researcher intensive. While guidance is now beginning to appear

to help those wanting to design UbiComp-based learning and playing experiences

[e.g., 9, 36] we need also to strive towards creating the next generation of physical-

digital spaces and toolkits that will be as easy, cheap and popular to construct as Lego

kits once were.

3.2 Scientific Practices

Another area where UbiComp has great potential for augmenting human activities is

the practice of scientific inquiry and research. Currently, the sciences are going

through a major transformation in terms of how they are studied and the computa-

tional tools that are used and needed. Microsoft’s 2020 Science report – a comprehen-

sive vision of science for the next 14 years written by a group of internationally

distinguished scientists – outlines this paradigm shift [31]. It points out how new

conceptual and technological tools are needed that scientists from different fields can

“understand and learn from each other’s solutions, and ultimately for scientists to

acquire a set of widely applicable complex problem solving capabilities”. These in-

clude new programming, computational, analysis and publication tools. There is much

scope, too, for utilizing UbiComp technologies to enhance computation thinking,

through integrating sensor-based instrumentation in the medical, environmental and

chemical sciences. The ability to deliver multiple streams of dynamic data to scien-

tists, however, needs to be matched by powerful interfaces that allow them to manipu-

late and share them in new ways, from any location whether in the lab or in the field.

Areas where there is likely to be obvious benefits to scientists through the integra-

tion of UbiComp and computational tools are environmental science and climate

change. These involve collaborative visualization of scientific data, mobile access to

data and capture of data from sensors deployed in the physical world. Being able to

gain a bigger, better and more accurate picture of the environmental processes may

help scientists make more accurate predictions and anticipate more effectively natural

disasters, such as tsunamis, volcanoes, earthquakes and flooding. However, it may not

simply be a case of more is more. New ways of managing the burgeoning datasets

needs to be developed, that can be largely automated, but which also allows scientists

to have effective windows, lenses etc., into so that they can interpret and make intelli-

gible inferences from them at relevant times.

The 2020 report notes how tomorrow’s scientists will need to make sense of the

masses of data by becoming more computationally literate – in the sense of knowing

how to make inferences from the emerging patterns and anomalies that the new

416 Y. Rogers

generation of software analysis tools provide. To this end, a quite different mindset is

needed in schools for how science is taught. The design of new learning experiences

that utilize UbiComp technologies, both indoors and outdoors, need to be developed

to seed in young children the sense of what is involved in practicing new forms of

complex, computational science. An example of how this can be achieved is the em-

bedded phenomena approach; scientific phenomena are simulated using UbiComp

technologies, for long periods of time, to create opportunities for groups of students to

explore ‘patient’ science [32]. Essentially, this involves the accumulation, analysis

and representation of data collected from multiple computational devices over ex-

tended periods of observation in the classroom or other sites. In so doing, it allows

students to engage in the collaborative practice of scientific investigation that requires

hard computational thinking but which is also exciting, creative and authentic. A core

challenge, therefore, is to find ways of designing novel science learning experiences

that capitalize on the benefits of combining UbiComp and PC technologies that can be

used over extended periods.

3.3 Persuasive Practices

The third area where there is much potential for using UbiComp technologies to en-

gage people is as part of self-monitoring and behavioral change programs. While a

range of persuasive technologies (e.g., adverts, websites, posters) has already been

developed to change people’s attitudes and behaviors, based on models of social

learning [18], UbiComp technologies provide opportunities for new techniques. Spe-

cifically, mobile devices, such as PDAs coupled with on-body sensors, can be de-

signed to enable people to take control and change their habits or lifestyles to be

healthier by taking account of and acting upon dynamically updated information pro-

vided by them. For example, Intille and his group are exploring how mobile computa-

tional tools for assessing behavioral change, based on social psychology models, can

be developed to motivate physical activity and healthy eating.

A key question that needs to be addressed is whether UbiComp technologies are

more (or less) effective compared with other technologies in changing behavior. A

diversity of media-based techniques (e.g., pop-up warning messages, reminders,

prompts, personalized messages) has been previously used to draw people’s attention

to certain kinds of information to change what they do or think at a given point. In

terms of helping people give up habits (e.g., smoking, excessive eating) they have had

mixed results since people often relapse. It is in the long-term context that UbiComp

technologies may prove to be most effective, being able to monitor certain aspects of

people’s behavior and represent this information at critically weak moments in a ca-

joling way. A constant but gentle ‘nagging’ mechanism may also be effective at per-

suading people to do something they might not have otherwise done or to not to do

something they are tempted to do. For example, a collaborative cell phone application

integrated with a pedometer was used to encourage cliques of teenage girls to monitor

their levels of exercise and learn more about nutrition in the context of their everyday

activities [44]. The software was designed to present the monitored process (e.g.,

walking) in a way that made it easy for the girls to compute and make inferences of

how well they were doing in terms of the number of steps taken relative to each other.

A preliminary study showed that such a collaborative self-monitoring system was

Moving on from Weiser’s Vision of Calm Computing 417

effective at increasing the girl’s awareness of their diet, level of exercise and enabling

them to understand the computations involved in burning food during different kinds

of exercise. But most significantly, it enabled the girls to share and discuss this infor-

mation with each other in their private clique, capitalizing on both the persuasive

technology and peer pressure.

Incorporating fun into the interface can also be an effective strategy; for example,

Nintendo’s Pocket Pikachu with pedometer attached was designed to motivate chil-

dren into being more physically active on a consistent basis. The owner of the digital

pet that ‘lives’ in the device is required to walk, run or jump each day to keep it alive.

If the owner does not exercise for a week the virtual pet becomes unhappy and even-

tually dies. This can be a powerful means of persuasion given that children often

become emotionally attached to their virtual pets, especially when they start to care

for them.

UbiComp technologies can also be used to reduce bad habits through explicitly

providing dynamic information that someone would not have been aware of other-

wise. In so doing, it can make them actively think about their behavior and modify it

accordingly. The WaterBot system was developed using a special monitoring and

feedback device to reduce householder’s usage of water in their homes – based on the

premise that many people are simply unaware of how wasteful they are [3]. A sensor-

based system was developed that provided positive auditory messages and chimes

when the tap was turned off. A central idea was to encourage members of the house-

hold to talk to one another about their relative levels of water usage provided by the

display and to try to out do one another in the amount of water used.

But to what extent do UbiComp technologies, designed for persuasive uses, differ

from the other forms of monitoring that were critiqued earlier in the paper? A main

difference is that there is more active involvement of those being monitored in attain-

ing their desired behavior change compared with those who were being monitored

and assisted in care homes. The objective is to enable people, themselves, to engage

with the collected information, by monitoring, understanding, interpreting and acting

upon it – and not the environment or others to act upon their behalf. Much of the

research to date in UbiComp and healthcare has focussed on automated bio-

monitoring of physiological processes, such as EEGs and heart rate, which others, i.e.,

specialists, examine and use to monitor their patient’s health. In contrast, persuasive

technologies are intended to provide dynamic information about a behavioral process

that will encourage people from doing or not doing something, by being alerted

and/or made aware of the consequences of what they are about to do. Moreover, de-

signing a device to be solely in the control of the users (and their social group) en-

ables them to be the owners of the collected data. This circumvents the need to be

centrally concerned with privacy issues, allowing the focus of the research to be more

oriented towards considering how best to design dynamically updated information to

support cognitive and social change. A challenge, however, in this area is for long

term studies to be conducted that can convincingly show that it is the perpetual and

time-sensitive nature of the sensed data and the type of feedback provided that con-

tributes to behavioral modification.

418 Y. Rogers

4 Conclusions

Many of the research projects that have followed in the footsteps of Weiser’s vision

of calm computing have been disappointing; their achievements being limited by the

extent to which they have been able to program computers to act on behalf of humans.

Just as ‘strong’ AI failed to achieve its goals – where it was assumed that “the com-

puter is not merely a tool in the study of the mind; rather, the appropriately

programmed computer really is a mind” [41], it appears that ‘strong’ UbiComp is

suffering from the same fate. And just as ‘weak’ AI 2 revived AI’s fortunes, so, too,

can ‘weak’ UbiComp bring success to the field. This will involve pursuing more prac-

tical goals and addressing less ambitious challenges; where ensembles of technologies

are designed for specific activities to be used by people in bounded locations. To

make this happen, however, requires moving from a mindset that wants to make the

environment smart and proactive to one that enables people, themselves, to be smarter

and proactive in their everyday and working practices. Three areas of research were

suggested as to how this could be achieved; but, equally, there are others where there

is much potential for enhancing and extending human activities (e.g., vineyard com-

puting [11], firefighting [24] and sports). As part of the expansion of UbiComp, a

wider range of human aspects should be considered, drawing upon alternative theory,

guiding frameworks and metaphors [c.f. 8, 15]. To enable other human concerns to

become more prominent, however, requires the hefty weight of privacy and other

related ethical issues on UbiComp’s shoulders to be lessoned.

The ‘excitement of interaction’ that Weiser suggested forsaking in the pursuit of a

vision of calm living should be embraced again, enabling users, designers and re-

searchers to participate in the creation of a new generation of user experiences that go

beyond what is currently possible with our existing bricolage of tools and media. We

should be provoking people in their scientific, learning, analytic, creative, playing and

personal activities and pursuit. Finally, while we have been privileged to have had such

a great visionary, whose legacy has done so much to help shape the field, it is timely

for a new set of ideas, challenges and goals to come to the fore and open up the field.

Acknowledgements

Thanks to Tom Rodden for his suggestions on an earlier draft and the anonymous

reviewers for their constructive comments.

References

1. Abowd, G.D., Mynatt. E.D.: Charting past, present, and future research in ubiquitous com-

puting. ACM Transactions on Computer-Human Interaction, 7 (2000) 29-58

2. Anderson, K., Dourish, P.: Situated Privacies: Do you know where you mother [trucker]

is? In Proceedings of the 11th International Conference on Human-Computer Interaction.

Las Vegas. July 22-27, 2005

2 Weak AI refers to the development of software programs to perform specific problem-solving

or reasoning tasks that do not have to match the way humans do them.

Moving on from Weiser’s Vision of Calm Computing 419

3. Arroyo, E., Bonnanni, L., Selker, T.: WaterBot: exploring feedback and persuasive tech-

niques at the sink. In CHI Proceedings, ACM, New York, 631-639, 2005

4. Ashley, S.: Cognitive Radio, Scientific American, (March 2006), 67-73

5. Augmented Cognition International Society. http://www.augmentedcognition.org/, Re-

trieved on 30/03/2006

6. Beckwith, R., Lederer, S.: Designing for one’s dotage: UbiComp and Residential Care fa-

cilities. Conference on the Networked Home and the Home of the Future (HOIT 2003), Ir-

vine, CA: April 2003

7. Bell, M., Chalmers, M., Barkhuus, L., Hall, M., Sherwood, S., Tennent, P., Brown. B.,

Rowland, D., Benford, S., Hampshire, A., Captra, M.: Interweaving mobile games with

everyday life. In Proceedings of CHI’06, Conference on Human Factors in Computing.

ACM Press, (2006) 417-426

8. Bellotti, V., Back, M., Edwards, K., Grinter, R., Henderson. A., Lopes, C.: Making sense

of sensing systems: five questions for designers and researchers. In Proceedings of

CHI’2002, ACM Press, (2002) 415-422

9. Benford, S., Schnädelbach, H., Koleva, B., Anastasi, R., Greenhalgh, C., Rodden, T.,

Green, J., Ghali, A., Pridmore, T., Gaver, B., Boucher, A., Walker, B., Pennington, S.,

Schmidt, A., Gellersen, H., Steed, A.: Expected, sensed, and desired: A framework for de-

signing sensing-based interaction. ACM Trans. Comput.-Hum. Interact. 12 (2005) 3-30

10. Bruner, J.S. The Relevance of Education. Harmondsworth, Middlesex, UK. (1972)

11. Burrell, J., Brooke, T., Beckwith, R.: Vineyard Computing: Sensor Networks in agricul-

tural production, Pervasive Computing, 3(1) (2004) 38-45

12. Chalmers, D., Chalmers, M., Crowcroft, J., Kwiatkowska, M., Milner, R., O’Neill, E.,

Rodden, T., Sassone, V., Sloman, M.: Ubiquitous Computing: Experience, design and sci-

ence. Version 4. http://www-dse.doc.ic.ac.uk/Projects/UbiNet/GC/index.html Retrieved on

30/03/2006

13. Consolvo, S., Roessler, P., Shelton, B., LaMarca, A., Schilit, B., Bly, S.: Technology for

care networks for elders. Pervasive Computing 3 (2004) 22-29

14. Digiens@U-City.: Korea moves into ubiquitous mode.

http://digiens.blogspot.com/2005/08/korea-moves-into-ubiquitous-mode.html. Retrieved

30/03/2006

15. Dourish, P.: Where the action is: the foundation of embodied interaction. MIT, Cambridge,

MA., (2001)

16. Dourish, P., Grinter, B., Delgado de la Flor, J., Joseph, M.: Security in the wild: user

strategies for managing security as an everyday, practical problem. Personal and Ubiqui-

tous Computing, 8 (6) (2004) 391-401

17. Facer, K., Joiner, R., Stanton, D., Reid, J., Hull, R., Kirk, D.: Savannah: mobile gaming

and learning. Journal of Computer Assisted Learning, 20 (2004) 399-409

18. Fogg, B.J.: Persuasive Technology: Using Computers to change what we think and do.

Morgan Kaufmann Publishers, San Fransisco. (2003)

19. Gaver, W. W., Bowers, J., Boucher, A., Gellersen, H., Pennington, S., Schmidt, A., Steed,

A., Villars, N., Walker, B.: The drift table: designing for ludic engagement. In Proceedings

of CHI Extended Abstracts (2004) 885-900.

20. Greenfield, A.: Everyware: The Dawning Age of Ubiquitous Computing. New Riders,

Berkeley, CA. (2006)

21. Intel Research at Intel: Research Seattle.

www.intel.com/research/network/seattle\_collab.htm. Retrieved on 20/03/2006.

420 Y. Rogers

22. Intille, S., Larson, K., Beaudin, J., Nawyn, J., Munguia Tapia, E., Kaushik, P.: A living

laboratory for the design and evaluation of ubiquitous computing technologies. In Proceed-

ings of CHI Extended Abstracts (2005) 1941-1944

23. Intille, S.S., Bao, L., Munguia Tapia, E., Rondoni, J.: Acquiring in situ training data for

context-aware ubiquitous computing applications. In Proceedings CHI (2004) 1-8

24. Jiang, X., Chen, N.Y., Hong, J.I., Wang, K., Takayama, L.A., Landay, J.A.: Siren: Con-

text-aware Computing for Firefighting. In Proceedings of Second International Conference

on Pervasive Computing. Lecture Notes in Computer Science, Springer Berlin Heidelberg

87-105 (2004)

25. Journal of Ubiquitous Computing and Intelligence. www.aspbs.com/juci.html Retrieved

20/03/2006/

26. Kindberg, T., Fox, A.: System Software for Ubiquitous Computing. IEEE Pervasive Com-

puting, 1 (1) (2002) 70-81

27. Klopfer, E., K. Squire.: Environmental Detectives – The Development of an Augmented

Reality Platform for Environmental Simulations. Educational Technology Research and

Development. (2005)

28. Krikke, J.: T-Engine: Japan’s Ubiquitous Computing Architecture is ready for prime time.

Pervasive Computing (2005) 4-9

29. LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., Sohn, T.,

Howard, J., Hughes, J., Potter, F., Tabert, J., Powledge, P., Borriello, G., Schilit. B.: Place

Lab: Device Positioning Using Radio Beacons in the Wild, Intel Research, IRS-TR-04-

016, (2004) http://placelab.org/publications/pubs/IRS-TR-04-016.pdf

30. Marmasse, N., Schmandt, C.: Location-aware information delivery with commotion, In

HUC 2000 Proceedings, Springer-Verlag, (2000) 157-171

31. Microsoft 2020 Science.: http://research.microsoft.com/towards2020science/. Retrieved

30/03/2006

32. Moher, T.: Embedded Phenomena: Supporting science learning with classroom-sized-

distribution simulations. In Proceedings of CHI 2006

33. Moher, T., Hussain, S., Halter, T., Kilb, D..: RoomQuake: embedding dynamic phenomena

within the physical space of an elementary school classroom. Extended Abstracts, In Pro-

ceedings of CHI’05, Conference on Human Factors in Computing Systems. ACM Press

(2005) 1655-1668

34. Mynatt, E., Melenhorst, A., Fisk, A.D., Rogers, W.: Aware technologies for aging in place:

Understanding user needs and attitudes. Pervasive Computing (2004) 36-41

35. Price, S. Rogers, Y. Let’s get physical: the learning benefits of interacting in digitally

augmented physical spaces. Journal of Computers and Education, 43 (2004) 137-151

36. Rogers, Y., Muller, H.: A framework for designing sensor-based interactions to promote

exploration and reflection. International Journal of Human-Computer Studies, 64 (1)

(2005) 1-15

37. Rogers, Y., Price, S., Fitzpatrick, G., Fleck, R., Harris, E., Smith, H., Randell, C., Muller,

H., O’Malley, C., Stanton, D., Thompson, M., Weal, M.: Ambient Wood: Designing new

forms of digital augmentation for learning outdoors. In Proceedings of Interaction Design

and Children, ACM (2004) 1-8

38. Rogers, Y., Scaife, M., Harris, E., Phelps, T., Price, S., Smith, H., Muller, H., Randall, C.,

Moss, A., Taylor, I., Stanton, D., O’Malley, C., Corke, G., Gabrielli, S.: Things aren’t

what they seem to be: innovation through technology inspiration. In Proceedings of

DIS’2002 Designing Interactive Systems, ACM Press, (2002) 373-379

39. Ross, D.A.: Cyber Crumbs for successful aging with vision loss. Pervasive Computing, 3

(2004) 30-35

Moving on from Weiser’s Vision of Calm Computing 421

40. Salvador, T., Anderson, K. Practical Considerations of Context for Context Based Sys-

tems: An Example from an Ethnographic Case Study of a Man Diagnosed with Early On-

set Alzheimer’s Disease. In UbiComp’03 Proceedings, A.K. Dey et al. (Eds.), LNCS 2864,

Springer-Verlag Berlin Heidelberg, 243-255, 2003

41. Seely Brown, J.: Ubiquitous Computing and beyond – an emerging new common sense

model. www.johnseelybrown.com/JSB.pdf. Retrieved 20/03/2006

42. Stirling, B.: Without Vision, the People Perish. Speech Given at CRA Conference on

Grand Research Challenges in Computer Science and Engineering. Airlie House, Warren-

ton, Virginia, June 23, 2002 www.cra.org/Activities/grand.challenges/sterling.html Re-

trieved 20/03/2006

43. Tennenhouse, D.L. “Proactive Computing,” Communications of the ACM 43, No. 5, 43–

50, 2000

44. Toscos, T., Faber, A., An, S., Gandhi, M.; Chick Clique: Persuasive Technology to Moti-

vate Teenage Girls to Exercise. In CHI’06 Extended Abstracts on Human Factor in Com-

puting Systems, ACM Press (2006) 1873-1878

45. Tran, Q., Calcaterra, G., Mynatt, E.: Cook's Collage: Deja Vu Display for a Home Kitchen.

In Proceedings of HOIT 2005, 15-32

46. Weiser, M., Brown, J.S.: The coming age of calm technology. (1996)

www.ubiq.com/hypertext/weiser/acmfuture2endnote.htm. Retrieved 20/03/2006/

47. Weiser, M.: The computer for the 21st century. Scientific American (1991) 94–104

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-10

The Lucent Web site is built hierarchically, in the sense that

pages deeper in the directory tree represent more detailed infor-

mation than those at shallower levels. At its busiest, there can

be as many as 300 people browsing www.lucent.com; while

during the pre-dawn hours there can be as few as 5 simultaneous

visitors. Our sonification is designed to convey qualitative infor-

mation about site usage, answering questions like:

Overall, is the site busy or quiet?

What proportion of the visitors are delving for specific in-

formation deep within the site, as compared to those visitors

who are “just passing through,” glancing briefly at the home

page and then moving on?

How are users distributed across the various content areas

of the site?

Which portions of the site are visited together? What kinds

of patterns do we find in user behavior?

We think of this sonification as one possible “background” infor-

mation stream that can inform content providers, Web designers

and even the visitors themselves.

2.1.1. Sonification design

Our audio display makes use of the hierarchical structure of the

content offered by www.lucent.com. First, a unique pitch was

used to identify each of five high-level subdomains within the site:

/micro, representing Lucent’s microelectronics design and man-

ufacturing business (now Agere Systems); /enterprise, for

the enterprise systems and software business (now Avaya Com-

munications); /minds, a corporate introduction to Bell Labs re-

search; /press, a collection of press releases and investor infor-

mation; and /search, the local search engine for the site.

The total number of visitors accessing any information from a

subdomain affects the loudness and tonal balance of a low-register

drone at the associated pitch. Visitors requesting content deeper in

the site are represented by higher-pitched pulsing tones (separated

by one or two octaves from the base pitch for the subdomain):

the faster the pulse, the more people are accessing that area, and

the greater the proportion of high-register sounds, the more de-

tailed the content. By assigning well-separated pitches to each

subdomain, shifts in activity both within and between the areas

can be heard. In Table 1 we present a simple mapping of data col-

lected by the Lucent Web server to a continuously time-varying

vector of usage statistics. In the category of Overall browsing, we

count any visitor accessing content pages (HTML, PostScript or

PDF) from the indicated subdomain. A Mid-Level access is a re-

quest for content two or more directories down. Simple examples

are /micro/K56flex/index.html (information on a brand

of 56K modem) and /press/0101/010118.nsb.html (a

press release for January 18, 2001). The final category, Deep

browsing, refers to pages that are four or more directories down

in the tree. One example is a paper from the April/June 2000

issue of the Bell Labs Technical Journal, located at /minds/

techjournal/apr-jun2000/pdf/paper02.pdf.

Then, the resulting 15 values in Table 1, A1–E3, were mapped

to sound as follows:

Overall activity Measured by A1–E1, voiced with a low-register

drone. The aggregate number of visitors accessing infor-

mation within each of the five areas modulates the loudness

of each of the five pitches.

/micro /enterprise /minds /press /search

Overall A1 B1 C1 D1 E1

Mid-Level A2 B2 C2 D2 E2

Deep A3 B3 C3 D3 E3

Table 1: Mapping used for Web site traffic example. Overall ac-

tivity records the movements of all users; Mid-Level counts users

2 or 3 directories into the site; Deep browsing consists of users 4+

directories down.

Mid-Level browsing Measured by A2–E2 and assigned a rhyth-

mic middle-register tone pulse; pulse loudness and repeti-

tion speed rises and the timbral brightness increases as the

volume of mid-level browsing increases. There are five in-

dependent pulses, each at a different fixed pitch, represent-

ing the five content areas.

Deep browsing Measured by A3–E3 and made audible via rhyth-

mic high-register “ting” sounds (plucked steel string sam-

ples). Loudness and repetition speed rises as the volume of

deep browsing increases. Again, there are five independent

“ting” sounds, each at a different fixed pitch, representing

the five content areas.

We used pitch groups that were consonant, and for the sounds that

incorporated rhythm (A2–E3), the phase and frequency of each

pulse in the matrix varies independently, yielding a sound with a

changing rhythmic texture but no fixed beat.

The purpose of this sonification is to make interpretable the

activities of users on a Web site. Therefore, the stream of hits be-

ing processed by a Web server (reduced to include only the HTML,

PostScript and PDF documents) needs to be transformed to extract

meaningful user-level data. A real-time monitoring tool was devel-

oped that maintains a bank of active visits (recording separately the

activities of all the people browsing the site at a given time) and

updates various statistics with each user request. When cookies or

some other authentication mechanism allows us to recognize re-

turning visitors, the monitor will update a more complicated user

profile that encapsulates previous browsing patterns. Our traffic

sonification as described above takes as input the location of each

visitor within a site at a given point in time. When constructing

more elaborate sound displays, our design will continue to focus

on user activities, drawing more heavily on the statistics culled

by the monitoring tool. This emphasis distinguishes our approach

from sonification methods that assess Web server performance by

making audible statistics relating to server load, HTTP errors, and

agent types [?].

2.1.2. Impressions and extensions

We have created three audio examples for the activity on the Lu-

cent site. Our data were captured on November 11, 1999 and we

created sonifications of the traffic at 6:00 am, an extremely slow

period for the site; noon, a relatively active time; and 2:30 pm,

the point at which the site was busiest. The samples are located at

our project Web site [6]. Even with this relatively straightforward

mapping, one finds compelling patterns. For example, the affinity

between the /enterprise subdomain and the /search facil-

ity can be heard as the pulses for these areas rise and fall together.3

3 While clearly audible, these shifts can really only be precisely associ-

ated with areas after a certain amount of experience with the mapping.

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-11

Also, when comparing moderately active to extremely busy peri-

ods, we find that the number of people digging deep into the site is

not a fixed fraction of the total number of visitors. That is, the vol-

ume of the low-register drones exhibits much more variation than

the components for the other two categories of accesses. Each of

these effects can be verified by examining the logs, reinforcing the

usefulness of our sonification as a tool for constructing hypotheses

about site traffic.

As mentioned at the beginning of this section, Web browsers

offer a rich set of data about the visitor when requesting data from

a server. This display makes use of only the most basic informa-

tion about a visit, namely the depth of pages accessed. In ongoing

work, we are augmenting our sonification with extra features de-

rived both directly from the server data as well as from statistical

navigation models [12] fit for the Web site under study. So far, we

have found that such extensions are most effective when developed

in the context of a particular monitoring application. For example,

an extended version of this ambient display can aid system archi-

tects of large, Web hosting services understand cache performance

and can aid in server provisioning. Another extension will make

greater user of our navigation models and can help designers and

usability engineers better architect Web sites. We will report on

these and other developments through the project Web site [4].

2.2. Chat rooms and bulletin boards

At any given moment, tens of thousands of real-time conversa-

tions are taking place across the Internet on public forums, bulletin

boards and chat sites. To imagine making these conversations si-

multaneously audible evokes an image of uproarious babble. And

yet, in the aggregate, this massive stream of live communication

could exhibit rich thematic structure. Can we find a meaningful

way to listen in to so many conversations, rendering them in a way

that is comprehensible and not overwhelming?

In some sense, a byproduct of our Web traffic sonification is

the creation of a kind of community from the informal gather-

ing of thousands of visitors to a given Web site. Traditionally,

informational Web sites like www.lucent.com have provided

us with very little sense of the other people who are requesting

data from the server. To attract and retain visitors, however, many

commercial sites recognize the potential of the Web to form so-

cial as well as informational networks. As a result, Web-based fo-

rums, message boards and a variety of chat services are common

components of current site designs. While Internet Relay Chat

(IRC) has been a widely used standard since the inception of the

Internet, the popularization of the Web has resulted in a virtual

explosion of chat applications.4 For example, www.yahoo.com

(a US-based Web portal) offers hundreds of separate chat rooms

attracting tens of thousands of visitors a day. Specialized sites

like www.style.com (the homepage for Vogue magazine) or

www.audiworld.com (an resource for Audi owners) have also

found their message boards to be the most frequently accessed

parts of their domains.

To get a sense of the amount of content that is available in

these dynamic formats, we examined sites contained in the DMOZ

Open Directory [3], an open source listing of over 2 million Web

sites compiled and categorized by 33,000 volunteer editors. From

the November 20, 2000 image of the directory, we counted 36,681

4 RC was developed by Jarkko Oikarinen in Finland in the late eighties,

and was originally intended to work as a better substitute for talk on his

bulletin board.

separate sites offering some kind of chat, bulletin board or other

public forum. While we did not examine the activity on all of

these sites, the number is staggering. If we include other peer-

to-peer communication technologies like instant messaging,5 the

amount of dialogue taking place on the Web at any point in time

is almost unfathomable. The goal of our second sonification is

to make interpretable the thousands of streams of dynamic infor-

mation being generated on the Web. In so doing, we attempt to

characterize a global dialogue, integrating political debates, dis-

cussions of current events, and casual exchanges between mem-

bers of virtual communities.

2.2.1. Content monitors and the statistics engine

Our starting point is text. Albeit diverse in style and dynamic in

character, the text (or transcript) of these data sources carries their

meaning. Therefore, any auditory display consisting only of gen-

erated tones would not be able to adequately represent the data

without a very complex codebook. The design of our sonifica-

tion then depends heavily on text-to-speech (TTS). As with the

traffic example in the previous section, we think of the audio out-

put as another background information stream. The incorporation

of spoken components in the sound design poses new challenges,

both practical and aesthetic. For example, simply voicing every

word taking place in a single chat room can produce too much text

to be intelligible when played in real-time and can quickly exhaust

the listener. Instead, we build a hierarchical representation of the

text streams that relies on statistical processing for content organi-

zation and summarization prior to display.

Before considering sonification design, we first had to cre-

ate specialized software agents that would both discover new chat

rooms and message boards, as well as harvest the content posted

to these sites. (See Figure 1 for an overview of our system ar-

chitecture.) Most bulletin boards and some chat applications use

standard HTML to store visitor contributions. In many cases, a

specific login name is required to gain access to the site. For

these situations, we constructed a content agent in Perl, as this

language provides us the most convenient platform for managing

access details (like cookies). The public chat rooms on sites like

chat.yahoo.com can be monitored in this way. For IRC we

built a configurable Java client that polls a particular server for

active channels. Web sites like www.cnn.com (a popular news

portal) and www.financialchat.com (a financial commu-

nity hosting chat services for day traders) offer several IRC rooms,

some of which are tightly moderated.

In addition to collecting content, each monitoring agent also

summarizes the chat stream, identifying basic topics and updating

statistics about the characteristics of the discussion: What percent-

age of visitors are contributing? How often to they contribute and

at what length? Is the room “on topic,” or are many visitors post-

ing comments on very different subjects? Topics are derived from

the chat stream using a variant of generalized sequence mining [7]

that incorporates tags for the different parts of speech. While the

exact details are beyond the scope of this abstract, a generalized se-

quence is a string of words possibly separated by a wildcard, “\*”.

For example, if we let A, B and C denote specific “contentful”

words (say, nouns, adjectives and adverbs), then AB C , A B C

and A B C are all generalized sequences. The wildcard al-

lows us to identify “Gore \* disputes \* election” from the sentences

5 AOL alone records tens of millions of people using their instant mes-

saging service each month.

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-12

Chat

BB

Chat Chat

BB

Sonification

Engine

Stats

Channel

Audio Right

Channel

Audio Left

Engine

Statistics

Text Feedback

Content

Monitor

Content

Monitor

Content

Monitor

Content

Monitor

Content

Monitor

Figure 1: System architecture overview. A large number of content

streams (Chat = chat rooms; BB = Bulletin boards) are gathered by

specialized agents that transmit them in a homogenized format to

the statistics engine. The statistics engine then distills the streams

into a much smaller number of configurable text streams as well

as a number of descriptive vectors. The sonification engine then

“plays” these text and data streams. The entire systems operates in

real-time.

“Vice President Gore filed papers to dispute the presidential elec-

tion,” “Aides for Gore indicated that he has every reason to dispute

the election”, and “Gore is still deciding whether or not to dispute

the election”.

As many posts to chat rooms contain spelling mistakes and

incorrect grammar, assigning words to different parts of speech is

error-prone. However, unlike most applications of statistical nat-

ural language processing, our content monitors update their sum-

maries each time new material is posted and downweight older

contributions. Because our sonification renders these sources in

real-time, small mistakes have little effect on the power of the over-

all display to convey the ideas being discussed.

Each of the content monitors are periodically polled by the

statistics engine (see Figure 1). This Java-application clusters the

different chat rooms and bulletin boards based on their topic and

numerical summaries. As the topic in a room changes over time,

the statistics engine is constantly updating and reformulating clus-

ter membership. Because a content stream can in fact support

a number of simultaneous discussions (the threads of a bulletin

board, say), we employ a soft-clustering technique. In our initial

work, we have used a mixture-based scheme that determines the

number of clusters with an MDL (Minimum Description Length)

criterion [9]. Each room is then assigned a probability that it be-

longs to the different groups. This model also provides for topic

summarization at the cluster-level. Next, a stochastic framework

was developed to sample representative sentences posted to the

chat or bulletin board. When a discussion is extremely unstruc-

tured, this selection is essentially random sampling from all the

contributions added to the chat since the last polling point. In ad-

dition to textual data streams, the statistics engine is also respon-

sible for communicating the various ingredients for the display to

our sonification engine, Max/MSP [2] (see Figure 1). We have

adopted the Open Sound Control [13] protocol from Center for

New Music and Audio Technologies to transfer data between the

statistics engine (running on a Macintosh with LinuxPPC) and the

sonification engine (running on a Macintosh with OS/9).

2.2.2. Sonification design

As with the previous example (Section 2), our goal is to create

a sonification that is both communicative and listenable. Here we

face the additional challenge of incorporating verbal content. With

TTS annotations, it becomes more difficult to intelligibly convey

more than one layer of information through the audio channel. Our

design incorporates spatialization, pitch and timbral differentia-

tion, and rhythm to achieve clarity in the presentation of the hi-

erarchically structured data coming from the statistics engine.

The auditory display cycles through topic clusters, spending

relatively more time on subjects being actively discussed by the

largest numbers of people. Each different topic is assigned a dif-

ferent pitch group, reinforcing subject changes when they occur.

For each cluster, the statistics engine sends three streams of infor-

mation to the sonification engine:

Topics A continuously updated list of up to ten “topics” (the most

frequently appearing words and phrases – generalized se-

quences – mined from the multiple chat streams associated

with the given cluster; the number of topics is configurable,

but ten was chosen based on timing considerations);

Content samples A selection of sample sentences, identified by

the statistics engine as typical or representative, in which

these topics appear;

Content entropy A vector that represents the changing level of

entropy in the source data.

The topics are spoken by the TTS system6 at regular intervals in

a pitched monotone, and are panned alternately hard left and hard

right in the stereo field, creating a sort of rhythmic “call and re-

sponse.” The sample sentences are panned center, and rendered

with limited inflection (as opposed to the pitched monotone of the

topics). The tonal, rhythmic and spatial qualities of the topics con-

trasts sufficiently with the sample sentences to create two distinctly

comprehensible streams of verbal information.

The entropy vector controls an algorithmic piano score. When

entropy is minimal and the discussion in the chat room or bulletin

board is very focused on one subject, chords are played rhythmi-

cally in time with the rhythmic recitation of the topics. As entropy

increases and the conversations diverge, a Gaussian distribution is

used to expand the number, range and dynamics of notes that fall

between the chords. With this audio component, one can easily

differentiate a well-moderated content source from a more free-

form, public chat without distracting from the TTS annotations.

The piano score also serves a secondary function as an accompa-

niment to the vocal foreground, enhancing the compositional bal-

ance and overall musicality of the sound design.

2.2.3. Sample sonification and impressions

On our project Web site [5], we have a sample chat room sonifi-

cation that cycles through three topics. In this sound file, we are

6 The built-in MacOS TTS capability controlled by Max/MSP.

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-13

listening to the output of only three content monitors. Hence, by

design, each topic is confined to a single site. The first portion of

this example (ending at 1:47 into the sample) concerns the recent

recall of Bridgestone tires and was based on a www.cnn.com

chat room. This discussion was heavily moderated and hence the

backing piano score frequently reduces to a simple rhythm. For our

second topic (from 1:47 to 3:21 of the sample) we recorded chat

exchanges on www.financialchat.com one morning when

Yahoo’s stock opened low. In this example, we hear day traders

frantically exchanging predictions about when Yahoo’s stock will

“bounce.” The final topic in this sample (from 3:21 to the end) is

again from www.cnn.com and treats a recent strike by the Screen

Actor’s Guild and the American Federation of Television and Ra-

dio Artists. This chat room was much less moderated than the

previous CNN chat, and the backing piano score reflects that.

Although this example does not make full use of the clustering

capabilities of the statistics engine, the essence of our sonification

design is clear. The audio display provides an informative and

accessible representation of dynamic, textual content. The topic

and content sample streams are easy to separate, and when placed

in the background, call our attention to important new subjects

being discussed on the Web.

2.2.4. Applications and Extensions

Our sonification provides an audible interface to the (now) massive

amount of dynamic content available on the Web. Given the pre-

processing that takes place in the content monitors and the statis-

tics engine, a simple extension is to provide search-like function-

ality. A user can register interest in a certain topic and “tune”

our display to present only rooms where this subject is being dis-

cussed. The necessary ingredients to implement this feature are

all currently available in the statistics engine. Similarly, one can

easily restrict the sites that are used for the display. When a new

subject appears that draws the user’s interest, it is also trivial to

add a feature that would direct the user’s browser to one or more

chats associated with the topic. As a final extension, we have pro-

vided the content monitors with a configurable list of Web sites

that can be used to help disambiguate elements in the chat stream.

For example, the day traders speak in ticker symbols. Providing

the content monitor with the URL for the ticker symbol look-up

service offered by Yahoo allows the content monitor to weave not

only company names but also recent company-related headlines

directly into the stream fed to the statistics engine.

While we have focused mainly on chat and bulletin boards,

this technology can be applied in other settings. We have begun

collaborating with the designers of a natural language interface for

Web-based help systems. Here, we give voice to the hundreds of

simultaneous conversations taking place between Web site visitors

and the automated help system. A similar display can be imagined

for other natural language interfaces, including search engines like

AskJeeves (www.jeeves.com). In general, the practical appli-

cations of this summarization and auditory display tool abound.

3. CONCLUSION AND COMMENTS ON

COLLABORATIVE RESEARCH

The two applications outlined in this paper are the first outcomes of

a collaboration sponsored by Bell Laboratories and the Brooklyn

Academy of Music under the Arts in Multimedia project (AIM).

The goal of AIM is to bring together researchers (in this case a

statistician) and artists (in this case a sound artist), with the ob-

jective of advancing our separate agendas through collaborative

projects. Our work together is predicated on the notion that so-

phistication both in data treatment and aesthetics are crucial to the

successful design of audio displays. Thus, in each of our exam-

ples, we have endeavored to create a result which communicates

information clearly, yet at the same time sounds well composed

and appealing. Moving forward, it is our intention to apply these

techniques both to practical applications, and also to create a series

of artworks. These artworks will use our sonification techniques

to establish a series of real-time listening posts, both on the Web

and in physical locations. The listening posts will tap in to various

points of interest on the Internet, using sound to reveal patterns and

trends that would otherwise remain hidden.

In terms of applications, we are exploring the use of sonifica-

tion to support the design, provisioning and monitoring of commu-

nication networks. A network operations center (NOC), for exam-

ple, routinely receives clues about the health of the system in the

form of text messages generated by routers and switches. An audio

display installed inside a NOC can act as an early warning system

for approaching bottlenecks as well as aid in troubleshooting. By

continued exposure to the sound of a “normally” functioning net-

work, operators will be alerted to system changes that could signal

problems.

Art emerges unexpectedly from experimentations with new

statistical methods or considerations involving practical applica-

tions; and new tools for data analysis and modeling develop in re-

sponse to artistic concerns. Each of us continues to be surprised by

the connections that emerge from rethinking familiar problems in

a new context. Through our project, we hope to illustrate both the

value of art-technology collaborations as well as their necessity,

especially when finding meaning in complex data.

4. REFERENCES

[1] Visual insights. www.visualinsights.com.

[2] Cycling74. Max/msp. www.cycling74.com.

[3] Open directory project. www.dmoz.com.

[4] Ear to the ground. cm.bell-labs.com/stat/ear.

[5] Ear to the ground, chat example.

cm.bell-labs.com/stat/ear/chat.html.

[6] Ear to the ground, web traffic samples.

cm.bell-labs.com/stat/ear/samples.html.

[7] W. Gaul and L. Schmidt-Thieme. Mining web navigation

path fragments. In Proceedings of the Workshop on Web

Mining for E-Commerce – Challenges and Opportunities,

Boston, MA, August 2000.

[8] M. H. Hansen and B. Rubin. The audiences would be the

artists and their life would be the arts. IEEE MultiMedia,

7(2), April 2000.

[9] M. H. Hansen and B. Yu. Model selection and the princi-

ple of minimum description length. Journal of the American

Statistical Association. To appear.

[10] G. Kramer. An introduction to auditory display. In

G. Kramer, editor, Auditory Display. Addison-Wesley, 1994.

[11] N. Minar and J. Donath. Visualizing the crowds at a web site.

In Proceedings of CHI 99.

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-14

[12] R. Sen and M. H. Hansen. Predicting a web user’s next re-

quest based on log data. Submitted to ASA Student Paper

Competition.

[13] M. Wright. Open sound control. cnmat.berkeley.edu.

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

TAXONOMY AND DEFINITIONS FOR SONIFICATION AND AUDITORY DISPLAY

Thomas Hermann

Neuroinformatics Group

Faculty of Technology, Bielefeld University,Bielefeld, Germany

thermann@techfak.uni-bielefeld.de

ABSTRACT

Sonification is still a relatively young research field and

many terms such as sonification, auditory display, aural-

ization, audification have been used without a precise def-

inition. Recent developments such as the introduction of

Model-Based Sonification, the establishment of interactive

sonification and the increased interest in sonification from

arts have raised the need to revisit the definitions in order

to move towards a clearer terminology. This paper intro-

duces a new definition for sonification and auditory display

that emphasizes the necessary and sufficient conditions for

organized sound to be called sonification. It furthermore

suggests a taxonomy, and discusses the relation between vi-

sualization and sonification. A hierarchy of closed-loop in-

teractions is furthermore introduced. This paper aims to ini-

tiate vivid discussion towards the establishment of a deeper

theory of sonification and auditory display.

1. INTRODUCTION

Auditory Display is still a young research field whose birth

may be perhaps best traced back to the first ICAD confer-

ence1 in 1992 organized by Kramer. The resulting proceed-

ings volume “Auditory Display” [1] is still one of the most

important books in the field. Since then a vast growth of in-

terest, research, and initiatives in auditory display and soni-

fication has occurred. The potential of sound to support hu-

man activity, communication with technical systems and to

explore complex data has been acknowledged [2] and the

field has been established and has clearly left its infancy.

As in every new scientific field, the initial use of terms

lacks coherence and terms are being used with diffuse defi-

nitions. As the field matures and new techniques are discov-

ered, old definitions may appear too narrow, or, in light of

interdisciplinary applications, too unspecific. This is what

motivates the redefinitions in this article.

The shortest accepted definition for sonification is from

Barrass and Kramer et al. [2]: “Sonification is the use of

non-speech audio to convey information”. This definition

excludes speech as this was the primary association in the

1see www.icad.org

auditory display of information at that time. The definition

is unclear about what is meant by conveyance of informa-

tion: are real-world interaction sounds sonifications, e.g. of

the properties of an object that is being hit? Is a computer

necessary for its rendition? As a more specific definition,

the definition in [2] continues:

“Sonification is the transformation of data re-

lations into perceived relations in an acoustic

signal for the purposes of facilitating commu-

nication or interpretation.”

It is significant that the emphasis here is put on the pur-

pose of the usage of sound. This automatically distinguishes

sonification from music, where the purpose is not on the

precise perception of what interactions are done with an in-

strument or what data caused the sound, but on an underly-

ing artistic level that operates on a different level. Often, the

word ‘mapping’ has been used interchangeably with ‘trans-

formation’ in the above definition. This, however, suggests

a severe limitation of sonification towards just mappings be-

tween data and sound – which was perfectly fine at the time

of the definition where such a ‘Parameter-Mapping Sonifi-

cation’ was the dominating paradigm.

However, the introduction of Model-Based Sonification

(MBS) [3, 4] demonstrates methods to explore data by us-

ing sound in a way that is very different from a mapping:

in Parameter-Mapping Sonification, data values are mapped

to acoustic attributes of a sound (in other words: the data

‘play’ an instrument), whereas in MBS sonification models

create and configure dynamic processes that do not make

sound at all without external interactions (in other words:

the data is used to build an instrument or sound-capable

object, while the playing is left to the user). The user ex-

cites the sonification model and receives acoustic responses

that are determined by the temporal evolution of the model.

By doing this, structural information is holistically encoded

into the sound signal, and is no longer a mere mapping of

data to sound. One can perhaps state that data are mapped

to the configurations of sound-capable objects, but not that

they are mapped to sound.

Clearly, sonification models implemented according to

MBS are very much in line with the original idea that sonifi-

ICAD08-1

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

cation allows for the discovery of structures in data through

sound. Therefore there is the need to reformulate or adapt

the definition for sonification to better include such uses of

sound, and beyond that hopefully other possible yet-to-be-

discovered linkages between data and sound.

Another challenge for the definition comes from the use

of sonification in the arts and music: recently more and

more artists incorporate methods from sonification in their

work. What implications does this have for the term sonifi-

cation? Think of scientific visualization vs. art: what is the

difference between a painting and a modern visualization?

Both are certainly organized colors on a surface, both may

have aesthetic qualities, yet they operate on a completely

different level: the painting is viewed for different layers

of interpretation than the visualization. The visualization

is expected to have a precise connection to the underlying

data, else it would be useless for the process of interpret-

ing the data. In viewing the painting, however, the focus

is set more on whether the observer is being touched by it

or what interpretation the painter wants to inspire than what

can be learnt about the underlying data. Analogies between

sonification and music are close-by.

Although music and sonification are both organized

sound, and sonifications can sound like music and vice

versa, and certainly sonifications can be ‘heard as’ music

as pointed out in [5], there are important differences which

are so far not manifest in the definition of sonification.

2. A DEFINITION FOR SONIFICATION

This section introduces a definition for sonification in light

of the aforementioned problems. The definition has been

refined thanks to many fruitful discussions with colleagues

as listed in the acknowledgements and shall be regarded as

a new working definition to foster ongoing discussion in the

community towards a solid terminology.

Definition: A technique that uses data as input, and gener-

ates sound signals (eventually in response to optional addi-

tional excitation or triggering) may be called sonification,

if and only if

(C1) The sound reflects objective properties or relations in

the input data.

(C2) The transformation is systematic. This means that

there is a precise definition provided of how the data

(and optional interactions) cause the sound to change.

(C3) The sonification is reproducible: given the same data

and identical interactions (or triggers) the resulting

sound has to be structurally identical.

(C4) The system can intentionally be used with different

data, and also be used in repetition with the same

data.Data Sonification

Algorithm

systematic

transformation reproducable exchangeability

of data

interactions (optional)

Definition: Sonification

Figure 1: Illustration of the general structure and necessary

conditions for sonification. The yellow box depicts besides

the sonification elements few other components of auditory

displays, see also Sec. 3.

This definition emphasizes important prerequisites for

the scientific utility of sonification. It has several partly un-

expected implications that are to be explored in the follow-

ing discussion.

2.1. Discussion

2.1.1. General Comments

Sonification Techniques: According to the above defini-

tion, the techniques Audification, Earcons, Auditory Icons,

Parameter-Mapping Sonification as well as Model-Based

Sonification are all covered by the definition – they all rep-

resent information/data by using sound in an organized and

well-structured way and they are therefore different sonifi-

cation technique.2 This may first appear unfamiliar in light

of the common parlance to see earcons/auditory icons as

different from sonification. However, imagine an auditory

display for biomedical data that uses auditory icons as sonic

events to represent different classes (e.g. auditory icons for

benign/malignant tissue). The sonification would then be

the superposition or mixture of all the auditory icons chosen

for instance according to the class label and organized prop-

erly on the time axis. If we sonify a data set consisting only

of a single data item we naturally obtain as an extreme case

a single auditory icon. The same can be said for earcons.

Although sonification originally has the connotation of rep-

resenting large and complex data sets, it makes sense for the

definition to also work for single data points.

Data vs. Information: A distinction between data and

information is – as far as the above definition – irrelevant.

Think of earcons to represent computer desktop interactions

such as “delete file”, “rename folder”. There can be a lexi-

2they are also covered by the definition of sonification as ‘non-speech

use of sound to convey information’!

ICAD08-2

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

con of terms (file, folder, link) and actions (delete, rename,

etc.), and in practical computer implementations these fea-

tures would be represented numerically, e.g. object = O1,

action = A3. By doing so, the information has been turned

into data, and this is generally done if there is more than

one signal type to give. Information like for instance a

verbal message can always be represented numerically and

thus be understood as data. On the other side, raw data

values often carry semantic interpretation: e.g. the outside

temperature data value -10◦C (a one-dimensional data set

of size 1) – this is cold, and clearly information! Assum-

ing that information is always encoded as data values for

its processing we can deal with both in a single definition.

How the data are then represented by using sound is another

question: whether sonification techniques use a more sym-

bolic or analogic representation according to the analogic-

symbolic continuum of Kramer [6] is secondary for the def-

inition.

Mapping as a specific case of sonification: Some

articles have used “sonification” to refer specifically

to mapping-based sonification, where data features are

mapped to acoustic features of sound events or streams. Yet

sonification is more generally the representation of data by

using sound. There may be times when a clear specifica-

tion of the sonification technique, e.g. as model-based, au-

dification or parameter-mapping sonification, may be help-

ful to avoid confusion with the general term of sonification.

It makes sense to always use the most specific term possi-

ble, that is to use the term Parameter Mapping Sonification,

Audification, Model-Based Sonification, etc. to convey ex-

actly what is meant. The term Sonification, however, is,

according to the definition, more general which is also sup-

ported by many online definitions3. In result we suggest

using sonification with the same level of generality as the

term visualization is used in visual display.

Sonification as algorithm and sound: Sonification

refers to the technique and the process, so basically it refers

to the algorithm that is at work between the data, the user

and the resulting sound. Often, and with equal right, the re-

sulting sounds are called sonifications. Algorithm means a

set of clear rules, independent of whether it is implemented

on a computer or any other way.

Sonification as scientific method: According to the

definition, sonification is an accurate scientific method

which leads to reproducible results, addressing the ear

rather then the eye (as visualization does). This does not

limit the use of sonifications to data from the sciences, but

only states that sonification can be used as a valid instru-

ment to gain insight. The subjectivity in human percep-

3http://en.wikipedia.org/wiki/Sonification,

http://wvvel.csee.wvu.edu/sepscor/sonification/lesson9.html,

http://www.techfak.uni-bielefeld.de/ags/ni/projects/datamining/datason/

datason e.html, http://www.cs.uiowa.edu/ kearney/22c296Fall02/ Critten-

donSpecialty.pdf, to name a few.

tion and interpretation is shared with other perceptualization

techniques that bridge the gap between data and the human

sensory system. Being a scientific method, a prefix like in

“scientific sonification” is not necessary.

Same as some data visualizations may be ‘viewed’ as

art, sonifications may be heard as ‘music’[5], yet this use

differs from the original intent.

2.1.2. Comments to (C1)

(C1) The sound reflects objective properties or

relations in the input data.

Real-world acoustics are typically not a sonification al-

though they often deliver object-property-specific system-

atic sound, since there is no external input data as requested

in C1. For instance, with a bursting bottle, one can identify

what is the data, the model and the sound, but the process

cannot be repeated with the same bottle. However, using

a bottle that fills with rain, hitting it with a spoon once a

minute can be seen as a sonification: The data here is the

amount of rainfall, which is here measured by the fill level,

and the other conditions are also fulfilled. Tuning a guitar

string might also be regarded as a sonification to adjust the

tension of a string4. These examples show that sonifications

are not limited to computer-implementations according to

the definition, which embraces the possibility of other non-

computer-implemented sonifications.

The borders of sonification and real-world acoustics are

fuzzy. It might be discussed how helpful it is to regard or

denote everyday sounds as sonifications.

2.1.3. Comments to (C2)

(C2) The transformation is systematic. This

means that there is a precise definition pro-

vided of how the data (and optional interac-

tions) cause the sound to change.

What exactly do we mean by “precise”? Some sound

generators use noise and thereby random elements so that

sound events will per se sound different on each rendering.

In Parameter-Mapping Sonifications, the intentional addi-

tion of noise (for instance as onset jitter to increase per-

ceptability of events that would otherwise coincide) is often

used and makes sense. In order to include such cases ran-

domness is allowed in the definition, yet it is important to

declare where and what random elements are used (e.g. by

describing the noise distribution). It is also helpful to give

a motivation for the use of such random elements. By us-

ing too much noise, it is possible to generate useless soni-

fications in the sense that they garble interpretation of the

underlying data. In the same way it is possible to create

useless scientific visualizations.

4thanks to the referee for this example!

ICAD08-3

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

2.1.4. Comments to (C3)

(C3) The sonification is reproducible: given the

same data and identical interactions (or trig-

gers) the resulting sound has to be structurally

identical.

The definition claims reproducibility. This may not

strictly be achieved for several reasons: the loudspeakers

may generate a different sound at different temperatures,

other factors such as introduced noise as discussed above

may have been added. The use of the term “structurally

identical” in the definition aims to weaken the stronger

claim of sample-based identity. Sample-based identity is

not necessary, yet all possible psychophysical tests should

come to identical conclusions.

2.1.5. Comments to (C4)

(C4) The system can intentionally be used with

different data, and also be used in repetition

with the same data.

Repeatability is essential for a technique to be scientif-

ically valid and useful – otherwise nobody could check the

results obtained by using sonification as instrument to gain

insight. However, there are some implications by claim-

ing repeatability for what can and cannot be called sonifi-

cation. It has for instance been suggested that a musician

improvising on his instrument produces ‘a sonification of

the musician’s emotional state’. With C4, however, “play-

ing a musical instrument” is not a sonification of the per-

former’s emotional state, since it can not be repeated with

the ‘identical’ data. However, the resulting sound may be

called a sonification of the interactions with the instrument

(regarded here as data), and in fact, music can be heard with

the focus to understand the systematic interaction patterns

with the instruments.

Some of these conditions have been set as constraints

for sonification, e.g. reproducibility in the ‘Listening to the

Mind Listening’ concert5, but not been connected to a defi-

nition of sonification.

In summary, the given definition provides a set of neces-

sary conditions for systems and methods to be called soni-

fication. The definition is neither exhaustive nor complete;

we hope it will serve as the core definition as we as commu-

nity work towards a complete one.

3. SONIFICATION AND AUDITORY DISPLAY

With the above definition, the term sonification takes the

role of a general term to express the method of rendering

5http://www.icad.org/websiteV2.0/Conferences/ICAD2004/concert call.htm

sound in an organized and well-structured way. This is in

good analogy with the term visualization which is also the

general term under which a variety of specific techniques

such as bar charts, scatter plots, graphs, etc. are subsumed.

Particularly there is an analogy between scatter plots where

graphical symbols (data-mapped color/size...) are orga-

nized in space to deliver the visualization, and Parameter-

Mapping Sonification, where in a structurally identical way

acoustic events (with data-mapped features) are organized

in time. It is helpful to have with sonification a term that

operates on the same level of generality as visualization.

This raises the question what then do we mean by au-

ditory displays? Interestingly, in the visual realm, the

term ‘display’ suggests a necessary but complementary part

of the interface chain: the device to generate structured

light/images, for instance a CRT or LCD display or a projec-

tor. So in visualization, the term visualization emphasizes

the way how data are rendered as an image while the display

is necessary for a user to actually see the information. For

auditory display, we suggest to include this aspect of con-

version of sound signals into audible sound, so that an au-

ditory display encompasses also the technical system used

to create sound waves, or more general: all possible trans-

missions which finally lead to audible perceptions for the

user. This could range from loudspeakers over headphones

to bone conduction devices. We suggest furthermore that

auditory display should also include the user context (user,

task, background sound, constraints) and the application

context, since these are all quite essential for the design and

implementation. Sonification is thereby an integral compo-

nent within an auditory display system which addresses the

actual rendering of sound signals which in turn depend on

the data and optional interactions, as illustrated in Fig. 2.

Auditory Displays are more comprehensive than sonifica-Components of Auditory Display Systems

User/Listener

Technical

Sound Display

Sonification

(Rendering)

0101

0100

Application

Context

Data

Usage Context

mobile?

PC?

office?

Interactions

Figure 2: Auditory Displays: systems that employ sonifica-

tion for structuring sound and furthermore include the trans-

mission chain leading to audible perceptions and the appli-

cation context.

ICAD08-4

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

tion since for instance dialogue systems and speech inter-

faces may also be regarded as auditory displays since they

use sound for communication. While such interfaces are not

the primary focus in this research field the terminology sug-

gests their inclusion. On the other hand, Auditory Display

may be seen as a subset of the more general term of Audi-

tory Interfaces which do not only include output interfaces

(auditory displays, sonification) but also auditory input in-

terfaces which engender bidirectional auditory control and

communication between a user and a (in most cases) tech-

nical system (e.g. voice control system, query-by humming

systems, etc.).

4. HIERARCHY FROM SOUND TO

SONIFICATION

So far we have dealt with the necessary conditions sur-

rounding sonification and thus narrowed sonification down

to a specific subset of using sound. In this section, we look

at sonification in a systemic manner to elucidate its super-

ordinate categories. Figure 3 shows how we suggest to or-

ganize the different classes of sound. On the highest level,Map of Sound

Organized Sound

Functional Sounds

Music &

Media Arts Sonification(a)

(b)

Figure 3: Systemic map of sound, showing sonification and

its relation to other categories.

sounds are here classified as Organized Sound and unorga-

nized sound. Organized sounds separate from random or

otherwise complex structured sounds in the fact that their

occurence and structure is shaped by intention. Environ-

mental sounds appear often to be very structured and could

thus also be organized sounds, however, if so, any sound

would match that category to some extent. It thus may be

useful to apply the term to sounds that are intentionally or-

ganized – in most cases by the sound/interface developer.

The set of organized sound comprises two large sets that

partially overlap: music and functional sounds. Music is

without question a complex structured signal, organized on

various levels, from the acoustic signal to its temporal orga-

nization in bars, motifs, parts, layers. It is not our purpose

to give a definition of music.

The second set is functional sounds. These are orga-

nized sounds that serve a certain function or goal [7]. The

function is the motivation for their creation and use. To give

an example, all signal sounds (such as telephones, door-

bells, horns and warning hooters) are functional sounds.

Certainly there are intersections with music, as music can

serve functional aspects. For instance, trombones and kettle

drums have been used to demonstrate kingship and power.

A more subtle function is the use of music in supermarkets

to enhance the ‘shopping mood’. For that reason these sets

overlap – the size of the overlap depends on what is regarded

as function.

Sonification in the sense of the above definition is cer-

tainly a subset of functional sounds. The sounds are ren-

dered to fulfill a certain function, be it communication of in-

formation (signals & alarms), the monitoring of processes,

or to support better understanding of structure in data under

analysis. So is there a difference between functional sounds

and sonification at all? The following example makes clear

that sonification is really a subset: Recently a new selec-

tive acoustic weapon has been used, the mosquito device6,

a loudspeaker that produces a HF-sound inaudible to older

people, which drives away teenagers hanging around in

front of shops. This sound is surely functional, yet it could

neither pass as sonification nor as music.

Finally, we discuss whether sonification has an intersec-

tion with music&media arts. Obviously there are many ex-

amples where data are used to drive aspects of musical per-

formances, e.g. data collected from motion tracking or bio-

sensors attached to a performer. This is, concerning the in-

volved techniques and implementations similar to mapping

sonifications. However, a closer look at our proposed defi-

nition shows that often the condition for the transformation

to be systematic C2 is violated and the exact rules are not

made explicit. But without making the relationship explicit,

the listener cannot use the sound to understand the underly-

ing data better. In addition, condition C4 may often be vio-

lated. If sonification-like techniques are employed to obtain

a specific musical or acoustic effect without transparency

between the used data and details of the sonification tech-

niques, it might, for the sake of clarity, better be denoted

as ‘data-inspired music’, or ‘data-controlled music’ than as

sonification. Iannis Xenakis, for instance, did not even want

the listener to be aware of the data source nor the rules of

sound generation.

6see http://www.compoundsecurity.co.uk/, last seen 2008-01-16

ICAD08-5

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

5. CLOSED INTERACTION LOOPS

IN AUDITORY DISPLAYS

This section emphasizes the role of interaction in sonifica-

tion. We propose different terms depending on the scope of

the closure of the interaction loop. The motivation for this

discussion is that it might be helpful to address how terms

such as biofeedback or interactive sonification relate to each

other.

We start the discussion with Fig. 4 that depicts closed

loop interactions. The sonification module in the upper cen-

ter playing rendered sonifications to the user. Data sources

for sonification enter the box on the left side and the most

important parts are (a) World/System: this comprises any

system in the world that is connected to the sonification

module, e.g. via sensors that measure its state, and (b) Data:

these are any data under analysis or represented information

to be displayed that are stored separately and accessible by

the sonification.World/System

Sonification

Interactive Sonification

Human Activity (supported by sonification)

Auditory Biofeedback

Data

Navigation

Monitoring

No Action

Figure 4: Illustration of Closed-Loop Auditory Systems.

In this setting, Process Monitoring is the least inter-

active sonification, where data recorded from the world (in

real-time) or read from the data repository is continuously

used as input for a sonification rendering process. Here, the

listener is merely passively listening to the sound with the

only active component being his/her focus of attention onto

parts of the sound. Certainly, certain changes in the sound

might attract attention and force the user to act (e.g. sell

stocks, stop a machine, etc...).

A higher degree of active involvement occurs when the

user actively changes and adjusts parameters of the sonifi-

cation module, or interacts otherwise with the sonification

system. We denote this case as Interactive Sonification.

There is a wide field of possibilities of why and how to do

so, and we discuss 3 different prototypical examples:

(a) Triggering: Consider a mapping sonification of a

given data set. An essential interaction for the user

is to issue the command to render/playback the soni-

fication for a selected dataset. Possibly he/she does

this several times in order to attend different parts of

the sound signal. This elementary case is an interac-

tion, however, a very basic one.

(b) Parameter Adjustment is done when the user changes

parameters, such as what data feature are mapped

to acoustic parameters, control ranges, compression

factors, etc. Often such adjustments happen sepa-

rate from the playback so that the changes are made

and afterwards the updated sound is rendered. How-

ever, interactive real-time control is feasible in many

cases and shows a higher degree of interactivity. The

user actively explores the data by generating different

‘views’ of the data [8]. In visualization a similar in-

teractivity is obtained by allowing the user to select

axes scalings, etc.

(c) Excitatory Interaction is the third sort of interaction

and is structurally similar to the case of triggering.

Particularly in Model-Based Sonification [4], usually

the data are used to configure a sound-capable vir-

tual object that in turn reacts on excitatory interac-

tions with acoustic responses whereby the user can

explore the data interactively. Excitation puts energy

into the dynamic system and thus initiates an audible

dynamical system behavior. Beyond a simple trigger-

ing, excitatory interactions can be designed to make

use of the fine-grained manipulation skills that human

hands allow, e.g. by enabling to shake, squeeze, tilt or

deform the virtual object, for instance using sensor-

equipped physical interfaces to interact with the soni-

fication model. A good example for MBS is Shoogle

by Williamson et al. [9], where short text messages

in a mobile phone can be overviewed by shaking a

mobile phone equipped with accelerometer sensors,

resulting in audible responses of the text messages

as objects moving virtually inside the phone. Excita-

tory interactions offer rich and complex interactions

for interactive sonification.

The next possibility for a closed loop is by interactions

that select or browse data. Since data are chosen, it may

best be referred to as Navigation. Navigation can also be

regarded as special case of Interactive Sonification, depend-

ing on where the data are selected and the borders are here

really soft. Navigation usually goes hand in hand with trig-

gering of sonification (explained above).

Auditory Biofeedback can be interpreted as a sonifi-

cation of measured sensor data. In contrast to the above

types, the user’s activity is not controlling an otherwise au-

tonomous sonification with independent data, but it pro-

duces the input data for the sonification system. The user

perceives a sound that depends on his/her own activity.

Such systems have applications that range from rehabilita-

tion training to movement training in sports, e.g. to perform

ICAD08-6

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

a complex motion sequence (e.g. a tennis serve) so that its

sonification is structurally more similar to the sonification

of an expert performing the action [10].

The final category is Human Activity, which means

that the interaction ranges beyond the sonification system

into the world, often driven by the goal to change a world

state in a specific way. In turn, any sensors that pick up the

change may lead to changes in the sonification. The differ-

ence between the loop types before is that the primary fo-

cus is to achieve a goal beyond the sonification system, and

not to interact with a closed-loop sonification system. Even

without attending the sonification consciously or primarily,

the sound can be helpful to reach the goal. For example,

imagine the real-world task to fill a thermos bottle with tea.

While your primary goal is to get the bottle filled you will

receive the ‘gluck-gluck’ sound with increasing pitch as a

by-product of the interaction. If this is consistently useful,

you subconsciously adapt your activity to exploit the cues in

the sound – but the sound is only periphery for the goal. In a

similar sense, sonifications may deliver helpful by-products

to actions that change the world state. We regard such in-

teraction add-ons where sonification is a non-obtrusive yet

helpful cue for goal attainment as inspiring design direc-

tion. Such sonifications might even become subliminal in

the sense that users, when asked about the sound, are not

even aware of the sound, yet they perform better with sound

than without.

6. DISCUSSION AND CONCLUSION

The definitions in this paper are given on the basis of

three goals: (i) to anchor sonification as a precise scien-

tific method so that it delivers reproducible results and thus

can be used and trusted as instrument to obtain insight into

data under analysis. (ii) to offer a generalization which does

not limit itself to the special case of mappings from data to

sound, but which introduces sonification as general system-

atic mediator between data and sound, whatever the repre-

sentation might be. (iii) to balance the definition so that the

often-seen pair of terms ‘visualization & sonification’ are at

the same level of generality.

The definition has several implications which have been

discussed in Sec. 2. We’d like to emphasize that this effort

is being done in hope that the definition inspires a general

discussion on the terminology and taxonomy of the research

field of auditory display. An online version of the definition

is provided at www.sonification.de with the aim to collect

comments and examples of sonifications as well as exam-

ples that are agreed not to be sonifications and which help

in turn to improve the definition.

In Section 3, we described integral parts for auditory

display so that sonification takes a key component as the

technical part involving the rendition of sound. Again, the

suggested modules are meant as working hypothesis to be

discussed at ICAD.

While the given definitions specified terms on a horizon-

tal level, Section 4 proposes a vertical organization of sound

in relation to often used terms. The intersections between

the different terms and categories have been addressed with

examples.

Finally, we have presented in Section 5 an integrative

scheme for organizing different classes of auditory closed

loops according to the loop closure scope. It proves help-

ful to clarify classes of interactive sonifications. We think

that grouping existing sonifications according to these cat-

egories can be helpful to better find alternative approaches

for a given task.

The suggested terminology and taxonomy is the result

of many discussions and a thorough search for helpful con-

cepts. We suggest it as working definitions to be discussed

at the interdisciplinary level of ICAD in hope to contribute

towards a maturing of the fields of auditory display and

sonification.

7. ACKNOWLEDGEMENT

Many collegues have been very helpful in discussions to

refine the definitions. Particularly, I thank Till Bover-

mann, Arne Wulf, Andy Hunt, Florian Grond, Georg

Spehr, Alberto de Campo, Gerold Baier, Camille Peres,

and in particular Gregory Kramer for the helpful discus-

sions on the definition for sonification. Thanks also to col-

legues of the COST IC0601 Sonic Interaction Design (SID)

WG4/Sonification. I also thank Arne Wulf for the inspiring

discussions on Closed-Loop Auditory Systems, and Louise

Nickerson for many language improvements.

8. REFERENCES

[1] G. Kramer, Ed., Auditory Display - Sonification, Au-

dification, and Auditory Interfaces. Addison-Wesley,

1994.

[2] G. Kramer, B. Walker, T. Bonebright, P. Cook,

J. Flowers, N. Miner, and J. Neuhoff, “Sonification

report: Status of the field and research agenda,” Tech.

Rep., International Community for Auditory Display,

1999, http://www.icad.org/websiteV2.

0/References/nsf.html.

[3] Thomas Hermann and Helge Ritter, “Listen to your

data: Model-based sonification for data analysis,” in

Advances in intelligent computing and multimedia sys-

tems, G. E. Lasker, Ed., Baden-Baden, Germany, 08

1999, pp. 189–194, Int. Inst. for Advanced Studies in

System research and cybernetics.

ICAD08-7

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

[4] Thomas Hermann, Sonification for Exploratory Data

Analysis, Ph.D. thesis, Bielefeld University, Bielefeld,

Germany, 02 2002.

[5] Paul Vickers and Bennett Hogg, “Sonification ab-

straite/sonification concr`ete: An ‘æsthetic perspective

space’ for classifying auditory displays in the ars mu-

sica domain,” in ICAD 2006 - The 12th Meeting of the

International Conference on Auditory Display, Alis-

tair D N Edwards and Tony Stockman, Eds., London,

UK, June 20-23 2006, pp. 210–216.

[6] G. Kramer, “An introduction to auditory display,” in

Auditory Display, G. Kramer, Ed. ICAD, 1994, pp. 1–

79, Addison-Wesley.

[7] Georg Spehr, SOUND STUDIES. Traditionen - Meth-

oden – Desiderate, chapter Funktionale Kl ̈ange - Mehr

als ein Ping, transcript Verlag, Bielefeld, Germany,

2008.

[8] Thomas Hermann and Andy Hunt, “The discipline of

interactive sonification,” in Proceedings of the Inter-

national Workshop on Interactive Sonification (ISon

2004), Thomas Hermann and Andy Hunt, Eds., Biele-

feld, Germany, 01 2004, Bielefeld University, Interac-

tive Sonification Community, peer-reviewed article.

[9] John Williamson, Rod Murray-Smith, and S. Hughes,

“Shoogle: excitatory multimodal interaction on mo-

bile devices,” in Proceedings of the SIGCHI confer-

ence on Human factors in computing systems, San

Jose, California, USA, 2007, pp. 121–124, ACM

Press.

[10] Thomas Hermann, Oliver H ̈oner, and Helge Rit-

ter, “Acoumotion - an interactive sonification sys-

tem for acoustic motion control,” in Gesture in

Human-Computer Interaction and Simulation: 6th In-

ternational Gesture Workshop, GW 2005, Berder Is-

land, France, May 18-20, 2005, Revised Selected Pa-

pers, Sylvie Gibet, Nicolas Courty, and Jean-Francois

Kamp, Eds., Berlin, Heidelberg, 2006, vol. 3881/2006

of Lecture Notes in Computer Science, pp. 312–323,

Springer.

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/221513907

Nomadic Radio: Scaleable and Contextual Notification for Wearable Audio

Messaging.

Conference Paper · January 1999

DOI: 10.1145/302979.303005 · Source: DBLP

CITATIONS

95

READS

153

2 authors, including:

Some of the authors of this publication are also working on these related projects:

Bayesian Modeling for Human Development Indicators View project

Aware Community Portals View project

Nitin Sawhney

Aalto University

55 PUBLICATIONS 1,560 CITATIONS

SEE PROFILE

All content following this page was uploaded by Nitin Sawhney on 30 March 2015.

The user has requested enhancement of the downloaded file.

Papers CHI 99 15-20 MAY 1999

Nomadic Radio: Scaleable and Contextual Notification

for Wearable Audio Messaging

Nitin Sawhney and Chris Schmandt

SpeechInterface Group, MIT Media Laboratory

20 Ames St., Cambridge, MA 02139

{ nitin, geek} @media.mit.edu

ABSTRACT

Mobile workers need seamless access to communication

and information services on portable devices. However

current solutions overwhelm users with intrusive and

ambiguous notifications. In this paper, we describe

scaleable auditory techniques and a contextual notification

model for providing timely information, while minimizing

interruptions. User’s actions influence local adaptation in

the model. These techniques are demonstrated in Nomadic

Radio, an audio-only wearable computing platform.

Keywords

Auditory I/O, passive awareness, wearable computing,

adaptive interfaces, interruptions, notifications

INTRODUCTION

In today’s information-rich environments, people use a

number of appliances and portable devices for a variety of

tasks in the home, workplace and on the run. Such devices

are ubiquitous and each plays a unique functional role in a

user’s lifestyle. To be effective, these devices need to notify

users of changes in their functional state, incoming

messages or exceptional conditions. In a typical office

environment, the user attends to a plethora of devices with

notifications such as calls on telephones, asynchronous

messages on pagers, email notification on desktop

computers, and reminders on personal organizers or

watches. This scenario poses a number of key problems.

Lack of Differentiation in Notification Cues

Every device provides some unique form of notification. In

many cases, these are distinct auditory cues. Yet, most cues

are generally binary in nature, i.e. they convey only the

occurrence of a notification and not its urgency or dynamic

state. This prevents users from making timely decisions

about received messages without having to shift focus of

attention (from the primary task) to interact with the device

and access the relevant information.

Permission to make digital or hard copies of all or part of this work for

personal or classroom use is granted without fee provided that copies

arc not made or distributed for profit or commercial advantage and that

copies bear this notice and the full citation on the first page. To copy

otherwise, to republish, to post on servers or to redistribute to lists.

requires prior specific permission and/or a fee.

CHI ‘99 Pittsburgh PA USA

Copyright ACM 1999 0-201-48559-1/99/05...$5.00

Minimal Awareness of the User and Environment

Such notifications occur without any regard to the user’s

engagement in her current activity or her focus of attention.

This interrupts a conversation or causes an annoying

disruption in the user’s task and flow of thoughts. To

prevent undue embarrassment in social environments, users

typically turn off cell-phones and pagers in meetings or

lectures. This prevents the user from getting notification of

timely messages and frustrates people trying to get in touch

with her.

No Learning from Prior Interactions with User

Such systems typically have no mechanism to adapt their

behavior based on the positive or negative actions of the

user. Pagers continue to buzz and cell-phones do not stop

ringing despite the fact that the user may be in a

conversation and ignoring the device for some time.

Lack of Coordinated Notifications

All devices compete for a user’s undivided attention without

any coordination and synchronization of their notifications.

If two or more notifications occur within a short time of

each other, the user gets confused or frustrated. As people

start carrying around many such portable devices, frequent

and uncoordinated interruptions inhibit their daily tasks and

interactions in social environments.

Given these problems, most devices fail to serve their

intended purpose of notification or communication, and

thus do not operate in an efficient manner for a majority of

their life cycle. New users choose not to adopt such

technologies, having observed the obvious problems

encountered with their usage. In addition, current users tend

to turn off the devices in many situations, inhibiting the

optimal operation of such personal devices.

Nature of Interruptions in the Workplace

A recent observational study [4] evaluated the effect of

interruptions on the activity of mobile professionals in their

workplace. An interruption, defined as an asynchronous and

unscheduled interaction, not initiated by the user, results in

the recipient discontinuing the current activity. The results

revealed several key issues. On average, sub.jects were

interrupted over 4 times per hour, for an average duration

slightly over 2 minutes. Hence, nearly 10 minutes per hour

96

C H I 9 9 1 5 - 2 0 M A Y 1 9 9 9 Papers

was spent on interruptions. Although a majority of the

interruptions occurred in a face-to-face setting, 20% were

due to telephone calls (no email or pager activity was

analyzed in this study). In 64% of the interruptions, the

recipient received some benefit from the interaction. This

suggests that a blanket approach to prevent interruptions,

such as holding all calls at certain times of the day, would

prevent beneficial interactions from occurring. However in

41% of the interruptions, the recipients did not resume the

work they were doing prior to it. But active use of new

communication technologies makes users easily vulnerable

to undesirable interruptions.

These interruptions constitute a significant problem for

mobile professionals using tools such as pagers, cell-phones

and PDAs, by disrupting their time-critical activities.

Improved synchronous access using these tools benefits

initiators but leaves recipients with little control over the

interactions. The study suggests development of improved

filtering techniques that are especially light-weight, i.e.

don’t require more attention from the user and are less

disruptive than the interruption itself. By moving

interruptions to asynchronous media, messages can be

stored for retrieval and delivery at more appropriate times.

NOMADIC RADIO: WEARABLE AUDIO MESSAGING

Personal messaging and communication, demonstrated in

Nomadic Radio, provides a simple and constrained problem

domain in which to develop and evaluate a contextual

notification model. Messaging requires development of a

model that dynamically selects a suitable notification

strategy based on message priority, usage level, and

environmental context. Such a system must infer the user’s

attention by monitoring her current activities such as

interactions with the device and conversations in the room.

The user’s prior responses to notifications must also be

taken into consideration to adapt the notifications over time.

In this paper, we will consider techniques for scaleable

auditory presentation and an appropriate parameterized

approach towards contextual notification.

Several recent projects utilized speech and audio I/O on

wearable devices to present information. A prototype

augmented audio tour guide [l] played digital audio

recordings indexed by the spatial location of visitors in a

museum. SpeechWear [11] enabled users to perform data

entry and retrieval using speech recognition and synthesis.

Audio Aura [10] explored the use of background auditory

cues to provide serendipitous information coupled with

people’s physical location in the workplace. In Nomadic

Radio, the user’s inferred context rather than actual location

is used to decide when and how to deliver scaleable audio

notifications. In a recent paper [13], researchers suggest the

use of sensors and user modeling to allow wearables to

infer when users should be interrupted by incoming

messages. They suggest waiting for a break in the

conversation to post a message summary on the user’s

heads-up display. In this paper we describe a primarily non-

visual approach to provide timely information to nomadic

listeners, based on a variety of contextual cues.

Nomadic Radio is a wearable computing platform that

provides a unified audio-only interface to remote services

and messages such as email, voice mail, hourly news

broadcasts, and personal calendar events. These messages

are automatically downloaded to the device throughout the

day and users can browse through them using voice

commands and tactile input. The system consists of Java-

based clients and remote servers (written in C and Perl) that

communicate over wireless LAN, and utilize the telephony

infrastructure in the Speech Interface group. Simultaneous

spatial audio streams are rendered using a HRTF-based

Java audio API. Speech I/O is provided via a networked

implementation of AT&T Watson Speech API.

To provide a hands-free and unobtrusive interface to a

nomadic user, the system primarily operates as a wearable

audio-only device. The SoundBeam Neckset, a research

prototype patented by Nortel for use in hands-free

telephony, was adapted as the primary wearable platform in

Nomadic Radio. It consists of two directional speakers

mounted on the user’s shoulders, and a directional

microphone placed on the chest (see figure 1). Here

information and feedback is provided to the user through a

combination of auditory cues, spatial audio rendering, and

synthetic speech. Integration of a variety of auditory

techniques on a wearable device provides hands-free access

and navigation as well as lightweight and expressive

notification.

An audio-only interface has been incorporated in Nomadic

Radio, and a networked infrastructure for unified messaging

has been developed for wearable access [12]. The system

currently operates on a Libretto 100 mini-portable PC worn

by the user. The key issue addressed in this paper is that of

handling interruptions to the listener in a manner that

reduces disruption, while providing timely notifications for

contextually relevant messages.

P a p e r s

USAGE AND NOTIFICATION SCENARIO

The following scenario demonstrates the audio interface

and presentation of notifications in Nomadic Radio (no

voice commands from the user are shown here).

CHI 99 15-20 MAY 1999

SCALEABLE AUDITORY PRESENTATION

A scaleable presentation is necessary for delivering

sufficient information while minimizing interruption to the

listener. Messages in Nomadic Radio are scaled

dynamically to unfold as seven increasing levels of

notification (see figure 3): silence, ambient cues, auditory

cues, message summary, preview, full body, and foreground

rendering. These are described further below:

Silence for Least Interruption and Conservation

In this mode all auditory cues and speech feedback are

turned-off. Messages can be scaled down to silence when

the message priority is inferred to be too low for the

message to be relevant for playback or awareness to a user,

based on her recent usage of the device and the

conversation level. This mode also serves to conserve

processing, power and memory resources on a portable

device or wearable computer.

Ambient Cues for Peripheral Awareness

In Nomadic Radio, ambient auditory cues are continuously

played in the background to provide an awareness of the

operational state of the system and ongoing status of

messages being downloaded (see figure 4). The sound of

flowing water provides an unobtrusive form of ambient

awareness that indicates the system is active (silence

indicates sleep mode). Such a sound tends to fade into the

perceptual background after a short time, so it does not

distract the listener. The pitch is increased during file

downloads, momentarily foregrounding the ambient sound.

A short e-mail message sounds like a splash while a two-

minute audio news summary is heard as faster flowing

water while being downloaded. This implicitly indicates

message size without the need for additional audio cues and

prepares the listener to hear (or deactivate) the message

before it becomes available. Such peripheral awareness

minimizes cognitive overhead of monitoring incoming

messages relative to notifications played as distinct auditory

cues, which incur a somewhat higher cost of attention on

part of the listener.

Related Work in Auditory Awareness

In ARKola [5], an audio/visual simulation of a bottling

factory, repetitive streams of sounds allowed people to keep

track of activity, rate, and functioning of running machines.

Without sounds people often overlooked problems; with

auditory cues, problems were indicated by the machine’s

sound ceasing (often ineffective) or via distinct alert

sounds. The various auditory cues (as many as 12 sounds

play simultaneously) merged as an auditory texture, allowed

people to hear the plant as a complex integrated process.

Background sounds were also explored in ShareMon [3], a

prototype application that notified users of file sharing

activity. Cohen found that pink noise used to indicate

%CPU time was considered “obnoxious”, even though

users understood the, pitch correlation. However,

preliminary reactions to wave sounds were considered

positive and even soothing. In Audio Aura [IO], alarm

CHI 9 9 1 5 - 2 0 M A Y 1 9 9 9 P a p e r s

sounds were eliminated and a number of “harmonically

coherent sonic ecologies” were explored, mapping events to

auditory, musical or voice-based feedback. Such techniques

were used to passively convey the number of email

messages received, identity of senders, and abstract

representations of group activity.

Auditory Cues for Notification and Identification

In Nomadic Radio, auditory cues are a crucial means for

conveying awareness, notification and providing necessary

assurances in its non-visual interface. Different types of

auditory techniques provide distinct feedback, awareness

and message information.

Feedback Cues

Several types of audio cues indicate feedback for a number

of operational events in Nomadic Radio:

1. Task completion and confirmations - button pressed,

speech understood, connected to servers, finished

playing or loaded/deleted messages.

2. Mode transitions - switching categories, going to

non-speech or ambient mode.

3. Exceptional conditions - message not found, lost

connection with servers, and errors.

Priority Cues for Notification

In a related project, “email glances” [7] were formulated as

a stream of short sounds indicating category, sender and

content flags (from keywords in the message). In Nomadic

Radio, message priority inferred from email content

filtering provides distinct auditory cues (assigned by the

user) for group, personal, timely, and important messages.

In addition, auditory cues such as telephone ringing indicate

voice mail, whereas an extracted sound of a station

identifier indicates a news summary.

VoiceCues for Identification

VoiceCues represent a novel approach for easy

identification of the sender of an email, based on a unique

auditory signature of the person. VoiceCues are created by

manually extracting a l-2 second audio sample from the

voice messages of callers and associating them with their

respective email login. When a new email message arrives,

the system queries its database for a related VoiceCue for

that person before playing it to the user as a notification,

along with the priority cues. The authors have found

VoiceCues to be a remarkably effective method for quickly

conveying the sender of the message in a very short

duration. This technique reduces the need for synthetic

speech feedback, which can often be distracting.

99

Papers CHI 99 15-20 MAY 1999

Message Summary Generation

A spoken description of an incoming message can present

relevant information in a concise manner. Such a

description typically utilizes header information in email

messages to convey the name of the sender and the subject

of the message. In Nomadic Radio, message summaries are

generated for all messages, including voice-mail, news and

calendar events. The summaries are augmented by

additional attributes of the message indicating category,

order, priority, and duration. For audio sources, like voice

messages and news broadcasts, the system plays the first

2.5 seconds of the audio. This identifies the caller and the

urgency of the call, inferred from intonation in the caller’s

voice or provides a station identifier for news summaries.

Message Previews using Content Summarization

Messages are scaled to allow listeners to quickly preview

the contents of an email or voice message. In Nomadic

Radio, a preview for text messages extracts the first 100

characters of the message (a default size that can be user

defined). This heuristic generally provides sufficient

context for the listener to anticipate the overall message

theme and urgency. For email messages, redundant headers

and previous replies are eliminated from the preview for

effective extraction. Use of text summarization techniques,

based on tools such as ProSum’ developed by British

Telecom, would allow more flexible means of scaling

message content. Natural language parsing techniques used

in ProSum permit a scaleable summary of an arbitrarily

large text document.

A preview for an audio source such as a voice message or

news broadcast presents a fifth of the message at a

gradually increasing playback rate of up to 1.3 times faster

than normal. There are a range of techniques for time-

compressing speech without modifying the pitch, however

twice the playback rate usually makes the audio

incomprehensible. A better representation for content

summarization requires a structural description of the audio,

based on annotated or automatically determined pauses in

speech, speaker and topic changes. Such an auditory

thumbnail must function similar to its visual counterpart. A

preview for a structured voice message would provide

pertinent aspects such as name of caller and phone number,

whereas a structured news preview would be heard as the

hourly headlines.

Full Body: Playing Complete Message Content

This mode plays the entire audio file or reads the full text of

the message at the original playback rate. Some parsing of

the text is necessary to eliminate redundant header

information and format tags. The message is augmented

with summary information indicating sender and subject.

This message is generally spoken or played in the

background of the listener’s audio space.

