Seeing Beyond Sound: Visualization and Abstraction in Audio Data Representation

Final Draft 1

Ashlae Blum, PhD Student, Math Department VUW
a.h.blum@gmail.com

Keywords: Audio information, data visualization, audio signal processing, design philosophy, human-computer interaction, bioacoustics, interface design, software design.

0. Abstract

See – Smell – Taste – Touch – Hear. The human capacity to interpret complex data is epistemically shaped by the perceptual frameworks of representation through which information is presented. In audio research, fields such as bioacoustics, music information retrieval, auditory science, and cognitive psychology employ a wide array of tools that transform theoretical knowledge into applied science. These tools predominantly arose from technologies originally developed by the experimental media, entertainment, communications, and defense industries, and carry a variety of inherited domain-specific technical conventions. As a result, knowledge gaps have formed in the adoption to their new scientific contexts. We argue that designing tools specifically grounded in modern audio research can allow a more natural and intuitive way of engaging with audio data analysis and visualization, inspiring new insights to emerge. Historically, shifts in representational paradigms – such as the transition from waveform editing to spectral manipulation, or the adoption of 3D sound field reconstructions in spatial audio – have required users to overcome initial cognitive dissonance before achieving fluency. As they are incorporated into standard workflows, such advancements frequently lead to improved analytical efficiency and creative flexibility. This paper explores the potentials associated with adding dimensionality back into visualizations to encode complex audio information in spatially and structurally richer formats.

I. Introduction

Advanced data visualization techniques enable scientists to interpret complex datasets by transforming high-dimensional data and metadata into abstract visual elements. This serves not only to reveal patterns in information, but also to build narratives that enhance our collective understanding of the world around us. Such representations are often mediated by software and tools designed for specific domains, embedding assumptions that, while optimized in one context, may inhibit another. For audio data, waveforms and spectrograms form the basis of our visual knowledge. These rely on two-dimensional visualizations of the time-frequency domain that are mathematically well-defined, but often lack intuitive correspondence with the multisensory nature of auditory perception. The advent of the digital audio workstation (DAW) provided users a familiar template for audio interaction. With origins in the software revolution of the 1970s, its design elements persist in today's interfaces that span from the film industry to scientific research. More recently, the rise of programming literacy and the expansion of audio research have evolved alongside the need and interest in low-level control. Libraries such as Librosa (Python), Web Audio API (JavaScript), and tuneR (R) have arrived on the scene, with enthusiastic online userbases that connect communities across the internet, and the world. The broadening scope of creative coding has bridged science and art to expand the worlds of the technical and the expressive into expansive layers of abstraction; apps and games built to facilitate music-making and sound exploration proliferate; sound art and sound design are now well-established as legitimate commercial fields. In short, the spectrum of use cases in which audio is being transformed from numbers into something else is ever-expanding, and so, too, must the ways in which we interact with it.

II. A Brief History of / The Historical Landscape of / Audio Visualization Software

Modern audio analysis software is an amalgamation of design principles, applied scientific theory, and physical constraints that has been continuously refined over the last century or so. Early hardware inventions that modeled sound signals were built using analog electronics to implement theoretical concepts from harmonic and

spectral analysis. Ranging from exploratory to practical, these devices were physical embodiments of the understanding of sound as a medium of the times. Due to their inherent physicality, they also carried with them necessary limitations and operational conventions that have persisted in the shift from analog to digital audio analysis. In today's software, such assumptions are now often overlooked, as analog origins have largely been superseded by their digital descendants. DAW-like analysis software, such as Audacity, Raven, and Sonic Visualiser, are but some examples at the heart of audio workflows that propel scientific inquiry. However, the presets embedded in such tools often assume specific use cases. Without knowledge of their existence, it can be easy to generate results using numerical parameters intended for another domain. To better assess the contemporary landscape, we first review the historical origins of modern audio visualization tools.