I http://transend.Iabs.bt.com/prosum/on-line/

Foreground Rendering via Spatial Proximity

An important message is played in the foreground of the

listening space. The audio source of the message is rapidly

moved closer to the listener, allowing it to be heard louder,

and played there for 415” of its duration. The message

gradually begins to fade away, moving back to its original

position and amplitude for the remaining l/S” of the

duration. The foregrounding algorithm ensures that the

messages are quickly brought into perceptual focus by

pulling them to the listener rapidly. However the messages

are pushed back slowly to provide an easy fading effect as

the next one is heard. As the message moves its spatial

direction is maintained so that the listener can retain a focus

on the audio source even if another begins to play.

Hence a range of techniques provide scaleable forms of

background awareness, auditory notification, spoken

feedback and foreground rendering of incoming messages.

CONTEXTUAL NOTIFICATION

In Nomadic Radio, context dynamically scales the

notifications for incoming messages. The primary

contextual cues used include: message priority from email

filtering, usage level based on time since last user action,

and the likelihood of conversation estimated from real-time

analysis of the auditory scene. In our experience these

parameters provide sufficient context to scale notifications,

however data from motion or location sensors can also be

integrated in such a model. A linear and scaleable auditory

notification model is utilized, based on the notion of

estimating costs of interruption and the value of information

to be delivered to the user. This approach is similar to

recent work [6] on using perceptual costs and a focus of

attention model for scaleable graphics rendering.

Message Priority

The priority of incoming messages is explicitly determined

via content-based email filtering using CLUES [9], a

filtering and prioritization system. CLUES has been

integrated into Nomadic Radio to determine the timely

nature of messages by finding correlation between a user’s

calendar, rolodex, to-do list, as well as a record of outgoing

messages and phone calls. These rules are integrated with

static rules created by the user for prioritizing specific

people or message subjects. When a new email message

arrives, keywords from its sender and. subject header

information are correlated with static and generated

filtering rules to assign a priority to the message. Email

messages are also prioritized if the user is traveling and

meeting others in the same geographic area (via area codes

in the rolodex). The current priorities include: group,

personal, very important, most important, and timely.

Priorities are parameterized by logarithmically scaling all

priorities within a range of 0 to 1. Logarithmic scaling

ensures that higher priority messages are weighted higher

relative to unimportant or uncategorized messages.

Priority ( i ) = ( log ( i ) / log (Priority Levels Mu ) )

100

C H I 9 9 1 5 - 2 0 M A Y 1 9 9 9 Papers

Usage Level

One problem with using last actions for setting usage levels

is that if a user deactivates an annoying message, that

action is again time-stamped. Such negative reinforcements

continue to increase the usage level and the related

notification. Therefore negative actions such as stopping

audio playback or deactivating speech are excluded from

generating actions for computing the usage.

Likelihood of Conversation

Conversation in the environment can be used to gauge

whether the user is in a social context where an

interruption is less appropriate. If the system detects the

occurrence of more than several speakers over a period of

time, that is an indication of a conversational situation.

Auditory events are first detected by adaptively

thresholding total energy and incorporating constraints on

event length and surrounding pauses. The system uses mel-

scaled filter-bank coefficients (MFCs) and pitch estimates

to discriminate, reasonably well, a variety of speech and

non-speech sounds. HMMs (Hidden Markov Models)

capture both the temporal characteristics and spectral

content of sound events. The techniques for feature

extraction and classification of the auditory scene using

HMMs are described in a recent workshop paper [2]. The

likelihood of speech detected in the environment is

computed for each event in a short window of time. In

addition, the probabilities are weighted, such that most

recent time periods in the window are considered more

relevant for computing the overall Speech Level. We are

evaluating the classifier’s effectiveness by training it with a

variety of speakers and background sounds.

Notification Level

A weighted average for all three contextual cues provides

level has an inversely proportional relationship with

notification i.e. a lower notification must be provided

during high conversation.

Presentation Latency

Latency represents the period of time to wait before

playing the message to the listener, after a notification cue

is delivered. Latency is computed as a function of the

notification level and the maximum window of time

(Latency,& that a lowest priority message can be delayed

for playback. The default maximum latency is set to 20

seconds, but can be modified by the user.

101

CHI99 15-20 MAY 1999

were increased. Jane was notified of a group message

shortly after the voice message, since the system detected

higher usage activity. Hence, the system correctly scaled

down notifications when Jane did not want to be bothered

whereas notifications were scaled up when Jane started to

use the system to browse her messages.

EFFECTIVENESS OF THE NOTIFICATION MODEL

The nature of peripheral awareness and unobtrusive

notification on a wearable device requires a usage

evaluation that must be conducted on an ongoing and long-

term basis. However, the predictive effectiveness of the

notification model must first be evaluated on a quantitative

basis. Hence, all message and notification parameters are

captured for such analysis. Lets consider two actual

examples of notification computed for email messages with

different priorities. Figure 7 shows an auditory cue

generated for a group message (low priority).

The timely message (in figure 8) received greater priority

and consequently a higher notification level for summary

playback. A moderate latency time (approx. 6 secs.) was

chosen. However when the user interrupted the notification

by a button press, the summary playback was aborted. The

user’s action reduced overall weights by 5%.

P a p e r s

Dynamic Adaptation of the Notification Model

The user can initially set the weights for the notification

model to high, medium, or low (interruption). These weight

settings were selected by experimenting with notifications

over time using an interactive visualization of message

parameters. This allowed us to observe the model, modify

weights and infer the effect on notification based on

different weighting strategies. Pre-defined weights provide

an approximate behavior for the model and help bootstrap

the system for novice users. The system also allows the user

to dynamically adjust these weights (changing the

interruption and notification levels) by their implicit actions

while playing or ignoring messages.

The system allows localized positive and negative

reinforcement of the weights by monitoring the actions of

the user during notifications. As a message arrives, the

system plays an auditory cue if its computed notification

level is above the necessary threshold for auditory cues. It

then uses the computed latency interval to wait before

playing the appropriate summary or preview of the

message. During that time, the user can request the message

be played earlier or abort any further notification for the

message via speech or button commands. If aborted, all

weights are reduced by a fixed percentage (default is 5%), a

negative reinforcement. If the user activates the message

(positive reinforcement) within 60 seconds after the

notification, the playback scale selected by the user is used

to increase all weights. If the message is ignored, no change

is made to the weights, but the message remains active for

60 seconds during which the user’s actions can continue to

influence the weights.

Figure 6 shows a zoomed view of the extended scenario

introduced earlier, focusing on Jane’s actions that reinforce

the model. Jane received several messages and ignored

most of the group messages and a recent personal message

(the weights remain unchanged). While in the meeting, Jane

interrupted a timely message to abort its playback. This

reduced the weights for future messages, and the ones with

low priority (group message) were not notified to Jane. The

voice message from Kathy, her daughter, prompted Jane to

reinforce the message by playing it. In this case, the weights

Continuous local reinforcement over time should allow the

system to reach a state where it is somewhat stable and

robust in converging to the user’s preferred notification.

Currently the user’s actions primarily adjust weights for

subsequent messages, however effective reinforcement

learning requires a model that generalizes a notification

policy that maximizes some long-term measure of

reinforcement [8]; this will be the focus of our future work.

Cl-II 99 15-20 MAY 1999 Papers

PRELIMINARY EVALUATION

Although the authors have been using and relining these

techniques during system development, a preliminary 2-day

evaluation was conducted with a novice user, who had prior

experience with mobile phones and 2-way pagers. The user

was able to listen to notifications while attending to tasks in

parallel such as reading or typing. He managed to have

casual discussions with others while hearing notifications;

however he preferred turning off all audio during an

important meeting with his advisor. People nearby

sometimes found the spoken feedback distracting if heard

louder, however that also cued them to wait before

interrupting the user. The volume on the device was

lowered to minimize any disruption to others and maintain

the privacy of messages. The user requested an automatic

volume gain that adapted to the environmental noise level.

In contrast to speech-only feedback, the user found the

unfolding presentation of ambient and auditory cues

allowed sufficient time to switch attention to the incoming

message. Familiarization with the auditory cues was

necessary. He preferred longer and gradual notifications

rather than distinct auditory tones. The priority cues were

the least useful indicator whereas VoiceCues provided

obvious benefit. Knowing the actual priority of a message

was less important than simply having it presented in the

right manner. The user suggested weaving message priority

into the ambient audio (as increased pitch). He found the

overall auditory scheme somewhat complex, preferring

instead a simple notification consisting of ambient

awareness, Voice&es and spoken text.

The user stressed that the ambient audio provided the most

benefit while requiring least cognitive effort. He wished to

hear ambient audio at all times to remain reassured that the

system was still operational. An unintended effect

discovered was that a “pulsating” audio stream indicated

low battery power on the wearable device. A “pause” button

was requested, to hold all messages while participating in a

conversation, along with subtle but periodic auditory alerts

for unread messages waiting in queue. The user felt that

Nomadic Radio provided appropriate awareness and its

expressive qualities justified its use over a pager. A long-

term trial with several nomadic users is necessary to further

validate these notification techniques.

CONCLUSIONS

We have demonstrated techniques for scaleable auditory

presentation and message notification using a variety of

contextual cues. The auditory techniques and notification

model have been refined based on continuous usage by the

authors, however we are currently conducting additional

evaluations with several users. Ongoing work explores

adaptation of the notification model based on reinforcement

from user behavior over time. Our efforts have focused on

wearable audio platforms, however these ideas can be

readily utilized in consumer devices such as pagers, PDAs

and mobile phones to minimize disruptions while providing

timely information to users on the move.

ACKNOWLEDGMENTS

Thanks to Brian Clarkson for ongoing work on the audio

classifier and Stefan Marti for help with user evaluations.

We also thank Lisa Fast and Andre Van Schyndel at Nortel

for their support of the project.

REFERENCES

1.

2.

3.

4.

5.

6.

7.

8.

9.

Bederson, Benjamin B. Audio Augmented Reality: A

Prototype Automated Tour Guide. Proceedings of CHI

‘95, May 1995, pp. 210-211.

Clarkson, Brian, Nitin Sawhney and Alex Pentland.

Auditory Context Awareness via Wearable Computing,

Workshop on Perceptual User Inte$aces, Nov. 1998.

Cohen, J. Monitoring Background Activities. Auditory

Display: Sonitication, Audification, and Auditory

Interfaces. Reading MA: Addison-Wesley, 1994.

Conaill, 0’ Brid and David Frohlich. Timespace in the

Workplace: Dealing with Interruptions. Proceedings of

CHI ‘95, 1995.

Gaver, W.W., R. B. Smith, T. OShea. Effective Sounds

in Complex Systems: The ARKola Simulation.

Proceedings of CHI ‘91, April 28-May 2, 1991.

Horvitz, Eric and Jed Lengyel. Perception, Attention,

and Resources: A Decision-Theoretic Approach to

Graphics Rendering. Proceedings of Uncertainty in

Artificial Intelligence, Aug. l-3, 1997, pp. 238-249.

Hudson, Scott E. and Ian Smith. Electronic Mail

Previews Using Non-Speech Audio. Proceedings of

CHI ‘96, April 1996, pp. 237-238.

Kaelbling, L.P. and Littman, M.L. Reinforcement

Learning: A Survey. Journal of Artijicial Intelligence

Research, vol. 4, 1996, pp. 237-285.

Marx, Matthew and Chris Schmandt. CLUES: Dynamic

Personalized Message Filtering. Proceedings of CSCW

‘96, pp. 113-121, November 1996.

lO.Mynatt, E.D., Back, M., Want, R. Baer, M., and Ellis

J.B. Designing Audio Aura. Proceedings of CHI ‘98,

April 1998.

1l.Rudnicky, Alexander, Reed, S. and Thayer, E.

SpeechWear: A mobile speech system. Proceedings of

ICSLP ‘96,1996.

12.Sawhney, Nitin and Chris Schmandt. Speaking and

Listening on the Run: Design for Wearable Audio

Computing. Proceedings of the International

Symposium on Wearable Computing, October 1998.

13. Stainer, Thad, Mann, S., Rhodes, B., Levine, J., Healey,

J., Kirsch, D., Picard, R., and Pentland, A. Augmented

Reality through Wearable Computing. Presence, Vol. 6,

No. 4, August 1997, pp. 386-398.

The Physics of Sound

Sound lies at the very center of speech communication. A sound wave is both the end product of the speech

production mechanism and the primary source of raw material used by the listener to recover the speaker's message.

Because of the central role played by sound in speech communication, it is important to have a good understanding

of how sound is produced, modified, and measured. The purpose of this chapter will be to review some basic

principles underlying the physics of sound, with a particular focus on two ideas that play an especially important

role in both speech and hearing: the concept of the spectrum and acoustic filtering. The speech production

mechanism is a kind of assembly line that operates by generating some relatively simple sounds consisting of

various combinations of buzzes, hisses, and pops, and then filtering those sounds by making a number of fine

adjustments to the tongue, lips, jaw, soft palate, and other articulators. We will also see that a crucial step at the

receiving end occurs when the ear breaks this complex sound into its individual frequency components in much the

same way that a prism breaks white light into components of different optical frequencies. Before getting into these

ideas it is first necessary to cover the basic principles of vibration and sound propagation.

Sound and Vibration

A sound wave is an air pressure disturbance that results from vibration. The vibration can come from a tuning

fork, a guitar string, the column of air in an organ pipe, the head (or rim) of a snare drum, steam escaping from a

radiator, the reed on a clarinet, the diaphragm of a loudspeaker, the vocal cords, or virtually anything that vibrates in

a frequency range that is audible to a listener (roughly 20 to 20,000 cycles per second for humans). The two

conditions that are required for the generation of a sound wave are a vibratory disturbance and an elastic medium,

the most familiar of which is air. We will begin by describing the characteristics of vibrating objects, and then see

what happens when vibratory motion occurs in an elastic medium such as air. We can begin by examining a simple

vibrating object such as the one shown in Figure 3-1. If we set this object into vibration by tapping it from the

bottom, the bar will begin an upward and downward oscillation until the internal resistance of the bar causes the

vibration to cease.

The graph to the right of Figure 3-1 is a visual representation of the upward and downward motion of the bar.

To see how this graph is created, imagine that we use a strobe light to take a series of snapshots of the bar as it

vibrates up and down. For each snapshot, we measure the instantaneous displacement of the bar, which is the

difference between the position of the bar at the split second that the snapshot is taken and the position of the bar at

rest. The rest position of the bar is arbitrarily given a displacement of zero; positive numbers are used for

displacements above the rest position, and negative numbers are used for displacements below the rest position. So,

the first snapshot, taken just as the bar is struck, will show an instantaneous displacement of zero; the next snapshot

will show a small positive displacement, the next will show a somewhat larger positive displacement, and so on. The

pattern that is traced out has a very specific shape to it. The type of vibratory motion that is produced by a simple

vibratory system of this kind is called simple harmonic motion or uniform circular motion, and the pattern that is

traced out in the graph is called a sine wave or a sinusoid.

Figure 3-1. A bar is fixed at one and is set into vibration by tapping it from the bottom. Imagine that

a strobe light is used to take a series of snapshots of the bar as it vibrates up and down. At each

snapshot the instantaneous displacement of the bar is measured. Instantaneous displacement is the

distance between the rest position of the bar (defined as zero displacement) and its position at any

particular instant in time. Positive numbers signify displacements that are above the rest position,

while negative numbers signify displacements that are below the rest position. The vibratory pattern

that is traced out when the sequence of displacements is graphed is called a sinusoid.

The Physics of Sound 2

Basic Terminology

We are now in a position to define some of the basic terminology that applies to sinusoidal vibration.

periodic: The vibratory pattern in Figure 3-1, and the waveform that is shown in the graph, are examples of

periodic vibration, which simply means that there is a pattern that repeats itself over time.

cycle: Cycle refers to one repetition of the pattern. The instantaneous displacement waveform in Figure 3-1 shows

four cycles, or four repetitions of the pattern.

period: Period is the time required to complete one cycle of vibration. For example, if 20 cycles are completed in 1

second, the period is 1/20th of a second (s), or 0.05 s. For speech applications, the most commonly used unit of

measurement for period is the millisecond (ms):

1 ms = 1/1,000 s = 0.001 s = 10 -3 s

A somewhat less commonly used unit is the microsecond (μs):

1 μs = 1/1,000,000 s = 0.000001 s = 10 -6 s

frequency: Frequency is defined as the number of cycles completed in one second. The unit of measurement for

frequency is hertz (Hz), and it is fully synonymous the older and more straightforward term cycles per second

(cps). Conceptually, frequency is simply the rate of vibration. The most crucial function of the auditory system is to

serve as a frequency analyzer – a system that determines how much energy is present at different signal frequencies.

Consequently, frequency is the single most important concept in hearing science. The formula for frequency is:

f = 1/t, where: f = frequency in Hz

t = period in seconds

So, for a period 0.05 s:

f = 1/t = 1/0.05 = 20 Hz

It is important to note that period must be represented in seconds in order to get the answer to come out in cycles per

second, or Hz. If the period is represented in milliseconds, which is very often the case, the period first has to be

converted from milliseconds into seconds by shifting the decimal point three places to the left. For example, for a

period of 10 ms:

f = 1/10 ms = 1/0.01 s = 100 Hz

Similarly, for a period of 100 μs:

f = 1/100 μs = 1/0.0001 s = 10,000 Hz

The period can also be calculated if the frequency is known. Since period and frequency are inversely related, t

= 1/f. So, for a 200 Hz frequency, t = 1/200 = 0.005 s = 5 ms.

Characteristics of Simple Vibratory Systems

Simple vibratory systems of this kind can differ from one another in just three dimensions: frequency,

amplitude, and phase. Figure 3-2 shows examples of signals that differ in frequency. The term amplitude is a bit

different from the other terms that have been discussed thus far, such as force and pressure. As we saw in the last

chapter, terms such as force and pressure have quite specific definitions as various combinations of the basic

dimensions of mass, time, and distance. Amplitude, on the other hand, will be used in this text as a generic term

meaning "how much." How much what? The term amplitude can be used to refer to the magnitude of displacement,

the magnitude of an air pressure disturbance, the magnitude of a force, the magnitude of power, and so on. In the

The Physics of Sound 3

0 5 10 15 20 25 30 35 40 45 50

-10

-5

0

5

10

Time (ms)

Instantaneous Amp.

-10

-5

0

5

10

Instantaneous Amp.

present context, the term amplitude refers to the magnitude of the displacement pattern. Figure 3-3 shows two

displacement waveforms that differ in amplitude. Although the concept of amplitude is as straightforward as the two

waveforms shown in the figure suggest, measuring amplitude is not as simple as it might seem. The reason is that

the instantaneous amplitude of the waveform (in this case, the displacement of the object at a particular split

second in time) is constantly changing. There are many ways to measure amplitude, but a very simple method called

peak-to-peak amplitude will serve our purposes well enough. Peak-to-peak amplitude is simply the difference in

amplitude between the maximum positive and maximum negative peaks in the signal. For example, the bottom

panel in Figure 3-3 has a peak-to-peak amplitude of 10 cm, and the top panel has a peak-to-peak amplitude of 20

cm. Figure 3-4 shows several signals that are identical in frequency and amplitude, but differ from one another in

phase. The waveform labeled 0 o phase would be produced if the bar were set into vibration by tapping it from the

bottom. The waveform labeled 180 o phase would be produced if the bar were set into vibration by tapping it from

the top, so that the initial movement of the bar was downward rather than upward. The waveforms labeled 90 o phase

and 270 o phase would be produced if the bar were set into vibration by pulling the bar to maximum displacement

and letting go -- beginning at maximum positive displacement for 90 o phase, and beginning at maximum negative

displacement for 270 o phase. So, the various vibratory patterns shown in Figure 3-4 are identical except with respect

to phase; that is, they begin at different points in the vibratory cycle. As can be seen in Figure 3-5, the system for

representing phase in degrees treats one cycle of the waveform as a circle; that is, one cycle equals 360 o. For

example, a waveform that begins at zero displacement and shows its initial movement upward has a phase of 0 o, a

waveform that begins at maximum positive displacement and shows its initial movement downward has a phase of

90 o, and so on.

Figure 3-2. Two vibratory patterns that differ in frequency. The panel on top is higher in frequency

than the panel on bottom.

The Physics of Sound 4

0 5 10 15 20 25 30 35 40 45 50

-10

-5

0

5

10

Time (ms)

Instantaneous Amp.

-10

-5

0

5

10

Instantaneous Amp.

Figure 3-3. Two vibratory patterns that differ in amplitude. The panel on top is higher in amplitude than the

panel on bottom.

Phase: 0

Phase: 90

Phase: 180

Phase: 270

Figure 3-4. Four vibratory patterns that differ in phase. Shown above are vibratory patterns with phases of 0 0, 90 0,

180 0, and 270 0.

The Physics of Sound 5

Springs and Masses

We have noted that objects can vibrate at different frequencies, but so far have not discussed the physical

characteristics that are responsible for variations in frequency. There are many factors that affect the natural

vibrating frequency of an object, but among the most important are the mass and stiffness of the object. The effects

of mass and stiffness on natural vibrating frequency can be illustrated with the simple spring-and-mass systems

shown in Figure 3-6. In the pair of spring-and-mass systems to the left, the masses are identical but one spring is

stiffer than the other. If these two spring-and-mass systems are set into vibration, the system with the stiffer spring

will vibrate at a higher frequency than the system with the looser spring. This effect is similar to the changes in

Time ->

Instantaneous Amplitude

0

90

180

270

0/360

Figure 3-5. The system for representing phase treats one cycle of the vibratory pattern as a circle,

consisting of 360 0

. A pattern that begins at zero amplitude heading toward positive values (i.e., heading

upward) is designated 0 0 phase; a waveform that begins at maximum positive displacement and shows

its initial movement downward has a phase of 90 o

; a waveform that begins at zero and heads

downward has a phase of 180 o; and a waveform that begins at maximum negative displacement and

shows its initial movement upward has a phase of 270 o. . The four phase angles that are shown above

are just examples. An infinite variety of phase angles are possible.

Figure 3-6. A spring and mass system whose natural vibrating frequency is controlled by two

parameters: (1) the stiffness of the spring (the stiffer the spring the higher the natural vibrating

frequency), and (2) the mass of the material that is suspended from the spring (the greater the mass, the

lower the natural vibrating frequency).

The Physics of Sound 6

frequency that occur when a guitarist turns the tuning key clockwise or counterclockwise to tune a guitar string by

altering its stiffness.1

The spring-and-mass systems to the right have identical springs but different masses. When these systems are

set into vibration, the system with the greater mass will show a lower natural vibrating frequency. The reason is that

the larger mass shows greater inertia and, consequently, shows greater opposition to changes in direction. Anyone

who has tried to push a car out of mud or snow by rocking it back and forth knows that this is much easier with a

light car than a heavy car. The reason is that the more massive car shows greater opposition to changes in direction.

In summary, the natural vibrating frequency of a spring-and-mass system is controlled by mass and stiffness.

Frequency is directly proportional to stiffness (S↑F↑) and inversely proportional to mass (M↑F↓). It is important to

recognize that these rules apply to all objects, and not just simple spring-and-mass systems. For example, we will

see that the frequency of vibration of the vocal folds is controlled to a very large extent by muscular forces that act

to alter the mass and stiffness of the folds. We will also see that the frequency analysis that is carried out by the

inner ear depends to a large extent on a tuned membrane whose stiffness varies systematically from one end of the

cochlea to the other.

Sound Propagation

As was mentioned at the beginning of this chapter, the generation of a sound wave requires not only vibration,

but also an elastic medium in which the disturbance created by that vibration can be transmitted (see Box 3-1 [bell

jar experiment described in Patrick's science book - not yet written]). To say that air is an elastic medium means that

air, like all other matter, tends to return to its original shape after it is deformed through the application of a force.

The prototypical example of an object that exhibits this kind of restoring force is a spring. To understand the

mechanism underlying sound propagation, it is useful to think of air as consisting of collection of particles that are

connected to one another by springs, with the springs representing the restoring forces associated with the elasticity

of the medium. Air pressure is related to particle density. When a volume of air is undisturbed, the individual

particles of air distribute themselves more-or-less evenly, and the elastic forces are at their resting state. A volume of

air that is in this undisturbed state it is said to be at atmospheric pressure. For our purposes, atmospheric pressure

can be defined in terms of two interrelated conditions: (1) the air molecules are approximately evenly spaced, and

(2) the elastic forces, represented by the interconnecting springs, are neither compressed nor stretched beyond their

resting state. When a vibratory disturbance causes the air particles to crowd together (i.e., producing an increase in

particle density), air pressure is higher than atmospheric, and the elastic forces are in a compressed state.

Conversely, when particle spacing is relatively large, air pressure is lower than atmospheric.

1The example of tuning a guitar string is imperfect since the mass of the vibrating portion of the string decreases slightly as the string is

tightened. This occurs because a portion of the string is wound onto the tuning key as it is tightened.

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

TIME

Figure 3-7. Shown above is a highly schematic illustration of the chain reaction that

results in the propagation of a sound wave (modeled after Denes and Pinson, 1963).

The Physics of Sound 7

When a vibrating object is placed in an elastic medium, an air pressure disturbance is created through a chain

reaction similar to that illustrated in Figure 3-7. As the vibrating object (a tuning fork in this case) moves to the

right, particle a, which is immediately adjacent to the tuning fork, is displaced to the right. The elastic force

generated between particles a and b (not shown in the figure) has the effect a split second later of displacing particle

b to the right. This disturbance will eventually reach particles c, d, e, and so on, and in each case the particles will be

momentarily crowded together. This crowding effect is called compression or condensation, and it is characterized

by dense particle spacing and, consequently, air pressure that is slightly higher than atmospheric pressure. The

propagation of the disturbance is analogous to the chain reaction that occurs when an arrangement of dominos is

toppled over. Figure 3-7 also shows that at some close distance to the left of a point of compression, particle spacing

will be greater than average, and the elastic forces will be in a stretched state. This effect is called rarefaction, and

it is characterized by relatively wide particle spacing and, consequently, air pressure that is slightly lower than

atmospheric pressure.

The compression wave, along with the rarefaction wave that immediately follows it, will be propagated outward

at the speed of sound. The speed of sound varies depending on the average elasticity and density of the medium in

which the sound is propagated, but a good working figure for air is about 35,000 centimeters per second, or

approximately 783 miles per hour. Although Figure 3-7 gives a reasonably good idea of how sound propagation

works, it is misleading in two respects. First, the scale is inaccurate to an absurd degree: a single cubic inch of air

contains approximately 400 billion molecules, and not the handful of particles shown in the figure. Consequently,

the compression and rarefaction effects are statistical rather than strictly deterministic as shown in Figure 3-7.

Second, although Figure 3-7 makes it appear that the air pressure disturbance is propagated in a simple straight line

from the vibrating object, it actually travels in all directions from the source. This idea is captured somewhat better

in Figure 3-8, which shows sound propagation in two of the three dimensions in which the disturbance will be

transmitted. The figure shows rod and piston connected to a wheel spinning at a constant speed. Connected to the

piston is a balloon that expands and contracts as the piston moves in and out of the cylinder. As the balloon expands

the air particles are compressed; i.e., air pressure is momentarily higher than atmospheric. Conversely, when the

balloon contracts the air particles are sucked inward, resulting in rarefaction. The alternating compression and

rarefaction waves are propagated outward in all directions form the source. Only two of the three dimensions are

shown here; that is, the shape of the pressure disturbance is actually spherical rather than the circular pattern that is

shown here. Superimposed on the figure, in the graph labeled “one line of propagation,” is the resulting air pressure

waveform. Note that the pressure waveform takes on a high value during instants of compression and a low value

during instants of rarefaction. The figure also gives some idea of where the term uniform circular motion comes

from. If one were to make a graph plotting the height of the connecting rod on the rotating wheel as a function of

time it would trace out a perfect sinusoid; i.e., with exactly the shape of the pressure waveform that is superimposed

on the figure.

The Sound Pressure Waveform

Returning to Figure 3-7 for a moment, imagine that we chose some specific distance from the tuning fork to

observe how the movement and density of air particles varied with time. We would see individual air particles

oscillating small distances back and forth, and if we monitored particle density we would find that high particle

density (high air pressure) would be followed a moment later by relatively even particle spacing (atmospheric

pressure), which would be followed by a moment later by wide particle spacing (low air pressure), and so on.

Therefore, for an object that is vibrating sinusoidally, a graph showing variations in instantaneous air pressure

over time would also be sinusoidal. This is illustrated in Figure 3-9.

The vibratory patterns that have been discussed so far have all been sinusoidal. The concept of a sinusoid has

not been formally defined, but for our purposes it is enough to know that a sinusoid has precisely the smooth shape

that is shown in Figures such as 3-4 and 3-5. While sinusoids, also known as pure tones, have a very special place

in acoustic theory, they are rarely encountered in nature. The sound produced by a tuning fork comes quite close to a

sinusoidal shape, as do the simple tones that are used in hearing tests. Much more common in both speech and music

are more complex, nonsinusoidal patterns, to be discussed below. As will be seen in later chapters, these complex

vibratory patterns play a very important role in speech.

The Physics of Sound 8

The Frequency Domain

We now arrive at what is probably the single most important concept for understanding both hearing and speech

acoustics. The graphs that we have used up to this point for representing either vibratory motion or the air pressure

disturbance created by this motion are called time domain representations. These graphs show how instantaneous

displacement (or instantaneous air pressure) varies over time. Another method for representing either sound or

vibration is called a frequency domain representation, also known as a spectrum. There are, in fact, two kinds of

frequency domain representations that are used to characterize sound. One is called an amplitude spectrum (also

known as a magnitude spectrum or a power spectrum, depending on how the level of the signal is represented)

and the other is called a phase spectrum. For reasons that will become clear soon, the amplitude spectrum is by far

the more important of the two. An amplitude spectrum is simply a graph showing what frequencies are present with

what amplitudes. Frequency is given along the x axis and some measure of amplitude is given on the y axis. A phase

spectrum is a graph showing what frequencies are present with what phases.

Figure 3-10 shows examples of the amplitude and phase spectra for several sinusoidal signals. The top panel

shows a time-domain representation of a sinusoid with a period of 10 ms and, consequently, a frequency of 100 Hz

(f = 1/t = 1/0.01 sec = 100 Hz). The peak-to-peak amplitude for this signal is 400 μPa, and the signal has a phase of

90 o. Since the amplitude spectrum is a graph showing what frequencies are present with what amplitudes, the

amplitude spectrum for this signal will show a single line at 100 Hz with a height of 400 μPa. The phase spectrum is

a graph showing what frequencies are present with what phases, so the phase spectrum for this signal will show a

single line at 100 Hz with a height of 90 o

. The second panel in Figure 3-10 shows a 200 Hz sinusoid with a peak-to-

peak amplitude of 200 μPa and a phase of 180 o

. Consequently, the amplitude spectrum will show a single line at 200

Hz with a height of 100 μPa, while the phase spectrum will show a line at 200 Hz with a height of 180 o.

Complex Periodic Sounds

Sinusoids are sometimes referred to as simple periodic signals. The term "periodic" means that there is a

pattern that repeats itself, and the term "simple" means that there is only one frequency component present. This is

confirmed in the frequency domain representations in Figure 3-10, which all show a single frequency component in

both the amplitude and phase spectra. Complex periodic signals involve the repetition of a nonsinusoidal pattern,

and in all cases, complex periodic signals consist of more than a single frequency component. All nonsinusoidal

periodic signals are considered complex periodic.

Figure 3-8 Illustration of the propagation of a sound wave in two dimensions.

The Physics of Sound 9

Figure 3-11 shows several examples of complex periodic signals, along with the amplitude spectra for these signals.

The time required to complete one cycle of the complex pattern is called the fundamental period. This is precisely

the same concept as the term period that was introduced earlier. The only reason for using the term "fundamental

period" instead of the simpler term "period" for complex periodic signals is to differentiate the fundamental period

(the time required to complete one cycle of the pattern as a whole) from other periods that may be present in the

signal (e.g., more rapid oscillations that might be observed within each cycle). The symbol for fundamental period is

t o. Fundamental frequency (f o) is calculated from fundamental period using the same kind of formula that we used

earlier for sinusoids:

fo = 1/to

The signal in the top panel of Figure 3-11 has a fundamental period of 5 ms, so fo = 1/0.005 = 200 Hz.

Examination of the amplitude spectra of the signals in Figure 3-11 confirms that they do, in fact, consist of

more than a single frequency. In fact, complex periodic signals show a very particular kind of amplitude spectrum

called a harmonic spectrum. A harmonic spectrum shows energy at the fundamental frequency and at whole

number multiples of the fundamental frequency. For example, the signal in the top panel of Figure 3-11 has energy

present at 200 Hz, 400 Hz, 600 Hz, 800 Hz, 1,000 Hz, 1200 Hz, and so on. Each frequency component in the

0 5 10 15 20 25 30

-200

-100

0

100

200

Inst. Air Pressure

Period: 10 ms, Freq: 100 Hz, Amp: 400, Phase: 90

0 5 10 15 20 25 30

-200

-100

0

100

200

Inst. Air Pressure

Period: 5 ms, Freq: 200 Hz, Amp: 200, Phase: 180

0 5 10 15 20 25 30

-200

-100

0

100

200

Time (msec)

Inst. Air Pressure

Period: 2.5 ms, Freq: 400 Hz, Amp: 200, Phase: 270

TIME DOMAIN FREQUENCY DOMAIN

0 100 200 300 400 500

0

100

200

300

400

Frequency (Hz)

Amplitude

Amplitude Spectrum

0 100 200 300 400 500

0

100

200

300

400

Frequency (Hz)

Amplitude

0 100 200 300 400 500

0

100

200

300

400

Frequency (Hz)

Amplitude

0 100 200 300 400 500

0

90

180

270

360

Frequency (Hz)

Phase

Phase Spectrum

0 100 200 300 400 500

0

90

180

270

360

Frequency (Hz)

Phase

0 100 200 300 400 500

0

90

180

270

360

Frequency (Hz)

Phase

Figure 3-10. Time and frequency domain representations of three sinusoids. The frequency domain

consists of two graphs: an amplitude spectrum and a phase spectrum. An amplitude spectrum is a

graph showing what frequencies are present with what amplitudes, and a phase spectrum is a graph

showing the phases of each frequency component.

The Physics of Sound 10

amplitude spectrum of a complex periodic signal is called a harmonic (also known as a partial). The fundamental

frequency, in this case 200 Hz, is also called the first harmonic, the 400 Hz component (2 ⋅ fo) is called the second

harmonic, the 600 Hz component (3 ⋅ fo) is called the third harmonic, and so on.

The second panel in Figure 3-11 shows a complex periodic signal with a fundamental period of 10 ms and,

consequently, a fundamental frequency of 100 Hz. The harmonic spectrum that is associated with this signal will

therefore show energy at 100 Hz, 200 Hz, 300 Hz, 400 Hz, 500 Hz, and so on. The bottom panel of Figure 3-11

shows a complex periodic signal with a fundamental period of 2.5 ms, a fundamental frequency of 400 Hz, and

harmonics at 400, 800, 1200, 1600, and so on. Notice that there two completely interchangeable ways to define the

term fundamental frequency. In the time domain, the fundamental frequency is the number of cycles of the complex

pattern that are completed in one second. In the frequency domain, except in the case of certain special signals, the

fundamental frequency is the lowest harmonic in the harmonic spectrum. Also, the fundamental frequency defines

the harmonic spacing; that is, when the fundamental frequency is 100 Hz, harmonics will be spaced at 100 Hz

Figure 3-11. Time and frequency domain representations of three complex periodic signals.

Complex periodic signals have harmonic spectra, with energy at the fundamental frequency (f0) and

at whole number multiples of f0 (f0. 2, f0. 3, f0. 4, etc.) For example, the signal in the upper left, with a

fundamental frequency of 200 Hz, shows energy at 200 Hz, 400 Hz, 600 Hz, etc. In the spectra on

the right, amplitude is measured in arbitrary units. The main point being made in this figure is the

distribution of harmonic frequencies at whole number multiples of f0 for complex periodic signals.

0 5 10 15 20 25 30

-200

-100

0

100

200

Inst. Air Pres. (UPa) t0: 5 ms, f0: 200 Hz

t0: 10 ms, f0: 100 Hz

0 5 10 15 20 25 30

-200

-100

0

100

200

Inst. Air Pres. (UPa)

t0: 2.5 ms, f0: 400 Hz

0 5 10 15 20 25 30

-200

-100

0

100

200

Time (msec)

Inst. Air Pres. (UPa)

0 200 400 600 800 1000 1200 1400 1600

0

20

40

60

80

100

120

Frequency (Hz)

Amplitude

0 200 400 600 800 1000 1200 1400 1600

0

20

40

60

80

100

120

Frequency (Hz)

Amplitude

0 200 400 600 800 1000 1200 1400 1600

0

20

40

60

80

100

120

Frequency (Hz)

Amplitude

TIME DOMAIN FREQUENCY DOMAIN

The Physics of Sound 11

0 10 20 30 40 50

-200

-100

0

100

200

Inst. Air Pres. (UPa)

White Noise

/s/

0 10 20 30 40 50

-200

-100

0

100

200

Inst. Air Pres. (UPa)

/f/

0 10 20 30 40 50

-200

-100

0

100

200

TIME (msec)

Inst. Air Pres. (UPa)

0 1 2 3 4 5 6 7 8 9 10

0

20

40

60

80

100

Amplitude

0 1 2 3 4 5 6 7 8 9 10

0

20

40

60

80

100

Amplitude

0 1 2 3 4 5 6 7 8 9 10

0

20

40

60

80

100

Frequency (kHz)

Amplitude

TIME DOMAIN FREQUENCY DOMAIN

intervals (i.e., 100, 200, 300 ...), when the fundamental frequency is 125 Hz, harmonics will be spaced at 125 Hz

intervals (i.e., 125, 250, 375...), and when the fundamental frequency is 200 Hz, harmonics will be spaced at 200 Hz

intervals (i.e., 200, 400, 600 ...). (For some special signals this will not be the case.2) So, when fo is low, harmonics

will be closely spaced, and when fo is high, harmonics will be widely spaced. This is clearly seen in Figure 3-11: the

signal with the lowest f0 (100 Hz, the middle signal) shows the narrowest harmonic spacing, while the signal with

the highest f0 (400 Hz, the bottom signal) shows the widest harmonic spacing.

There are certain characteristics of the spectra of complex periodic sounds that can be determined by making simple

measurements of the time domain signal, and there are certain other characteristics that require a more complex

analysis. For example, simply by examining the signal in the bottom panel of Figure 3-11 we can determine that it is

complex periodic (i.e., it is periodic but not sinusoidal) and therefore it will show a harmonic spectrum with energy

at whole number multiples of the fundamental frequency. Further, by measuring the fundamental period (2.5 ms)

2There are some complex periodic signals that have energy at odd multiples of the fundamental frequency only. A square wave, for

example, is a signal that alternates between maximum positive amplitude and maximum negative amplitude. The spectrum of square wave shows

energy at odd multiples of the fundamental frequency only. Also, a variety of simple signal processing tricks can be used to create signals with

harmonics at any arbitrary set of frequencies. For example, it is a simple matter to create a signal with energy at 400, 500, and 600 Hz only.

While these kinds of signals can be quite useful for conducting auditory perception experiments, it remains true that most naturally occurring

complex periodic signals have energy at all whole number multiples of the fundamental frequency.

Figure 3-12. Time and frequency domain representations of three non-transient complex aperiodic

signals. Unlike complex periodic signals, complex aperiodic signals show energy that is spread

across the spectrum. This type of spectrum is called dense or continuous. These spectra have a very

different appearance from the “picket fence” look that is associated with the discrete, harmonic

spectra of complex periodic signals.

The Physics of Sound 12

and converting it into fundamental frequency (400 Hz), we are able to determine that the signal will have energy at

400, 800, 1200, 1600, etc. But how do we know the amplitude of each of these frequency components? And how do

we know the phase of each component? The answer is that you cannot determine harmonic amplitudes or phases

simply by inspecting the signal or by making simple measurements of the time domain signals with a ruler. We will

see soon that a technique called Fourier analysis is able to determine both the amplitude spectrum and the phase

spectrum of any signal. We will also see that the inner ears of humans and many other animals have developed a

trick that is able to produce a neural representation that is comparable in some respects to an amplitude spectrum.

We will also see that the ear has no comparable trick for deriving a representation that is equivalent to a phase

spectrum. This explains why the amplitude spectrum is far more important for speech and hearing applications than

the phase spectrum. We will return to this point later.

To summarize: (1) a complex periodic signal is any periodic signal that is not sinusoidal, (2) complex periodic

signals have energy at the fundamental frequency (fo) and at whole number multiples of the fundamental frequency

(2 ⋅ fo, 3 ⋅ fo , 4 ⋅ fo ...), and (3) although measuring the fundamental frequency allows us to determine the frequency

locations of harmonics, there is no simple measurement that can tell us harmonic amplitudes or phases. For this,

Fourier analysis or some other spectrum analysis technique is needed.

Figure 3-13. Time and frequency domain representations of three transients. Transients are complex

aperiodic signals that are defined by their brief duration. Pops, clicks, and the sound gun fire are

examples of transients. In common with longer duration complex aperiodic signals, transients show

dense or continuous spectra, very unlike the discrete, harmonic spectra associated with complex periodic

d

0 10 20 30 40 50 60 70 80 90 100

-200

-100

0

100

200

Inst. Amp. (UPa)

Rap on Desk

Clap

0 10 20 30 40 50 60 70 80 90 100

-200

-100

0

100

200

Inst. Amp. (UPa)

Tap on Cheek

0 10 20 30 40 50 60 70 80 90 100

-200

-100

0

100

200

TIME (msec)

Inst. Amp. (UPa)

0 1 2 3 4 5

0

20

40

60

80

100

Amplitude

0 1 2 3 4 5

0

20

40

60

80

100

Amplitude

0 1 2 3 4 5

0

20

40

60

80

100

Frequency (kHz)

Amplitude

TIME DOMAIN FREQUENCY DOMAIN

The Physics of Sound 13

-200

0

200

Inst. Air Pres.

(a)

-200

0

200

Inst. Air Pres.

(b)

-200

0

200

Inst. Air Pres.

(c)

-200

0

200

Inst. Air Pres.

(d)

-300

0

300

Inst. Air Pres.

(e)

Time ->

Aperiodic Sounds

An aperiodic sound is any sound that does not show a repeating pattern in its time domain representation. There are

many aperiodic sounds in speech. Examples include the hissy sounds associated with fricatives such as /f/ and /s/,

and the various hisses and pops associated with articulatory release for the stop consonants /b,d,g,p,t,k/. Examples of

non-speech aperiodic sounds include a drummer's cymbal or snare drum, the hiss produced by a radiator, and static

sound produced by a poorly tuned radio. There are two types of aperiodic sounds: (1) continuous aperiodic sounds

(also known as noise) and (2) transients. Although there is no sharp cutoff, the distinction between continuous

aperiodic sounds and transients is based on duration. Transients (also "pops" and "clicks") are defined by their very

brief duration, and continuous aperiodic sounds are of longer duration. Figure 3-12 shows several examples of time

domain representations and amplitude spectra for continuous aperiodic sounds. The lack of periodicity in the time

Figure 3-14. Illustration of the principle underlying Fourier analysis. The complex periodic signal

shown in panel e was derived by point-for-point summation of the sinusoidal signals shown in

panels a-d. Point-for-point summation simply means beginning at time zero (i.e., the start of the

signal) and adding the instantaneous amplitude of signal a to the instantaneous amplitude of signal b

at time zero, then adding that sum to the instantaneous amplitude of signal c, also at time zero, then

adding that sum to instantaneous amplitude of signal d at time zero. The sum of instantaneous

amplitudes at time zero of signals a-d is the instantaneous amplitude of the composite signal e at

time zero. For example, at time zero the amplitudes of sinusoids a-d are 0, +100, -200, and 0,

respectively, producing a sum of -100. This agrees with the instantaneous amplitude at the very

beginning of composite signal e. The same summation procedure is followed for all time points.

The Physics of Sound 14

domain is quite evident; that is, unlike the periodic sounds we have seen, there is no pattern that repeats itself over

time.

All aperiodic sounds -- both continuous and transient -- are complex in the sense that they always consist of

energy at more than one frequency. The characteristic feature of aperiodic sounds in the frequency domain is a

dense or continuous spectrum, which stands in contrast to the harmonic spectrum that is associated with complex

periodic sounds. In a harmonic spectrum, there is energy at the fundamental frequency, followed by a gap with little

or no energy, followed by energy at the second harmonic, followed by another gap, and so on. The spectra of

aperiodic sounds do not share this "picket fence" appearance. Instead, energy is smeared more-or-less continuously

across the spectrum. The top panel in Figure 3-12 shows a specific type of continuous aperiodic sound called white

noise. By analogy to white light, white noise has a flat amplitude spectrum; that is, approximately equal amplitude at

all frequencies. The middle panel in Figure 3-12 shows the sound /s/, and the bottom panel shows sound /f/. Notice

that the spectra for all three sounds are dense; that is, they do not show the "picket fence" look that reveals harmonic

structure. As was the case for complex periodic sounds, there is no way to tell how much energy there will be at

different frequencies by inspecting the time domain signal or by making any simple measures with a ruler. Likewise,

there is no simple way to determine the phase spectrum. So, after inspecting a time-domain signal and determining

that it is aperiodic, all we know for sure is that it will have a dense spectrum rather than a harmonic spectrum.

Figure 3-13 shows time domain representations and amplitude spectra for three transients. The transient in the

top panel was produced by rapping on a wooden desk, the second is a single clap of the hands, and the third was

produced by holding the mouth in position for the vowel /o/, and tapping the cheek with an index finger. Note the

brief durations of the signals. Also, as with continuous aperiodic sounds, the spectra associated with transients are

dense; that is, there is no evidence of harmonic organization. In speech, transients occur at the instant of articulatory

release for stop consonants. There are also some languages, such as the South African languages Zulu, Hottentot,

and Xhosa, that contain mouth clicks as part of their phonemic inventory (MacKay, 1986). Fourier Analysis

TIME DOMAIN

Time ->

Inst. Air Pres.

Fourier

Analyzer

0 200 400 600 800

Frequency (Hz)

Amplitude

FREQUENCY DOMAIN

0 200 400 600 800

Frequency (Hz)

Phase

Figure 3-15. A signal enters a Fourier analyzer in the time domain and exits in the frequency domain.

As outputs, the Fourier analyzer produces two frequency-domain representations: an amplitude

spectrum that shows the amplitude of each sinusoidal component that is present in the input signal, and

a phase spectrum that shows the phase of each of the sinusoids. The input signal can be reconstructed

perfectly by summing sinusoids at frequencies, amplitudes, and phase that are shown in the Fourier

amplitude and phase spectra, using the summing method that is illustrated in Figure 3-14..

The Physics of Sound 15

Fourier analysis is an extremely powerful tool that has widespread applications in nearly every major branch

of physics and engineering. The method was developed by the 19 th century mathematician Joseph Fourier, and

although Fourier was studying thermal waves at the time, the technique can be applied to the frequency analysis of

any kind of wave. Fourier's great insight was the discovery that all complex waves can be derived by adding

sinusoids together, so long as the sinusoids are of the appropriate frequencies, amplitudes, and phases. For example,

the complex periodic signal at the bottom of Figure 3-14 can be derived by summing sinusoids at 100, 200, 300, and

400 Hz, with each sinusoidal component having the amplitude and phase that is shown in the figure (see the caption

of Figure 3-14 for an explanation of what is meant by summing the sinusoidal components). The assumption that all

complex waves can be derived by adding sinusoids together is called Fourier's theorem, and the analysis technique

that Fourier developed from this theorem is called Fourier analysis. Fourier analysis is a mathematical technique that

takes a time domain signal as its input and determines: (1) the amplitude of each sinusoidal component that is

present in the input signal, and (2) the phase of each sinusoidal component that is present in the input signal.

Another way of stating this is that Fourier analysis takes a time domain signal as its input and produces two

frequency domain representations as output: (1) an amplitude spectrum, and (2) a phase spectrum.

The basic concept is illustrated in Figure 3-15, which shows a time domain signal entering the Fourier analyzer.

Emerging at the output of the Fourier analyzer is an amplitude spectrum (a graph showing the amplitude of each

sinusoid that is present in the input signal) and a phase spectrum (a graph showing the phase of each sinusoid that is

present in the input signal). The amplitude spectrum tells us that the input signal contains: (1) 200 Hz sinusoid with

an amplitude of 100 μPa, a 400 Hz sinusoid with an amplitude of 200 μPa, and a 600 Hz sinusoid with an amplitude

of 50 μPa. Similarly, the phase spectrum tells us that the 200 Hz sinusoid has a phase of 90 o, the 400 Hz sinusoid

has a phase of 180 o, and the 600 Hz sinusoid has a phase of 270 o. If Fourier's theorem is correct, we should be able

to reconstruct the input signal by summing sinusoids at 200, 400, and 600 Hz, using the amplitudes and phases that

are shown. In fact, summing these three sinusoids in this way would precisely reproduce the original time domain

signal; that is, we would get back an exact replica of our original signal, and not just a rough approximation to it.

For our purposes it is not important to understand how Fourier analysis works. The most important point about

Fourier's idea is that, visual appearances aside, all complex waves consist of sinusoids of varying frequencies,

amplitudes, and phases. In fact, Fourier analysis applies not only to periodic signals such as those shown in Figure

3-15, but also to noise and transients. In fact, the amplitude spectra of the aperiodic signals shown in Figure 3-13

were calculated using Fourier analysis. In later chapters we will see that the auditory system is able to derive a

neural representation that is roughly comparable to a Fourier amplitude spectrum. However, as was mentioned

earlier, the auditory system does not derive a representation comparable to a Fourier phase spectrum. As a result,

listeners are very sensitive to changes in the amplitude spectrum but are relatively insensitive to changes in phase.

Some Additional Terminology

Overtones vs. Harmonics: The term overtone and the term harmonic refer to the same concept; they are just

counted differently. As we have seen, in a harmonic series such as 100, 200, 300, 400, etc., the 100 Hz component

can be referred to as either the fundamental frequency or the first harmonic; the 200 Hz component is the second

harmonic, the 300 Hz component is the third harmonic, and so on. An alternative set of terminology would refer to

the 100 Hz component as the fundamental frequency, the 200 Hz component as the first overtone, the 300 Hz

component as the second overtone, and so on. Use of the term overtone tends to be favored by those interested in

musical acoustics, while most other acousticians tend to use the term harmonic.

Octaves vs. Harmonics: An octave refers to a doubling of frequency. So, if we begin at 100 Hz, the next octave up

would 200 Hz, the next would be 400 Hz, the next would be 800 Hz, and so on. Note that this is quite different from

a harmonic progression. A harmonic progression beginning at 300 Hz would be 300, 600, 900, 1200, 1500, etc.,

while an octave progression would be 300, 600, 1200, 2400, 4800, etc. There is something auditorilly natural about

octave spacing, and octaves play a very important role in the organization of musical scales. For example, on a piano

keyboard, middle A (A5) is 440 Hz, A above middle A (A 6) is 880 Hz, A7 is 1,760 and so on. (See Box 3-2).

Wavelength: The concept of wavelength is best illustrated with an example given by Small (1973). Small asks us

to imagine dipping a finger repeatedly into a puddle of water at a perfectly regular interval. Each time the finger hits

the water, a wave is propagated outward, and we would see a pattern formed consisting of a series of concentric

The Physics of Sound 16

circles (see Figure 3-16). Wavelength is simply the distance between the adjacent waves. Precisely the same concept

can be applied to sound waves: wavelength is simply the distance between one compression wave and the next (or

one rarefaction wave and the next or, more generally, the distance between any two corresponding points in adjacent

waves). For our purposes, the most important point to be made about wavelength is that there is a simple

relationship between frequency and wavelength. Using the puddle example, imagine that we begin by dipping our

finger into the puddle at a very slow rate; that is, with a low "dipping frequency." Since the waves have a long

period of time to travel from one dip to the next, the wavelength will be large. By the same reasoning, the

wavelength becomes smaller as the "dipping frequency" is increased; that is, the time allowed for the wave to travel

at high "dipping frequency" is small, so the wavelength is small. Wavelength is a measure of distance, and the

formula for calculating wavelength is a straightforward algebraic rearrangement of the familiar "distance = rate ⋅

time" formula from junior high school.

λ = c/f, where: λ = wavelength

c = the speed of sound

f = frequency

By rearranging the formula, frequency can be calculated if wavelength and the speed of sound are known:

f = c/λ

Lower Frequency

(Longer Wavelength)

Higher Frequency

(Shorter Wavelength)

Figure 3-16. Wavelength is a measure of the distance between the crest of one cycle of a wave and the

crest of the next cycle (or trough to trough or, in fact, the distance between any two corresponding

points in the wave). Wavelength and frequency are related to one another. Because the wave has only a

short time to travel from one cycle to the next, high frequencies produce short wavelengths.

Conversely, because of the longer travel times, low frequencies produce long wavelengths.

The Physics of Sound 17

Spectrum Envelope: The term spectrum envelope refers to an imaginary smooth line drawn to enclose an

amplitude spectrum. Figure 3-17 shows several examples. This is a rather simple concept that will play a very

important role in understanding certain aspects of auditory perception. For example, we will see that our perception

of a perceptual attribute called timbre (also called sound quality) is controlled primarily by the shape of the

spectrum envelope, and not by the fine details of the amplitude spectrum. The examples in Figure 3-17 show how

differences in spectrum envelope play a role in signaling differences in one specific example of timbre called

vowel quality (i.e., whether a vowel sounds like /i/ vs. /a/ vs. /u/, etc.). For example, panels a and b in Figure 3-17

show the vowel /å/ produced at two different fundamental frequencies. (We know that the fundamental frequencies

are different because one spectrum shows wide harmonic spacing and the other shows narrow harmonic spacing.)

The fact that the two vowels are heard as /a/ despite the difference in fundamental frequency can be attributed to the

fact that these two signals have similar spectrum envelopes. Panels c and d in Figure 3-17 show the spectra of two

signals with different spectrum envelopes but the same fundamental frequency (i.e., with the same harmonic

spacing). As we will see in the chapter on auditory perception, differences in fundamental frequency are perceived

as differences in pitch. So, for signals (a) and (b) in Figure 3-17, the listener will hear the same vowel produced at

two different pitches. Conversely, for signals (c) and (d) in Figure 3-17, the listener will hear two different vowels

produced at the same pitch. We will return to the concept of spectrum envelope in the chapter on auditory

perception.

Amplitude Envelope: The term amplitude envelope refers to an imaginary smooth line that is drawn on top of a

time domain signal. Figure 3-18 shows sinusoids that are identical except for their amplitude envelopes. It can be

seen that the different amplitude envelopes reflect differences in the way the sounds are turned on and off. For

example, panel a shows a signal that is turned on abruptly and turned off abruptly; panel b shows a signal that is

turned on gradually and turned off abruptly; and so on. Differences in amplitude envelope have an important effect

on the quality of a sound. As we will see in the chapter on auditory perception, amplitude envelope, along with

spectrum envelope discussed above, is another physical parameter that affects timbre or sound quality. For

0 1 2 3

0

10

20

30

40

50

60

70

Frequency (kHz)

Amplitude

(a) Vowel: /a/, f0: 100 Hz

0 1 2 3

0

10

20

30

40

50

60

70

Frequency (kHz)

Amplitude

(b) Vowel: /a/, f0: 200 Hz

0 1 2 3

0

10

20

30

40

50

60

70

Frequency (kHz)

Amplitude

(c)

Vowel: /i/, f0: 150 Hz

0 1 2 3

0

10

20

30

40

50

60

70

Frequency (kHz)

Amplitude

(d)

Vowel: /u/, f0: 150 Hz

Figure 3-17. A spectrum envelope is an imaginary smooth line drawn to enclose an amplitude

spectrum. Panels a and b show the spectra of two signals (the vowel /ɑ/) with different fundamental

frequencies (note the differences in harmonic spacing) but very similar spectrum envelopes. Panels c

and d show the spectra of two signals with different spectrum envelopes (the vowels /i/ and /u/ in this

case) but the same fundamental frequencies (i.e., the same harmonic spacing).

The Physics of Sound 18

example, piano players know that a given note will sound different depending on whether or not the damping pedal

is used. Similarly, notes played on a stringed instrument such as a violin or cello will sound different depending on

whether the note is plucked or bowed. In both cases, the underlying acoustic difference is amplitude envelope.

Acoustic Filters

As will be seen in subsequent chapters, acoustic filtering plays a central role in the processing of sound by the

inner ear. The human vocal tract also serves as an acoustic filter that modifies and shapes the sounds that are created

by the larynx and other articulators. For this reason, it is quite important to understand how acoustic filters work. In

the most general sense, the term filter refers to a device or system that is selective about the kinds of things that are

allowed to pass through versus the kinds of things that are blocked. An oil filter, for example, is designed to allow

oil to pass through while blocking particles of dirt. Of special interest to speech and hearing science are frequency

selective filters. These are devices that allow some frequencies to pass through while blocking or attenuating other

frequencies. (The term attenuate means to weaken or reduce in amplitude).

A simple example of a frequency selective filter from the world of optics is a pair of tinted sunglasses. A piece

of white paper that is viewed through red tinted sunglasses will appear red. Since the original piece of paper is

white, and since we know that white light consists of all of the visible optical frequencies mixed in equal amounts,

the reason that the paper appears red through the red tinted glasses is that optical frequencies other than those

corresponding to red are being blocked or attenuated by the optical filter. As a result, it is primarily the red light that

is being allowed to pass through. (Starting at the lowest optical frequency and going to the highest, light will appear

red, orange, yellow, green, blue, indigo, and violet.)

Inst. Air Pres.

(a)

Signals Differing in Amplitude Envelope

Inst. Air Pres.

(b)

Inst. Air Pres.

(c)

Time ->

Inst. Air Pres.

(d)

Figure 3-18. Amplitude envelope is an imaginary smooth line drawn to enclose a time-domain signal.

This feature describes how a sound is turned on and turned off; for example, whether the sound is

turned on abruptly and turned off abruptly (panel a), turned on gradually and turned off abruptly (panel

b), turned on abruptly and turned off gradually (panel c), or turned on and off gradually (panel d).

The Physics of Sound 19

A graph called a frequency response curve is used to describe how a frequency selective filter will behave. A

frequency response curve is a graph showing how energy at different frequencies will be affected by the filter.

Specifically, a frequency response curve plots a variable called "gain" as a function of variations in the frequency of

the input signal. Gain is the amount of amplification provided by the filter at different signal frequencies. Gains are

interpreted as amplitude multipliers; for example, suppose that the gain of a filter at 100 Hz is 1.3. If a 100 Hz

sinusoid enters the filter measuring 10 uPa, the amplitude at the output of the filter at 100 Hz will measure 13 μPa

(10 μPa x 1.3 = 13 μPa). The only catch in this scheme is that gains can and very frequently are less than 1, meaning

that the effect of the filter will be to attenuate the signal. For example, if the gain at 100 Hz is 0.5, a 10 μPa input

signal at 100 Hz will measure 5 μPa at the output of the filter. When the filter gain is 1.0, the signal is unaffected by

the filter; i.e., a 10 μPa input signal will measure 10 μPa at the output of the filter.

Figure 3-19 shows frequency response curves for several optical filters. Panel a shows a frequency response

curve for the red optical filter discussed in the example above. If we put white light into the filter in panel a, the

signal amplitude at the output of the filter will be high only when the frequency of the input signal is low. This is

because the gain of the filter is high only in the low-frequency portion of the frequency-response curve. This is an

example of a lowpass filter; that is, a filter that allows low frequencies to pass through. Panel b shows an optical

filter that has precisely the reverse effect on an input signal; that is, this filter will allow high frequencies to pass

through while attenuating low- and mid-frequency signals. A white surface viewed through this filter would

therefore appear violet. This is an example of a highpass filter. Panel c shows the frequency response curve for a

filter that allows a band of energy in the center of the spectrum to pass through while attenuating signal components

of higher and lower frequency. A white surface viewed through this filter would appear green. This is called a

bandpass filter.

Acoustic filters do for sound exactly what optical filters do for light; that is, they allow some frequencies to pass

through while attenuating other frequencies. To get a better idea of how a frequency response curve is measured,

imagine that we ask a singer to attempt to shatter a crystal wine glass with a voice signal alone. To see how the

frequency response curve is created we have to make two rather unrealistic assumptions: (1) we need to assume that

the singer is able to produce a series of pure tones of various frequencies (the larynx, in fact, produces a complex

periodic sound and not a sinusoid), and (2) the amplitudes of these pure tones are always exactly the same. The wine

glass will serve as the filter whose frequency response curve we wish to measure. As shown in Figure 3-20, we

attach a vibration meter to the wine glass, and the reading on this meter will serve as our measure of output

Figure 3-19. Frequency response curves for three optical filters. The lowpass filter on the left allows

low frequencies to pass through, while attenuating or blocking optical energy at higher frequencies.

The highpass filter in the middle has the opposite effect, allowing high frequencies to pass through,

while attenuating or blocking optical energy at lower frequencies. The bandpass filter on the right

allows a band of optical frequencies in the center of the spectrum to pass through, while attenuating or

blocking energy at higher and lower frequencies.

The Physics of Sound 20

amplitude for the filter. For the purpose of this example, will assume that the signal frequency needed to break the

glass is 500 Hz. We now ask the singer to produce a low frequency signal, say 50 Hz. Since this frequency is quite

remote from the 500 Hz needed to break the glass, the output amplitude measured by the vibration meter will be

quite low. As the singer gets closer and closer to the required 500 Hz, the measured output amplitude will increase

systematically until the glass finally breaks. If we assume that the glass does not break but rather reaches a

maximum amplitude just short of that required to shatter the glass, we can continue our measurement of the

frequency response curve by asking the singer to produce signals that are increasingly high in frequency. We would

find that the output amplitude would become lower and lower the further we got from the 500 Hz natural vibrating

frequency of the wine glass. The pattern that is traced by our measures of output amplitude at each signal frequency

would resemble the frequency response curve we saw earlier for green sunglasses; that is, we would see the

frequency response curve for a bandpass filter.

Additional Comments on Filters

Cutoff Frequency, Center Frequency, Bandwidth. The top panel of Figure 3-21 shows frequency response curves

for two lowpass filters that differ in a parameter called cutoff frequency. Both filters allow low frequencies to pass

through while attenuating high frequencies; the filters differ only in the frequency at which the attenuation begins.

The bottom panel of Figure 3-21 shows two highpass filters that differ in cutoff frequency. There are two additional

terms that apply only to bandpass filters. In our wineglass example above, the natural vibrating frequency of the

wine glass was 300 Hz. For this reason, when the frequency response curve is measured, we find that the wine glass

reaches its maximum output amplitude at 300 Hz. This is called the center frequency or resonance of the filter. It

is possible for two bandpass filters to have the same center frequency but differ with respect to a property called

Figure 3-20. Illustration of how the frequency response curve of a crystal wine glass

might be measured. Our singer produces a series of sinusoids that are identical in

amplitude but cover a wide range of frequencies. (This part of the example is

unrealistic: the human larynx produces a complex sound rather than a sinusoid.) The

gain of the wine glass filter can be traced out by measuring the amplitude of

vibration at the different signal frequencies.)

The Physics of Sound 21

bandwidth. Figure 3-22 shows two filters that differ in bandwidth. The tall, thin frequency response curve describes

a narrow band filter. For this type of filter, output amplitude reaches a very sharp peak at the center frequency and

drops off abruptly on either side of the peak. The other frequency response curve describes a wide band filter (also

called broad band). For the wide band filter, the peak that occurs at the resonance of the filter is less sharp and the

drop in output amplitude on either side of the center frequency is more gradual.

Fixed vs. Variable Filters. A fixed filter is a filter whose frequency response curve cannot be altered. For example,

an engineer might design a lowpass filter that attenuates at frequencies above 500 Hz, or a bandpass filter that passes

with a center frequency of 1,000 Hz. It is also possible to create a filter whose characteristics can be varied. For

example, the tuning dial on a radio controls the center frequency of a narrow bandpass filter that allows a single

radio channel to pass through while blocking channels at all other frequencies. The human vocal tract is an example

0 1000 2000 3000 4000

0.0

0.2

0.4

0.6

0.8

1.0

Frequency (Hz)

Gain

Lowpass Filters with Different

Cutoff Frequencies

0 1000 2000 3000 4000

0.0

0.2

0.4

0.6

0.8

1.0

Frequency (Hz)

Gain

Highpass Filters with Different

Cutoff Frequencies

Figure 3-21. Lowpass and highpass filters differing in cutoff frequency.

0 1000 2000 3000 4000

0.0

0.2

0.4

0.6

0.8

1.0

Frequency (Hz)

Gain

Bandpass Filters Differing

in Bandwidth

Narrow Band Filter

Wide Band Filter

Figure 3-22. Frequency response curves for two bandpass filters with identical center

frequencies but different bandwidths. Both filters pass a band of energy centered around

2000 Hz, but the narrow band filter is more selective than the wide band filter; that is,

gain decreases at a higher rate above and below the center frequency for the narrow band

filter than for the wide band filter

The Physics of Sound 22

of a variable filter of the most spectacular sort. For example: (1) during the occlusion interval that occurs in the

production of a sound like /b/, the vocal tract behaves like a lowpass filter; (2) in the articulatory posture for sounds

like /s/ and /sh/ the vocal tract behaves like a highpass filter; and (3) in the production of vowels, the vocal tract

behaves like a series of bandpass filters connected to one another, and the center frequencies of these filters can be

adjusted by changing the positions of the tongue, lips, and jaw. To a very great extent, the production of speech

involves making adjustments to the articulators that have the effect of setting the vocal tract filter in differ modes to

produce the desired sound quality. We will have much more to say about this in later chapters.

Frequency Response Curves vs. Amplitude Spectra. It is not uncommon for students to confuse a frequency

response curve with an amplitude spectrum. The axis labels are rather similar: an amplitude spectrum plots

amplitude on the y axis and frequency on the x axis, while a frequency response curve plots gain on the y axis and

frequency on the x axis. The apparent similarities are deceiving, however, since a frequency response curve and an

amplitude spectrum display very different kinds of information. The difference is that an amplitude spectrum

describes a sound while a frequency response curve describes a filter. For any given sound wave, an amplitude

spectrum tells us what frequencies are present with what amplitudes. A frequency response curve, on the other hand,

describes a filter, and for that filter, it tells us what frequencies will be allowed to pass through and what frequencies

will be attenuated. Keeping these two ideas separate will be quite important for understanding the key role played by

filters in both hearing and speech science.

Resonance

The concept of resonance has been alluded to on several occasions but has not been formally defined. The term

resonance is used in two different but very closely related ways. The term resonance refers to: (1) the phenomenon

of forced vibration, and (2) natural vibrating frequency (also resonant frequency or resonance frequency) To

gain an appreciation for both uses of this term, imagine the following experiment. We begin with two identical

tuning forks, each tuned to 435 Hz. Tuning fork A is set into vibration and placed one centimeter from tuning fork

B, but not touching it. If we now hold tuning fork B to a healthy ear, we will find that it is producing a 435 Hz tone

that is faint but quite audible, despite the fact that it was not struck and did not come into physical contact with

tuning fork A. The explanation for this "action-at-a-distance" phenomenon is that the sound wave generated by

tuning fork A forces tuning fork B into vibration; that is, the series of compression and rarefaction waves will

alternately push and pull the tuning fork, resulting in vibration at the frequency being generated by tuning fork A.

The phenomenon of forced vibration is not restricted to this "action-at-a-distance" case. The same effect can be

demonstrated by placing a vibrating tuning fork in contact with a desk or some other hard surface. The intensity of

the signal will increase dramatically because the tuning fork is forcing the desk to vibrate, resulting in a larger

volume of air being compressed and rarefied.3

Returning to our original tuning fork experiment, suppose that we repeat this test using two mismatched tuning

forks; for example, tuning fork A with a natural frequency of 256 Hz and tuning fork B with a natural vibrating

frequency of 435 Hz. If we repeat the experiment – setting tuning fork A into vibration and holding it one centimeter

from tuning fork B – we will find that tuning fork B does not produce an audible tone. The reason is that forced

vibration is most efficient when the frequency of the driving force is closest to the natural vibration frequency of the

object that is being forced to vibrate. Another way to think about this is that tuning fork B in these experiments is

behaving like a filter that is being driven by the signal produced by tuning fork A. Tuning forks, in fact, behave like

rather narrow bandpass filters. In the experiment with matched tuning forks, the filter was being driven by a signal

frequency corresponding to the peak in the filter's frequency response curve. Consequently, the filter produced a

great deal of energy at its output. In the experiment with mismatched tuning forks, the filter is being driven by a

signal that is remote from the peak in the filter's frequency response curve, producing a low amplitude output signal.

To summarize, resonance refers to the ability of one vibrating system to force another system into vibration.

Further, the amplitude of this forced vibration will be greater as the frequency of the driving force approaches the

natural vibrating frequency (resonance) of the system that is being forced into vibration.

3The increase in intensity that would occur as the tuning fork is placed in contact with a hard surface does not mean that additional energy is

created. The increase in intensity would be offset by a decrease in the duration of the tone, so the total amount of energy would not increase

relative to a freely vibrating tuning fork.

The Physics of Sound 23

Cavity Resonators

An air-filled cavity exhibits frequency selective properties and should be considered a filter in precisely the way

that the tuning forks and wine glasses mentioned above are filters. The human vocal tract is an air-filled cavity that

behaves like a filter whose frequency response curve varies depending on the positions of the articulators. Tuning

forks and other simple filters have a single resonant frequency. (Note that we will be using the terms "natural

vibrating frequency" and "resonant frequency" interchangeably.) Cavity resonators, on the other hand, can have an

infinite number of resonant frequencies.

A simple but very important cavity resonator is the uniform tube. This is a tube whose cross-sectional area is

the same (uniform) at all points along its length. A simple water glass is an example of a uniform tube. The method

for determining the resonant frequency pattern for a uniform tube will vary depending on whether the tube is closed

at both ends, open at both ends, or closed at just one end. The configuration that is most directly applicable to

problems in speech and hearing is the uniform tube that is closed at one end and open at the other end. The ear canal,

for example, is approximately uniform in cross-sectional area and is closed medially by the ear drum and open

0.0

0.2

0.4

0.6

0.8

1.0

Gain

500 1500 2500 3500 4500

17.5 cm Uniform Tube

0.0

0.2

0.4

0.6

0.8

1.0

Gain

437.5 1312.5 2187.5 3062.5 3937.5

20 cm Uniform Tube

0 1000 2000 3000 4000 5000

0.0

0.2

0.4

0.6

0.8

1.0

Frequency (Hz)

Gain

583.3 1750.0 2916.7 4083.3 5225.0

15 cm Uniform Tube

Figure 3-23. Frequency response curves for three uniform tubes open at one end and closed at the

other. These kinds of tubes have an infinite number of resonances at odd multiples of the lowest

resonance. As the figure shows, shortening the tube shifts all resonances to higher frequencies while

lengthening the tube shifts all resonances to lower frequencies.

The Physics of Sound 24

laterally. Also, in certain configurations the vocal tract is approximately uniform in cross-sectional area and is

effectively closed from below by the vocal folds and open at the lips. The resonant frequencies for a uniform tube

closed at one end are determined by its length. The lowest resonant frequency (F1) for this kind of tube is given by:

F1 = c/4L, where: c = the speed of sound

L = the length of the tube

For example, for a 17.5 cm tube, F1 = c/4L = 35000/70 = 500 Hz. This tube will also have an infinite number of

higher frequency resonances at odd multiples of the lowest resonance:

F1 = F1 . 1 = 500 Hz

F2 = F1 ⋅ 3 = 1,500 Hz

F3 = F1 . 5 = 2,500 Hz

F4 = F1 ⋅ 7 = 3,500 Hz

The frequency response curve for this tube for frequencies below 4000 Hz is shown in the solid curve in Figure

3-23. Notice that the frequency response curve shows peaks at 500, 1500, 2500, and 3500 Hz, and valleys in

between these peaks. The frequency response curve, in fact, looks like a number of bandpass filters connected in

series with one another. It is important to appreciate that what we have calculated here is a series of natural vibrating

frequencies of a tube. What this means is that the tube will respond best to forced vibration if the tube is driven by

signals with frequencies at or near 500 Hz, 1500 Hz, 2500 Hz, and so on. Also, the resonant frequencies that were

just calculated should not be confused with harmonics. Harmonics are frequency components that are present in the

amplitude spectra of complex periodic sounds; resonant frequencies are peaks in the frequency response curve of

filters.

We next need to see what will happen to the resonant frequency pattern of the tube when the tube length

changes. If the tube is lengthened to 20 cm:

F1 = c/4L = 35,000/80 = 437.5 Hz

F2 = F1 ⋅ 3 = 1,312.5 Hz

F3 = F1 ⋅ 5 = 2,187.5 Hz

F4 = F1 ⋅ 7 = 3,062.5 Hz

It can be seen that lengthening the tube from 17.5 cm to 20 cm has the effect of shifting all of the resonant

frequencies downward (see Figure 3-23). Similarly, shortening the tube has the effect of shifting all of the resonant

frequencies upward. For example, the resonant frequency pattern for a 15 cm tube would be:

F1 = c/4L = 35,000/60 = 583.3 Hz

F2 = F1⋅ 3 = 1,750 Hz

F3 = F1 ⋅ 5 = 2,916.7 Hz

F4 = F1 ⋅ 7 = 4,083.3 Hz

The general rule is quite simple: all else being equal, long tubes have low resonant frequencies and short tubes

have high resonant frequencies. This can be demonstrated easily by blowing into bottles of various lengths. The

longer bottles will produce lower tones than shorter bottles. This effect is also demonstrated every time a water glass

is filled. The increase in the frequency of the sound that is produced as the glass is filled occurs because the

resonating cavity becomes shorter and shorter as more air is displaced by water. This simple rule will be quite

useful. For example, it can be applied directly to the differences that are observed in the acoustic properties of

speech produced by men, women, and children, who have vocal tracts that are quite different in length.

Resonant Frequencies and Formant Frequencies

The term "resonant frequency" refers to natural vibrating frequency or, equivalently, to a peak in a frequency

response curve. For reasons that are entirely historical, if the filter that is being described happens to be a human

vocal tract, the term formant frequency is generally used. So, one typically refers to the formant frequencies of the

The Physics of Sound 25

vocal tract but to the resonant frequencies of a plastic tube, the body of a guitar, the diaphragm of a loudspeaker, or

most any other type of filter other than the vocal tract. This is unfortunate since it is possible to get the mistaken idea

that formant frequencies and resonant frequencies are different sorts of things. The two terms are, in fact, fully

synonymous.

The Decibel Scale

The final topic that we need to address in this chapter is the representation of signal amplitude using the decibel

scale. The decibel scale is a powerful and immensely flexible scale for representing the amplitude of a sound wave.

The scale can sometimes cause students difficulty because it differs from most other measurement scales in not just

one but two ways. Most of the measurement scales with which we are familiar are absolute and linear. The decibel

scale, however, is relative rather than absolute, and logarithmic rather than linear. Neither of these characteristics is

terribly complicated, but in combination they can make the decibel scale appear far more obscure than it is. We will

examine these features one at a time, and then see how they are put together in building the decibel scale.

Linear vs. Logarithmic Measurement Scales

Most measurement scales are linear. To say that a measurement scale is linear means that it is based on equal

additive distances. This is such a common feature of measurement scales that we do not give it much thought. For

example, on a centigrade (or Fahrenheit) scale for measuring temperature, going from a temperature of 90 o to a

temperature of 91 o involves adding one 1 o. One rather obvious consequence of this simple additivity rule is that the

difference in temperature between 10 o and 11 o is the same as the difference in temperature between 90 o and 91 o.

However, there are scales for which this additivity rule does not apply. One of the best known examples is the

Richter scale that is used for measuring seismic intensity. The difference in seismic intensity between Richter values

of 4.0 and 5.0, 5.0 and 6.0, 6.0 and 7.0 is not some constant amount of seismic intensity, but rather a constant

multiple. Specifically, a 7.0 on the Richter scale indicates an earthquake that is 10 times greater in intensity than an

earthquake that measures 6.0 on the Richter scale. Similarly, an 8.0 on the Richter scale is 10 times greater in

intensity than a 7.0. Whenever jumping from one scale value to the next involves multiplying by a constant rather

than adding a constant, the scale is called logarithmic. (The multiplicative constant need not be 10. See Box 3-2 for

an example of a logarithmic scale – an octave progression – that uses 2 as the constant.) Another way of making the

same point is to note that the values along the Richter scale are exponents rather than ordinary numbers; for

example, a Richter value of 6 indicates a seismic intensity of 10 6

, a Richter value of 7 indicates a seismic intensity of

10 7, etc. The Richter values can, of course, just as well be referred to as powers or logarithms since both of these

terms are synonyms for exponent. The decibel scale is an example of a logarithmic scale, meaning that it is based on

equal multiples rather than equal additive distances.

Absolute vs. Relative Measurement Scales

A simple example of a relative measurement scale is the Mach scale that is used by rocket scientists to measure

speed. The Mach scale measures speed not in absolute terms but in relation to the speed of sound. For example, a

missile at Mach 2.0 is traveling at twice the speed of sound, while a missile at Mach 0.9 is traveling at 90% of the

speed of sound. So, the Mach scale does not represent a measured speed (Sm) in absolute terms, but rather,

represents a measured speed in relation to a reference speed (Sm/Sr ). The reference that is used for the Mach scale is

the speed of sound, so a measured absolute speed can be converted to a relative speed on the Mach scale by simple

division. For example, taking 783 mph as the speed of sound, 1,200 mph = 1200/783 = Mach 1.53. The decibel scale

also exploits this relative measurement scheme. The decibel scale does not represent a measured intensity (I m) in

absolute terms, but rather, represents the ratio of a measured intensity to a reference intensity (Im/I r ).

The decibel scale is trickier than the Mach scale in one important respect. For the Mach scale, the reference is

always the speed of sound, but for the decibel scale, many different references can be used. In explaining how the

decibel scale works, we will begin with the commonly used intensity reference of 10 -12 w/m2 (watts per square

meter), which is approximately the intensity that is required for an average normal hearing listener to barely detect a

1,000 Hz pure tone. So, for our initial pass through the decibel scale, 10 -12 w/m2 will serve as I r , and will perform the

same function that the speed of sound does for the Mach scale. Table 3-1 lists several sounds that cover a very broad

range of intensities. The second column shows the measured intensities of those sounds, and the third column shows

the ratio of those intensities to our reference intensity. Whispered speech, for example, measures approximately 10 -8

The Physics of Sound 26

w/m2, which is 10,000 times more intense than the reference intensity (10 -8 /10 -12 = 10 4 = 10,000). The main point to

be made about column 3 is that the ratios become very large very soon. Even a moderately intense sound like

conversational speech is 1,000,000 times more intense than the reference intensity. The awkwardness of dealing

with these very large ratios has a very simple solution. Column 4 shows the ratios written in exponential notation,

and column 5 simplifies the situation even further by recording the exponent only. The term exponent and the term

logarithm are synonymous, so the measurement scheme that is expressed by the numbers in column 5 can be

summarized as follows: (1) divide a measured intensity by a reference intensity (in this case, 10 -12 w/m2), (2) take the

logarithm of this ratio (i.e., write the number in exponential notation and keep the exponent only). This method, in

fact, is a completely legitimate way to represent signal intensity. The unit of measure is called the bel, after A.G.

Bell, and the formula is:

bel = log10 I m/I r , where: I m = a measured intensity

I r = a reference intensity

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Table 3-1. Sound intensities and intensity ratios showing how the decibel scale is created. Column 2 shows the

measured intensities (Im) of several sounds. Column 3 shows the ratio of these intensities to a reference intensity of

10 -12 w/m2 . Column 4 shows the ratio written in exponential notation while column 5 shows the exponent only. The

last column shows the intensity ratio expressed in decibels, which is simply the logarithm of the intensity ratio

multiplied by 10.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Measured Ratio Ratio in Exponent Decibel

Sound Intensity (I m) (Im/Ir ) Exp. Not. (log 10) (10 x log 10)

Threshold 10 -12 w/m2 1 10 0 0 0

@ 1 kHz

Whisper 10 -8 w/m2 10,000 10 4 4 40

Conversational 10 -6 w/m2 1,000,000 10 6 6 60

Speech

City Traffic 10 -4 w/m2 100,000,000 10 8 8 80

Rock & Roll 10 -2 w/m2 10,000,000,000 10 10 10 100

Jet Engine 10 0 w/m2 1,000,000,000,000 10 12 12 120

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Legitimate or not, the bel finds its sole application in textbooks attempting to explain the decibel. For reasons that

are purely historical, the log10 of the intensity ratio is multiplied by 10, changing bel into the decibel (dB). As shown

in the last column of Table 3-1, this has the very simple effect of turning 4 bels into 40 decibels, 8 bels into 80

decibels, etc. The formula for the decibel, then, is:

dB IL = 10 log10 I m/I r , where:

I m = a measured intensity

I r = a reference intensity

The designation "IL" stands for intensity level, and it indicates that the underlying measurements are of sound

intensity and not sound pressure. As will be seen below, a different version of this formula is needed if sound

pressure measurements are used. The multiplication by 10 in the dB IL formula is a simple operation, but it can

The Physics of Sound 27

sometimes have the unfortunate effect of making the formula appear more obscure that it is. The decibel values that

are calculated, however, should be readily interpretable. For example, 30 dB IL means 3 factors of 10 more intense

than I r , 60 dB IL means 6 factors of 10 more intense than I r , and 90 dB IL means 9 factors of 10 more intense than I r .

Deriving a Pressure Version of the dB Formula

In a simple world, we would be finished with the decibel scale. The problem is that the formula is based on

measurements of sound intensity, but as a purely practical matter sound intensity is difficult to measure. Sound

pressure, on the other hand, is quite easy to measure. An ordinary microphone, for example, is a pressure sensitive

device. The problem, then, is that the decibel is defined in terms of intensity measurements, but the measurements

that are actually used will nearly always be measures of sound pressure. This problem can be addressed since there

is a predictable relationship between intensity (I) and pressure (E): intensity is proportional to pressure squared:

I ο⊂ Ε 2

Knowing this relationship allows us to create a completely equivalent version of the decibel formula that will work

when sound pressure measurements are used instead of sound intensity measurements. All we need to do is

substitute squared pressure measurements in place of the intensity measurements:

dB IL = 10 log10 I m/I r (intensity version of formula)

dB SPL= 10 log10 E 2m/E 2 (pressure version of formula)

The designation "SPL" stands for sound pressure level, and it indicates that measures of sound pressure have been

used and not measures of sound intensity. Although the dB SPL formula shown here will work fine, it will almost

never be seen in this form. The reason is that the formula is algebraically rearranged so that the squaring operation is

not needed. The algebra is shown below:

(1) dB IL = 10 log10 I m/I r (the intensity version of the formula)

(2) dB SPL = 10 log10 E 2m/E 2 (measures of E 2 replace measures of I because I ο⊂ E 2)

(3) dB SPL = 10 log10 (E m/E r ) 2 (a 2

/b 2 = (a/b) 2)

(4) dB SPL = 10 ⋅ 2 log10 E m/E r (this is the only tricky step: log a b = b log a)

(5) dB SPL = 20 log10 E m/E r (2 ⋅ 10 = 20)

With the possible exception of the fourth step,4 the algebra is straightforward, but the details of the derivation

are less important than the following general points:

1. The decibel formula is defined in terms of intensity ratios. The basic formula is;

dB IL = 10 log10 I m/I r .

2. While sound intensity is difficult to measure, sound pressure is easy to measure. It is therefore necessary to

derive a version of the decibel formula that works when measures of sound pressure are used instead of sound

intensity.

4 Step 4 is the only tricky part of derivation. The reason it works is that squaring a number and then taking a log is the same as taking

the log first, and then multiplying the log by 2. For example, note that the two calculations below produce the same result:

log 10 100 2 = log 10 10,000 = 4 (square first, then take the log)

log 10 100 2 = (log 10 100) x 2 = 2 x 2 = 4 (take the log, then multiply by 2)

The Physics of Sound 28

3. The derivation of the pressure version of the formula is based entirely on the fact that intensity is proportional to

pressure squared (I ο⊂ Ε 2). This allows measures of E 2 to replace measures of I, turning: dB IL = 10 log10 I m/I r into

dB SPL = 10 log10 E 2m/E r2 . A few algebra tricks are applied to turn this formula into the more aesthetically pleasing

final version: dB SPL = 20 log10 E m/E r .

4. The two versions of the formula are fully equivalent to one another (see Box 3-3).

This last point about the equivalence of the intensity and sound pressure versions of the formula is explained in

some detail in Box 3-3, but the basic point is quite simple. The pressure version of the dB formula was derived from

the intensity version of the formula through algebraic manipulations (based on this relationship: I ο⊂ Ε 2). The whole

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Box 3-2

HARMONICS, OCTAVES, LINEAR SCALES, AND LOGARITHMIC SCALES

As we will see when the decibel scale is introduced, there is an important distinction to be made between

linear scales, which are quite common, and logarithmic scales, which are less common but quite important.

This distinction can be illustrated by examining the difference between a harmonic progression and an octave

progression. Notice that in a harmonic progression, the spacing between the harmonics is always the same; that

is, the difference between H1 and H 2 is the same as the difference between H2 and H 3, and so on. This is because

increases in frequency between one harmonic and the next involve adding a constant, with the constant being

the fundamental frequency. For example:

H 1 500

H 2 1000 (add 500)

H 3 1500 (add 500)

H 4 2000 (add 500)

. .

. .

. .

To get from one scale value to another on an octave progression involves multiplying by a constant rather

than adding a constant. For example, an octave progression starting at 500 Hz looks like this:

O 1 500

O 2 1000 (multiply by 2)

O 3 2000 (multiply by 2)

O 4 4000 (multiply by 2)

. .

. .

. .

As a result of the fact that we are multiplying by a constant rather than adding a constant, the spacing is no

longer even (i.e., the spacing between O 1 and O 2 is 500 Hz, the spacing between O2 and O 3 is 1000 Hz, and so

on). The point to be made of this is that there are two fundamentally different kinds of scales: (1) scales like

harmonic progressions that are created by adding a constant, which are by far the more common, and (2) scales

like octave progressions that are created by multiplying by a constant. Scales that are created by adding a

constant are called linear scales, while scales that are created by multiplying by a constant are called

logarithmic scales. Note that for an octave progression, the multiplier happens to be 2, meaning that progressing

from one frequency to an octave above that frequency involves multiplication by 2. However, a logarithmic

scale can be built using any multiplier. We will return to the distinction between linear and logarithmic scales

when we talk about the decibel scale, and there we will see that a logarithmic scale is built around multiplication

by a constant value of 10 rather than 2.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The Physics of Sound 29

point of algebra, of course, is to keep the expression on the left equal to the expression on the right. The simple and

useful point that emerges from this is this: If an intensity meter shows that a given sound measures 60 dB IL, for

example, a pressure meter will show that the same sound measures exactly 60 dB SPL. (This may seem

counterintuitive due to the differences in the formulas, but see Box 3-3 for the explanation.) The equivalence of the

two versions of the dB formula greatly simplifies the interpretation of sound levels that are expressed in decibels.

References

The reference that is used for the Mach scale is always the speed of sound. One of the virtues of the decibel

scale is that any reference can be used as long as it is clearly specified. The only reference that has been mentioned

so far is 10 -12 w/m2

, which is roughly the audibility threshold for a 1,000 Hz pure tone. This is a standard reference

intensity, and unless otherwise stated it should be assumed that this is used when a signal level is reported in dB IL.

The standard reference that is used for dB SPL is 20 μPa, so when a signal level is reported in dB SPL it should be

assumed that this reference is used unless otherwise stated.5

Many references besides these two standard references can be used. For example, suppose that a speech signal

is presented to a listener at an average level of 3500 μPa in the presence of a noise signal whose average sound

pressure is 1400 μPa. The speech-to-noise ratio (S/N) can be represented on a decibel scale, using the level of the

speech as E m and the level of the noise as E r :

dB s/n = 20 log10 E m/ E r

= 20 log10 3500/1400

= 20 log10 2.5

= 20 (0.39794)

= 7.96 dB

To take one more example, assume that a voice patient prior to treatment produces sustained vowels that

average 2300 μPa. Following treatment the average sound pressures increase to 8890 μPa. The improvement in

sound pressure (post-treatment relative to pre-treatment) can be represented on a decibel scale:

dB Improvement = 20 log10 E post/E pre

= 20 log10 8890/2300

= 20 log10 (3.86522)

= 20 (0.58717)

= 11.74 dB

A final example can be used to make the point that the decibel scale can be used to represent intensity ratios for

any type of energy, not just sound. Bright sunlight has a luminance measuring 100,000 cd/m2 (candela per square

meter). Light from a barely visible star, on the other hand, has a luminance measuring 0.0001 cd/m2. We can now

ask how much more luminous bright sunlight is in relation to barely visible star light, and the dB scale can be used

to represent this value. Since the underlying physical quanities here are measures of electromagnetic intensity, we

want the intensity version of the formula rather than the pressure version.

dB = 10 log10 I sunlight/I starlight

= 10 log10 100000/0.0001

= 10 log10 10 5/10 -4

= 10 log10 10 9 (division is done by subtracting exponents: 5 – (-4) = 9)

= 10 (9)

= 90 dB

5The standard pressure reference for dB SPL is sometimes given as 0.0002 dynes/cm2 rather than 20 μPa. These two sound pressures are

identical, however, in exactly the same sense that 4 quarts and 1 gallon are identical. Likewise, the standard reference for dB IL is often given as

10 -16 w/cm2 instead of 10 -12 w/m2

. These two intensities are also identical.

The Physics of Sound 30

The fact that we are measuring light rather than sound makes no difference: a decibel is 10 log10 I m/I r (or,

equivalently, 20 log10 E m/E r ), regardless of whether the energy comes from sound, light, electrical current, or any

other type of energy.

dB Hearing Level (dB HL)

The dB Hearing Level (dB HL) scale was developed specifically for testing hearing sensitivity for pure tones

of different frequencies. The sound-level dials on clinical audiometers,6 for example, are calibrated in dB HL rather

than dB SPL. To understand the motivation for the dB HL scale examine Figure 3-24, which shows the sound level (in

dB SPL) required for the average, normal-hearing listener to barely detect pure tones at frequencies between 125 and

8000 Hz. This is called the audibility curve and the simple but very important point to notice about this graph is

that the curve is not a flat line; that is, the ear is clearly more sensitive at some frequencies than others. The

differences in sensitivity are quite large in some cases. For example, the average normal-hearing listener will barely

detect a 1000 Hz pure tone at 7 dB SPL, but at 125 Hz the sound level needs to be cranked all the way up to 45 dB SPL,

an increase in intensity of nearly 4000:1. Now suppose we were to test pure-tone sensitivity using an audiometer that

is calibrated in dB SPL. Imagine that a listener barely detects a 1000 Hz pure tone at 25 dB SPL. Does this listener have

a hearing loss, and if so how large? The only way to answer this question is to consult the data in Figure 3-24, which

shows that the threshold of audibility for the average normal hearing listener at 1000 Hz is 7 dB SPL. This means that

the hypothetical listener in this example has a hearing loss of 25-7 = 18 dB. Suppose further that the same listener

detects a 250 Hz tone at 20 dBSPL. The table in Figure 3-24 shows that normal hearing sensitivity at 250 Hz is 25.5

dB SPL, meaning that the listener has slightly better than normal hearing at this frequency. As a final example,

imagine that this listener barely detects a 500 Hz tone at 30 dB SPL. Since the table shows that normal hearing

sensitivity at 500 Hz is 11.5 dB SPL, the listener has a hearing loss of 30.0-11.5 = 18.5 dB. The simple point to be

made about these examples is that, with an audiometer dial that is calibrated in dB SPL, it is not possible to determine

whether a listener has a hearing loss, or to measure the size of that loss, without doing some arithmetic involving the

normative data in Figure 3-24. The dB HL scale, however, provides a simple solution to this problem that avoids this

arithmetic entirely. The solution involves calibrating the audiometer in such a way that, when the level dial is set to

0 dB HL, sound level is set to the threshold of audibility for the average normal-hearing listener for that signal

frequency. For example, when the level dial is set to 0 dB HL at 125 Hz the level of tone will be 45 dB SPL – the

threshold of audibility for the average normal hearing listener at this frequency. Now if a listener barely detects the

125 Hz tone at 0 dB HL, no arithmetic is needed; the listener has normal hearing at this frequency. Further, if the

listener barely detects this 125 Hz tone at 40 dB HL, for example, the listener must have a 40 dB loss at this frequency

– and again it is not necessary to consult the data in Figure 3-24. Similarly, when the level dial is set to 0 dB HL at

250 Hz the level of the tone will be 25.5 dB SPL, which is the audibility threshold at 250 Hz. If this tone is barely

detected at 0 dB HL, the listener has normal hearing at this frequency. However, if the tone is not heard until the dial

is increased to 50 dB dB HL, for example, the listener has a 50 dB hearing loss at this frequency. The same system is

used for all signal frequencies: in all cases, the 0 dB HL reference is not a fixed number as it is for dB SPL (a constant

value of 20 μPa, no matter what the signal frequency is) or dB IL (a constant value of 10 -12 watts/m2, again

independent of signal frequency), but rather a family of numbers. In each case the reference for the dB HL scale is the

threshold of audibility for an average, normal-hearing listener at a particular signal frequency. What this means is

that values in dB HL are a fixed distance above the audibility curve, although they may be very different levels in

dB SPL. For illustration, Figure 3-25 shows the audibility curve (the filled symbols) and, above that in the unfilled

symbols, a collection of values that all measure 30 dB HL. Although the sound levels on the 30 dB HL curve vary

considerably in dB SPL (i.e. measured using 20 μPa as the reference), every data point on this curve is a constant 3

factors of 10, or 30 dB, above the audibility curve. The value of 30 dB in this figure is just an example. All values in

dB HL and dB SPL are interpreted in the same way: 50 dB SPL means that the signal being measured is 100,000 times

(i.e., 5 factors of 10) more intense than the fixed reference of 20 μPa, independent of frequency; 50 dB HL, on the

other hand, means that the signal being measured is 100,000 times (again, 5 factors of 10) more intense than a tone

that is barely audible to a normal-hearing listener at that signal frequency. Similarly, 20 dB SPL means that the signal

is 20 dB (2 factors of 10) more intense than the fixed reference of 20 μPa, while 20 dB HL means that the signal is 20

dB (again, 2 factors of 10) above the audibility curve.

6A clinical audiometer is an instrument with, among other things, one dial (for each ear) that controls pure-tone frequency and another dial

that controls the intensity of the tone. The listener is asked to raise a hand when the tone is barely audible.

The Physics of Sound 31

Summary

The decibel is a powerful scale for representing signal amplitude. The scale has two important properties: (1)

similar to the Mach scale, it represents signal level not in absolute terms but as a measured level divided by a

reference level; and (2) like the Richter scale, the dB scale is logarithmic rather than linear, meaning that it is based

on equal multiplicative distances rather than equal additive distances. While the decibel is defined in terms of

intensity ratios, for practical reasons, measures of sound pressure are far more common than measures of sound

intensity. Consequently, a version of the decibel formula was derived that makes use of pressure ratios rather than

intensity ratios. The derivation was based on the fact that intensity is proportional to pressure squared. The two

versions of the decibel formula (dB IL = 10 log 10 I m/I r and dB SPL = 20 log 10 E m/E r ) are fully equivalent, meaning that

if a sound measures 60 dB IL that same sound will measure 60 dB SPL. Unlike the Mach scale, which always uses the

speed of sound as a reference, any number of references can be used with the decibel scale. The standard reference

for the dB IL scale is 10 -12 w/m2 and the standard reference for the dB SPL scale is 20 μPa. However, any level can be

used as a reference as long as it is specified. The dB HL scale, widely used in audiological assessment, was developed

specifically for measuring sensitivity to pure tones of difference frequencies. The reference that is used for the dB HL

scale is the threshold of audibility at a particular signal frequency for the average, normal-hearing listener. Sound

levels in dB SPL and dB HL are interpreted quite differently. For example, a pure tone measuring 40 dB SPL is 4 factors

of 10 (i.e., 40 dB) greater than the fixed SPL reference of 20 μPa, while a pure tone measuring 40 dB HL is 4 factors

of 10 (again, 40 dB) greater than a tone of that same frequency that is barely audible to an average, normal-hearing

listener.

Frequency Threshold

125 45.0

250 25.5

500 11.5

750 8.0

1000 7.0

1500 6.5

2000 9.0

3000 10.0

4000 9.5

6000 15.5

8000 13.0

Figure 3-24. The threshold of audibility for the average, normal-hearing listener for pure tones varying between

125 and 8000 Hz. The audibility threshold is the sound level in dBSPL that is required for a listener to barely detect

a tone. Values on this curve are shown in the table to the right. The most important point to note about this graph

is that the curve is not flat, meaning that the ear is more sensitive at some frequencies than others. In particular,

the ear is more sensitive in a range of mid-frequencies between about 1000 and 4000 Hz than it is at lower and

higher frequencies. The complex shape of this curve provides the underlying motivation for the dB HL scale. See

text for details.

The Physics of Sound 32

Figure 3-25. The lower function is the audibility curve – the sound level in dB SPL that is required for an average

normal hearing listener to barely detect pure tones of different frequencies. The upper function shows sound levels for

a set of tones that all measure 30 dB HL. These tones vary quite a bit in dB SPL (i.e., relative to the constant value of 20

μPa) but in all cases the tones are a constant 3 factors of 10 in intensity (i.e., 30 dB) above the audibility curve.

The Physics of Sound 33

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Box 3-3

THE EQUIVALENCE OF THE INTENSITY AND PRESSURE

VERSIONS OF THE DECIBEL FORMULA

One fact about the two versions of the dB formula that is not always well understood is that the dB IL and

dB SPL formulas are fully equivalent. By "fully equivalent" we mean the following: suppose that a sound intensity

meter is used to measure the level of some sound, and we find that this sound is 1,000 times more intense than

the standard intensity reference of 10 -12 w/m2. The sound would then measure 30 dB IL (10 log10 1,000 = 10 (3) =

30 dB IL). Now suppose that we put the sound intensity meter away and use a sound pressure meter to measure

the same sound. You might think that the sound would measure 60 dB SPL since now we are multiplying by 20

instead of 10, but the trick is that the ratio is no longer 1,000. Recall that intensity is proportional to pressure

squared, which means that pressure is proportional to the square root of intensity. This means that if the intensity

ratio is 1,000, the pressure ratio must be the square root of 1,000, or 31.6. So, the formula now becomes 20 log

31.6 = 20 (1.5) = 30 dB SPL, which is exactly what we obtained originally. It will always work out this way: if a

sound measures 50 dB IL, that same sound will measure 50 dB SPL.

Table 3-2 might help to make this more clear. The first column shows an intensity ratio, the second column

shows the corresponding pressure ratio (this is always the square root of the intensity ratio), the third column

shows the dB IL value (10 log of the intensity ratio), and the fourth column shows dB SPL value (20 log of the

pressure ratio). As you can see, they are always the same.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Table 3-2. Intensity ratios, equivalent pressure ratios, dB IL values and

dB SPL values showing the equivalence of the intensity and pressure versions

of the dB formula.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Intensity Pressure dB IL dB SPL

Ratio Ratio (10 log10 I m/I r ) (20 log10 E m/E r )

10 3.16 10.00 10.00

20 4.47 13.01 13.01

40 6.32 16.02 16.02

50 7.07 16.99 16.99

60 7.75 17.78 17.78

70 8.37 18.45 18.45

80 8.94 19.03 19.03

90 9.49 19.54 19.54

100 10.00 20.00 20.00

200 14.14 23.01 23.01

300 17.32 24.77 24.77

400 20.00 26.02 26.02

500 22.36 26.99 26.99

1000 31.62 30.00 30.00

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The Physics of Sound 34

Study Questions: Physical Acoustics

1. Explain the basic processes that are involved in the propagation of a sound wave.

2. Draw time- and frequency-domain representations of simple periodic, complex periodic, complex aperiodic, and

transient sounds.

3. Draw time- and frequency-domain representations of two complex periodic sounds with different fundamental

frequencies.

4. Draw time-domain representations of two simple periodic sounds with the same frequency and phase, but different

amplitudes.

5. Draw time-domain representations of two simple periodic sounds with the same frequency and different amplitudes but

different phases.

6. Draw amplitude spectra of two sounds with the same fundamental frequencies but different spectrum envelopes.

7. Draw amplitude spectra of two sounds with different fundamental frequencies but similar spectrum envelopes.

8. Calculate signal frequencies for sinusoids with the following values:

a. period = 0.34 s

b. period = 2 s

c. period = 10 ms

d. period = 2 ms

e. wavelength = 20 cm

f. wavelength = 100 cm

Answers:

a. f = 1/0.34 = 2.94 Hz

b. f = 1/2 = 0.5 Hz

c. f = 1/0.01 = 100 Hz

d. f = 1/.002 = 500 Hz

e. f = c/WL (speed of sound/wavelength) = 35000/20 = 1750 Hz

f. f = c/WL (speed of sound/wavelength) = 35000/100 = 350 Hz

9. Calculate the three lowest resonant frequencies of the following uniform tubes that are closed at one end and open at

the other end:

a. 10 cm

b. 30 cm

c. 40 cm

Answers:

a. wavelength of lowest resonance = 40 cm (10 x 4)

f = 35000/40 = 875

R1 = 875 (R1 = frequency of resonance number 1)

R2 = 2625

R3 = 4375

b. wavelength of lowest resonance = 120 cm (30 x 4)

f = 35000/120 = 291.7

The Physics of Sound 35

R1 = 291.7

R2 = 875.0

R3 = 1458.3

c. wavelength of lowest resonance = 160 cm (40 x 4)

f = 35000/160 = 218.75

R1 = 218.75

R2 = 656.25

R3 = 1093.75

10. Show what the frequency-response curves look like for the tubes in the problem above.

11. A complex periodic signal has a fundamental period of 4 msec. What is the fundamental frequency of the signal? At

what frequencies would we expect to find energy?

12. How are the terms octave and harmonic different?

13. Give examples of the following kinds of graphs, being sure to label both axes:

a. amplitude spectrum

b. phase spectrum

c. frequency-response curve

d. time-domain representation

14. Give a brief explanation of the basic idea behind Fourier analysis. What is the input to Fourier analysis and what kind

of output(s) does it produce?

15. Draw and label frequency-response curves for low-pass, high-pass, and band-pass filters.

16. What parameters control the frequency of vibration of a spring and mass system?

17. Draw the time domain representation of one cycle of a sinusoid as variations in instantaneous air pressure over time

and one cycle of that same sinusoid as variations in instantaneous velocity over time.

18. How, if at all, are the terms resonant frequency and harmonic different?

19. How, if at all, are the terms resonant frequency and formant different?

20. A harmonic is a peak in: (a) a frequency response curve, (b) an amplitude spectrum, or (c) either a frequency response

curve or an amplitude spectrum.

21. A resonance is a peak in: (a) a frequency response curve, (b) an amplitude spectrum, or (c) either a frequency response

curve or an amplitude spectrum.

22. A formant is a peak in: (a) a frequency response curve, (b) an amplitude spectrum, or (c) either a frequency response

curve or an amplitude spectrum.

23. A frequency response curve describes a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

24. An amplitude spectrum describes a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

The Physics of Sound 36

Frequency Response Problems

The Physics of Sound 37

Answers to Frequency Response Problems

The Physics of Sound 38

Decibel Study Questions

1. What reference is used for the dB IL scale?

2. What reference is used for the dB SPL scale?

3. What reference is used for the dB HL scale?

4. What reference is used for the dB SL scale?

5. A listener barely detects a 125 Hz pure tone at 55 dB SPL. Does this listener have a hearing loss at 125 Hz, and if

so, what is the size of the hearing loss?

6. A listener barely detects a 1,000 Hz pure tone at 55 dB SPL. Does this listener have a hearing loss at 1,000 Hz,

and if so, what is the size of the hearing loss?

7. A listener barely detects a 125 Hz pure tone at 55 dB HL. Does this listener have a hearing loss at 125 Hz, and if

so, what is the size of the hearing loss?

8. A listener barely detects a 1,000 Hz pure tone at 55 dB HL. Does this listener have a hearing loss at 1,000 Hz,

and if so, what is the size of the hearing loss?

9. 60 dB SPL at 1,000 Hz means \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ more intense than \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

10. 60 dB IL at 1,000 Hz means \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ more intense than \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

11. 60 dB HL at 1,000 Hz means \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ more intense than \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

12. The reference that is used for the dB SPL scale is:

a. a number

b. a sentence

13. If the answer to the question above is a number, give the number; if it’s a sentence, give the sentence.

14. The reference that is used for the dB HL scale is:

a. a number

b. a sentence

15. If the answer to the question above is a number, give the number; if it’s a sentence, give the sentence.

16. A specific individual has a 70 dB hearing loss in the left ear at 1,000 Hz. A 90 dB HL, 1,000 Hz tone that is

presented to this listener’s left ear would measure \_\_\_\_\_\_ dB SL.

17. A sound measures 42 dB IL. On the dB SPL scale, that same sound will measure:

a. 84 dB SPL because with the dB SPL formula we are now are multiplying the ratio by 20 instead of 10.

b. 42 dB SPL because the two versions of the formula are equivalent

18. A sound measures 60 dB IL. (a) The measured intensity (I M) must therefore be \_\_\_\_\_\_\_\_\_ times

greater than the reference intensity (I R). (b) What would the pressure ratio (E M/E R) be for this same sound? (c)

Do the arithmetic to show what this sound would measure in dB SPL.

The Physics of Sound 39

19. A sound measures 40 dB IL. (a) The measured intensity (I M) must therefore be \_\_\_\_\_\_\_\_\_ times

greater than the reference intensity (I R). (b) What would the pressure ratio (E M/E R) be for this same sound? (c)

Do the arithmetic to show what this sound would measure in dB SPL.

20. On the graph below, put a mark at: (a) 3,000 Hz, 20 dB SPL, and (b) 3,000 Hz, 20 dB HL (the

grid lines on the y axis are spaced at 2 dB intervals).

Frequency Threshold

in Hz in dB SPL

125 45.0

250 25.5

500 11.5

750 8.0

1000 7.0

1500 6.5

2000 9.0

3000 10.0

4000 9.5

6000 15.5

8000 13.0

The Physics of Sound 40

Answers to Decibel Study Questions

1. 10 -12 watts/m2

2. 20 μPa (or, equivalently, 0.0002 dynes/cm2)

3. The threshold of audibility for an average, normal-hearing listener at a particular signal frequency.

4. 3. The threshold of audibility for a particular listener at a particular signal frequency.

5. Consulting the attached figure and table showing the audibility curve for average, normal-hearing listeners, we

find that the threshold of audibility at 125 Hz is 45 dB SPL. A listener who barely detected a 125 Hz tone at 55

dB SPL would therefore have hearing loss of 55-45=10 dB; that is, the hearing sensitivity of this listener would be

10 dB worse than normal.

6. Consulting the attached figure and table showing the audibility curve for average, normal-hearing listeners, we

find that the threshold of audibility at 1,000 Hz is 7 dB SPL. A listener who barely detected a 1,000 Hz tone at

55 dB SPL would therefore have a hearing loss of 55-7=48 dB; that is, the hearing sensitivity of this listener

would be 48 dB worse than normal.

7. The reference for dB HL is the audibility threshold, so this listener would have a 55 dB hearing loss at 125 Hz.

There is no need to consult the table.

8. The reference for dB HL is the audibility threshold, so this listener would have a 55 dB hearing loss at 1,000 Hz.

There is no need to consult the table.

9. 6 factors of 10 (i.e., 1,000,000 times) more intense than 20 μPa)

10. 6 factors of 10 (i.e., 1,000,000 times) more intense than 10 -12 watts/m2

11. 6 factors of 10 (i.e., 1,000,000 times) more intense than a 1,000 Hz tone that is barely audible to an average,

normal-hearing listener.

12. a number

13. 20 μPa

14. a sentence

15. The threshold of audibility for an average, normal-hearing listener at a particular signal frequency.

16. 20 dB SL. The reference for the dB SL (SL=sensation level) is the threshold of audibility for a specific listener. So,

what we want to know here very simply is where this 90 dB HL tone is in relation to this particular listener’s

threshold. This listener has a 70 dB hearing loss at this frequency, so the 90 dBHL tone, which would be 90 dB

above a normal-hearing listener’s threshold, is only 20 dB above this particular listener’s threshold.

17. 42 dB SPL: The pressure version of the formula was derived from the intensity version through algebraic

manipulations, so they have to be equivalent to one another. The next problem was designed to illustrate how

this can be the case.

18. (a) 1,000,000 times (6 factors of 10) more intense than I R. (b) If the intensity ratio is 1,000,000, the pressure

ratio has to be the square root of 1,000,000, which is 1,000. (c) dB SPL = 20 log 1,000 = 20 . 3 = 60 dB SPL. This is

exactly what we got for the same sound measured in dB IL. It will always be the same. If a sound measures 60

dB IL, that same sound will measure 60 dB SPL.

The Physics of Sound 41

19. (a) 10,000 times (4 factors of 10) more intense than I R. (b) If the intensity ratio is 10,000, the pressure ratio has

to be the square root of 10,000, which is 100. (c) dB SPL = 20 log 100 = 20 . 2 = 40 dB SPL. This is exactly what

we got for the same sound measured in dB IL. It will always be the same. If a sound measures 40 dB IL, that same

sound will measure 40 dB SPL.

20. See below. The lower of the two marks is 20 dB (2 factors of 10) above the constant reference line of 20 μPa.

The higher of the two marks is 20 dB (also 2 factors of 10) above the curvey line, which is the threshold of

audibility for the average normal-hearing listener.

The Physics of Sound 42

The Physics of Sound 43

The Physics of Sound 44

A Tutorial on Digital Sound Synthesis Techniques

Author(s): Giovanni de Poli

Source:

Computer Music Journal, Vol. 7, No. 4 (Winter, 1983), pp. 8-26

Published by: The MIT Press

Stable URL: https://www.jstor.org/stable/3679529

Accessed: 08-06-2020 20:13 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide

range of content in a trusted digital archive. We use information technology and tools to increase productivity and

facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at

https://about.jstor.org/terms

The MIT Press is collaborating with JSTOR to digitize, preserve and extend access to

Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Giovanni De Poli

Centro di Sonologia Computazionale

Istituto di Elettrotecnica ed Elettronica

Universith di Padova, Italy

A Tutorial on Digital

Sound Synthesis

Techniques

Introduction

Progress in electronics and computer technology

has led to an ever-increasing utilization of digital

techniques for musical sound production. Some of

these are the digital equivalents of techniques em-

ployed in analog synthesizers and in other fields of

electrical engineering. Other techniques have been

specifically developed for digital music devices and

are peculiar to these.

This paper introduces the fundamentals of the

main digital synthesis techniques. Mathematical

developments have been restricted in the exposi-

tion and can be found in the papers listed in the

references. To simplify the discussion, whenever

possible, the techniques are presented with refer-

ence to continuous signals.

Sound synthesis is a procedure used to produce a

sound without the help of acoustic instruments. In

digital synthesis, a sound is represented by a se-

quence of numbers (samples). Hence, a digital syn-

thesis technique consists of a computing procedure

or mathematical formula, which computes each

sample value.

Normally, the synthesis formula depends on

some values, that is, parameters. Frequency and

amplitude are examples of such parameters. Param-

eters can be constant or slowly time variant during

the sound. Time-variant parameters are also called

control functions.

Synthesis techniques can be classified as (1) gen-

eration techniques (Fig. la), which directly produce

the signal from given data, and (2) transformation

techniques (Fig. Ib), which can be divided into two

stages, the generation of one or more simple signals

and their modification. Often, more or less elabo-

rate combinations of these techniques are employed.

Fixed-Waveform Synthesis

In many musical sounds, pitch is a characteristic to

which we are quite sensitive. In examining the tem-

poral shape of pitched sounds, we see a periodic rep-

etition of the waveform without great variations.

The simplest synthesis method attempts to re-

produce this characteristic, generating a periodic

signal through continuous repetition of the wave-

form. This method is called fixed-waveform

synthesis.

The technique is carried out by a module called

an oscillator (Fig. 2), which repeats the waveform

with a specified amplitude and frequency. In certain

cases, the waveform is characteristic of the os-

cillator and cannot be changed. But often it can be

chosen in a predetermined set of options or given

explicitly when required.

Usually, in digital synthesis the waveform value

at a particular instant is not computed anew for

each sample. Rather, a table, containing the period

values computed in equally spaced points, is built

beforehand. Obviously, the more numerous the

points in the table, the better the approximation

will be. To produce a sample, the oscillator requires

the waveform value at that precise instant. It cy-

clically searches the table to get the point nearest

to the required one. Sometimes a finer precision is

achieved by interpolation between two adjacent

points.

The distance in the table between two samples

read at subsequent instants is called the sam-

pling\_increment. The sampling \_increment is pro-

portional to the frequency f of the generated signal

according to the following formula (Mathews 1969):

N

samplingincrement = SRf,

where N is the table length and SR the sampling

rate.

In the oscillator, the frequency is usually speci-

Computer Music Journal, Volume 7, No. 4,

Winter, 1983, 0148-9267/83/040008-19 $04.00/0,

? 1983 Massachusetts Institute of Technology.

8 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 1. Classification of

synthesis techniques. Gen-

eration techniques (a) and

transformation techniques

(b).

Fig. 2. Fixed-waveform

synthesis oscillator.

(a) Parameters

'[ M SoundGeneration S

signal

Complex

(b) Parameters sound

I signal

Generation Transformation

Simple

signals

A(t) f(t)

s(t)

fled as a sampling- increment and the algorithm

that realizes it is as follows:

signal [t] := amplitude \* table [phase],

(Relation 1)

and

phase := mod(n, phase + samplingincrement),

(Relation 2)

where

Table contains one period of the waveform;

Phase is the theoretical position in the table of

the sample to be extracted at the instant; and

Amplitude is the signal amplitude.

Relation 2 computes the phase value in the subse-

quent instant, approximating the frequency integra-

tion by a summation. The modulus operation keeps

the phase inside the table length n.

It is noteworthy that the signal generated in this

way is an approximation of the desired one (Mail-

liard 1976). The approximation depends on the

table length, the interpolation method, and the sig-

nal frequency. For a sufficiently long table, it is

fully satisfactory.

The results of fixed-waveform synthesis are of

poor musical quality, as the sound does not present

any variation along its duration. This technique can

be changed by allowing the amplitude to vary in

time. In real sounds, the amplitude is rarely con-

stant: it starts from zero, reaches a maximum after

a certain time (attack), remains nearly constant

(steady state) and, after a certain evolution, it re-

turns to zero (decay). This sequence of amplitude

behavior is called the envelope. Thus, when the

amplitude varies according to a control function,

we have fixed-waveform synthesis with an ampli-

tude envelope.

The envelope can be generated in many ways.

In software-based synthesis, the most frequent

method uses an oscillator module, seen previously,

using a very low frequency equal to the inverse of

the duration. In this case, it performs a single cycle

and its waveform corresponds to the amplitude

envelope.

By carefully analyzing natural periodic sounds, it

has been shown that even the most stable ones con-

tain small frequency fluctuations. These improve

the sound quality and avoid unpleasant beatings

when more sounds are present at the same time.

The fixed-waveform technique can also be modi-

fied so that the oscillator frequency can slowly vary

around a value. This enables the production of a

tremolo and, with wider variations, of a glissando

or melodies.

The combination of these two variations consti-

tutes fixed-waveform synthesis with time-varying

amplitude and frequency. The waveform is fixed,

while the amplitude and frequency vary. The par-

tials are exact multiples of the fundamental, and

they all behave the same.

Fixed-waveform synthesis is realized rather sim-

ply. Hence, it is often employed when good sound

quality is not required. The constant waveform

gives the sound a mechanical, dull, and unnatural

character, which soon annoys the audience. Thus,

DePoli 9

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

in musical applications, fixed-waveform synthesis

is not very effective when used alone. It is em-

ployed for its simplicity when timbral variety is not

required, for example, for real-time synthesis on

very limited hardware.

For economy, other methods of generating wave-

forms that do not use tables or multiplications have

been devised. The simplest generates a square or

(more generally) a rectangular wave, alternating

sequences of positive and negative samples of the

same value. The frequencies that can be obtained

are submultiples of the sampling rate.

A sawtooth signal can also be generated by an ac-

cumulator to which a constant value is continu-

ously added. The output increases linearly until it

overflows and starts from the beginning. The signal

frequency is proportional to the constant value.

This method is used to produce linearly variable

control signals. Every time the additive constant

changes, the slope changes. Hence, functions com-

posed of straight segments, such as envelopes, can

be obtained.

This technique has been generalized recently by

Mitsuhashi (1982a). A polynomial of degree N can

be generated by putting N accumulators in cascade.

The accumulators are initialized by the value of the

forward differences, in decreasing order, of the poly-

nomial to be generated (Cerruti and Rodeghiero

1983). The waveforms obtained exhibit great vari-

ety and, in certain conditions, they are periodic.

Granular Synthesis

The technique of fixed-waveform synthesis pro-

duces rather static sounds in time. Yet a fundamen-

tal characteristic of musical sound is its timbral

evolution in time. A sound can be thought of as a

sequence of elementary sounds of constant dura-

tion, analogous to a film, in which a moving image

is produced by a sequence of images.

In computer music, the elementary sounds are

called grains, and the technique of exploiting this

facility is granular synthesis (Roads 1978). The

grains can be produced by a simple oscillator or by

other methods. The duration of each grain is very

short, on the order of 5-20 msec.

There are two ways to implement granular syn-

thesis. The first is to organize the grains into

frames, like the frames of a film. At each frame,

the parameters of all the grains are updated. This is

the approach sketched by Xenakis (1971). The sec-

ond way involves scattering the grains within a

mask, which bounds a particular frequency/ampli-

tude/time region. The density of the grains may

vary within the mask. This is the method imple-

mented by Roads (1978).

A problem with granular synthesis is the large

amount of parameter data to be specified. In some

other types of synthesis (additive and subtractive, to

be discussed shortly), these data can be obtained by

analyzing natural sounds. However, no analysis sys-

tem for granular synthesis has been developed. An-

other possibility is to obtain the parameter data

from an interactive composition system, which al-

lows the composer to work with high-level musical

concepts while automatically generating the thou-

sands of grain parameters needed.

Additive Synthesis

In additive synthesis, complex sounds are produced

by the superimposition of elementary sounds. In

certain conditions, the constituent sounds fuse

together and the result is perceived as a unique

sound. This procedure is used in some traditional

instruments, too. In an organ, the pipes generally

produce relatively simple sounds; to obtain a richer

spectrum in some registers, notes are created by

using more pipes sounding at different pitches at

the same time. The piano uses a different proce-

dure. Many notes are obtained by the simultaneous

percussion of two or three strings, each oscillating

at a slightly different frequency. This improves the

sound intensity and enriches it with beatings.

In order to choose the elementary sounds of addi-

tive synthesis, we first note that the Fourier analy-

sis model enables us to analyze sounds in a way

similar to the human ear and so to extract param-

eters that are perceptually significant. When we

analyze a real, almost-periodic sound, we imme-

diately notice that each partial amplitude is not

proportionally constant, but that it varies in time

10 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 3. Additive synthesis.

AI(t) f,(t) A2(t) f2(t) AM(t) fM(t)

s(t)

according to different laws. In the attack portion of

a note, some partials, which in the steady state are

negligible, are often significant.

Any almost-periodic sound can be approximated

as a sum of sinusoids. Each sinusoid's frequency is

nearly multiple that of the fundamental, and each

sinusoid evolves in time. For higher precision, the

frequency of each component can be considered as

slowly varying. Thus, additive synthesis consists of

the addition of some sinusoidal oscillators, whose

amplitude, and at times frequency, is time varying

(Fig. 3).

The additive-synthesis technique also provides

good reproduction of nonperiodic sounds, present-

ing in the spectrum the energy concentrated in

some spectral lines. For example, Risset (1969) imi-

tated a bell sound by summing sinusoidal compo-

nents of harmonically unrelated frequencies, some

of which were beating. In Risset's example, the ex-

ponential envelope was longer for the lower partials.

Additive synthesis provides great generality. But a

problem arises because of the large amount of data

to be specified for each note. Two control functions

for each component have to be specified, and nor-

mally they are different for each sound, depending

on its duration, intensity, and frequency. The pos-

sibility of data reduction has been investigated. At

Stanford University, a first result has been obtained

by representing the control functions of the ampli-

tude and the frequency of each component by line

segments, without affecting "naturalness" of the

sound (Grey and Moorer 1977).

The next step has been to investigate the rela-

tions between these functions (Risset and Mathews

1969; Beauchamp 1975) or their relation to others

of more general character (Charbonneau 1981). Ad-

ditive synthesis is most practically used either in

synthesis based on analysis (analysis/synthesis),

often transforming the extracted parameters, or

when a sound of a precise and well-determined

characteristic is required, as in psychoacoustic ex-

periments. In any case, in order to familiarize musi-

cians with sound characteristics and frequency

representations, the technique is also useful from

a pedagogical point of view.

Additive synthesis can be generalized by using

waveform components of other shapes besides si-

nusoids. To allow the reproduction of any sound,

these waveforms have to satisfy specific mathe-

matical properties. Walsh functions are an example

of this kind of function; they are used for their sim-

ple hardware realization (Rozenberg 1979).

VOSIM

In the synthesis techniques already discussed, os-

cillators that periodically reproduce a given wave-

form are employed. Other synthesis techniques,

instead of continuously repeating a given wave-

form, calculate it anew each period, with minor

variations. The control of this calculation process

allows continuous spectral variations. A common

method of this type is the voice simulation (VOSIM)

technique. A VOSIM oscillator has been devised in

a project at the Institute of Sonology in Utrecht

(Kaegi 1973, 1974; Kaegi and Tempelaars 1978).

The VOSIM waveform (Fig. 4) consists of a se-

quence of N pulses of shape sin2, of the same dura-

tion T, and of decreasing amplitude. The sequence

is followed by a pause M. Each pulse's amplitude

is smaller than the preceding one, by a constant

factor b.

The VOSIM spectrum (Fig. 5a) is described as the

product of two terms (Tempelaars 1976; De Poli

and De Poli 1979). The first term S, (Fig. 5b) de-

pends only on the pulse shape and limits the signal

bandwidth to 2F (being F = 1/T). The second term

S2 (Fig. 5c) depends on the relationship between the

individual pulse amplitudes. S2 is periodic in the

frequency domain with a period F, and it is sym-

DePoli 11

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 4. VOSIM oscillator: T

is the duration of single

pulse, M the rest between

two sequences of pulses.

Fig. 5. Spectral envelope of

a VOSIM oscillator (N =

5, b = 0, 8) (a). The enve-

lope is the product of the

terms S, (b) and S2 (c).

1.5

1

0.5

0 5 M- 10 15

T

metric with respect to F/2. When b 1, its ampli-

tude will be greater around the extremes of the

period 0 and F. When b - -1, its amplitude will be

greater in the central position around F/2. Thus, a

characteristic formant in F or F/2 will result. The

number of pulses N produces N oscillations in the

S2 term between 0 and F, with strong signals for b

near +F.

This constitutes the spectral envelope of the re-

peated waveform. Taking a as the ratio between the

signal period and a single pulse duration, the num-

ber of the harmonic corresponding to the formant is

a if b is positive, and a/2 if b is negative. Thus, by

varying a, the formant shifts, and the relative am-

plitude of all the harmonics vary continuously but

not homogeneously, following the spectral enve-

lope. The signal and the formant frequencies can be

separately controlled.

More kinds of sounds can be obtained by modu-

lating (sinusoidally or randomly) the value of the

time interval M between two consecutive pulse se-

quences. This means that a varies independently

from T. In this case, the formant frequency remains

constant while the harmonic amplitudes vary. Then

the ear can easily perceive the spectral envelope

and fuse the components together. This property

makes the VOSIM oscillator effective in musical

applications.

If a variation is strong, practically aperiodic

sounds or colored noises are obtained. Adding sev-

eral VOSIM oscillators allows one to control the

position of the formants. This results in an additive

(a)

5s(f)l

6

4

2

0

0 0.5F IF 1.5F 2F 2.5F

(b)

2.5

IsI(f)l

2

1.5

1

0.5

0

0 0.5F iF 1.5F 2F 2.5F

(c)

Is2(f)l

3

2.5

2

1.5

1

0.5

0 0.5F 1F 1.5F 2F 2.5F

12 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

synthesis of already complex sounds rather than of

sinuosidal components. Instead of the frequency of

partials, the position of the formants is controlled.

This is a more relevant parameter, from an acoustic

standpoint.

The formant-wave-function synthesis of Rodet

(1980) is analogous to VOSIM, but it allows over-

lapping of single waveforms. This provides better

control and generally richer sounds. Mitsuhashi

(1982a) and Bass and Goeddel (1981) generalized

the VOSIM model by including the case of pulses

of any amplitude and using different elementary

waveforms.

Synthesis by Random Signals

Up to now, we have considered signals whose be-

havior at any instant is supposed to be perfectly

knowable. These signals are called deterministic

signals. Besides these signals, random signals, of

unknown or only partly known behavior, may be

considered. For random signals, only some general

characteristics, called statistical properties, are

known or are of interest. The statistical properties

are characteristic of an entire signal class rather

than of a single signal. A set of random signals is

represented by a random process. Particular numer-

ical procedures simulate random processes, pro-

ducing sequences of random (or more precisely,

pseudorandom) numbers. The linear congruential

method is commonly used to produce uniformly

distributed numbers. From a starting value X0, a

sequence of random integers X0, X1, . . . , XK...

is generated according to the relation

XK+ = (a XK + C)modm,

where m is the modulus and the maximum se-

quence period, and a and c are two specific integer

constants.

The modulus operation can be avoided by choos-

ing m as the maximum number representable in

the computer, that is, m = 2b, where b is the word

length (bit number in a binary computer). So the

numbers are automatically truncated. The choice

of X0, a, and c greatly affects the statistical charac-

teristics of the generated sequence, and its accept-

ability has to be accurately verified by statisti

tests. A general discussion of various distribut

and the methods used to generate them can be

in Lorrain's paper (1980).

Random sequences can be used both as sign

(i.e., to produce white or colored noise used as

put to a filter) and as control functions to pro

a variety in the synthesis parameters most per

tible by the listener.

In the analysis of natural sounds, some chara

teristics vary in an unpredictable way; their m

statistical properties are perceptibly more sign

cant than their exact behavior. Hence, the add

of a random component to the deterministic f

tions controlling the synthesis parameters is o

desirable.

In general, a combination of random processe

is used because the temporal organization of th

musical parameters often has a hierarchical asp

It cannot be well described by a single random

process, but rather by a combination of rando

cesses evolving at different rates.

Linear Transformations

Let us now examine techniques for signal modifica-

tion. A transformation is a set of rules and proce-

dures transforming a signal called input to another

signal called output. A transformation is linear if

the superimposition principle is valid, that is, if the

effect of the transformation caused by a two-signal

addition is equal to the addition of the individual

signal transformations applied separately. In partic-

ular, in a linear transformation a signal can be mul-

tiplied by a constant but not by another signal.

Digital filters are linear transformations that can

be described by the following difference equation:

N M

Sa, yK=0 i1=0

where aK and b-

and y(i) are the i

signal. The value

ear combination

with the precedi

DePoli 13

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 6. Finite-impulse-

response (FIR) filter with

two zeros described by the

equation y(n) = x(n) +

a,x(n - 1) + a2x(n - 2)

(a). Infinite-impulse-

response (IIR) filter with

two poles described by the

equation y(n) = x(n) +

P1y(n - 1) + 32y(n - 2)

(b).

x(n) (a)y(n)

Z-1

x(n - 2)

----'--

x(n) (b)y(n)

Z-1

- ~ y(n- 2)

the input is sinusoidal, the steady-state output is

sinusoidal with the same frequency. The amplitude

and phase of the frequencies are determined by the

system. That is why this transformation is called

a filter.

Subtractive Synthesis

Sound produced by filtering a complex waveform is

called, sometimes inappropriately, subtractive syn-

thesis. First, a periodic or aleatoric signal rich in

harmonics is generated by the previously examined

techniques or others. This signal must contain

energy in all frequencies required in the output

sound. Second, one or more filters are used to alter

selectively the specific frequency components. The

undesired components are attenuated (subtracted)

and others are eventually amplified. When the filter

coefficients change, the frequency response changes,

too. Thus, it is possible to vary characteristics of

the output sound.

In modular diagrams, filters are usually repre-

sented by rectangles and the difference equation or

the transfer function is given as a label near the

rectangle. Two examples of simple digital filters,

showing their internal structure, are shown in

Fig. 6. The first filter (Fig. 6a) has a finite-impulse

response (FIR). This structure is useful to produce

transmission zeros: that is, it can nullify some fre-

quencies that depend on a1, a2 values and on the

sampling rate. The second filter (Fig. 6b) is recur-

sive, or has an infinite-impulse response (IIR). Feed-

back in the structure amplifies certain frequencies,

that is, produces transmission poles. When used as

bandpass filter, in general terms, the coefficient P,

controls the center frequency and the coefficient 32

the bandwidth.

One of the most attractive aspects of digital filter-

ing is that it is analogous to the functioning of

many acoustic musical instruments. Indeed, instru-

ment physics can be used as a model for synthesis.

For example, in the brasses and woodwind instru-

ments, the lips or vibrating reed generate a periodic

signal rich in harmonics. The various cavities and

the shape of the instrument act as resonators, en-

hancing some spectral components and attenuating

others. In the human voice, the excitation signals

are periodic pulses of the glottis (in the case of

voiced sounds) or white noise (in the case of un-

voiced sounds-for example, the consonants s and

z). The throat, the mouth, and the nose are the fil-

tering cavities, and their dimensions vary in time.

Their great variability makes the human voice the

most rich and interesting musical instrument.

Today, subtractive synthesis is the standard means

of speech synthesis. An analysis procedure, called

linear predictive coding (LPC), allows us to obtain

14 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 7. Elementary filters

used in reverberators.

Comb filters (a). All-pass

filter (b).

the pitch and the coefficients of a recursive (poles

only) filter (see Cann's [1979-1980] tutorial and

Moorer's paper [1979a]). These data can be utilized

to synthesize the sound directly or following modi-

fication. For example, speech can be accelerated or

slowed down, and pitch can be varied. An instru-

ment or orchestral sound can be used as input to the

filter, producing the effect of a "talking orchestra."

Interesting possibilities for musique concrete

sound processing arise. Not only simple filtering of

sounds is possible, but the modification of their

most intrinsic characteristics is also made possible

by varying the parameters of the deduced sound-

production model.

Generally, LPC is relatively difficult to use. In-

tuitively, the filter characteristics depend on the

position of the zeros and the poles in the transfer

function. These characteristics are affected in a

complex and nonintuitive way by the filter coeffi-

cients. In some simple cases, approximate formulas

give the coefficients as functions of significant pa-

rameters, that is, center frequency and bandwidth,

or cutoff frequency and slope. The filters can be

used in series or in parallel. In the most complex

cases, a precise analysis is obtained by using spe-

cific programs for digital filter design and analysis.

Such digital filters can be very stable and precise,

but only at the cost of a large amount of calcula-

tion. Simple linear digital networks can also be

used as oscillators (Tempelaars 1982) by applying

a pulse sequence to the input and choosing an

impulse response equal to the signal function to

be generated.

Reverberation

One application of digital filters is sound reverbera-

tion. An acoustic environment can be simulated by

distributing sound among different loudspeakers

and by adjusting the ratio between direct and rever-

berated sound (Chowning 1971). Most of the studio

reverberators sold today use digital technology.

The two elementary filters used in reverberation

are shown in Fig. 7. The first filter is called a comb

filter; in it, the signal is delayed a certain number of

samples, attenuated, and added to the input. An ex-

(a)

+ )( Delay

(b)

- G

ponentially decaying, repeated echo is so obtained.

The frequency response is characterized by equi-

spaced peaks-hence this filter's name. The peaks'

amplitude increases as G approaches 1.

The second filter is called an all-pass filter, since

the frequency response is flat and there is only a

phase shift. The input signal is attenuated and sub-

tracted from the delayed signal so that the feedback

effect is compensated and the echoes are main-

tained. The all-pass property is valid only in the

steady state with stationary sounds, not in tran-

sient states. Thus, it has a well-defined sound qual-

ity that a skilled listener can easily distinguish.

Reverberators are built combining some of these

filters (Moorer 1979b). Distinguishable signal repe-

titions should not occur in them, since the rever-

berated result should consist of a diffused sound.

The delay time of each elementary filter has to be

chosen very carefully. Sometimes a nonrecursive

echo generator is added to produce the first aperi-

odic echoes, which are the main perceptual deter-

minants of the characteristics of the room.

Nonlinear Techniques

In addition to linear transformations, which are

used in other fields and have a rather developed the-

ory, nonlinear transformations are used more and

DePoli 15

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 8. Waveshaping.

more commonly in musical applications. They

derive mainly from electrical communication the-

ory, and they have proved to be promising and effec-

tive. One use of nonlinear synthesis is in the large

amount of computer music generated by frequency

modulation (FM) synthesis (Chowning 1973).

In the classic case, nonlinear techniques use sim-

ple sinusoids as input signals. The output is com-

posed of many sinusoids, whose frequency and

amplitude depend mostly on the input ones.

Two main types of nonlinear techniques can be

distinguished, waveshaping and modulation. In

waveshaping, one input is shaped by a function de-

pending only on the input value in that instant. In

modulation (with two or more inputs), a simple pa-

rameter of one signal, called the carrier, is varied

according to the behavior of another signal, called

the modulator. In electrical communications (e.g.,

radio) the spectra of the signals are clearly distin-

guished and therefore easily separable. The origi-

nality in computer music application is the utiliza-

tion of signals in the same frequency range. Thus,

the two signals interact in a complex way, and

simple input variation affects all the resultant

components.

Often, the input amplitudes are varied by multi-

plying them by a constant or time-dependent pa-

rameter I, called the modulation index. Thus,

acting only on one parameter, the sound charac-

teristics are substantially varied. Dynamic and

variable spectra are easily obtainable. In additive

synthesis, similar variations require a much larger

amount of data.

Waveshaping

A linear filter can change the amplitude and phase

of a sinusoid, but not its waveform, whereas the

aim of waveshaping is to change the waveform. The

distortion of a signal heard from a nonlinear ampli-

fier is common. The output from a nonlinear am-

plifier of a sinusoidal signal is a signal with the

same period, but with a different waveform. The

various harmonics are present, and their amplitude

depends on the input and on the distortion. In stereo

systems, these distortions are usually avoided,

x(t)

y(t)y~t)

while waveshaping (Arfib 1979; Le Brun 1979;

Roads 1979) exploits them to generate periodic

sounds, rich in harmonics, from a simple sinusoid.

The function F(x), describing distortion, is called

the shaping function, and it associates with each in-

put value the corresponding output value indepen-

dent of time. If the input is x(t) = cos(27r ft), the

output is

s(t) = F(x(t)) = F(cos[2rr ft]).

In analog synthesis, it is difficult to have an am-

plifier with a precise and variable distortion char-

acteristic. In digital synthesis, this technique is

extremely easy to implement (Fig. 8). As in the case

of the oscillator, the shaping function can be previ-

ously computed and stored in a table. All that is

necessary is to look up the proper value from the

table.

Generally, if F(x) = F,(x) + F2(x), the distortion

produced by F is equal to the sum of those pro-

duced by F, and F2 separately. Usually, the shaping

produces infinite harmonics. But when a poly-

nomial of degree N is chosen as shaping function,

only the first N harmonics are present. Thus, fold-

over is easily avoided. Arfib and Le Brun deal exten-

sively with the mathematical relations among the

coefficients di of the shaping polynomial and the

amplitudes hi of harmonics generated when the am-

plitude I of the cosinusoidal input varies.

The shaping function, producing the jth har-

monic, is the Chebychev polynomial 7T(x) of degree

i (Fig. 9). Thus, to obtain the various harmonics of

16 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 9. Chebychev poly-

nomial of degree K used as

shaping function produces

only the Kth harmonic. In

the figure, K = 3.

T3(cos[ot]) = cos(3ot)

T3(x) = 4X3 - 3x

X = cos(wt) wt

01 1

-rr/6 -

2I/37T/3\_ os( ut)

73/2

27r/3

1lT/6

77r/6

137T/6

ot

amplitude h,, it is sufficient to add the correspon-

dent Chebychev polynomials, each multiplied

by hi:

N N

F(x) = h(x) = d;x.=0 i=O0

From these relations, it follow

monics are comp

even polynomial

harmonics. In th

only the odd ha

cient of x7 affect

enth harmonics

harmonic of ord

odd) coefficients

DePoli 17

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

example, the seventh harmonic is affected by the

odd coefficients from the seventh up to the degree

of the polynomial.

When the input amplitude I varies, the distortion

and the output spectrum vary. This is similar to an

expansion or contraction of the function, since

greater or smaller range of the function is employed.

From a mathematical point of view, the amplitude

variation corresponds to the multiplication of each

polynomial coefficient d, by II. The amplitudes of

the even or odd harmonics depend on I according

to the even (or odd) polynomials, which contain

the terms from the harmonic order up to the

polynomial degree.

If the spectrum is rather smooth, the number of

significant harmonics increases with the index.

Thus, a typical characteristic of real instruments is

reproduced, in that amplitude and spectrum are cor-

related. The amplitude and loudness of the output

vary with the input amplitude. In simple cases, this

effect can be compensated for by multiplying the

output by a suitable normalization function. But in

musical applications, the amplitude of the signal is

rarely constant, and it is multiplied by an envelope.

Normalization can be avoided by combining it with

the amplitude envelope in experimental or intuitive

ways after considering the normalization function.

It is also advisable to choose the even (or odd)

polynomial coefficients with alternating signs, that

is, according to the following model: + + - -

+ + - -. It is also advisable that the hi amplitude

not decrease abruptly, sharply limiting the band.

Otherwise, a spectrum would result that varied

very irregularly with I.

Dynamic spectral behavior cannot be easily an-

ticipated from the coefficients or from the static

spectrum. Moreover, the same (absolute-value)

spectrum can be produced by many polynomials

with different dynamic behaviors (Forin 1982). With

waveshaping, listening and graphic considerations

have more relevance than purely mathematical

formulations.

Another dynamic variation of waveshaping that

is easy to implement occurs when a constant is

added to the input; the shaping function shifts hori-

zontally. Even in this case, the spectrum varies.

The signal is periodic, with the same number of

harmonics. But in this case, the harmonic behavior

depends on both the even and the odd coefficients.

Generalizations of waveshaping technique are

possible. Reinhard (1981) studied the relations that

produce the partials generated by the polynomial

distortion of two cosine waves of frequency f, and

f2. All the components of frequency 1K f, ? Jf2

with IK + j|l N, where N is the polynomial de-

gree, are present.

Shaping functions that are not polynomial can

be used if the spectra produced by them are almost

band limited. Of particular interest is the use of

trigonometric and exponential functions (Moorer

1977) and of those where the input also appears in

the denominator (Winham and Steiglitz 1970;

Moorer 1976; Lehmann and Brown 1976; De Poli

1981).

Due to the wide spectral variation induced by

only one parameter (amplitude or shift), wave-

shaping is particularly convenient in musical ap-

plications, especially in combination with multi-

plicative synthesis. Moreover, it is suitable for

modeling the sound production of some acoustic

instruments (Beauchamp 1979, 1982). There is a

large and not intuitive problem in choosing the co-

efficients, however, and further research is required.

Multiplicative Synthesis (Ring Modulation)

The simplest nonlinear transformation consists of

the multiplication of two signals. In analog syn-

thesizers, it is called ring modulation (RM). Some-

times it is also called amplitude modulation (AM),

but the two differ, especially in their realization.

With two inputs x,(t) and x2(t), the output is

s(t) = Xl(t) . x2(t). Obviously, when the inputs inter-

change, the result does not vary. The resulting spec-

trum is obtained from the convolution of the two

signals' spectra. Usually, one of the two signals,

called the carrier, is sinusoidal; the result is not

too complex and noisy.

When x1 is the sinusoidal carrier of frequency f/,

and x2 (modulator) is sinusoidal with frequency f2,

from cos(a) , cos(3) = 12lcos(a + 3) + cos(a - /)1,

18 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 10. Multiplicative syn-

thesis. Spectrum of a peri-

odic signal X2 with four

harmonics (a). Resulting

spectrum when d2 is mul-

tiplied by a sinusoid of fre-

quency f, greater than its

bandwidth (f, = 7f2) (b).

Resulting spectrum when

x2 is multiplied by a sinu-

soid of frequency inferior

to its bandwidth (f, =

26f2) (c). The components

deriving from the folding

of negative frequencies are

shown as dashed lines.

the output consists of two sinusoidal partials of fre-

quency fI + f2 and f, - f2. The phases of the output

are also the sum and the difference of the phases of

the two inputs. For example, if x, and x2 frequen-

cies are 400 Hz and 100 Hz, the output has two par-

tials of frequency 500 Hz and 300 Hz.

Negative frequencies may occur, for example,

when f, = 100 Hz and f, = 400 Hz. This often hap-

pens in modulations (foldunder) and can be ex-

plained by the trigonometric relation cos(a)

= cos(-a), from which cos(27r ft + (4) = cos(7r[-flt

- 4). The alteration of the frequency sign only

changes the sign of the phase with respect to the

cosine. In particular, a cosine signal is unaffected,

while a sine wave changes its sign. In the inter-

pretation of the results, only absolute frequency

values have to be considered. Usually, the phase is

not significant, as the ear is not terribly sensitive to

it. But the phase has to be taken into account while

summing the amplitude of components of identical

frequencies.

In multiplicative synthesis, usually x2 is periodic

with frequency f2. The multiplication causes every

harmonic spectral line of frequency K . f2 in the

original signal to be replaced by two spectral lines

(called sidebands) of frequency f, + K f2 and

f, - K f2. The resulting spectrum has components

of frequency If +? K f2 , where K is equal to the or-

der of the different harmonics in x2 (Fig. 10).

Thus two sidebands, symmetric with respect to

the carrier, occur. When f, is less than the greatest

frequency in x2, then the negative frequencies fold

around zero, as discussed above.

The possibility of shifting the spectrum is very

intriguing in musical applications. From simple

components, harmonic and inharmonic sounds can

be created, and various harmonic relations among

the partials can be established. If x2 is a signal with

spectrum X2, the signal obtained from its multi-

plication with a sinusoid of frequency f, has two

sidebands symmetric with respect to f, and shaped

like X2.

A periodic signal x, can be expanded in Fourier

series. Each x, partial will have sidebands of ampli-

tude proportional to its own. If f1 is less than the

bandwidth of x,, then the sidebands overlap with

(a) jx2(f)l

(b)

I|S(f)l

ff2

(c)

S(f)I

Iu I li I fl

eventual component superimposition. In this case,

the phases have to be taken in account while sum-

ming. Dashow (1978, 1980) describes some general-

ization of this technique and employs the generated

spectra for particular "harmonizations" of pitches

specified by the composer.

Amplitude Modulation

In RM, the carrier does not appear in the spectrum

created by the product of a sinusoidal carrier with

another signal, except when the modulator has a di-

rect current (dc) component. In carrying out the

modulation in AM (Fig. 11), the carrier is present in

the output, with an amplitude independent of the

sidebands. The formula for AM is as follows:

s(t) = xl(t) . (K + x2(t)).

The result is RM with carrier added. When the

carrier is sinusoidal and the modulator is periodic,

the spectrum is composed of partials of frequency

If1 1 K f,2, with K = 0, 1, . It is useful to distin-

DePoli 19

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 11. Amplitude

modulation.

K A2 f2 f1

X2(t)

x1(t)

s(t)

guish between the two modulations because they

have different realization schemes.

Spectra of Type If , K f2

The following considerations are valid for all spec-

tra whose components are of type If, + K f2h, with

K = 0, 1, .... The spectrum is characterized by the

ratio f,/f2. (This is often referred to as the carrier-

to-modulator [c:m] ratio.) When this ratio is ra-

tional, it can be expressed as an irreducible fraction

fI/f2 = NI/N2, with N, and N2 as integers that are

prime between themselves. In this case, the result-

ing sound is harmonic, since the various compo-

nents are a multiple of a fundamental according to

integer factors. The fundamental frequency is

fo = - /

N1 N2 '

and the carrier coincides with the NIth harmonic.

If N2 = 1, all the harmonics are present and the

sideband components coincide. If N2 = 2, only odd

harmonics are present and the sidebands superim-

pose. If N2 = 3, the harmonics that are multiples of

3 are missing. The c: m ratio is also an index of the

harmonicity of the spectrum. The sound is mo

"harmonious" intuitively when the N /N2 rati

simple and formally when the N, N2 productsmaller.

The ratios can be grouped in families (Truax

All ratios of the type If, K f2l/f2 can produce

same components that flf, produces. Only th

tial coinciding with the carrier (f,) changes. Fo

ample, the ratios 2/3, 5/3, 1/3, 4/3, 7/3 and s

all belong to the same family. Only the harmo

that are multiples of 3 are missing (see N2 = 3

the carrier is respectively the second, fifth, fir

fourth, seventh, and so on harmonic.

The ratio that distinguishes a family is defin

normal form when it is - 1/2. In the previous

ample, it is 1/3. Each family is characterized b

ratio in normal form. Similar spectra can be p

duced using ratios from the same family. Diffe

spectra are obtained by sounds of different fa

When the fl/f2 ratio is irrational, the resulti

sound is aperiodic and hence, inharmonic. Of p

ticular interest is the case of an f,/f2 ratio app

mating a simple value, that is,

fl/f2 = N1/N2 + e.

Here the sound is no longer rigorously periodi

The fundamental frequency fo is still f2/N2,

the harmonics are shifted from their exact valu

by +e / f2. When N2 is equal to 1 or 2, the po

and negative components are not superimposed

beat with a frequency of 2e / f2. Hence, a smal

of the carrier does not change the pitch, even

slightly spreads the partials and makes the sou

more lively. But the same shift of the modula

frequency f, changes the sound's pitch.

Frequency and Phase Modulation

Another type of modulation, suggested by Cho

ing (1973), has become one of the most widely

synthesis techniques. In general, it consists of

modulation and it can be realized both as ph

modulation (4M) or as FM. This technique do

not derive from models of production of phys

sounds, but only from the mathematical prope

ties of a formula. It has some of the advantage

20 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 12. The number of sig-

nificant sidebands in FM.

waveshaping and RM, and it avoids some of their

drawbacks.

The technique consists of the modulation of the

instantaneous phase or frequency of a sinusoidal

carrier according to the behavior of another signal

(modulator), which is usually sinusoidal. It can be

expressed as follows:

s(t) = sin(2zT f t + I sin[27r fmt]) =

=>K JK (I)sinj27T(f, + Kfj)tI.

The resulting spectrum is of the type If, +? K f,,. All

the spectral considerations discussed previously are

applicable, particularly those regarding negative fre-

quency, foldunder, fclf, ratios, and harmonic andinharmonic sounds.

The amplitude of each Kth side component of

the FM technique is given by the Bessel function of

Kth order computed in I. To plot the spectrum, a

table of Bessel functions has to be referenced to ob-

tain the amplitudes of the carrier and of the side

frequencies in the upper sideband. The odd-order

side frequencies in the lower sideband have signs

opposite to those in the upper one, and the even-

order side frequencies have the same sign. The

negative frequencies, being sine waves, are folded,

changing the sign. When superimposition occurs,

the amplitudes are added algebraically.

When I (called the modulation index) varies, the

amplitude of each component varies as well. Thus,

dynamic spectra can be obtained simply by varying

this index. Each component varies its amplitude

by following the corresponding Bessel function. A

Bessel function can be asymptotically approxi-

mated by a damped sinuosoid. So when the index

varies, some components increase and others de-

crease, all without sharp variations.

In Eq. (1), the sum includes infinite terms, so the-

oretically the signal bandwidth is not limited. But,

practically, it is limited. In the Bessel function's

behavior, only a few low-order functions are sig-

nificant for small index values. When the index in-

creases, the number and the order of the significant

functions increase. For a given index, the side am-

plitudes oscillate with gradually increasing ampli-

tude and slowly increasing period all the way from

25

20 /

15 /

5

M ' I I

I 5 10 15 20

the origin to a

toward zero. Th

slightly below

Usually, in the

signal, all side f

than /loo of th

ered. The numb

M = I + 2.4 J10.27

(See Fig. 12.) Often, as a rule of thumb, it is roughly

considered as

M = I+ 1.

In Eq. (1), the sum can be performed for K fr

-M to +M. For a harmonic sound, that is, w

the ratio fc/fm = NI/N2 is simple, the maxim

der of significant harmonics is N, + M ' N2.

For wide index variations, the sounds produc

are characteristic of the FM technique. A typi

timbre of FM sound is easily recognizable an

well defined. This does not happen for small i

variations or for compound carriers or modul

Frequency modulation synthesis has another p

DePoli 21

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 13. Frequency

modulation.

A(t) fc d(t) fm

s(t)

Fig. 14. Frequency modu-

lation with N carriers

modulated by the same

oscillator.

a, f/ a2 f2 aN fN d f,

+ + +

s(t)

erty that is very important in musical applications:

the maximum amplitude and the signal power do

not vary with the index I. Unlike the situation in

waveshaping, normalization of the output is not

necessary.

Let us now examine the difference between OM

and FM. Phase modulation is defined as follows:

s(t) = sin (2M ft + 0[ t]),

and it corresponds to Eq. (1) if the modulating sig-

nal is 0(t) = I sin(2rr f, t).

Frequency modulation occurs when the instanta-

neous frequency varies around the carrier value ac-

cording to the behavior of the modulating wave. For

a signal s(t) = sin(p[ t]), the instantaneous frequency

is fj = (1/27r) (di[t]l/dt). Thus, the instantaneous

frequency of the signal in Eq. (1) is as follows:

fi = fc + I fm , cos(2r f,, t).

The frequency varies around f, with a maximum

deviation d = I f,. Thus, with a modulating wave

I - f, cos(27r f, t), an FM equivalent to OM is ob-

tained. Both phase and frequency modulations are

special cases of angle modulation.

In sound synthesis programs, frequency-driven

oscillators are provided. The integration involved in

calculating the instantaneous phase is therefore

computed automatically. Frequency modulation is

normally implemented as in Fig. 13. A change of

the phase between the carrier and the modulating

wave in Eq. (1) only changes the reciprocal phase of

the partials. If components superimpose, their total

amplitude changes, and a direct-current component

may appear. The next sections examine some use-

ful extensions of the basic algorithm.

Nonsinusoidal Carrier

Here we consider a periodic nonsinusoidal carrier.

The result of its modulation is the modulation of

each of its harmonics by the same wave. Sidebands

of amplitude proportional to each harmonic will be

present around the carrier. The result is a spectrum

with components of frequency In f, + K . fmi,, with

K = 0,... , M and n = 1,..., N, when Nis the

number of significant harmonics. The maximum

frequency present is N - f, + M . f,. In general, there

may be various independent carriers modulated by

the same wave (Fig. 14) or by different modulating

signals. This is like additive synthesis, only instead

of sinusoidal addends, more complex addends are

used. For example, harmonic sounds can be gener-

ated by controlling the various spectral ranges with a

few significant and independent parameters. Sounds

of the same "family" are possible.

The frequency of each carrier determines the

22 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 15. Frequency modu-

lation with two modu-

lators.

Fig. 16. Frequency modu-

lation with N modulators.

A fc dl(t) f, d2(t) f2

s(t)

location of the formant position, the amplitude

determines its energy, and the modulation index

specifies its bandwidth. Chowning (1981) demon-

strated these facilities in synthesis of the singing

voice of a soprano.

Compound Modulation

Let us examine the case of a modulation composed

of two sinusoids (Fig. 15), each with its own modu-

lation index, applied to a sinusoidal carrier. The for-

mula for two-sine-wave OM (Le Brun 1977) is as

follows:

s(t) = sin(121T fct + I, sin[27r f t] + 12 sin[27r f2t])

= K n, JK(I1) Jn(12) . sin(I27T[fc + K f, + n f2ltI).

The same result can be obtained with FM using as

modulating signal the following expression:

I,f,/cos(27r f, t) + I2f2cos(2r f2t).

The resulting spectrum is much more complex

than in the one-modulator case. All the compo-

nents of frequency If/ ? K f, n f,2 are present, and

their amplitude is JK(I) " Jn(I2).

To interpret the effect, let us consider f1 > f2. If

only f, were present, the resulting spectrum would

have a certain number of components of amplitude

JK(I1) and frequency fc + K f1. When the modulator

A fc d1(t) f, d2(t) f2 dM(t) fM

s(t)

f, is applied, these components become carriers,

with sidebands produced by f,. The resulting band-

width is approximately equal to the sum of the two

bandwidths.

If the frequencies have simple ratios, the spec-

trum is of the type If, -K f,,, where now f, is the

greatest common divisor of f, and f,. For example,

with f, = 700 Hz, f = 300 Hz, and f2 = 200 Hz, the

components are 1700 ? K 1001. Thus, by choosing

f, and f2 multiples of f,, sounds belonging to the

same family as a simple modulation, but with a

more complex spectral structure, can be generated.

In general, if the modulating signal is composed of

N sinusoids (Fig. 16), the following relations hold:

s(t) = sin (2rT fct + EI sin[2r fs t]

N

= KS, HKS (Is) sin|2ir(fct + Y Ifs)t1.

Thus, all the components of frequency If , Klf/ -. . ? KNfNI, with amplitudes given by the product

of N Bessel functions, are obtained. A very complex

spectrum results. If the relations among the fre-

quencies f, are simple, that is, if the modulating

wave is periodic, then the spectrum is of the type

Ifc ? K fm, where f,m is the greatest common divisor

among the modulating components. Otherwise, the

sonorities are definitely inharmonic and particu-

larly noisy for high indexes.

DePoli 23

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 17. Nested FM.

Nested or Complex Modulation

Let us examine the case of a sinusoidal modulator

that is phase modulated by another sinusoid. The

signal is defined as follows:

s(t) = sin(I27r ft + I sin[2ir f, t + 12 sin{2ir f2t}]))

= K JK(Il)sin(I2zr[fc + K f,]t

+ K 12 sin[27r f2t]I)

= 1K J(I,)" Jn(KI2)sin(2r[fc + K,f, + n f2]t).

The result can be interpreted as if each partial pro-

duced by the modulator f, were modulated in its

turn by f2 with modulation index K I,. Thus, all

the partials of frequency If +? K f, n f2j, with ap-

proximately 0 - K - I, Os n In , 1 I2, are present.

The maximum frequency is fc + I1I(f + 12/2).

The structure of the spectrum is similar to that

produced by the two-sinusoid modulation, but with

a larger bandwidth. Even where fm is the greatest

common divisor between f, and f2, the spectrum is

of the type If, + K fn-.

In the equivalent realization by FM (Fig. 17), the

spectrum is of the same type, but with slightly dif-

ferent amplitudes. A direct-current component in

the resulting modulating wave added to the carrier

is avoided by choosing a sine wave modulated by a

cosine wave.

This technique is made more interesting by an

algorithm suggested by Justice (1979), which en-

ables an analysis of a sound according to this model,

with the frequency and the index behavior of two

or more nested modulators being deducible.

Other Two-Input, Nonlinear Transformations

Mitsuhashi (1980) proposed a more complex two-

input, nonlinear transformation, in which the in-

stantaneous phase and amplitude of an approx-

imately sinusoidal signal are simultaneously varied.

In another paper, Mitsuhashi (1982c) generalized

this technique while discussing some criteria in

choosing the two-input, nonlinear function and

suggesting two examples. The function is time in-

dependent, bidimensional, and considered periodic

outside the definition field. Thus, it can be imple-

mented with a two-dimensional table, with analogy

to an oscillator. This technique appears very inter-

A fc d (t) f/ d2(t) f,

+

s(t)

esting, even if it seems to be difficult to find a sim-

ple expression that bounds significant parameters of

the resulting spectrum to the input and function

characteristics. Another promising modulation

technique is linear sweep synthesis, recently sug-

gested by Rozenberg (1982).

Conclusion

As a consequence of progress in digital hardware

and software, the initial antithesis between com-

puting efficiency and timbral richness is lessening.

Digital sound quality largely depends on the amount

of introduced or controlled detail; excessive sim-

plifications lead often to trivial results. It follows

that increased computing power can generate more

sophisticated results.

A musically interesting sound can be obtained in

two ways. The first consists of the utilization of

more complex techniques or of the combination of

24 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

many of the techniques described here. Many linear

and nonlinear transformations are possible. Most of

the parameters do not have to be constant and can

be varied by control functions and random signals.

The other synthesis approach consists of the su-

perimposition of many simple sounds produced by

basic techniques. The evolution of the individual

sounds is not complex, and the richness of the re-

sult essentially depends on their combination. In

this approach, the parameters of many elementary

sounds have to be given. Specific programs are often

used to define these parameters.

Sound evolution can be regulated either by con-

trol functions in the synthesis or by programs com-

puting the parameters for the synthesis. In any

case, many details of the sound have to be accu-

rately controlled. Their coherence both within the

sound and in the context of adjacent and simul-

taneous notes has to be guaranteed. The relations

among sounds can be more easily highlighted when

they are reflected not only in macroscopic param-

eter variations but also in internal structure.

The extensive utilization of a single technique re-

veals its peculiar characteristics. This derives from

the finite repertoire of obtainable sounds and, more

specifically, from the more easily producible dy-

namic variations associated with it. Thus, it is wise

to use different techniques, the better to exploit

their different potential. Moreover, the musician

must study and experiment with a technique. This

is essential in order to determine all its charac-

teristics and to acquire a feeling for the parameter

choices necessary for nontrivial use. In any case, a

synthesis technique is simply a tool to produce

sound, and sound is not yet music.

References

Arfib, D. 1979. "Digital Synthesis of Complex Spectra by

Means of Multiplication of Nonlinear Distorted Sine

Waves." Journal of the Audio Engineering Society

27(10): 757-768.

Bass, S. C., and T. W. Goeddel. 1981. "The Efficient Digi-

tal Implementation of Subtractive Music Synthesis."

IEEE Micro 1(3) :24-37.

Beauchamp, J. W. 1975. "Analysis and Synthesis of Cor-

net Tones Using Non Linear Interharmonic Rela-

tionships." Journal of the Audio Engineering Society

23(10): 778-795.

Beauchamp, J. W. 1979. "Brass Tone Synthesis by Spec-

trum Evolution Matching with Nonlinear Functions."

Computer Music Journal 3(2): 35-43.

Beauchamp, J. W. 1982. "Synthesis by Spectral Amplitude

and 'Brightness' Matching of Analyzed Musical Instru-

mental Tones." Journal of the Audio Engineering So-

ciety 30(6):396-406.

Cann, R. 1979-1980. "An Analysis Synthesis Tutorial."

Part 1, Computer Music Journal 3(3):6-11; Part 2,

Computer Music Journal 3(4):9-13; Part 3, Computer

Music Journal 4(1):36-42.

Cerruti, R., and G. Rodeghiero. 1983. "Comments on

'Musical' Sound Synthesis by Forward Differences."

Journal of the Audio Engineering Society 31(6).

Charbonneau, G. 1981. "Three Types of Data Reduction."

Computer Music Journal 5(2): 10-19.

Chowning, J. M. 1971. "The Simulation of Moving Sound

Sources." Journal of the Audio Engineering Society

19(1): 2-6. (Reprinted in Computer Music Journal

1[3]:48-52, 1977.)

Chowning, J. M. 1973. "The Synthesis of Complex Audio

Spectra by Means of Frequency Modulation." Journal of

the Audio Engineering Society 21(7): 526-534. (Re-

printed in Computer Music Journal 1[2]:46-54, 1977.)

Chowning, J. M. 1981. "Computer Synthesis of the Sing-

ing Voice." In Sound Generation in Winds Strings

Computers. Stockholm: KTH Skriftserie 29, pp. 4-13.

Dashow, J. 1978. "Three Methods for the Digital Syn-

thesis of Chordal Structure with Non-Harmonic Par-

tials." Interface 7(2/3):69-94.

Dashow, J. 1980. "Spectra as Chords." Computer Music

Journal 4(1):43-52.

De Poli, G. 1981. "Sintesi di suoni mediante funzione dis-

torcente con poli complessi coniugati." Atti del IV Col-

loquio di Informatica Musicale 1, Pisa, pp. 103-130.

De Poli, E., and G. De Poli. 1979. "Identificazione di

parametri di un oscillatore VOSIM a partire da una

descrizione spettrale." Atti del III Colloquio di

Informatica Musicale, Pisa, pp. 161-177.

Forin, A. 1982. "Spettri dinamici prodotti mediante dis-

torsione con polinomi equivalenti in un punto." Bollet-

tino LIMB 2:62-76.

Grey, J. M., and J. A. Moorer. 1977. "Perceptual Eva

tion of Synthesized Musical Instrument Tones." Jou

nal of the Acoustical Society of America 62:434-

Justice, J. M. 1979. "Analytic Signal Processing in M

Computation." IEEE Transactions on Acoustics, S

and Signal Processing (ASSP) 27(6):670-684.

DePoli 25

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Kaegi, W. 1973. "A Minimum Description of the Linguis-

tic Sign Repertoire (part 1)." Interface 2:141-156.

Kaegi, W. 1974. "A Minimum Description of the Linguis-

tic Sign Repertoire (part 2)." Interface 3: 132-158.

Kaegi, W., and S. Tempelaars. 1978. "VOSIM-A New

Sound Synthesis System." Journal of the Audio Engi-

neering Society 26(6): 418-424.

Le Brun, M. 1977. "A Derivation of the Spectrum of FM

with Complex Modulating Wave." Computer Music

Journal 1(4):51-52.

Le Brun, M. 1979. "Digital Waveshaping Synthesis." Jour-

nal of the Audio Engineering Society 27(4):250-265.

Lehmann, R., and F. Brown. 1976. "Synthese rapide des

sons musicaux." Revue d'Acoustique 38:211-215.

Lorrain, D. 1980. "A Panoply of Stochastic 'Cannons.'"

Computer Music Journal 4(1):53-81.

Mailliard, R. 1976. "Les distorsions de Music V." Cahiers

recherche/musique 3:207-246.

Mathews, M. V. 1969. The Technology of Computer Mu-

sic. Cambridge, Massachusetts: MIT Press.

Mitsuhashi, Y. 1980. "Waveshape Parameter Modulation

in Producing Complex Audio Spectra." Journal of the

Audio Engineering Society 28(12): 879-895.

Mitsuhashi, Y. 1982a. "Musical Sound Synthesis by For-

ward Differences." Journal of the Audio Engineering

Society 30(1/2): 2-9.

Mitsuhashi, Y. 1982b. "Piecewise Interpolation Tech-

nique for Audio Signal Synthesis." Journal of the Audio

Engineering Society 30(4): 192-202.

Mitsuhashi, Y. 1982c. "Audio Signal Synthesis by Func-

tions of Two Variables." Journal of the Audio Engineer-

ing Society 30( 10): 701 - 706.

Moorer, J. A. 1976. "The Synthesis of Complex Audio

Spectra by Means of Discrete Summation Formulae."

Journal of the Audio Engineering Society 24(9):

717-727.

Moorer, J. A. 1977. "Signal Processing Aspects of Com-

puter Music: A Survey." Proceedings of the IEEE

65(8): 1108-1132. (Reprinted in Computer Music Jour-

nal 1[1]:4-37, 1977.)

Moorer, J. A. 1979a. "The Use of Linear Prediction of

Speech in Computer Music Applications." Journal of

the Audio Engineering Society 27(3): 134-140.

Moorer, J. A. 1979b. "About This Reverberation Busi-

ness." Computer Music Journal 3(2): 13-28.

Reinhard, P. 1981. "Distorsione non lineare della somm

di due cosinusoidi: analisi dello spettro tramite ma-

trici." Atti del IV Colloquio di Informatica Musicale

Pisa, pp. 160-183.

Risset, J.-C. 1969. "An Introductory Catalog of Comput

Synthesized Sounds." Murray Hill, New Jersey: Bell

Laboratories.

Risset, J.-C., and M. V. Mathews. 1969. "Analysis of Mu-

sical Instrument Tones." Physics Today 22(2):23-30.

Roads, C. 1978. "Automated Granular Synthesis of

Sounds." Computer Music Journal 2(2):61-62. Revised

and updated version forthcoming in C. Roads and

J. Strawn, eds., Foundations of Computer Music. Cam-

bridge, Massachusetts: MIT Press.

Roads, C. 1979. "A Tutorial on Non-linear Distortion or

Waveshaping Synthesis." Computer Music Journal

3(2): 21-34.

Rodet, X. 1980. "Time Domain Formant Wave-Function

Synthesis." In Spoken Language Generation and Un-

derstanding, ed. J. G. Simon. Dordrecht: D. Reidel.

Rozenberg, M. 1979. "Microcomputer-controlled Sound

Processing Using Walsh Functions." Computer Music

Journal 3(1):42-47.

Rozenberg, M. 1982. "Linear Sweep Synthesis." Com-

puter Music Journal 6(3): 65-71.

Tempelaars, S. 1976. "The VOSIM Signal Spectrum." In-

terface 6:81-86.

Tempelaars, S. 1982. "Linear Digital Oscillators." Inter-

face 11(2): 109-130.

Truax, B. 1977. "Organizational Techniques for C: M

Ratios in Frequency Modulation." Computer Music

Journal 1(4):39-45.

Winham, G., and K. Steiglitz. 1970. "Input Generators for

Digital Sound Synthesis" (Part 2). Journal of the Acous-

tical Society of America 47(2):665-666.

Xenakis, I. 1971. Formalized Music. Bloomington, In-

diana: Indiana University Press.

26 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Designing Calm Technology http://www.ubiq.com/weiser/calmtech/calmtech.htm

1 of 5 2/4/04 4:21 PM

Designing Calm Technology

Mark Weiser and John Seely Brown

Xerox PARC

December 21, 1995

Introduction

Bits flowing through the wires of a computer network are ordinarily invisible. But a radically new tool

shows those bits through motion, sound, and even touch. It communicates both light and heavy network

traffic. Its output is so beautifully integrated with human information processing that one does not even

need to be looking at it or near it to take advantage of its peripheral clues. It takes no space on your

existing computer screen, and in fact does not use or contain a computer at all. It uses no software, only

a few dollars in hardware, and can be shared by many people at the same time. It is called the "Dangling

String".

Created by artist Natalie Jeremijenko, the "Dangling String" is an 8 foot piece of plastic spaghetti that

hangs from a small electric motor mounted in the ceiling. The motor is electrically connected to a nearby

Ethernet cable, so that each bit of information that goes past causes a tiny twitch of the motor. A very

busy network causes a madly whirling string with a characteristic noise; a quiet network causes only a

small twitch every few seconds. Placed in an unused corner of a hallway, the long string is visible and

audible from many offices without being obtrusive. It is fun and useful. The Dangling String meets a

key challenge in technology design for the next decade: how to create calm technology.

We have struggled for some time to understand the design of calm technology, and our thoughts are still

incomplete and perhaps even a bit confused. Nonetheless, we believe that calm technology may be the

most important design problem of the twenty-first century, and it is time to begin the dialogue.

The Periphery

Designs that encalm and inform meet two human needs not usually met together. Information

technology is more often the enemy of calm. Pagers, cellphones, newservices, the World-Wide-Web,

email, TV, and radio bombard us frenetically. Can we really look to technology itself for a solution?

But some technology does lead to true calm and comfort. There is no less technology involved in a

Designing Calm Technology http://www.ubiq.com/weiser/calmtech/calmtech.htm

2 of 5 2/4/04 4:21 PM

comfortable pair of shoes, in a fine writing pen, or in delivering the New York Times on a Sunday

morning, than in a home PC. Why is one often enraging, the others frequently encalming? We believe

the difference is in how they engage our attention. Calm technology engages both the center and the

periphery of our attention, and in fact moves back and forth between the two.

We use "periphery" to name what we are attuned to without attending to explicitly. Ordinarily when

driving our attention is centered on the road, the radio, our passenger, but not the noise of the engine.

But an unusual noise is noticed immediately, showing that we were attuned to the noise in the periphery,

and could come quickly to attend to it.

It should be clear that what we mean by the periphery is anything but on the fringe or unimportant. What

is in the periphery at one moment may in the next moment come to be at the center of our attention and

so be crucial. The same physical form may even have elements in both the center and periphery. The ink

that communicates the central words of a text also, though choice of font and layout, peripherally clues

us into the genre of the text.

A calm technology will move easily from the periphery of our attention, to the center, and back. This is

fundamentally encalming, for two reasons.

First, by placing things in the periphery we are able to attune to many more things than we could if

everything had to be at the center. Things in the periphery are attuned to by the large portion of our

brains devoted to peripheral (sensory) processing. Thus the periphery is informing without

overburdening.

Second, by recentering something formerly in the periphery we take control of it. Peripherally we may

become aware that something is not quite right, as when awkward sentences leave a reader tired and

discomforted without knowing why. By moving sentence construction from periphery to center we are

empowered to act, either by finding better literature or accepting the source of the unease and continuing.

Without centering the periphery might be a source of frantic following of fashion; with centering the

periphery is a fundamental enabler of calm through increased awareness and power.

Not all technology need be calm. A calm videogame would get little use; the point is to be excited. But

too much design focuses on the object itself and its surface features without regard for context. We must

learn to design for the periphery so that we can most fully command technology without being

dominated by it.

Our notion of technology in the periphery is related to the notion of affordances, due to Gibson by

popularized by Norman. An affordance is a relationship between an object in the world and the

intentions, perceptions, and capabilities of a person. The side of a door that only pushes out affords this

action by offering a flat pushplate. The idea of affordance, powerful as it is, tends to describe the surface

of a design. For us the term "affordance" does not reach far enough into the periphery where a design

must be attuned to but not attended to.