1. The Steampunk Origins of Sound Science

The steampunk origins of sound science grew out of the electromechanical age, in which the use of electricity to process and transform information revolutionized all aspects of society. While Fourier's seminal works on harmonic analysis in 1807 and 1822 laid the mathematical foundations for audio signal processing, practical applications of these theories took time to crystallize. Devices such as the telegraph (1800), telephone (1876), sound spectrograph, wave analyzer, and graphic equalizer were built from analog components, their design and use constrained by both material and human limitations. Friction and inertia of mechanical components, short-circuits, overheating, fixed-bandwidth filter banks, and slow-burning electrochemical paper are but a few. These limitations were far from hidden; they were explicit, tactile, and fundamentally affected the user's interaction with and interpretation of sound.

2. Theoretical Foundations: Let's Get Digital

The development of the Fast-Fourier Transform (FFT) in 1965 formed the backbone of signal processing algorithms as digital computing became ubiquitous through the rest of the century, and beyond. FFT-based methods proliferated through a wide variety of industries, for example, telecommunications, medicine, and music, their implementations often remaining domain-specific. Now essential, Digital Signal Processing (DSP) algorithms form the building blocks of audio analysis software, and are intricately linked to our fundamental understanding of sound. Yet they, too, are built upon necessary parametric limits and assumptions inherited from their origins. For example, there is always a tradeoff of knowable information about a signal, as described by the Gabor uncertainty principle. This places a lower limit on the amount time-frequency uncertainty, and affects nfft and hop length parameter choices. Music production, speech analysis, and sonar engineering serve as specific examples where innovative uses of DSP algorithms left a lasting impact. Many of these pivotal technologies were gradually incorporated into the greater lexicon of digital audio analysis software, where they now live side-by-side as part of an unassuming digital toolkit.

3. How We Interact With Sound: Interface Design, Use Cases, and The Rise of the DAW

Like many of its digital audio counterparts, the modern DAW was also originally a piece of hardware. Arguably, the first DAW was the Soundstream Digital Editing System (1977), which operated on a minicomputer that ran custom software called the Digital Audio Processor (DAP). It was used to edit master tapes in the commercial audio industry, and featured hard disk recording, an interactive screen for waveform editing, and both analog and digital interfaces. The Fairlight CMI (1979) was another groundbreaking technology that became famous for its "Page R" sequencing environment, displaying rows of blocks that represented notes and audio – a precursor to today's MIDI sequencing capabilities. Text-based DAWs of the 1980s, such as the Commodore 64 and Keyboard Computer System (KCS), supported multiple MIDI tracks using lists and drop-down menus. The visual layout of the Steinberg Pro-16 (1987), developed for the Atari, was the predecessor to today's DAW interface. It was a MIDI sequencer that visually mirrored physical hardware mixing consoles, complete with playback and routing controls, and horizontal arrangement views to allow for multiple perspectives within the audio environment. This took the abstract concept of sequencing, previously done using manual list entry, and made it look and feel like working with physical audio hardware. Since computer processors at the time could not yet power multi-track recording or playback, these early workstations were MIDI-only. As computers became more powerful

throughout the 1990s, computationally-expensive audio functions such hard-disk audio processing could live side-by-side with sequencing. Prominent examples include Sound Tools (1992), with its limited audio recording; Cubase (1992), with its MIDI and audio visible in the same interface; and the invention of the Virtual Studio Technology (VST, 1996) plugin, which allowed digital effects to be applied to individual channels.

4. How We Perceive Sound: Sensory, Perceptual, and Cognitive Considerations

How we view sound is profoundly affected by not only the physical experience of perceiving an image on a screen, but by a broader sense of cognition and perception about its fundamental nature. For most humans, sound is one of five core senses we experience throughout our lives. Our relationship with it changes as we age, and as we add information to our sensory network through lived experiences. A number of tools are used to visualize sound, some of which strive to depict spatialized relationships between its components, and others which employ layers of abstraction to expand its sphere of perceptible information. Oscilloscopes plot time-amplitude waveforms by reading the voltage from a transducer (microphone) to display pressure oscillations. A spectrogram uses the Short-Time Fourier Transform (STFT) to sum windowed segments of a signal, trading temporal precision for frequency resolution: lower time-resolution allows the calculation of finely-grained frequency evolution, and vice-versa. Mel-Frequency Cepstral Coefficients (MFCCs) represent spectral energy as a series of coefficients scaled exponentially to align with the human auditory system. These types of audio tools are optimized for quantitative feature extraction and machine readability, however, they can obscure higher-order perceptual structures such as timbral nuance, microtonality, or the fullness of polyphonic sound .