Three signs of calm technology

Technologies encalm as they empower our periphery. This happens in two ways. First, as already

mentioned, a calming technology may be one that easily moves from center to periphery and back.

Second, a technology may enhance our peripheral reach by bringing more details into the periphery. An

example is a video conference that, by comparison to a telephone conference, enables us to attune to

nuances of body posture and facial expression that would otherwise be inaccessible. This is encalming

when the enhanced peripheral reach increases our knowledge and so our ability to act without increasing

information overload.

The result of calm technology is to put us at home, in a familiar place. When our periphery is

Designing Calm Technology http://www.ubiq.com/weiser/calmtech/calmtech.htm

3 of 5 2/4/04 4:21 PM

functioning well we are tuned into what is happening around us, and so also to what is going to happen,

and what has just happened. We are connected effortlessly to a myriad of familiar details. This

connection to the world around we called "locatedness", and it is the fundamental gift that the periphery

gives us.

Examples of calm technology

To deepen the dialogue we now examine a few designs in terms of their motion between center and

periphery, peripheral reach, and locatedness. Below we consider inner office windows, Internet

Multicast, and once again the Dangling String.

inner office windows

We do not know who invented the concept of glass windows from offices out to hallways. But these

inner windows are a beautifully simple design that enhances peripheral reach and locatedness.

The hallway window extends our periphery by creating a two-way channel for clues about the

environment. Whether it is motion of other people down the hall (its time for a lunch; the big meeting is

starting), or noticing the same person peeking in for the third time while you are on the phone (they

really want to see me; I forgot an appointment), the window connects the person inside to the nearby

world.

Inner windows also connect with those who are outside the office. A light shining out into the hall

means someone is working late; someone picking up their office means this might be a good time for a

casual chat. These small clues become part of the periphery of a calm and comfortable workplace.

Office windows illustrate a fundamental property of motion between center and periphery. Contrast

them with an open office plan in which desks are separated only by low or no partitions. Open offices

force too much to the center. For example, a person hanging out near an open cubicle demands attention

by social conventions of privacy and politeness. There is less opportunity for the subtle clue of peeking

through a window without eavesdropping on a conversation. The individual, not the environment, must

be in charge of moving things from center to periphery and back.

The inner office window is a metaphor for what is most exciting about the Internet, namely the ability to

locate and be located by people passing by on the information highway.

Internet Multicast

A technology called Internet Multicast may become the next World Wide Web (WWW) phenomenon.

Sometimes called the MBone (for Multicast backBONE), multicasting was invented by a then graduate

Designing Calm Technology http://www.ubiq.com/weiser/calmtech/calmtech.htm

4 of 5 2/4/04 4:21 PM

student at Stanford University, Steve Deering.

Whereas the World Wide Web (WWW) connects only two computers at a time, and then only for the

few moments that information is being downloaded, the MBone continuously connects many computers

at the same time. To use the familiar highway metaphor, for any one person the WWW only lets one car

on the road at a time, and it must travel straight to its destination with no stops or side trips. By contrast,

the MBone opens up streams of traffic between multiple people and so enables the flow of activities that

constitute a neighborhood. Where the WWW ventures timidly to one location at a time before scurrying

back home again, the MBone sustains ongoing relationships between machines, places, and people.

Multicast is fundamentally about increasing peripheral reach, derived from its ability to cheaply support

multiple multimedia (video, audio, etc.) connections all day long. Continuous video from another place

is no longer television, and no longer video-conferencing, but more like a window of awareness. A

continuous video stream brings new details into the periphery: the room is cleaned up, something

important may be about to happen; everyone got in late today on the east coast, must be a big snowstorm

or traffic tie-up.

Multicast shares with videoconferencing and television an increased opportunity to attune to additional

details. Compared to a telephone or fax, the broader channel of full multimedia better projects the person

through the wire. The presence is enhanced by the responsiveness that full two-way (or multiway)

interaction brings.

Like the inner windows, Multicast enables control of the periphery to remain with the individual, not the

environment. A properly designed real-time Multicast tool will offer, but not demand. The MBone

provides the necessary partial separation for moving between center and periphery that a high bandwidth

world alone does not. Less is more, when less bandwidth provides more calmness.

Multicast at the moment is not an easy technology to use, and only a few applications have been

developed by some very smart people. This could also be said of the digital computer in 1945, and of

the Internet in 1975. Multicast in our periphery will utterly change our world in twenty years.

Dangling String

Let's return to the dangling string. At first it creates a new center of attention just by being unique. But

this center soon becomes peripheral as the gentle waving of the string moves easily to the background.

That the string can be both seen and heard helps by increasing the clues for peripheral attunement.

Designing Calm Technology http://www.ubiq.com/weiser/calmtech/calmtech.htm

5 of 5 2/4/04 4:21 PM

The dangling string increases our peripheral reach to the formerly inaccessible network traffic. While

screen displays of traffic are common, their symbols require interpretation and attention, and do not

peripheralize well. The string, in part because it is actually in the physical world, has a better impedance

match with our brain's peripheral nerve centers.

In Conclusion

It seems contradictory to say, in the face of frequent complaints about information overload, that more

information could be encalming. It seems almost nonsensical to say that the way to become attuned to

more information is to attend to it less. It is these apparently bizarre features that may account for why

so few designs properly take into account center and periphery to achieve an increased sense of

locatedness. But such designs are crucial. Once we are located in a world, the door is opened to social

interactions among shared things in that world. As we learn to design calm technology, we will enrich

not only our space of artifacts, but also our opportunities for being with other people. Thus may design

of calm technology come to play a central role in a more humanly empowered twenty-first century.

Bibliography

Gibson, J. The Ecological Approach to Visual Perception. New York: Houghton Mifflin, 1979.

Norman, D.A. The Psychology of Everyday Things. New York: Basic Books, 1988.

MBone. http://www.best.com/~prince/techinfo/mbone.html

Brown, J.S. and Duguid, P. Keeping It Simple: Investigating Resources in the Periphery. To appear in Solving the

Software Puzzle. Ed. T. Winograd, Stanford University. Spring 1996.

Weiser, M. The Computer for the Twenty-First Century. Scientific American. September 1991.

Brown, J.S. http://www.startribune.com/digage/seelybro.htm

Communication, Emotion & EngagementIDC 2018, June 19–22, 2018, Trondheim, Norway300

speech interface versions and a qualitative analysis of their

interactions towards those interfaces. Finally, speech inter-

faces implemented as information appliances (e.g., Amazon

Echo, Google Home) may need to account for the prefer-

ences and needs of both children and adults [3]. As such,

our third research question is (RQ3) How do the experienc-

es and preferences of adults compare to those of children?

We address this question by comparing both qualitative and

quantitative aspects of children’s interaction with our

speech interfaces from those of 27 adults answering the

same informational queries.

In the remainder of this paper, we begin by situating our

work in the body of previous investigations on spoken dia-

log systems, personification and personalization in speech

interfaces, and children’s informational queries. We de-

scribe our methods, including the demographics of our par-

ticipants, the speech interfaces used, and the procedure of

the investigation. We discuss our analysis and findings in

response to each of the above research questions. Finally,

we conclude with a discussion that provides specific impli-

cations for the design of future speech interfaces for homes.

RELATED WORK

We outline previous work in spoken dialog systems and

speech interfaces for children and adults to situate the three

research questions we are exploring in this investigation.

Spoken Dialog Systems & Speech Recognition

Spoken dialog systems and speech recognition systems are

becoming more and more sophisticated, with systems like

Apple Siri, Amazon Alexa, Microsoft Cortana, and Google

Assistant available as off-the-shelf products. Recent ad-

vances in the field focus on personalizing to the needs of

specific users [6], maintaining an understanding of context

over multiple turns of interaction [8], and asking appropri-

ate clarifying questions to guide the user [30]. However,

most of these advances target adult users.

A number of papers provide evidence that speech is a prom-

ising mode of computer interaction for children (e.g., [41]).

Proposed applications of such systems are as diverse as

play/entertainment (e.g., [5,22,28,32,33]), social skill prac-

tice (e.g., [15]), literacy education (e.g., [21,39,45]), and in-

car entertainment (e.g., [18]). However, spoken dialog sys-

tems for children is an area with significant potential for

improvement. First, speech recognition with children is

notoriously difficult (e.g., [17,26,50,55]), especially for

spontaneous utterances which have been described as “dis-

fluent and ungrammatical” [21]. Second, recognizing intent

is significantly more difficult than just recognizing speech,

even with adults [53]. While several investigations have

sought to generate datasets of children’s speech (e.g., [29]),

datasets of children’s spontaneous utterances are lacking

[41]. Our work begins addressing this gap, however we are

not seeking to improve the state of the art in speech recog-

nition or spoken dialog systems, but rather answer several

open questions in speech interface design.

Personification & Personalization in Speech Interfaces

One open question in speech interfaces for the home is the

role of personalization and personification. HCI researchers

have pursued personalized and personified interaction with

speech interfaces since the early 1990s [38], though wheth-

er these kinds of agents lead to the best user experience

remains an open question even after decades of investiga-

tion (e.g., [52]). For example, one previous study found that

personified agent-like output from a speech device led

adults to interact with it in ways that were not well support-

ed by their system (e.g., asking the device direct questions)

[24]. Adults interacting with Siri reported negative respons-

es to a similar mismatch between the level of personifica-

tion and actual capabilities of the speech interface [37]. An

investigation of adults’ interaction with a virtual reception-

ist found that people attributed different amounts of person-

ification to the receptionist and interacted differently with it

based on that attribution [31]. Another investigation fo-

cused on voice output found no effects of output “embodi-

ment” (speech coming directly from smart objects vs. a

disembodied home control agent vs. a home controlled

agent that was also embodied as an on-screen avatar) on

users’ experience, though the disembodied voice was slight-

ly preferred overall [47]. Nonetheless, most current com-

mercial speech interfaces (e.g., Siri, Amazon Alexa, Google

Assistant, Cortana) are both personified and personalized.

The role of the user’s age in personification and personali-

zation preferences is even less clear. One paper suggested

that “factual” versus “social” interaction with a dialog sys-

tem may be an inherent user preference rather than an age-

dependent characteristic [56]. Other preliminary work with

speech interfaces in the home shows that children interact

with voice systems differently than adults. For example,

children are more likely to include social exchanges with

the system (e.g., “bye!”) [4]. Older participants are more

likely than younger children to have a negative affective

response to speech interfaces that violate privacy expecta-

tions (e.g., know information that the child didn’t explicitly

tell them) [34]. Outside of speech interfaces, at least a few

investigations have suggested that personified search agents

may help children interpret query results and found that this

approach worked best with 8- and 9-year-olds (older chil-

dren found it to be too “childish”) [20]. Systems can be

fairly accurate at distinguishing between adult and chil-

dren’s speech [43], however little is known about how a

system could then adjust its personification and personali-

zation in the most appropriate way to ages or preferences of

users. Our work addresses this gap, with the goal of leading

to more tailored speech appliances for families (e.g., as

suggested in [3]).

Children’s Informational Query Practices

While there has been substantial previous work on speech

interfaces for children’s entertainment (e.g., [18,22,32,33]),

we focus on informational queries as the specific context of

interaction with speech interfaces. There are two reasons

for this decision. First, these interactions represent an im-Communication, Emotion & EngagementIDC 2018, June 19–22, 2018, Trondheim, Norway301

portant use case for speech interfaces. Previous work re-

vealed that the plurality (30%) of family speech interactions

with a home voice kiosk [4] and the plurality (45%) of chil-

dren’s interactions with Apple’s Siri [36] fell into the in-

formation seeking and web search categories. Second, there

is substantial potential for leveraging speech interfaces to

address a number of challenges that children face with cur-

rent text-based search practices. Log analysis has found that

children’s web searches are frequently “unsuccessful” and

“confused” [13,14]. The major challenges identified in pre-

vious investigations of children’s web search include

spelling, typing, and query formulation [10,11]. Spelling

and typing difficulties may be amplified by children’s ten-

dency to use longer, natural language queries [12,27]

(though this may be culturally dependent, as a study of

German children’s search found that they were more likely

to have shorter queries [19]). Speech interfaces may remove

the barriers of spelling and typing, allowing children to

focus on the task of query formulation.

Query reformulation in response to a misrecognition or

misunderstanding is an important aspect of interacting with

speech interfaces. This has been found to be a challenging

task even for adults, who employed strategies as diverse as

word substitution, phrase re-ordering, and phonetic empha-

sis [25,37]. Other previous work on speech-based home

automation controls investigated adults’ responses to differ-

ent types of errors (i.e., ones leading to stagnation, regres-

sion, or partial progress towards a goal) [47], again reiterat-

ing that reformulation may be challenging. It may be even

more difficult for children, but little is known about chil-

dren’s practices with speech interface query reformulation.

Previous exploratory work that has examined children’s

interaction with Apple’s Siri through a content review of

YouTube videos found that children had substantial trouble

dealing with speech interface errors, relying mostly on pho-

netic emphasis in reformulation [36]. Additionally, in a

study with a similar WoZ design to ours, Oviatt et al. found

that children change the prosodic (e.g., speed, pitch) quali-

ties of their speech to match an animated agent’s [48], but

did not investigate semantic reformulations or adaptations.

One of our goals in this work is to understand how children

deal with restating or restructuring queries when they can-

not get to an answer.

METHODS

In this section, we describe our setting, participants, and

detail the system setup and procedure to support replication.

Setting

The study took place in the research outreach building at

the Minnesota (MN) State Fair. The Driven2Discover

building is a permanent facility on the State Fairgrounds,

visited daily by thousands of fair attendees who are repre-

sentative of the population in the Minnesota.1 The perma-

nent facility provided a relatively private and quiet space

1 http://d2d.umn.edu/

for the study to take place (e.g., other studies were separat-

ed by curtains and the building was protected from the gen-

eral hustle and bustle of the Fair). There were two study

stations set up in the building booth (see Figure 1), as well

as another station for gathering initial information and con-

sent from the children and parents.

Participants

We recruited child participants (ages 5–12), along with a

parent or guardian who could provide consent, give infor-

mation about the child, and potentially participate in this

study. One benefit of our chosen study setting was the abil-

ity to recruit and include a greater diversity of families than

most lab-based investigations. Families passing through the

research facility and choosing to participate in the study

were representative of Fair attendees as a whole. For exam-

ple, 25% of the visitors were from rural counties (consistent

with state population) and 28% of the parents did not have a

college degree (consistent with published demographics

statistics from the 2016 State Fair).

During the course of the study, 87 children completed the

procedure (57% female; M = 9 years old, SD = 1.99). Par-

ents or guardians were given one of three options:

1. If there was an empty station, they could attempt the

tasks themselves (otherwise, we prioritized children).

2. They could sit next to their child to help read the ques-

tions and write answers (any assistance given by par-

ents beyond reading and writing help, such as hints or

prompts, was logged and included in our analysis).

3. They could do neither and wait for their child to com-

plete the study.

Given these options, 27 adults completed option one,

providing us with a base of comparison for adult partici-

pants (48% female; M = 47 years old, SD = 10.61). All

Figure 1. A child points to one of the interfaces in the study

setup. Each interface is represented as a plastic bin with a

speaker. The Wizard-of-Oz (visible behind the cardboard

divider) controls the text-to-speech output of each device.Communication, Emotion & EngagementIDC 2018, June 19–22, 2018, Trondheim, Norway302

participants (adults and children) were either native English

speakers or agreed with the statement “I am comfortable

speaking English.” We also requested parents to tell us

which (if any) voice assistants they and their children have

used in the past (e.g., Apple Siri, Amazon Alexa). 96% of

the adults and 94% of the children had used one or more

different voice assistants in the past. Parents were more

likely to have used these interfaces infrequently (mode re-

sponse was “less than once a week”), while children were

more likely to use these interfaces frequently (mode re-

sponse was “multiple times per day”).

Systems

To address our research questions, we developed multiple

versions of a voice assistant. We discuss the physical hard-

ware employed, the software and Wizard-of-Oz training,

and the specific modifications made in each condition.

Equipment

Each study station housed three variations of speech inter-

faces, each represented as a different plastic housing and an

AISBR 3W wired speaker (see Figure 1) to allow partici-

pants to more easily distinguish between and refer to each

interface. All participants were audio recorded using a Blue

Yeti USB microphone, which contains a tri-capsule array to

support field recording of human voices at a 48kHz sample

rate and 16bit bit rate. All of these peripherals were con-

nected to a laptop running the Wizard-of-Oz’s control soft-

ware and Audacity recording software.

Wizard-of-Oz Controls and Training

To focus our investigation on the role of question reformu-

lations (rather than recognition factors), we chose to use a

Wizard-of-Oz (WoZ) technique to simulate a voice assis-

tant. This allowed us to have human-quality speech recog-

nition, removing this confound from the investigation. As in

other WoZ studies (e.g., [9]), to allow the Wizards to pro-

vide near real-time response to the participants, we had to

develop custom WoZ control software. The team created a

Python GUI (see Figure 2) to allow the Wizard to quickly

and consistently select common responses and statements,

as well as directly edit response text when modifications

were necessary. The Wizards followed a specific script to

ensure that participants received consistent responses from

the system. Response text was converted to speech using

Microsoft’s Zira voice (female voice with an American

accent) and the pyttsx Python text-to-speech library.

Five Wizards supported this investigation, allowing work to

occur in shorter shifts to avoid inconsistencies due to fa-

tigue. All Wizards used a common protocol and guide for

contingency responses and trained together through multi-

ple piloting sessions to increase consistency in Wizard re-

sponse. Wizards were visible to the participants, though a

privacy screen hid their immediate actions (see Figure 1).

Participants were told that the Wizards were there to help

the voice assistants. Due to our significant GUI shortcuts

and piloting, most responses were quick and required no

typing, revealing little about the Wizard’s role. Due to the

increasingly common use of human computation as a tech-

nique in computing systems, the IRB did not view WoZ to

be an example of deception so no debriefing was required

with participants who did not inquire about the specific

functioning of the system. Only two adults expressed suspi-

cion that the researcher had a greater role and they were

debriefed after the study. None of the children expressed

suspicion or inquired about how the system worked.

Three Conditions

To answer our questions regarding the role of personifica-

tion and naming personalization, we built three variations

of the speech system (supplementary materials include ex-

amples of full scripts of interaction with each system):

• Voice Search System – the non-personified and non-

personalized system never referred to itself in first per-

son, gave only task-related responses, and did not men-

tion the participant’s name, e.g.: “Welcome to the voice

search system. Please, say your question.”

• Fraga2 – the personified and non-personalized system

referred to itself in first person, gave some responses

that were not task-related, but did not mention the par-

ticipant’s name, e.g.: “Hello, I am Fraga. Do you have

a question for me?”

• Swali2 – the personified and personalized system re-

ferred to itself in first person, gave some responses that

were not task-related, and periodically referred to the

participant’s name and age, e.g.: “Hello Jake, I am

Swali. I see you are 6 years old. That makes you 5.5

years older than me. Do you have a question for me?”

2 “Fraga” and “Swali” mean “question” in Swedish and Swahili,

respectively.

Figure 2. A custom software interface guided the Wizard.

Many common interactions (e.g., greetings, hints) were

pre-programmed. Ad-hoc interactions could be typed di-

rectly into the response box or edited from an existing

scripted response. After a participant stated their question,

a “ding!” sound provided feedback that the “system”

heard it. The interface guided the Wizard through condi-

tion order and logged interactions.Communication, Emotion & EngagementIDC 2018, June 19–22, 2018, Trondheim, Norway303

Given the limited nature of children’s attention spans and

that reliable preference measures for children require with-

in-subjects comparisons [57], we minimized the number of

conditions. We omitted the non-personified, personalized

condition as least natural and least interesting.

The Wizards were directed by the GUI as to the order of

these interfaces and would switch speakers appropriately.

The control interface automatically applied interface-

specific wording and formatting to the dialog (e.g., adding

the name to the greeting in the personalized condition). We

chose to use the same voice for all three variations of the

interface in order to avoid user preference of certain voices

or dialects from influencing their selection.

Procedure

Potential participants were solicited by a research team

member as they passed through the Driven2Discover Build-

ing. If they were still interested in the study after the brief

pitch, the families were led to a table where the study was

explained in more detail and a researcher answered their

questions. The child wrote their first name on an assent

form, if able. The parent filled out consent for himself or

herself and for the child to participate, as well as a short

demographic and background questionnaire.

The study represents a within-subjects design, with each

participant asking at least one question to each interface, in

counterbalanced order. It is important to note that we had

specific ethical considerations and constraints that took

priority over procedural consistency in certain cases. In our

discussions with our IRB, it became clear that it was im-

portant to make sure that every child left the study feeling

that they “succeeded” at the assigned task. This considera-

tion led to four study design decisions. First, members of

the family could sit next to the participant, potentially offer-

ing advice (this was used to allow siblings under 5 to serve

as “special helpers” and get a toy at the end). Second, if a

child was not making any progress towards a question (e.g.,

continuing to say it the same way), they were offered pro-

gressively more significant hints by the system or by a re-

searcher. Third, if more than three minutes passed without a

child arriving at an answer to a question or the child was

getting increasingly dejected, the Wizard would “lower the

bar,” giving an answer in situations where the protocol

would otherwise require the system to request additional

clarification. Fourth, there were two “levels” of questions

(with the second question labeled “bonus”). If a participant

was able to arrive at an answer to the first question within

one minute of using the system, they were presented with a

harder bonus question for the same interface.

The questions were presented to participants on a sheet of

paper and they were asked to write a response once they

arrived at an answer with the voice systems. The easier first

question typically provided some context and then a ques-

tion that referred to that context (see Table 1). To arrive at

an answer, participants had to ask their question in a way

that integrated the provided context. For example, for the

second question in Table 1, the participants had to ask the

system about the number of mini-donuts sold in a particular

year and compare the amount to the one stated in the ques-

tion. To answer the more difficult “bonus” questions, the

participants had to decompose a given question into two

parts. For example, for the first “bonus” question in Table

1, the participant had to first ask what the newest ride was

and then ask where that ride was located at the fair. The six

questions were always presented in the same order, howev-

er the order of the interfaces used was counterbalanced

(Latin square). This was done to control for the role of both

Table 2. Descriptive statistics of the comparisons between conditions for children and adults. Average hints and exchanges were

calculated across the first question in each condition (as that was the one that was completed by all participants).

Single Condition Results Grouped by Personification Grouped by Personalization

Voice Search Fraga Swali Non-Personified Personified Non-Personalized Personalized

Children Preferred by #

(# expected if random)

17

(27)

34

(27)

31

(27)

17

(27)

65

(55)

51

(55)

31

(27)

Avg. Hints 0.87 0.96 1.02 0.87 0.99 0.95 1.02

Avg. Exchanges 2.73 2.85 3.10 2.73 2.98 2.79 3.10

Got to Bonus? 33% 32% 26% 33% 29% 33% 26%

Adults

Preferred by

(expected if random)

5

(7)

6

(7)

11

(7)

5

(7)

17

(15)

11

(15)

11

(7)

Avg. Rating 3.63 3.40 3.80 3.63 3.60 3.52 3.80

Avg. Hints 0.04 0.12 0.12 0.04 0.12 0.08 0.12

Avg. Exchanges 2.00 2.46 2.23 2.00 2.35 2.23 2.23

Got to Bonus? 85% 61% 75% 85% 68% 73% 75%

Table 1. Questions and “bonus” questions given as tasks.

Three Initial Questions

The biggest Pig in the history of the MN State Fair was Reggie

the Pig in 2010. How much did he weigh?

500 thousand corn dogs were sold at the MN State Fair in

2016. Were more or fewer mini donuts sold there that year?

The oldest ride at the MN State Fair is “Ye Old Mill.” What

year was it new to the MN State Fair?

Three “Bonus” Questions

Where can you find the newest ride at the MN State Fair?

Did more people attend the MN State Fair in 2015 or 2016?

Which State Fair is older, the MN State Fair or TX State Fair?Communication, Emotion & EngagementIDC 2018, June 19–22, 2018, Trondheim, Norway304

order effects and natural variations in question difficulty on

participant preferences.

After each question, children were asked to rate the voice

system they used on the smiley-o-meter scale [51]. As sug-

gested in previous work, we used this scale as an opportuni-

ty for children to pause and reflect on the experience rather

than as a reliable metric of preference [51,57]. Adults speci-

fied their ratings on a similar 5-point scale. After trying the

three interfaces, we asked all participants to pick their fa-

vorite interface of the three (a three-way variation of the

“This or That” method [57]), explain why they liked it, and

ask it one other question on any topic. All questions and

sections of the study were optional—for each question, we

only report results for the subset of participants who an-

swered it. At the end of the study, parents and children were

given a choice of a university-branded drawstring pack or a

small stuffed animal as compensation for their time.

RESULTS

In this section, we describe the analysis and discuss our

findings to each of our three guiding research questions.

RQ1: How do children restructure informational queries

towards a speech interface?

We reviewed the logs and audio recordings of all the partic-

ipants, coding the exchanges initiated by each participant to

answer each question. Since our IRB required that all chil-

dren arrive at an answer by the end of each interaction, we

had to use an alternative measure of effectiveness than an-

swer accuracy. To estimate this effectiveness, we coded the

number of hints and/or prompts participants required to

reach an answer (which could come from the system, the

researcher, or the parent when the child was stuck). We also

noted when a child attempted the bonus question (meaning

that they arrived at the answer to the first question within

one minute) as a signal of success with the interface. Table

2 provides the descriptive statistics of each of these

measures across conditions. Generally, children struggled

with the reformulation task, requiring one or more hints and

with less than half of the children reaching the “bonus”

question. These difficulties reduced with age, with the

chance of getting to the bonus question significantly in-

creasing (r = 0.30, p = 0.006) and the number of needed

hints significantly decreasing (r = -0.33, p = 0.004).

A team of four researchers reviewed the recorded audio of

each interaction, taking notes, transcribing (example tran-

scripts available in supplementary materials), and describ-

ing each instance of a reformulation. Thus four researchers

transcribed, memoed, and open coded each transcript indi-

vidually following the process described by Lofland et al

[35]. Then the four researchers took part in a workshop led

by the lead author to arrive at clusters of codes through

multiple rounds of constant comparison, using “abductive”

analysis [40] to develop our categories. Once our codebook

was developed, we each reviewed all the transcripts again,

applying those codes to the original dataset. The following

categories of reformulations emerged through this data-

driven inductive process:

Off-Course – changing the question to something relevant

to the topic that the system can answer, but that does not

get the asker closer to the target answer. Examples:

Child repeatedly asks “Was Reggie the Pig the fattest

pig in 2010 at the MN State Fair?” receiving “Yes” as

a response from the system but not knowing how to fol-

low up about the pig’s weight.

Child gets frustrated with repeatedly getting the re-

sponse “That question is too complicated for me” and

changes the topic, asking the system: “Who is the

strongest superhero?”

Restating or Repeating – changing how a question is pro-

nounced or emphasized, without changing any words or

structure of the question. Examples:

System asks the child to “Please, ask the question in a

different way.” Child sings the question to the system.

System asks the child to “Please, ask the question in a

different way.” Child repeats question louder.

Substituting Words – changing a word or phrase in a ques-

tion without adding any additional information or removing

any complexity from the previously stated question. Exam-

ples:

Child rereads the whole question about which State

Fair is older, substituting “which is younger” for

“which is older.”

Child rephrases question as “How many pounds was

Reggie the Pig?”

Reordering – reordering the components of a question

without adding any additional information or removing any

complexity. Examples:

Child rephrases question as “True or False: the MN

State Fair is older than the Texas State Fair.”

Table 3. Prevalence of categories of query reformulation by age.

We excluded 3 children whose audio data was incomplete (one or

more condition was inaudible) and 5 children who did not make

any independent reformulations (all reformulations were prompt-

ed). In this study, starred categories helped participants get closer

to an answer; others were not effective.

Children, 5–7

(N = 19)

Children, 8–12

(N = 58)

Adults

(N = 27)

Off-Course 5% 9% 0%

Restate 53% 64% 30%

Substitute Words 53% 52% 63%

Reorder 37% 53% 37%

State Context 26% 22% 30%

Expand Pronouns\* 32% 62% 70%

Add Context\* 58% 66% 44%Communication, Emotion & EngagementIDC 2018, June 19–22, 2018, Trondheim, Norway305

Child rephrases question as “How many fewer mini-

donuts were there sold than corn dogs in 2016?

Stating Context – adding additional information before

asking a question, but phrasing this addition as keywords or

statements (not integrated into the question). Examples:

Child states “Reggie was the fattest pig in 2010 at the

MN State Fair. How many pounds was he?”

Child states “The Great Big Wheel is the newest ride at

the MN State Fair. Where is it located?”

Expanding Pronouns in Question – replacing the non-

specific pronoun in a question with a specific noun. Exam-

ples:

Child expands the pronoun “he” to “How much did

Reggie the Pig weigh?”

Child expands the pronoun “it” to “What year was Ye

Old Mill first introduced?”

Adding Context Phrases – adding context following or

preceding the noun in a question to narrow to the specific

case of interest. Examples:

Child asks the questions one at a time until the system is

able to answer: “How much did Reggie the pig weigh?”

“How much did Reggie the pig from the State Fair

weigh?” “How much did Reggie the pig ... from 2010

... from the fair ... from the MN State Fair weigh?”

Child first asks about the number of mini donuts and

corn dogs without specifying a year or location. System

asks him to be more specific. Child expands: “At the

MN fair were there ... in 2016 ... were there more or

less mini donuts sold than corn dogs?”

It is important to note that in this study, the first five types

of reformulations did not help the participants arrive at an

answer. However, when interacting with other speech inter-

faces, some of these may be helpful strategies. Table 3 re-

ports the percent of children who employed each reformula-

tion. The most common unsuccessful reformulations were

restating without changing the question (53% and 64%,

among 5–7 year-olds and 8–12 year-olds respectively) and

changing the words in the question without changing its

structure (53% and 52% in the two age buckets). The most

common reformulation strategies that got the participant

closer to the answer included adding context into the ques-

tion (58% and 66% in the two age buckets) and expanding

the pronouns in a question (32% and 62% in the two age

buckets). Most children tried a number of different strate-

gies (on average, three or more), but many were stuck in a

single strategy until a hint or prompt was provided.

After the children had the opportunity to use each interface,

they could ask one more question on any topic to the inter-

face of their choice. We categorized these based on the

structure and assumed intent of the question. While children

were told that they could ask a question about anything at

all, many seemed to be quite influenced by the previous

tasks (this was expected given previous preliminary work

[53]). 32% of their questions were quantitative questions

about other state fair topics (e.g., “What year did the Ferris

wheel open?”) and another 34% were quantitative questions

unrelated to the state fair (e.g., “Are there more dogs in the

world than cats?”). Qualitative questions (e.g., “How do

staplers work?”) accounted for 8% of the dataset. Two oth-

er interesting categories emerged. In 18% of the questions,

children wanted to learn more about the experiences and

interests of the interface (e.g., “What is your favorite foot-

ball team?”). In 9% of the questions, children tried to test

the interface (e.g., “What color bucket is under you?” or

“What is zero divided by zero?”).

RQ2: How does a speech interface’s personification and

naming personalization affect children’s experience?

We asked each child to interact with all three of the inter-

faces, thereby gauging their response to personified and

personalized systems. Table 2 provides descriptive statistics

of metrics gathered in each of the conditions. We compared

children’s preferences for personified vs. non-personified

conditions using Chi-Square Goodness-of-Fit test, finding

that children did indeed prefer personified interfaces at a

greater and statistically significant rate (p = 0.015). A simi-

lar comparison between personalized and non-personalized

conditions was not statistically significant. We did not ob-

serve a relationship between the child’s age and their inter-

face preference, though it is possible that one could emerge

in a study with more participants of each age.

Effectiveness measures were similar across the conditions.

When compared using a Repeated Measured ANOVA, the

difference in the number of hints required and exchanges

made across conditions was not statistically significant.

Similarly, comparing the number of children getting to the

bonus question in each condition versus an even distribu-

tion across conditions using the Chi-Square Goodness-of-

Fit test did not show a statistically significant difference.

Therefore, we conclude that personification and naming

personalization do not influence children’s effectiveness

with speech interfaces in a statistically significant way.

We asked each child about why they liked a particular inter-

face best. Of those that could give an answer beyond the

tautological (e.g., “I liked it better”), 62% said that they

thought that their favorite interface understood them better

or was less confused about their questions (this was mostly

an order effect—children developed more effective refor-

mulation strategies by the third condition). Eleven percent

of the answers were related to irrelevant surface considera-

tion (e.g., favorite color), however removing these respons-

es from the analysis did not change the direction or magni-

tude of the result. Two children mentioned liking the per-

sonalized interface best because of its personality (“more

friendly” and “polite”) and another four children explicitly

mentioned liking the interface that “knew my name.” On the

other hand, another four children were explicitly turned offCommunication, Emotion & EngagementIDC 2018, June 19–22, 2018, Trondheim, Norway306

by the naming personalization saying that it was “creepy.”

One of these children loudly exclaimed, “Are you stalking

me?!” when the system referred to him by name and age.

However, it is important to interpret this reaction in con-

text—this information was solicited by the researcher not

directly by the robot, so this reaction may be moderated in

other arrangements. Fewer children provided interface-

specific reasons for liking the non-personified one, though

one did mention liking it because it “spoke faster and was

more efficient.” While the qualitative experience for this

individual was more efficient with the non-personified in-

terface and there was a descriptive difference between the

conditions in this direction, it was not statistically signifi-

cant so it does not generalize across our sample.

When given the opportunity to ask their favorite interface

any question they wanted, it was interesting to note that

questions about the interests and experiences of the inter-

face and qualitative questions were directed almost exclu-

sively (all but one) towards the personified conditions.

RQ3: How do the experiences and preferences of adults

compare to those of children?

As can be expected, adults faced fewer struggles with ques-

tion reformulations, with only a few requiring hints and

with the majority reaching the bonus question (see Table 2).

We did not see any examples where adults went so off-

course in their reformulations that they asked questions that

did not get them closer to the answer (though in two cases,

adults asked additional irrelevant questions after completing

their task). Fewer adults (30%) than children (53% or 64%,

depending on age) focused on the unsuccessful strategy of

simply restating the question with a different emphasis or

pronunciation. Adults employed some successful strategies

more often than children. For example, older participants

were more likely to expand pronouns (32% of 5–7-year-

olds, 62% of 8–12-year-olds, and 70% of adults). They

were less likely to reformulate the question with additional

context, but only because many included all the necessary

context phrases from the first formulation.

Adult preferences for naming personalization and personifi-

cation were not statistically significantly different from

random when tested with the Chi-Square Goodness-of-Fit

test (p = 0.095). There was no statistically significant dif-

ference between their ratings of each condition on a 5-point

Likert-type scale when compared with a Repeated

Measures ANOVA. Like the children’s results, there was

no statistically significant differences between conditions

for adults’ effectiveness with the systems in terms of the

number of hints required, number of exchanges, or the like-

lihood of getting to the bonus question. While descriptively,

adults seemed to prefer the naming personalization condi-

tion at greater rates than children, the difference in prefer-

ence distributions between adults and children was not sta-

tistically significant when compared with a Chi-Square test.

Fourteen of the 22 adults who had a preferred interface pro-

vided a reason for that selection. Interestingly, the most

common (36%) reason explicitly referred to liking the “per-

sonality” of the naming personalization interface best, de-

scribing it as “like a friend,” “witty,” and “more personal.”

However, one adult did mention being “creeped out” and

disliking the same interface for saying their name (even

though they understood that they provided their name to the

system). One adult also mentioned that they liked the non-

personified interface best because of its efficiency (“not a

lot of extra talking”). Finally, adults generally recognized

that all three interfaces understood them equally poorly and

were more likely to reflect on the role of order in their sys-

tem preferences, acknowledging that they became better at

asking questions by the end of the study.

While almost all children took the opportunity to ask their

favorite interface another question, only 13 out of the 27

adults did so. As with the child participants, 38% of these

were quantitative questions related to the state fair and an-

other 38% were quantitative questions on other topics.

None of the adults asked qualitative questions or attempted

to learn more about the experiences and interests of the

interface. However, as children did, 23% of the adults did

try to ask questions that tested the interface, usually by ask-

ing a question where they already knew the answer (e.g.,

“Who was the first president of the United States?”).

DISCUSSION

In this section, we discuss the implications of our findings

on design and research in speech interfaces for families.

Limitations and Future Directions

All study designs have inherent limitations. We had to

make specific decisions regarding the wording of questions,

responses, and specific operationalization of concepts like

“personification” and “naming personalization.” Children

may have personified the non-personified condition despite

the language used by the system, as previous work shows

that personal pronouns are not necessary for perspective-

taking [2]. Similarly, personalization can be much more

nuanced and useful than merely referring to a person by

their first name [16]. For example, it would be a useful per-

sonalization to use the participants’ location to provide con-

text for the questions or to tailor the systems responses

based on a child’s ability. Given that many of the concerns

children expressed against personalization were privacy-

based, the privacy theory of “proportionality” [23] suggests

that some of those concerns would be ameliorated if the

personalization provided a functional benefit in interpreting

and answering questions. Also, it may not have been trans-

parent to the child how the systems learned their name,

since these were entered by the researchers. But, this paral-

lels the real-world situation where a parent would generally

provide information for a child’s voice assistant account.

Certainly, we do not consider our investigation to be the

final word on personalization of speech interfaces, but ra-

ther a point of evidence towards the idea that a parent enter-

ing a child’s name into a speech system does not support a

personalization that provides measurable value to children.Communication, Emotion & EngagementIDC 2018, June 19–22, 2018, Trondheim, Norway307

We made study design choices that allowed us to collect

data from a large, diverse sample in a field setting. While

this has many advantages, it also introduced several limita-

tions. First, our study was not tightly controlled or in a lab

setting. This may mean that there were times when siblings,

parents, etc. distracted the children from their tasks. How-

ever, this also represents a more ecologically valid situa-

tion. Second, we provided most of the questions asked by

the participants. We selected questions that were challeng-

ing and that would necessitate reformulation (similar to the

approach of previous work on children’s online search, e.g.,

[9]). We picked questions that were relevant to the setting

and similar in style to the kinds of questions children may

be asked to answer on a homework worksheet or similar

assignment. However, we do not know to what extent these

compound complex questions appear in the lexicon of chil-

dren’s interactions with speech interfaces in the wild.

Finally, each participant had a relatively brief interaction

with a speech interface that belonged to the researcher.

There are two possible biases that this introduced in the

dataset. First, we could not observe long-term learning ef-

fects. We did observe that participants learned from earlier

interactions and their performance improved even in the

short time of the study. It is important to conduct a follow-

up study with a home or mobile speech device in the wild,

to understand which reformulations remain problematic in

the long run and which are quicker adaptations. Second,

privacy concerns may have been exacerbated by interacting

with a speech assistant that belonged to the researcher, ra-

ther than a personal device. It is possible that some of the

“creepiness” factor of naming personalization may have

been ameliorated if participants interacted with their own

personal speech device. This was not possible in the context

of our controlled investigation, but would be an important

and promising follow-up study if participants could be giv-

en personal devices to keep and interact with long-term.

Considerations for User-Friendly Speech Interfaces

Observing the struggles and relatively low success rate of

our child participants, we reflect that current speech inter-

faces may be considered poor designs from a classic design

perspective (e.g., [46]). They provide no constraints on the

kinds of statements that may be directed to them, leading to

the observed high number of exchanges before arriving at

an answer. They provide little feedback regarding what they

heard and understood, leading most children to spend sig-

nificant time on phonetic reformulations. They give almost

no visibility as to why the system may be struggling with a

particular question (e.g., Amazon Alexa simply answers “I

don’t know” if it is confused by any part of the question).

Many of our participants struggled but took reasonable

paths—first assuming that a question was misheard, then

assuming that a word was not understood, before moving

on to more effective strategies. However, the process was

undeniably frustrating and many of the children required

help to get past the first two strategies and to arrive at ques-

tions that would yield an appropriate answer.

Currently available systems rarely (if ever) provide hints or

clarification. Based on each of the common reformulation

types and the types of hints and prompts that were most

useful to the participants in our study, we recommend that

systems integrate the following types of clarifications:

• Restate what was heard if the system fails to meet a

confidence threshold. Children are used to being mis-

heard by speech interfaces and most of them focused

on phonetic reformulations. With feedback that the sys-

tem heard the question correctly, they may proceed to

semantic reformulations.

• If a particular piece of context is needed but missing

(e.g., year, identity of a particular pronoun), follow up

requesting this information (e.g., “What is “it” in your

question?). It may also be reasonable to make a “best

guess” about missing information from previous ques-

tions or by assuming current location, year, etc.

• If a part of the question is known but the entire ques-

tion is too complex, provide some information on what

is known and request clarification for the rest (e.g., “I

know that the newest ride at the fair is the ‘The Great

Big Wheel,’ what would you like to know about it?”).

• If a particular style of question is difficult for the sys-

tem, clarify this and provide guidance (e.g., “Compari-

sons are hard for me. Can you ask about each part of

your question separately?”).

Additionally, sometimes a significant amount of contextual

information is needed to provide the correct answer to a

question. It can be difficult and awkward for children to

construct complex questions that integrate all of the neces-

sary components. While many children (58%-66%) eventu-

ally attempted integrating contextual information directly

into the question, they struggled with the compound sen-

tences that resulted, pausing after each informational ele-

ment. We also saw that stating information up-front rather

than integrating it directly into the question was a strategy

employed by both adults (30%) and children (22%-26%).

Speech systems could become more usable in that regard by

allowing contextual information for a particular question to

be provided in the form of a sentence or multiple sentences.

For example, instead of having to ask a system “What will

the weather be like this Thursday and Friday in Trond-

heim?” it may be easier for users to give upfront context (“I

am traveling to Trondheim on Thursday.”) followed by the

particular question (“What will the weather be like?”).

Costs and Benefit of Personifying & Personalizing

Previous work in the field of speech interfaces had provided

divergent findings regarding the potential roles of personi-

fication and personalization. This investigation helped ad-

dress some aspects of this open question.

None of the participants had specific objections to personi-

fication of interfaces, though several participants did note

that non-personified interfaces covered less non-task con-

tent and thus were more efficient at quickly providing an

answer. Children responded to personification, preferringCommunication, Emotion & EngagementIDC 2018, June 19–22, 2018, Trondheim, Norway308

personified conditions to the non-personified. Children also

seemed more willing to direct a broader variety of ques-

tions, including qualitative ones, to personified interfaces.

Whether because of actual confusion or as a playful act,

children asked the personified interfaces question about

their experience and preferences (echoing previous studies

[3]). This may not be entirely positive, as it is difficult to

balance engaging in the play to provide a “fun” answer and

answering questions honestly without misleading the child

as to the nature of the interaction (e.g., if asked about its

favorite food, should the system lie or should it say that it

does not eat?). It is worth noting that the same considera-

tion need not be extended to adults—we saw no examples

of adults asking these kinds of questions. While not stated

as a concern by any participants, researchers in other do-

mains have noted that increased personification of interfac-

es may create a “robotic moment,” in which children may

become confused the role and agency of humans compared

to machines [54]. This may be a trade-off of personification

if empirically confirmed in future research.

We did not find statistically significant evidence favoring

naming personalization as operationalized in our study. One

common objection cited by several children and one adult

was that it was a violation of privacy expectations to have

the speech device know specifics like their age and name,

even though they themselves provided this information to

the researcher. This objection complements previous inves-

tigations of children interacting with robotic agents [34].

On the other hand, some children and adults did find the

personalized system to be more friendly, witty, and polite.

This may be an individual difference. However, it is also

worth noting that even the kind of surface naming personal-

ization mentioned in this study may not be easy to accom-

plish in the field. First, it may require interfaces to distin-

guish between users based on voice, which is possible but

not trivial [43]. Second, it is increasingly common for chil-

dren to have names that are non-traditional, unusually

spelled, or names from non-western cultures. This increases

the likelihood that a speech system will mispronounce a

child’s name, which may reverse perceptions of friendliness

and politeness. Given that benefits are not evidenced in this

investigation, naming personalization should not be a high

priority feature. However, it is again important to reiterate

that there are many other possible types of personalization

beyond “naming,” which may provide a different set of

costs and benefits to the participants.

Building for the Whole Family

One of the surprising descriptive findings of our study is

that 93% of children had used one or more speech interfac-

es prior to the study and the plurality of these children used

such interfaces multiple times every day. While we thought

of this interface modality as “emerging,” it became clear

that it was already well-entrenched in the lives of children.

There are three ways children typically interact with speech

interfaces. Some children may have their own smart phone

or tablet. In this case, it may make sense to default to a per-

sonified interface, letting the child control their level of

personalization. Other children may experience so-called

“pass-back”—the opportunity to interact with speech inter-

faces on their parents’ smart phone. It may be useful for

such devices to detect children’s voices (e.g., [44]) and re-

move any personalization features aimed at the parent while

retaining personification. Third, children may live in a

home that has a household speech information appliance

(e.g., Google Home). Such devices need to account for the

specific preferences of multiple users [2]. Our study com-

plements previous investigations showing that interface

personification preference may be more related to individu-

al differences than age [56]. We did not see specific age

effects found in other previous work (e.g., older children

considering personification to be too childish) [20]. There

was a substantial split regarding personalization and there

were also a fair number of participants who preferred the

non-personified condition (i.e., a family of four is not un-

likely to have one person in this category). It is an open

question whether it would provide the best experience for

the whole family to adapt to individual preferences or

whether families would prefer to have a consistent experi-

ence even if some individuals’ preferences are violated.

CONCLUSION

Most children in our study used voice assistants and many

of them used such interfaces multiple times every day. In-

formational queries are a common use case, but prior to this

investigation little was known about how children formu-

late questions towards speech interfaces. Our work address-

es this gap by observing 87 children asking questions of

three variations of Wizard-of-Oz speech interfaces. We

found that children struggled with reformulating questions,

with most of them requiring hints to complete the task.

Though most children eventually tried effective reformula-

tions such as substituting objects for pronouns and provid-

ing context within the question, many children first began

with surface reformulations such as repeating the question

or substituting synonym words. Older children and adults

were more effective than younger children at informational

query reformulation. By comparing variations of the inter-

face, we discovered that children preferred personified in-

terfaces, but showed no preference towards the interface

that was personalized with their name and age. We suggest

several considerations for future speech interfaces. First,

personified interfaces are well indicated, while naming per-

sonalization (especially, when that name is provided by

others) is not. Second, we point to five strategies to support

more effective reformulations: providing feedback on what

was heard, asking for missing context, clarifying what is

known, specifying formulations that are difficult for the

system, and allowing context to be provided as a statement.

ACKNOWLEDGMENTS

We gratefully acknowledge our funding from a 2017H2

Mozilla Research Grant and a 2017 Google Faculty Re-

search Award. We also thank the D2D State Fair team.Communication, Emotion & EngagementIDC 2018, June 19–22, 2018, Trondheim, Norway309

SELECTION AND PARTICIPATION OF CHILDREN

This study was reviewed and approved by a University

IRB. We invited families with children between the ages of

5 and 12 to participate at the MN State Fair event. A re-

searcher explained the study and its risks to both the child

and parent and answered any questions. If interested, par-

ents signed informed consent and a parental permission

form. A researcher read a paper assent form out loud to

each child. If assenting to the study, the child wrote their

name on the form (as able). Both parental permission and

the child’s explicit assent were necessary for the study to

proceed and either person could ask us to stop the study at

any time (the child still received a toy for their help).

REFERENCES

1. Nadia Aram. 2017. Hey, Alexa: What’s New in Chil-

dren’s Privacy?... FTC Updates COPPA Guidance. Re-

trieved August 28, 2017 from

http://www.wcsr.com/Insights/Alerts/2017/June/Hey-

Alexa-Whats-New-in-Childrens-Privacy-FTC-Updates-

COPPA-Guidance

2. Tad T. Brunyé, Tali Ditman, Grace E. Giles, Amanda

Holmes, and Holly A. Taylor. 2016. Mentally simulat-

ing narrative perspective is not universal or necessary

for language comprehension. Journal of Experimental

Psychology. Learning, Memory, and Cognition 42, 10:

1592–1605. https://doi.org/10.1037/xlm0000250

3. A. J. Bernheim Brush and Kori M. Inkpen. 2007. Yours,

Mine and Ours? Sharing and Use of Technology in Do-

mestic Environments. In Proceedings of the 9th Interna-

tional Conference on Ubiquitous Computing (UbiComp

’07), 109–126. Retrieved December 27, 2016 from

http://dl.acm.org/citation.cfm?id=1771592.1771599

4. A. J. Brush, Paul Johns, Kori Inkpen, and Brian Meyers.

2011. Speech@Home: An Exploratory Study. In CHI

’11 Extended Abstracts on Human Factors in Compu-

ting Systems (CHI EA ’11), 617–632.

https://doi.org/10.1145/1979742.1979657

5. Theodora Chaspari, Samer Al Moubayed, and Jill Fain

Lehman. 2015. Exploring Children’s Verbal and Acous-

tic Synchrony: Towards Promoting Engagement in

Speech-Controlled Robot-Companion Games. In Pro-

ceedings of the 1st Workshop on Modeling INTERPER-

sonal SynchrONy And infLuence (INTERPERSONAL

’15), 21–24. https://doi.org/10.1145/2823513.2823518

6. Yun-Nung Chen, Ming Sun, Alexander I. Rudnicky, and

Anatole Gershman. 2015. Leveraging Behavioral Pat-

terns of Mobile Applications for Personalized Spoken

Language Understanding. In Proceedings of the 2015

ACM on International Conference on Multimodal Inter-

action (ICMI ’15), 83–86.

https://doi.org/10.1145/2818346.2820781

7. Christopher K. Cowley and Dylan M. Jones. 1993. Talk-

ing to Machines (Abstract). In Proceedings of the

INTERACT ’93 and CHI ’93 Conference on Human

Factors in Computing Systems (CHI ’93), 522–.

https://doi.org/10.1145/169059.169512

8. Dan Bohus and Alexander I. Rudnicky. 2003.

RavenClaw: Dialog Management Using Hierarchical

Task Decomposition and an Expectation Agenda. Re-

trieved from http://repository.cmu.edu/compsci/1392/

9. Steven P. Dow, Manish Mehta, Blair MacIntyre, and

Michael Mateas. 2010. Eliza Meets the Wizard-of-oz:

Blending Machine and Human Control of Embodied

Characters. In Proceedings of the SIGCHI Conference

on Human Factors in Computing Systems (CHI ’10),

547–556. https://doi.org/10.1145/1753326.1753408

10. Allison Druin, Elizabeth Foss, Leshell Hatley, Evan

Golub, Mona Leigh Guha, Jerry Fails, and Hilary

Hutchinson. 2009. How Children Search the Internet

with Keyword Interfaces. In Proceedings of the 8th In-

ternational Conference on Interaction Design and Chil-

dren (IDC ’09), 89–96.

https://doi.org/10.1145/1551788.1551804

11. Allison Druin, Elizabeth Foss, Hilary Hutchinson, Evan

Golub, and Leshell Hatley. 2010. Children’s Roles Us-

ing Keyword Search Interfaces at Home. In Proceedings

of the SIGCHI Conference on Human Factors in Com-

puting Systems (CHI ’10), 413–422.

https://doi.org/10.1145/1753326.1753388

12. Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Ser-

dyukov. 2010. An Analysis of Queries Intended to

Search Information for Children. In Proceedings of the

Third Symposium on Information Interaction in Context

(IIiX ’10), 235–244.

https://doi.org/10.1145/1840784.1840819

13. Sergio Duarte Torres, Ingmar Weber, and Djoerd Hi-

emstra. 2014. Analysis of Search and Browsing Behav-

ior of Young Users on the Web. ACM Trans. Web 8, 2:

7:1–7:54. https://doi.org/10.1145/2555595

14. Carsten Eickhoff, Pieter Dekker, and Arjen P. de Vries.

2012. Supporting Children’s Web Search in School En-

vironments. In Proceedings of the 4th Information In-

teraction in Context Symposium (IIIX ’12), 129–137.

https://doi.org/10.1145/2362724.2362748

15. Pedro Fialho and Luísa Coheur. 2015. ChatWoz: Chat-

ting Through a Wizard of Oz. In Proceedings of the 17th

International ACM SIGACCESS Conference on Com-

puters & Accessibility (ASSETS ’15), 423–424.

https://doi.org/10.1145/2700648.2811334

16. FitzGerald Elizabeth, Kucirkova Natalia, Jones Ann,

Cross Simon, Ferguson Rebecca, Herodotou Chris-

tothea, Hillaire Garron, and Scanlon Eileen. 2017. Di-

mensions of personalisation in technology‐enhanced

learning: A framework and implications for design. Brit-

ish Journal of Educational Technology 49, 1: 165–181.

https://doi.org/10.1111/bjet.12534

17. D. Giuliani and M. Gerosa. 2003. Investigating recogni-

tion of children’s speech. In 2003 IEEE International

Conference on Acoustics, Speech, and Signal Pro-

cessing, 2003. Proceedings. (ICASSP ’03), II-137–40

vol.2. https://doi.org/10.1109/ICASSP.2003.1202313

18. Michal Gordon and Cynthia Breazeal. 2015. Designing a

Virtual Assistant for In-car Child Entertainment. In Pro-Communication, Emotion & EngagementIDC 2018, June 19–22, 2018, Trondheim, Norway310

ceedings of the 14th International Conference on Inter-

action Design and Children (IDC ’15), 359–362.

https://doi.org/10.1145/2771839.2771916

19. Tatiana Gossen, Thomas Low, and Andreas Nürnberger.

2011. What Are the Real Differences of Children’s and

Adults’ Web Search. In Proceedings of the 34th Interna-

tional ACM SIGIR Conference on Research and Devel-

opment in Information Retrieval (SIGIR ’11), 1115–

1116. https://doi.org/10.1145/2009916.2010076

20. Tatiana Gossen, Rene Müller, Sebastian Stober, and

Andreas Nürnberger. 2014. Search Result Visualization

with Characters for Children. In Proceedings of the 2014

Conference on Interaction Design and Children (IDC

’14), 125–134.

https://doi.org/10.1145/2593968.2593983

21. A. Hagen, B. Pellom, and R. Cole. 2003. Children’s

speech recognition with application to interactive books

and tutors. In 2003 IEEE Workshop on Automatic

Speech Recognition and Understanding (IEEE Cat.

No.03EX721), 186–191.

https://doi.org/10.1109/ASRU.2003.1318426

22. Hannaneh Hajishirzi, Jill F. Lehman, and Jessica K.

Hodgins. 2012. Using Group History to Identify Charac-

ter-directed Utterances in Multi-child Interactions. In

Proceedings of the 13th Annual Meeting of the Special

Interest Group on Discourse and Dialogue (SIGDIAL

’12), 207–216. Retrieved from

http://dl.acm.org/citation.cfm?id=2392800.2392838

23. Giovanni Iachello and Gregory Abowd. 2005. Privacy

and proportionality: adapting legal evaluation techniques

to inform design in ubiquitous computing. In Proc. of

CHI, 91–100.

24. Pradthana Jarusriboonchai, Thomas Olsson, and Kaisa

Väänänen-Vainio-Mattila. 2014. User Experience of

Proactive Audio-based Social Devices: A Wizard-of-oz

Study. In Proceedings of the 13th International Confer-

ence on Mobile and Ubiquitous Multimedia (MUM ’14),

98–106. https://doi.org/10.1145/2677972.2677995

25. Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How Do

Users Respond to Voice Input Errors?: Lexical and Pho-

netic Query Reformulation in Voice Search. In Proceed-

ings of the 36th International ACM SIGIR Conference

on Research and Development in Information Retrieval

(SIGIR ’13), 143–152.

https://doi.org/10.1145/2484028.2484092

26. Oliver Jokisch, Horst-Udo Hain, Rico Petrick, and

Rüdiger Hoffmann. 2009. Robustness Optimization of a

Speech Interface for Child-directed Embedded Lan-

guage Tutoring. In Proceedings of the 2Nd Workshop on

Child, Computer and Interaction (WOCCI ’09), 10:1–

10:4. https://doi.org/10.1145/1640377.1640387

27. Yvonne Kammerer and Maja Bohnacker. 2012. Chil-

dren’s Web Search with Google: The Effectiveness of

Natural Language Queries. In Proceedings of the 11th

International Conference on Interaction Design and

Children (IDC ’12), 184–187.

https://doi.org/10.1145/2307096.2307121

28. Theofanis Kannetis, Alexandros Potamianos, and Geor-

gios N. Yannakakis. 2009. Fantasy, Curiosity and Chal-

lenge As Adaptation Indicators in Multimodal Dialogue

Systems for Preschoolers. In Proceedings of the 2Nd

Workshop on Child, Computer and Interaction (WOCCI

’09), 1:1–1:6. https://doi.org/10.1145/1640377.1640378

29. Sawit Kasuriya and Alistair D. N. Edwards. 2009. Pilot

Experiments on Children’s Voice Recording. In Pro-

ceedings of the 2Nd Workshop on Child, Computer and

Interaction (WOCCI ’09), 13:1–13:5.

https://doi.org/10.1145/1640377.1640390

30. Lin-shan Lee, James Glass, Hung-yi Lee, and Chun-an

Chan. 2015. Spoken Content Retrieval: Beyond Cascad-

ing Speech Recognition with Text Retrieval. IEEE/ACM

Trans. Audio, Speech and Lang. Proc. 23, 9: 1389–

1420. https://doi.org/10.1109/TASLP.2015.2438543

31. Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2010.

Receptionist or Information Kiosk: How Do People Talk

with a Robot? In Proceedings of the 2010 ACM Confer-

ence on Computer Supported Cooperative Work (CSCW

’10), 31–40. https://doi.org/10.1145/1718918.1718927

32. Jill Fain Lehman and Iolanda Leite. 2017. Turn-Taking,

Children, and the Unpredictability of Fun. AI Magazine

37, 4: 55–62. https://doi.org/10.1609/aimag.v37i4.2685

33. Iolanda Leite, Hannaneh Hajishirzi, Sean Andrist, and

Jill Lehman. 2013. Managing Chaos: Models of Turn-

taking in Character-multichild Interactions. In Proceed-

ings of the 15th ACM on International Conference on

Multimodal Interaction (ICMI ’13), 43–50.

https://doi.org/10.1145/2522848.2522871

34. Iolanda Leite and Jill Fain Lehman. 2016. The Robot

Who Knew Too Much: Toward Understanding the Pri-

vacy/Personalization Trade-Off in Child-Robot Conver-

sation. In Proceedings of the The 15th International

Conference on Interaction Design and Children (IDC

’16), 379–387.

https://doi.org/10.1145/2930674.2930687

35. John Lofland, David A. Snow, Leon Anderson, and Lyn

H. Lofland. 2005. Analyzing Social Settings: A Guide to

Qualitative Observation and Analysis. Cengage Learn-

ing, Belmont, CA.

36. Silvia Lovato and Anne Marie Piper. 2015. “Siri, is This

You?”: Understanding Young Children’s Interactions

with Voice Input Systems. In Proceedings of the 14th

International Conference on Interaction Design and

Children (IDC ’15), 335–338.

https://doi.org/10.1145/2771839.2771910

37. Ewa Luger and Abigail Sellen. 2016. “Like Having a

Really Bad PA”: The Gulf Between User Expectation

and Experience of Conversational Agents. In Proceed-

ings of the 2016 CHI Conference on Human Factors in

Computing Systems (CHI ’16), 5286–5297.

https://doi.org/10.1145/2858036.2858288

38. Matt Marx and Chris Schmandt. 1994. Putting People

First: Specifying Proper Names in Speech Interfaces. In

Proceedings of the 7th Annual ACM Symposium on UserCommunication, Emotion & EngagementIDC 2018, June 19–22, 2018, Trondheim, Norway311

Interface Software and Technology (UIST ’94), 29–37.

https://doi.org/10.1145/192426.192439

39. Jack Mostow and Steven F. Roth. 1995. Demonstration

of a Reading Coach That Listens. In Proceedings of the

8th Annual ACM Symposium on User Interface and

Software Technology (UIST ’95), 77–78.

https://doi.org/10.1145/215585.215665

40. Michael Muller. Curiosity, Creativity, and Surprise as

Analytic Tools: Grounded Theory Method. In Ways of

Knowing in HCI, Judith S. Olson and Wendy A. Kellogg

(eds.). Springer, 25–48.

41. S. Narayanan and A. Potamianos. 2002. Creating con-

versational interfaces for children. IEEE Transactions

on Speech and Audio Processing 10, 2: 65–78.

https://doi.org/10.1109/89.985544

42. Clifford Nass and Li Gong. 2000. Speech Interfaces

from an Evolutionary Perspective. Commun. ACM 43, 9:

36–43. https://doi.org/10.1145/348941.348976

43. R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano.

2004. Public speech-oriented guidance system with

adult and child discrimination capability. In 2004 IEEE

International Conference on Acoustics, Speech, and

Signal Processing, I-433–6 vol.1.

https://doi.org/10.1109/ICASSP.2004.1326015

44. R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano.

2004. Public speech-oriented guidance system with

adult and child discrimination capability. In 2004 IEEE

International Conference on Acoustics, Speech, and

Signal Processing, I-433–6 vol.1.

https://doi.org/10.1109/ICASSP.2004.1326015

45. Don Nix, Peter Fairweather, and Bill Adams. 1998.

Speech Recognition, Children, and Reading. In CHI 98

Conference Summary on Human Factors in Computing

Systems (CHI ’98), 245–246.

https://doi.org/10.1145/286498.286730

46. Donald A. Norman. 1990. The Design of Everyday

Things. Basic Books, New York.

47. A. Oulasvirta, K. P. Engelbrecht, A. Jameson, and S.

Moller. 2007. Communication failures in the speech-

based control of smart home systems. In 2007 3rd IET

International Conference on Intelligent Environments,

135–143. https://doi.org/10.1049/cp:20070358

48. Sharon Oviatt, Courtney Darves, and Rachel Coulston.

2004. Toward Adaptive Conversational Interfaces:

Modeling Speech Convergence with Animated Per-

sonas. ACM Trans. Comput.-Hum. Interact. 11, 3: 300–

328. https://doi.org/10.1145/1017494.1017498

49. Aasish Pappu and Alexander Rudnicky. 2013. Deploy-

ing Speech Interfaces to the Masses. In Proceedings of

the Companion Publication of the 2013 International

Conference on Intelligent User Interfaces Companion

(IUI ’13 Companion), 41–42.

https://doi.org/10.1145/2451176.2451189

50. A. Potamianos and S. Narayanan. 2003. Robust recogni-

tion of children’s speech. IEEE Transactions on Speech

and Audio Processing 11, 6: 603–616.

https://doi.org/10.1109/TSA.2003.818026

51. Janet C. Read and Stuart MacFarlane. 2006. Using the

fun toolkit and other survey methods to gather opinions

in child computer interaction. In Proc. of IDC, 81–88.

52. Ben Shneiderman and Pattie Maes. 1997. Direct Manip-

ulation vs. Interface Agents. interactions 4, 6: 42–61.

https://doi.org/10.1145/267505.267514

53. Lisa Stifelman, Adam Elman, and Anne Sullivan. 2013.

Designing Natural Speech Interactions for the Living

Room. In CHI ’13 Extended Abstracts on Human Fac-

tors in Computing Systems (CHI EA ’13), 1215–1220.

https://doi.org/10.1145/2468356.2468574

54. Sherry Turkle. 2011. Alone Together: Why We Expect

More from Technology and Less from Each Other.

55. J. G. Wilpon and C. N. Jacobsen. 1996. A study of

speech recognition for children and the elderly. In 1996

IEEE International Conference on Acoustics, Speech,

Moving on from Weiser’s Vision of Calm Computing:

Engaging UbiComp Experiences

Yvonne Rogers

School of Informatics, Indiana University, 901 East 10th Street,

Bloomington, IN47408, USA

yrogers@indiana.edu

Abstract. A motivation behind much UbiComp research has been to make our

lives convenient, comfortable and informed, following in the footsteps of

Weiser’s calm computing vision. Three themes that have dominated are context

awareness, ambient intelligence and monitoring/tracking. While these avenues

of research have been fruitful their accomplishments do not match up to any-

thing like Weiser’s world. This paper discusses why this is so and argues that is

time for a change of direction in the field. An alternative agenda is outlined that

focuses on engaging rather than calming people. Humans are very resourceful

at exploiting their environments and extending their capabilities using existing

strategies and tools. I describe how pervasive technologies can be added to the

mix, outlining three areas of practice where there is much potential for profes-

sionals and laypeople alike to combine, adapt and use them in creative and con-

structive ways.

Keywords: calm computing, Weiser, user experiences, engaged living, Ubi-

Comp history, pervasive technologies, proactive computing.

1 Introduction

Mark Weiser’s vision of ubiquitous computing has had an enormous impact on the

directions that the nascent field of UbiComp has taken. A central thesis was that while

“computers for personal use have focused on the excitement of interaction...the most

potentially interesting, challenging and profound change implied by the ubiquitous

computing era is a focus on calm.” [46]. Given the likelihood that computers will be

everywhere, in our environments and even embedded in our bodies, he argued that

they better “stay out of the way” and not overburden us in our everyday lives. In con-

trast, his picture of calm technology portrayed a world of serenity, comfort and aware-

ness, where we are kept perpetually informed of what is happening around us, what is

going to happen and what has just happened. Information would appear in the centre

of our attention when needed and effortlessly disappear into the periphery of our at-

tention when not.

Now regarded as the forefather of UbiComp, Weiser has inspired governments, re-

searchers and developers across the globe. Most prominent was the European Com-

munity’s Disappearing Computer initiative in the late 90s and early 2000s, that

funded a large number of research projects to investigate how information technology

could be diffused into everyday objects and settings and to see how this could lead to

Moving on from Weiser’s Vision of Calm Computing 405

new ways of supporting and enhancing people’s lives that went above and beyond

what was possible using desktop machines. Other ambitious and far-reaching projects

included MIT’s Oxygen, HP’s CoolTown, IBM’s BlueEyes, Philips Vision of the

Future and attempts by various telecom companies and academia to create the ulti-

mate ‘smart home’, e.g., Orange-at-Home and Aware Home. A central aspiration

running through these early efforts was that the environment, the home, and our pos-

sessions would be aware, adapt and respond to our varying comfort needs, individual

moods and information requirements. We would only have to walk into a room, make

a gesture or speak aloud and the environment would bend to our will and respond or

react as deemed appropriate for that point in time.

Considerable effort has gone into realizing Weiser’s vision in terms of developing

frameworks, technologies and infrastructures. Proactive computing was put forward

as an approach to determine how to program computers to take the initiative to act on

people’s behalf [43]. The environment has been augmented with various computa-

tional resources to provide information and services, when and where desired, with

the implicit goal of “assisting everyday life and not overwhelming it” [1]. An assort-

ment of sensors have been experimented with in our homes, hospitals, public build-

ings, physical environments and even our bodies to detect trends and anomalies, pro-

viding a dizzying array of data about our health, movements, changes in the environ-

ment and so on. Algorithms have been developed to analyze the data in order for

inferences to be made about what actions to take for people. In addition, sensed data

is increasingly being used to automate mundane operations and actions that we would

have done in our everyday worlds using conventional knobs, buttons and other physi-

cal controls. For example, our favorite kind of music or TV show that we like to exer-

cise to will automatically play as we enter a gym. Sensed data is also being used to

remind us of things we often forget to do at salient times, such as detecting the ab-

sence of milk in the fridge and messaging us to buy a carton when passing the grocery

store.

But, as advanced and impressive as these endeavors have been they still do not

match up to anything like a world of calm computing. There is an enormous gap be-

tween the dream of comfortable, informed and effortless living and the accomplish-

ments of UbiComp research. As pointed out by Greenfield [20] “we simply don’t do

‘smart’ very well yet” because it involves solving very hard artificial intelligence

problems that in many ways are more challenging than creating an artificial human

[26]. A fundamental stumbling block has been harnessing the huge variability in what

people do, their motives for doing it, when they do it and how they do it. Ethno-

graphic studies of how people manage their lives – ranging from those suffering from

Alzheimer’s Disease to high-powered professionals – have revealed that the specifics

of the context surrounding people’s day-to-day living are much more subtle, fluid and

idiosyncratic than theories of context have led us to believe [40]. This makes it diffi-

cult, if not impossible, to try to implement context in any practical sense and from

which to make sensible predictions about what someone is feeling, wanting or need-

ing at a given moment. Hence, while it has been possible to develop a range of simple

UbiComp systems that can offer relevant information at opportune moments (e.g.,

reminding and recommending to us things that are considered useful and important) it

is proving to be much more difficult to build truly smart systems that can understand

or accurately model people’s behaviors, moods and intentions.

406 Y. Rogers

The very idea of calm computing has also raised a number of ethical and social

concerns. Even if it was possible for Weiser’s dream to be fulfilled would we want to

live in such a world? In particular, is it desirable to depend on computers to take on

our day-to-day decision-making and planning activities? Will our abilities to learn,

remember and think for ourselves suffer if we begin to rely increasingly on the envi-

ronment to do them for us? Furthermore, how do designers decide which activities

should be left for humans to control and which are acceptable and valuable for the

environment to take over responsibility for?

In this paper I argue that progress in UbiComp research has been hampered by in-

tractable computational and ethical problems and that we need to begin taking stock

of both the dream and developments in the field. In particular, we need to rethink the

value and role of calm and proactive computing as main driving forces. It is without

question that Weiser’s enormous legacy will (and should) continue to have an impact

on UbiComp developments. However, sufficient time has passed since his untimely

death and it should be possible now for researchers to take a critical stance. As part of

this exercise, I propose that the field needs to broaden its scope, setting and address-

ing other goals that are more attainable and down-to-earth. New agendas need also to

be outlined that can guide, stimulate and challenge UbiComp (and other) researchers

and developers, building upon the growing body of research in the field.

To this end, I propose one such alternative agenda which focuses on designing

UbiComp technologies for engaging user experiences. It argues for a significant shift

from proactive computing to proactive people; where UbiComp technologies are

designed not to do things for people but to engage them more actively in what they

currently do. Rather than calm living it promotes engaged living, where technology is

designed to enable people to do what they want, need or never even considered before

by acting in and upon the environment. Instead of embedding pervasive computing

everywhere in the environment it considers how UbiComp technologies can be cre-

ated as ensembles or ecologies of resources, that can be mobile and/or fixed, to serve

specific purposes and be situated in particular places. Furthermore, it argues that peo-

ple rather than computers should take the initiative to be constructive, creative and,

ultimately, in control of their interactions with the world – in novel and extensive

ways.

While this agenda might appear to be a regressive step and even an anathema to

some ardent followers of Weiser’s vision, I argue that it (and other agendas) will turn

out to be more beneficial for society than persisting with following an unrealistic

goal. Current technological developments together with emerging findings from user

studies, showing how human activities have been positively extended by ‘bounded’

(as opposed to pervasive) technologies, suggest that much can be gained from re-

conceptualizing UbiComp in terms of designing user experiences that creatively,

excitedly, and constructively extend what people currently do. This does not mean

that the main tenet of Weiser’s vision be discarded (i.e., computers appearing when

needed and disappearing when not) but rather we begin to entertain other possibilities

– besides calmness – for steering UbiComp research. Examples include extending and

supporting personal, cognitive and social processes such as habit-changing, problem-

solving, creating, analyzing, learning or performing a skill. Ultimately, research and

development should be driven by a better understanding of human activity rather than

Moving on from Weiser’s Vision of Calm Computing 407

what has tended to happen, namely, “daring to intervene, clumsily, in situations that

already work reasonably well” [20, p231].

In the remainder of this paper I offer a constructive critique of Weiser’s vision and

the subsequent research that has followed in its footsteps. I then outline an alternative

agenda for UbiComp, highlighting pertinent questions, concerns and illustrative ex-

amples of how it can be achieved.

2 Weiser’s Vision Revisited and Early Research

To illustrate how his early vision of ubiquitous computing could work, Weiser [47]

presented a detailed scenario about a day in the life of Sal, an executive single mother.

The scenario describes what Sal gets up to, as she moves from her domestic world to

her work place, during which she is perpetually informed of the goings on of her

family, neighbors, fellow citizens and work colleagues. With this knowledge she is

able to keep up-to-date, avoid obstacles, make the most of her time and conduct her

work – all in smooth and effective ways. The scenario emphasizes coziness, comfort

and effortlessness:

“Sal awakens: she smells coffee. A few minutes ago her alarm clock, alerted by her

restless rolling before waking, had quietly asked “coffee?”, and she had mumbled

“yes.” “Yes” and “no” are the only words it knows.

Sal looks out her windows at her neighborhood. Sunlight and a fence are visible

through one, but through others she sees electronic trails that have been kept for her

of neighbors’ coming and going during the early morning. Privacy conventions and

practical data rates prevent displaying video footage, but time markers electronic

tracks on the neighborhood map let Sal feel cozy in her street.”

In this small excerpt we see how the world evolves around Sal’s assumed needs,

where computers, cameras and sensors are embedded into her world to make her life

super efficient, smooth and calm. It is as if she glides through life, where everything

is done or laid out for her and whenever there is potential for frustration, such as a

traffic jam or parking problem, the invisible computers come to her rescue and gently

inform her of what to do and where to go. It is worth drawing an analogy here with

the world of the landed aristocracy in Victorian England who’s day-to-day live was

supported by a raft of servants that were deemed to be invisible to them. This scenario

also highlights the ethical issues that such an informed world needs to address,

namely the importance of establishing appropriate levels of privacy that are consid-

ered acceptable by a community (e.g., having abstract digital trails rather than video

footage to ensure anonymity).

The core topics raised in Weiser’s seminal papers have motivated much subsequent

UbiComp research. Most prominent themes are context-aware computing, ambi-

ent/ubiquitous intelligence and recording/tracking and monitoring. (N.B. It should be

noted that these are not mutually exclusive but overlap in the aims and methods used.)

2.1 Context-Aware Computing

Context-aware computing focuses on detecting, identifying and locating people’s

movements, routines or actions with a view to using this information to provide

408 Y. Rogers

relevant information that may augment or assist a person or persons. Many projects

have been conducted under this heading to the extent that it has been noted that ubiq-

uitous computing is sometimes called context-aware computing [12]. In a nutshell,

context is viewed as something that can be sensed and measured using location, time,

person, activity type and other dimensions. An example of an early context-sensitive

application was comMotion that used location information and a speech output sys-

tem to inform people when they were driving or cycling past a store to buy the grocer-

ies they needed [30].

A motivation behind much context-aware computing is to find ways of compensat-

ing for limitations in human cognition, e.g., attention, memory, learning, comprehen-

sion, and decision-making, through the use of sensor-based and computational tools.

For example, augmented cognition – originating in military research – seeks to de-

velop methods “to open bottlenecks and address the biases and deficits in human

cognition” by continually sensing the ongoing context and inferring what strategies to

employ to help people in their tasks [5].

Key questions in context-aware computing concern what to sense, what form and

what kind of information to represent to augment ongoing activities. A number of

location and tagging technologies have been developed, such as RFID, satellite, GPS

and ultrasonics, to enable certain categories of information to be tracked and detected.

Many of these, however, have been beset with detection and precision limitations,

sometimes resulting in unreliable and inaccurate data. Recent advances in cognitive

radio technology that is software defined (SDR), promises to be more powerful; wire-

less systems will be able to locate and link to locally unused radio frequency, based

on the ability to sense and remember various factors, such as human behavior, making

them more dependable and more aware of their surroundings [4]. The advocates of

this new technology portray its potential for highly complex settings, such as combat

war zones to help commanders from different friendly forces stay appraised of the

latest situation, through voice, data and video links, thereby reducing collateral dam-

age [4].

While newer technological developments may enable more accurate data to be de-

tected and collected it is questionable as to how effectively it can be used. It still in-

volves Herculean efforts to understand, interpret and act upon in real-time and in

meaningful ways. Context-aware systems that attempt to guide a person through cer-

tain activities require models of human behavior and intentionality that are based on

rationality and predictability [40]. However, as already mentioned, people often be-

have in unpredictable and subtle ways in their day-to-day contexts. Therefore, it is

likely that context-aware systems will only ever be successful in highly constrained

settings.

2.2 Ambient and Ubiquitous Intelligence

Another dominant theme that has emerged in the field of UbiComp is ubiquitous or

ambient intelligence, i.e., computational intelligence that is part of both the physical

and the digital worlds. This approach follows on from work in artificial intelligence.

The phrase ‘right place/right time/right means’ has been sloganized with visions of

smart worlds and smart things, embedded with intelligence, that will predict people’s

needs and react accordingly [25]. Instead of reaching for the remote to change the TV

Moving on from Weiser’s Vision of Calm Computing 409

channel the smart entertainment system will do it for us, instead of browsing the web

the smart internet will find the information we need and so on. Just as it is becoming

increasingly common place for supermarkets to automatically open their doors as we

walk towards them, toilets to flush when we stand up and taps to release water as we

wave our hands under them it is envisioned that information will appear on our TVs,

watches, walls, and other displays as and when needed (e.g., children will be alerted

of dangers and tourists will be informed of points of interest when walking through an

unfamiliar city).

However, similar to context-aware computing, ambient intelligence is proving to

be a hard nut to crack. While there have been significant advances in computer vision,

speech recognition and gesture-based detection, the reality of multimodal interfaces –

that can predict and deliver with accuracy and sensitivity what is assumed people

want or need – is a long way off. One of the most well known attempts at implement-

ing ambient intelligence was IBM’s BlueEyes project, that sought to develop com-

puters that could “see” and “feel” like humans. Sensing technology was used to iden-

tify a person’s actions and to extract key information that was then analyzed to deter-

mine the person’s physical, emotional, or informational state. This was intended to be

used to help make people “more productive by performing expected actions or by

providing expected information.” The success of the BlueEyes project, however, was

limited; an example of an achievement that is posted on its website is of a television

that would turn itself on when a person in the room made eye contact with it. To turn

it off, the person could ‘tell’ it to switch off.

Such meager accomplishments in both context-aware computing and ambient intel-

ligence reflect just how difficult it can be to get a machine to behave like a human.

But it is essential that such systems be accurate for them to be accepted by humans in

their everyday context. Reading, interpreting and acting upon people’s moods, inten-

tions, desires, etc, at any given moment in an appropriate way is a highly developed

human skill that when humans get it wrong can lead to misunderstanding. When a

ubiquitous computing system gets it wrong – which is likely to be considerably more

frequent – it is likely to be more frustrating and we are likely to be less forgiving. For

example, when the system decides to switch on the TV because we happen momen-

tarily to stare into space while reading a book, it is likely to be unnerving and ex-

tremely annoying, especially if ‘it’ persistently gets it wrong.

2.3 Recording, Tracking and Monitoring

The push towards developing assistive applications through sensing and alerting has

been most marked for vulnerable people; a number of UbiComp systems have been

built to constantly check up on the elderly, the physically and mentally disabled [34].

The movements, habits, health and mishaps of such people are recorded, tracked and

presented via remote monitors to the families, carers and other people responsible for

them, who can then use the information to make decisions about whether to intervene

or administer alternative forms of medical care or help. In particular, there has been a

move towards developing ubiquitous computing systems to aid elderly people, who

need to be cared for, by helping them take their medicines regularly, checking up on

their physical health, monitoring their whereabouts and detecting when they have

fallen over [e.g., 13].

410 Y. Rogers

A number of assisted living applications and services has also been developed to

help people with loss of vision or deteriorating memory to be more independent in

their lives. For example, Cyber Crumbs was designed to help people with progressive

vision loss find their way around a building using a reader badge system that reads out

directions and warns of obstacles, such as fire hydrants [39]. Cook’s Collage was

developed as an aid for people with memory loss. It replays a series of digital still

images in a comic strip reel format depicting people’s cooking actions in situ, in-

tended to help them remember if they have forgotten a step (e.g., adding a particular

ingredient) after being distracted [45].

A reason for there being so much interest in helping the less able in UbiComp is

that explicit needs and benefits can be readily identified for these user groups. More-

over, there is an assumption that pervasive technologies offer more flexibility and

scope for providing solutions compared with other computing technologies since they

can sense, monitor and detect people’s movements, bodily functions, etc., in ways not

possible before. There is a danger, however, that such techniques may probe too far

into the lives of less able people resulting in – albeit unintentionally – ‘extreme’

forms of recording, tracking and monitoring that these people may have no control

over. For example, consider the extent to which a group of researchers went to in

order to help with the care of old people in a residential care home [6]. A variety of

monitoring devices were installed in the home, including badges on the patients and

the caregivers and switches on the room doors that detected when they were open or

closed. Load sensors were also used to measure and monitor weight changes of peo-

ple while in their beds; the primary aim was to track trends in weight gain or loss over

time. But the sensors could also be used to infer how well someone was sleeping. If

significant movement was detected during the night this could enable a caregiver to

see whether the person was having trouble sleeping (and if there was a huge increase

in weight this could be inferred as someone else getting in or on the bed).

Such panopticon developments elicit a knee-jerk reaction of horror in us. While the

motives behind such projects are altruistic they can also be naïve, overlooking how

vulnerable people’s privacy and self-respect may be being violated. Not surprisingly,

there has been enormous concern by the media and other social scientists about the

social implications of recording, tracking and re-representing people’s movements,

conversations, actions and transactions. Inevitably, a focus has been on the negative

aspects, namely a person’s right to privacy being breached. Is it right to be videoing

and sensing people when sleeping, eating, etc., especially when they are not at their

best [2]? Is it right to be providing information to other family members about their

granny’s sleeping habits, especially if it can be inferred from the sensed data that she

might have got into bed with another patient, which none of the vested parties might

want to share or let the others know about.

While most projects are sensitive to the privacy and ethical problems surrounding

the monitoring of people, they are not easy to solve and have ended up overwhelming

UbiComp research. Indeed, much of the discussion about the human aspects in the

field has been primarily about the trade-offs between security and privacy, conven-

ience and privacy, and informedness and privacy. This focus has often been at the

expense of other human concerns receiving less airing, such as how recording, track-

ing and re-representing movements and other information can be used to facilitate

social and cognitive processes.

Moving on from Weiser’s Vision of Calm Computing 411

My intention here is not to diminish the importance of awareness, ambience and

monitoring to detect and inform people in their everyday lives, together with the ethi-

cal and social issues they raise. Rather, my overview of the projects in these areas has

revealed how difficult it is to build calm computing systems and yet the attempts have

largely dominated the field of UbiComp. Those that have tried have fallen short, re-

sulting in prototype systems that can sometimes appear to be trivial or demeaning.

Conversely, there has been less focus on other areas of research that could prove to be

easier to achieve and potentially of more benefit to society. The time is ripe for other

directions to take center stage in UbiComp. One such avenue promoted here is to

consider how humankind’s evolved practices of science, learning, health, work and

play can be enhanced. This involves thinking about UbiComp not in terms of embed-

ding the environment with all manner of pervasive technologies but instead as

bounded ensembles of entities (e.g., tools, surfaces and lenses) that can be mobile,

collaborative or remote, through which information, other people and the environment

are viewed and interacted with when needed. Importantly, it argues for rethinking the

nature of our relationship with the computer.

3 A New Agenda for UbiComp: Engaging User Experiences

I suggest here that it is highly profitable to recast UbiComp research in the context of

a central motivation that computers were originally designed for, namely, as tools,

devices and systems that can extend and engage people in their activities and pursuits.

My reason for proposing this is based on the success of researchers who have started

to take this approach. In particular, a number of user studies, exploring how UbiComp

technologies are being appropriated, are revealing how the ‘excitement of interaction’

can be brought back in innovative ways; that is not frustrating and which is quite

different from that experienced with desktop applications. For example, various

mixed reality, physical-digital spaces and sensor-rich physical environments have

been developed to enable people to engage and use multiple dynamic representations

in novel ways: in scientific and working practices and in collaborative learning and

experimental games. More extensive inquiries and decisions have been enabled in

situ, e.g., determining the effects of deforestation in different continents and working

out when is the best time to spray or pick grapes in a vineyard.

Recently, world famous computer scientist John Seely Brown put forward his up-

dated vision of UbiComp 1 in a keynote, outlining ‘a common sense’ model that em-

phasizes how UbiComp can help to catalyze creativity [41]. He proposed that creating

and learning be seen as integral to our work and leisure that are formed through re-

creation and appropriation activities. In a similar vein, I argue that it is timely to

switch from a reactive view of people towards a more proactive one. Instead of aug-

menting the environment to reduce the need for humans to think for themselves about

what to do, what to select, etc., and doing it for them, we should consider how Ubi-

Comp technologies can be designed to augment the human intellect so that people can

perform ever greater feats, extending their ability to learn, make decisions, reason,

create, solve complex problems and generate innovative ideas. Weiser’s idea that

1 John Seely Brown was a co-author of the paper written by Weiser on calm technology.

412 Y. Rogers

technologies be designed to be ‘so embedded, so fitting and so natural’ that we use

them without thinking about them needs to be counter-balanced; we should also be

designing them to be exciting, stimulating and even provocative – causing us to re-

flect upon and think about our interactions with them. While Weiser promoted the

advantages of calm computing I advocate the benefits of engaging UbiComp experi-

ences that provoke us to learn, understand and reflect more upon our interactions with

technologies and each other.

A central concern of the engaging UbiComp experiences agenda is to fathom out

how best to represent and present information that is accessible via different surfaces,

devices and tools for the activity at hand. This requires determining how to make

intelligible, usable and useful, the recordings of science, medicine, etc., that are

streaming from an increasing array of sensors placed throughout the world. It also

entails figuring out how to integrate and replay, in meaningful and powerful ways, the

masses of digital recordings that are begin gathered and archived such that profes-

sionals and researchers can perform new forms of computation and problem-solving,

leading to novel insights. In addition, it involves experimenting more with creative

and constructive uses of UbiComp technologies and archived digital material that will

excite and even make people feel uncomfortable.

In terms of who should benefit, it is useful to think of how UbiComp technologies

can be developed not for the Sal’s of the world, but for particular domains that can be

set up and customized by an individual firm or organization, such as for agriculture

production, environmental restoration or retailing. At a smaller scale, it is important to

consider how suitable combinations of sensors, mobile devices, shared displays, and

computational devices can be assembled by non-UbiComp experts (such as scientists,

teachers, doctors) that they can learn, customize and ‘mash’ (i.e., combine together

different components to create a new use). Such toolkits should not need an army of

computer scientists to set up and maintain, rather the inhabitants of ubiquitous worlds

should be able to take an active part in controlling their set up, evolution and destruc-

tion. Their benefits should be clear: enabling quite different forms of information flow

(i.e., ways and means of accessing information) and information management (i.e.,

ways of storing, recording, and re-using information) from older technologies, making

it possible for non-UbiCompers to begin to see how to and subsequently develop their

own systems that can make a difference to their worlds. In so doing, there should be

an emphasis on providing the means by which to augment and extend existing prac-

tices of working, learning and science.

As quoted by Bruner [10] “to assist the development of the powers of the mind is

to provide amplification systems to which human beings, equipped with appropriate

skills, can link themselves” (p.53). To enable this to happen requires a better under-

standing of existing human practices, be it learning, working, communicating, etc.

Part of this reconceptualization should be to examine the interplay between technolo-

gies and their settings in terms of practice and appropriation [15]. “Practices develop

around technologies, and technologies are adapted and incorporated into practices.”

(Dourish, 2001, p. 204). More studies are needed that examine what people do with

their current tools and devices in their surrounding environments. In addition, more

studies are needed of UbiComp technologies being used in situ or the wild – to help

illuminate how people can construct, appropriate and use them [e.g., 16, 22, 23, 29].

Moving on from Weiser’s Vision of Calm Computing 413

With respect to interaction design issues, we need to consider how to represent and

present data and information that will enable people to more extensively compute,

analyze, integrate, inquire and make decisions; how to design appropriate kinds of

interfaces and interaction styles for combinations of devices, displays and tools; and

how to provide transparent systems that people can understand sufficiently to know

how to control and interact with them. We also need to find ways of enabling profes-

sionals and laypeople alike to build, adapt and leverage UbiComp technologies in

ways that extend and map onto their activities and identified needs.

A more engaging and bounded approach to UbiComp is beginning to happen but in

a scattered way. Three of the most promising areas are described below: (i) playful

and learning practices, (ii) scientific practices and (iii) persuasive practices. They

show how UbiComp technologies can be developed to extend or change human ac-

tivities together with the pertinent issues that need to be addressed. Quite different

practices are covered, reflecting how the scope of UbiComp can be broad but at the

same time targeted at specific users and uses.

3.1 Playful and Learning Practices

One promising approach is to develop small-scale toolkits and sandboxes, comprising

interlinked tools, digital representations and physical artifacts that offer the means by

which to facilitate creative authoring, designing, learning, thinking and playing. By a

sandbox it is not meant the various senses it has been used in computing but more

literally as a physical-digital place, kitted out with objects and tangibles to play and

interact with. Importantly, these should allow different groups of people to participate

in novel activities that will provoke and extend existing repertoires of technology-

augmented learning, playing, improvising and creating. An example of a promising

UbiComp technology toolkit is PicoCrickets, developed at MIT Media Lab, arising

from the work of Mitch Resnick and his colleagues. The toolkit comprises sensors,

motors, lights, microcomputers, and other physical and electrical devices that can be

easily programmed and assembled to make them react, interact and communicate,

enabling “musical sculptures, interactive jewelry, dancing creatures and other playful

inventions” to be created by children and adults alike. An advantage of such light-

weight, off-the-shelf tangible toolkits is that they offer many opportunities for differ-

ent user groups (e.g., educators, consultants) to assemble and appropriate in a range of

settings, such as schools, waiting rooms, playgrounds, national parks, and museums.

A nagging question, however, is how do the benefits of such UbiComp toolkits and

sand boxes compare with those offered by more conventional ones – that are much

cheaper and more practical to make? Is it not the case that children can be highly

creative and imaginative when given simply a cardboard box to play with? If so, why

go to such lengths to provide them with new tools? The debate is redolent of whether

it is better for children to read a book or watch a 3D Imax movie. One is not necessar-

ily better than the other: the two provide quite different experiences, triggering differ-

ent forms of imagination, enjoyment and reflection. Likewise, UbiComp and physical

toys can both provoke and stimulate, but promote different kinds of learning and

collaboration among children. However, a benefit of UbiComp toolkits over physical

artifacts is that they offer new opportunities to combine physical interaction, through

manipulation of objects or tools or through physical body postural movement and

414 Y. Rogers

location, with new ways of interacting, through digital technology. In particular, they

provide different ways of thinking about the world than interacting solely with digital

representations or solely with the physical world. In turn, this can encourage or even

enhance further exploration, discovery, reflection and collaboration [35].

Examples of projects that have pioneered the design of novel physical-digital

spaces to facilitate creativity and reflection include the Hunting of the Snark [32],

Ambient Wood [36], RoomQuake [33] Savannah [17], Environmental Detectives

[27], Drift Table [19] and Feeding Yoshi [7]. Each of these have experimented with

the use of mobile, sensor and fixed technologies in combination with wireless infra-

structures to encourage exploration, invention, and out of the box thinking.

The Hunting of the Snark adventure game provoked young children into observing,

wondering, understanding, and integrating their fragmented experiences of novel

physical-digital spaces that subsequently they reflected upon and shared as a narrative

with each other. A combination of sensor-based, tangible, handheld and wireless

technologies was used to create the physical-digital spaces, where an imaginary vir-

tual creature was purported to be roaming around in. The children had to work out

how to entice the creature to appear in them and then gather evidence about its per-

sonality, moods, etc, by walking with it, feeding it and flying with it. Similarly, Sa-

vannah was designed as a physical-digital game to encourage the development of

children’s conceptual understanding of animal behavior and interactions in an imagi-

nary virtual world. The project used GPS and handheld computers to digitally overlay

a school playing field with a virtual plain. Children took on the roles of lions, had to

hunt animals in the virtual savannah and capture them to maintain energy levels. After

the game, the children reflected on their experiences by interacting with a visualiza-

tion on a large interactive whiteboard, that showed the trails they made in the Savan-

nah and the sounds and images that they encountered at specific place.

The Ambient Wood project used an assortment of UbiComp technologies to en-

courage more self-initiation in inquiry and reflective learning. Various wireless and

sensor technologies, devices and representational media were combined, designed and

choreographed to appear and be used in an ‘ambient’ woodland. Several handcrafted

listening, recording and viewing devices were created to present certain kinds of digi-

tal augmentations, such as sounds of biological processes, images of organisms, and

video clips of life cycles. Some of these were triggered by the children’s exploratory

movements, others were collected by the children, while still others were aggregated

and represented as composite information visualizations of their exploratory behavior.

RoomQuake was designed to encourage children to practice scientific investigatory

practices: an earthquake was simulated in a classroom using a combination of inter-

connected ambient media, string and physical styrofoam balls. The ambient media

provided dynamic readings of the simulated earthquakes, which students then re-

represented as physical models using the physical artifacts. The combination of com-

puter-based simulations and physical-based artifacts enabled the whole class to take

part in the measuring, modeling, interpreting, sparking much debate and reflection

among the children about the seismic events.

As part of the Equator collaboration, a number of innovative ‘seamful games’ have

been developed. The inherent limitations of ubiquitous technologies have been delib-

erately exploited to provoke the players into thinking about and acting upon their

significance to the ongoing activity. Two examples are Treasure in which players had

Moving on from Weiser’s Vision of Calm Computing 415

to move in and out of a wireless network connectivity to collect and then deposit gold

tokens and Feeding Yoshi where the players were required to feed virtual creatures

scattered around a city with virtual fruits that popped up on their displays as a result

of their location and activity therein.

Evaluations of this emerging genre of physical-digital spaces for learning and play-

ing have been positive, highlighting enhanced understanding and an immense sense of

engagement. Children and adults have been able to step back and think about what

they are doing when taking part in the game or learning experience, examining the

rationale behind their choices when acting out and interacting with the UbiComp-

based technologies in the space. However, many of the pioneering projects were tech-

nology, resource and researcher intensive. While guidance is now beginning to appear

to help those wanting to design UbiComp-based learning and playing experiences

[e.g., 9, 36] we need also to strive towards creating the next generation of physical-

digital spaces and toolkits that will be as easy, cheap and popular to construct as Lego

kits once were.

3.2 Scientific Practices

Another area where UbiComp has great potential for augmenting human activities is

the practice of scientific inquiry and research. Currently, the sciences are going

through a major transformation in terms of how they are studied and the computa-

tional tools that are used and needed. Microsoft’s 2020 Science report – a comprehen-

sive vision of science for the next 14 years written by a group of internationally

distinguished scientists – outlines this paradigm shift [31]. It points out how new

conceptual and technological tools are needed that scientists from different fields can

“understand and learn from each other’s solutions, and ultimately for scientists to

acquire a set of widely applicable complex problem solving capabilities”. These in-

clude new programming, computational, analysis and publication tools. There is much

scope, too, for utilizing UbiComp technologies to enhance computation thinking,

through integrating sensor-based instrumentation in the medical, environmental and

chemical sciences. The ability to deliver multiple streams of dynamic data to scien-

tists, however, needs to be matched by powerful interfaces that allow them to manipu-

late and share them in new ways, from any location whether in the lab or in the field.

Areas where there is likely to be obvious benefits to scientists through the integra-

tion of UbiComp and computational tools are environmental science and climate

change. These involve collaborative visualization of scientific data, mobile access to

data and capture of data from sensors deployed in the physical world. Being able to

gain a bigger, better and more accurate picture of the environmental processes may

help scientists make more accurate predictions and anticipate more effectively natural

disasters, such as tsunamis, volcanoes, earthquakes and flooding. However, it may not

simply be a case of more is more. New ways of managing the burgeoning datasets

needs to be developed, that can be largely automated, but which also allows scientists

to have effective windows, lenses etc., into so that they can interpret and make intelli-

gible inferences from them at relevant times.

The 2020 report notes how tomorrow’s scientists will need to make sense of the

masses of data by becoming more computationally literate – in the sense of knowing

how to make inferences from the emerging patterns and anomalies that the new

416 Y. Rogers

generation of software analysis tools provide. To this end, a quite different mindset is

needed in schools for how science is taught. The design of new learning experiences

that utilize UbiComp technologies, both indoors and outdoors, need to be developed

to seed in young children the sense of what is involved in practicing new forms of

complex, computational science. An example of how this can be achieved is the em-

bedded phenomena approach; scientific phenomena are simulated using UbiComp

technologies, for long periods of time, to create opportunities for groups of students to

explore ‘patient’ science [32]. Essentially, this involves the accumulation, analysis

and representation of data collected from multiple computational devices over ex-

tended periods of observation in the classroom or other sites. In so doing, it allows

students to engage in the collaborative practice of scientific investigation that requires

hard computational thinking but which is also exciting, creative and authentic. A core

challenge, therefore, is to find ways of designing novel science learning experiences

that capitalize on the benefits of combining UbiComp and PC technologies that can be

used over extended periods.

3.3 Persuasive Practices

The third area where there is much potential for using UbiComp technologies to en-

gage people is as part of self-monitoring and behavioral change programs. While a

range of persuasive technologies (e.g., adverts, websites, posters) has already been

developed to change people’s attitudes and behaviors, based on models of social

learning [18], UbiComp technologies provide opportunities for new techniques. Spe-

cifically, mobile devices, such as PDAs coupled with on-body sensors, can be de-

signed to enable people to take control and change their habits or lifestyles to be

healthier by taking account of and acting upon dynamically updated information pro-

vided by them. For example, Intille and his group are exploring how mobile computa-

tional tools for assessing behavioral change, based on social psychology models, can

be developed to motivate physical activity and healthy eating.

A key question that needs to be addressed is whether UbiComp technologies are

more (or less) effective compared with other technologies in changing behavior. A

diversity of media-based techniques (e.g., pop-up warning messages, reminders,

prompts, personalized messages) has been previously used to draw people’s attention

to certain kinds of information to change what they do or think at a given point. In

terms of helping people give up habits (e.g., smoking, excessive eating) they have had

mixed results since people often relapse. It is in the long-term context that UbiComp

technologies may prove to be most effective, being able to monitor certain aspects of

people’s behavior and represent this information at critically weak moments in a ca-

joling way. A constant but gentle ‘nagging’ mechanism may also be effective at per-

suading people to do something they might not have otherwise done or to not to do

something they are tempted to do. For example, a collaborative cell phone application

integrated with a pedometer was used to encourage cliques of teenage girls to monitor

their levels of exercise and learn more about nutrition in the context of their everyday

activities [44]. The software was designed to present the monitored process (e.g.,

walking) in a way that made it easy for the girls to compute and make inferences of

how well they were doing in terms of the number of steps taken relative to each other.

A preliminary study showed that such a collaborative self-monitoring system was

Moving on from Weiser’s Vision of Calm Computing 417

effective at increasing the girl’s awareness of their diet, level of exercise and enabling

them to understand the computations involved in burning food during different kinds

of exercise. But most significantly, it enabled the girls to share and discuss this infor-

mation with each other in their private clique, capitalizing on both the persuasive

technology and peer pressure.

Incorporating fun into the interface can also be an effective strategy; for example,

Nintendo’s Pocket Pikachu with pedometer attached was designed to motivate chil-

dren into being more physically active on a consistent basis. The owner of the digital

pet that ‘lives’ in the device is required to walk, run or jump each day to keep it alive.

If the owner does not exercise for a week the virtual pet becomes unhappy and even-

tually dies. This can be a powerful means of persuasion given that children often

become emotionally attached to their virtual pets, especially when they start to care

for them.

UbiComp technologies can also be used to reduce bad habits through explicitly

providing dynamic information that someone would not have been aware of other-

wise. In so doing, it can make them actively think about their behavior and modify it

accordingly. The WaterBot system was developed using a special monitoring and

feedback device to reduce householder’s usage of water in their homes – based on the

premise that many people are simply unaware of how wasteful they are [3]. A sensor-

based system was developed that provided positive auditory messages and chimes

when the tap was turned off. A central idea was to encourage members of the house-

hold to talk to one another about their relative levels of water usage provided by the

display and to try to out do one another in the amount of water used.

But to what extent do UbiComp technologies, designed for persuasive uses, differ

from the other forms of monitoring that were critiqued earlier in the paper? A main

difference is that there is more active involvement of those being monitored in attain-

ing their desired behavior change compared with those who were being monitored

and assisted in care homes. The objective is to enable people, themselves, to engage

with the collected information, by monitoring, understanding, interpreting and acting

upon it – and not the environment or others to act upon their behalf. Much of the

research to date in UbiComp and healthcare has focussed on automated bio-

monitoring of physiological processes, such as EEGs and heart rate, which others, i.e.,

specialists, examine and use to monitor their patient’s health. In contrast, persuasive

technologies are intended to provide dynamic information about a behavioral process

that will encourage people from doing or not doing something, by being alerted

and/or made aware of the consequences of what they are about to do. Moreover, de-

signing a device to be solely in the control of the users (and their social group) en-

ables them to be the owners of the collected data. This circumvents the need to be

centrally concerned with privacy issues, allowing the focus of the research to be more

oriented towards considering how best to design dynamically updated information to

support cognitive and social change. A challenge, however, in this area is for long

term studies to be conducted that can convincingly show that it is the perpetual and

time-sensitive nature of the sensed data and the type of feedback provided that con-

tributes to behavioral modification.

418 Y. Rogers

4 Conclusions

Many of the research projects that have followed in the footsteps of Weiser’s vision

of calm computing have been disappointing; their achievements being limited by the

extent to which they have been able to program computers to act on behalf of humans.

Just as ‘strong’ AI failed to achieve its goals – where it was assumed that “the com-

puter is not merely a tool in the study of the mind; rather, the appropriately

programmed computer really is a mind” [41], it appears that ‘strong’ UbiComp is

suffering from the same fate. And just as ‘weak’ AI 2 revived AI’s fortunes, so, too,

can ‘weak’ UbiComp bring success to the field. This will involve pursuing more prac-

tical goals and addressing less ambitious challenges; where ensembles of technologies

are designed for specific activities to be used by people in bounded locations. To

make this happen, however, requires moving from a mindset that wants to make the

environment smart and proactive to one that enables people, themselves, to be smarter

and proactive in their everyday and working practices. Three areas of research were

suggested as to how this could be achieved; but, equally, there are others where there

is much potential for enhancing and extending human activities (e.g., vineyard com-

puting [11], firefighting [24] and sports). As part of the expansion of UbiComp, a

wider range of human aspects should be considered, drawing upon alternative theory,

guiding frameworks and metaphors [c.f. 8, 15]. To enable other human concerns to

become more prominent, however, requires the hefty weight of privacy and other

related ethical issues on UbiComp’s shoulders to be lessoned.

The ‘excitement of interaction’ that Weiser suggested forsaking in the pursuit of a

vision of calm living should be embraced again, enabling users, designers and re-

searchers to participate in the creation of a new generation of user experiences that go

beyond what is currently possible with our existing bricolage of tools and media. We

should be provoking people in their scientific, learning, analytic, creative, playing and

personal activities and pursuit. Finally, while we have been privileged to have had such

a great visionary, whose legacy has done so much to help shape the field, it is timely

for a new set of ideas, challenges and goals to come to the fore and open up the field.

Acknowledgements

Thanks to Tom Rodden for his suggestions on an earlier draft and the anonymous

reviewers for their constructive comments.

References

1. Abowd, G.D., Mynatt. E.D.: Charting past, present, and future research in ubiquitous com-

puting. ACM Transactions on Computer-Human Interaction, 7 (2000) 29-58

2. Anderson, K., Dourish, P.: Situated Privacies: Do you know where you mother [trucker]

is? In Proceedings of the 11th International Conference on Human-Computer Interaction.

Las Vegas. July 22-27, 2005

2 Weak AI refers to the development of software programs to perform specific problem-solving

or reasoning tasks that do not have to match the way humans do them.

Moving on from Weiser’s Vision of Calm Computing 419

3. Arroyo, E., Bonnanni, L., Selker, T.: WaterBot: exploring feedback and persuasive tech-

niques at the sink. In CHI Proceedings, ACM, New York, 631-639, 2005

4. Ashley, S.: Cognitive Radio, Scientific American, (March 2006), 67-73

5. Augmented Cognition International Society. http://www.augmentedcognition.org/, Re-

trieved on 30/03/2006

6. Beckwith, R., Lederer, S.: Designing for one’s dotage: UbiComp and Residential Care fa-

cilities. Conference on the Networked Home and the Home of the Future (HOIT 2003), Ir-

vine, CA: April 2003

7. Bell, M., Chalmers, M., Barkhuus, L., Hall, M., Sherwood, S., Tennent, P., Brown. B.,

Rowland, D., Benford, S., Hampshire, A., Captra, M.: Interweaving mobile games with

everyday life. In Proceedings of CHI’06, Conference on Human Factors in Computing.

ACM Press, (2006) 417-426

8. Bellotti, V., Back, M., Edwards, K., Grinter, R., Henderson. A., Lopes, C.: Making sense

of sensing systems: five questions for designers and researchers. In Proceedings of

CHI’2002, ACM Press, (2002) 415-422

9. Benford, S., Schnädelbach, H., Koleva, B., Anastasi, R., Greenhalgh, C., Rodden, T.,

Green, J., Ghali, A., Pridmore, T., Gaver, B., Boucher, A., Walker, B., Pennington, S.,

Schmidt, A., Gellersen, H., Steed, A.: Expected, sensed, and desired: A framework for de-

signing sensing-based interaction. ACM Trans. Comput.-Hum. Interact. 12 (2005) 3-30

10. Bruner, J.S. The Relevance of Education. Harmondsworth, Middlesex, UK. (1972)

11. Burrell, J., Brooke, T., Beckwith, R.: Vineyard Computing: Sensor Networks in agricul-

tural production, Pervasive Computing, 3(1) (2004) 38-45

12. Chalmers, D., Chalmers, M., Crowcroft, J., Kwiatkowska, M., Milner, R., O’Neill, E.,

Rodden, T., Sassone, V., Sloman, M.: Ubiquitous Computing: Experience, design and sci-

ence. Version 4. http://www-dse.doc.ic.ac.uk/Projects/UbiNet/GC/index.html Retrieved on

30/03/2006

13. Consolvo, S., Roessler, P., Shelton, B., LaMarca, A., Schilit, B., Bly, S.: Technology for

care networks for elders. Pervasive Computing 3 (2004) 22-29

14. Digiens@U-City.: Korea moves into ubiquitous mode.

http://digiens.blogspot.com/2005/08/korea-moves-into-ubiquitous-mode.html. Retrieved

30/03/2006

15. Dourish, P.: Where the action is: the foundation of embodied interaction. MIT, Cambridge,

MA., (2001)

16. Dourish, P., Grinter, B., Delgado de la Flor, J., Joseph, M.: Security in the wild: user

strategies for managing security as an everyday, practical problem. Personal and Ubiqui-

tous Computing, 8 (6) (2004) 391-401

17. Facer, K., Joiner, R., Stanton, D., Reid, J., Hull, R., Kirk, D.: Savannah: mobile gaming

and learning. Journal of Computer Assisted Learning, 20 (2004) 399-409

18. Fogg, B.J.: Persuasive Technology: Using Computers to change what we think and do.

Morgan Kaufmann Publishers, San Fransisco. (2003)

19. Gaver, W. W., Bowers, J., Boucher, A., Gellersen, H., Pennington, S., Schmidt, A., Steed,

A., Villars, N., Walker, B.: The drift table: designing for ludic engagement. In Proceedings

of CHI Extended Abstracts (2004) 885-900.

20. Greenfield, A.: Everyware: The Dawning Age of Ubiquitous Computing. New Riders,

Berkeley, CA. (2006)

21. Intel Research at Intel: Research Seattle.

www.intel.com/research/network/seattle\_collab.htm. Retrieved on 20/03/2006.

420 Y. Rogers

22. Intille, S., Larson, K., Beaudin, J., Nawyn, J., Munguia Tapia, E., Kaushik, P.: A living

laboratory for the design and evaluation of ubiquitous computing technologies. In Proceed-

ings of CHI Extended Abstracts (2005) 1941-1944

23. Intille, S.S., Bao, L., Munguia Tapia, E., Rondoni, J.: Acquiring in situ training data for

context-aware ubiquitous computing applications. In Proceedings CHI (2004) 1-8

24. Jiang, X., Chen, N.Y., Hong, J.I., Wang, K., Takayama, L.A., Landay, J.A.: Siren: Con-

text-aware Computing for Firefighting. In Proceedings of Second International Conference

on Pervasive Computing. Lecture Notes in Computer Science, Springer Berlin Heidelberg

87-105 (2004)

25. Journal of Ubiquitous Computing and Intelligence. www.aspbs.com/juci.html Retrieved

20/03/2006/

26. Kindberg, T., Fox, A.: System Software for Ubiquitous Computing. IEEE Pervasive Com-

puting, 1 (1) (2002) 70-81

27. Klopfer, E., K. Squire.: Environmental Detectives – The Development of an Augmented

Reality Platform for Environmental Simulations. Educational Technology Research and

Development. (2005)

28. Krikke, J.: T-Engine: Japan’s Ubiquitous Computing Architecture is ready for prime time.

Pervasive Computing (2005) 4-9

29. LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., Sohn, T.,

Howard, J., Hughes, J., Potter, F., Tabert, J., Powledge, P., Borriello, G., Schilit. B.: Place

Lab: Device Positioning Using Radio Beacons in the Wild, Intel Research, IRS-TR-04-

016, (2004) http://placelab.org/publications/pubs/IRS-TR-04-016.pdf

30. Marmasse, N., Schmandt, C.: Location-aware information delivery with commotion, In

HUC 2000 Proceedings, Springer-Verlag, (2000) 157-171

31. Microsoft 2020 Science.: http://research.microsoft.com/towards2020science/. Retrieved

30/03/2006

32. Moher, T.: Embedded Phenomena: Supporting science learning with classroom-sized-

distribution simulations. In Proceedings of CHI 2006

33. Moher, T., Hussain, S., Halter, T., Kilb, D..: RoomQuake: embedding dynamic phenomena

within the physical space of an elementary school classroom. Extended Abstracts, In Pro-

ceedings of CHI’05, Conference on Human Factors in Computing Systems. ACM Press

(2005) 1655-1668

34. Mynatt, E., Melenhorst, A., Fisk, A.D., Rogers, W.: Aware technologies for aging in place:

Understanding user needs and attitudes. Pervasive Computing (2004) 36-41

35. Price, S. Rogers, Y. Let’s get physical: the learning benefits of interacting in digitally

augmented physical spaces. Journal of Computers and Education, 43 (2004) 137-151

36. Rogers, Y., Muller, H.: A framework for designing sensor-based interactions to promote

exploration and reflection. International Journal of Human-Computer Studies, 64 (1)

(2005) 1-15

37. Rogers, Y., Price, S., Fitzpatrick, G., Fleck, R., Harris, E., Smith, H., Randell, C., Muller,

H., O’Malley, C., Stanton, D., Thompson, M., Weal, M.: Ambient Wood: Designing new

forms of digital augmentation for learning outdoors. In Proceedings of Interaction Design

and Children, ACM (2004) 1-8

38. Rogers, Y., Scaife, M., Harris, E., Phelps, T., Price, S., Smith, H., Muller, H., Randall, C.,

Moss, A., Taylor, I., Stanton, D., O’Malley, C., Corke, G., Gabrielli, S.: Things aren’t

what they seem to be: innovation through technology inspiration. In Proceedings of

DIS’2002 Designing Interactive Systems, ACM Press, (2002) 373-379

39. Ross, D.A.: Cyber Crumbs for successful aging with vision loss. Pervasive Computing, 3

(2004) 30-35

Moving on from Weiser’s Vision of Calm Computing 421

40. Salvador, T., Anderson, K. Practical Considerations of Context for Context Based Sys-

tems: An Example from an Ethnographic Case Study of a Man Diagnosed with Early On-

set Alzheimer’s Disease. In UbiComp’03 Proceedings, A.K. Dey et al. (Eds.), LNCS 2864,

Springer-Verlag Berlin Heidelberg, 243-255, 2003

41. Seely Brown, J.: Ubiquitous Computing and beyond – an emerging new common sense

model. www.johnseelybrown.com/JSB.pdf. Retrieved 20/03/2006

42. Stirling, B.: Without Vision, the People Perish. Speech Given at CRA Conference on

Grand Research Challenges in Computer Science and Engineering. Airlie House, Warren-

ton, Virginia, June 23, 2002 www.cra.org/Activities/grand.challenges/sterling.html Re-

trieved 20/03/2006

43. Tennenhouse, D.L. “Proactive Computing,” Communications of the ACM 43, No. 5, 43–

50, 2000

44. Toscos, T., Faber, A., An, S., Gandhi, M.; Chick Clique: Persuasive Technology to Moti-

vate Teenage Girls to Exercise. In CHI’06 Extended Abstracts on Human Factor in Com-

puting Systems, ACM Press (2006) 1873-1878

45. Tran, Q., Calcaterra, G., Mynatt, E.: Cook's Collage: Deja Vu Display for a Home Kitchen.

In Proceedings of HOIT 2005, 15-32

46. Weiser, M., Brown, J.S.: The coming age of calm technology. (1996)

www.ubiq.com/hypertext/weiser/acmfuture2endnote.htm. Retrieved 20/03/2006/

47. Weiser, M.: The computer for the 21st century. Scientific American (1991) 94–104

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-10

The Lucent Web site is built hierarchically, in the sense that

pages deeper in the directory tree represent more detailed infor-

mation than those at shallower levels. At its busiest, there can

be as many as 300 people browsing www.lucent.com; while

during the pre-dawn hours there can be as few as 5 simultaneous

visitors. Our sonification is designed to convey qualitative infor-

mation about site usage, answering questions like:

Overall, is the site busy or quiet?

What proportion of the visitors are delving for specific in-

formation deep within the site, as compared to those visitors

who are “just passing through,” glancing briefly at the home

page and then moving on?

How are users distributed across the various content areas

of the site?

Which portions of the site are visited together? What kinds

of patterns do we find in user behavior?

We think of this sonification as one possible “background” infor-

mation stream that can inform content providers, Web designers

and even the visitors themselves.

2.1.1. Sonification design

Our audio display makes use of the hierarchical structure of the

content offered by www.lucent.com. First, a unique pitch was

used to identify each of five high-level subdomains within the site:

/micro, representing Lucent’s microelectronics design and man-

ufacturing business (now Agere Systems); /enterprise, for

the enterprise systems and software business (now Avaya Com-

munications); /minds, a corporate introduction to Bell Labs re-

search; /press, a collection of press releases and investor infor-

mation; and /search, the local search engine for the site.

The total number of visitors accessing any information from a

subdomain affects the loudness and tonal balance of a low-register

drone at the associated pitch. Visitors requesting content deeper in

the site are represented by higher-pitched pulsing tones (separated

by one or two octaves from the base pitch for the subdomain):

the faster the pulse, the more people are accessing that area, and

the greater the proportion of high-register sounds, the more de-

tailed the content. By assigning well-separated pitches to each

subdomain, shifts in activity both within and between the areas

can be heard. In Table 1 we present a simple mapping of data col-

lected by the Lucent Web server to a continuously time-varying

vector of usage statistics. In the category of Overall browsing, we

count any visitor accessing content pages (HTML, PostScript or

PDF) from the indicated subdomain. A Mid-Level access is a re-

quest for content two or more directories down. Simple examples

are /micro/K56flex/index.html (information on a brand

of 56K modem) and /press/0101/010118.nsb.html (a

press release for January 18, 2001). The final category, Deep

browsing, refers to pages that are four or more directories down

in the tree. One example is a paper from the April/June 2000

issue of the Bell Labs Technical Journal, located at /minds/

techjournal/apr-jun2000/pdf/paper02.pdf.

Then, the resulting 15 values in Table 1, A1–E3, were mapped

to sound as follows:

Overall activity Measured by A1–E1, voiced with a low-register

drone. The aggregate number of visitors accessing infor-

mation within each of the five areas modulates the loudness

of each of the five pitches.

/micro /enterprise /minds /press /search

Overall A1 B1 C1 D1 E1

Mid-Level A2 B2 C2 D2 E2

Deep A3 B3 C3 D3 E3

Table 1: Mapping used for Web site traffic example. Overall ac-

tivity records the movements of all users; Mid-Level counts users

2 or 3 directories into the site; Deep browsing consists of users 4+

directories down.

Mid-Level browsing Measured by A2–E2 and assigned a rhyth-

mic middle-register tone pulse; pulse loudness and repeti-

tion speed rises and the timbral brightness increases as the

volume of mid-level browsing increases. There are five in-

dependent pulses, each at a different fixed pitch, represent-

ing the five content areas.

Deep browsing Measured by A3–E3 and made audible via rhyth-

mic high-register “ting” sounds (plucked steel string sam-

ples). Loudness and repetition speed rises as the volume of

deep browsing increases. Again, there are five independent

“ting” sounds, each at a different fixed pitch, representing

the five content areas.

We used pitch groups that were consonant, and for the sounds that

incorporated rhythm (A2–E3), the phase and frequency of each

pulse in the matrix varies independently, yielding a sound with a

changing rhythmic texture but no fixed beat.

The purpose of this sonification is to make interpretable the

activities of users on a Web site. Therefore, the stream of hits be-

ing processed by a Web server (reduced to include only the HTML,

PostScript and PDF documents) needs to be transformed to extract

meaningful user-level data. A real-time monitoring tool was devel-

oped that maintains a bank of active visits (recording separately the

activities of all the people browsing the site at a given time) and

updates various statistics with each user request. When cookies or

some other authentication mechanism allows us to recognize re-

turning visitors, the monitor will update a more complicated user

profile that encapsulates previous browsing patterns. Our traffic

sonification as described above takes as input the location of each

visitor within a site at a given point in time. When constructing

more elaborate sound displays, our design will continue to focus

on user activities, drawing more heavily on the statistics culled

by the monitoring tool. This emphasis distinguishes our approach

from sonification methods that assess Web server performance by

making audible statistics relating to server load, HTTP errors, and

agent types [?].

2.1.2. Impressions and extensions

We have created three audio examples for the activity on the Lu-

cent site. Our data were captured on November 11, 1999 and we

created sonifications of the traffic at 6:00 am, an extremely slow

period for the site; noon, a relatively active time; and 2:30 pm,

the point at which the site was busiest. The samples are located at

our project Web site [6]. Even with this relatively straightforward

mapping, one finds compelling patterns. For example, the affinity

between the /enterprise subdomain and the /search facil-

ity can be heard as the pulses for these areas rise and fall together.3

3 While clearly audible, these shifts can really only be precisely associ-

ated with areas after a certain amount of experience with the mapping.

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-11

Also, when comparing moderately active to extremely busy peri-

ods, we find that the number of people digging deep into the site is

not a fixed fraction of the total number of visitors. That is, the vol-

ume of the low-register drones exhibits much more variation than

the components for the other two categories of accesses. Each of

these effects can be verified by examining the logs, reinforcing the

usefulness of our sonification as a tool for constructing hypotheses

about site traffic.

As mentioned at the beginning of this section, Web browsers

offer a rich set of data about the visitor when requesting data from

a server. This display makes use of only the most basic informa-

tion about a visit, namely the depth of pages accessed. In ongoing

work, we are augmenting our sonification with extra features de-

rived both directly from the server data as well as from statistical

navigation models [12] fit for the Web site under study. So far, we

have found that such extensions are most effective when developed

in the context of a particular monitoring application. For example,

an extended version of this ambient display can aid system archi-

tects of large, Web hosting services understand cache performance

and can aid in server provisioning. Another extension will make

greater user of our navigation models and can help designers and

usability engineers better architect Web sites. We will report on

these and other developments through the project Web site [4].

2.2. Chat rooms and bulletin boards

At any given moment, tens of thousands of real-time conversa-

tions are taking place across the Internet on public forums, bulletin

boards and chat sites. To imagine making these conversations si-

multaneously audible evokes an image of uproarious babble. And

yet, in the aggregate, this massive stream of live communication

could exhibit rich thematic structure. Can we find a meaningful

way to listen in to so many conversations, rendering them in a way

that is comprehensible and not overwhelming?

In some sense, a byproduct of our Web traffic sonification is

the creation of a kind of community from the informal gather-

ing of thousands of visitors to a given Web site. Traditionally,

informational Web sites like www.lucent.com have provided

us with very little sense of the other people who are requesting

data from the server. To attract and retain visitors, however, many

commercial sites recognize the potential of the Web to form so-

cial as well as informational networks. As a result, Web-based fo-

rums, message boards and a variety of chat services are common

components of current site designs. While Internet Relay Chat

(IRC) has been a widely used standard since the inception of the

Internet, the popularization of the Web has resulted in a virtual

explosion of chat applications.4 For example, www.yahoo.com

(a US-based Web portal) offers hundreds of separate chat rooms

attracting tens of thousands of visitors a day. Specialized sites

like www.style.com (the homepage for Vogue magazine) or

www.audiworld.com (an resource for Audi owners) have also

found their message boards to be the most frequently accessed

parts of their domains.

To get a sense of the amount of content that is available in

these dynamic formats, we examined sites contained in the DMOZ

Open Directory [3], an open source listing of over 2 million Web

sites compiled and categorized by 33,000 volunteer editors. From

the November 20, 2000 image of the directory, we counted 36,681

4 RC was developed by Jarkko Oikarinen in Finland in the late eighties,

and was originally intended to work as a better substitute for talk on his

bulletin board.

separate sites offering some kind of chat, bulletin board or other

public forum. While we did not examine the activity on all of

these sites, the number is staggering. If we include other peer-

to-peer communication technologies like instant messaging,5 the

amount of dialogue taking place on the Web at any point in time

is almost unfathomable. The goal of our second sonification is

to make interpretable the thousands of streams of dynamic infor-

mation being generated on the Web. In so doing, we attempt to

characterize a global dialogue, integrating political debates, dis-

cussions of current events, and casual exchanges between mem-

bers of virtual communities.

2.2.1. Content monitors and the statistics engine

Our starting point is text. Albeit diverse in style and dynamic in

character, the text (or transcript) of these data sources carries their

meaning. Therefore, any auditory display consisting only of gen-

erated tones would not be able to adequately represent the data

without a very complex codebook. The design of our sonifica-

tion then depends heavily on text-to-speech (TTS). As with the

traffic example in the previous section, we think of the audio out-

put as another background information stream. The incorporation

of spoken components in the sound design poses new challenges,

both practical and aesthetic. For example, simply voicing every

word taking place in a single chat room can produce too much text

to be intelligible when played in real-time and can quickly exhaust

the listener. Instead, we build a hierarchical representation of the

text streams that relies on statistical processing for content organi-

zation and summarization prior to display.

Before considering sonification design, we first had to cre-

ate specialized software agents that would both discover new chat

rooms and message boards, as well as harvest the content posted

to these sites. (See Figure 1 for an overview of our system ar-

chitecture.) Most bulletin boards and some chat applications use

standard HTML to store visitor contributions. In many cases, a

specific login name is required to gain access to the site. For

these situations, we constructed a content agent in Perl, as this

language provides us the most convenient platform for managing

access details (like cookies). The public chat rooms on sites like

chat.yahoo.com can be monitored in this way. For IRC we

built a configurable Java client that polls a particular server for

active channels. Web sites like www.cnn.com (a popular news

portal) and www.financialchat.com (a financial commu-

nity hosting chat services for day traders) offer several IRC rooms,

some of which are tightly moderated.

In addition to collecting content, each monitoring agent also

summarizes the chat stream, identifying basic topics and updating

statistics about the characteristics of the discussion: What percent-

age of visitors are contributing? How often to they contribute and

at what length? Is the room “on topic,” or are many visitors post-

ing comments on very different subjects? Topics are derived from

the chat stream using a variant of generalized sequence mining [7]

that incorporates tags for the different parts of speech. While the

exact details are beyond the scope of this abstract, a generalized se-

quence is a string of words possibly separated by a wildcard, “\*”.

For example, if we let A, B and C denote specific “contentful”

words (say, nouns, adjectives and adverbs), then AB C , A B C

and A B C are all generalized sequences. The wildcard al-

lows us to identify “Gore \* disputes \* election” from the sentences

5 AOL alone records tens of millions of people using their instant mes-

saging service each month.

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-12

Chat

BB

Chat Chat

BB

Sonification

Engine

Stats

Channel

Audio Right

Channel

Audio Left

Engine

Statistics

Text Feedback

Content

Monitor

Content

Monitor

Content

Monitor

Content

Monitor

Content

Monitor

Figure 1: System architecture overview. A large number of content

streams (Chat = chat rooms; BB = Bulletin boards) are gathered by

specialized agents that transmit them in a homogenized format to

the statistics engine. The statistics engine then distills the streams

into a much smaller number of configurable text streams as well

as a number of descriptive vectors. The sonification engine then

“plays” these text and data streams. The entire systems operates in

real-time.

“Vice President Gore filed papers to dispute the presidential elec-

tion,” “Aides for Gore indicated that he has every reason to dispute

the election”, and “Gore is still deciding whether or not to dispute

the election”.

As many posts to chat rooms contain spelling mistakes and

incorrect grammar, assigning words to different parts of speech is

error-prone. However, unlike most applications of statistical nat-

ural language processing, our content monitors update their sum-

maries each time new material is posted and downweight older

contributions. Because our sonification renders these sources in

real-time, small mistakes have little effect on the power of the over-

all display to convey the ideas being discussed.

Each of the content monitors are periodically polled by the

statistics engine (see Figure 1). This Java-application clusters the

different chat rooms and bulletin boards based on their topic and

numerical summaries. As the topic in a room changes over time,

the statistics engine is constantly updating and reformulating clus-

ter membership. Because a content stream can in fact support

a number of simultaneous discussions (the threads of a bulletin

board, say), we employ a soft-clustering technique. In our initial

work, we have used a mixture-based scheme that determines the

number of clusters with an MDL (Minimum Description Length)

criterion [9]. Each room is then assigned a probability that it be-

longs to the different groups. This model also provides for topic

summarization at the cluster-level. Next, a stochastic framework

was developed to sample representative sentences posted to the

chat or bulletin board. When a discussion is extremely unstruc-

tured, this selection is essentially random sampling from all the

contributions added to the chat since the last polling point. In ad-

dition to textual data streams, the statistics engine is also respon-

sible for communicating the various ingredients for the display to

our sonification engine, Max/MSP [2] (see Figure 1). We have

adopted the Open Sound Control [13] protocol from Center for

New Music and Audio Technologies to transfer data between the

statistics engine (running on a Macintosh with LinuxPPC) and the

sonification engine (running on a Macintosh with OS/9).

2.2.2. Sonification design

As with the previous example (Section 2), our goal is to create

a sonification that is both communicative and listenable. Here we

face the additional challenge of incorporating verbal content. With

TTS annotations, it becomes more difficult to intelligibly convey

more than one layer of information through the audio channel. Our

design incorporates spatialization, pitch and timbral differentia-

tion, and rhythm to achieve clarity in the presentation of the hi-

erarchically structured data coming from the statistics engine.

The auditory display cycles through topic clusters, spending

relatively more time on subjects being actively discussed by the

largest numbers of people. Each different topic is assigned a dif-

ferent pitch group, reinforcing subject changes when they occur.

For each cluster, the statistics engine sends three streams of infor-

mation to the sonification engine:

Topics A continuously updated list of up to ten “topics” (the most

frequently appearing words and phrases – generalized se-

quences – mined from the multiple chat streams associated

with the given cluster; the number of topics is configurable,

but ten was chosen based on timing considerations);

Content samples A selection of sample sentences, identified by

the statistics engine as typical or representative, in which

these topics appear;

Content entropy A vector that represents the changing level of

entropy in the source data.

The topics are spoken by the TTS system6 at regular intervals in

a pitched monotone, and are panned alternately hard left and hard

right in the stereo field, creating a sort of rhythmic “call and re-

sponse.” The sample sentences are panned center, and rendered

with limited inflection (as opposed to the pitched monotone of the

topics). The tonal, rhythmic and spatial qualities of the topics con-

trasts sufficiently with the sample sentences to create two distinctly

comprehensible streams of verbal information.

The entropy vector controls an algorithmic piano score. When

entropy is minimal and the discussion in the chat room or bulletin

board is very focused on one subject, chords are played rhythmi-

cally in time with the rhythmic recitation of the topics. As entropy

increases and the conversations diverge, a Gaussian distribution is

used to expand the number, range and dynamics of notes that fall

between the chords. With this audio component, one can easily

differentiate a well-moderated content source from a more free-

form, public chat without distracting from the TTS annotations.

The piano score also serves a secondary function as an accompa-

niment to the vocal foreground, enhancing the compositional bal-

ance and overall musicality of the sound design.

2.2.3. Sample sonification and impressions

On our project Web site [5], we have a sample chat room sonifi-

cation that cycles through three topics. In this sound file, we are

6 The built-in MacOS TTS capability controlled by Max/MSP.

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-13

listening to the output of only three content monitors. Hence, by

design, each topic is confined to a single site. The first portion of

this example (ending at 1:47 into the sample) concerns the recent

recall of Bridgestone tires and was based on a www.cnn.com

chat room. This discussion was heavily moderated and hence the

backing piano score frequently reduces to a simple rhythm. For our

second topic (from 1:47 to 3:21 of the sample) we recorded chat

exchanges on www.financialchat.com one morning when

Yahoo’s stock opened low. In this example, we hear day traders

frantically exchanging predictions about when Yahoo’s stock will

“bounce.” The final topic in this sample (from 3:21 to the end) is

again from www.cnn.com and treats a recent strike by the Screen

Actor’s Guild and the American Federation of Television and Ra-

dio Artists. This chat room was much less moderated than the

previous CNN chat, and the backing piano score reflects that.

Although this example does not make full use of the clustering

capabilities of the statistics engine, the essence of our sonification

design is clear. The audio display provides an informative and

accessible representation of dynamic, textual content. The topic

and content sample streams are easy to separate, and when placed

in the background, call our attention to important new subjects

being discussed on the Web.

2.2.4. Applications and Extensions

Our sonification provides an audible interface to the (now) massive

amount of dynamic content available on the Web. Given the pre-

processing that takes place in the content monitors and the statis-

tics engine, a simple extension is to provide search-like function-

ality. A user can register interest in a certain topic and “tune”

our display to present only rooms where this subject is being dis-

cussed. The necessary ingredients to implement this feature are

all currently available in the statistics engine. Similarly, one can

easily restrict the sites that are used for the display. When a new

subject appears that draws the user’s interest, it is also trivial to

add a feature that would direct the user’s browser to one or more

chats associated with the topic. As a final extension, we have pro-

vided the content monitors with a configurable list of Web sites

that can be used to help disambiguate elements in the chat stream.

For example, the day traders speak in ticker symbols. Providing

the content monitor with the URL for the ticker symbol look-up

service offered by Yahoo allows the content monitor to weave not

only company names but also recent company-related headlines

directly into the stream fed to the statistics engine.

While we have focused mainly on chat and bulletin boards,

this technology can be applied in other settings. We have begun

collaborating with the designers of a natural language interface for

Web-based help systems. Here, we give voice to the hundreds of

simultaneous conversations taking place between Web site visitors

and the automated help system. A similar display can be imagined

for other natural language interfaces, including search engines like

AskJeeves (www.jeeves.com). In general, the practical appli-

cations of this summarization and auditory display tool abound.

3. CONCLUSION AND COMMENTS ON

COLLABORATIVE RESEARCH

The two applications outlined in this paper are the first outcomes of

a collaboration sponsored by Bell Laboratories and the Brooklyn

Academy of Music under the Arts in Multimedia project (AIM).

The goal of AIM is to bring together researchers (in this case a

statistician) and artists (in this case a sound artist), with the ob-

jective of advancing our separate agendas through collaborative

projects. Our work together is predicated on the notion that so-

phistication both in data treatment and aesthetics are crucial to the

successful design of audio displays. Thus, in each of our exam-

ples, we have endeavored to create a result which communicates

information clearly, yet at the same time sounds well composed

and appealing. Moving forward, it is our intention to apply these

techniques both to practical applications, and also to create a series

of artworks. These artworks will use our sonification techniques

to establish a series of real-time listening posts, both on the Web

and in physical locations. The listening posts will tap in to various

points of interest on the Internet, using sound to reveal patterns and

trends that would otherwise remain hidden.

In terms of applications, we are exploring the use of sonifica-

tion to support the design, provisioning and monitoring of commu-

nication networks. A network operations center (NOC), for exam-

ple, routinely receives clues about the health of the system in the

form of text messages generated by routers and switches. An audio

display installed inside a NOC can act as an early warning system

for approaching bottlenecks as well as aid in troubleshooting. By

continued exposure to the sound of a “normally” functioning net-

work, operators will be alerted to system changes that could signal

problems.

Art emerges unexpectedly from experimentations with new

statistical methods or considerations involving practical applica-

tions; and new tools for data analysis and modeling develop in re-

sponse to artistic concerns. Each of us continues to be surprised by

the connections that emerge from rethinking familiar problems in

a new context. Through our project, we hope to illustrate both the

value of art-technology collaborations as well as their necessity,

especially when finding meaning in complex data.

4. REFERENCES

[1] Visual insights. www.visualinsights.com.

[2] Cycling74. Max/msp. www.cycling74.com.

[3] Open directory project. www.dmoz.com.

[4] Ear to the ground. cm.bell-labs.com/stat/ear.

[5] Ear to the ground, chat example.

cm.bell-labs.com/stat/ear/chat.html.

[6] Ear to the ground, web traffic samples.

cm.bell-labs.com/stat/ear/samples.html.

[7] W. Gaul and L. Schmidt-Thieme. Mining web navigation

path fragments. In Proceedings of the Workshop on Web

Mining for E-Commerce – Challenges and Opportunities,

Boston, MA, August 2000.

[8] M. H. Hansen and B. Rubin. The audiences would be the

artists and their life would be the arts. IEEE MultiMedia,

7(2), April 2000.

[9] M. H. Hansen and B. Yu. Model selection and the princi-

ple of minimum description length. Journal of the American

Statistical Association. To appear.

[10] G. Kramer. An introduction to auditory display. In

G. Kramer, editor, Auditory Display. Addison-Wesley, 1994.

[11] N. Minar and J. Donath. Visualizing the crowds at a web site.

In Proceedings of CHI 99.

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-14

[12] R. Sen and M. H. Hansen. Predicting a web user’s next re-

quest based on log data. Submitted to ASA Student Paper

Competition.

[13] M. Wright. Open sound control. cnmat.berkeley.edu.

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-15

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

TAXONOMY AND DEFINITIONS FOR SONIFICATION AND AUDITORY DISPLAY

Thomas Hermann

Neuroinformatics Group

Faculty of Technology, Bielefeld University,Bielefeld, Germany

thermann@techfak.uni-bielefeld.de

ABSTRACT

Sonification is still a relatively young research field and

many terms such as sonification, auditory display, aural-

ization, audification have been used without a precise def-

inition. Recent developments such as the introduction of

Model-Based Sonification, the establishment of interactive

sonification and the increased interest in sonification from

arts have raised the need to revisit the definitions in order

to move towards a clearer terminology. This paper intro-

duces a new definition for sonification and auditory display

that emphasizes the necessary and sufficient conditions for

organized sound to be called sonification. It furthermore

suggests a taxonomy, and discusses the relation between vi-

sualization and sonification. A hierarchy of closed-loop in-

teractions is furthermore introduced. This paper aims to ini-

tiate vivid discussion towards the establishment of a deeper

theory of sonification and auditory display.

1. INTRODUCTION

Auditory Display is still a young research field whose birth

may be perhaps best traced back to the first ICAD confer-

ence1 in 1992 organized by Kramer. The resulting proceed-

ings volume “Auditory Display” [1] is still one of the most

important books in the field. Since then a vast growth of in-

terest, research, and initiatives in auditory display and soni-

fication has occurred. The potential of sound to support hu-

man activity, communication with technical systems and to

explore complex data has been acknowledged [2] and the

field has been established and has clearly left its infancy.

As in every new scientific field, the initial use of terms

lacks coherence and terms are being used with diffuse defi-

nitions. As the field matures and new techniques are discov-

ered, old definitions may appear too narrow, or, in light of

interdisciplinary applications, too unspecific. This is what

motivates the redefinitions in this article.

The shortest accepted definition for sonification is from

Barrass and Kramer et al. [2]: “Sonification is the use of

non-speech audio to convey information”. This definition

excludes speech as this was the primary association in the

1see www.icad.org

auditory display of information at that time. The definition

is unclear about what is meant by conveyance of informa-

tion: are real-world interaction sounds sonifications, e.g. of

the properties of an object that is being hit? Is a computer

necessary for its rendition? As a more specific definition,

the definition in [2] continues:

“Sonification is the transformation of data re-

lations into perceived relations in an acoustic

signal for the purposes of facilitating commu-

nication or interpretation.”

It is significant that the emphasis here is put on the pur-

pose of the usage of sound. This automatically distinguishes

sonification from music, where the purpose is not on the

precise perception of what interactions are done with an in-

strument or what data caused the sound, but on an underly-

ing artistic level that operates on a different level. Often, the

word ‘mapping’ has been used interchangeably with ‘trans-

formation’ in the above definition. This, however, suggests

a severe limitation of sonification towards just mappings be-

tween data and sound – which was perfectly fine at the time

of the definition where such a ‘Parameter-Mapping Sonifi-

cation’ was the dominating paradigm.

However, the introduction of Model-Based Sonification

(MBS) [3, 4] demonstrates methods to explore data by us-

ing sound in a way that is very different from a mapping:

in Parameter-Mapping Sonification, data values are mapped

to acoustic attributes of a sound (in other words: the data

‘play’ an instrument), whereas in MBS sonification models

create and configure dynamic processes that do not make

sound at all without external interactions (in other words:

the data is used to build an instrument or sound-capable

object, while the playing is left to the user). The user ex-

cites the sonification model and receives acoustic responses

that are determined by the temporal evolution of the model.

By doing this, structural information is holistically encoded

into the sound signal, and is no longer a mere mapping of

data to sound. One can perhaps state that data are mapped

to the configurations of sound-capable objects, but not that

they are mapped to sound.

Clearly, sonification models implemented according to

MBS are very much in line with the original idea that sonifi-

ICAD08-1

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

cation allows for the discovery of structures in data through

sound. Therefore there is the need to reformulate or adapt

the definition for sonification to better include such uses of

sound, and beyond that hopefully other possible yet-to-be-

discovered linkages between data and sound.

Another challenge for the definition comes from the use

of sonification in the arts and music: recently more and

more artists incorporate methods from sonification in their

work. What implications does this have for the term sonifi-

cation? Think of scientific visualization vs. art: what is the

difference between a painting and a modern visualization?

Both are certainly organized colors on a surface, both may

have aesthetic qualities, yet they operate on a completely

different level: the painting is viewed for different layers

of interpretation than the visualization. The visualization

is expected to have a precise connection to the underlying

data, else it would be useless for the process of interpret-

ing the data. In viewing the painting, however, the focus

is set more on whether the observer is being touched by it

or what interpretation the painter wants to inspire than what

can be learnt about the underlying data. Analogies between

sonification and music are close-by.

Although music and sonification are both organized

sound, and sonifications can sound like music and vice

versa, and certainly sonifications can be ‘heard as’ music

as pointed out in [5], there are important differences which

are so far not manifest in the definition of sonification.

2. A DEFINITION FOR SONIFICATION

This section introduces a definition for sonification in light

of the aforementioned problems. The definition has been

refined thanks to many fruitful discussions with colleagues

as listed in the acknowledgements and shall be regarded as

a new working definition to foster ongoing discussion in the

community towards a solid terminology.

Definition: A technique that uses data as input, and gener-

ates sound signals (eventually in response to optional addi-

tional excitation or triggering) may be called sonification,

if and only if

(C1) The sound reflects objective properties or relations in

the input data.

(C2) The transformation is systematic. This means that

there is a precise definition provided of how the data

(and optional interactions) cause the sound to change.

(C3) The sonification is reproducible: given the same data

and identical interactions (or triggers) the resulting

sound has to be structurally identical.

(C4) The system can intentionally be used with different

data, and also be used in repetition with the same

data.Data Sonification

Algorithm

systematic

transformation reproducable exchangeability

of data

interactions (optional)

Definition: Sonification

Figure 1: Illustration of the general structure and necessary

conditions for sonification. The yellow box depicts besides

the sonification elements few other components of auditory

displays, see also Sec. 3.

This definition emphasizes important prerequisites for

the scientific utility of sonification. It has several partly un-

expected implications that are to be explored in the follow-

ing discussion.

2.1. Discussion

2.1.1. General Comments

Sonification Techniques: According to the above defini-

tion, the techniques Audification, Earcons, Auditory Icons,

Parameter-Mapping Sonification as well as Model-Based

Sonification are all covered by the definition – they all rep-

resent information/data by using sound in an organized and

well-structured way and they are therefore different sonifi-

cation technique.2 This may first appear unfamiliar in light

of the common parlance to see earcons/auditory icons as

different from sonification. However, imagine an auditory

display for biomedical data that uses auditory icons as sonic

events to represent different classes (e.g. auditory icons for

benign/malignant tissue). The sonification would then be

the superposition or mixture of all the auditory icons chosen

for instance according to the class label and organized prop-

erly on the time axis. If we sonify a data set consisting only

of a single data item we naturally obtain as an extreme case

a single auditory icon. The same can be said for earcons.

Although sonification originally has the connotation of rep-

resenting large and complex data sets, it makes sense for the

definition to also work for single data points.

Data vs. Information: A distinction between data and

information is – as far as the above definition – irrelevant.

Think of earcons to represent computer desktop interactions

such as “delete file”, “rename folder”. There can be a lexi-

2they are also covered by the definition of sonification as ‘non-speech

use of sound to convey information’!

ICAD08-2

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

con of terms (file, folder, link) and actions (delete, rename,

etc.), and in practical computer implementations these fea-

tures would be represented numerically, e.g. object = O1,

action = A3. By doing so, the information has been turned

into data, and this is generally done if there is more than

one signal type to give. Information like for instance a

verbal message can always be represented numerically and

thus be understood as data. On the other side, raw data

values often carry semantic interpretation: e.g. the outside

temperature data value -10◦C (a one-dimensional data set

of size 1) – this is cold, and clearly information! Assum-

ing that information is always encoded as data values for

its processing we can deal with both in a single definition.

How the data are then represented by using sound is another

question: whether sonification techniques use a more sym-

bolic or analogic representation according to the analogic-

symbolic continuum of Kramer [6] is secondary for the def-

inition.

Mapping as a specific case of sonification: Some

articles have used “sonification” to refer specifically

to mapping-based sonification, where data features are

mapped to acoustic features of sound events or streams. Yet

sonification is more generally the representation of data by

using sound. There may be times when a clear specifica-

tion of the sonification technique, e.g. as model-based, au-

dification or parameter-mapping sonification, may be help-

ful to avoid confusion with the general term of sonification.

It makes sense to always use the most specific term possi-

ble, that is to use the term Parameter Mapping Sonification,

Audification, Model-Based Sonification, etc. to convey ex-

actly what is meant. The term Sonification, however, is,

according to the definition, more general which is also sup-

ported by many online definitions3. In result we suggest

using sonification with the same level of generality as the

term visualization is used in visual display.

Sonification as algorithm and sound: Sonification

refers to the technique and the process, so basically it refers

to the algorithm that is at work between the data, the user

and the resulting sound. Often, and with equal right, the re-

sulting sounds are called sonifications. Algorithm means a

set of clear rules, independent of whether it is implemented

on a computer or any other way.

Sonification as scientific method: According to the

definition, sonification is an accurate scientific method

which leads to reproducible results, addressing the ear

rather then the eye (as visualization does). This does not

limit the use of sonifications to data from the sciences, but

only states that sonification can be used as a valid instru-

ment to gain insight. The subjectivity in human percep-

3http://en.wikipedia.org/wiki/Sonification,

http://wvvel.csee.wvu.edu/sepscor/sonification/lesson9.html,

http://www.techfak.uni-bielefeld.de/ags/ni/projects/datamining/datason/

datason e.html, http://www.cs.uiowa.edu/ kearney/22c296Fall02/ Critten-

donSpecialty.pdf, to name a few.

tion and interpretation is shared with other perceptualization

techniques that bridge the gap between data and the human

sensory system. Being a scientific method, a prefix like in

“scientific sonification” is not necessary.

Same as some data visualizations may be ‘viewed’ as

art, sonifications may be heard as ‘music’[5], yet this use

differs from the original intent.

2.1.2. Comments to (C1)

(C1) The sound reflects objective properties or

relations in the input data.

Real-world acoustics are typically not a sonification al-

though they often deliver object-property-specific system-

atic sound, since there is no external input data as requested

in C1. For instance, with a bursting bottle, one can identify

what is the data, the model and the sound, but the process

cannot be repeated with the same bottle. However, using

a bottle that fills with rain, hitting it with a spoon once a

minute can be seen as a sonification: The data here is the

amount of rainfall, which is here measured by the fill level,

and the other conditions are also fulfilled. Tuning a guitar

string might also be regarded as a sonification to adjust the

tension of a string4. These examples show that sonifications

are not limited to computer-implementations according to

the definition, which embraces the possibility of other non-

computer-implemented sonifications.

The borders of sonification and real-world acoustics are

fuzzy. It might be discussed how helpful it is to regard or

denote everyday sounds as sonifications.

2.1.3. Comments to (C2)

(C2) The transformation is systematic. This

means that there is a precise definition pro-

vided of how the data (and optional interac-

tions) cause the sound to change.

What exactly do we mean by “precise”? Some sound

generators use noise and thereby random elements so that

sound events will per se sound different on each rendering.

In Parameter-Mapping Sonifications, the intentional addi-

tion of noise (for instance as onset jitter to increase per-

ceptability of events that would otherwise coincide) is often

used and makes sense. In order to include such cases ran-

domness is allowed in the definition, yet it is important to

declare where and what random elements are used (e.g. by

describing the noise distribution). It is also helpful to give

a motivation for the use of such random elements. By us-

ing too much noise, it is possible to generate useless soni-

fications in the sense that they garble interpretation of the

underlying data. In the same way it is possible to create

useless scientific visualizations.

4thanks to the referee for this example!

ICAD08-3

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

2.1.4. Comments to (C3)

(C3) The sonification is reproducible: given the

same data and identical interactions (or trig-

gers) the resulting sound has to be structurally

identical.

The definition claims reproducibility. This may not

strictly be achieved for several reasons: the loudspeakers

may generate a different sound at different temperatures,

other factors such as introduced noise as discussed above

may have been added. The use of the term “structurally

identical” in the definition aims to weaken the stronger

claim of sample-based identity. Sample-based identity is

not necessary, yet all possible psychophysical tests should

come to identical conclusions.

2.1.5. Comments to (C4)

(C4) The system can intentionally be used with

different data, and also be used in repetition

with the same data.

Repeatability is essential for a technique to be scientif-

ically valid and useful – otherwise nobody could check the

results obtained by using sonification as instrument to gain

insight. However, there are some implications by claim-

ing repeatability for what can and cannot be called sonifi-

cation. It has for instance been suggested that a musician

improvising on his instrument produces ‘a sonification of

the musician’s emotional state’. With C4, however, “play-

ing a musical instrument” is not a sonification of the per-

former’s emotional state, since it can not be repeated with

the ‘identical’ data. However, the resulting sound may be

called a sonification of the interactions with the instrument

(regarded here as data), and in fact, music can be heard with

the focus to understand the systematic interaction patterns

with the instruments.

Some of these conditions have been set as constraints

for sonification, e.g. reproducibility in the ‘Listening to the

Mind Listening’ concert5, but not been connected to a defi-

nition of sonification.

In summary, the given definition provides a set of neces-

sary conditions for systems and methods to be called soni-

fication. The definition is neither exhaustive nor complete;

we hope it will serve as the core definition as we as commu-

nity work towards a complete one.

3. SONIFICATION AND AUDITORY DISPLAY

With the above definition, the term sonification takes the

role of a general term to express the method of rendering

5http://www.icad.org/websiteV2.0/Conferences/ICAD2004/concert call.htm

sound in an organized and well-structured way. This is in

good analogy with the term visualization which is also the

general term under which a variety of specific techniques

such as bar charts, scatter plots, graphs, etc. are subsumed.

Particularly there is an analogy between scatter plots where

graphical symbols (data-mapped color/size...) are orga-

nized in space to deliver the visualization, and Parameter-

Mapping Sonification, where in a structurally identical way

acoustic events (with data-mapped features) are organized

in time. It is helpful to have with sonification a term that

operates on the same level of generality as visualization.

This raises the question what then do we mean by au-

ditory displays? Interestingly, in the visual realm, the

term ‘display’ suggests a necessary but complementary part

of the interface chain: the device to generate structured

light/images, for instance a CRT or LCD display or a projec-

tor. So in visualization, the term visualization emphasizes

the way how data are rendered as an image while the display

is necessary for a user to actually see the information. For

auditory display, we suggest to include this aspect of con-

version of sound signals into audible sound, so that an au-

ditory display encompasses also the technical system used

to create sound waves, or more general: all possible trans-

missions which finally lead to audible perceptions for the

user. This could range from loudspeakers over headphones

to bone conduction devices. We suggest furthermore that

auditory display should also include the user context (user,

task, background sound, constraints) and the application

context, since these are all quite essential for the design and

implementation. Sonification is thereby an integral compo-

nent within an auditory display system which addresses the

actual rendering of sound signals which in turn depend on

the data and optional interactions, as illustrated in Fig. 2.

Auditory Displays are more comprehensive than sonifica-Components of Auditory Display Systems

User/Listener

Technical

Sound Display

Sonification

(Rendering)

0101

0100

Application

Context

Data

Usage Context

mobile?

PC?

office?

Interactions

Figure 2: Auditory Displays: systems that employ sonifica-

tion for structuring sound and furthermore include the trans-

mission chain leading to audible perceptions and the appli-

cation context.

ICAD08-4

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

tion since for instance dialogue systems and speech inter-

faces may also be regarded as auditory displays since they

use sound for communication. While such interfaces are not

the primary focus in this research field the terminology sug-

gests their inclusion. On the other hand, Auditory Display

may be seen as a subset of the more general term of Audi-

tory Interfaces which do not only include output interfaces

(auditory displays, sonification) but also auditory input in-

terfaces which engender bidirectional auditory control and

communication between a user and a (in most cases) tech-

nical system (e.g. voice control system, query-by humming

systems, etc.).

4. HIERARCHY FROM SOUND TO

SONIFICATION

So far we have dealt with the necessary conditions sur-

rounding sonification and thus narrowed sonification down

to a specific subset of using sound. In this section, we look

at sonification in a systemic manner to elucidate its super-

ordinate categories. Figure 3 shows how we suggest to or-

ganize the different classes of sound. On the highest level,Map of Sound

Organized Sound

Functional Sounds

Music &

Media Arts Sonification(a)

(b)

Figure 3: Systemic map of sound, showing sonification and

its relation to other categories.

sounds are here classified as Organized Sound and unorga-

nized sound. Organized sounds separate from random or

otherwise complex structured sounds in the fact that their

occurence and structure is shaped by intention. Environ-

mental sounds appear often to be very structured and could

thus also be organized sounds, however, if so, any sound

would match that category to some extent. It thus may be

useful to apply the term to sounds that are intentionally or-

ganized – in most cases by the sound/interface developer.

The set of organized sound comprises two large sets that

partially overlap: music and functional sounds. Music is

without question a complex structured signal, organized on

various levels, from the acoustic signal to its temporal orga-

nization in bars, motifs, parts, layers. It is not our purpose

to give a definition of music.

The second set is functional sounds. These are orga-

nized sounds that serve a certain function or goal [7]. The

function is the motivation for their creation and use. To give

an example, all signal sounds (such as telephones, door-

bells, horns and warning hooters) are functional sounds.

Certainly there are intersections with music, as music can

serve functional aspects. For instance, trombones and kettle

drums have been used to demonstrate kingship and power.

A more subtle function is the use of music in supermarkets

to enhance the ‘shopping mood’. For that reason these sets

overlap – the size of the overlap depends on what is regarded

as function.

Sonification in the sense of the above definition is cer-

tainly a subset of functional sounds. The sounds are ren-

dered to fulfill a certain function, be it communication of in-

formation (signals & alarms), the monitoring of processes,

or to support better understanding of structure in data under

analysis. So is there a difference between functional sounds

and sonification at all? The following example makes clear

that sonification is really a subset: Recently a new selec-

tive acoustic weapon has been used, the mosquito device6,

a loudspeaker that produces a HF-sound inaudible to older

people, which drives away teenagers hanging around in

front of shops. This sound is surely functional, yet it could

neither pass as sonification nor as music.

Finally, we discuss whether sonification has an intersec-

tion with music&media arts. Obviously there are many ex-

amples where data are used to drive aspects of musical per-

formances, e.g. data collected from motion tracking or bio-

sensors attached to a performer. This is, concerning the in-

volved techniques and implementations similar to mapping

sonifications. However, a closer look at our proposed defi-

nition shows that often the condition for the transformation

to be systematic C2 is violated and the exact rules are not

made explicit. But without making the relationship explicit,

the listener cannot use the sound to understand the underly-

ing data better. In addition, condition C4 may often be vio-

lated. If sonification-like techniques are employed to obtain

a specific musical or acoustic effect without transparency

between the used data and details of the sonification tech-

niques, it might, for the sake of clarity, better be denoted

as ‘data-inspired music’, or ‘data-controlled music’ than as

sonification. Iannis Xenakis, for instance, did not even want

the listener to be aware of the data source nor the rules of

sound generation.

6see http://www.compoundsecurity.co.uk/, last seen 2008-01-16

ICAD08-5

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

5. CLOSED INTERACTION LOOPS

IN AUDITORY DISPLAYS

This section emphasizes the role of interaction in sonifica-

tion. We propose different terms depending on the scope of

the closure of the interaction loop. The motivation for this

discussion is that it might be helpful to address how terms

such as biofeedback or interactive sonification relate to each

other.

We start the discussion with Fig. 4 that depicts closed

loop interactions. The sonification module in the upper cen-

ter playing rendered sonifications to the user. Data sources

for sonification enter the box on the left side and the most

important parts are (a) World/System: this comprises any

system in the world that is connected to the sonification

module, e.g. via sensors that measure its state, and (b) Data:

these are any data under analysis or represented information

to be displayed that are stored separately and accessible by

the sonification.World/System

Sonification

Interactive Sonification

Human Activity (supported by sonification)

Auditory Biofeedback

Data

Navigation

Monitoring

No Action

Figure 4: Illustration of Closed-Loop Auditory Systems.

In this setting, Process Monitoring is the least inter-

active sonification, where data recorded from the world (in

real-time) or read from the data repository is continuously

used as input for a sonification rendering process. Here, the

listener is merely passively listening to the sound with the

only active component being his/her focus of attention onto

parts of the sound. Certainly, certain changes in the sound

might attract attention and force the user to act (e.g. sell

stocks, stop a machine, etc...).

A higher degree of active involvement occurs when the

user actively changes and adjusts parameters of the sonifi-

cation module, or interacts otherwise with the sonification

system. We denote this case as Interactive Sonification.

There is a wide field of possibilities of why and how to do

so, and we discuss 3 different prototypical examples:

(a) Triggering: Consider a mapping sonification of a

given data set. An essential interaction for the user

is to issue the command to render/playback the soni-

fication for a selected dataset. Possibly he/she does

this several times in order to attend different parts of

the sound signal. This elementary case is an interac-

tion, however, a very basic one.

(b) Parameter Adjustment is done when the user changes

parameters, such as what data feature are mapped

to acoustic parameters, control ranges, compression

factors, etc. Often such adjustments happen sepa-

rate from the playback so that the changes are made

and afterwards the updated sound is rendered. How-

ever, interactive real-time control is feasible in many

cases and shows a higher degree of interactivity. The

user actively explores the data by generating different

‘views’ of the data [8]. In visualization a similar in-

teractivity is obtained by allowing the user to select

axes scalings, etc.

(c) Excitatory Interaction is the third sort of interaction

and is structurally similar to the case of triggering.

Particularly in Model-Based Sonification [4], usually

the data are used to configure a sound-capable vir-

tual object that in turn reacts on excitatory interac-

tions with acoustic responses whereby the user can

explore the data interactively. Excitation puts energy

into the dynamic system and thus initiates an audible

dynamical system behavior. Beyond a simple trigger-

ing, excitatory interactions can be designed to make

use of the fine-grained manipulation skills that human

hands allow, e.g. by enabling to shake, squeeze, tilt or

deform the virtual object, for instance using sensor-

equipped physical interfaces to interact with the soni-

fication model. A good example for MBS is Shoogle

by Williamson et al. [9], where short text messages

in a mobile phone can be overviewed by shaking a

mobile phone equipped with accelerometer sensors,

resulting in audible responses of the text messages

as objects moving virtually inside the phone. Excita-

tory interactions offer rich and complex interactions

for interactive sonification.

The next possibility for a closed loop is by interactions

that select or browse data. Since data are chosen, it may

best be referred to as Navigation. Navigation can also be

regarded as special case of Interactive Sonification, depend-

ing on where the data are selected and the borders are here

really soft. Navigation usually goes hand in hand with trig-

gering of sonification (explained above).

Auditory Biofeedback can be interpreted as a sonifi-

cation of measured sensor data. In contrast to the above

types, the user’s activity is not controlling an otherwise au-

tonomous sonification with independent data, but it pro-

duces the input data for the sonification system. The user

perceives a sound that depends on his/her own activity.

Such systems have applications that range from rehabilita-

tion training to movement training in sports, e.g. to perform

ICAD08-6

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

a complex motion sequence (e.g. a tennis serve) so that its

sonification is structurally more similar to the sonification

of an expert performing the action [10].

The final category is Human Activity, which means

that the interaction ranges beyond the sonification system

into the world, often driven by the goal to change a world

state in a specific way. In turn, any sensors that pick up the

change may lead to changes in the sonification. The differ-

ence between the loop types before is that the primary fo-

cus is to achieve a goal beyond the sonification system, and

not to interact with a closed-loop sonification system. Even

without attending the sonification consciously or primarily,

the sound can be helpful to reach the goal. For example,

imagine the real-world task to fill a thermos bottle with tea.

While your primary goal is to get the bottle filled you will

receive the ‘gluck-gluck’ sound with increasing pitch as a

by-product of the interaction. If this is consistently useful,

you subconsciously adapt your activity to exploit the cues in

the sound – but the sound is only periphery for the goal. In a

similar sense, sonifications may deliver helpful by-products

to actions that change the world state. We regard such in-

teraction add-ons where sonification is a non-obtrusive yet

helpful cue for goal attainment as inspiring design direc-

tion. Such sonifications might even become subliminal in

the sense that users, when asked about the sound, are not

even aware of the sound, yet they perform better with sound

than without.

6. DISCUSSION AND CONCLUSION

The definitions in this paper are given on the basis of

three goals: (i) to anchor sonification as a precise scien-

tific method so that it delivers reproducible results and thus

can be used and trusted as instrument to obtain insight into

data under analysis. (ii) to offer a generalization which does

not limit itself to the special case of mappings from data to

sound, but which introduces sonification as general system-

atic mediator between data and sound, whatever the repre-

sentation might be. (iii) to balance the definition so that the

often-seen pair of terms ‘visualization & sonification’ are at

the same level of generality.

The definition has several implications which have been

discussed in Sec. 2. We’d like to emphasize that this effort

is being done in hope that the definition inspires a general

discussion on the terminology and taxonomy of the research

field of auditory display. An online version of the definition

is provided at www.sonification.de with the aim to collect

comments and examples of sonifications as well as exam-

ples that are agreed not to be sonifications and which help

in turn to improve the definition.

In Section 3, we described integral parts for auditory

display so that sonification takes a key component as the

technical part involving the rendition of sound. Again, the

suggested modules are meant as working hypothesis to be

discussed at ICAD.

While the given definitions specified terms on a horizon-

tal level, Section 4 proposes a vertical organization of sound

in relation to often used terms. The intersections between

the different terms and categories have been addressed with

examples.

Finally, we have presented in Section 5 an integrative

scheme for organizing different classes of auditory closed

loops according to the loop closure scope. It proves help-

ful to clarify classes of interactive sonifications. We think

that grouping existing sonifications according to these cat-

egories can be helpful to better find alternative approaches

for a given task.

The suggested terminology and taxonomy is the result

of many discussions and a thorough search for helpful con-

cepts. We suggest it as working definitions to be discussed

at the interdisciplinary level of ICAD in hope to contribute

towards a maturing of the fields of auditory display and

sonification.

7. ACKNOWLEDGEMENT

Many collegues have been very helpful in discussions to

refine the definitions. Particularly, I thank Till Bover-

mann, Arne Wulf, Andy Hunt, Florian Grond, Georg

Spehr, Alberto de Campo, Gerold Baier, Camille Peres,

and in particular Gregory Kramer for the helpful discus-

sions on the definition for sonification. Thanks also to col-

legues of the COST IC0601 Sonic Interaction Design (SID)

WG4/Sonification. I also thank Arne Wulf for the inspiring

discussions on Closed-Loop Auditory Systems, and Louise

Nickerson for many language improvements.

8. REFERENCES

[1] G. Kramer, Ed., Auditory Display - Sonification, Au-

dification, and Auditory Interfaces. Addison-Wesley,

1994.

[2] G. Kramer, B. Walker, T. Bonebright, P. Cook,

J. Flowers, N. Miner, and J. Neuhoff, “Sonification

report: Status of the field and research agenda,” Tech.

Rep., International Community for Auditory Display,

1999, http://www.icad.org/websiteV2.

0/References/nsf.html.

[3] Thomas Hermann and Helge Ritter, “Listen to your

data: Model-based sonification for data analysis,” in

Advances in intelligent computing and multimedia sys-

tems, G. E. Lasker, Ed., Baden-Baden, Germany, 08

1999, pp. 189–194, Int. Inst. for Advanced Studies in

System research and cybernetics.

ICAD08-7

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

[4] Thomas Hermann, Sonification for Exploratory Data

Analysis, Ph.D. thesis, Bielefeld University, Bielefeld,

Germany, 02 2002.

[5] Paul Vickers and Bennett Hogg, “Sonification ab-

straite/sonification concr`ete: An ‘æsthetic perspective

space’ for classifying auditory displays in the ars mu-

sica domain,” in ICAD 2006 - The 12th Meeting of the

International Conference on Auditory Display, Alis-

tair D N Edwards and Tony Stockman, Eds., London,

UK, June 20-23 2006, pp. 210–216.

[6] G. Kramer, “An introduction to auditory display,” in

Auditory Display, G. Kramer, Ed. ICAD, 1994, pp. 1–

79, Addison-Wesley.

[7] Georg Spehr, SOUND STUDIES. Traditionen - Meth-

oden – Desiderate, chapter Funktionale Kl ̈ange - Mehr

als ein Ping, transcript Verlag, Bielefeld, Germany,

2008.

[8] Thomas Hermann and Andy Hunt, “The discipline of

interactive sonification,” in Proceedings of the Inter-

national Workshop on Interactive Sonification (ISon

2004), Thomas Hermann and Andy Hunt, Eds., Biele-

feld, Germany, 01 2004, Bielefeld University, Interac-

tive Sonification Community, peer-reviewed article.

[9] John Williamson, Rod Murray-Smith, and S. Hughes,

“Shoogle: excitatory multimodal interaction on mo-

bile devices,” in Proceedings of the SIGCHI confer-

ence on Human factors in computing systems, San

Jose, California, USA, 2007, pp. 121–124, ACM

Press.

[10] Thomas Hermann, Oliver H ̈oner, and Helge Rit-

ter, “Acoumotion - an interactive sonification sys-

tem for acoustic motion control,” in Gesture in

Human-Computer Interaction and Simulation: 6th In-

ternational Gesture Workshop, GW 2005, Berder Is-

land, France, May 18-20, 2005, Revised Selected Pa-

pers, Sylvie Gibet, Nicolas Courty, and Jean-Francois

Kamp, Eds., Berlin, Heidelberg, 2006, vol. 3881/2006

of Lecture Notes in Computer Science, pp. 312–323,

Springer.

ICAD08-8

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/221513907

Nomadic Radio: Scaleable and Contextual Notification for Wearable Audio

Messaging.

Conference Paper · January 1999

DOI: 10.1145/302979.303005 · Source: DBLP

CITATIONS

95

READS

153

2 authors, including:

Some of the authors of this publication are also working on these related projects:

Bayesian Modeling for Human Development Indicators View project

Aware Community Portals View project

Nitin Sawhney

Aalto University

55 PUBLICATIONS 1,560 CITATIONS

SEE PROFILE

All content following this page was uploaded by Nitin Sawhney on 30 March 2015.

The user has requested enhancement of the downloaded file.

Papers CHI 99 15-20 MAY 1999

Nomadic Radio: Scaleable and Contextual Notification

for Wearable Audio Messaging

Nitin Sawhney and Chris Schmandt

SpeechInterface Group, MIT Media Laboratory

20 Ames St., Cambridge, MA 02139

{ nitin, geek} @media.mit.edu

ABSTRACT

Mobile workers need seamless access to communication

and information services on portable devices. However

current solutions overwhelm users with intrusive and

ambiguous notifications. In this paper, we describe

scaleable auditory techniques and a contextual notification

model for providing timely information, while minimizing

interruptions. User’s actions influence local adaptation in

the model. These techniques are demonstrated in Nomadic

Radio, an audio-only wearable computing platform.

Keywords

Auditory I/O, passive awareness, wearable computing,

adaptive interfaces, interruptions, notifications

INTRODUCTION

In today’s information-rich environments, people use a

number of appliances and portable devices for a variety of

tasks in the home, workplace and on the run. Such devices

are ubiquitous and each plays a unique functional role in a

user’s lifestyle. To be effective, these devices need to notify

users of changes in their functional state, incoming

messages or exceptional conditions. In a typical office

environment, the user attends to a plethora of devices with

notifications such as calls on telephones, asynchronous

messages on pagers, email notification on desktop

computers, and reminders on personal organizers or

watches. This scenario poses a number of key problems.

Lack of Differentiation in Notification Cues

Every device provides some unique form of notification. In

many cases, these are distinct auditory cues. Yet, most cues

are generally binary in nature, i.e. they convey only the

occurrence of a notification and not its urgency or dynamic

state. This prevents users from making timely decisions

about received messages without having to shift focus of

attention (from the primary task) to interact with the device

and access the relevant information.

Permission to make digital or hard copies of all or part of this work for

personal or classroom use is granted without fee provided that copies

arc not made or distributed for profit or commercial advantage and that

copies bear this notice and the full citation on the first page. To copy

otherwise, to republish, to post on servers or to redistribute to lists.

requires prior specific permission and/or a fee.

CHI ‘99 Pittsburgh PA USA

Copyright ACM 1999 0-201-48559-1/99/05...$5.00

Minimal Awareness of the User and Environment

Such notifications occur without any regard to the user’s

engagement in her current activity or her focus of attention.

This interrupts a conversation or causes an annoying

disruption in the user’s task and flow of thoughts. To

prevent undue embarrassment in social environments, users

typically turn off cell-phones and pagers in meetings or

lectures. This prevents the user from getting notification of

timely messages and frustrates people trying to get in touch

with her.

No Learning from Prior Interactions with User

Such systems typically have no mechanism to adapt their

behavior based on the positive or negative actions of the

user. Pagers continue to buzz and cell-phones do not stop

ringing despite the fact that the user may be in a

conversation and ignoring the device for some time.

Lack of Coordinated Notifications

All devices compete for a user’s undivided attention without

any coordination and synchronization of their notifications.

If two or more notifications occur within a short time of

each other, the user gets confused or frustrated. As people

start carrying around many such portable devices, frequent

and uncoordinated interruptions inhibit their daily tasks and

interactions in social environments.

Given these problems, most devices fail to serve their

intended purpose of notification or communication, and

thus do not operate in an efficient manner for a majority of

their life cycle. New users choose not to adopt such

technologies, having observed the obvious problems

encountered with their usage. In addition, current users tend

to turn off the devices in many situations, inhibiting the

optimal operation of such personal devices.

Nature of Interruptions in the Workplace

A recent observational study [4] evaluated the effect of

interruptions on the activity of mobile professionals in their

workplace. An interruption, defined as an asynchronous and

unscheduled interaction, not initiated by the user, results in

the recipient discontinuing the current activity. The results

revealed several key issues. On average, sub.jects were

interrupted over 4 times per hour, for an average duration

slightly over 2 minutes. Hence, nearly 10 minutes per hour

96

C H I 9 9 1 5 - 2 0 M A Y 1 9 9 9 Papers

was spent on interruptions. Although a majority of the

interruptions occurred in a face-to-face setting, 20% were

due to telephone calls (no email or pager activity was

analyzed in this study). In 64% of the interruptions, the

recipient received some benefit from the interaction. This

suggests that a blanket approach to prevent interruptions,

such as holding all calls at certain times of the day, would

prevent beneficial interactions from occurring. However in

41% of the interruptions, the recipients did not resume the

work they were doing prior to it. But active use of new

communication technologies makes users easily vulnerable

to undesirable interruptions.

These interruptions constitute a significant problem for

mobile professionals using tools such as pagers, cell-phones

and PDAs, by disrupting their time-critical activities.

Improved synchronous access using these tools benefits

initiators but leaves recipients with little control over the

interactions. The study suggests development of improved

filtering techniques that are especially light-weight, i.e.

don’t require more attention from the user and are less

disruptive than the interruption itself. By moving

interruptions to asynchronous media, messages can be

stored for retrieval and delivery at more appropriate times.

NOMADIC RADIO: WEARABLE AUDIO MESSAGING

Personal messaging and communication, demonstrated in

Nomadic Radio, provides a simple and constrained problem

domain in which to develop and evaluate a contextual

notification model. Messaging requires development of a

model that dynamically selects a suitable notification

strategy based on message priority, usage level, and

environmental context. Such a system must infer the user’s

attention by monitoring her current activities such as

interactions with the device and conversations in the room.