One major challenge in data visualization is mapping high-dimensional features to visual variables in a way that intuitively makes sense when you look at it. Many tools from statistics, such as scatter plots and time-series graphs, are precise and well-established, yet they require an input of low-dimensional data. Audio features, which are highly multidimensional (e.g. dozens of MFCCs, spectral and temporal centroids, entropy scores), require correspondingly advanced encodings. There are innovative efforts across many domains that strive to expand and explore the nature of data visualization, and the unification of multidimensional and interactive visualizations with cognition. The adoption of topological data analysis can reveal the underlying shape of a dataset of whale songs. Activation maps in convolutional neural networks trained on audio show the features a model learns to detect. Through exposure, use, and familiarization, these visual innovations have become part of standard audio data visualization workflows.

As experimental graphics research continues to push the boundaries of technology, media domains such as virtual reality (VR), augmented reality (AR), and 360 video offer expanded formats for multisensory immersion. Such tools and techniques prioritize enhanced perceptual experiences and intuitive interactive elements. 3D time-frequency embeddings attempt to visualize timbral similarity by projecting features into a spatial manifold, allowing researchers to see clusters of similar bird calls or phonetic units. Similarly, sonic labyrinths use interactive 3D structures to represent sound, where navigation corresponds to spectral exploration. From science to media, innovations in audio data visualization proliferate as technology facilitates the accessible transformation of multisensory information.

III. Addressing Specific Knowledge Gaps

Using 1) the setup from the historical origins, 2) the transition from physical and analog to perceptual and digital, and 3) an overview of physical limitations and inherited conventions, we now illustrate some very specific examples found in software that embody these limitations.

a. Hidden assumptions: software as a black-box

The metaphor of the black-box comes from the aviation industry, where flight data recorders in airplanes were housed in a literal metal box that had been painted black. These were comprised of analog and electromechanical components that engraved flight metrics onto metal foil []. The black-box metaphor has since become an

analogy for the study of a closed system without prior knowledge of its inner workings, relying solely on knowledge of input, and observation of output, to evaluate its structure and evolution [].

With software comprising anywhere from hundreds to ten-thousands of lines of code and more, it becomes necessary to treat it as a black-box, or we would never get anything done. Since code is more often read than it is written, especially for free, libre, and open-source software (FLOSS), it is seen as a best practice to leave a clear, well-documented paper trail in the form of in-line notes, for posterity. Along with a (hopefully) clear set of instructions on how to use the software, these notes, known colloquially as documentation, are essential so that others who use it thereafter can follow the design and flow of logic, and possibly to understand features that may be only partially implemented, or future scaling intentions. This facilitates not only a deeper understanding of such tools, but also the ability to change, edit, or repurpose the software for either similar or far-flung or imaginitive notions (use cases). Also, in an area of development where people are often working independently, documentation serves as a form of communication and connectedness between developers who may never meet each other in real life, adding an additional layer of cognition aside from just a functional or utilitarian need.

b) Parameters, presets, and

Design transparency openly acknowledges such choices, providing access to customization that may liberate the user from the constraints of domain-specific applications. Knowledge of equations from signal processing, population dynamics, or neuroscience can allow for backtracking through the lines of code. These equations are often direct, if dense, translations into formal logic through layers of abstraction that take the form of standard software libraries. As with all equations that govern the empirical sciences, numerical parameters must be chosen to allow mathematical computation to occur. This is the starting point, from which it is assumed that values will be changed to suit the particular needs of a specific application at-hand. However, as meta-uses compound, the reliance on presets or parameters can become buried, obscured, or forgotten. Therein runs a risk of making assumptions that may not be appropriate for a specific domain’s application. In the following section, we focus primarily on a comparison of FLOSS tools and their hardcoded assumptions that have been noticed firsthand while reading through source code. See Appendix for a more complete list of audio-specific software and libraries that incorporate presets.