The user’s prior responses to notifications must also be

taken into consideration to adapt the notifications over time.

In this paper, we will consider techniques for scaleable

auditory presentation and an appropriate parameterized

approach towards contextual notification.

Several recent projects utilized speech and audio I/O on

wearable devices to present information. A prototype

augmented audio tour guide [l] played digital audio

recordings indexed by the spatial location of visitors in a

museum. SpeechWear [11] enabled users to perform data

entry and retrieval using speech recognition and synthesis.

Audio Aura [10] explored the use of background auditory

cues to provide serendipitous information coupled with

people’s physical location in the workplace. In Nomadic

Radio, the user’s inferred context rather than actual location

is used to decide when and how to deliver scaleable audio

notifications. In a recent paper [13], researchers suggest the

use of sensors and user modeling to allow wearables to

infer when users should be interrupted by incoming

messages. They suggest waiting for a break in the

conversation to post a message summary on the user’s

heads-up display. In this paper we describe a primarily non-

visual approach to provide timely information to nomadic

listeners, based on a variety of contextual cues.

Nomadic Radio is a wearable computing platform that

provides a unified audio-only interface to remote services

and messages such as email, voice mail, hourly news

broadcasts, and personal calendar events. These messages

are automatically downloaded to the device throughout the

day and users can browse through them using voice

commands and tactile input. The system consists of Java-

based clients and remote servers (written in C and Perl) that

communicate over wireless LAN, and utilize the telephony

infrastructure in the Speech Interface group. Simultaneous

spatial audio streams are rendered using a HRTF-based

Java audio API. Speech I/O is provided via a networked

implementation of AT&T Watson Speech API.

To provide a hands-free and unobtrusive interface to a

nomadic user, the system primarily operates as a wearable

audio-only device. The SoundBeam Neckset, a research

prototype patented by Nortel for use in hands-free

telephony, was adapted as the primary wearable platform in

Nomadic Radio. It consists of two directional speakers

mounted on the user’s shoulders, and a directional

microphone placed on the chest (see figure 1). Here

information and feedback is provided to the user through a

combination of auditory cues, spatial audio rendering, and

synthetic speech. Integration of a variety of auditory

techniques on a wearable device provides hands-free access

and navigation as well as lightweight and expressive

notification.

An audio-only interface has been incorporated in Nomadic

Radio, and a networked infrastructure for unified messaging

has been developed for wearable access [12]. The system

currently operates on a Libretto 100 mini-portable PC worn

by the user. The key issue addressed in this paper is that of

handling interruptions to the listener in a manner that

reduces disruption, while providing timely notifications for

contextually relevant messages.

P a p e r s

USAGE AND NOTIFICATION SCENARIO

The following scenario demonstrates the audio interface

and presentation of notifications in Nomadic Radio (no

voice commands from the user are shown here).

CHI 99 15-20 MAY 1999

SCALEABLE AUDITORY PRESENTATION

A scaleable presentation is necessary for delivering

sufficient information while minimizing interruption to the

listener. Messages in Nomadic Radio are scaled

dynamically to unfold as seven increasing levels of

notification (see figure 3): silence, ambient cues, auditory

cues, message summary, preview, full body, and foreground

rendering. These are described further below:

Silence for Least Interruption and Conservation

In this mode all auditory cues and speech feedback are

turned-off. Messages can be scaled down to silence when

the message priority is inferred to be too low for the

message to be relevant for playback or awareness to a user,

based on her recent usage of the device and the

conversation level. This mode also serves to conserve

processing, power and memory resources on a portable

device or wearable computer.

Ambient Cues for Peripheral Awareness

In Nomadic Radio, ambient auditory cues are continuously

played in the background to provide an awareness of the

operational state of the system and ongoing status of

messages being downloaded (see figure 4). The sound of

flowing water provides an unobtrusive form of ambient

awareness that indicates the system is active (silence

indicates sleep mode). Such a sound tends to fade into the

perceptual background after a short time, so it does not

distract the listener. The pitch is increased during file

downloads, momentarily foregrounding the ambient sound.

A short e-mail message sounds like a splash while a two-

minute audio news summary is heard as faster flowing

water while being downloaded. This implicitly indicates

message size without the need for additional audio cues and

prepares the listener to hear (or deactivate) the message

before it becomes available. Such peripheral awareness

minimizes cognitive overhead of monitoring incoming

messages relative to notifications played as distinct auditory

cues, which incur a somewhat higher cost of attention on

part of the listener.

Related Work in Auditory Awareness

In ARKola [5], an audio/visual simulation of a bottling

factory, repetitive streams of sounds allowed people to keep

track of activity, rate, and functioning of running machines.

Without sounds people often overlooked problems; with

auditory cues, problems were indicated by the machine’s

sound ceasing (often ineffective) or via distinct alert

sounds. The various auditory cues (as many as 12 sounds

play simultaneously) merged as an auditory texture, allowed

people to hear the plant as a complex integrated process.

Background sounds were also explored in ShareMon [3], a

prototype application that notified users of file sharing

activity. Cohen found that pink noise used to indicate

%CPU time was considered “obnoxious”, even though

users understood the, pitch correlation. However,

preliminary reactions to wave sounds were considered

positive and even soothing. In Audio Aura [IO], alarm

CHI 9 9 1 5 - 2 0 M A Y 1 9 9 9 P a p e r s

sounds were eliminated and a number of “harmonically

coherent sonic ecologies” were explored, mapping events to

auditory, musical or voice-based feedback. Such techniques

were used to passively convey the number of email

messages received, identity of senders, and abstract

representations of group activity.

Auditory Cues for Notification and Identification

In Nomadic Radio, auditory cues are a crucial means for

conveying awareness, notification and providing necessary

assurances in its non-visual interface. Different types of

auditory techniques provide distinct feedback, awareness

and message information.

Feedback Cues

Several types of audio cues indicate feedback for a number

of operational events in Nomadic Radio:

1. Task completion and confirmations - button pressed,

speech understood, connected to servers, finished

playing or loaded/deleted messages.

2. Mode transitions - switching categories, going to

non-speech or ambient mode.

3. Exceptional conditions - message not found, lost

connection with servers, and errors.

Priority Cues for Notification

In a related project, “email glances” [7] were formulated as

a stream of short sounds indicating category, sender and

content flags (from keywords in the message). In Nomadic

Radio, message priority inferred from email content

filtering provides distinct auditory cues (assigned by the

user) for group, personal, timely, and important messages.

In addition, auditory cues such as telephone ringing indicate

voice mail, whereas an extracted sound of a station

identifier indicates a news summary.

VoiceCues for Identification

VoiceCues represent a novel approach for easy

identification of the sender of an email, based on a unique

auditory signature of the person. VoiceCues are created by

manually extracting a l-2 second audio sample from the

voice messages of callers and associating them with their

respective email login. When a new email message arrives,

the system queries its database for a related VoiceCue for

that person before playing it to the user as a notification,

along with the priority cues. The authors have found

VoiceCues to be a remarkably effective method for quickly

conveying the sender of the message in a very short

duration. This technique reduces the need for synthetic

speech feedback, which can often be distracting.

99

Papers CHI 99 15-20 MAY 1999

Message Summary Generation

A spoken description of an incoming message can present

relevant information in a concise manner. Such a

description typically utilizes header information in email

messages to convey the name of the sender and the subject

of the message. In Nomadic Radio, message summaries are

generated for all messages, including voice-mail, news and

calendar events. The summaries are augmented by

additional attributes of the message indicating category,

order, priority, and duration. For audio sources, like voice

messages and news broadcasts, the system plays the first

2.5 seconds of the audio. This identifies the caller and the

urgency of the call, inferred from intonation in the caller’s

voice or provides a station identifier for news summaries.

Message Previews using Content Summarization

Messages are scaled to allow listeners to quickly preview

the contents of an email or voice message. In Nomadic

Radio, a preview for text messages extracts the first 100

characters of the message (a default size that can be user

defined). This heuristic generally provides sufficient

context for the listener to anticipate the overall message

theme and urgency. For email messages, redundant headers

and previous replies are eliminated from the preview for

effective extraction. Use of text summarization techniques,

based on tools such as ProSum’ developed by British

Telecom, would allow more flexible means of scaling

message content. Natural language parsing techniques used

in ProSum permit a scaleable summary of an arbitrarily

large text document.

A preview for an audio source such as a voice message or

news broadcast presents a fifth of the message at a

gradually increasing playback rate of up to 1.3 times faster

than normal. There are a range of techniques for time-

compressing speech without modifying the pitch, however

twice the playback rate usually makes the audio

incomprehensible. A better representation for content

summarization requires a structural description of the audio,

based on annotated or automatically determined pauses in

speech, speaker and topic changes. Such an auditory

thumbnail must function similar to its visual counterpart. A

preview for a structured voice message would provide

pertinent aspects such as name of caller and phone number,

whereas a structured news preview would be heard as the

hourly headlines.

Full Body: Playing Complete Message Content

This mode plays the entire audio file or reads the full text of

the message at the original playback rate. Some parsing of

the text is necessary to eliminate redundant header

information and format tags. The message is augmented

with summary information indicating sender and subject.

This message is generally spoken or played in the

background of the listener’s audio space.

I http://transend.Iabs.bt.com/prosum/on-line/

Foreground Rendering via Spatial Proximity

An important message is played in the foreground of the

listening space. The audio source of the message is rapidly

moved closer to the listener, allowing it to be heard louder,

and played there for 415” of its duration. The message

gradually begins to fade away, moving back to its original

position and amplitude for the remaining l/S” of the

duration. The foregrounding algorithm ensures that the

messages are quickly brought into perceptual focus by

pulling them to the listener rapidly. However the messages

are pushed back slowly to provide an easy fading effect as

the next one is heard. As the message moves its spatial

direction is maintained so that the listener can retain a focus

on the audio source even if another begins to play.

Hence a range of techniques provide scaleable forms of

background awareness, auditory notification, spoken

feedback and foreground rendering of incoming messages.

CONTEXTUAL NOTIFICATION

In Nomadic Radio, context dynamically scales the

notifications for incoming messages. The primary

contextual cues used include: message priority from email

filtering, usage level based on time since last user action,

and the likelihood of conversation estimated from real-time

analysis of the auditory scene. In our experience these

parameters provide sufficient context to scale notifications,

however data from motion or location sensors can also be

integrated in such a model. A linear and scaleable auditory

notification model is utilized, based on the notion of

estimating costs of interruption and the value of information

to be delivered to the user. This approach is similar to

recent work [6] on using perceptual costs and a focus of

attention model for scaleable graphics rendering.

Message Priority

The priority of incoming messages is explicitly determined

via content-based email filtering using CLUES [9], a

filtering and prioritization system. CLUES has been

integrated into Nomadic Radio to determine the timely

nature of messages by finding correlation between a user’s

calendar, rolodex, to-do list, as well as a record of outgoing

messages and phone calls. These rules are integrated with

static rules created by the user for prioritizing specific

people or message subjects. When a new email message

arrives, keywords from its sender and. subject header

information are correlated with static and generated

filtering rules to assign a priority to the message. Email

messages are also prioritized if the user is traveling and

meeting others in the same geographic area (via area codes

in the rolodex). The current priorities include: group,

personal, very important, most important, and timely.

Priorities are parameterized by logarithmically scaling all

priorities within a range of 0 to 1. Logarithmic scaling

ensures that higher priority messages are weighted higher

relative to unimportant or uncategorized messages.

Priority ( i ) = ( log ( i ) / log (Priority Levels Mu ) )

100

C H I 9 9 1 5 - 2 0 M A Y 1 9 9 9 Papers

Usage Level

One problem with using last actions for setting usage levels

is that if a user deactivates an annoying message, that

action is again time-stamped. Such negative reinforcements

continue to increase the usage level and the related

notification. Therefore negative actions such as stopping

audio playback or deactivating speech are excluded from

generating actions for computing the usage.

Likelihood of Conversation

Conversation in the environment can be used to gauge

whether the user is in a social context where an

interruption is less appropriate. If the system detects the

occurrence of more than several speakers over a period of

time, that is an indication of a conversational situation.

Auditory events are first detected by adaptively

thresholding total energy and incorporating constraints on

event length and surrounding pauses. The system uses mel-

scaled filter-bank coefficients (MFCs) and pitch estimates

to discriminate, reasonably well, a variety of speech and

non-speech sounds. HMMs (Hidden Markov Models)

capture both the temporal characteristics and spectral

content of sound events. The techniques for feature

extraction and classification of the auditory scene using

HMMs are described in a recent workshop paper [2]. The

likelihood of speech detected in the environment is

computed for each event in a short window of time. In

addition, the probabilities are weighted, such that most

recent time periods in the window are considered more

relevant for computing the overall Speech Level. We are

evaluating the classifier’s effectiveness by training it with a

variety of speakers and background sounds.

Notification Level

A weighted average for all three contextual cues provides

level has an inversely proportional relationship with

notification i.e. a lower notification must be provided

during high conversation.

Presentation Latency

Latency represents the period of time to wait before

playing the message to the listener, after a notification cue

is delivered. Latency is computed as a function of the

notification level and the maximum window of time

(Latency,& that a lowest priority message can be delayed

for playback. The default maximum latency is set to 20

seconds, but can be modified by the user.

101

CHI99 15-20 MAY 1999

were increased. Jane was notified of a group message

shortly after the voice message, since the system detected

higher usage activity. Hence, the system correctly scaled

down notifications when Jane did not want to be bothered

whereas notifications were scaled up when Jane started to

use the system to browse her messages.

EFFECTIVENESS OF THE NOTIFICATION MODEL

The nature of peripheral awareness and unobtrusive

notification on a wearable device requires a usage

evaluation that must be conducted on an ongoing and long-

term basis. However, the predictive effectiveness of the

notification model must first be evaluated on a quantitative

basis. Hence, all message and notification parameters are

captured for such analysis. Lets consider two actual

examples of notification computed for email messages with

different priorities. Figure 7 shows an auditory cue

generated for a group message (low priority).

The timely message (in figure 8) received greater priority

and consequently a higher notification level for summary

playback. A moderate latency time (approx. 6 secs.) was

chosen. However when the user interrupted the notification

by a button press, the summary playback was aborted. The

user’s action reduced overall weights by 5%.

P a p e r s

Dynamic Adaptation of the Notification Model

The user can initially set the weights for the notification

model to high, medium, or low (interruption). These weight

settings were selected by experimenting with notifications

over time using an interactive visualization of message

parameters. This allowed us to observe the model, modify

weights and infer the effect on notification based on

different weighting strategies. Pre-defined weights provide

an approximate behavior for the model and help bootstrap

the system for novice users. The system also allows the user

to dynamically adjust these weights (changing the

interruption and notification levels) by their implicit actions

while playing or ignoring messages.

The system allows localized positive and negative

reinforcement of the weights by monitoring the actions of

the user during notifications. As a message arrives, the

system plays an auditory cue if its computed notification

level is above the necessary threshold for auditory cues. It

then uses the computed latency interval to wait before

playing the appropriate summary or preview of the

message. During that time, the user can request the message

be played earlier or abort any further notification for the

message via speech or button commands. If aborted, all

weights are reduced by a fixed percentage (default is 5%), a

negative reinforcement. If the user activates the message

(positive reinforcement) within 60 seconds after the

notification, the playback scale selected by the user is used

to increase all weights. If the message is ignored, no change

is made to the weights, but the message remains active for

60 seconds during which the user’s actions can continue to

influence the weights.

Figure 6 shows a zoomed view of the extended scenario

introduced earlier, focusing on Jane’s actions that reinforce

the model. Jane received several messages and ignored

most of the group messages and a recent personal message

(the weights remain unchanged). While in the meeting, Jane

interrupted a timely message to abort its playback. This

reduced the weights for future messages, and the ones with

low priority (group message) were not notified to Jane. The

voice message from Kathy, her daughter, prompted Jane to

reinforce the message by playing it. In this case, the weights

Continuous local reinforcement over time should allow the

system to reach a state where it is somewhat stable and

robust in converging to the user’s preferred notification.

Currently the user’s actions primarily adjust weights for

subsequent messages, however effective reinforcement

learning requires a model that generalizes a notification

policy that maximizes some long-term measure of

reinforcement [8]; this will be the focus of our future work.

Cl-II 99 15-20 MAY 1999 Papers

PRELIMINARY EVALUATION

Although the authors have been using and relining these

techniques during system development, a preliminary 2-day

evaluation was conducted with a novice user, who had prior

experience with mobile phones and 2-way pagers. The user

was able to listen to notifications while attending to tasks in

parallel such as reading or typing. He managed to have

casual discussions with others while hearing notifications;

however he preferred turning off all audio during an

important meeting with his advisor. People nearby

sometimes found the spoken feedback distracting if heard

louder, however that also cued them to wait before

interrupting the user. The volume on the device was

lowered to minimize any disruption to others and maintain

the privacy of messages. The user requested an automatic

volume gain that adapted to the environmental noise level.

In contrast to speech-only feedback, the user found the

unfolding presentation of ambient and auditory cues

allowed sufficient time to switch attention to the incoming

message. Familiarization with the auditory cues was

necessary. He preferred longer and gradual notifications

rather than distinct auditory tones. The priority cues were

the least useful indicator whereas VoiceCues provided

obvious benefit. Knowing the actual priority of a message

was less important than simply having it presented in the

right manner. The user suggested weaving message priority

into the ambient audio (as increased pitch). He found the

overall auditory scheme somewhat complex, preferring

instead a simple notification consisting of ambient

awareness, Voice&es and spoken text.

The user stressed that the ambient audio provided the most

benefit while requiring least cognitive effort. He wished to

hear ambient audio at all times to remain reassured that the

system was still operational. An unintended effect

discovered was that a “pulsating” audio stream indicated

low battery power on the wearable device. A “pause” button

was requested, to hold all messages while participating in a

conversation, along with subtle but periodic auditory alerts

for unread messages waiting in queue. The user felt that

Nomadic Radio provided appropriate awareness and its

expressive qualities justified its use over a pager. A long-

term trial with several nomadic users is necessary to further

validate these notification techniques.

CONCLUSIONS

We have demonstrated techniques for scaleable auditory

presentation and message notification using a variety of

contextual cues. The auditory techniques and notification

model have been refined based on continuous usage by the

authors, however we are currently conducting additional

evaluations with several users. Ongoing work explores

adaptation of the notification model based on reinforcement

from user behavior over time. Our efforts have focused on

wearable audio platforms, however these ideas can be

readily utilized in consumer devices such as pagers, PDAs

and mobile phones to minimize disruptions while providing

timely information to users on the move.

ACKNOWLEDGMENTS

Thanks to Brian Clarkson for ongoing work on the audio

classifier and Stefan Marti for help with user evaluations.

We also thank Lisa Fast and Andre Van Schyndel at Nortel

for their support of the project.

REFERENCES

The Physics of Sound

Sound lies at the very center of speech communication. A sound wave is both the end product of the speech

production mechanism and the primary source of raw material used by the listener to recover the speaker's message.

Because of the central role played by sound in speech communication, it is important to have a good understanding

of how sound is produced, modified, and measured. The purpose of this chapter will be to review some basic

principles underlying the physics of sound, with a particular focus on two ideas that play an especially important

role in both speech and hearing: the concept of the spectrum and acoustic filtering. The speech production

mechanism is a kind of assembly line that operates by generating some relatively simple sounds consisting of

various combinations of buzzes, hisses, and pops, and then filtering those sounds by making a number of fine

adjustments to the tongue, lips, jaw, soft palate, and other articulators. We will also see that a crucial step at the

receiving end occurs when the ear breaks this complex sound into its individual frequency components in much the

same way that a prism breaks white light into components of different optical frequencies. Before getting into these

ideas it is first necessary to cover the basic principles of vibration and sound propagation.

Sound and Vibration

A sound wave is an air pressure disturbance that results from vibration. The vibration can come from a tuning

fork, a guitar string, the column of air in an organ pipe, the head (or rim) of a snare drum, steam escaping from a

radiator, the reed on a clarinet, the diaphragm of a loudspeaker, the vocal cords, or virtually anything that vibrates in

a frequency range that is audible to a listener (roughly 20 to 20,000 cycles per second for humans). The two

conditions that are required for the generation of a sound wave are a vibratory disturbance and an elastic medium,

the most familiar of which is air. We will begin by describing the characteristics of vibrating objects, and then see

what happens when vibratory motion occurs in an elastic medium such as air. We can begin by examining a simple

vibrating object such as the one shown in Figure 3-1. If we set this object into vibration by tapping it from the

bottom, the bar will begin an upward and downward oscillation until the internal resistance of the bar causes the

vibration to cease.

The graph to the right of Figure 3-1 is a visual representation of the upward and downward motion of the bar.

To see how this graph is created, imagine that we use a strobe light to take a series of snapshots of the bar as it

vibrates up and down. For each snapshot, we measure the instantaneous displacement of the bar, which is the

difference between the position of the bar at the split second that the snapshot is taken and the position of the bar at

rest. The rest position of the bar is arbitrarily given a displacement of zero; positive numbers are used for

displacements above the rest position, and negative numbers are used for displacements below the rest position. So,

the first snapshot, taken just as the bar is struck, will show an instantaneous displacement of zero; the next snapshot

will show a small positive displacement, the next will show a somewhat larger positive displacement, and so on. The

pattern that is traced out has a very specific shape to it. The type of vibratory motion that is produced by a simple

vibratory system of this kind is called simple harmonic motion or uniform circular motion, and the pattern that is

traced out in the graph is called a sine wave or a sinusoid.

Figure 3-1. A bar is fixed at one and is set into vibration by tapping it from the bottom. Imagine that

a strobe light is used to take a series of snapshots of the bar as it vibrates up and down. At each

snapshot the instantaneous displacement of the bar is measured. Instantaneous displacement is the

distance between the rest position of the bar (defined as zero displacement) and its position at any

particular instant in time. Positive numbers signify displacements that are above the rest position,

while negative numbers signify displacements that are below the rest position. The vibratory pattern

that is traced out when the sequence of displacements is graphed is called a sinusoid.

The Physics of Sound 2

Basic Terminology

We are now in a position to define some of the basic terminology that applies to sinusoidal vibration.

periodic: The vibratory pattern in Figure 3-1, and the waveform that is shown in the graph, are examples of

periodic vibration, which simply means that there is a pattern that repeats itself over time.

cycle: Cycle refers to one repetition of the pattern. The instantaneous displacement waveform in Figure 3-1 shows

four cycles, or four repetitions of the pattern.

period: Period is the time required to complete one cycle of vibration. For example, if 20 cycles are completed in 1

second, the period is 1/20th of a second (s), or 0.05 s. For speech applications, the most commonly used unit of

measurement for period is the millisecond (ms):

1 ms = 1/1,000 s = 0.001 s = 10 -3 s

A somewhat less commonly used unit is the microsecond (μs):

1 μs = 1/1,000,000 s = 0.000001 s = 10 -6 s

frequency: Frequency is defined as the number of cycles completed in one second. The unit of measurement for

frequency is hertz (Hz), and it is fully synonymous the older and more straightforward term cycles per second

(cps). Conceptually, frequency is simply the rate of vibration. The most crucial function of the auditory system is to

serve as a frequency analyzer – a system that determines how much energy is present at different signal frequencies.

Consequently, frequency is the single most important concept in hearing science. The formula for frequency is:

f = 1/t, where: f = frequency in Hz

t = period in seconds

So, for a period 0.05 s:

f = 1/t = 1/0.05 = 20 Hz

It is important to note that period must be represented in seconds in order to get the answer to come out in cycles per

second, or Hz. If the period is represented in milliseconds, which is very often the case, the period first has to be

converted from milliseconds into seconds by shifting the decimal point three places to the left. For example, for a

period of 10 ms:

f = 1/10 ms = 1/0.01 s = 100 Hz

Similarly, for a period of 100 μs:

f = 1/100 μs = 1/0.0001 s = 10,000 Hz

The period can also be calculated if the frequency is known. Since period and frequency are inversely related, t

= 1/f. So, for a 200 Hz frequency, t = 1/200 = 0.005 s = 5 ms.

Characteristics of Simple Vibratory Systems

Simple vibratory systems of this kind can differ from one another in just three dimensions: frequency,

amplitude, and phase. Figure 3-2 shows examples of signals that differ in frequency. The term amplitude is a bit

different from the other terms that have been discussed thus far, such as force and pressure. As we saw in the last

chapter, terms such as force and pressure have quite specific definitions as various combinations of the basic

dimensions of mass, time, and distance. Amplitude, on the other hand, will be used in this text as a generic term

meaning "how much." How much what? The term amplitude can be used to refer to the magnitude of displacement,

the magnitude of an air pressure disturbance, the magnitude of a force, the magnitude of power, and so on. In the

The Physics of Sound 3

0 5 10 15 20 25 30 35 40 45 50

-10

-5

0

5

10

Time (ms)

Instantaneous Amp.

-10

-5

0

5

10

Instantaneous Amp.

present context, the term amplitude refers to the magnitude of the displacement pattern. Figure 3-3 shows two

displacement waveforms that differ in amplitude. Although the concept of amplitude is as straightforward as the two

waveforms shown in the figure suggest, measuring amplitude is not as simple as it might seem. The reason is that

the instantaneous amplitude of the waveform (in this case, the displacement of the object at a particular split

second in time) is constantly changing. There are many ways to measure amplitude, but a very simple method called

peak-to-peak amplitude will serve our purposes well enough. Peak-to-peak amplitude is simply the difference in

amplitude between the maximum positive and maximum negative peaks in the signal. For example, the bottom

panel in Figure 3-3 has a peak-to-peak amplitude of 10 cm, and the top panel has a peak-to-peak amplitude of 20

cm. Figure 3-4 shows several signals that are identical in frequency and amplitude, but differ from one another in

phase. The waveform labeled 0 o phase would be produced if the bar were set into vibration by tapping it from the

bottom. The waveform labeled 180 o phase would be produced if the bar were set into vibration by tapping it from

the top, so that the initial movement of the bar was downward rather than upward. The waveforms labeled 90 o phase

and 270 o phase would be produced if the bar were set into vibration by pulling the bar to maximum displacement

and letting go -- beginning at maximum positive displacement for 90 o phase, and beginning at maximum negative

displacement for 270 o phase. So, the various vibratory patterns shown in Figure 3-4 are identical except with respect

to phase; that is, they begin at different points in the vibratory cycle. As can be seen in Figure 3-5, the system for

representing phase in degrees treats one cycle of the waveform as a circle; that is, one cycle equals 360 o. For

example, a waveform that begins at zero displacement and shows its initial movement upward has a phase of 0 o, a

waveform that begins at maximum positive displacement and shows its initial movement downward has a phase of

90 o, and so on.

Figure 3-2. Two vibratory patterns that differ in frequency. The panel on top is higher in frequency

than the panel on bottom.

The Physics of Sound 4

0 5 10 15 20 25 30 35 40 45 50

-10

-5

0

5

10

Time (ms)

Instantaneous Amp.

-10

-5

0

5

10

Instantaneous Amp.

Figure 3-3. Two vibratory patterns that differ in amplitude. The panel on top is higher in amplitude than the

panel on bottom.

Phase: 0

Phase: 90

Phase: 180

Phase: 270

Figure 3-4. Four vibratory patterns that differ in phase. Shown above are vibratory patterns with phases of 0 0, 90 0,

180 0, and 270 0.

The Physics of Sound 5

Springs and Masses

We have noted that objects can vibrate at different frequencies, but so far have not discussed the physical

characteristics that are responsible for variations in frequency. There are many factors that affect the natural

vibrating frequency of an object, but among the most important are the mass and stiffness of the object. The effects

of mass and stiffness on natural vibrating frequency can be illustrated with the simple spring-and-mass systems

shown in Figure 3-6. In the pair of spring-and-mass systems to the left, the masses are identical but one spring is

stiffer than the other. If these two spring-and-mass systems are set into vibration, the system with the stiffer spring

will vibrate at a higher frequency than the system with the looser spring. This effect is similar to the changes in

Time ->

Instantaneous Amplitude

0

90

180

270

0/360

Figure 3-5. The system for representing phase treats one cycle of the vibratory pattern as a circle,

consisting of 360 0

. A pattern that begins at zero amplitude heading toward positive values (i.e., heading

upward) is designated 0 0 phase; a waveform that begins at maximum positive displacement and shows

its initial movement downward has a phase of 90 o

; a waveform that begins at zero and heads

downward has a phase of 180 o; and a waveform that begins at maximum negative displacement and

shows its initial movement upward has a phase of 270 o. . The four phase angles that are shown above

are just examples. An infinite variety of phase angles are possible.

Figure 3-6. A spring and mass system whose natural vibrating frequency is controlled by two

parameters: (1) the stiffness of the spring (the stiffer the spring the higher the natural vibrating

frequency), and (2) the mass of the material that is suspended from the spring (the greater the mass, the

lower the natural vibrating frequency).

The Physics of Sound 6

frequency that occur when a guitarist turns the tuning key clockwise or counterclockwise to tune a guitar string by

altering its stiffness.1

The spring-and-mass systems to the right have identical springs but different masses. When these systems are

set into vibration, the system with the greater mass will show a lower natural vibrating frequency. The reason is that

the larger mass shows greater inertia and, consequently, shows greater opposition to changes in direction. Anyone

who has tried to push a car out of mud or snow by rocking it back and forth knows that this is much easier with a

light car than a heavy car. The reason is that the more massive car shows greater opposition to changes in direction.

In summary, the natural vibrating frequency of a spring-and-mass system is controlled by mass and stiffness.

Frequency is directly proportional to stiffness (S↑F↑) and inversely proportional to mass (M↑F↓). It is important to

recognize that these rules apply to all objects, and not just simple spring-and-mass systems. For example, we will

see that the frequency of vibration of the vocal folds is controlled to a very large extent by muscular forces that act

to alter the mass and stiffness of the folds. We will also see that the frequency analysis that is carried out by the

inner ear depends to a large extent on a tuned membrane whose stiffness varies systematically from one end of the

cochlea to the other.

Sound Propagation

As was mentioned at the beginning of this chapter, the generation of a sound wave requires not only vibration,

but also an elastic medium in which the disturbance created by that vibration can be transmitted (see Box 3-1 [bell

jar experiment described in Patrick's science book - not yet written]). To say that air is an elastic medium means that

air, like all other matter, tends to return to its original shape after it is deformed through the application of a force.

The prototypical example of an object that exhibits this kind of restoring force is a spring. To understand the

mechanism underlying sound propagation, it is useful to think of air as consisting of collection of particles that are

connected to one another by springs, with the springs representing the restoring forces associated with the elasticity

of the medium. Air pressure is related to particle density. When a volume of air is undisturbed, the individual

particles of air distribute themselves more-or-less evenly, and the elastic forces are at their resting state. A volume of

air that is in this undisturbed state it is said to be at atmospheric pressure. For our purposes, atmospheric pressure

can be defined in terms of two interrelated conditions: (1) the air molecules are approximately evenly spaced, and

(2) the elastic forces, represented by the interconnecting springs, are neither compressed nor stretched beyond their

resting state. When a vibratory disturbance causes the air particles to crowd together (i.e., producing an increase in

particle density), air pressure is higher than atmospheric, and the elastic forces are in a compressed state.

Conversely, when particle spacing is relatively large, air pressure is lower than atmospheric.

1The example of tuning a guitar string is imperfect since the mass of the vibrating portion of the string decreases slightly as the string is

tightened. This occurs because a portion of the string is wound onto the tuning key as it is tightened.

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

a b c d e f g h i

TIME

Figure 3-7. Shown above is a highly schematic illustration of the chain reaction that

results in the propagation of a sound wave (modeled after Denes and Pinson, 1963).

The Physics of Sound 7

When a vibrating object is placed in an elastic medium, an air pressure disturbance is created through a chain

reaction similar to that illustrated in Figure 3-7. As the vibrating object (a tuning fork in this case) moves to the

right, particle a, which is immediately adjacent to the tuning fork, is displaced to the right. The elastic force

generated between particles a and b (not shown in the figure) has the effect a split second later of displacing particle

b to the right. This disturbance will eventually reach particles c, d, e, and so on, and in each case the particles will be

momentarily crowded together. This crowding effect is called compression or condensation, and it is characterized

by dense particle spacing and, consequently, air pressure that is slightly higher than atmospheric pressure. The

propagation of the disturbance is analogous to the chain reaction that occurs when an arrangement of dominos is

toppled over. Figure 3-7 also shows that at some close distance to the left of a point of compression, particle spacing

will be greater than average, and the elastic forces will be in a stretched state. This effect is called rarefaction, and

it is characterized by relatively wide particle spacing and, consequently, air pressure that is slightly lower than

atmospheric pressure.

The compression wave, along with the rarefaction wave that immediately follows it, will be propagated outward

at the speed of sound. The speed of sound varies depending on the average elasticity and density of the medium in

which the sound is propagated, but a good working figure for air is about 35,000 centimeters per second, or

approximately 783 miles per hour. Although Figure 3-7 gives a reasonably good idea of how sound propagation

works, it is misleading in two respects. First, the scale is inaccurate to an absurd degree: a single cubic inch of air

contains approximately 400 billion molecules, and not the handful of particles shown in the figure. Consequently,

the compression and rarefaction effects are statistical rather than strictly deterministic as shown in Figure 3-7.

Second, although Figure 3-7 makes it appear that the air pressure disturbance is propagated in a simple straight line

from the vibrating object, it actually travels in all directions from the source. This idea is captured somewhat better

in Figure 3-8, which shows sound propagation in two of the three dimensions in which the disturbance will be

transmitted. The figure shows rod and piston connected to a wheel spinning at a constant speed. Connected to the

piston is a balloon that expands and contracts as the piston moves in and out of the cylinder. As the balloon expands

the air particles are compressed; i.e., air pressure is momentarily higher than atmospheric. Conversely, when the

balloon contracts the air particles are sucked inward, resulting in rarefaction. The alternating compression and

rarefaction waves are propagated outward in all directions form the source. Only two of the three dimensions are

shown here; that is, the shape of the pressure disturbance is actually spherical rather than the circular pattern that is

shown here. Superimposed on the figure, in the graph labeled “one line of propagation,” is the resulting air pressure

waveform. Note that the pressure waveform takes on a high value during instants of compression and a low value

during instants of rarefaction. The figure also gives some idea of where the term uniform circular motion comes

from. If one were to make a graph plotting the height of the connecting rod on the rotating wheel as a function of

time it would trace out a perfect sinusoid; i.e., with exactly the shape of the pressure waveform that is superimposed

on the figure.

The Sound Pressure Waveform

Returning to Figure 3-7 for a moment, imagine that we chose some specific distance from the tuning fork to

observe how the movement and density of air particles varied with time. We would see individual air particles

oscillating small distances back and forth, and if we monitored particle density we would find that high particle

density (high air pressure) would be followed a moment later by relatively even particle spacing (atmospheric

pressure), which would be followed by a moment later by wide particle spacing (low air pressure), and so on.

Therefore, for an object that is vibrating sinusoidally, a graph showing variations in instantaneous air pressure

over time would also be sinusoidal. This is illustrated in Figure 3-9.

The vibratory patterns that have been discussed so far have all been sinusoidal. The concept of a sinusoid has

not been formally defined, but for our purposes it is enough to know that a sinusoid has precisely the smooth shape

that is shown in Figures such as 3-4 and 3-5. While sinusoids, also known as pure tones, have a very special place

in acoustic theory, they are rarely encountered in nature. The sound produced by a tuning fork comes quite close to a

sinusoidal shape, as do the simple tones that are used in hearing tests. Much more common in both speech and music

are more complex, nonsinusoidal patterns, to be discussed below. As will be seen in later chapters, these complex

vibratory patterns play a very important role in speech.

The Physics of Sound 8

The Frequency Domain

We now arrive at what is probably the single most important concept for understanding both hearing and speech

acoustics. The graphs that we have used up to this point for representing either vibratory motion or the air pressure

disturbance created by this motion are called time domain representations. These graphs show how instantaneous

displacement (or instantaneous air pressure) varies over time. Another method for representing either sound or

vibration is called a frequency domain representation, also known as a spectrum. There are, in fact, two kinds of

frequency domain representations that are used to characterize sound. One is called an amplitude spectrum (also

known as a magnitude spectrum or a power spectrum, depending on how the level of the signal is represented)

and the other is called a phase spectrum. For reasons that will become clear soon, the amplitude spectrum is by far

the more important of the two. An amplitude spectrum is simply a graph showing what frequencies are present with

what amplitudes. Frequency is given along the x axis and some measure of amplitude is given on the y axis. A phase

spectrum is a graph showing what frequencies are present with what phases.

Figure 3-10 shows examples of the amplitude and phase spectra for several sinusoidal signals. The top panel

shows a time-domain representation of a sinusoid with a period of 10 ms and, consequently, a frequency of 100 Hz

(f = 1/t = 1/0.01 sec = 100 Hz). The peak-to-peak amplitude for this signal is 400 μPa, and the signal has a phase of

90 o. Since the amplitude spectrum is a graph showing what frequencies are present with what amplitudes, the

amplitude spectrum for this signal will show a single line at 100 Hz with a height of 400 μPa. The phase spectrum is

a graph showing what frequencies are present with what phases, so the phase spectrum for this signal will show a

single line at 100 Hz with a height of 90 o

. The second panel in Figure 3-10 shows a 200 Hz sinusoid with a peak-to-

peak amplitude of 200 μPa and a phase of 180 o

. Consequently, the amplitude spectrum will show a single line at 200

Hz with a height of 100 μPa, while the phase spectrum will show a line at 200 Hz with a height of 180 o.

Complex Periodic Sounds

Sinusoids are sometimes referred to as simple periodic signals. The term "periodic" means that there is a

pattern that repeats itself, and the term "simple" means that there is only one frequency component present. This is

confirmed in the frequency domain representations in Figure 3-10, which all show a single frequency component in

both the amplitude and phase spectra. Complex periodic signals involve the repetition of a nonsinusoidal pattern,

and in all cases, complex periodic signals consist of more than a single frequency component. All nonsinusoidal

periodic signals are considered complex periodic.

Figure 3-8 Illustration of the propagation of a sound wave in two dimensions.

The Physics of Sound 9

Figure 3-11 shows several examples of complex periodic signals, along with the amplitude spectra for these signals.

The time required to complete one cycle of the complex pattern is called the fundamental period. This is precisely

the same concept as the term period that was introduced earlier. The only reason for using the term "fundamental

period" instead of the simpler term "period" for complex periodic signals is to differentiate the fundamental period

(the time required to complete one cycle of the pattern as a whole) from other periods that may be present in the

signal (e.g., more rapid oscillations that might be observed within each cycle). The symbol for fundamental period is

t o. Fundamental frequency (f o) is calculated from fundamental period using the same kind of formula that we used

earlier for sinusoids:

fo = 1/to

The signal in the top panel of Figure 3-11 has a fundamental period of 5 ms, so fo = 1/0.005 = 200 Hz.

Examination of the amplitude spectra of the signals in Figure 3-11 confirms that they do, in fact, consist of

more than a single frequency. In fact, complex periodic signals show a very particular kind of amplitude spectrum

called a harmonic spectrum. A harmonic spectrum shows energy at the fundamental frequency and at whole

number multiples of the fundamental frequency. For example, the signal in the top panel of Figure 3-11 has energy

present at 200 Hz, 400 Hz, 600 Hz, 800 Hz, 1,000 Hz, 1200 Hz, and so on. Each frequency component in the

0 5 10 15 20 25 30

-200

-100

0

100

200

Inst. Air Pressure

Period: 10 ms, Freq: 100 Hz, Amp: 400, Phase: 90

0 5 10 15 20 25 30

-200

-100

0

100

200

Inst. Air Pressure

Period: 5 ms, Freq: 200 Hz, Amp: 200, Phase: 180

0 5 10 15 20 25 30

-200

-100

0

100

200

Time (msec)

Inst. Air Pressure

Period: 2.5 ms, Freq: 400 Hz, Amp: 200, Phase: 270

TIME DOMAIN FREQUENCY DOMAIN

0 100 200 300 400 500

0

100

200

300

400

Frequency (Hz)

Amplitude

Amplitude Spectrum

0 100 200 300 400 500

0

100

200

300

400

Frequency (Hz)

Amplitude

0 100 200 300 400 500

0

100

200

300

400

Frequency (Hz)

Amplitude

0 100 200 300 400 500

0

90

180

270

360

Frequency (Hz)

Phase

Phase Spectrum

0 100 200 300 400 500

0

90

180

270

360

Frequency (Hz)

Phase

0 100 200 300 400 500

0

90

180

270

360

Frequency (Hz)

Phase

Figure 3-10. Time and frequency domain representations of three sinusoids. The frequency domain

consists of two graphs: an amplitude spectrum and a phase spectrum. An amplitude spectrum is a

graph showing what frequencies are present with what amplitudes, and a phase spectrum is a graph

showing the phases of each frequency component.

The Physics of Sound 10

amplitude spectrum of a complex periodic signal is called a harmonic (also known as a partial). The fundamental

frequency, in this case 200 Hz, is also called the first harmonic, the 400 Hz component (2 ⋅ fo) is called the second

harmonic, the 600 Hz component (3 ⋅ fo) is called the third harmonic, and so on.

The second panel in Figure 3-11 shows a complex periodic signal with a fundamental period of 10 ms and,

consequently, a fundamental frequency of 100 Hz. The harmonic spectrum that is associated with this signal will

therefore show energy at 100 Hz, 200 Hz, 300 Hz, 400 Hz, 500 Hz, and so on. The bottom panel of Figure 3-11

shows a complex periodic signal with a fundamental period of 2.5 ms, a fundamental frequency of 400 Hz, and

harmonics at 400, 800, 1200, 1600, and so on. Notice that there two completely interchangeable ways to define the

term fundamental frequency. In the time domain, the fundamental frequency is the number of cycles of the complex

pattern that are completed in one second. In the frequency domain, except in the case of certain special signals, the

fundamental frequency is the lowest harmonic in the harmonic spectrum. Also, the fundamental frequency defines

the harmonic spacing; that is, when the fundamental frequency is 100 Hz, harmonics will be spaced at 100 Hz

Figure 3-11. Time and frequency domain representations of three complex periodic signals.

Complex periodic signals have harmonic spectra, with energy at the fundamental frequency (f0) and

at whole number multiples of f0 (f0. 2, f0. 3, f0. 4, etc.) For example, the signal in the upper left, with a

fundamental frequency of 200 Hz, shows energy at 200 Hz, 400 Hz, 600 Hz, etc. In the spectra on

the right, amplitude is measured in arbitrary units. The main point being made in this figure is the

distribution of harmonic frequencies at whole number multiples of f0 for complex periodic signals.

0 5 10 15 20 25 30

-200

-100

0

100

200

Inst. Air Pres. (UPa) t0: 5 ms, f0: 200 Hz

t0: 10 ms, f0: 100 Hz

0 5 10 15 20 25 30

-200

-100

0

100

200

Inst. Air Pres. (UPa)

t0: 2.5 ms, f0: 400 Hz

0 5 10 15 20 25 30

-200

-100

0

100

200

Time (msec)

Inst. Air Pres. (UPa)

0 200 400 600 800 1000 1200 1400 1600

0

20

40

60

80

100

120

Frequency (Hz)

Amplitude

0 200 400 600 800 1000 1200 1400 1600

0

20

40

60

80

100

120

Frequency (Hz)

Amplitude

0 200 400 600 800 1000 1200 1400 1600

0

20

40

60

80

100

120

Frequency (Hz)

Amplitude

TIME DOMAIN FREQUENCY DOMAIN

The Physics of Sound 11

0 10 20 30 40 50

-200

-100

0

100

200

Inst. Air Pres. (UPa)

White Noise

/s/

0 10 20 30 40 50

-200

-100

0

100

200

Inst. Air Pres. (UPa)

/f/

0 10 20 30 40 50

-200

-100

0

100

200

TIME (msec)

Inst. Air Pres. (UPa)

0 1 2 3 4 5 6 7 8 9 10

0

20

40

60

80

100

Amplitude

0 1 2 3 4 5 6 7 8 9 10

0

20

40

60

80

100

Amplitude

0 1 2 3 4 5 6 7 8 9 10

0

20

40

60

80

100

Frequency (kHz)

Amplitude

TIME DOMAIN FREQUENCY DOMAIN

intervals (i.e., 100, 200, 300 ...), when the fundamental frequency is 125 Hz, harmonics will be spaced at 125 Hz

intervals (i.e., 125, 250, 375...), and when the fundamental frequency is 200 Hz, harmonics will be spaced at 200 Hz

intervals (i.e., 200, 400, 600 ...). (For some special signals this will not be the case.2) So, when fo is low, harmonics

will be closely spaced, and when fo is high, harmonics will be widely spaced. This is clearly seen in Figure 3-11: the

signal with the lowest f0 (100 Hz, the middle signal) shows the narrowest harmonic spacing, while the signal with

the highest f0 (400 Hz, the bottom signal) shows the widest harmonic spacing.

There are certain characteristics of the spectra of complex periodic sounds that can be determined by making simple

measurements of the time domain signal, and there are certain other characteristics that require a more complex

analysis. For example, simply by examining the signal in the bottom panel of Figure 3-11 we can determine that it is

complex periodic (i.e., it is periodic but not sinusoidal) and therefore it will show a harmonic spectrum with energy

at whole number multiples of the fundamental frequency. Further, by measuring the fundamental period (2.5 ms)

2There are some complex periodic signals that have energy at odd multiples of the fundamental frequency only. A square wave, for

example, is a signal that alternates between maximum positive amplitude and maximum negative amplitude. The spectrum of square wave shows

energy at odd multiples of the fundamental frequency only. Also, a variety of simple signal processing tricks can be used to create signals with

harmonics at any arbitrary set of frequencies. For example, it is a simple matter to create a signal with energy at 400, 500, and 600 Hz only.

While these kinds of signals can be quite useful for conducting auditory perception experiments, it remains true that most naturally occurring

complex periodic signals have energy at all whole number multiples of the fundamental frequency.

Figure 3-12. Time and frequency domain representations of three non-transient complex aperiodic

signals. Unlike complex periodic signals, complex aperiodic signals show energy that is spread

across the spectrum. This type of spectrum is called dense or continuous. These spectra have a very

different appearance from the “picket fence” look that is associated with the discrete, harmonic

spectra of complex periodic signals.

The Physics of Sound 12

and converting it into fundamental frequency (400 Hz), we are able to determine that the signal will have energy at

400, 800, 1200, 1600, etc. But how do we know the amplitude of each of these frequency components? And how do

we know the phase of each component? The answer is that you cannot determine harmonic amplitudes or phases

simply by inspecting the signal or by making simple measurements of the time domain signals with a ruler. We will

see soon that a technique called Fourier analysis is able to determine both the amplitude spectrum and the phase

spectrum of any signal. We will also see that the inner ears of humans and many other animals have developed a

trick that is able to produce a neural representation that is comparable in some respects to an amplitude spectrum.

We will also see that the ear has no comparable trick for deriving a representation that is equivalent to a phase

spectrum. This explains why the amplitude spectrum is far more important for speech and hearing applications than

the phase spectrum. We will return to this point later.

To summarize: (1) a complex periodic signal is any periodic signal that is not sinusoidal, (2) complex periodic

signals have energy at the fundamental frequency (fo) and at whole number multiples of the fundamental frequency

(2 ⋅ fo, 3 ⋅ fo , 4 ⋅ fo ...), and (3) although measuring the fundamental frequency allows us to determine the frequency

locations of harmonics, there is no simple measurement that can tell us harmonic amplitudes or phases. For this,

Fourier analysis or some other spectrum analysis technique is needed.

Figure 3-13. Time and frequency domain representations of three transients. Transients are complex

aperiodic signals that are defined by their brief duration. Pops, clicks, and the sound gun fire are

examples of transients. In common with longer duration complex aperiodic signals, transients show

dense or continuous spectra, very unlike the discrete, harmonic spectra associated with complex periodic

d

0 10 20 30 40 50 60 70 80 90 100

-200

-100

0

100

200

Inst. Amp. (UPa)

Rap on Desk

Clap

0 10 20 30 40 50 60 70 80 90 100

-200

-100

0

100

200

Inst. Amp. (UPa)

Tap on Cheek

0 10 20 30 40 50 60 70 80 90 100

-200

-100

0

100

200

TIME (msec)

Inst. Amp. (UPa)

0 1 2 3 4 5

0

20

40

60

80

100

Amplitude

0 1 2 3 4 5

0

20

40

60

80

100

Amplitude

0 1 2 3 4 5

0

20

40

60

80

100

Frequency (kHz)

Amplitude

TIME DOMAIN FREQUENCY DOMAIN

The Physics of Sound 13

-200

0

200

Inst. Air Pres.

(a)

-200

0

200

Inst. Air Pres.

(b)

-200

0

200

Inst. Air Pres.

(c)

-200

0

200

Inst. Air Pres.

(d)

-300

0

300

Inst. Air Pres.

(e)

Time ->

Aperiodic Sounds

An aperiodic sound is any sound that does not show a repeating pattern in its time domain representation. There are

many aperiodic sounds in speech. Examples include the hissy sounds associated with fricatives such as /f/ and /s/,

and the various hisses and pops associated with articulatory release for the stop consonants /b,d,g,p,t,k/. Examples of

non-speech aperiodic sounds include a drummer's cymbal or snare drum, the hiss produced by a radiator, and static

sound produced by a poorly tuned radio. There are two types of aperiodic sounds: (1) continuous aperiodic sounds

(also known as noise) and (2) transients. Although there is no sharp cutoff, the distinction between continuous

aperiodic sounds and transients is based on duration. Transients (also "pops" and "clicks") are defined by their very

brief duration, and continuous aperiodic sounds are of longer duration. Figure 3-12 shows several examples of time

domain representations and amplitude spectra for continuous aperiodic sounds. The lack of periodicity in the time

Figure 3-14. Illustration of the principle underlying Fourier analysis. The complex periodic signal

shown in panel e was derived by point-for-point summation of the sinusoidal signals shown in

panels a-d. Point-for-point summation simply means beginning at time zero (i.e., the start of the

signal) and adding the instantaneous amplitude of signal a to the instantaneous amplitude of signal b

at time zero, then adding that sum to the instantaneous amplitude of signal c, also at time zero, then

adding that sum to instantaneous amplitude of signal d at time zero. The sum of instantaneous

amplitudes at time zero of signals a-d is the instantaneous amplitude of the composite signal e at

time zero. For example, at time zero the amplitudes of sinusoids a-d are 0, +100, -200, and 0,

respectively, producing a sum of -100. This agrees with the instantaneous amplitude at the very

beginning of composite signal e. The same summation procedure is followed for all time points.

The Physics of Sound 14

domain is quite evident; that is, unlike the periodic sounds we have seen, there is no pattern that repeats itself over

time.

All aperiodic sounds -- both continuous and transient -- are complex in the sense that they always consist of

energy at more than one frequency. The characteristic feature of aperiodic sounds in the frequency domain is a

dense or continuous spectrum, which stands in contrast to the harmonic spectrum that is associated with complex

periodic sounds. In a harmonic spectrum, there is energy at the fundamental frequency, followed by a gap with little

or no energy, followed by energy at the second harmonic, followed by another gap, and so on. The spectra of

aperiodic sounds do not share this "picket fence" appearance. Instead, energy is smeared more-or-less continuously

across the spectrum. The top panel in Figure 3-12 shows a specific type of continuous aperiodic sound called white

noise. By analogy to white light, white noise has a flat amplitude spectrum; that is, approximately equal amplitude at

all frequencies. The middle panel in Figure 3-12 shows the sound /s/, and the bottom panel shows sound /f/. Notice

that the spectra for all three sounds are dense; that is, they do not show the "picket fence" look that reveals harmonic

structure. As was the case for complex periodic sounds, there is no way to tell how much energy there will be at

different frequencies by inspecting the time domain signal or by making any simple measures with a ruler. Likewise,

there is no simple way to determine the phase spectrum. So, after inspecting a time-domain signal and determining

that it is aperiodic, all we know for sure is that it will have a dense spectrum rather than a harmonic spectrum.

Figure 3-13 shows time domain representations and amplitude spectra for three transients. The transient in the

top panel was produced by rapping on a wooden desk, the second is a single clap of the hands, and the third was

produced by holding the mouth in position for the vowel /o/, and tapping the cheek with an index finger. Note the

brief durations of the signals. Also, as with continuous aperiodic sounds, the spectra associated with transients are

dense; that is, there is no evidence of harmonic organization. In speech, transients occur at the instant of articulatory

release for stop consonants. There are also some languages, such as the South African languages Zulu, Hottentot,

and Xhosa, that contain mouth clicks as part of their phonemic inventory (MacKay, 1986). Fourier Analysis

TIME DOMAIN

Time ->

Inst. Air Pres.

Fourier

Analyzer

0 200 400 600 800

Frequency (Hz)

Amplitude

FREQUENCY DOMAIN

0 200 400 600 800

Frequency (Hz)

Phase

Figure 3-15. A signal enters a Fourier analyzer in the time domain and exits in the frequency domain.

As outputs, the Fourier analyzer produces two frequency-domain representations: an amplitude

spectrum that shows the amplitude of each sinusoidal component that is present in the input signal, and

a phase spectrum that shows the phase of each of the sinusoids. The input signal can be reconstructed

perfectly by summing sinusoids at frequencies, amplitudes, and phase that are shown in the Fourier

amplitude and phase spectra, using the summing method that is illustrated in Figure 3-14..

The Physics of Sound 15

Fourier analysis is an extremely powerful tool that has widespread applications in nearly every major branch

of physics and engineering. The method was developed by the 19 th century mathematician Joseph Fourier, and

although Fourier was studying thermal waves at the time, the technique can be applied to the frequency analysis of

any kind of wave. Fourier's great insight was the discovery that all complex waves can be derived by adding

sinusoids together, so long as the sinusoids are of the appropriate frequencies, amplitudes, and phases. For example,

the complex periodic signal at the bottom of Figure 3-14 can be derived by summing sinusoids at 100, 200, 300, and

400 Hz, with each sinusoidal component having the amplitude and phase that is shown in the figure (see the caption

of Figure 3-14 for an explanation of what is meant by summing the sinusoidal components). The assumption that all

complex waves can be derived by adding sinusoids together is called Fourier's theorem, and the analysis technique

that Fourier developed from this theorem is called Fourier analysis. Fourier analysis is a mathematical technique that

takes a time domain signal as its input and determines: (1) the amplitude of each sinusoidal component that is

present in the input signal, and (2) the phase of each sinusoidal component that is present in the input signal.

Another way of stating this is that Fourier analysis takes a time domain signal as its input and produces two

frequency domain representations as output: (1) an amplitude spectrum, and (2) a phase spectrum.

The basic concept is illustrated in Figure 3-15, which shows a time domain signal entering the Fourier analyzer.

Emerging at the output of the Fourier analyzer is an amplitude spectrum (a graph showing the amplitude of each

sinusoid that is present in the input signal) and a phase spectrum (a graph showing the phase of each sinusoid that is

present in the input signal). The amplitude spectrum tells us that the input signal contains: (1) 200 Hz sinusoid with

an amplitude of 100 μPa, a 400 Hz sinusoid with an amplitude of 200 μPa, and a 600 Hz sinusoid with an amplitude

of 50 μPa. Similarly, the phase spectrum tells us that the 200 Hz sinusoid has a phase of 90 o, the 400 Hz sinusoid

has a phase of 180 o, and the 600 Hz sinusoid has a phase of 270 o. If Fourier's theorem is correct, we should be able

to reconstruct the input signal by summing sinusoids at 200, 400, and 600 Hz, using the amplitudes and phases that

are shown. In fact, summing these three sinusoids in this way would precisely reproduce the original time domain

signal; that is, we would get back an exact replica of our original signal, and not just a rough approximation to it.

For our purposes it is not important to understand how Fourier analysis works. The most important point about

Fourier's idea is that, visual appearances aside, all complex waves consist of sinusoids of varying frequencies,

amplitudes, and phases. In fact, Fourier analysis applies not only to periodic signals such as those shown in Figure

3-15, but also to noise and transients. In fact, the amplitude spectra of the aperiodic signals shown in Figure 3-13

were calculated using Fourier analysis. In later chapters we will see that the auditory system is able to derive a

neural representation that is roughly comparable to a Fourier amplitude spectrum. However, as was mentioned

earlier, the auditory system does not derive a representation comparable to a Fourier phase spectrum. As a result,

listeners are very sensitive to changes in the amplitude spectrum but are relatively insensitive to changes in phase.

Some Additional Terminology

Overtones vs. Harmonics: The term overtone and the term harmonic refer to the same concept; they are just

counted differently. As we have seen, in a harmonic series such as 100, 200, 300, 400, etc., the 100 Hz component

can be referred to as either the fundamental frequency or the first harmonic; the 200 Hz component is the second

harmonic, the 300 Hz component is the third harmonic, and so on. An alternative set of terminology would refer to

the 100 Hz component as the fundamental frequency, the 200 Hz component as the first overtone, the 300 Hz

component as the second overtone, and so on. Use of the term overtone tends to be favored by those interested in

musical acoustics, while most other acousticians tend to use the term harmonic.

Octaves vs. Harmonics: An octave refers to a doubling of frequency. So, if we begin at 100 Hz, the next octave up

would 200 Hz, the next would be 400 Hz, the next would be 800 Hz, and so on. Note that this is quite different from

a harmonic progression. A harmonic progression beginning at 300 Hz would be 300, 600, 900, 1200, 1500, etc.,

while an octave progression would be 300, 600, 1200, 2400, 4800, etc. There is something auditorilly natural about

octave spacing, and octaves play a very important role in the organization of musical scales. For example, on a piano

keyboard, middle A (A5) is 440 Hz, A above middle A (A 6) is 880 Hz, A7 is 1,760 and so on. (See Box 3-2).

Wavelength: The concept of wavelength is best illustrated with an example given by Small (1973). Small asks us

to imagine dipping a finger repeatedly into a puddle of water at a perfectly regular interval. Each time the finger hits

the water, a wave is propagated outward, and we would see a pattern formed consisting of a series of concentric

The Physics of Sound 16

circles (see Figure 3-16). Wavelength is simply the distance between the adjacent waves. Precisely the same concept

can be applied to sound waves: wavelength is simply the distance between one compression wave and the next (or

one rarefaction wave and the next or, more generally, the distance between any two corresponding points in adjacent

waves). For our purposes, the most important point to be made about wavelength is that there is a simple

relationship between frequency and wavelength. Using the puddle example, imagine that we begin by dipping our

finger into the puddle at a very slow rate; that is, with a low "dipping frequency." Since the waves have a long

period of time to travel from one dip to the next, the wavelength will be large. By the same reasoning, the

wavelength becomes smaller as the "dipping frequency" is increased; that is, the time allowed for the wave to travel

at high "dipping frequency" is small, so the wavelength is small. Wavelength is a measure of distance, and the

formula for calculating wavelength is a straightforward algebraic rearrangement of the familiar "distance = rate ⋅

time" formula from junior high school.

λ = c/f, where: λ = wavelength

c = the speed of sound

f = frequency

By rearranging the formula, frequency can be calculated if wavelength and the speed of sound are known:

f = c/λ

Lower Frequency

(Longer Wavelength)

Higher Frequency

(Shorter Wavelength)

Figure 3-16. Wavelength is a measure of the distance between the crest of one cycle of a wave and the

crest of the next cycle (or trough to trough or, in fact, the distance between any two corresponding

points in the wave). Wavelength and frequency are related to one another. Because the wave has only a

short time to travel from one cycle to the next, high frequencies produce short wavelengths.

Conversely, because of the longer travel times, low frequencies produce long wavelengths.

The Physics of Sound 17

Spectrum Envelope: The term spectrum envelope refers to an imaginary smooth line drawn to enclose an

amplitude spectrum. Figure 3-17 shows several examples. This is a rather simple concept that will play a very

important role in understanding certain aspects of auditory perception. For example, we will see that our perception

of a perceptual attribute called timbre (also called sound quality) is controlled primarily by the shape of the

spectrum envelope, and not by the fine details of the amplitude spectrum. The examples in Figure 3-17 show how

differences in spectrum envelope play a role in signaling differences in one specific example of timbre called

vowel quality (i.e., whether a vowel sounds like /i/ vs. /a/ vs. /u/, etc.). For example, panels a and b in Figure 3-17

show the vowel /å/ produced at two different fundamental frequencies. (We know that the fundamental frequencies

are different because one spectrum shows wide harmonic spacing and the other shows narrow harmonic spacing.)

The fact that the two vowels are heard as /a/ despite the difference in fundamental frequency can be attributed to the

fact that these two signals have similar spectrum envelopes. Panels c and d in Figure 3-17 show the spectra of two

signals with different spectrum envelopes but the same fundamental frequency (i.e., with the same harmonic

spacing). As we will see in the chapter on auditory perception, differences in fundamental frequency are perceived

as differences in pitch. So, for signals (a) and (b) in Figure 3-17, the listener will hear the same vowel produced at

two different pitches. Conversely, for signals (c) and (d) in Figure 3-17, the listener will hear two different vowels

produced at the same pitch. We will return to the concept of spectrum envelope in the chapter on auditory

perception.

Amplitude Envelope: The term amplitude envelope refers to an imaginary smooth line that is drawn on top of a

time domain signal. Figure 3-18 shows sinusoids that are identical except for their amplitude envelopes. It can be

seen that the different amplitude envelopes reflect differences in the way the sounds are turned on and off. For

example, panel a shows a signal that is turned on abruptly and turned off abruptly; panel b shows a signal that is

turned on gradually and turned off abruptly; and so on. Differences in amplitude envelope have an important effect

on the quality of a sound. As we will see in the chapter on auditory perception, amplitude envelope, along with

spectrum envelope discussed above, is another physical parameter that affects timbre or sound quality. For

0 1 2 3

0

10

20

30

40

50

60

70

Frequency (kHz)

Amplitude

(a) Vowel: /a/, f0: 100 Hz

0 1 2 3

0

10

20

30

40

50

60

70

Frequency (kHz)

Amplitude

(b) Vowel: /a/, f0: 200 Hz

0 1 2 3

0

10

20

30

40

50

60

70

Frequency (kHz)

Amplitude

(c)

Vowel: /i/, f0: 150 Hz

0 1 2 3

0

10

20

30

40

50

60

70

Frequency (kHz)

Amplitude

(d)

Vowel: /u/, f0: 150 Hz

Figure 3-17. A spectrum envelope is an imaginary smooth line drawn to enclose an amplitude

spectrum. Panels a and b show the spectra of two signals (the vowel /ɑ/) with different fundamental

frequencies (note the differences in harmonic spacing) but very similar spectrum envelopes. Panels c

and d show the spectra of two signals with different spectrum envelopes (the vowels /i/ and /u/ in this

case) but the same fundamental frequencies (i.e., the same harmonic spacing).

The Physics of Sound 18

example, piano players know that a given note will sound different depending on whether or not the damping pedal

is used. Similarly, notes played on a stringed instrument such as a violin or cello will sound different depending on

whether the note is plucked or bowed. In both cases, the underlying acoustic difference is amplitude envelope.

Acoustic Filters

As will be seen in subsequent chapters, acoustic filtering plays a central role in the processing of sound by the

inner ear. The human vocal tract also serves as an acoustic filter that modifies and shapes the sounds that are created

by the larynx and other articulators. For this reason, it is quite important to understand how acoustic filters work. In

the most general sense, the term filter refers to a device or system that is selective about the kinds of things that are

allowed to pass through versus the kinds of things that are blocked. An oil filter, for example, is designed to allow

oil to pass through while blocking particles of dirt. Of special interest to speech and hearing science are frequency

selective filters. These are devices that allow some frequencies to pass through while blocking or attenuating other

frequencies. (The term attenuate means to weaken or reduce in amplitude).

A simple example of a frequency selective filter from the world of optics is a pair of tinted sunglasses. A piece

of white paper that is viewed through red tinted sunglasses will appear red. Since the original piece of paper is

white, and since we know that white light consists of all of the visible optical frequencies mixed in equal amounts,

the reason that the paper appears red through the red tinted glasses is that optical frequencies other than those

corresponding to red are being blocked or attenuated by the optical filter. As a result, it is primarily the red light that

is being allowed to pass through. (Starting at the lowest optical frequency and going to the highest, light will appear

red, orange, yellow, green, blue, indigo, and violet.)

Inst. Air Pres.

(a)

Signals Differing in Amplitude Envelope

Inst. Air Pres.

(b)

Inst. Air Pres.

(c)

Time ->

Inst. Air Pres.

(d)

Figure 3-18. Amplitude envelope is an imaginary smooth line drawn to enclose a time-domain signal.

This feature describes how a sound is turned on and turned off; for example, whether the sound is

turned on abruptly and turned off abruptly (panel a), turned on gradually and turned off abruptly (panel

b), turned on abruptly and turned off gradually (panel c), or turned on and off gradually (panel d).

The Physics of Sound 19

A graph called a frequency response curve is used to describe how a frequency selective filter will behave. A

frequency response curve is a graph showing how energy at different frequencies will be affected by the filter.

Specifically, a frequency response curve plots a variable called "gain" as a function of variations in the frequency of

the input signal. Gain is the amount of amplification provided by the filter at different signal frequencies. Gains are

interpreted as amplitude multipliers; for example, suppose that the gain of a filter at 100 Hz is 1.3. If a 100 Hz

sinusoid enters the filter measuring 10 uPa, the amplitude at the output of the filter at 100 Hz will measure 13 μPa

(10 μPa x 1.3 = 13 μPa). The only catch in this scheme is that gains can and very frequently are less than 1, meaning

that the effect of the filter will be to attenuate the signal. For example, if the gain at 100 Hz is 0.5, a 10 μPa input

signal at 100 Hz will measure 5 μPa at the output of the filter. When the filter gain is 1.0, the signal is unaffected by

the filter; i.e., a 10 μPa input signal will measure 10 μPa at the output of the filter.

Figure 3-19 shows frequency response curves for several optical filters. Panel a shows a frequency response

curve for the red optical filter discussed in the example above. If we put white light into the filter in panel a, the

signal amplitude at the output of the filter will be high only when the frequency of the input signal is low. This is

because the gain of the filter is high only in the low-frequency portion of the frequency-response curve. This is an

example of a lowpass filter; that is, a filter that allows low frequencies to pass through. Panel b shows an optical

filter that has precisely the reverse effect on an input signal; that is, this filter will allow high frequencies to pass

through while attenuating low- and mid-frequency signals. A white surface viewed through this filter would

therefore appear violet. This is an example of a highpass filter. Panel c shows the frequency response curve for a

filter that allows a band of energy in the center of the spectrum to pass through while attenuating signal components

of higher and lower frequency. A white surface viewed through this filter would appear green. This is called a

bandpass filter.

Acoustic filters do for sound exactly what optical filters do for light; that is, they allow some frequencies to pass

through while attenuating other frequencies. To get a better idea of how a frequency response curve is measured,

imagine that we ask a singer to attempt to shatter a crystal wine glass with a voice signal alone. To see how the

frequency response curve is created we have to make two rather unrealistic assumptions: (1) we need to assume that

the singer is able to produce a series of pure tones of various frequencies (the larynx, in fact, produces a complex

periodic sound and not a sinusoid), and (2) the amplitudes of these pure tones are always exactly the same. The wine

glass will serve as the filter whose frequency response curve we wish to measure. As shown in Figure 3-20, we

attach a vibration meter to the wine glass, and the reading on this meter will serve as our measure of output

Figure 3-19. Frequency response curves for three optical filters. The lowpass filter on the left allows

low frequencies to pass through, while attenuating or blocking optical energy at higher frequencies.

The highpass filter in the middle has the opposite effect, allowing high frequencies to pass through,

while attenuating or blocking optical energy at lower frequencies. The bandpass filter on the right

allows a band of optical frequencies in the center of the spectrum to pass through, while attenuating or

blocking energy at higher and lower frequencies.

The Physics of Sound 20

amplitude for the filter. For the purpose of this example, will assume that the signal frequency needed to break the

glass is 500 Hz. We now ask the singer to produce a low frequency signal, say 50 Hz. Since this frequency is quite

remote from the 500 Hz needed to break the glass, the output amplitude measured by the vibration meter will be

quite low. As the singer gets closer and closer to the required 500 Hz, the measured output amplitude will increase

systematically until the glass finally breaks. If we assume that the glass does not break but rather reaches a

maximum amplitude just short of that required to shatter the glass, we can continue our measurement of the

frequency response curve by asking the singer to produce signals that are increasingly high in frequency. We would

find that the output amplitude would become lower and lower the further we got from the 500 Hz natural vibrating

frequency of the wine glass. The pattern that is traced by our measures of output amplitude at each signal frequency

would resemble the frequency response curve we saw earlier for green sunglasses; that is, we would see the

frequency response curve for a bandpass filter.

Additional Comments on Filters

Cutoff Frequency, Center Frequency, Bandwidth. The top panel of Figure 3-21 shows frequency response curves

for two lowpass filters that differ in a parameter called cutoff frequency. Both filters allow low frequencies to pass

through while attenuating high frequencies; the filters differ only in the frequency at which the attenuation begins.

The bottom panel of Figure 3-21 shows two highpass filters that differ in cutoff frequency. There are two additional

terms that apply only to bandpass filters. In our wineglass example above, the natural vibrating frequency of the

wine glass was 300 Hz. For this reason, when the frequency response curve is measured, we find that the wine glass

reaches its maximum output amplitude at 300 Hz. This is called the center frequency or resonance of the filter. It

is possible for two bandpass filters to have the same center frequency but differ with respect to a property called

Figure 3-20. Illustration of how the frequency response curve of a crystal wine glass

might be measured. Our singer produces a series of sinusoids that are identical in

amplitude but cover a wide range of frequencies. (This part of the example is

unrealistic: the human larynx produces a complex sound rather than a sinusoid.) The

gain of the wine glass filter can be traced out by measuring the amplitude of

vibration at the different signal frequencies.)

The Physics of Sound 21

bandwidth. Figure 3-22 shows two filters that differ in bandwidth. The tall, thin frequency response curve describes

a narrow band filter. For this type of filter, output amplitude reaches a very sharp peak at the center frequency and

drops off abruptly on either side of the peak. The other frequency response curve describes a wide band filter (also

called broad band). For the wide band filter, the peak that occurs at the resonance of the filter is less sharp and the

drop in output amplitude on either side of the center frequency is more gradual.

Fixed vs. Variable Filters. A fixed filter is a filter whose frequency response curve cannot be altered. For example,

an engineer might design a lowpass filter that attenuates at frequencies above 500 Hz, or a bandpass filter that passes

with a center frequency of 1,000 Hz. It is also possible to create a filter whose characteristics can be varied. For

example, the tuning dial on a radio controls the center frequency of a narrow bandpass filter that allows a single

radio channel to pass through while blocking channels at all other frequencies. The human vocal tract is an example

0 1000 2000 3000 4000

0.0

0.2

0.4

0.6

0.8

1.0

Frequency (Hz)

Gain

Lowpass Filters with Different

Cutoff Frequencies

0 1000 2000 3000 4000

0.0

0.2

0.4

0.6

0.8

1.0

Frequency (Hz)

Gain

Highpass Filters with Different

Cutoff Frequencies

Figure 3-21. Lowpass and highpass filters differing in cutoff frequency.

0 1000 2000 3000 4000

0.0

0.2

0.4

0.6

0.8

1.0

Frequency (Hz)

Gain

Bandpass Filters Differing

in Bandwidth

Narrow Band Filter

Wide Band Filter

Figure 3-22. Frequency response curves for two bandpass filters with identical center

frequencies but different bandwidths. Both filters pass a band of energy centered around

2000 Hz, but the narrow band filter is more selective than the wide band filter; that is,

gain decreases at a higher rate above and below the center frequency for the narrow band

filter than for the wide band filter

The Physics of Sound 22

of a variable filter of the most spectacular sort. For example: (1) during the occlusion interval that occurs in the

production of a sound like /b/, the vocal tract behaves like a lowpass filter; (2) in the articulatory posture for sounds

like /s/ and /sh/ the vocal tract behaves like a highpass filter; and (3) in the production of vowels, the vocal tract

behaves like a series of bandpass filters connected to one another, and the center frequencies of these filters can be

adjusted by changing the positions of the tongue, lips, and jaw. To a very great extent, the production of speech

involves making adjustments to the articulators that have the effect of setting the vocal tract filter in differ modes to

produce the desired sound quality. We will have much more to say about this in later chapters.

Frequency Response Curves vs. Amplitude Spectra. It is not uncommon for students to confuse a frequency

response curve with an amplitude spectrum. The axis labels are rather similar: an amplitude spectrum plots

amplitude on the y axis and frequency on the x axis, while a frequency response curve plots gain on the y axis and

frequency on the x axis. The apparent similarities are deceiving, however, since a frequency response curve and an

amplitude spectrum display very different kinds of information. The difference is that an amplitude spectrum

describes a sound while a frequency response curve describes a filter. For any given sound wave, an amplitude

spectrum tells us what frequencies are present with what amplitudes. A frequency response curve, on the other hand,

describes a filter, and for that filter, it tells us what frequencies will be allowed to pass through and what frequencies

will be attenuated. Keeping these two ideas separate will be quite important for understanding the key role played by

filters in both hearing and speech science.

Resonance

The concept of resonance has been alluded to on several occasions but has not been formally defined. The term

resonance is used in two different but very closely related ways. The term resonance refers to: (1) the phenomenon

of forced vibration, and (2) natural vibrating frequency (also resonant frequency or resonance frequency) To

gain an appreciation for both uses of this term, imagine the following experiment. We begin with two identical

tuning forks, each tuned to 435 Hz. Tuning fork A is set into vibration and placed one centimeter from tuning fork

B, but not touching it. If we now hold tuning fork B to a healthy ear, we will find that it is producing a 435 Hz tone

that is faint but quite audible, despite the fact that it was not struck and did not come into physical contact with

tuning fork A. The explanation for this "action-at-a-distance" phenomenon is that the sound wave generated by

tuning fork A forces tuning fork B into vibration; that is, the series of compression and rarefaction waves will

alternately push and pull the tuning fork, resulting in vibration at the frequency being generated by tuning fork A.

The phenomenon of forced vibration is not restricted to this "action-at-a-distance" case. The same effect can be

demonstrated by placing a vibrating tuning fork in contact with a desk or some other hard surface. The intensity of

the signal will increase dramatically because the tuning fork is forcing the desk to vibrate, resulting in a larger

volume of air being compressed and rarefied.3

Returning to our original tuning fork experiment, suppose that we repeat this test using two mismatched tuning

forks; for example, tuning fork A with a natural frequency of 256 Hz and tuning fork B with a natural vibrating

frequency of 435 Hz. If we repeat the experiment – setting tuning fork A into vibration and holding it one centimeter

from tuning fork B – we will find that tuning fork B does not produce an audible tone. The reason is that forced

vibration is most efficient when the frequency of the driving force is closest to the natural vibration frequency of the

object that is being forced to vibrate. Another way to think about this is that tuning fork B in these experiments is

behaving like a filter that is being driven by the signal produced by tuning fork A. Tuning forks, in fact, behave like

rather narrow bandpass filters. In the experiment with matched tuning forks, the filter was being driven by a signal

frequency corresponding to the peak in the filter's frequency response curve. Consequently, the filter produced a

great deal of energy at its output. In the experiment with mismatched tuning forks, the filter is being driven by a

signal that is remote from the peak in the filter's frequency response curve, producing a low amplitude output signal.

To summarize, resonance refers to the ability of one vibrating system to force another system into vibration.

Further, the amplitude of this forced vibration will be greater as the frequency of the driving force approaches the

natural vibrating frequency (resonance) of the system that is being forced into vibration.

3The increase in intensity that would occur as the tuning fork is placed in contact with a hard surface does not mean that additional energy is

created. The increase in intensity would be offset by a decrease in the duration of the tone, so the total amount of energy would not increase

relative to a freely vibrating tuning fork.

The Physics of Sound 23

Cavity Resonators

An air-filled cavity exhibits frequency selective properties and should be considered a filter in precisely the way

that the tuning forks and wine glasses mentioned above are filters. The human vocal tract is an air-filled cavity that

behaves like a filter whose frequency response curve varies depending on the positions of the articulators. Tuning

forks and other simple filters have a single resonant frequency. (Note that we will be using the terms "natural

vibrating frequency" and "resonant frequency" interchangeably.) Cavity resonators, on the other hand, can have an

infinite number of resonant frequencies.

A simple but very important cavity resonator is the uniform tube. This is a tube whose cross-sectional area is

the same (uniform) at all points along its length. A simple water glass is an example of a uniform tube. The method

for determining the resonant frequency pattern for a uniform tube will vary depending on whether the tube is closed

at both ends, open at both ends, or closed at just one end. The configuration that is most directly applicable to

problems in speech and hearing is the uniform tube that is closed at one end and open at the other end. The ear canal,

for example, is approximately uniform in cross-sectional area and is closed medially by the ear drum and open

0.0

0.2

0.4

0.6

0.8

1.0

Gain

500 1500 2500 3500 4500

17.5 cm Uniform Tube

0.0

0.2

0.4

0.6

0.8

1.0

Gain

437.5 1312.5 2187.5 3062.5 3937.5

20 cm Uniform Tube

0 1000 2000 3000 4000 5000

0.0

0.2

0.4

0.6

0.8

1.0

Frequency (Hz)

Gain

583.3 1750.0 2916.7 4083.3 5225.0

15 cm Uniform Tube

Figure 3-23. Frequency response curves for three uniform tubes open at one end and closed at the

other. These kinds of tubes have an infinite number of resonances at odd multiples of the lowest

resonance. As the figure shows, shortening the tube shifts all resonances to higher frequencies while

lengthening the tube shifts all resonances to lower frequencies.

The Physics of Sound 24

laterally. Also, in certain configurations the vocal tract is approximately uniform in cross-sectional area and is

effectively closed from below by the vocal folds and open at the lips. The resonant frequencies for a uniform tube

closed at one end are determined by its length. The lowest resonant frequency (F1) for this kind of tube is given by:

F1 = c/4L, where: c = the speed of sound

L = the length of the tube

For example, for a 17.5 cm tube, F1 = c/4L = 35000/70 = 500 Hz. This tube will also have an infinite number of

higher frequency resonances at odd multiples of the lowest resonance:

F1 = F1 . 1 = 500 Hz

F2 = F1 ⋅ 3 = 1,500 Hz

F3 = F1 . 5 = 2,500 Hz

F4 = F1 ⋅ 7 = 3,500 Hz

The frequency response curve for this tube for frequencies below 4000 Hz is shown in the solid curve in Figure

3-23. Notice that the frequency response curve shows peaks at 500, 1500, 2500, and 3500 Hz, and valleys in

between these peaks. The frequency response curve, in fact, looks like a number of bandpass filters connected in

series with one another. It is important to appreciate that what we have calculated here is a series of natural vibrating

frequencies of a tube. What this means is that the tube will respond best to forced vibration if the tube is driven by

signals with frequencies at or near 500 Hz, 1500 Hz, 2500 Hz, and so on. Also, the resonant frequencies that were

just calculated should not be confused with harmonics. Harmonics are frequency components that are present in the

amplitude spectra of complex periodic sounds; resonant frequencies are peaks in the frequency response curve of

filters.

We next need to see what will happen to the resonant frequency pattern of the tube when the tube length

changes. If the tube is lengthened to 20 cm:

F1 = c/4L = 35,000/80 = 437.5 Hz

F2 = F1 ⋅ 3 = 1,312.5 Hz

F3 = F1 ⋅ 5 = 2,187.5 Hz

F4 = F1 ⋅ 7 = 3,062.5 Hz

It can be seen that lengthening the tube from 17.5 cm to 20 cm has the effect of shifting all of the resonant

frequencies downward (see Figure 3-23). Similarly, shortening the tube has the effect of shifting all of the resonant

frequencies upward. For example, the resonant frequency pattern for a 15 cm tube would be:

F1 = c/4L = 35,000/60 = 583.3 Hz

F2 = F1⋅ 3 = 1,750 Hz

F3 = F1 ⋅ 5 = 2,916.7 Hz

F4 = F1 ⋅ 7 = 4,083.3 Hz

The general rule is quite simple: all else being equal, long tubes have low resonant frequencies and short tubes

have high resonant frequencies. This can be demonstrated easily by blowing into bottles of various lengths. The

longer bottles will produce lower tones than shorter bottles. This effect is also demonstrated every time a water glass

is filled. The increase in the frequency of the sound that is produced as the glass is filled occurs because the

resonating cavity becomes shorter and shorter as more air is displaced by water. This simple rule will be quite

useful. For example, it can be applied directly to the differences that are observed in the acoustic properties of

speech produced by men, women, and children, who have vocal tracts that are quite different in length.

Resonant Frequencies and Formant Frequencies

The term "resonant frequency" refers to natural vibrating frequency or, equivalently, to a peak in a frequency

response curve. For reasons that are entirely historical, if the filter that is being described happens to be a human

vocal tract, the term formant frequency is generally used. So, one typically refers to the formant frequencies of the

The Physics of Sound 25

vocal tract but to the resonant frequencies of a plastic tube, the body of a guitar, the diaphragm of a loudspeaker, or

most any other type of filter other than the vocal tract. This is unfortunate since it is possible to get the mistaken idea

that formant frequencies and resonant frequencies are different sorts of things. The two terms are, in fact, fully

synonymous.

The Decibel Scale

The final topic that we need to address in this chapter is the representation of signal amplitude using the decibel

scale. The decibel scale is a powerful and immensely flexible scale for representing the amplitude of a sound wave.

The scale can sometimes cause students difficulty because it differs from most other measurement scales in not just

one but two ways. Most of the measurement scales with which we are familiar are absolute and linear. The decibel

scale, however, is relative rather than absolute, and logarithmic rather than linear. Neither of these characteristics is

terribly complicated, but in combination they can make the decibel scale appear far more obscure than it is. We will

examine these features one at a time, and then see how they are put together in building the decibel scale.

Linear vs. Logarithmic Measurement Scales

Most measurement scales are linear. To say that a measurement scale is linear means that it is based on equal

additive distances. This is such a common feature of measurement scales that we do not give it much thought. For

example, on a centigrade (or Fahrenheit) scale for measuring temperature, going from a temperature of 90 o to a

temperature of 91 o involves adding one 1 o. One rather obvious consequence of this simple additivity rule is that the

difference in temperature between 10 o and 11 o is the same as the difference in temperature between 90 o and 91 o.

However, there are scales for which this additivity rule does not apply. One of the best known examples is the

Richter scale that is used for measuring seismic intensity. The difference in seismic intensity between Richter values

of 4.0 and 5.0, 5.0 and 6.0, 6.0 and 7.0 is not some constant amount of seismic intensity, but rather a constant

multiple. Specifically, a 7.0 on the Richter scale indicates an earthquake that is 10 times greater in intensity than an

earthquake that measures 6.0 on the Richter scale. Similarly, an 8.0 on the Richter scale is 10 times greater in

intensity than a 7.0. Whenever jumping from one scale value to the next involves multiplying by a constant rather

than adding a constant, the scale is called logarithmic. (The multiplicative constant need not be 10. See Box 3-2 for

an example of a logarithmic scale – an octave progression – that uses 2 as the constant.) Another way of making the

same point is to note that the values along the Richter scale are exponents rather than ordinary numbers; for

example, a Richter value of 6 indicates a seismic intensity of 10 6

, a Richter value of 7 indicates a seismic intensity of

10 7, etc. The Richter values can, of course, just as well be referred to as powers or logarithms since both of these

terms are synonyms for exponent. The decibel scale is an example of a logarithmic scale, meaning that it is based on

equal multiples rather than equal additive distances.

Absolute vs. Relative Measurement Scales

A simple example of a relative measurement scale is the Mach scale that is used by rocket scientists to measure

speed. The Mach scale measures speed not in absolute terms but in relation to the speed of sound. For example, a

missile at Mach 2.0 is traveling at twice the speed of sound, while a missile at Mach 0.9 is traveling at 90% of the

speed of sound. So, the Mach scale does not represent a measured speed (Sm) in absolute terms, but rather,

represents a measured speed in relation to a reference speed (Sm/Sr ). The reference that is used for the Mach scale is

the speed of sound, so a measured absolute speed can be converted to a relative speed on the Mach scale by simple

division. For example, taking 783 mph as the speed of sound, 1,200 mph = 1200/783 = Mach 1.53. The decibel scale

also exploits this relative measurement scheme. The decibel scale does not represent a measured intensity (I m) in

absolute terms, but rather, represents the ratio of a measured intensity to a reference intensity (Im/I r ).

The decibel scale is trickier than the Mach scale in one important respect. For the Mach scale, the reference is

always the speed of sound, but for the decibel scale, many different references can be used. In explaining how the

decibel scale works, we will begin with the commonly used intensity reference of 10 -12 w/m2 (watts per square

meter), which is approximately the intensity that is required for an average normal hearing listener to barely detect a

1,000 Hz pure tone. So, for our initial pass through the decibel scale, 10 -12 w/m2 will serve as I r , and will perform the

same function that the speed of sound does for the Mach scale. Table 3-1 lists several sounds that cover a very broad

range of intensities. The second column shows the measured intensities of those sounds, and the third column shows

the ratio of those intensities to our reference intensity. Whispered speech, for example, measures approximately 10 -8

The Physics of Sound 26

w/m2, which is 10,000 times more intense than the reference intensity (10 -8 /10 -12 = 10 4 = 10,000). The main point to

be made about column 3 is that the ratios become very large very soon. Even a moderately intense sound like

conversational speech is 1,000,000 times more intense than the reference intensity. The awkwardness of dealing

with these very large ratios has a very simple solution. Column 4 shows the ratios written in exponential notation,

and column 5 simplifies the situation even further by recording the exponent only. The term exponent and the term

logarithm are synonymous, so the measurement scheme that is expressed by the numbers in column 5 can be

summarized as follows: (1) divide a measured intensity by a reference intensity (in this case, 10 -12 w/m2), (2) take the

logarithm of this ratio (i.e., write the number in exponential notation and keep the exponent only). This method, in

fact, is a completely legitimate way to represent signal intensity. The unit of measure is called the bel, after A.G.

Bell, and the formula is:

bel = log10 I m/I r , where: I m = a measured intensity

I r = a reference intensity

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Table 3-1. Sound intensities and intensity ratios showing how the decibel scale is created. Column 2 shows the

measured intensities (Im) of several sounds. Column 3 shows the ratio of these intensities to a reference intensity of

10 -12 w/m2 . Column 4 shows the ratio written in exponential notation while column 5 shows the exponent only. The

last column shows the intensity ratio expressed in decibels, which is simply the logarithm of the intensity ratio

multiplied by 10.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Measured Ratio Ratio in Exponent Decibel

Sound Intensity (I m) (Im/Ir ) Exp. Not. (log 10) (10 x log 10)

Threshold 10 -12 w/m2 1 10 0 0 0

@ 1 kHz

Whisper 10 -8 w/m2 10,000 10 4 4 40

Conversational 10 -6 w/m2 1,000,000 10 6 6 60

Speech

City Traffic 10 -4 w/m2 100,000,000 10 8 8 80

Rock & Roll 10 -2 w/m2 10,000,000,000 10 10 10 100

Jet Engine 10 0 w/m2 1,000,000,000,000 10 12 12 120

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Legitimate or not, the bel finds its sole application in textbooks attempting to explain the decibel. For reasons that

are purely historical, the log10 of the intensity ratio is multiplied by 10, changing bel into the decibel (dB). As shown

in the last column of Table 3-1, this has the very simple effect of turning 4 bels into 40 decibels, 8 bels into 80

decibels, etc. The formula for the decibel, then, is:

dB IL = 10 log10 I m/I r , where:

I m = a measured intensity

I r = a reference intensity

The designation "IL" stands for intensity level, and it indicates that the underlying measurements are of sound

intensity and not sound pressure. As will be seen below, a different version of this formula is needed if sound

pressure measurements are used. The multiplication by 10 in the dB IL formula is a simple operation, but it can

The Physics of Sound 27

sometimes have the unfortunate effect of making the formula appear more obscure that it is. The decibel values that

are calculated, however, should be readily interpretable. For example, 30 dB IL means 3 factors of 10 more intense

than I r , 60 dB IL means 6 factors of 10 more intense than I r , and 90 dB IL means 9 factors of 10 more intense than I r .

Deriving a Pressure Version of the dB Formula

In a simple world, we would be finished with the decibel scale. The problem is that the formula is based on

measurements of sound intensity, but as a purely practical matter sound intensity is difficult to measure. Sound

pressure, on the other hand, is quite easy to measure. An ordinary microphone, for example, is a pressure sensitive

device. The problem, then, is that the decibel is defined in terms of intensity measurements, but the measurements

that are actually used will nearly always be measures of sound pressure. This problem can be addressed since there

is a predictable relationship between intensity (I) and pressure (E): intensity is proportional to pressure squared:

I ο⊂ Ε 2

Knowing this relationship allows us to create a completely equivalent version of the decibel formula that will work

when sound pressure measurements are used instead of sound intensity measurements. All we need to do is

substitute squared pressure measurements in place of the intensity measurements:

dB IL = 10 log10 I m/I r (intensity version of formula)

dB SPL= 10 log10 E 2m/E 2 (pressure version of formula)

The designation "SPL" stands for sound pressure level, and it indicates that measures of sound pressure have been

used and not measures of sound intensity. Although the dB SPL formula shown here will work fine, it will almost

never be seen in this form. The reason is that the formula is algebraically rearranged so that the squaring operation is

not needed. The algebra is shown below:

(1) dB IL = 10 log10 I m/I r (the intensity version of the formula)

(2) dB SPL = 10 log10 E 2m/E 2 (measures of E 2 replace measures of I because I ο⊂ E 2)

(3) dB SPL = 10 log10 (E m/E r ) 2 (a 2

/b 2 = (a/b) 2)

(4) dB SPL = 10 ⋅ 2 log10 E m/E r (this is the only tricky step: log a b = b log a)

(5) dB SPL = 20 log10 E m/E r (2 ⋅ 10 = 20)

With the possible exception of the fourth step,4 the algebra is straightforward, but the details of the derivation

are less important than the following general points:

1. The decibel formula is defined in terms of intensity ratios. The basic formula is;

dB IL = 10 log10 I m/I r .

2. While sound intensity is difficult to measure, sound pressure is easy to measure. It is therefore necessary to

derive a version of the decibel formula that works when measures of sound pressure are used instead of sound

intensity.

4 Step 4 is the only tricky part of derivation. The reason it works is that squaring a number and then taking a log is the same as taking

the log first, and then multiplying the log by 2. For example, note that the two calculations below produce the same result:

log 10 100 2 = log 10 10,000 = 4 (square first, then take the log)

log 10 100 2 = (log 10 100) x 2 = 2 x 2 = 4 (take the log, then multiply by 2)

The Physics of Sound 28

3. The derivation of the pressure version of the formula is based entirely on the fact that intensity is proportional to

pressure squared (I ο⊂ Ε 2). This allows measures of E 2 to replace measures of I, turning: dB IL = 10 log10 I m/I r into

dB SPL = 10 log10 E 2m/E r2 . A few algebra tricks are applied to turn this formula into the more aesthetically pleasing

final version: dB SPL = 20 log10 E m/E r .

4. The two versions of the formula are fully equivalent to one another (see Box 3-3).

This last point about the equivalence of the intensity and sound pressure versions of the formula is explained in

some detail in Box 3-3, but the basic point is quite simple. The pressure version of the dB formula was derived from

the intensity version of the formula through algebraic manipulations (based on this relationship: I ο⊂ Ε 2). The whole

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Box 3-2

HARMONICS, OCTAVES, LINEAR SCALES, AND LOGARITHMIC SCALES

As we will see when the decibel scale is introduced, there is an important distinction to be made between

linear scales, which are quite common, and logarithmic scales, which are less common but quite important.

This distinction can be illustrated by examining the difference between a harmonic progression and an octave

progression. Notice that in a harmonic progression, the spacing between the harmonics is always the same; that

is, the difference between H1 and H 2 is the same as the difference between H2 and H 3, and so on. This is because

increases in frequency between one harmonic and the next involve adding a constant, with the constant being

the fundamental frequency. For example:

H 1 500

H 2 1000 (add 500)

H 3 1500 (add 500)

H 4 2000 (add 500)

. .

. .

. .

To get from one scale value to another on an octave progression involves multiplying by a constant rather

than adding a constant. For example, an octave progression starting at 500 Hz looks like this:

O 1 500

O 2 1000 (multiply by 2)

O 3 2000 (multiply by 2)

O 4 4000 (multiply by 2)

. .

. .

. .

As a result of the fact that we are multiplying by a constant rather than adding a constant, the spacing is no

longer even (i.e., the spacing between O 1 and O 2 is 500 Hz, the spacing between O2 and O 3 is 1000 Hz, and so

on). The point to be made of this is that there are two fundamentally different kinds of scales: (1) scales like

harmonic progressions that are created by adding a constant, which are by far the more common, and (2) scales

like octave progressions that are created by multiplying by a constant. Scales that are created by adding a

constant are called linear scales, while scales that are created by multiplying by a constant are called

logarithmic scales. Note that for an octave progression, the multiplier happens to be 2, meaning that progressing

from one frequency to an octave above that frequency involves multiplication by 2. However, a logarithmic

scale can be built using any multiplier. We will return to the distinction between linear and logarithmic scales

when we talk about the decibel scale, and there we will see that a logarithmic scale is built around multiplication

by a constant value of 10 rather than 2.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The Physics of Sound 29

point of algebra, of course, is to keep the expression on the left equal to the expression on the right. The simple and

useful point that emerges from this is this: If an intensity meter shows that a given sound measures 60 dB IL, for

example, a pressure meter will show that the same sound measures exactly 60 dB SPL. (This may seem

counterintuitive due to the differences in the formulas, but see Box 3-3 for the explanation.) The equivalence of the

two versions of the dB formula greatly simplifies the interpretation of sound levels that are expressed in decibels.

References

The reference that is used for the Mach scale is always the speed of sound. One of the virtues of the decibel

scale is that any reference can be used as long as it is clearly specified. The only reference that has been mentioned

so far is 10 -12 w/m2

, which is roughly the audibility threshold for a 1,000 Hz pure tone. This is a standard reference

intensity, and unless otherwise stated it should be assumed that this is used when a signal level is reported in dB IL.

The standard reference that is used for dB SPL is 20 μPa, so when a signal level is reported in dB SPL it should be

assumed that this reference is used unless otherwise stated.5

Many references besides these two standard references can be used. For example, suppose that a speech signal

is presented to a listener at an average level of 3500 μPa in the presence of a noise signal whose average sound

pressure is 1400 μPa. The speech-to-noise ratio (S/N) can be represented on a decibel scale, using the level of the

speech as E m and the level of the noise as E r :

dB s/n = 20 log10 E m/ E r

= 20 log10 3500/1400

= 20 log10 2.5

= 20 (0.39794)

= 7.96 dB

To take one more example, assume that a voice patient prior to treatment produces sustained vowels that

average 2300 μPa. Following treatment the average sound pressures increase to 8890 μPa. The improvement in

sound pressure (post-treatment relative to pre-treatment) can be represented on a decibel scale:

dB Improvement = 20 log10 E post/E pre

= 20 log10 8890/2300

= 20 log10 (3.86522)

= 20 (0.58717)

= 11.74 dB

A final example can be used to make the point that the decibel scale can be used to represent intensity ratios for

any type of energy, not just sound. Bright sunlight has a luminance measuring 100,000 cd/m2 (candela per square

meter). Light from a barely visible star, on the other hand, has a luminance measuring 0.0001 cd/m2. We can now

ask how much more luminous bright sunlight is in relation to barely visible star light, and the dB scale can be used

to represent this value. Since the underlying physical quanities here are measures of electromagnetic intensity, we

want the intensity version of the formula rather than the pressure version.

dB = 10 log10 I sunlight/I starlight

= 10 log10 100000/0.0001

= 10 log10 10 5/10 -4

= 10 log10 10 9 (division is done by subtracting exponents: 5 – (-4) = 9)

= 10 (9)

= 90 dB

5The standard pressure reference for dB SPL is sometimes given as 0.0002 dynes/cm2 rather than 20 μPa. These two sound pressures are

identical, however, in exactly the same sense that 4 quarts and 1 gallon are identical. Likewise, the standard reference for dB IL is often given as

10 -16 w/cm2 instead of 10 -12 w/m2

. These two intensities are also identical.

The Physics of Sound 30

The fact that we are measuring light rather than sound makes no difference: a decibel is 10 log10 I m/I r (or,

equivalently, 20 log10 E m/E r ), regardless of whether the energy comes from sound, light, electrical current, or any

other type of energy.

dB Hearing Level (dB HL)

The dB Hearing Level (dB HL) scale was developed specifically for testing hearing sensitivity for pure tones

of different frequencies. The sound-level dials on clinical audiometers,6 for example, are calibrated in dB HL rather

than dB SPL. To understand the motivation for the dB HL scale examine Figure 3-24, which shows the sound level (in

dB SPL) required for the average, normal-hearing listener to barely detect pure tones at frequencies between 125 and

8000 Hz. This is called the audibility curve and the simple but very important point to notice about this graph is

that the curve is not a flat line; that is, the ear is clearly more sensitive at some frequencies than others. The

differences in sensitivity are quite large in some cases. For example, the average normal-hearing listener will barely

detect a 1000 Hz pure tone at 7 dB SPL, but at 125 Hz the sound level needs to be cranked all the way up to 45 dB SPL,

an increase in intensity of nearly 4000:1. Now suppose we were to test pure-tone sensitivity using an audiometer that

is calibrated in dB SPL. Imagine that a listener barely detects a 1000 Hz pure tone at 25 dB SPL. Does this listener have

a hearing loss, and if so how large? The only way to answer this question is to consult the data in Figure 3-24, which

shows that the threshold of audibility for the average normal hearing listener at 1000 Hz is 7 dB SPL. This means that

the hypothetical listener in this example has a hearing loss of 25-7 = 18 dB. Suppose further that the same listener

detects a 250 Hz tone at 20 dBSPL. The table in Figure 3-24 shows that normal hearing sensitivity at 250 Hz is 25.5

dB SPL, meaning that the listener has slightly better than normal hearing at this frequency. As a final example,

imagine that this listener barely detects a 500 Hz tone at 30 dB SPL. Since the table shows that normal hearing

sensitivity at 500 Hz is 11.5 dB SPL, the listener has a hearing loss of 30.0-11.5 = 18.5 dB. The simple point to be

made about these examples is that, with an audiometer dial that is calibrated in dB SPL, it is not possible to determine

whether a listener has a hearing loss, or to measure the size of that loss, without doing some arithmetic involving the

normative data in Figure 3-24. The dB HL scale, however, provides a simple solution to this problem that avoids this

arithmetic entirely. The solution involves calibrating the audiometer in such a way that, when the level dial is set to

0 dB HL, sound level is set to the threshold of audibility for the average normal-hearing listener for that signal

frequency. For example, when the level dial is set to 0 dB HL at 125 Hz the level of tone will be 45 dB SPL – the

threshold of audibility for the average normal hearing listener at this frequency. Now if a listener barely detects the

125 Hz tone at 0 dB HL, no arithmetic is needed; the listener has normal hearing at this frequency. Further, if the

listener barely detects this 125 Hz tone at 40 dB HL, for example, the listener must have a 40 dB loss at this frequency

– and again it is not necessary to consult the data in Figure 3-24. Similarly, when the level dial is set to 0 dB HL at

250 Hz the level of the tone will be 25.5 dB SPL, which is the audibility threshold at 250 Hz. If this tone is barely

detected at 0 dB HL, the listener has normal hearing at this frequency. However, if the tone is not heard until the dial

is increased to 50 dB dB HL, for example, the listener has a 50 dB hearing loss at this frequency. The same system is

used for all signal frequencies: in all cases, the 0 dB HL reference is not a fixed number as it is for dB SPL (a constant

value of 20 μPa, no matter what the signal frequency is) or dB IL (a constant value of 10 -12 watts/m2, again

independent of signal frequency), but rather a family of numbers. In each case the reference for the dB HL scale is the

threshold of audibility for an average, normal-hearing listener at a particular signal frequency. What this means is

that values in dB HL are a fixed distance above the audibility curve, although they may be very different levels in

dB SPL. For illustration, Figure 3-25 shows the audibility curve (the filled symbols) and, above that in the unfilled

symbols, a collection of values that all measure 30 dB HL. Although the sound levels on the 30 dB HL curve vary

considerably in dB SPL (i.e. measured using 20 μPa as the reference), every data point on this curve is a constant 3

factors of 10, or 30 dB, above the audibility curve. The value of 30 dB in this figure is just an example. All values in

dB HL and dB SPL are interpreted in the same way: 50 dB SPL means that the signal being measured is 100,000 times

(i.e., 5 factors of 10) more intense than the fixed reference of 20 μPa, independent of frequency; 50 dB HL, on the

other hand, means that the signal being measured is 100,000 times (again, 5 factors of 10) more intense than a tone

that is barely audible to a normal-hearing listener at that signal frequency. Similarly, 20 dB SPL means that the signal

is 20 dB (2 factors of 10) more intense than the fixed reference of 20 μPa, while 20 dB HL means that the signal is 20

dB (again, 2 factors of 10) above the audibility curve.

6A clinical audiometer is an instrument with, among other things, one dial (for each ear) that controls pure-tone frequency and another dial

that controls the intensity of the tone. The listener is asked to raise a hand when the tone is barely audible.

The Physics of Sound 31

Summary

The decibel is a powerful scale for representing signal amplitude. The scale has two important properties: (1)

similar to the Mach scale, it represents signal level not in absolute terms but as a measured level divided by a

reference level; and (2) like the Richter scale, the dB scale is logarithmic rather than linear, meaning that it is based

on equal multiplicative distances rather than equal additive distances. While the decibel is defined in terms of

intensity ratios, for practical reasons, measures of sound pressure are far more common than measures of sound

intensity. Consequently, a version of the decibel formula was derived that makes use of pressure ratios rather than

intensity ratios. The derivation was based on the fact that intensity is proportional to pressure squared. The two

versions of the decibel formula (dB IL = 10 log 10 I m/I r and dB SPL = 20 log 10 E m/E r ) are fully equivalent, meaning that

if a sound measures 60 dB IL that same sound will measure 60 dB SPL. Unlike the Mach scale, which always uses the

speed of sound as a reference, any number of references can be used with the decibel scale. The standard reference

for the dB IL scale is 10 -12 w/m2 and the standard reference for the dB SPL scale is 20 μPa. However, any level can be

used as a reference as long as it is specified. The dB HL scale, widely used in audiological assessment, was developed

specifically for measuring sensitivity to pure tones of difference frequencies. The reference that is used for the dB HL

scale is the threshold of audibility at a particular signal frequency for the average, normal-hearing listener. Sound

levels in dB SPL and dB HL are interpreted quite differently. For example, a pure tone measuring 40 dB SPL is 4 factors

of 10 (i.e., 40 dB) greater than the fixed SPL reference of 20 μPa, while a pure tone measuring 40 dB HL is 4 factors

of 10 (again, 40 dB) greater than a tone of that same frequency that is barely audible to an average, normal-hearing

listener.

Frequency Threshold

125 45.0

250 25.5

500 11.5

750 8.0

1000 7.0

1500 6.5

2000 9.0

3000 10.0

4000 9.5

6000 15.5

8000 13.0

Figure 3-24. The threshold of audibility for the average, normal-hearing listener for pure tones varying between

125 and 8000 Hz. The audibility threshold is the sound level in dBSPL that is required for a listener to barely detect

a tone. Values on this curve are shown in the table to the right. The most important point to note about this graph

is that the curve is not flat, meaning that the ear is more sensitive at some frequencies than others. In particular,

the ear is more sensitive in a range of mid-frequencies between about 1000 and 4000 Hz than it is at lower and

higher frequencies. The complex shape of this curve provides the underlying motivation for the dB HL scale. See

text for details.

The Physics of Sound 32

Figure 3-25. The lower function is the audibility curve – the sound level in dB SPL that is required for an average

normal hearing listener to barely detect pure tones of different frequencies. The upper function shows sound levels for

a set of tones that all measure 30 dB HL. These tones vary quite a bit in dB SPL (i.e., relative to the constant value of 20

μPa) but in all cases the tones are a constant 3 factors of 10 in intensity (i.e., 30 dB) above the audibility curve.

The Physics of Sound 33

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Box 3-3

THE EQUIVALENCE OF THE INTENSITY AND PRESSURE

VERSIONS OF THE DECIBEL FORMULA

One fact about the two versions of the dB formula that is not always well understood is that the dB IL and

dB SPL formulas are fully equivalent. By "fully equivalent" we mean the following: suppose that a sound intensity

meter is used to measure the level of some sound, and we find that this sound is 1,000 times more intense than

the standard intensity reference of 10 -12 w/m2. The sound would then measure 30 dB IL (10 log10 1,000 = 10 (3) =

30 dB IL). Now suppose that we put the sound intensity meter away and use a sound pressure meter to measure

the same sound. You might think that the sound would measure 60 dB SPL since now we are multiplying by 20

instead of 10, but the trick is that the ratio is no longer 1,000. Recall that intensity is proportional to pressure

squared, which means that pressure is proportional to the square root of intensity. This means that if the intensity

ratio is 1,000, the pressure ratio must be the square root of 1,000, or 31.6. So, the formula now becomes 20 log

31.6 = 20 (1.5) = 30 dB SPL, which is exactly what we obtained originally. It will always work out this way: if a

sound measures 50 dB IL, that same sound will measure 50 dB SPL.

Table 3-2 might help to make this more clear. The first column shows an intensity ratio, the second column

shows the corresponding pressure ratio (this is always the square root of the intensity ratio), the third column

shows the dB IL value (10 log of the intensity ratio), and the fourth column shows dB SPL value (20 log of the

pressure ratio). As you can see, they are always the same.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Table 3-2. Intensity ratios, equivalent pressure ratios, dB IL values and

dB SPL values showing the equivalence of the intensity and pressure versions

of the dB formula.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Intensity Pressure dB IL dB SPL

Ratio Ratio (10 log10 I m/I r ) (20 log10 E m/E r )

10 3.16 10.00 10.00

20 4.47 13.01 13.01

40 6.32 16.02 16.02

50 7.07 16.99 16.99

60 7.75 17.78 17.78

70 8.37 18.45 18.45

80 8.94 19.03 19.03

90 9.49 19.54 19.54

100 10.00 20.00 20.00

200 14.14 23.01 23.01

300 17.32 24.77 24.77

400 20.00 26.02 26.02

500 22.36 26.99 26.99

1000 31.62 30.00 30.00

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The Physics of Sound 34

Study Questions: Physical Acoustics

1. Explain the basic processes that are involved in the propagation of a sound wave.

2. Draw time- and frequency-domain representations of simple periodic, complex periodic, complex aperiodic, and

transient sounds.

3. Draw time- and frequency-domain representations of two complex periodic sounds with different fundamental

frequencies.

4. Draw time-domain representations of two simple periodic sounds with the same frequency and phase, but different

amplitudes.

5. Draw time-domain representations of two simple periodic sounds with the same frequency and different amplitudes but

different phases.

6. Draw amplitude spectra of two sounds with the same fundamental frequencies but different spectrum envelopes.

7. Draw amplitude spectra of two sounds with different fundamental frequencies but similar spectrum envelopes.

8. Calculate signal frequencies for sinusoids with the following values:

a. period = 0.34 s

b. period = 2 s

c. period = 10 ms

d. period = 2 ms

e. wavelength = 20 cm

f. wavelength = 100 cm

Answers:

a. f = 1/0.34 = 2.94 Hz

b. f = 1/2 = 0.5 Hz

c. f = 1/0.01 = 100 Hz

d. f = 1/.002 = 500 Hz

e. f = c/WL (speed of sound/wavelength) = 35000/20 = 1750 Hz

f. f = c/WL (speed of sound/wavelength) = 35000/100 = 350 Hz

9. Calculate the three lowest resonant frequencies of the following uniform tubes that are closed at one end and open at

the other end:

a. 10 cm

b. 30 cm

c. 40 cm

Answers:

a. wavelength of lowest resonance = 40 cm (10 x 4)

f = 35000/40 = 875

R1 = 875 (R1 = frequency of resonance number 1)

R2 = 2625

R3 = 4375

b. wavelength of lowest resonance = 120 cm (30 x 4)

f = 35000/120 = 291.7

The Physics of Sound 35

R1 = 291.7

R2 = 875.0

R3 = 1458.3

c. wavelength of lowest resonance = 160 cm (40 x 4)

f = 35000/160 = 218.75

R1 = 218.75

R2 = 656.25

R3 = 1093.75

10. Show what the frequency-response curves look like for the tubes in the problem above.

11. A complex periodic signal has a fundamental period of 4 msec. What is the fundamental frequency of the signal? At

what frequencies would we expect to find energy?

12. How are the terms octave and harmonic different?

13. Give examples of the following kinds of graphs, being sure to label both axes:

a. amplitude spectrum

b. phase spectrum

c. frequency-response curve

d. time-domain representation

14. Give a brief explanation of the basic idea behind Fourier analysis. What is the input to Fourier analysis and what kind

of output(s) does it produce?

15. Draw and label frequency-response curves for low-pass, high-pass, and band-pass filters.

16. What parameters control the frequency of vibration of a spring and mass system?

17. Draw the time domain representation of one cycle of a sinusoid as variations in instantaneous air pressure over time

and one cycle of that same sinusoid as variations in instantaneous velocity over time.

18. How, if at all, are the terms resonant frequency and harmonic different?

19. How, if at all, are the terms resonant frequency and formant different?

20. A harmonic is a peak in: (a) a frequency response curve, (b) an amplitude spectrum, or (c) either a frequency response

curve or an amplitude spectrum.

21. A resonance is a peak in: (a) a frequency response curve, (b) an amplitude spectrum, or (c) either a frequency response

curve or an amplitude spectrum.

22. A formant is a peak in: (a) a frequency response curve, (b) an amplitude spectrum, or (c) either a frequency response

curve or an amplitude spectrum.

23. A frequency response curve describes a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

24. An amplitude spectrum describes a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

The Physics of Sound 36

Frequency Response Problems

The Physics of Sound 37

Answers to Frequency Response Problems

The Physics of Sound 38

Decibel Study Questions

1. What reference is used for the dB IL scale?

2. What reference is used for the dB SPL scale?

3. What reference is used for the dB HL scale?

4. What reference is used for the dB SL scale?

5. A listener barely detects a 125 Hz pure tone at 55 dB SPL. Does this listener have a hearing loss at 125 Hz, and if

so, what is the size of the hearing loss?

6. A listener barely detects a 1,000 Hz pure tone at 55 dB SPL. Does this listener have a hearing loss at 1,000 Hz,

and if so, what is the size of the hearing loss?

7. A listener barely detects a 125 Hz pure tone at 55 dB HL. Does this listener have a hearing loss at 125 Hz, and if

so, what is the size of the hearing loss?

8. A listener barely detects a 1,000 Hz pure tone at 55 dB HL. Does this listener have a hearing loss at 1,000 Hz,

and if so, what is the size of the hearing loss?

9. 60 dB SPL at 1,000 Hz means \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ more intense than \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

10. 60 dB IL at 1,000 Hz means \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ more intense than \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

11. 60 dB HL at 1,000 Hz means \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ more intense than \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

12. The reference that is used for the dB SPL scale is:

a. a number

b. a sentence

13. If the answer to the question above is a number, give the number; if it’s a sentence, give the sentence.

14. The reference that is used for the dB HL scale is:

a. a number

b. a sentence

15. If the answer to the question above is a number, give the number; if it’s a sentence, give the sentence.

16. A specific individual has a 70 dB hearing loss in the left ear at 1,000 Hz. A 90 dB HL, 1,000 Hz tone that is

presented to this listener’s left ear would measure \_\_\_\_\_\_ dB SL.

17. A sound measures 42 dB IL. On the dB SPL scale, that same sound will measure:

a. 84 dB SPL because with the dB SPL formula we are now are multiplying the ratio by 20 instead of 10.

b. 42 dB SPL because the two versions of the formula are equivalent

18. A sound measures 60 dB IL. (a) The measured intensity (I M) must therefore be \_\_\_\_\_\_\_\_\_ times

greater than the reference intensity (I R). (b) What would the pressure ratio (E M/E R) be for this same sound? (c)

Do the arithmetic to show what this sound would measure in dB SPL.

The Physics of Sound 39

19. A sound measures 40 dB IL. (a) The measured intensity (I M) must therefore be \_\_\_\_\_\_\_\_\_ times

greater than the reference intensity (I R). (b) What would the pressure ratio (E M/E R) be for this same sound? (c)

Do the arithmetic to show what this sound would measure in dB SPL.

20. On the graph below, put a mark at: (a) 3,000 Hz, 20 dB SPL, and (b) 3,000 Hz, 20 dB HL (the

grid lines on the y axis are spaced at 2 dB intervals).

Frequency Threshold

in Hz in dB SPL

125 45.0

250 25.5

500 11.5

750 8.0

1000 7.0

1500 6.5

2000 9.0

3000 10.0

4000 9.5

6000 15.5

8000 13.0

The Physics of Sound 40

Answers to Decibel Study Questions

1. 10 -12 watts/m2

2. 20 μPa (or, equivalently, 0.0002 dynes/cm2)

3. The threshold of audibility for an average, normal-hearing listener at a particular signal frequency.

4. 3. The threshold of audibility for a particular listener at a particular signal frequency.

5. Consulting the attached figure and table showing the audibility curve for average, normal-hearing listeners, we

find that the threshold of audibility at 125 Hz is 45 dB SPL. A listener who barely detected a 125 Hz tone at 55

dB SPL would therefore have hearing loss of 55-45=10 dB; that is, the hearing sensitivity of this listener would be

10 dB worse than normal.

6. Consulting the attached figure and table showing the audibility curve for average, normal-hearing listeners, we

find that the threshold of audibility at 1,000 Hz is 7 dB SPL. A listener who barely detected a 1,000 Hz tone at

55 dB SPL would therefore have a hearing loss of 55-7=48 dB; that is, the hearing sensitivity of this listener

would be 48 dB worse than normal.

7. The reference for dB HL is the audibility threshold, so this listener would have a 55 dB hearing loss at 125 Hz.

There is no need to consult the table.

8. The reference for dB HL is the audibility threshold, so this listener would have a 55 dB hearing loss at 1,000 Hz.

There is no need to consult the table.

9. 6 factors of 10 (i.e., 1,000,000 times) more intense than 20 μPa)

10. 6 factors of 10 (i.e., 1,000,000 times) more intense than 10 -12 watts/m2

11. 6 factors of 10 (i.e., 1,000,000 times) more intense than a 1,000 Hz tone that is barely audible to an average,

normal-hearing listener.

12. a number

13. 20 μPa

14. a sentence

15. The threshold of audibility for an average, normal-hearing listener at a particular signal frequency.

16. 20 dB SL. The reference for the dB SL (SL=sensation level) is the threshold of audibility for a specific listener. So,

what we want to know here very simply is where this 90 dB HL tone is in relation to this particular listener’s

threshold. This listener has a 70 dB hearing loss at this frequency, so the 90 dBHL tone, which would be 90 dB

above a normal-hearing listener’s threshold, is only 20 dB above this particular listener’s threshold.

17. 42 dB SPL: The pressure version of the formula was derived from the intensity version through algebraic

manipulations, so they have to be equivalent to one another. The next problem was designed to illustrate how

this can be the case.

18. (a) 1,000,000 times (6 factors of 10) more intense than I R. (b) If the intensity ratio is 1,000,000, the pressure

ratio has to be the square root of 1,000,000, which is 1,000. (c) dB SPL = 20 log 1,000 = 20 . 3 = 60 dB SPL. This is

exactly what we got for the same sound measured in dB IL. It will always be the same. If a sound measures 60

dB IL, that same sound will measure 60 dB SPL.

The Physics of Sound 41

19. (a) 10,000 times (4 factors of 10) more intense than I R. (b) If the intensity ratio is 10,000, the pressure ratio has

to be the square root of 10,000, which is 100. (c) dB SPL = 20 log 100 = 20 . 2 = 40 dB SPL. This is exactly what

we got for the same sound measured in dB IL. It will always be the same. If a sound measures 40 dB IL, that same

sound will measure 40 dB SPL.

20. See below. The lower of the two marks is 20 dB (2 factors of 10) above the constant reference line of 20 μPa.

The higher of the two marks is 20 dB (also 2 factors of 10) above the curvey line, which is the threshold of

audibility for the average normal-hearing listener.

The Physics of Sound 42

The Physics of Sound 43

The Physics of Sound 44

A Tutorial on Digital Sound Synthesis Techniques

Author(s): Giovanni de Poli

Source:

Computer Music Journal, Vol. 7, No. 4 (Winter, 1983), pp. 8-26

Published by: The MIT Press

Stable URL: https://www.jstor.org/stable/3679529

Accessed: 08-06-2020 20:13 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide

range of content in a trusted digital archive. We use information technology and tools to increase productivity and

facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at

https://about.jstor.org/terms

The MIT Press is collaborating with JSTOR to digitize, preserve and extend access to

Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Giovanni De Poli

Centro di Sonologia Computazionale

Istituto di Elettrotecnica ed Elettronica

Universith di Padova, Italy

A Tutorial on Digital

Sound Synthesis

Techniques

Introduction

Progress in electronics and computer technology

has led to an ever-increasing utilization of digital

techniques for musical sound production. Some of

these are the digital equivalents of techniques em-

ployed in analog synthesizers and in other fields of

electrical engineering. Other techniques have been

specifically developed for digital music devices and

are peculiar to these.

This paper introduces the fundamentals of the

main digital synthesis techniques. Mathematical

developments have been restricted in the exposi-

tion and can be found in the papers listed in the

references. To simplify the discussion, whenever

possible, the techniques are presented with refer-

ence to continuous signals.

Sound synthesis is a procedure used to produce a

sound without the help of acoustic instruments. In

digital synthesis, a sound is represented by a se-

quence of numbers (samples). Hence, a digital syn-

thesis technique consists of a computing procedure

or mathematical formula, which computes each

sample value.

Normally, the synthesis formula depends on

some values, that is, parameters. Frequency and

amplitude are examples of such parameters. Param-

eters can be constant or slowly time variant during

the sound. Time-variant parameters are also called

control functions.

Synthesis techniques can be classified as (1) gen-

eration techniques (Fig. la), which directly produce

the signal from given data, and (2) transformation

techniques (Fig. Ib), which can be divided into two

stages, the generation of one or more simple signals

and their modification. Often, more or less elabo-

rate combinations of these techniques are employed.

Fixed-Waveform Synthesis

In many musical sounds, pitch is a characteristic to

which we are quite sensitive. In examining the tem-

poral shape of pitched sounds, we see a periodic rep-

etition of the waveform without great variations.

The simplest synthesis method attempts to re-

produce this characteristic, generating a periodic

signal through continuous repetition of the wave-

form. This method is called fixed-waveform

synthesis.

The technique is carried out by a module called

an oscillator (Fig. 2), which repeats the waveform

with a specified amplitude and frequency. In certain

cases, the waveform is characteristic of the os-

cillator and cannot be changed. But often it can be

chosen in a predetermined set of options or given

explicitly when required.

Usually, in digital synthesis the waveform value

at a particular instant is not computed anew for

each sample. Rather, a table, containing the period

values computed in equally spaced points, is built

beforehand. Obviously, the more numerous the

points in the table, the better the approximation

will be. To produce a sample, the oscillator requires

the waveform value at that precise instant. It cy-

clically searches the table to get the point nearest

to the required one. Sometimes a finer precision is

achieved by interpolation between two adjacent

points.

The distance in the table between two samples

read at subsequent instants is called the sam-

pling\_increment. The sampling \_increment is pro-

portional to the frequency f of the generated signal

according to the following formula (Mathews 1969):

N

samplingincrement = SRf,

where N is the table length and SR the sampling

rate.

In the oscillator, the frequency is usually speci-

Computer Music Journal, Volume 7, No. 4,

Winter, 1983, 0148-9267/83/040008-19 $04.00/0,

? 1983 Massachusetts Institute of Technology.

8 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 1. Classification of

synthesis techniques. Gen-

eration techniques (a) and

transformation techniques

(b).

Fig. 2. Fixed-waveform

synthesis oscillator.

(a) Parameters

'[ M SoundGeneration S

signal

Complex

(b) Parameters sound

I signal

Generation Transformation

Simple

signals

A(t) f(t)

s(t)

fled as a sampling- increment and the algorithm

that realizes it is as follows:

signal [t] := amplitude \* table [phase],

(Relation 1)

and

phase := mod(n, phase + samplingincrement),

(Relation 2)

where

Table contains one period of the waveform;

Phase is the theoretical position in the table of

the sample to be extracted at the instant; and

Amplitude is the signal amplitude.

Relation 2 computes the phase value in the subse-

quent instant, approximating the frequency integra-

tion by a summation. The modulus operation keeps

the phase inside the table length n.

It is noteworthy that the signal generated in this

way is an approximation of the desired one (Mail-

liard 1976). The approximation depends on the

table length, the interpolation method, and the sig-

nal frequency. For a sufficiently long table, it is

fully satisfactory.

The results of fixed-waveform synthesis are of

poor musical quality, as the sound does not present

any variation along its duration. This technique can

be changed by allowing the amplitude to vary in

time. In real sounds, the amplitude is rarely con-

stant: it starts from zero, reaches a maximum after

a certain time (attack), remains nearly constant

(steady state) and, after a certain evolution, it re-

turns to zero (decay). This sequence of amplitude

behavior is called the envelope. Thus, when the

amplitude varies according to a control function,

we have fixed-waveform synthesis with an ampli-

tude envelope.

The envelope can be generated in many ways.

In software-based synthesis, the most frequent

method uses an oscillator module, seen previously,

using a very low frequency equal to the inverse of

the duration. In this case, it performs a single cycle

and its waveform corresponds to the amplitude

envelope.

By carefully analyzing natural periodic sounds, it

has been shown that even the most stable ones con-

tain small frequency fluctuations. These improve

the sound quality and avoid unpleasant beatings

when more sounds are present at the same time.

The fixed-waveform technique can also be modi-

fied so that the oscillator frequency can slowly vary

around a value. This enables the production of a

tremolo and, with wider variations, of a glissando

or melodies.

The combination of these two variations consti-

tutes fixed-waveform synthesis with time-varying

amplitude and frequency. The waveform is fixed,

while the amplitude and frequency vary. The par-

tials are exact multiples of the fundamental, and

they all behave the same.

Fixed-waveform synthesis is realized rather sim-

ply. Hence, it is often employed when good sound

quality is not required. The constant waveform

gives the sound a mechanical, dull, and unnatural

character, which soon annoys the audience. Thus,

DePoli 9

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

in musical applications, fixed-waveform synthesis

is not very effective when used alone. It is em-

ployed for its simplicity when timbral variety is not

required, for example, for real-time synthesis on

very limited hardware.

For economy, other methods of generating wave-

forms that do not use tables or multiplications have

been devised. The simplest generates a square or

(more generally) a rectangular wave, alternating

sequences of positive and negative samples of the

same value. The frequencies that can be obtained

are submultiples of the sampling rate.

A sawtooth signal can also be generated by an ac-

cumulator to which a constant value is continu-

ously added. The output increases linearly until it

overflows and starts from the beginning. The signal

frequency is proportional to the constant value.

This method is used to produce linearly variable

control signals. Every time the additive constant

changes, the slope changes. Hence, functions com-

posed of straight segments, such as envelopes, can

be obtained.

This technique has been generalized recently by

Mitsuhashi (1982a). A polynomial of degree N can

be generated by putting N accumulators in cascade.

The accumulators are initialized by the value of the

forward differences, in decreasing order, of the poly-

nomial to be generated (Cerruti and Rodeghiero

1983). The waveforms obtained exhibit great vari-

ety and, in certain conditions, they are periodic.

Granular Synthesis

The technique of fixed-waveform synthesis pro-

duces rather static sounds in time. Yet a fundamen-

tal characteristic of musical sound is its timbral

evolution in time. A sound can be thought of as a

sequence of elementary sounds of constant dura-

tion, analogous to a film, in which a moving image

is produced by a sequence of images.

In computer music, the elementary sounds are

called grains, and the technique of exploiting this

facility is granular synthesis (Roads 1978). The

grains can be produced by a simple oscillator or by

other methods. The duration of each grain is very

short, on the order of 5-20 msec.

There are two ways to implement granular syn-

thesis. The first is to organize the grains into

frames, like the frames of a film. At each frame,

the parameters of all the grains are updated. This is

the approach sketched by Xenakis (1971). The sec-

ond way involves scattering the grains within a

mask, which bounds a particular frequency/ampli-

tude/time region. The density of the grains may

vary within the mask. This is the method imple-

mented by Roads (1978).

A problem with granular synthesis is the large

amount of parameter data to be specified. In some

other types of synthesis (additive and subtractive, to

be discussed shortly), these data can be obtained by

analyzing natural sounds. However, no analysis sys-

tem for granular synthesis has been developed. An-

other possibility is to obtain the parameter data

from an interactive composition system, which al-

lows the composer to work with high-level musical

concepts while automatically generating the thou-

sands of grain parameters needed.

Additive Synthesis

In additive synthesis, complex sounds are produced

by the superimposition of elementary sounds. In

certain conditions, the constituent sounds fuse

together and the result is perceived as a unique

sound. This procedure is used in some traditional

instruments, too. In an organ, the pipes generally

produce relatively simple sounds; to obtain a richer

spectrum in some registers, notes are created by

using more pipes sounding at different pitches at

the same time. The piano uses a different proce-

dure. Many notes are obtained by the simultaneous

percussion of two or three strings, each oscillating

at a slightly different frequency. This improves the

sound intensity and enriches it with beatings.

In order to choose the elementary sounds of addi-

tive synthesis, we first note that the Fourier analy-

sis model enables us to analyze sounds in a way

similar to the human ear and so to extract param-

eters that are perceptually significant. When we

analyze a real, almost-periodic sound, we imme-

diately notice that each partial amplitude is not

proportionally constant, but that it varies in time

10 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 3. Additive synthesis.

AI(t) f,(t) A2(t) f2(t) AM(t) fM(t)

s(t)

according to different laws. In the attack portion of

a note, some partials, which in the steady state are

negligible, are often significant.

Any almost-periodic sound can be approximated

as a sum of sinusoids. Each sinusoid's frequency is

nearly multiple that of the fundamental, and each

sinusoid evolves in time. For higher precision, the

frequency of each component can be considered as

slowly varying. Thus, additive synthesis consists of

the addition of some sinusoidal oscillators, whose

amplitude, and at times frequency, is time varying

(Fig. 3).

The additive-synthesis technique also provides

good reproduction of nonperiodic sounds, present-

ing in the spectrum the energy concentrated in

some spectral lines. For example, Risset (1969) imi-

tated a bell sound by summing sinusoidal compo-

nents of harmonically unrelated frequencies, some

of which were beating. In Risset's example, the ex-

ponential envelope was longer for the lower partials.

Additive synthesis provides great generality. But a

problem arises because of the large amount of data

to be specified for each note. Two control functions

for each component have to be specified, and nor-

mally they are different for each sound, depending

on its duration, intensity, and frequency. The pos-

sibility of data reduction has been investigated. At

Stanford University, a first result has been obtained

by representing the control functions of the ampli-

tude and the frequency of each component by line

segments, without affecting "naturalness" of the

sound (Grey and Moorer 1977).

The next step has been to investigate the rela-

tions between these functions (Risset and Mathews

1969; Beauchamp 1975) or their relation to others

of more general character (Charbonneau 1981). Ad-

ditive synthesis is most practically used either in

synthesis based on analysis (analysis/synthesis),

often transforming the extracted parameters, or

when a sound of a precise and well-determined

characteristic is required, as in psychoacoustic ex-

periments. In any case, in order to familiarize musi-

cians with sound characteristics and frequency

representations, the technique is also useful from

a pedagogical point of view.

Additive synthesis can be generalized by using

waveform components of other shapes besides si-

nusoids. To allow the reproduction of any sound,

these waveforms have to satisfy specific mathe-

matical properties. Walsh functions are an example

of this kind of function; they are used for their sim-

ple hardware realization (Rozenberg 1979).

VOSIM

In the synthesis techniques already discussed, os-

cillators that periodically reproduce a given wave-

form are employed. Other synthesis techniques,

instead of continuously repeating a given wave-

form, calculate it anew each period, with minor

variations. The control of this calculation process

allows continuous spectral variations. A common

method of this type is the voice simulation (VOSIM)

technique. A VOSIM oscillator has been devised in

a project at the Institute of Sonology in Utrecht

(Kaegi 1973, 1974; Kaegi and Tempelaars 1978).

The VOSIM waveform (Fig. 4) consists of a se-

quence of N pulses of shape sin2, of the same dura-

tion T, and of decreasing amplitude. The sequence

is followed by a pause M. Each pulse's amplitude

is smaller than the preceding one, by a constant

factor b.

The VOSIM spectrum (Fig. 5a) is described as the

product of two terms (Tempelaars 1976; De Poli

and De Poli 1979). The first term S, (Fig. 5b) de-

pends only on the pulse shape and limits the signal

bandwidth to 2F (being F = 1/T). The second term

S2 (Fig. 5c) depends on the relationship between the

individual pulse amplitudes. S2 is periodic in the

frequency domain with a period F, and it is sym-

DePoli 11

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 4. VOSIM oscillator: T

is the duration of single

pulse, M the rest between

two sequences of pulses.

Fig. 5. Spectral envelope of

a VOSIM oscillator (N =

5, b = 0, 8) (a). The enve-

lope is the product of the

terms S, (b) and S2 (c).

1.5

1

0.5

0 5 M- 10 15

T

metric with respect to F/2. When b 1, its ampli-

tude will be greater around the extremes of the

period 0 and F. When b - -1, its amplitude will be

greater in the central position around F/2. Thus, a

characteristic formant in F or F/2 will result. The

number of pulses N produces N oscillations in the

S2 term between 0 and F, with strong signals for b

near +F.

This constitutes the spectral envelope of the re-

peated waveform. Taking a as the ratio between the

signal period and a single pulse duration, the num-

ber of the harmonic corresponding to the formant is

a if b is positive, and a/2 if b is negative. Thus, by

varying a, the formant shifts, and the relative am-

plitude of all the harmonics vary continuously but

not homogeneously, following the spectral enve-

lope. The signal and the formant frequencies can be

separately controlled.

More kinds of sounds can be obtained by modu-

lating (sinusoidally or randomly) the value of the

time interval M between two consecutive pulse se-

quences. This means that a varies independently

from T. In this case, the formant frequency remains

constant while the harmonic amplitudes vary. Then

the ear can easily perceive the spectral envelope

and fuse the components together. This property

makes the VOSIM oscillator effective in musical

applications.

If a variation is strong, practically aperiodic

sounds or colored noises are obtained. Adding sev-

eral VOSIM oscillators allows one to control the

position of the formants. This results in an additive

(a)

5s(f)l

6

4

2

0

0 0.5F IF 1.5F 2F 2.5F

(b)

2.5

IsI(f)l

2

1.5

1

0.5

0

0 0.5F iF 1.5F 2F 2.5F

(c)

Is2(f)l

3

2.5

2

1.5

1

0.5

0 0.5F 1F 1.5F 2F 2.5F

12 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

synthesis of already complex sounds rather than of

sinuosidal components. Instead of the frequency of

partials, the position of the formants is controlled.

This is a more relevant parameter, from an acoustic

standpoint.

The formant-wave-function synthesis of Rodet

(1980) is analogous to VOSIM, but it allows over-

lapping of single waveforms. This provides better

control and generally richer sounds. Mitsuhashi

(1982a) and Bass and Goeddel (1981) generalized

the VOSIM model by including the case of pulses

of any amplitude and using different elementary

waveforms.

Synthesis by Random Signals

Up to now, we have considered signals whose be-

havior at any instant is supposed to be perfectly

knowable. These signals are called deterministic

signals. Besides these signals, random signals, of

unknown or only partly known behavior, may be

considered. For random signals, only some general

characteristics, called statistical properties, are

known or are of interest. The statistical properties

are characteristic of an entire signal class rather

than of a single signal. A set of random signals is

represented by a random process. Particular numer-

ical procedures simulate random processes, pro-

ducing sequences of random (or more precisely,

pseudorandom) numbers. The linear congruential

method is commonly used to produce uniformly

distributed numbers. From a starting value X0, a

sequence of random integers X0, X1, . . . , XK...

is generated according to the relation

XK+ = (a XK + C)modm,

where m is the modulus and the maximum se-

quence period, and a and c are two specific integer

constants.

The modulus operation can be avoided by choos-

ing m as the maximum number representable in

the computer, that is, m = 2b, where b is the word

length (bit number in a binary computer). So the

numbers are automatically truncated. The choice

of X0, a, and c greatly affects the statistical charac-

teristics of the generated sequence, and its accept-

ability has to be accurately verified by statisti

tests. A general discussion of various distribut

and the methods used to generate them can be

in Lorrain's paper (1980).

Random sequences can be used both as sign

(i.e., to produce white or colored noise used as

put to a filter) and as control functions to pro

a variety in the synthesis parameters most per

tible by the listener.

In the analysis of natural sounds, some chara

teristics vary in an unpredictable way; their m

statistical properties are perceptibly more sign

cant than their exact behavior. Hence, the add

of a random component to the deterministic f

tions controlling the synthesis parameters is o

desirable.

In general, a combination of random processe

is used because the temporal organization of th

musical parameters often has a hierarchical asp

It cannot be well described by a single random

process, but rather by a combination of rando

cesses evolving at different rates.

Linear Transformations

Let us now examine techniques for signal modifica-

tion. A transformation is a set of rules and proce-

dures transforming a signal called input to another

signal called output. A transformation is linear if

the superimposition principle is valid, that is, if the

effect of the transformation caused by a two-signal

addition is equal to the addition of the individual

signal transformations applied separately. In partic-

ular, in a linear transformation a signal can be mul-

tiplied by a constant but not by another signal.

Digital filters are linear transformations that can

be described by the following difference equation:

N M

Sa, yK=0 i1=0

where aK and b-

and y(i) are the i

signal. The value

ear combination

with the precedi

DePoli 13

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 6. Finite-impulse-

response (FIR) filter with

two zeros described by the

equation y(n) = x(n) +

a,x(n - 1) + a2x(n - 2)

(a). Infinite-impulse-

response (IIR) filter with

two poles described by the

equation y(n) = x(n) +

P1y(n - 1) + 32y(n - 2)

(b).

x(n) (a)y(n)

Z-1

x(n - 2)

----'--

x(n) (b)y(n)

Z-1

- ~ y(n- 2)

the input is sinusoidal, the steady-state output is

sinusoidal with the same frequency. The amplitude

and phase of the frequencies are determined by the

system. That is why this transformation is called

a filter.

Subtractive Synthesis

Sound produced by filtering a complex waveform is

called, sometimes inappropriately, subtractive syn-

thesis. First, a periodic or aleatoric signal rich in

harmonics is generated by the previously examined

techniques or others. This signal must contain

energy in all frequencies required in the output

sound. Second, one or more filters are used to alter

selectively the specific frequency components. The

undesired components are attenuated (subtracted)

and others are eventually amplified. When the filter

coefficients change, the frequency response changes,

too. Thus, it is possible to vary characteristics of

the output sound.

In modular diagrams, filters are usually repre-

sented by rectangles and the difference equation or

the transfer function is given as a label near the

rectangle. Two examples of simple digital filters,

showing their internal structure, are shown in

Fig. 6. The first filter (Fig. 6a) has a finite-impulse

response (FIR). This structure is useful to produce

transmission zeros: that is, it can nullify some fre-

quencies that depend on a1, a2 values and on the

sampling rate. The second filter (Fig. 6b) is recur-

sive, or has an infinite-impulse response (IIR). Feed-

back in the structure amplifies certain frequencies,

that is, produces transmission poles. When used as

bandpass filter, in general terms, the coefficient P,

controls the center frequency and the coefficient 32

the bandwidth.

One of the most attractive aspects of digital filter-

ing is that it is analogous to the functioning of

many acoustic musical instruments. Indeed, instru-

ment physics can be used as a model for synthesis.

For example, in the brasses and woodwind instru-

ments, the lips or vibrating reed generate a periodic

signal rich in harmonics. The various cavities and

the shape of the instrument act as resonators, en-

hancing some spectral components and attenuating

others. In the human voice, the excitation signals

are periodic pulses of the glottis (in the case of

voiced sounds) or white noise (in the case of un-

voiced sounds-for example, the consonants s and

z). The throat, the mouth, and the nose are the fil-

tering cavities, and their dimensions vary in time.

Their great variability makes the human voice the

most rich and interesting musical instrument.

Today, subtractive synthesis is the standard means

of speech synthesis. An analysis procedure, called

linear predictive coding (LPC), allows us to obtain

14 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 7. Elementary filters

used in reverberators.

Comb filters (a). All-pass

filter (b).

the pitch and the coefficients of a recursive (poles

only) filter (see Cann's [1979-1980] tutorial and

Moorer's paper [1979a]). These data can be utilized

to synthesize the sound directly or following modi-

fication. For example, speech can be accelerated or

slowed down, and pitch can be varied. An instru-

ment or orchestral sound can be used as input to the

filter, producing the effect of a "talking orchestra."

Interesting possibilities for musique concrete

sound processing arise. Not only simple filtering of

sounds is possible, but the modification of their

most intrinsic characteristics is also made possible

by varying the parameters of the deduced sound-

production model.

Generally, LPC is relatively difficult to use. In-

tuitively, the filter characteristics depend on the

position of the zeros and the poles in the transfer

function. These characteristics are affected in a

complex and nonintuitive way by the filter coeffi-

cients. In some simple cases, approximate formulas

give the coefficients as functions of significant pa-

rameters, that is, center frequency and bandwidth,

or cutoff frequency and slope. The filters can be

used in series or in parallel. In the most complex

cases, a precise analysis is obtained by using spe-

cific programs for digital filter design and analysis.

Such digital filters can be very stable and precise,

but only at the cost of a large amount of calcula-

tion. Simple linear digital networks can also be

used as oscillators (Tempelaars 1982) by applying

a pulse sequence to the input and choosing an

impulse response equal to the signal function to

be generated.

Reverberation

One application of digital filters is sound reverbera-

tion. An acoustic environment can be simulated by

distributing sound among different loudspeakers

and by adjusting the ratio between direct and rever-

berated sound (Chowning 1971). Most of the studio

reverberators sold today use digital technology.

The two elementary filters used in reverberation

are shown in Fig. 7. The first filter is called a comb

filter; in it, the signal is delayed a certain number of

samples, attenuated, and added to the input. An ex-

(a)

+ )( Delay

(b)

- G

ponentially decaying, repeated echo is so obtained.

The frequency response is characterized by equi-

spaced peaks-hence this filter's name. The peaks'

amplitude increases as G approaches 1.

The second filter is called an all-pass filter, since

the frequency response is flat and there is only a

phase shift. The input signal is attenuated and sub-

tracted from the delayed signal so that the feedback

effect is compensated and the echoes are main-

tained. The all-pass property is valid only in the

steady state with stationary sounds, not in tran-

sient states. Thus, it has a well-defined sound qual-

ity that a skilled listener can easily distinguish.

Reverberators are built combining some of these

filters (Moorer 1979b). Distinguishable signal repe-

titions should not occur in them, since the rever-

berated result should consist of a diffused sound.

The delay time of each elementary filter has to be

chosen very carefully. Sometimes a nonrecursive

echo generator is added to produce the first aperi-

odic echoes, which are the main perceptual deter-

minants of the characteristics of the room.

Nonlinear Techniques

In addition to linear transformations, which are

used in other fields and have a rather developed the-

ory, nonlinear transformations are used more and

DePoli 15

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 8. Waveshaping.

more commonly in musical applications. They

derive mainly from electrical communication the-

ory, and they have proved to be promising and effec-

tive. One use of nonlinear synthesis is in the large

amount of computer music generated by frequency

modulation (FM) synthesis (Chowning 1973).

In the classic case, nonlinear techniques use sim-

ple sinusoids as input signals. The output is com-

posed of many sinusoids, whose frequency and

amplitude depend mostly on the input ones.

Two main types of nonlinear techniques can be

distinguished, waveshaping and modulation. In

waveshaping, one input is shaped by a function de-

pending only on the input value in that instant. In

modulation (with two or more inputs), a simple pa-

rameter of one signal, called the carrier, is varied

according to the behavior of another signal, called

the modulator. In electrical communications (e.g.,

radio) the spectra of the signals are clearly distin-

guished and therefore easily separable. The origi-

nality in computer music application is the utiliza-

tion of signals in the same frequency range. Thus,

the two signals interact in a complex way, and

simple input variation affects all the resultant

components.

Often, the input amplitudes are varied by multi-

plying them by a constant or time-dependent pa-

rameter I, called the modulation index. Thus,

acting only on one parameter, the sound charac-

teristics are substantially varied. Dynamic and

variable spectra are easily obtainable. In additive

synthesis, similar variations require a much larger

amount of data.

Waveshaping

A linear filter can change the amplitude and phase

of a sinusoid, but not its waveform, whereas the

aim of waveshaping is to change the waveform. The

distortion of a signal heard from a nonlinear ampli-

fier is common. The output from a nonlinear am-

plifier of a sinusoidal signal is a signal with the

same period, but with a different waveform. The

various harmonics are present, and their amplitude

depends on the input and on the distortion. In stereo

systems, these distortions are usually avoided,

x(t)

y(t)y~t)

while waveshaping (Arfib 1979; Le Brun 1979;

Roads 1979) exploits them to generate periodic

sounds, rich in harmonics, from a simple sinusoid.

The function F(x), describing distortion, is called

the shaping function, and it associates with each in-

put value the corresponding output value indepen-

dent of time. If the input is x(t) = cos(27r ft), the

output is

s(t) = F(x(t)) = F(cos[2rr ft]).

In analog synthesis, it is difficult to have an am-

plifier with a precise and variable distortion char-

acteristic. In digital synthesis, this technique is

extremely easy to implement (Fig. 8). As in the case

of the oscillator, the shaping function can be previ-

ously computed and stored in a table. All that is

necessary is to look up the proper value from the

table.

Generally, if F(x) = F,(x) + F2(x), the distortion

produced by F is equal to the sum of those pro-

duced by F, and F2 separately. Usually, the shaping

produces infinite harmonics. But when a poly-

nomial of degree N is chosen as shaping function,

only the first N harmonics are present. Thus, fold-

over is easily avoided. Arfib and Le Brun deal exten-

sively with the mathematical relations among the

coefficients di of the shaping polynomial and the

amplitudes hi of harmonics generated when the am-

plitude I of the cosinusoidal input varies.

The shaping function, producing the jth har-

monic, is the Chebychev polynomial 7T(x) of degree

i (Fig. 9). Thus, to obtain the various harmonics of

16 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 9. Chebychev poly-

nomial of degree K used as

shaping function produces

only the Kth harmonic. In

the figure, K = 3.

T3(cos[ot]) = cos(3ot)

T3(x) = 4X3 - 3x

X = cos(wt) wt

01 1

-rr/6 -

2I/37T/3\_ os( ut)

73/2

27r/3

1lT/6

77r/6

137T/6

ot

amplitude h,, it is sufficient to add the correspon-

dent Chebychev polynomials, each multiplied

by hi:

N N

F(x) = h(x) = d;x.=0 i=O0

From these relations, it follow

monics are comp

even polynomial

harmonics. In th

only the odd ha

cient of x7 affect

enth harmonics

harmonic of ord

odd) coefficients

DePoli 17

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

example, the seventh harmonic is affected by the

odd coefficients from the seventh up to the degree

of the polynomial.

When the input amplitude I varies, the distortion

and the output spectrum vary. This is similar to an

expansion or contraction of the function, since

greater or smaller range of the function is employed.

From a mathematical point of view, the amplitude

variation corresponds to the multiplication of each

polynomial coefficient d, by II. The amplitudes of

the even or odd harmonics depend on I according

to the even (or odd) polynomials, which contain

the terms from the harmonic order up to the

polynomial degree.

If the spectrum is rather smooth, the number of

significant harmonics increases with the index.

Thus, a typical characteristic of real instruments is

reproduced, in that amplitude and spectrum are cor-

related. The amplitude and loudness of the output

vary with the input amplitude. In simple cases, this

effect can be compensated for by multiplying the

output by a suitable normalization function. But in

musical applications, the amplitude of the signal is

rarely constant, and it is multiplied by an envelope.

Normalization can be avoided by combining it with

the amplitude envelope in experimental or intuitive

ways after considering the normalization function.

It is also advisable to choose the even (or odd)

polynomial coefficients with alternating signs, that

is, according to the following model: + + - -

+ + - -. It is also advisable that the hi amplitude

not decrease abruptly, sharply limiting the band.

Otherwise, a spectrum would result that varied

very irregularly with I.

Dynamic spectral behavior cannot be easily an-

ticipated from the coefficients or from the static

spectrum. Moreover, the same (absolute-value)

spectrum can be produced by many polynomials

with different dynamic behaviors (Forin 1982). With

waveshaping, listening and graphic considerations

have more relevance than purely mathematical

formulations.

Another dynamic variation of waveshaping that

is easy to implement occurs when a constant is

added to the input; the shaping function shifts hori-

zontally. Even in this case, the spectrum varies.

The signal is periodic, with the same number of

harmonics. But in this case, the harmonic behavior

depends on both the even and the odd coefficients.

Generalizations of waveshaping technique are

possible. Reinhard (1981) studied the relations that

produce the partials generated by the polynomial

distortion of two cosine waves of frequency f, and

f2. All the components of frequency 1K f, ? Jf2

with IK + j|l N, where N is the polynomial de-

gree, are present.

Shaping functions that are not polynomial can

be used if the spectra produced by them are almost

band limited. Of particular interest is the use of

trigonometric and exponential functions (Moorer

1977) and of those where the input also appears in

the denominator (Winham and Steiglitz 1970;

Moorer 1976; Lehmann and Brown 1976; De Poli

1981).

Due to the wide spectral variation induced by

only one parameter (amplitude or shift), wave-

shaping is particularly convenient in musical ap-

plications, especially in combination with multi-

plicative synthesis. Moreover, it is suitable for

modeling the sound production of some acoustic

instruments (Beauchamp 1979, 1982). There is a

large and not intuitive problem in choosing the co-

efficients, however, and further research is required.

Multiplicative Synthesis (Ring Modulation)

The simplest nonlinear transformation consists of

the multiplication of two signals. In analog syn-

thesizers, it is called ring modulation (RM). Some-

times it is also called amplitude modulation (AM),

but the two differ, especially in their realization.

With two inputs x,(t) and x2(t), the output is

s(t) = Xl(t) . x2(t). Obviously, when the inputs inter-

change, the result does not vary. The resulting spec-

trum is obtained from the convolution of the two

signals' spectra. Usually, one of the two signals,

called the carrier, is sinusoidal; the result is not

too complex and noisy.

When x1 is the sinusoidal carrier of frequency f/,

and x2 (modulator) is sinusoidal with frequency f2,

from cos(a) , cos(3) = 12lcos(a + 3) + cos(a - /)1,

18 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 10. Multiplicative syn-

thesis. Spectrum of a peri-

odic signal X2 with four

harmonics (a). Resulting

spectrum when d2 is mul-

tiplied by a sinusoid of fre-

quency f, greater than its

bandwidth (f, = 7f2) (b).

Resulting spectrum when

x2 is multiplied by a sinu-

soid of frequency inferior

to its bandwidth (f, =

26f2) (c). The components

deriving from the folding

of negative frequencies are

shown as dashed lines.

the output consists of two sinusoidal partials of fre-

quency fI + f2 and f, - f2. The phases of the output

are also the sum and the difference of the phases of

the two inputs. For example, if x, and x2 frequen-

cies are 400 Hz and 100 Hz, the output has two par-

tials of frequency 500 Hz and 300 Hz.

Negative frequencies may occur, for example,

when f, = 100 Hz and f, = 400 Hz. This often hap-

pens in modulations (foldunder) and can be ex-

plained by the trigonometric relation cos(a)

= cos(-a), from which cos(27r ft + (4) = cos(7r[-flt

- 4). The alteration of the frequency sign only

changes the sign of the phase with respect to the

cosine. In particular, a cosine signal is unaffected,

while a sine wave changes its sign. In the inter-

pretation of the results, only absolute frequency

values have to be considered. Usually, the phase is

not significant, as the ear is not terribly sensitive to

it. But the phase has to be taken into account while

summing the amplitude of components of identical

frequencies.

In multiplicative synthesis, usually x2 is periodic

with frequency f2. The multiplication causes every

harmonic spectral line of frequency K . f2 in the

original signal to be replaced by two spectral lines

(called sidebands) of frequency f, + K f2 and

f, - K f2. The resulting spectrum has components

of frequency If +? K f2 , where K is equal to the or-

der of the different harmonics in x2 (Fig. 10).

Thus two sidebands, symmetric with respect to

the carrier, occur. When f, is less than the greatest

frequency in x2, then the negative frequencies fold

around zero, as discussed above.

The possibility of shifting the spectrum is very

intriguing in musical applications. From simple

components, harmonic and inharmonic sounds can

be created, and various harmonic relations among

the partials can be established. If x2 is a signal with

spectrum X2, the signal obtained from its multi-

plication with a sinusoid of frequency f, has two

sidebands symmetric with respect to f, and shaped

like X2.

A periodic signal x, can be expanded in Fourier

series. Each x, partial will have sidebands of ampli-

tude proportional to its own. If f1 is less than the

bandwidth of x,, then the sidebands overlap with

(a) jx2(f)l

(b)

I|S(f)l

ff2

(c)

S(f)I

Iu I li I fl

eventual component superimposition. In this case,

the phases have to be taken in account while sum-

ming. Dashow (1978, 1980) describes some general-

ization of this technique and employs the generated

spectra for particular "harmonizations" of pitches

specified by the composer.

Amplitude Modulation

In RM, the carrier does not appear in the spectrum

created by the product of a sinusoidal carrier with

another signal, except when the modulator has a di-

rect current (dc) component. In carrying out the

modulation in AM (Fig. 11), the carrier is present in

the output, with an amplitude independent of the

sidebands. The formula for AM is as follows:

s(t) = xl(t) . (K + x2(t)).

The result is RM with carrier added. When the

carrier is sinusoidal and the modulator is periodic,

the spectrum is composed of partials of frequency

If1 1 K f,2, with K = 0, 1, . It is useful to distin-

DePoli 19

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 11. Amplitude

modulation.

K A2 f2 f1

X2(t)

x1(t)

s(t)

guish between the two modulations because they

have different realization schemes.

Spectra of Type If , K f2

The following considerations are valid for all spec-

tra whose components are of type If, + K f2h, with

K = 0, 1, .... The spectrum is characterized by the

ratio f,/f2. (This is often referred to as the carrier-

to-modulator [c:m] ratio.) When this ratio is ra-

tional, it can be expressed as an irreducible fraction

fI/f2 = NI/N2, with N, and N2 as integers that are

prime between themselves. In this case, the result-

ing sound is harmonic, since the various compo-

nents are a multiple of a fundamental according to

integer factors. The fundamental frequency is

fo = - /

N1 N2 '

and the carrier coincides with the NIth harmonic.

If N2 = 1, all the harmonics are present and the

sideband components coincide. If N2 = 2, only odd

harmonics are present and the sidebands superim-

pose. If N2 = 3, the harmonics that are multiples of

3 are missing. The c: m ratio is also an index of the

harmonicity of the spectrum. The sound is mo

"harmonious" intuitively when the N /N2 rati

simple and formally when the N, N2 productsmaller.

The ratios can be grouped in families (Truax

All ratios of the type If, K f2l/f2 can produce

same components that flf, produces. Only th

tial coinciding with the carrier (f,) changes. Fo

ample, the ratios 2/3, 5/3, 1/3, 4/3, 7/3 and s

all belong to the same family. Only the harmo

that are multiples of 3 are missing (see N2 = 3

the carrier is respectively the second, fifth, fir

fourth, seventh, and so on harmonic.

The ratio that distinguishes a family is defin

normal form when it is - 1/2. In the previous

ample, it is 1/3. Each family is characterized b

ratio in normal form. Similar spectra can be p

duced using ratios from the same family. Diffe

spectra are obtained by sounds of different fa

When the fl/f2 ratio is irrational, the resulti

sound is aperiodic and hence, inharmonic. Of p

ticular interest is the case of an f,/f2 ratio app

mating a simple value, that is,

fl/f2 = N1/N2 + e.

Here the sound is no longer rigorously periodi

The fundamental frequency fo is still f2/N2,

the harmonics are shifted from their exact valu

by +e / f2. When N2 is equal to 1 or 2, the po

and negative components are not superimposed

beat with a frequency of 2e / f2. Hence, a smal

of the carrier does not change the pitch, even

slightly spreads the partials and makes the sou

more lively. But the same shift of the modula

frequency f, changes the sound's pitch.

Frequency and Phase Modulation

Another type of modulation, suggested by Cho

ing (1973), has become one of the most widely

synthesis techniques. In general, it consists of

modulation and it can be realized both as ph

modulation (4M) or as FM. This technique do

not derive from models of production of phys

sounds, but only from the mathematical prope

ties of a formula. It has some of the advantage

20 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 12. The number of sig-

nificant sidebands in FM.

waveshaping and RM, and it avoids some of their

drawbacks.

The technique consists of the modulation of the

instantaneous phase or frequency of a sinusoidal

carrier according to the behavior of another signal

(modulator), which is usually sinusoidal. It can be

expressed as follows:

s(t) = sin(2zT f t + I sin[27r fmt]) =

=>K JK (I)sinj27T(f, + Kfj)tI.

The resulting spectrum is of the type If, +? K f,,. All

the spectral considerations discussed previously are

applicable, particularly those regarding negative fre-

quency, foldunder, fclf, ratios, and harmonic andinharmonic sounds.

The amplitude of each Kth side component of

the FM technique is given by the Bessel function of

Kth order computed in I. To plot the spectrum, a

table of Bessel functions has to be referenced to ob-

tain the amplitudes of the carrier and of the side

frequencies in the upper sideband. The odd-order

side frequencies in the lower sideband have signs

opposite to those in the upper one, and the even-

order side frequencies have the same sign. The

negative frequencies, being sine waves, are folded,

changing the sign. When superimposition occurs,

the amplitudes are added algebraically.

When I (called the modulation index) varies, the

amplitude of each component varies as well. Thus,

dynamic spectra can be obtained simply by varying

this index. Each component varies its amplitude

by following the corresponding Bessel function. A

Bessel function can be asymptotically approxi-

mated by a damped sinuosoid. So when the index

varies, some components increase and others de-

crease, all without sharp variations.

In Eq. (1), the sum includes infinite terms, so the-

oretically the signal bandwidth is not limited. But,

practically, it is limited. In the Bessel function's

behavior, only a few low-order functions are sig-

nificant for small index values. When the index in-

creases, the number and the order of the significant

functions increase. For a given index, the side am-

plitudes oscillate with gradually increasing ampli-

tude and slowly increasing period all the way from

25

20 /

15 /

5

M ' I I

I 5 10 15 20

the origin to a

toward zero. Th

slightly below

Usually, in the

signal, all side f

than /loo of th

ered. The numb

M = I + 2.4 J10.27

(See Fig. 12.) Often, as a rule of thumb, it is roughly

considered as

M = I+ 1.

In Eq. (1), the sum can be performed for K fr

-M to +M. For a harmonic sound, that is, w

the ratio fc/fm = NI/N2 is simple, the maxim

der of significant harmonics is N, + M ' N2.

For wide index variations, the sounds produc

are characteristic of the FM technique. A typi

timbre of FM sound is easily recognizable an

well defined. This does not happen for small i

variations or for compound carriers or modul

Frequency modulation synthesis has another p

DePoli 21

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 13. Frequency

modulation.

A(t) fc d(t) fm

s(t)

Fig. 14. Frequency modu-

lation with N carriers

modulated by the same

oscillator.

a, f/ a2 f2 aN fN d f,

+ + +

s(t)

erty that is very important in musical applications:

the maximum amplitude and the signal power do

not vary with the index I. Unlike the situation in

waveshaping, normalization of the output is not

necessary.

Let us now examine the difference between OM

and FM. Phase modulation is defined as follows:

s(t) = sin (2M ft + 0[ t]),

and it corresponds to Eq. (1) if the modulating sig-

nal is 0(t) = I sin(2rr f, t).

Frequency modulation occurs when the instanta-

neous frequency varies around the carrier value ac-

cording to the behavior of the modulating wave. For

a signal s(t) = sin(p[ t]), the instantaneous frequency

is fj = (1/27r) (di[t]l/dt). Thus, the instantaneous

frequency of the signal in Eq. (1) is as follows:

fi = fc + I fm , cos(2r f,, t).

The frequency varies around f, with a maximum

deviation d = I f,. Thus, with a modulating wave

I - f, cos(27r f, t), an FM equivalent to OM is ob-

tained. Both phase and frequency modulations are

special cases of angle modulation.

In sound synthesis programs, frequency-driven

oscillators are provided. The integration involved in

calculating the instantaneous phase is therefore

computed automatically. Frequency modulation is

normally implemented as in Fig. 13. A change of

the phase between the carrier and the modulating

wave in Eq. (1) only changes the reciprocal phase of

the partials. If components superimpose, their total

amplitude changes, and a direct-current component

may appear. The next sections examine some use-

ful extensions of the basic algorithm.

Nonsinusoidal Carrier

Here we consider a periodic nonsinusoidal carrier.

The result of its modulation is the modulation of

each of its harmonics by the same wave. Sidebands

of amplitude proportional to each harmonic will be

present around the carrier. The result is a spectrum

with components of frequency In f, + K . fmi,, with

K = 0,... , M and n = 1,..., N, when Nis the

number of significant harmonics. The maximum

frequency present is N - f, + M . f,. In general, there

may be various independent carriers modulated by

the same wave (Fig. 14) or by different modulating

signals. This is like additive synthesis, only instead

of sinusoidal addends, more complex addends are

used. For example, harmonic sounds can be gener-

ated by controlling the various spectral ranges with a

few significant and independent parameters. Sounds

of the same "family" are possible.

The frequency of each carrier determines the

22 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 15. Frequency modu-

lation with two modu-

lators.

Fig. 16. Frequency modu-

lation with N modulators.

A fc dl(t) f, d2(t) f2

s(t)

location of the formant position, the amplitude

determines its energy, and the modulation index

specifies its bandwidth. Chowning (1981) demon-

strated these facilities in synthesis of the singing

voice of a soprano.

Compound Modulation

Let us examine the case of a modulation composed

of two sinusoids (Fig. 15), each with its own modu-

lation index, applied to a sinusoidal carrier. The for-

mula for two-sine-wave OM (Le Brun 1977) is as

follows:

s(t) = sin(121T fct + I, sin[27r f t] + 12 sin[27r f2t])

= K n, JK(I1) Jn(12) . sin(I27T[fc + K f, + n f2ltI).

The same result can be obtained with FM using as

modulating signal the following expression:

I,f,/cos(27r f, t) + I2f2cos(2r f2t).

The resulting spectrum is much more complex

than in the one-modulator case. All the compo-

nents of frequency If/ ? K f, n f,2 are present, and

their amplitude is JK(I) " Jn(I2).

To interpret the effect, let us consider f1 > f2. If

only f, were present, the resulting spectrum would

have a certain number of components of amplitude

JK(I1) and frequency fc + K f1. When the modulator

A fc d1(t) f, d2(t) f2 dM(t) fM

s(t)

f, is applied, these components become carriers,

with sidebands produced by f,. The resulting band-

width is approximately equal to the sum of the two

bandwidths.

If the frequencies have simple ratios, the spec-

trum is of the type If, -K f,,, where now f, is the

greatest common divisor of f, and f,. For example,

with f, = 700 Hz, f = 300 Hz, and f2 = 200 Hz, the

components are 1700 ? K 1001. Thus, by choosing

f, and f2 multiples of f,, sounds belonging to the

same family as a simple modulation, but with a

more complex spectral structure, can be generated.

In general, if the modulating signal is composed of

N sinusoids (Fig. 16), the following relations hold:

s(t) = sin (2rT fct + EI sin[2r fs t]

N

= KS, HKS (Is) sin|2ir(fct + Y Ifs)t1.

Thus, all the components of frequency If , Klf/ -. . ? KNfNI, with amplitudes given by the product

of N Bessel functions, are obtained. A very complex

spectrum results. If the relations among the fre-

quencies f, are simple, that is, if the modulating

wave is periodic, then the spectrum is of the type

Ifc ? K fm, where f,m is the greatest common divisor

among the modulating components. Otherwise, the

sonorities are definitely inharmonic and particu-

larly noisy for high indexes.

DePoli 23

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Fig. 17. Nested FM.

Nested or Complex Modulation

Let us examine the case of a sinusoidal modulator

that is phase modulated by another sinusoid. The

signal is defined as follows:

s(t) = sin(I27r ft + I sin[2ir f, t + 12 sin{2ir f2t}]))

= K JK(Il)sin(I2zr[fc + K f,]t

+ K 12 sin[27r f2t]I)

= 1K J(I,)" Jn(KI2)sin(2r[fc + K,f, + n f2]t).

The result can be interpreted as if each partial pro-

duced by the modulator f, were modulated in its

turn by f2 with modulation index K I,. Thus, all

the partials of frequency If +? K f, n f2j, with ap-

proximately 0 - K - I, Os n In , 1 I2, are present.

The maximum frequency is fc + I1I(f + 12/2).

The structure of the spectrum is similar to that

produced by the two-sinusoid modulation, but with

a larger bandwidth. Even where fm is the greatest

common divisor between f, and f2, the spectrum is

of the type If, + K fn-.

In the equivalent realization by FM (Fig. 17), the

spectrum is of the same type, but with slightly dif-

ferent amplitudes. A direct-current component in

the resulting modulating wave added to the carrier

is avoided by choosing a sine wave modulated by a

cosine wave.

This technique is made more interesting by an

algorithm suggested by Justice (1979), which en-

ables an analysis of a sound according to this model,

with the frequency and the index behavior of two

or more nested modulators being deducible.

Other Two-Input, Nonlinear Transformations

Mitsuhashi (1980) proposed a more complex two-

input, nonlinear transformation, in which the in-

stantaneous phase and amplitude of an approx-

imately sinusoidal signal are simultaneously varied.

In another paper, Mitsuhashi (1982c) generalized

this technique while discussing some criteria in

choosing the two-input, nonlinear function and

suggesting two examples. The function is time in-

dependent, bidimensional, and considered periodic

outside the definition field. Thus, it can be imple-

mented with a two-dimensional table, with analogy

to an oscillator. This technique appears very inter-

A fc d (t) f/ d2(t) f,

+

s(t)

esting, even if it seems to be difficult to find a sim-

ple expression that bounds significant parameters of

the resulting spectrum to the input and function

characteristics. Another promising modulation

technique is linear sweep synthesis, recently sug-

gested by Rozenberg (1982).

Conclusion

As a consequence of progress in digital hardware

and software, the initial antithesis between com-

puting efficiency and timbral richness is lessening.

Digital sound quality largely depends on the amount

of introduced or controlled detail; excessive sim-

plifications lead often to trivial results. It follows

that increased computing power can generate more

sophisticated results.

A musically interesting sound can be obtained in

two ways. The first consists of the utilization of

more complex techniques or of the combination of

24 Computer Music Journal

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

many of the techniques described here. Many linear

and nonlinear transformations are possible. Most of

the parameters do not have to be constant and can

be varied by control functions and random signals.

The other synthesis approach consists of the su-

perimposition of many simple sounds produced by

basic techniques. The evolution of the individual

sounds is not complex, and the richness of the re-

sult essentially depends on their combination. In

this approach, the parameters of many elementary

sounds have to be given. Specific programs are often

used to define these parameters.

Sound evolution can be regulated either by con-

trol functions in the synthesis or by programs com-

puting the parameters for the synthesis. In any

case, many details of the sound have to be accu-

rately controlled. Their coherence both within the

sound and in the context of adjacent and simul-

taneous notes has to be guaranteed. The relations

among sounds can be more easily highlighted when

they are reflected not only in macroscopic param-

eter variations but also in internal structure.

The extensive utilization of a single technique re-

veals its peculiar characteristics. This derives from

the finite repertoire of obtainable sounds and, more

specifically, from the more easily producible dy-

namic variations associated with it. Thus, it is wise

to use different techniques, the better to exploit

their different potential. Moreover, the musician

must study and experiment with a technique. This

is essential in order to determine all its charac-

teristics and to acquire a feeling for the parameter

choices necessary for nontrivial use. In any case, a

synthesis technique is simply a tool to produce

sound, and sound is not yet music.

References

Arfib, D. 1979. "Digital Synthesis of Complex Spectra by

Means of Multiplication of Nonlinear Distorted Sine

Waves." Journal of the Audio Engineering Society

27(10): 757-768.

Bass, S. C., and T. W. Goeddel. 1981. "The Efficient Digi-

tal Implementation of Subtractive Music Synthesis."

IEEE Micro 1(3) :24-37.

Beauchamp, J. W. 1975. "Analysis and Synthesis of Cor-

net Tones Using Non Linear Interharmonic Rela-

tionships." Journal of the Audio Engineering Society

23(10): 778-795.

Beauchamp, J. W. 1979. "Brass Tone Synthesis by Spec-

trum Evolution Matching with Nonlinear Functions."

Computer Music Journal 3(2): 35-43.

Beauchamp, J. W. 1982. "Synthesis by Spectral Amplitude

and 'Brightness' Matching of Analyzed Musical Instru-

mental Tones." Journal of the Audio Engineering So-

ciety 30(6):396-406.

Cann, R. 1979-1980. "An Analysis Synthesis Tutorial."

Part 1, Computer Music Journal 3(3):6-11; Part 2,

Computer Music Journal 3(4):9-13; Part 3, Computer

Music Journal 4(1):36-42.

Cerruti, R., and G. Rodeghiero. 1983. "Comments on

'Musical' Sound Synthesis by Forward Differences."

Journal of the Audio Engineering Society 31(6).

Charbonneau, G. 1981. "Three Types of Data Reduction."

Computer Music Journal 5(2): 10-19.

Chowning, J. M. 1971. "The Simulation of Moving Sound

Sources." Journal of the Audio Engineering Society

19(1): 2-6. (Reprinted in Computer Music Journal

1[3]:48-52, 1977.)

Chowning, J. M. 1973. "The Synthesis of Complex Audio

Spectra by Means of Frequency Modulation." Journal of

the Audio Engineering Society 21(7): 526-534. (Re-

printed in Computer Music Journal 1[2]:46-54, 1977.)

Chowning, J. M. 1981. "Computer Synthesis of the Sing-

ing Voice." In Sound Generation in Winds Strings

Computers. Stockholm: KTH Skriftserie 29, pp. 4-13.

Dashow, J. 1978. "Three Methods for the Digital Syn-

thesis of Chordal Structure with Non-Harmonic Par-

tials." Interface 7(2/3):69-94.

Dashow, J. 1980. "Spectra as Chords." Computer Music

Journal 4(1):43-52.

De Poli, G. 1981. "Sintesi di suoni mediante funzione dis-

torcente con poli complessi coniugati." Atti del IV Col-

loquio di Informatica Musicale 1, Pisa, pp. 103-130.

De Poli, E., and G. De Poli. 1979. "Identificazione di

parametri di un oscillatore VOSIM a partire da una

descrizione spettrale." Atti del III Colloquio di

Informatica Musicale, Pisa, pp. 161-177.

Forin, A. 1982. "Spettri dinamici prodotti mediante dis-

torsione con polinomi equivalenti in un punto." Bollet-

tino LIMB 2:62-76.

Grey, J. M., and J. A. Moorer. 1977. "Perceptual Eva

tion of Synthesized Musical Instrument Tones." Jou

nal of the Acoustical Society of America 62:434-

Justice, J. M. 1979. "Analytic Signal Processing in M

Computation." IEEE Transactions on Acoustics, S

and Signal Processing (ASSP) 27(6):670-684.

DePoli 25

This content downloaded from 143.215.137.43 on Mon, 08 Jun 2020 20:13:38 UTC

All use subject to https://about.jstor.org/terms

Kaegi, W. 1973. "A Minimum Description of the Linguis-

tic Sign Repertoire (part 1)." Interface 2:141-156.

Kaegi, W. 1974. "A Minimum Description of the Linguis-

tic Sign Repertoire (part 2)." Interface 3: 132-158.

Kaegi, W., and S. Tempelaars. 1978. "VOSIM-A New

Sound Synthesis System." Journal of the Audio Engi-

neering Society 26(6): 418-424.

Le Brun, M. 1977. "A Derivation of the Spectrum of FM

with Complex Modulating Wave." Computer Music

Journal 1(4):51-52.

Le Brun, M. 1979. "Digital Waveshaping Synthesis." Jour-

nal of the Audio Engineering Society 27(4):250-265.

Lehmann, R., and F. Brown. 1976. "Synthese rapide des

sons musicaux." Revue d'Acoustique 38:211-215.

Lorrain, D. 1980. "A Panoply of Stochastic 'Cannons.'"

Computer Music Journal 4(1):53-81.

Mailliard, R. 1976. "Les distorsions de Music V." Cahiers

recherche/musique 3:207-246.

Mathews, M. V. 1969. The Technology of Computer Mu-

sic. Cambridge, Massachusetts: MIT Press.

Mitsuhashi, Y. 1980. "Waveshape Parameter Modulation

in Producing Complex Audio Spectra." Journal of the

Audio Engineering Society 28(12): 879-895.

Mitsuhashi, Y. 1982a. "Musical Sound Synthesis by For-

ward Differences." Journal of the Audio Engineering

Society 30(1/2): 2-9.

Mitsuhashi, Y. 1982b. "Piecewise Interpolation Tech-

nique for Audio Signal Synthesis." Journal of the Audio

Engineering Society 30(4): 192-202.

Mitsuhashi, Y. 1982c. "Audio Signal Synthesis by Func-

tions of Two Variables." Journal of the Audio Engineer-

ing Society 30( 10): 701 - 706.

Moorer, J. A. 1976. "The Synthesis of Complex Audio

Spectra by Means of Discrete Summation Formulae."

Journal of the Audio Engineering Society 24(9):

717-727.

Moorer, J. A. 1977. "Signal Processing Aspects of Com-

puter Music: A Survey." Proceedings of the IEEE

65(8): 1108-1132. (Reprinted in Computer Music Jour-

nal 1[1]:4-37, 1977.)

Moorer, J. A. 1979a. "The Use of Linear Prediction of

Speech in Computer Music Applications." Journal of

the Audio Engineering Society 27(3): 134-140.

Moorer, J. A. 1979b. "About This Reverberation Busi-

ness." Computer Music Journal 3(2): 13-28.

Reinhard, P. 1981. "Distorsione non lineare della somm

di due cosinusoidi: analisi dello spettro tramite ma-

trici." Atti del IV Colloquio di Informatica Musicale

Pisa, pp. 160-183.

Risset, J.-C. 1969. "An Introductory Catalog of Comput

Synthesized Sounds." Murray Hill, New Jersey: Bell

Laboratories.

Risset, J.-C., and M. V. Mathews. 1969. "Analysis of Mu-

sical Instrument Tones." Physics Today 22(2):23-30.

Roads, C. 1978. "Automated Granular Synthesis of

Sounds." Computer Music Journal 2(2):61-62. Revised

and updated version forthcoming in C. Roads and

J. Strawn, eds., Foundations of Computer Music. Cam-

bridge, Massachusetts: MIT Press.

Roads, C. 1979. "A Tutorial on Non-linear Distortion or

Waveshaping Synthesis." Computer Music Journal

3(2): 21-34.

Rodet, X. 1980. "Time Domain Formant Wave-Function

Synthesis." In Spoken Language Generation and Un-

derstanding, ed. J. G. Simon. Dordrecht: D. Reidel.

Rozenberg, M. 1979. "Microcomputer-controlled Sound

Processing Using Walsh Functions." Computer Music

Journal 3(1):42-47.

Rozenberg, M. 1982. "Linear Sweep Synthesis." Com-

puter Music Journal 6(3): 65-71.

Tempelaars, S. 1976. "The VOSIM Signal Spectrum." In-

terface 6:81-86.

Tempelaars, S. 1982. "Linear Digital Oscillators." Inter-

face 11(2): 109-130.

Truax, B. 1977. "Organizational Techniques for C: M

Ratios in Frequency Modulation." Computer Music

Journal 1(4):39-45.

Winham, G., and K. Steiglitz. 1970. "Input Generators for

Digital Sound Synthesis" (Part 2). Journal of the Acous-

tical Society of America 47(2):665-666.

Xenakis, I. 1971. Formalized Music. Bloomington, In-

diana: Indiana University Press

BABBLE ONLINE: APPLYING STATISTICS AND DESIGN TO SONIFY THE INTERNET

M. H. Hansen

Bell Laboratories

Murray Hill, New Jersey

cocteau@bell-labs.com

B. Rubin

EAR Studio

New York City, New York

benrubin@earstudio.com

ABSTRACT

A statistician (Hansen) and a media artist (Rubin) investigate the

application of statistical methods and sound-design principles to

the real-time sonification of Internet communications. This paper

presents results from two applications: the sonification of brows-

ing activity on Lucent’s Web site, and the sonification of a large

number of Internet chat sites in real-time. These experiments sug-

gest new ways to experience the diverse and dynamic data streams

generated by modern data networks. As an art-technology collab-

oration, the project outcomes range from the creation of art instal-

lations to the development of practical monitoring platforms. This

paper discusses the interplay between these two perspectives, and

suggests that each is motivated by a common interest in generating

meaningful experiences with dynamic data.

1. INTRODUCTION

Modern work in sonification emerged in the literature on computer-

human interfaces and over the last decade has matured into its own

field of scientific inquiry. The use of sound in exploring the in-

formation hidden in data, the principles and broad application of

auditory displays are eloquently described in [10]. Early applica-

tion areas included real-time monitoring of financial data, medical

diagnostics, and even air traffic control systems. Computer sim-

ulations also provided extensive data for sonification. Since that

point, a virtual explosion has taken place in our ability to collect

data relating to human communication and social systems.

Today, almost every aspect of our lives can be “rendered” dig-

itally. Advances in data collection technologies have made com-

monplace continuous, high-resolution measurements of our phys-

ical environment (weather patterns, seismic events, ecological in-

dicators). Equally open to observation are our routine movements

through and interactions with our physical surroundings (automo-

bile and air traffic, large-scale land use). In computer-mediated

settings, our activities either depend crucially on or consist en-

tirely of complex digital data (financial transactions, accesses to

global information systems, Web site and Internet usage). As a

reflection of the diversity and variety of the systems under study,

these data-based descriptions of our daily lives tend to be massive

in size, dynamic in character, and repleat with rich structures.

The advent of these enormous repositories of digital informa-

tion presents us with an interesting challenge: How can we rep-

resent and interpret such complex, abstract and socially important

data? In a new collaboration, Ear to the Ground [4], we have begun

an exploration into ways of creating experiential encounters with

otherwise abstract data streams, especially through sound.1 In [8],

1 Ear to the Ground is part of the Arts in Multimedia project co-

we discuss the broad goals of our collaboration and examine soni-

fication from both an artistic as well as a data analytic perspective.

In this paper, we examine the use of auditory displays in under-

standing large-scale Internet communications.

2. EXAMPLES

2.1. Web site traffic

Every day, large Web sites like Yahoo attract hundreds of thou-

sands of visitors. During active periods, there can be thousands

of people accessing data simultaneously from a Web site. For

users of information portals like Yahoo, the speed of the servers

(as reflected by rapid or sluggish responses) provides the only clue

about the number of other people accessing information. While

attempts have been made to visually assess browsing patterns in

real-time [11, 1], the effectiveness of these displays deteriorates

for high-traffic domains. For our first sonification example, we

create an ambient display to characterize certain aspects of the ac-

tivity on a busy Web site.

As you navigate the Web, your browser requests various kinds

of data from one or more Web servers. As part of part of the pro-

cess of delivering content, most Web servers will record informa-

tion about the visitor and the items they requested. These items

include HTML pages, images, Java class files and PostScript doc-

uments. The information available to the Web server about each

request includes a timestamp, the IP address of the visitor’s com-

puter, the type of browser they are using, the URL of the requested

item, the “referral page” (the URL that directed the visitor to the

requested item), and perhaps a “cookie” to recognize returning vis-

itors. These details are typically stored as a single line of a poten-

tially enormous log file.2 The data for this experiment came from

the Lucent Technologies corporate site, www.lucent.com. On

a typical day, 60K visitors to this site will generate a 15Mb (com-

pressed) log file, consisting of 700K entries. Given that our interest

is on how users navigate the content of a site, we restrict our at-

tention to HTML files, PostScript and PDF documents. All other

requests made to the Web server are ignored, reducing the data

by a factor of 10. We then further process the data to extract user

paths or “visits,” where a visit is a contiguous sequence of requests

made by a user while browsing the site. Over 70% of the visitors

to www.lucent.com look at just three pages or less, and hence

a minority of the visits exhibit “interesting” navigational patterns.

sponsored by Lucent Technologies and the Brooklyn Academy of Music

(BAM). The authors gratefully acknowledge the help of project managers

Wayne Ashley (BAM) and Marah Rosenberg (Lucent; now with Avaya

Communications).

2 Each line is commonly referred to as a “hit.”

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-10

The Lucent Web site is built hierarchically, in the sense that

pages deeper in the directory tree represent more detailed infor-

mation than those at shallower levels. At its busiest, there can

be as many as 300 people browsing www.lucent.com; while

during the pre-dawn hours there can be as few as 5 simultaneous

visitors. Our sonification is designed to convey qualitative infor-

mation about site usage, answering questions like:

Overall, is the site busy or quiet?

What proportion of the visitors are delving for specific in-

formation deep within the site, as compared to those visitors

who are “just passing through,” glancing briefly at the home

page and then moving on?

How are users distributed across the various content areas

of the site?

Which portions of the site are visited together? What kinds

of patterns do we find in user behavior?

We think of this sonification as one possible “background” infor-

mation stream that can inform content providers, Web designers

and even the visitors themselves.

2.1.1. Sonification design

Our audio display makes use of the hierarchical structure of the

content offered by www.lucent.com. First, a unique pitch was

used to identify each of five high-level subdomains within the site:

/micro, representing Lucent’s microelectronics design and man-

ufacturing business (now Agere Systems); /enterprise, for

the enterprise systems and software business (now Avaya Com-

munications); /minds, a corporate introduction to Bell Labs re-

search; /press, a collection of press releases and investor infor-

mation; and /search, the local search engine for the site.

The total number of visitors accessing any information from a

subdomain affects the loudness and tonal balance of a low-register

drone at the associated pitch. Visitors requesting content deeper in

the site are represented by higher-pitched pulsing tones (separated

by one or two octaves from the base pitch for the subdomain):

the faster the pulse, the more people are accessing that area, and

the greater the proportion of high-register sounds, the more de-

tailed the content. By assigning well-separated pitches to each

subdomain, shifts in activity both within and between the areas

can be heard. In Table 1 we present a simple mapping of data col-

lected by the Lucent Web server to a continuously time-varying

vector of usage statistics. In the category of Overall browsing, we

count any visitor accessing content pages (HTML, PostScript or

PDF) from the indicated subdomain. A Mid-Level access is a re-

quest for content two or more directories down. Simple examples

are /micro/K56flex/index.html (information on a brand

of 56K modem) and /press/0101/010118.nsb.html (a

press release for January 18, 2001). The final category, Deep

browsing, refers to pages that are four or more directories down

in the tree. One example is a paper from the April/June 2000

issue of the Bell Labs Technical Journal, located at /minds/

techjournal/apr-jun2000/pdf/paper02.pdf.

Then, the resulting 15 values in Table 1, A1–E3, were mapped

to sound as follows:

Overall activity Measured by A1–E1, voiced with a low-register

drone. The aggregate number of visitors accessing infor-

mation within each of the five areas modulates the loudness

of each of the five pitches.

/micro /enterprise /minds /press /search

Overall A1 B1 C1 D1 E1

Mid-Level A2 B2 C2 D2 E2

Deep A3 B3 C3 D3 E3

Table 1: Mapping used for Web site traffic example. Overall ac-

tivity records the movements of all users; Mid-Level counts users

2 or 3 directories into the site; Deep browsing consists of users 4+

directories down.

Mid-Level browsing Measured by A2–E2 and assigned a rhyth-

mic middle-register tone pulse; pulse loudness and repeti-

tion speed rises and the timbral brightness increases as the

volume of mid-level browsing increases. There are five in-

dependent pulses, each at a different fixed pitch, represent-

ing the five content areas.

Deep browsing Measured by A3–E3 and made audible via rhyth-

mic high-register “ting” sounds (plucked steel string sam-

ples). Loudness and repetition speed rises as the volume of

deep browsing increases. Again, there are five independent

“ting” sounds, each at a different fixed pitch, representing

the five content areas.

We used pitch groups that were consonant, and for the sounds that

incorporated rhythm (A2–E3), the phase and frequency of each

pulse in the matrix varies independently, yielding a sound with a

changing rhythmic texture but no fixed beat.

The purpose of this sonification is to make interpretable the

activities of users on a Web site. Therefore, the stream of hits be-

ing processed by a Web server (reduced to include only the HTML,

PostScript and PDF documents) needs to be transformed to extract

meaningful user-level data. A real-time monitoring tool was devel-

oped that maintains a bank of active visits (recording separately the

activities of all the people browsing the site at a given time) and

updates various statistics with each user request. When cookies or

some other authentication mechanism allows us to recognize re-

turning visitors, the monitor will update a more complicated user

profile that encapsulates previous browsing patterns. Our traffic

sonification as described above takes as input the location of each

visitor within a site at a given point in time. When constructing

more elaborate sound displays, our design will continue to focus

on user activities, drawing more heavily on the statistics culled

by the monitoring tool. This emphasis distinguishes our approach

from sonification methods that assess Web server performance by

making audible statistics relating to server load, HTTP errors, and

agent types [?].

2.1.2. Impressions and extensions

We have created three audio examples for the activity on the Lu-

cent site. Our data were captured on November 11, 1999 and we

created sonifications of the traffic at 6:00 am, an extremely slow

period for the site; noon, a relatively active time; and 2:30 pm,

the point at which the site was busiest. The samples are located at

our project Web site [6]. Even with this relatively straightforward

mapping, one finds compelling patterns. For example, the affinity

between the /enterprise subdomain and the /search facil-

ity can be heard as the pulses for these areas rise and fall together.3

3 While clearly audible, these shifts can really only be precisely associ-

ated with areas after a certain amount of experience with the mapping.

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-11

Also, when comparing moderately active to extremely busy peri-

ods, we find that the number of people digging deep into the site is

not a fixed fraction of the total number of visitors. That is, the vol-

ume of the low-register drones exhibits much more variation than

the components for the other two categories of accesses. Each of

these effects can be verified by examining the logs, reinforcing the

usefulness of our sonification as a tool for constructing hypotheses

about site traffic.

As mentioned at the beginning of this section, Web browsers

offer a rich set of data about the visitor when requesting data from

a server. This display makes use of only the most basic informa-

tion about a visit, namely the depth of pages accessed. In ongoing

work, we are augmenting our sonification with extra features de-

rived both directly from the server data as well as from statistical

navigation models [12] fit for the Web site under study. So far, we

have found that such extensions are most effective when developed

in the context of a particular monitoring application. For example,

an extended version of this ambient display can aid system archi-

tects of large, Web hosting services understand cache performance

and can aid in server provisioning. Another extension will make

greater user of our navigation models and can help designers and

usability engineers better architect Web sites. We will report on

these and other developments through the project Web site [4].

2.2. Chat rooms and bulletin boards

At any given moment, tens of thousands of real-time conversa-

tions are taking place across the Internet on public forums, bulletin

boards and chat sites. To imagine making these conversations si-

multaneously audible evokes an image of uproarious babble. And

yet, in the aggregate, this massive stream of live communication

could exhibit rich thematic structure. Can we find a meaningful

way to listen in to so many conversations, rendering them in a way

that is comprehensible and not overwhelming?

In some sense, a byproduct of our Web traffic sonification is

the creation of a kind of community from the informal gather-

ing of thousands of visitors to a given Web site. Traditionally,

informational Web sites like www.lucent.com have provided

us with very little sense of the other people who are requesting

data from the server. To attract and retain visitors, however, many

commercial sites recognize the potential of the Web to form so-

cial as well as informational networks. As a result, Web-based fo-

rums, message boards and a variety of chat services are common

components of current site designs. While Internet Relay Chat

(IRC) has been a widely used standard since the inception of the

Internet, the popularization of the Web has resulted in a virtual

explosion of chat applications.4 For example, www.yahoo.com

(a US-based Web portal) offers hundreds of separate chat rooms

attracting tens of thousands of visitors a day. Specialized sites

like www.style.com (the homepage for Vogue magazine) or

www.audiworld.com (an resource for Audi owners) have also

found their message boards to be the most frequently accessed

parts of their domains.

To get a sense of the amount of content that is available in

these dynamic formats, we examined sites contained in the DMOZ

Open Directory [3], an open source listing of over 2 million Web

sites compiled and categorized by 33,000 volunteer editors. From

the November 20, 2000 image of the directory, we counted 36,681

4 RC was developed by Jarkko Oikarinen in Finland in the late eighties,

and was originally intended to work as a better substitute for talk on his

bulletin board.

separate sites offering some kind of chat, bulletin board or other

public forum. While we did not examine the activity on all of

these sites, the number is staggering. If we include other peer-

to-peer communication technologies like instant messaging,5 the

amount of dialogue taking place on the Web at any point in time

is almost unfathomable. The goal of our second sonification is

to make interpretable the thousands of streams of dynamic infor-

mation being generated on the Web. In so doing, we attempt to

characterize a global dialogue, integrating political debates, dis-

cussions of current events, and casual exchanges between mem-

bers of virtual communities.

2.2.1. Content monitors and the statistics engine

Our starting point is text. Albeit diverse in style and dynamic in

character, the text (or transcript) of these data sources carries their

meaning. Therefore, any auditory display consisting only of gen-

erated tones would not be able to adequately represent the data

without a very complex codebook. The design of our sonifica-

tion then depends heavily on text-to-speech (TTS). As with the

traffic example in the previous section, we think of the audio out-

put as another background information stream. The incorporation

of spoken components in the sound design poses new challenges,

both practical and aesthetic. For example, simply voicing every

word taking place in a single chat room can produce too much text

to be intelligible when played in real-time and can quickly exhaust

the listener. Instead, we build a hierarchical representation of the

text streams that relies on statistical processing for content organi-

zation and summarization prior to display.

Before considering sonification design, we first had to cre-

ate specialized software agents that would both discover new chat

rooms and message boards, as well as harvest the content posted

to these sites. (See Figure 1 for an overview of our system ar-

chitecture.) Most bulletin boards and some chat applications use

standard HTML to store visitor contributions. In many cases, a

specific login name is required to gain access to the site. For

these situations, we constructed a content agent in Perl, as this

language provides us the most convenient platform for managing

access details (like cookies). The public chat rooms on sites like

chat.yahoo.com can be monitored in this way. For IRC we

built a configurable Java client that polls a particular server for

active channels. Web sites like www.cnn.com (a popular news

portal) and www.financialchat.com (a financial commu-

nity hosting chat services for day traders) offer several IRC rooms,

some of which are tightly moderated.

In addition to collecting content, each monitoring agent also

summarizes the chat stream, identifying basic topics and updating

statistics about the characteristics of the discussion: What percent-

age of visitors are contributing? How often to they contribute and

at what length? Is the room “on topic,” or are many visitors post-

ing comments on very different subjects? Topics are derived from

the chat stream using a variant of generalized sequence mining [7]

that incorporates tags for the different parts of speech. While the

exact details are beyond the scope of this abstract, a generalized se-

quence is a string of words possibly separated by a wildcard, “\*”.

For example, if we let A, B and C denote specific “contentful”

words (say, nouns, adjectives and adverbs), then AB C , A B C

and A B C are all generalized sequences. The wildcard al-

lows us to identify “Gore \* disputes \* election” from the sentences

5 AOL alone records tens of millions of people using their instant mes-

saging service each month.

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-12

Chat

BB

Chat Chat

BB

Sonification

Engine

Stats

Channel

Audio Right

Channel

Audio Left

Engine

Statistics

Text Feedback

Content

Monitor

Content

Monitor

Content

Monitor

Content

Monitor

Content

Monitor

Figure 1: System architecture overview. A large number of content

streams (Chat = chat rooms; BB = Bulletin boards) are gathered by

specialized agents that transmit them in a homogenized format to

the statistics engine. The statistics engine then distills the streams

into a much smaller number of configurable text streams as well

as a number of descriptive vectors. The sonification engine then

“plays” these text and data streams. The entire systems operates in

real-time.

“Vice President Gore filed papers to dispute the presidential elec-

tion,” “Aides for Gore indicated that he has every reason to dispute

the election”, and “Gore is still deciding whether or not to dispute

the election”.

As many posts to chat rooms contain spelling mistakes and

incorrect grammar, assigning words to different parts of speech is

error-prone. However, unlike most applications of statistical nat-

ural language processing, our content monitors update their sum-

maries each time new material is posted and downweight older

contributions. Because our sonification renders these sources in

real-time, small mistakes have little effect on the power of the over-

all display to convey the ideas being discussed.

Each of the content monitors are periodically polled by the

statistics engine (see Figure 1). This Java-application clusters the

different chat rooms and bulletin boards based on their topic and

numerical summaries. As the topic in a room changes over time,

the statistics engine is constantly updating and reformulating clus-

ter membership. Because a content stream can in fact support

a number of simultaneous discussions (the threads of a bulletin

board, say), we employ a soft-clustering technique. In our initial

work, we have used a mixture-based scheme that determines the

number of clusters with an MDL (Minimum Description Length)

criterion [9]. Each room is then assigned a probability that it be-

longs to the different groups. This model also provides for topic

summarization at the cluster-level. Next, a stochastic framework

was developed to sample representative sentences posted to the

chat or bulletin board. When a discussion is extremely unstruc-

tured, this selection is essentially random sampling from all the

contributions added to the chat since the last polling point. In ad-

dition to textual data streams, the statistics engine is also respon-

sible for communicating the various ingredients for the display to

our sonification engine, Max/MSP [2] (see Figure 1). We have

adopted the Open Sound Control [13] protocol from Center for

New Music and Audio Technologies to transfer data between the

statistics engine (running on a Macintosh with LinuxPPC) and the

sonification engine (running on a Macintosh with OS/9).

2.2.2. Sonification design

As with the previous example (Section 2), our goal is to create

a sonification that is both communicative and listenable. Here we

face the additional challenge of incorporating verbal content. With

TTS annotations, it becomes more difficult to intelligibly convey

more than one layer of information through the audio channel. Our

design incorporates spatialization, pitch and timbral differentia-

tion, and rhythm to achieve clarity in the presentation of the hi-

erarchically structured data coming from the statistics engine.

The auditory display cycles through topic clusters, spending

relatively more time on subjects being actively discussed by the

largest numbers of people. Each different topic is assigned a dif-

ferent pitch group, reinforcing subject changes when they occur.

For each cluster, the statistics engine sends three streams of infor-

mation to the sonification engine:

Topics A continuously updated list of up to ten “topics” (the most

frequently appearing words and phrases – generalized se-

quences – mined from the multiple chat streams associated

with the given cluster; the number of topics is configurable,

but ten was chosen based on timing considerations);

Content samples A selection of sample sentences, identified by

the statistics engine as typical or representative, in which

these topics appear;

Content entropy A vector that represents the changing level of

entropy in the source data.

The topics are spoken by the TTS system6 at regular intervals in

a pitched monotone, and are panned alternately hard left and hard

right in the stereo field, creating a sort of rhythmic “call and re-

sponse.” The sample sentences are panned center, and rendered

with limited inflection (as opposed to the pitched monotone of the

topics). The tonal, rhythmic and spatial qualities of the topics con-

trasts sufficiently with the sample sentences to create two distinctly

comprehensible streams of verbal information.

The entropy vector controls an algorithmic piano score. When

entropy is minimal and the discussion in the chat room or bulletin

board is very focused on one subject, chords are played rhythmi-

cally in time with the rhythmic recitation of the topics. As entropy

increases and the conversations diverge, a Gaussian distribution is

used to expand the number, range and dynamics of notes that fall

between the chords. With this audio component, one can easily

differentiate a well-moderated content source from a more free-

form, public chat without distracting from the TTS annotations.

The piano score also serves a secondary function as an accompa-

niment to the vocal foreground, enhancing the compositional bal-

ance and overall musicality of the sound design.

2.2.3. Sample sonification and impressions

On our project Web site [5], we have a sample chat room sonifi-

cation that cycles through three topics. In this sound file, we are

6 The built-in MacOS TTS capability controlled by Max/MSP.

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-13

listening to the output of only three content monitors. Hence, by

design, each topic is confined to a single site. The first portion of

this example (ending at 1:47 into the sample) concerns the recent

recall of Bridgestone tires and was based on a www.cnn.com

chat room. This discussion was heavily moderated and hence the

backing piano score frequently reduces to a simple rhythm. For our

second topic (from 1:47 to 3:21 of the sample) we recorded chat

exchanges on www.financialchat.com one morning when

Yahoo’s stock opened low. In this example, we hear day traders

frantically exchanging predictions about when Yahoo’s stock will

“bounce.” The final topic in this sample (from 3:21 to the end) is

again from www.cnn.com and treats a recent strike by the Screen

Actor’s Guild and the American Federation of Television and Ra-

dio Artists. This chat room was much less moderated than the

previous CNN chat, and the backing piano score reflects that.

Although this example does not make full use of the clustering

capabilities of the statistics engine, the essence of our sonification

design is clear. The audio display provides an informative and

accessible representation of dynamic, textual content. The topic

and content sample streams are easy to separate, and when placed

in the background, call our attention to important new subjects

being discussed on the Web.

2.2.4. Applications and Extensions

Our sonification provides an audible interface to the (now) massive

amount of dynamic content available on the Web. Given the pre-

processing that takes place in the content monitors and the statis-

tics engine, a simple extension is to provide search-like function-

ality. A user can register interest in a certain topic and “tune”

our display to present only rooms where this subject is being dis-

cussed. The necessary ingredients to implement this feature are

all currently available in the statistics engine. Similarly, one can

easily restrict the sites that are used for the display. When a new

subject appears that draws the user’s interest, it is also trivial to

add a feature that would direct the user’s browser to one or more

chats associated with the topic. As a final extension, we have pro-

vided the content monitors with a configurable list of Web sites

that can be used to help disambiguate elements in the chat stream.

For example, the day traders speak in ticker symbols. Providing

the content monitor with the URL for the ticker symbol look-up

service offered by Yahoo allows the content monitor to weave not

only company names but also recent company-related headlines

directly into the stream fed to the statistics engine.

While we have focused mainly on chat and bulletin boards,

this technology can be applied in other settings. We have begun

collaborating with the designers of a natural language interface for

Web-based help systems. Here, we give voice to the hundreds of

simultaneous conversations taking place between Web site visitors

and the automated help system. A similar display can be imagined

for other natural language interfaces, including search engines like

AskJeeves (www.jeeves.com). In general, the practical appli-

cations of this summarization and auditory display tool abound.

3. CONCLUSION AND COMMENTS ON

COLLABORATIVE RESEARCH

The two applications outlined in this paper are the first outcomes of

a collaboration sponsored by Bell Laboratories and the Brooklyn

Academy of Music under the Arts in Multimedia project (AIM).

The goal of AIM is to bring together researchers (in this case a

statistician) and artists (in this case a sound artist), with the ob-

jective of advancing our separate agendas through collaborative

projects. Our work together is predicated on the notion that so-

phistication both in data treatment and aesthetics are crucial to the

successful design of audio displays. Thus, in each of our exam-

ples, we have endeavored to create a result which communicates

information clearly, yet at the same time sounds well composed

and appealing. Moving forward, it is our intention to apply these

techniques both to practical applications, and also to create a series

of artworks. These artworks will use our sonification techniques

to establish a series of real-time listening posts, both on the Web

and in physical locations. The listening posts will tap in to various

points of interest on the Internet, using sound to reveal patterns and

trends that would otherwise remain hidden.

In terms of applications, we are exploring the use of sonifica-

tion to support the design, provisioning and monitoring of commu-

nication networks. A network operations center (NOC), for exam-

ple, routinely receives clues about the health of the system in the

form of text messages generated by routers and switches. An audio

display installed inside a NOC can act as an early warning system

for approaching bottlenecks as well as aid in troubleshooting. By

continued exposure to the sound of a “normally” functioning net-

work, operators will be alerted to system changes that could signal

problems.

Art emerges unexpectedly from experimentations with new

statistical methods or considerations involving practical applica-

tions; and new tools for data analysis and modeling develop in re-

sponse to artistic concerns. Each of us continues to be surprised by

the connections that emerge from rethinking familiar problems in

a new context. Through our project, we hope to illustrate both the

value of art-technology collaborations as well as their necessity,

especially when finding meaning in complex data.

4. REFERENCES

[1] Visual insights. www.visualinsights.com.

[2] Cycling74. Max/msp. www.cycling74.com.

[3] Open directory project. www.dmoz.com.

[4] Ear to the ground. cm.bell-labs.com/stat/ear.

[5] Ear to the ground, chat example.

cm.bell-labs.com/stat/ear/chat.html.

[6] Ear to the ground, web traffic samples.

cm.bell-labs.com/stat/ear/samples.html.

[7] W. Gaul and L. Schmidt-Thieme. Mining web navigation

path fragments. In Proceedings of the Workshop on Web

Mining for E-Commerce – Challenges and Opportunities,

Boston, MA, August 2000.

[8] M. H. Hansen and B. Rubin. The audiences would be the

artists and their life would be the arts. IEEE MultiMedia,

7(2), April 2000.

[9] M. H. Hansen and B. Yu. Model selection and the princi-

ple of minimum description length. Journal of the American

Statistical Association. To appear.

[10] G. Kramer. An introduction to auditory display. In

G. Kramer, editor, Auditory Display. Addison-Wesley, 1994.

[11] N. Minar and J. Donath. Visualizing the crowds at a web site.

In Proceedings of CHI 99.

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

ICAD01-14

[12] R. Sen and M. H. Hansen. Predicting a web user’s next re-

quest based on log data. Submitted to ASA Student Paper

Competition.

[13] M. Wright. Open sound control. cnmat.berkeley.edu.

Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001

TAXONOMY AND DEFINITIONS FOR SONIFICATION AND AUDITORY DISPLAY

Thomas Hermann

Neuroinformatics Group

Faculty of Technology, Bielefeld University,Bielefeld, Germany

thermann@techfak.uni-bielefeld.de

ABSTRACT

Sonification is still a relatively young research field and

many terms such as sonification, auditory display, aural-

ization, audification have been used without a precise def-

inition. Recent developments such as the introduction of

Model-Based Sonification, the establishment of interactive

sonification and the increased interest in sonification from

arts have raised the need to revisit the definitions in order

to move towards a clearer terminology. This paper intro-

duces a new definition for sonification and auditory display

that emphasizes the necessary and sufficient conditions for

organized sound to be called sonification. It furthermore

suggests a taxonomy, and discusses the relation between vi-

sualization and sonification. A hierarchy of closed-loop in-

teractions is furthermore introduced. This paper aims to ini-

tiate vivid discussion towards the establishment of a deeper

theory of sonification and auditory display.

1. INTRODUCTION

Auditory Display is still a young research field whose birth

may be perhaps best traced back to the first ICAD confer-

ence1 in 1992 organized by Kramer. The resulting proceed-

ings volume “Auditory Display” [1] is still one of the most

important books in the field. Since then a vast growth of in-

terest, research, and initiatives in auditory display and soni-

fication has occurred. The potential of sound to support hu-

man activity, communication with technical systems and to

explore complex data has been acknowledged [2] and the

field has been established and has clearly left its infancy.

As in every new scientific field, the initial use of terms

lacks coherence and terms are being used with diffuse defi-

nitions. As the field matures and new techniques are discov-

ered, old definitions may appear too narrow, or, in light of

interdisciplinary applications, too unspecific. This is what

motivates the redefinitions in this article.

The shortest accepted definition for sonification is from

Barrass and Kramer et al. [2]: “Sonification is the use of

non-speech audio to convey information”. This definition

excludes speech as this was the primary association in the

1see www.icad.org

auditory display of information at that time. The definition

is unclear about what is meant by conveyance of informa-

tion: are real-world interaction sounds sonifications, e.g. of

the properties of an object that is being hit? Is a computer

necessary for its rendition? As a more specific definition,

the definition in [2] continues:

“Sonification is the transformation of data re-

lations into perceived relations in an acoustic

signal for the purposes of facilitating commu-

nication or interpretation.”

It is significant that the emphasis here is put on the pur-

pose of the usage of sound. This automatically distinguishes

sonification from music, where the purpose is not on the

precise perception of what interactions are done with an in-

strument or what data caused the sound, but on an underly-

ing artistic level that operates on a different level. Often, the

word ‘mapping’ has been used interchangeably with ‘trans-

formation’ in the above definition. This, however, suggests

a severe limitation of sonification towards just mappings be-

tween data and sound – which was perfectly fine at the time

of the definition where such a ‘Parameter-Mapping Sonifi-

cation’ was the dominating paradigm.

However, the introduction of Model-Based Sonification

(MBS) [3, 4] demonstrates methods to explore data by us-

ing sound in a way that is very different from a mapping:

in Parameter-Mapping Sonification, data values are mapped

to acoustic attributes of a sound (in other words: the data

‘play’ an instrument), whereas in MBS sonification models

create and configure dynamic processes that do not make

sound at all without external interactions (in other words:

the data is used to build an instrument or sound-capable

object, while the playing is left to the user). The user ex-

cites the sonification model and receives acoustic responses

that are determined by the temporal evolution of the model.

By doing this, structural information is holistically encoded

into the sound signal, and is no longer a mere mapping of

data to sound. One can perhaps state that data are mapped

to the configurations of sound-capable objects, but not that

they are mapped to sound.

Clearly, sonification models implemented according to

MBS are very much in line with the original idea that sonifi-

ICAD08-1

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

cation allows for the discovery of structures in data through

sound. Therefore there is the need to reformulate or adapt

the definition for sonification to better include such uses of

sound, and beyond that hopefully other possible yet-to-be-

discovered linkages between data and sound.

Another challenge for the definition comes from the use

of sonification in the arts and music: recently more and

more artists incorporate methods from sonification in their

work. What implications does this have for the term sonifi-

cation? Think of scientific visualization vs. art: what is the

difference between a painting and a modern visualization?

Both are certainly organized colors on a surface, both may

have aesthetic qualities, yet they operate on a completely

different level: the painting is viewed for different layers

of interpretation than the visualization. The visualization

is expected to have a precise connection to the underlying

data, else it would be useless for the process of interpret-

ing the data. In viewing the painting, however, the focus

is set more on whether the observer is being touched by it

or what interpretation the painter wants to inspire than what

can be learnt about the underlying data. Analogies between

sonification and music are close-by.

Although music and sonification are both organized

sound, and sonifications can sound like music and vice

versa, and certainly sonifications can be ‘heard as’ music

as pointed out in [5], there are important differences which

are so far not manifest in the definition of sonification.

2. A DEFINITION FOR SONIFICATION

This section introduces a definition for sonification in light

of the aforementioned problems. The definition has been

refined thanks to many fruitful discussions with colleagues

as listed in the acknowledgements and shall be regarded as

a new working definition to foster ongoing discussion in the

community towards a solid terminology.

Definition: A technique that uses data as input, and gener-

ates sound signals (eventually in response to optional addi-

tional excitation or triggering) may be called sonification,

if and only if

(C1) The sound reflects objective properties or relations in

the input data.

(C2) The transformation is systematic. This means that

there is a precise definition provided of how the data

(and optional interactions) cause the sound to change.

(C3) The sonification is reproducible: given the same data

and identical interactions (or triggers) the resulting

sound has to be structurally identical.

(C4) The system can intentionally be used with different

data, and also be used in repetition with the same

data.Data Sonification

Algorithm

systematic

transformation reproducable exchangeability

of data

interactions (optional)

Definition: Sonification

Figure 1: Illustration of the general structure and necessary

conditions for sonification. The yellow box depicts besides

the sonification elements few other components of auditory

displays, see also Sec. 3.

This definition emphasizes important prerequisites for

the scientific utility of sonification. It has several partly un-

expected implications that are to be explored in the follow-

ing discussion.

2.1. Discussion

2.1.1. General Comments

Sonification Techniques: According to the above defini-

tion, the techniques Audification, Earcons, Auditory Icons,

Parameter-Mapping Sonification as well as Model-Based

Sonification are all covered by the definition – they all rep-

resent information/data by using sound in an organized and

well-structured way and they are therefore different sonifi-

cation technique.2 This may first appear unfamiliar in light

of the common parlance to see earcons/auditory icons as

different from sonification. However, imagine an auditory

display for biomedical data that uses auditory icons as sonic

events to represent different classes (e.g. auditory icons for

benign/malignant tissue). The sonification would then be

the superposition or mixture of all the auditory icons chosen

for instance according to the class label and organized prop-

erly on the time axis. If we sonify a data set consisting only

of a single data item we naturally obtain as an extreme case

a single auditory icon. The same can be said for earcons.

Although sonification originally has the connotation of rep-

resenting large and complex data sets, it makes sense for the

definition to also work for single data points.

Data vs. Information: A distinction between data and

information is – as far as the above definition – irrelevant.

Think of earcons to represent computer desktop interactions

such as “delete file”, “rename folder”. There can be a lexi-

2they are also covered by the definition of sonification as ‘non-speech

use of sound to convey information’!

ICAD08-2

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

con of terms (file, folder, link) and actions (delete, rename,

etc.), and in practical computer implementations these fea-

tures would be represented numerically, e.g. object = O1,

action = A3. By doing so, the information has been turned

into data, and this is generally done if there is more than

one signal type to give. Information like for instance a

verbal message can always be represented numerically and

thus be understood as data. On the other side, raw data

values often carry semantic interpretation: e.g. the outside

temperature data value -10◦C (a one-dimensional data set

of size 1) – this is cold, and clearly information! Assum-

ing that information is always encoded as data values for

its processing we can deal with both in a single definition.

How the data are then represented by using sound is another

question: whether sonification techniques use a more sym-

bolic or analogic representation according to the analogic-

symbolic continuum of Kramer [6] is secondary for the def-

inition.

Mapping as a specific case of sonification: Some

articles have used “sonification” to refer specifically

to mapping-based sonification, where data features are

mapped to acoustic features of sound events or streams. Yet

sonification is more generally the representation of data by

using sound. There may be times when a clear specifica-

tion of the sonification technique, e.g. as model-based, au-

dification or parameter-mapping sonification, may be help-

ful to avoid confusion with the general term of sonification.

It makes sense to always use the most specific term possi-

ble, that is to use the term Parameter Mapping Sonification,

Audification, Model-Based Sonification, etc. to convey ex-

actly what is meant. The term Sonification, however, is,

according to the definition, more general which is also sup-

ported by many online definitions3. In result we suggest

using sonification with the same level of generality as the

term visualization is used in visual display.

Sonification as algorithm and sound: Sonification

refers to the technique and the process, so basically it refers

to the algorithm that is at work between the data, the user

and the resulting sound. Often, and with equal right, the re-

sulting sounds are called sonifications. Algorithm means a

set of clear rules, independent of whether it is implemented

on a computer or any other way.

Sonification as scientific method: According to the

definition, sonification is an accurate scientific method

which leads to reproducible results, addressing the ear

rather then the eye (as visualization does). This does not

limit the use of sonifications to data from the sciences, but

only states that sonification can be used as a valid instru-

ment to gain insight. The subjectivity in human percep-

3http://en.wikipedia.org/wiki/Sonification,

http://wvvel.csee.wvu.edu/sepscor/sonification/lesson9.html,

http://www.techfak.uni-bielefeld.de/ags/ni/projects/datamining/datason/

datason e.html, http://www.cs.uiowa.edu/ kearney/22c296Fall02/ Critten-

donSpecialty.pdf, to name a few.

tion and interpretation is shared with other perceptualization

techniques that bridge the gap between data and the human

sensory system. Being a scientific method, a prefix like in

“scientific sonification” is not necessary.

Same as some data visualizations may be ‘viewed’ as

art, sonifications may be heard as ‘music’[5], yet this use

differs from the original intent.

2.1.2. Comments to (C1)

(C1) The sound reflects objective properties or

relations in the input data.

Real-world acoustics are typically not a sonification al-

though they often deliver object-property-specific system-

atic sound, since there is no external input data as requested

in C1. For instance, with a bursting bottle, one can identify

what is the data, the model and the sound, but the process

cannot be repeated with the same bottle. However, using

a bottle that fills with rain, hitting it with a spoon once a

minute can be seen as a sonification: The data here is the

amount of rainfall, which is here measured by the fill level,

and the other conditions are also fulfilled. Tuning a guitar

string might also be regarded as a sonification to adjust the

tension of a string4. These examples show that sonifications

are not limited to computer-implementations according to

the definition, which embraces the possibility of other non-

computer-implemented sonifications.

The borders of sonification and real-world acoustics are

fuzzy. It might be discussed how helpful it is to regard or

denote everyday sounds as sonifications.

2.1.3. Comments to (C2)

(C2) The transformation is systematic. This

means that there is a precise definition pro-

vided of how the data (and optional interac-

tions) cause the sound to change.

What exactly do we mean by “precise”? Some sound

generators use noise and thereby random elements so that

sound events will per se sound different on each rendering.

In Parameter-Mapping Sonifications, the intentional addi-

tion of noise (for instance as onset jitter to increase per-

ceptability of events that would otherwise coincide) is often

used and makes sense. In order to include such cases ran-

domness is allowed in the definition, yet it is important to

declare where and what random elements are used (e.g. by

describing the noise distribution). It is also helpful to give

a motivation for the use of such random elements. By us-

ing too much noise, it is possible to generate useless soni-

fications in the sense that they garble interpretation of the

underlying data. In the same way it is possible to create

useless scientific visualizations.

4thanks to the referee for this example!

ICAD08-3

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

2.1.4. Comments to (C3)

(C3) The sonification is reproducible: given the

same data and identical interactions (or trig-

gers) the resulting sound has to be structurally

identical.

The definition claims reproducibility. This may not

strictly be achieved for several reasons: the loudspeakers

may generate a different sound at different temperatures,

other factors such as introduced noise as discussed above

may have been added. The use of the term “structurally

identical” in the definition aims to weaken the stronger

claim of sample-based identity. Sample-based identity is

not necessary, yet all possible psychophysical tests should

come to identical conclusions.

2.1.5. Comments to (C4)

(C4) The system can intentionally be used with

different data, and also be used in repetition

with the same data.

Repeatability is essential for a technique to be scientif-

ically valid and useful – otherwise nobody could check the

results obtained by using sonification as instrument to gain

insight. However, there are some implications by claim-

ing repeatability for what can and cannot be called sonifi-

cation. It has for instance been suggested that a musician

improvising on his instrument produces ‘a sonification of

the musician’s emotional state’. With C4, however, “play-

ing a musical instrument” is not a sonification of the per-

former’s emotional state, since it can not be repeated with

the ‘identical’ data. However, the resulting sound may be

called a sonification of the interactions with the instrument

(regarded here as data), and in fact, music can be heard with

the focus to understand the systematic interaction patterns

with the instruments.

Some of these conditions have been set as constraints

for sonification, e.g. reproducibility in the ‘Listening to the

Mind Listening’ concert5, but not been connected to a defi-

nition of sonification.

In summary, the given definition provides a set of neces-

sary conditions for systems and methods to be called soni-

fication. The definition is neither exhaustive nor complete;

we hope it will serve as the core definition as we as commu-

nity work towards a complete one.

3. SONIFICATION AND AUDITORY DISPLAY

With the above definition, the term sonification takes the

role of a general term to express the method of rendering

5http://www.icad.org/websiteV2.0/Conferences/ICAD2004/concert call.htm

sound in an organized and well-structured way. This is in

good analogy with the term visualization which is also the

general term under which a variety of specific techniques

such as bar charts, scatter plots, graphs, etc. are subsumed.

Particularly there is an analogy between scatter plots where

graphical symbols (data-mapped color/size...) are orga-

nized in space to deliver the visualization, and Parameter-

Mapping Sonification, where in a structurally identical way

acoustic events (with data-mapped features) are organized

in time. It is helpful to have with sonification a term that

operates on the same level of generality as visualization.

This raises the question what then do we mean by au-

ditory displays? Interestingly, in the visual realm, the

term ‘display’ suggests a necessary but complementary part

of the interface chain: the device to generate structured

light/images, for instance a CRT or LCD display or a projec-

tor. So in visualization, the term visualization emphasizes

the way how data are rendered as an image while the display

is necessary for a user to actually see the information. For

auditory display, we suggest to include this aspect of con-

version of sound signals into audible sound, so that an au-

ditory display encompasses also the technical system used

to create sound waves, or more general: all possible trans-

missions which finally lead to audible perceptions for the

user. This could range from loudspeakers over headphones

to bone conduction devices. We suggest furthermore that

auditory display should also include the user context (user,

task, background sound, constraints) and the application

context, since these are all quite essential for the design and

implementation. Sonification is thereby an integral compo-

nent within an auditory display system which addresses the

actual rendering of sound signals which in turn depend on

the data and optional interactions, as illustrated in Fig. 2.

Auditory Displays are more comprehensive than sonifica-Components of Auditory Display Systems

User/Listener

Technical

Sound Display

Sonification

(Rendering)

0101

0100

Application

Context

Data

Usage Context

mobile?

PC?

office?

Interactions

Figure 2: Auditory Displays: systems that employ sonifica-

tion for structuring sound and furthermore include the trans-

mission chain leading to audible perceptions and the appli-

cation context.

ICAD08-4

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

tion since for instance dialogue systems and speech inter-

faces may also be regarded as auditory displays since they

use sound for communication. While such interfaces are not

the primary focus in this research field the terminology sug-

gests their inclusion. On the other hand, Auditory Display

may be seen as a subset of the more general term of Audi-

tory Interfaces which do not only include output interfaces

(auditory displays, sonification) but also auditory input in-

terfaces which engender bidirectional auditory control and

communication between a user and a (in most cases) tech-

nical system (e.g. voice control system, query-by humming

systems, etc.).

4. HIERARCHY FROM SOUND TO

SONIFICATION

So far we have dealt with the necessary conditions sur-

rounding sonification and thus narrowed sonification down

to a specific subset of using sound. In this section, we look

at sonification in a systemic manner to elucidate its super-

ordinate categories. Figure 3 shows how we suggest to or-

ganize the different classes of sound. On the highest level,Map of Sound

Organized Sound

Functional Sounds

Music &

Media Arts Sonification(a)

(b)

Figure 3: Systemic map of sound, showing sonification and

its relation to other categories.

sounds are here classified as Organized Sound and unorga-

nized sound. Organized sounds separate from random or

otherwise complex structured sounds in the fact that their

occurence and structure is shaped by intention. Environ-

mental sounds appear often to be very structured and could

thus also be organized sounds, however, if so, any sound

would match that category to some extent. It thus may be

useful to apply the term to sounds that are intentionally or-

ganized – in most cases by the sound/interface developer.

The set of organized sound comprises two large sets that

partially overlap: music and functional sounds. Music is

without question a complex structured signal, organized on

various levels, from the acoustic signal to its temporal orga-

nization in bars, motifs, parts, layers. It is not our purpose

to give a definition of music.

The second set is functional sounds. These are orga-

nized sounds that serve a certain function or goal [7]. The

function is the motivation for their creation and use. To give

an example, all signal sounds (such as telephones, door-

bells, horns and warning hooters) are functional sounds.

Certainly there are intersections with music, as music can

serve functional aspects. For instance, trombones and kettle

drums have been used to demonstrate kingship and power.

A more subtle function is the use of music in supermarkets

to enhance the ‘shopping mood’. For that reason these sets

overlap – the size of the overlap depends on what is regarded

as function.

Sonification in the sense of the above definition is cer-

tainly a subset of functional sounds. The sounds are ren-

dered to fulfill a certain function, be it communication of in-

formation (signals & alarms), the monitoring of processes,

or to support better understanding of structure in data under

analysis. So is there a difference between functional sounds

and sonification at all? The following example makes clear

that sonification is really a subset: Recently a new selec-

tive acoustic weapon has been used, the mosquito device6,

a loudspeaker that produces a HF-sound inaudible to older

people, which drives away teenagers hanging around in

front of shops. This sound is surely functional, yet it could

neither pass as sonification nor as music.

Finally, we discuss whether sonification has an intersec-

tion with music&media arts. Obviously there are many ex-

amples where data are used to drive aspects of musical per-

formances, e.g. data collected from motion tracking or bio-

sensors attached to a performer. This is, concerning the in-

volved techniques and implementations similar to mapping

sonifications. However, a closer look at our proposed defi-

nition shows that often the condition for the transformation

to be systematic C2 is violated and the exact rules are not

made explicit. But without making the relationship explicit,

the listener cannot use the sound to understand the underly-

ing data better. In addition, condition C4 may often be vio-

lated. If sonification-like techniques are employed to obtain

a specific musical or acoustic effect without transparency

between the used data and details of the sonification tech-

niques, it might, for the sake of clarity, better be denoted

as ‘data-inspired music’, or ‘data-controlled music’ than as

sonification. Iannis Xenakis, for instance, did not even want

the listener to be aware of the data source nor the rules of

sound generation.

6see http://www.compoundsecurity.co.uk/, last seen 2008-01-16

ICAD08-5

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

5. CLOSED INTERACTION LOOPS

IN AUDITORY DISPLAYS

This section emphasizes the role of interaction in sonifica-

tion. We propose different terms depending on the scope of

the closure of the interaction loop. The motivation for this

discussion is that it might be helpful to address how terms

such as biofeedback or interactive sonification relate to each

other.

We start the discussion with Fig. 4 that depicts closed

loop interactions. The sonification module in the upper cen-

ter playing rendered sonifications to the user. Data sources

for sonification enter the box on the left side and the most

important parts are (a) World/System: this comprises any

system in the world that is connected to the sonification

module, e.g. via sensors that measure its state, and (b) Data:

these are any data under analysis or represented information

to be displayed that are stored separately and accessible by

the sonification.World/System

Sonification

Interactive Sonification

Human Activity (supported by sonification)

Auditory Biofeedback

Data

Navigation

Monitoring

No Action

Figure 4: Illustration of Closed-Loop Auditory Systems.

In this setting, Process Monitoring is the least inter-

active sonification, where data recorded from the world (in

real-time) or read from the data repository is continuously

used as input for a sonification rendering process. Here, the

listener is merely passively listening to the sound with the

only active component being his/her focus of attention onto

parts of the sound. Certainly, certain changes in the sound

might attract attention and force the user to act (e.g. sell

stocks, stop a machine, etc...).

A higher degree of active involvement occurs when the

user actively changes and adjusts parameters of the sonifi-

cation module, or interacts otherwise with the sonification

system. We denote this case as Interactive Sonification.

There is a wide field of possibilities of why and how to do

so, and we discuss 3 different prototypical examples:

(a) Triggering: Consider a mapping sonification of a

given data set. An essential interaction for the user

is to issue the command to render/playback the soni-

fication for a selected dataset. Possibly he/she does

this several times in order to attend different parts of

the sound signal. This elementary case is an interac-

tion, however, a very basic one.

(b) Parameter Adjustment is done when the user changes

parameters, such as what data feature are mapped

to acoustic parameters, control ranges, compression

factors, etc. Often such adjustments happen sepa-

rate from the playback so that the changes are made

and afterwards the updated sound is rendered. How-

ever, interactive real-time control is feasible in many

cases and shows a higher degree of interactivity. The

user actively explores the data by generating different

‘views’ of the data [8]. In visualization a similar in-

teractivity is obtained by allowing the user to select

axes scalings, etc.

(c) Excitatory Interaction is the third sort of interaction

and is structurally similar to the case of triggering.

Particularly in Model-Based Sonification [4], usually

the data are used to configure a sound-capable vir-

tual object that in turn reacts on excitatory interac-

tions with acoustic responses whereby the user can

explore the data interactively. Excitation puts energy

into the dynamic system and thus initiates an audible

dynamical system behavior. Beyond a simple trigger-

ing, excitatory interactions can be designed to make

use of the fine-grained manipulation skills that human

hands allow, e.g. by enabling to shake, squeeze, tilt or

deform the virtual object, for instance using sensor-

equipped physical interfaces to interact with the soni-

fication model. A good example for MBS is Shoogle

by Williamson et al. [9], where short text messages

in a mobile phone can be overviewed by shaking a

mobile phone equipped with accelerometer sensors,

resulting in audible responses of the text messages

as objects moving virtually inside the phone. Excita-

tory interactions offer rich and complex interactions

for interactive sonification.

The next possibility for a closed loop is by interactions

that select or browse data. Since data are chosen, it may

best be referred to as Navigation. Navigation can also be

regarded as special case of Interactive Sonification, depend-

ing on where the data are selected and the borders are here

really soft. Navigation usually goes hand in hand with trig-

gering of sonification (explained above).

Auditory Biofeedback can be interpreted as a sonifi-

cation of measured sensor data. In contrast to the above

types, the user’s activity is not controlling an otherwise au-

tonomous sonification with independent data, but it pro-

duces the input data for the sonification system. The user

perceives a sound that depends on his/her own activity.

Such systems have applications that range from rehabilita-

tion training to movement training in sports, e.g. to perform

ICAD08-6

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

a complex motion sequence (e.g. a tennis serve) so that its

sonification is structurally more similar to the sonification

of an expert performing the action [10].

The final category is Human Activity, which means

that the interaction ranges beyond the sonification system

into the world, often driven by the goal to change a world

state in a specific way. In turn, any sensors that pick up the

change may lead to changes in the sonification. The differ-

ence between the loop types before is that the primary fo-

cus is to achieve a goal beyond the sonification system, and

not to interact with a closed-loop sonification system. Even

without attending the sonification consciously or primarily,

the sound can be helpful to reach the goal. For example,

imagine the real-world task to fill a thermos bottle with tea.

While your primary goal is to get the bottle filled you will

receive the ‘gluck-gluck’ sound with increasing pitch as a

by-product of the interaction. If this is consistently useful,

you subconsciously adapt your activity to exploit the cues in

the sound – but the sound is only periphery for the goal. In a

similar sense, sonifications may deliver helpful by-products

to actions that change the world state. We regard such in-

teraction add-ons where sonification is a non-obtrusive yet

helpful cue for goal attainment as inspiring design direc-

tion. Such sonifications might even become subliminal in

the sense that users, when asked about the sound, are not

even aware of the sound, yet they perform better with sound

than without.

6. DISCUSSION AND CONCLUSION

The definitions in this paper are given on the basis of

three goals: (i) to anchor sonification as a precise scien-

tific method so that it delivers reproducible results and thus

can be used and trusted as instrument to obtain insight into

data under analysis. (ii) to offer a generalization which does

not limit itself to the special case of mappings from data to

sound, but which introduces sonification as general system-

atic mediator between data and sound, whatever the repre-

sentation might be. (iii) to balance the definition so that the

often-seen pair of terms ‘visualization & sonification’ are at

the same level of generality.

The definition has several implications which have been

discussed in Sec. 2. We’d like to emphasize that this effort

is being done in hope that the definition inspires a general

discussion on the terminology and taxonomy of the research

field of auditory display. An online version of the definition

is provided at www.sonification.de with the aim to collect

comments and examples of sonifications as well as exam-

ples that are agreed not to be sonifications and which help

in turn to improve the definition.

In Section 3, we described integral parts for auditory

display so that sonification takes a key component as the

technical part involving the rendition of sound. Again, the

suggested modules are meant as working hypothesis to be

discussed at ICAD.

While the given definitions specified terms on a horizon-

tal level, Section 4 proposes a vertical organization of sound

in relation to often used terms. The intersections between

the different terms and categories have been addressed with

examples.

Finally, we have presented in Section 5 an integrative

scheme for organizing different classes of auditory closed

loops according to the loop closure scope. It proves help-

ful to clarify classes of interactive sonifications. We think

that grouping existing sonifications according to these cat-

egories can be helpful to better find alternative approaches

for a given task.

The suggested terminology and taxonomy is the result

of many discussions and a thorough search for helpful con-

cepts. We suggest it as working definitions to be discussed

at the interdisciplinary level of ICAD in hope to contribute

towards a maturing of the fields of auditory display and

sonification.

7. ACKNOWLEDGEMENT

Many collegues have been very helpful in discussions to

refine the definitions. Particularly, I thank Till Bover-

mann, Arne Wulf, Andy Hunt, Florian Grond, Georg

Spehr, Alberto de Campo, Gerold Baier, Camille Peres,

and in particular Gregory Kramer for the helpful discus-

sions on the definition for sonification. Thanks also to col-

legues of the COST IC0601 Sonic Interaction Design (SID)

WG4/Sonification. I also thank Arne Wulf for the inspiring

discussions on Closed-Loop Auditory Systems, and Louise

Nickerson for many language improvements.

8. REFERENCES

[1] G. Kramer, Ed., Auditory Display - Sonification, Au-

dification, and Auditory Interfaces. Addison-Wesley,

1994.

[2] G. Kramer, B. Walker, T. Bonebright, P. Cook,

J. Flowers, N. Miner, and J. Neuhoff, “Sonification

report: Status of the field and research agenda,” Tech.

Rep., International Community for Auditory Display,

1999, http://www.icad.org/websiteV2.

0/References/nsf.html.

[3] Thomas Hermann and Helge Ritter, “Listen to your

data: Model-based sonification for data analysis,” in

Advances in intelligent computing and multimedia sys-

tems, G. E. Lasker, Ed., Baden-Baden, Germany, 08

1999, pp. 189–194, Int. Inst. for Advanced Studies in

System research and cybernetics.

ICAD08-7

Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008

[4] Thomas Hermann, Sonification for Exploratory Data

Analysis, Ph.D. thesis, Bielefeld University, Bielefeld,

Germany, 02 2002.

[5] Paul Vickers and Bennett Hogg, “Sonification ab-

straite/sonification concr`ete: An ‘æsthetic perspective

space’ for classifying auditory displays in the ars mu-

sica domain,” in ICAD 2006 - The 12th Meeting of the

International Conference on Auditory Display, Alis-

tair D N Edwards and Tony Stockman, Eds., London,

UK, June 20-23 2006, pp. 210–216.

[6] G. Kramer, “An introduction to auditory display,” in

Auditory Display, G. Kramer, Ed. ICAD, 1994, pp. 1–

79, Addison-Wesley.

[7] Georg Spehr, SOUND STUDIES. Traditionen - Meth-

oden – Desiderate, chapter Funktionale Kl ̈ange - Mehr

als ein Ping, transcript Verlag, Bielefeld, Germany,

2008.

[8] Thomas Hermann and Andy Hunt, “The discipline of

interactive sonification,” in Proceedings of the Inter-

national Workshop on Interactive Sonification (ISon

2004), Thomas Hermann and Andy Hunt, Eds., Biele-

feld, Germany, 01 2004, Bielefeld University, Interac-

tive Sonification Community, peer-reviewed article.

[9] John Williamson, Rod Murray-Smith, and S. Hughes,

“Shoogle: excitatory multimodal interaction on mo-

bile devices,” in Proceedings of the SIGCHI confer-

ence on Human factors in computing systems, San

Jose, California, USA, 2007, pp. 121–124, ACM

Press.

[10] Thomas Hermann, Oliver H ̈oner, and Helge Rit-

ter, “Acoumotion - an interactive sonification sys-

tem for acoustic motion control,” in Gesture in

Human-Computer Interaction and Simulation: 6th In-

ternational Gesture Workshop, GW 2005, Berder Is-

land, France, May 18-20, 2005, Revised Selected Pa-

pers, Sylvie Gibet, Nicolas Courty, and Jean-Francois

Kamp, Eds., Berlin, Heidelberg, 2006, vol. 3881/2006

of Lecture Notes in Computer Science, pp. 312–323,

Springer.

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/221513907

Nomadic Radio: Scaleable and Contextual Notification for Wearable Audio

Messaging.

Conference Paper · January 1999

DOI: 10.1145/302979.303005 · Source: DBLP

CITATIONS

95

READS

153

2 authors, including:

Some of the authors of this publication are also working on these related projects:

Bayesian Modeling for Human Development Indicators View project

Aware Community Portals View project

Nitin Sawhney

Aalto University

55 PUBLICATIONS 1,560 CITATIONS

SEE PROFILE

All content following this page was uploaded by Nitin Sawhney on 30 March 2015.

The user has requested enhancement of the downloaded file.

Papers CHI 99 15-20 MAY 1999

Nomadic Radio: Scaleable and Contextual Notification

for Wearable Audio Messaging

Nitin Sawhney and Chris Schmandt

SpeechInterface Group, MIT Media Laboratory

20 Ames St., Cambridge, MA 02139

{ nitin, geek} @media.mit.edu

ABSTRACT

Mobile workers need seamless access to communication

and information services on portable devices. However

current solutions overwhelm users with intrusive and

ambiguous notifications. In this paper, we describe

scaleable auditory techniques and a contextual notification

model for providing timely information, while minimizing

interruptions. User’s actions influence local adaptation in

the model. These techniques are demonstrated in Nomadic

Radio, an audio-only wearable computing platform.

Keywords

Auditory I/O, passive awareness, wearable computing,

adaptive interfaces, interruptions, notifications

INTRODUCTION

In today’s information-rich environments, people use a

number of appliances and portable devices for a variety of

tasks in the home, workplace and on the run. Such devices

are ubiquitous and each plays a unique functional role in a

user’s lifestyle. To be effective, these devices need to notify

users of changes in their functional state, incoming

messages or exceptional conditions. In a typical office

environment, the user attends to a plethora of devices with

notifications such as calls on telephones, asynchronous

messages on pagers, email notification on desktop

computers, and reminders on personal organizers or

watches. This scenario poses a number of key problems.

Lack of Differentiation in Notification Cues

Every device provides some unique form of notification. In

many cases, these are distinct auditory cues. Yet, most cues

are generally binary in nature, i.e. they convey only the

occurrence of a notification and not its urgency or dynamic

state. This prevents users from making timely decisions

about received messages without having to shift focus of

attention (from the primary task) to interact with the device

and access the relevant information.

Permission to make digital or hard copies of all or part of this work for

personal or classroom use is granted without fee provided that copies

arc not made or distributed for profit or commercial advantage and that

copies bear this notice and the full citation on the first page. To copy

otherwise, to republish, to post on servers or to redistribute to lists.

requires prior specific permission and/or a fee.

CHI ‘99 Pittsburgh PA USA

Copyright ACM 1999 0-201-48559-1/99/05...$5.00

Minimal Awareness of the User and Environment

Such notifications occur without any regard to the user’s

engagement in her current activity or her focus of attention.

This interrupts a conversation or causes an annoying

disruption in the user’s task and flow of thoughts. To

prevent undue embarrassment in social environments, users

typically turn off cell-phones and pagers in meetings or

lectures. This prevents the user from getting notification of

timely messages and frustrates people trying to get in touch

with her.

No Learning from Prior Interactions with User

Such systems typically have no mechanism to adapt their

behavior based on the positive or negative actions of the

user. Pagers continue to buzz and cell-phones do not stop

ringing despite the fact that the user may be in a

conversation and ignoring the device for some time.

Lack of Coordinated Notifications

All devices compete for a user’s undivided attention without

any coordination and synchronization of their notifications.

If two or more notifications occur within a short time of

each other, the user gets confused or frustrated. As people

start carrying around many such portable devices, frequent

and uncoordinated interruptions inhibit their daily tasks and

interactions in social environments.

Given these problems, most devices fail to serve their

intended purpose of notification or communication, and

thus do not operate in an efficient manner for a majority of

their life cycle. New users choose not to adopt such

technologies, having observed the obvious problems

encountered with their usage. In addition, current users tend

to turn off the devices in many situations, inhibiting the

optimal operation of such personal devices.

Nature of Interruptions in the Workplace

A recent observational study [4] evaluated the effect of

interruptions on the activity of mobile professionals in their

workplace. An interruption, defined as an asynchronous and

unscheduled interaction, not initiated by the user, results in

the recipient discontinuing the current activity. The results

revealed several key issues. On average, sub.jects were

interrupted over 4 times per hour, for an average duration

slightly over 2 minutes. Hence, nearly 10 minutes per hour

96

C H I 9 9 1 5 - 2 0 M A Y 1 9 9 9 Papers

was spent on interruptions. Although a majority of the

interruptions occurred in a face-to-face setting, 20% were

due to telephone calls (no email or pager activity was

analyzed in this study). In 64% of the interruptions, the

recipient received some benefit from the interaction. This

suggests that a blanket approach to prevent interruptions,

such as holding all calls at certain times of the day, would

prevent beneficial interactions from occurring. However in

41% of the interruptions, the recipients did not resume the

work they were doing prior to it. But active use of new

communication technologies makes users easily vulnerable

to undesirable interruptions.

These interruptions constitute a significant problem for

mobile professionals using tools such as pagers, cell-phones

and PDAs, by disrupting their time-critical activities.

Improved synchronous access using these tools benefits

initiators but leaves recipients with little control over the

interactions. The study suggests development of improved

filtering techniques that are especially light-weight, i.e.

don’t require more attention from the user and are less

disruptive than the interruption itself. By moving

interruptions to asynchronous media, messages can be

stored for retrieval and delivery at more appropriate times.

NOMADIC RADIO: WEARABLE AUDIO MESSAGING

Personal messaging and communication, demonstrated in

Nomadic Radio, provides a simple and constrained problem

domain in which to develop and evaluate a contextual

notification model. Messaging requires development of a

model that dynamically selects a suitable notification

strategy based on message priority, usage level, and

environmental context. Such a system must infer the user’s

attention by monitoring her current activities such as

interactions with the device and conversations in the room.

The user’s prior responses to notifications must also be

taken into consideration to adapt the notifications over time.

In this paper, we will consider techniques for scaleable

auditory presentation and an appropriate parameterized

approach towards contextual notification.

Several recent projects utilized speech and audio I/O on

wearable devices to present information. A prototype

augmented audio tour guide [l] played digital audio

recordings indexed by the spatial location of visitors in a

museum. SpeechWear [11] enabled users to perform data

entry and retrieval using speech recognition and synthesis.

Audio Aura [10] explored the use of background auditory

cues to provide serendipitous information coupled with

people’s physical location in the workplace. In Nomadic

Radio, the user’s inferred context rather than actual location

is used to decide when and how to deliver scaleable audio

notifications. In a recent paper [13], researchers suggest the

use of sensors and user modeling to allow wearables to

infer when users should be interrupted by incoming

messages. They suggest waiting for a break in the

conversation to post a message summary on the user’s

heads-up display. In this paper we describe a primarily non-

visual approach to provide timely information to nomadic

listeners, based on a variety of contextual cues.

Nomadic Radio is a wearable computing platform that

provides a unified audio-only interface to remote services

and messages such as email, voice mail, hourly news

broadcasts, and personal calendar events. These messages

are automatically downloaded to the device throughout the

day and users can browse through them using voice

commands and tactile input. The system consists of Java-

based clients and remote servers (written in C and Perl) that

communicate over wireless LAN, and utilize the telephony

infrastructure in the Speech Interface group. Simultaneous

spatial audio streams are rendered using a HRTF-based

Java audio API. Speech I/O is provided via a networked

implementation of AT&T Watson Speech API.

To provide a hands-free and unobtrusive interface to a

nomadic user, the system primarily operates as a wearable

audio-only device. The SoundBeam Neckset, a research

prototype patented by Nortel for use in hands-free

telephony, was adapted as the primary wearable platform in

Nomadic Radio. It consists of two directional speakers

mounted on the user’s shoulders, and a directional

microphone placed on the chest (see figure 1). Here

information and feedback is provided to the user through a

combination of auditory cues, spatial audio rendering, and

synthetic speech. Integration of a variety of auditory

techniques on a wearable device provides hands-free access

and navigation as well as lightweight and expressive

notification.

An audio-only interface has been incorporated in Nomadic

Radio, and a networked infrastructure for unified messaging

has been developed for wearable access [12]. The system

currently operates on a Libretto 100 mini-portable PC worn

by the user. The key issue addressed in this paper is that of

handling interruptions to the listener in a manner that

reduces disruption, while providing timely notifications for

contextually relevant messages.

P a p e r s

USAGE AND NOTIFICATION SCENARIO

The following scenario demonstrates the audio interface

and presentation of notifications in Nomadic Radio (no

voice commands from the user are shown here).

CHI 99 15-20 MAY 1999

SCALEABLE AUDITORY PRESENTATION

A scaleable presentation is necessary for delivering

sufficient information while minimizing interruption to the

listener. Messages in Nomadic Radio are scaled

dynamically to unfold as seven increasing levels of

notification (see figure 3): silence, ambient cues, auditory

cues, message summary, preview, full body, and foreground

rendering. These are described further below:

Silence for Least Interruption and Conservation

In this mode all auditory cues and speech feedback are

turned-off. Messages can be scaled down to silence when

the message priority is inferred to be too low for the

message to be relevant for playback or awareness to a user,

based on her recent usage of the device and the

conversation level. This mode also serves to conserve

processing, power and memory resources on a portable

device or wearable computer.

Ambient Cues for Peripheral Awareness

In Nomadic Radio, ambient auditory cues are continuously

played in the background to provide an awareness of the

operational state of the system and ongoing status of

messages being downloaded (see figure 4). The sound of

flowing water provides an unobtrusive form of ambient

awareness that indicates the system is active (silence

indicates sleep mode). Such a sound tends to fade into the

perceptual background after a short time, so it does not

distract the listener. The pitch is increased during file

downloads, momentarily foregrounding the ambient sound.

A short e-mail message sounds like a splash while a two-

minute audio news summary is heard as faster flowing

water while being downloaded. This implicitly indicates

message size without the need for additional audio cues and

prepares the listener to hear (or deactivate) the message

before it becomes available. Such peripheral awareness

minimizes cognitive overhead of monitoring incoming

messages relative to notifications played as distinct auditory

cues, which incur a somewhat higher cost of attention on

part of the listener.

Related Work in Auditory Awareness

In ARKola [5], an audio/visual simulation of a bottling

factory, repetitive streams of sounds allowed people to keep

track of activity, rate, and functioning of running machines.

Without sounds people often overlooked problems; with

auditory cues, problems were indicated by the machine’s

sound ceasing (often ineffective) or via distinct alert

sounds. The various auditory cues (as many as 12 sounds

play simultaneously) merged as an auditory texture, allowed

people to hear the plant as a complex integrated process.

Background sounds were also explored in ShareMon [3], a

prototype application that notified users of file sharing

activity. Cohen found that pink noise used to indicate

%CPU time was considered “obnoxious”, even though

users understood the, pitch correlation. However,

preliminary reactions to wave sounds were considered

positive and even soothing. In Audio Aura [IO], alarm

CHI 9 9 1 5 - 2 0 M A Y 1 9 9 9 P a p e r s

sounds were eliminated and a number of “harmonically

coherent sonic ecologies” were explored, mapping events to

auditory, musical or voice-based feedback. Such techniques

were used to passively convey the number of email

messages received, identity of senders, and abstract

representations of group activity.

Auditory Cues for Notification and Identification

In Nomadic Radio, auditory cues are a crucial means for

conveying awareness, notification and providing necessary

assurances in its non-visual interface. Different types of

auditory techniques provide distinct feedback, awareness

and message information.

Feedback Cues

Several types of audio cues indicate feedback for a number

of operational events in Nomadic Radio:

1. Task completion and confirmations - button pressed,

speech understood, connected to servers, finished

playing or loaded/deleted messages.

2. Mode transitions - switching categories, going to

non-speech or ambient mode.

3. Exceptional conditions - message not found, lost

connection with servers, and errors.

Priority Cues for Notification

In a related project, “email glances” [7] were formulated as

a stream of short sounds indicating category, sender and

content flags (from keywords in the message). In Nomadic

Radio, message priority inferred from email content

filtering provides distinct auditory cues (assigned by the

user) for group, personal, timely, and important messages.

In addition, auditory cues such as telephone ringing indicate

voice mail, whereas an extracted sound of a station

identifier indicates a news summary.

VoiceCues for Identification

VoiceCues represent a novel approach for easy

identification of the sender of an email, based on a unique

auditory signature of the person. VoiceCues are created by

manually extracting a l-2 second audio sample from the

voice messages of callers and associating them with their

respective email login. When a new email message arrives,

the system queries its database for a related VoiceCue for

that person before playing it to the user as a notification,

along with the priority cues. The authors have found

VoiceCues to be a remarkably effective method for quickly

conveying the sender of the message in a very short

duration. This technique reduces the need for synthetic

speech feedback, which can often be distracting.

99

Papers CHI 99 15-20 MAY 1999

Message Summary Generation

A spoken description of an incoming message can present

relevant information in a concise manner. Such a

description typically utilizes header information in email

messages to convey the name of the sender and the subject

of the message. In Nomadic Radio, message summaries are

generated for all messages, including voice-mail, news and

calendar events. The summaries are augmented by

additional attributes of the message indicating category,

order, priority, and duration. For audio sources, like voice

messages and news broadcasts, the system plays the first

2.5 seconds of the audio. This identifies the caller and the

urgency of the call, inferred from intonation in the caller’s

voice or provides a station identifier for news summaries.

Message Previews using Content Summarization

Messages are scaled to allow listeners to quickly preview

the contents of an email or voice message. In Nomadic

Radio, a preview for text messages extracts the first 100

characters of the message (a default size that can be user

defined). This heuristic generally provides sufficient

context for the listener to anticipate the overall message

theme and urgency. For email messages, redundant headers

and previous replies are eliminated from the preview for

effective extraction. Use of text summarization techniques,

based on tools such as ProSum’ developed by British

Telecom, would allow more flexible means of scaling

message content. Natural language parsing techniques used

in ProSum permit a scaleable summary of an arbitrarily

large text document.

A preview for an audio source such as a voice message or

news broadcast presents a fifth of the message at a

gradually increasing playback rate of up to 1.3 times faster

than normal. There are a range of techniques for time-

compressing speech without modifying the pitch, however

twice the playback rate usually makes the audio

incomprehensible. A better representation for content

summarization requires a structural description of the audio,

based on annotated or automatically determined pauses in

speech, speaker and topic changes. Such an auditory

thumbnail must function similar to its visual counterpart. A

preview for a structured voice message would provide

pertinent aspects such as name of caller and phone number,

whereas a structured news preview would be heard as the

hourly headlines.

Full Body: Playing Complete Message Content

This mode plays the entire audio file or reads the full text of

the message at the original playback rate. Some parsing of

the text is necessary to eliminate redundant header

information and format tags. The message is augmented

with summary information indicating sender and subject.

This message is generally spoken or played in the

background of the listener’s audio space.

I http://transend.Iabs.bt.com/prosum/on-line/

Foreground Rendering via Spatial Proximity

An important message is played in the foreground of the

listening space. The audio source of the message is rapidly

moved closer to the listener, allowing it to be heard louder,

and played there for 415” of its duration. The message

gradually begins to fade away, moving back to its original

position and amplitude for the remaining l/S” of the

duration. The foregrounding algorithm ensures that the

messages are quickly brought into perceptual focus by

pulling them to the listener rapidly. However the messages

are pushed back slowly to provide an easy fading effect as

the next one is heard. As the message moves its spatial

direction is maintained so that the listener can retain a focus

on the audio source even if another begins to play.

Hence a range of techniques provide scaleable forms of

background awareness, auditory notification, spoken

feedback and foreground rendering of incoming messages.

CONTEXTUAL NOTIFICATION

In Nomadic Radio, context dynamically scales the

notifications for incoming messages. The primary

contextual cues used include: message priority from email

filtering, usage level based on time since last user action,

and the likelihood of conversation estimated from real-time

analysis of the auditory scene. In our experience these

parameters provide sufficient context to scale notifications,

however data from motion or location sensors can also be

integrated in such a model. A linear and scaleable auditory

notification model is utilized, based on the notion of

estimating costs of interruption and the value of information

to be delivered to the user. This approach is similar to

recent work [6] on using perceptual costs and a focus of

attention model for scaleable graphics rendering.

Message Priority

The priority of incoming messages is explicitly determined

via content-based email filtering using CLUES [9], a

filtering and prioritization system. CLUES has been

integrated into Nomadic Radio to determine the timely

nature of messages by finding correlation between a user’s

calendar, rolodex, to-do list, as well as a record of outgoing

messages and phone calls. These rules are integrated with

static rules created by the user for prioritizing specific

people or message subjects. When a new email message

arrives, keywords from its sender and. subject header

information are correlated with static and generated

filtering rules to assign a priority to the message. Email

messages are also prioritized if the user is traveling and

meeting others in the same geographic area (via area codes

in the rolodex). The current priorities include: group,

personal, very important, most important, and timely.

Priorities are parameterized by logarithmically scaling all

priorities within a range of 0 to 1. Logarithmic scaling

ensures that higher priority messages are weighted higher

relative to unimportant or uncategorized messages.

Priority ( i ) = ( log ( i ) / log (Priority Levels Mu ) )

100

C H I 9 9 1 5 - 2 0 M A Y 1 9 9 9 Papers

Usage Level

One problem with using last actions for setting usage levels

is that if a user deactivates an annoying message, that

action is again time-stamped. Such negative reinforcements

continue to increase the usage level and the related

notification. Therefore negative actions such as stopping

audio playback or deactivating speech are excluded from

generating actions for computing the usage.

Likelihood of Conversation

Conversation in the environment can be used to gauge

whether the user is in a social context where an

interruption is less appropriate. If the system detects the

occurrence of more than several speakers over a period of

time, that is an indication of a conversational situation.

Auditory events are first detected by adaptively

thresholding total energy and incorporating constraints on

event length and surrounding pauses. The system uses mel-

scaled filter-bank coefficients (MFCs) and pitch estimates

to discriminate, reasonably well, a variety of speech and

non-speech sounds. HMMs (Hidden Markov Models)

capture both the temporal characteristics and spectral

content of sound events. The techniques for feature

extraction and classification of the auditory scene using

HMMs are described in a recent workshop paper [2]. The

likelihood of speech detected in the environment is

computed for each event in a short window of time. In

addition, the probabilities are weighted, such that most

recent time periods in the window are considered more

relevant for computing the overall Speech Level. We are

evaluating the classifier’s effectiveness by training it with a

variety of speakers and background sounds.

Notification Level

A weighted average for all three contextual cues provides

level has an inversely proportional relationship with

notification i.e. a lower notification must be provided

during high conversation.

Presentation Latency

Latency represents the period of time to wait before

playing the message to the listener, after a notification cue

is delivered. Latency is computed as a function of the

notification level and the maximum window of time

(Latency,& that a lowest priority message can be delayed

for playback. The default maximum latency is set to 20

seconds, but can be modified by the user.

101

CHI99 15-20 MAY 1999

were increased. Jane was notified of a group message

shortly after the voice message, since the system detected

higher usage activity. Hence, the system correctly scaled

down notifications when Jane did not want to be bothered

whereas notifications were scaled up when Jane started to

use the system to browse her messages.

EFFECTIVENESS OF THE NOTIFICATION MODEL

The nature of peripheral awareness and unobtrusive

notification on a wearable device requires a usage

evaluation that must be conducted on an ongoing and long-

term basis. However, the predictive effectiveness of the

notification model must first be evaluated on a quantitative

basis. Hence, all message and notification parameters are

captured for such analysis. Lets consider two actual

examples of notification computed for email messages with

different priorities. Figure 7 shows an auditory cue

generated for a group message (low priority).

The timely message (in figure 8) received greater priority

and consequently a higher notification level for summary

playback. A moderate latency time (approx. 6 secs.) was

chosen. However when the user interrupted the notification

by a button press, the summary playback was aborted. The

user’s action reduced overall weights by 5%.

P a p e r s

Dynamic Adaptation of the Notification Model

The user can initially set the weights for the notification

model to high, medium, or low (interruption). These weight

settings were selected by experimenting with notifications

over time using an interactive visualization of message

parameters. This allowed us to observe the model, modify

weights and infer the effect on notification based on

different weighting strategies. Pre-defined weights provide

an approximate behavior for the model and help bootstrap

the system for novice users. The system also allows the user

to dynamically adjust these weights (changing the

interruption and notification levels) by their implicit actions

while playing or ignoring messages.

The system allows localized positive and negative

reinforcement of the weights by monitoring the actions of

the user during notifications. As a message arrives, the

system plays an auditory cue if its computed notification

level is above the necessary threshold for auditory cues. It

then uses the computed latency interval to wait before

playing the appropriate summary or preview of the

message. During that time, the user can request the message

be played earlier or abort any further notification for the

message via speech or button commands. If aborted, all

weights are reduced by a fixed percentage (default is 5%), a

negative reinforcement. If the user activates the message

(positive reinforcement) within 60 seconds after the

notification, the playback scale selected by the user is used

to increase all weights. If the message is ignored, no change

is made to the weights, but the message remains active for

60 seconds during which the user’s actions can continue to

influence the weights.

Figure 6 shows a zoomed view of the extended scenario

introduced earlier, focusing on Jane’s actions that reinforce

the model. Jane received several messages and ignored

most of the group messages and a recent personal message

(the weights remain unchanged). While in the meeting, Jane

interrupted a timely message to abort its playback. This

reduced the weights for future messages, and the ones with

low priority (group message) were not notified to Jane. The

voice message from Kathy, her daughter, prompted Jane to

reinforce the message by playing it. In this case, the weights

Continuous local reinforcement over time should allow the

system to reach a state where it is somewhat stable and

robust in converging to the user’s preferred notification.

Currently the user’s actions primarily adjust weights for

subsequent messages, however effective reinforcement

learning requires a model that generalizes a notification

policy that maximizes some long-term measure of

reinforcement [8]; this will be the focus of our future work.

Cl-II 99 15-20 MAY 1999 Papers

PRELIMINARY EVALUATION

Although the authors have been using and relining these

techniques during system development, a preliminary 2-day

evaluation was conducted with a novice user, who had prior

experience with mobile phones and 2-way pagers. The user

was able to listen to notifications while attending to tasks in

parallel such as reading or typing. He managed to have

casual discussions with others while hearing notifications;

however he preferred turning off all audio during an

important meeting with his advisor. People nearby

sometimes found the spoken feedback distracting if heard

louder, however that also cued them to wait before

interrupting the user. The volume on the device was

lowered to minimize any disruption to others and maintain

the privacy of messages. The user requested an automatic

volume gain that adapted to the environmental noise level.

In contrast to speech-only feedback, the user found the

unfolding presentation of ambient and auditory cues

allowed sufficient time to switch attention to the incoming

message. Familiarization with the auditory cues was

necessary. He preferred longer and gradual notifications

rather than distinct auditory tones. The priority cues were

the least useful indicator whereas VoiceCues provided

obvious benefit. Knowing the actual priority of a message

was less important than simply having it presented in the

right manner. The user suggested weaving message priority

into the ambient audio (as increased pitch). He found the

overall auditory scheme somewhat complex, preferring

instead a simple notification consisting of ambient

awareness, Voice&es and spoken text.

The user stressed that the ambient audio provided the most

benefit while requiring least cognitive effort. He wished to

hear ambient audio at all times to remain reassured that the

system was still operational. An unintended effect

discovered was that a “pulsating” audio stream indicated

low battery power on the wearable device. A “pause” button

was requested, to hold all messages while participating in a

conversation, along with subtle but periodic auditory alerts

for unread messages waiting in queue. The user felt that

Nomadic Radio provided appropriate awareness and its

expressive qualities justified its use over a pager. A long-

term trial with several nomadic users is necessary to further

validate these notification techniques.

CONCLUSIONS

We have demonstrated techniques for scaleable auditory

presentation and message notification using a variety of

contextual cues. The auditory techniques and notification

model have been refined based on continuous usage by the

authors, however we are currently conducting additional

evaluations with several users. Ongoing work explores

adaptation of the notification model based on reinforcement

from user behavior over time. Our efforts have focused on

wearable audio platforms, however these ideas can be

readily utilized in consumer devices such as pagers, PDAs

and mobile phones to minimize disruptions while providing

timely information to users on the move.

ACKNOWLEDGMENTS

Thanks to Brian Clarkson for ongoing work on the audio

classifier and Stefan Marti for help with user evaluations.

We also thank Lisa Fast and Andre Van Schyndel at Nortel

for their support of the project.

REFERENCES

1.

2.

3.

4.

5.

6.

7.

8.

9.

Bederson, Benjamin B. Audio Augmented Reality: A

Prototype Automated Tour Guide. Proceedings of CHI

‘95, May 1995, pp. 210-211.

Clarkson, Brian, Nitin Sawhney and Alex Pentland.

Auditory Context Awareness via Wearable Computing,

Workshop on Perceptual User Inte$aces, Nov. 1998.

Cohen, J. Monitoring Background Activities. Auditory

Display: Sonitication, Audification, and Auditory

Interfaces. Reading MA: Addison-Wesley, 1994.

Conaill, 0’ Brid and David Frohlich. Timespace in the

Workplace: Dealing with Interruptions. Proceedings of

CHI ‘95, 1995.

Gaver, W.W., R. B. Smith, T. OShea. Effective Sounds

in Complex Systems: The ARKola Simulation.

Proceedings of CHI ‘91, April 28-May 2, 1991.

Horvitz, Eric and Jed Lengyel. Perception, Attention,

and Resources: A Decision-Theoretic Approach to

Graphics Rendering. Proceedings of Uncertainty in

Artificial Intelligence, Aug. l-3, 1997, pp. 238-249.

Hudson, Scott E. and Ian Smith. Electronic Mail

Previews Using Non-Speech Audio. Proceedings of

CHI ‘96, April 1996, pp. 237-238.

Kaelbling, L.P. and Littman, M.L. Reinforcement

Learning: A Survey. Journal of Artijicial Intelligence

Research, vol. 4, 1996, pp. 237-285.

Marx, Matthew and Chris Schmandt. CLUES: Dynamic

Personalized Message Filtering. Proceedings of CSCW

‘96, pp. 113-121, November 1996.

lO.Mynatt, E.D., Back, M., Want, R. Baer, M., and Ellis

J.B. Designing Audio Aura. Proceedings of CHI ‘98,

April 1998.

1l.Rudnicky, Alexander, Reed, S. and Thayer, E.

SpeechWear: A mobile speech system. Proceedings of

ICSLP ‘96,1996.

12.Sawhney, Nitin and Chris Schmandt. Speaking and

Listening on the Run: Design for Wearable Audio

Computing. Proceedings of the International

Symposium on Wearable Computing, October 1998.

13. Stainer, Thad, Mann, S., Rhodes, B., Levine, J., Healey,

J., Kirsch, D., Picard, R., and Pentland, A. Augmented

Reality through Wearable Computing. Presence, Vol. 6,

No. 4, August 1997, pp. 386-398.