- Praat was developed specifically to study the human voice, and has pre-emphasis filtering that boosts frequencies above 50 Hz. This alters the relationship between frequency content in the signal, and can be problematic for many animals that communicate using low-frequency information. Specific examples include elephants, rhinos, whales, large ungulates, big cats, bears, wolves, seals, sea lions, manatees, and some fish. []

- Praat’s preset limits the visual display of audio clips greater than a certain duration of time.

- More fully-featured software, such as Audacity, Sonic Visualiser, Avisoft, and Raven, represent a spectrum of graphical DAW-like tools that have developed specialized use cases in audio information domains. Their workflows are rooted in temporal manipulation, which is often (but not always) a stepping-stone in audio information science. For example, the purpose of cutting audio at annotation points is to then perform other calculations on that audio slice, i.e. feature extraction.

- Horizontal vs. vertical layouts are tied to workflows from the audio recording industry. For scientific use cases, comparing many small files along horizontal timelines feels clunky when looking to broadly assess their similarities and differences. This is different from when we want to view the audio as a time sequence, where (horizontal) temporal continuity may be useful.

- Interacting with all files (or annotated slices) at once can be labor-intensive, often requiring manual interaction with each one. There is not always a way to batch import many files vertically along independent channels. Files may be required to be loaded individually, or the batching of such files might be for a calculation or analysis that is hidden in the software’s algorithms.

- If batch loading and viewing is indeed possible, interacting with all files simultaneously can require the manual labor of clicking each single track to turn such a feature on. Repetitive clicking with a mouse or trackpad is not physically ergonomic and can cause physical harm over time if done too frequently.

- Effects batching further exemplifies the problems associated with individual selection. If, for example, a bandpass filter is required to eliminate some machine noise or a natural event such as an earthquake, it is far more efficient to apply this same effect to all files at the same time. Instead, at times one must add a VST device onto every audio channel – a task that, when required for thousands of files, quickly becomes tiresome.

- In scikit-maad, a 4th-order Butterworth (infinite-impulse response) filter is the preset for automated feature and region of interest (roi) selection. It preferences frequency precision with a flat passband and -24dB/octave rolloff, but limits temporal precision due to its phase-nonlinearity. Since different frequency components of a signal travel at different rates, this type of filtering shifts the timing of low- and high-frequency information differently within the same acoustic event. The filter response can also create acausal pre-event artifacts that interfere with the detection of onset transients. To mitigate these effects, maad defaults to the zero-phase filtfilt, but these choices may be an oversight in cases that require high temporal precision. Examples include measuring intervals between syllables (such as echolocation clicks), sample-accurate onset detection, or fine-scale waveform comparison. Using scipy.signal can allow for more finely-grained control.

- Librosa's native sample rate is set to 22.05 kHz, and its STFT parameter defaults are set to a nfft value of 2048 and hop length of 512. Unless you know about this, you may be performing calculations with incorrect assumptions.

- Audacity's power spectrum requires users to use only certain nfft parameters that are dependent upon the signal length; as such, uniform nfft values can't be chosen for all files in a batch if they are of non-uniform lengths. Also, spectral analysis can only be performed by clicking through a series of sub-menus, and can only be done on one sound clip at a time. The low-level libraries that supposedly allow for batch processing of files to do this task don't actually work.

- Audacity's Fourier transform (pffft) relies on a translation of Fortran 77 code from FFTPACK that was written in 1985 []. These algorithms are very powerful, but may be difficult to implement with other modern software, and may not behave as expected, since they were designed to operate on hardware that had different limitations.

- The number of different FFT algorithms that have been written and re-written for specific uses is at this point an unofficial meme in signal processing. This is evident across many different packages with amusing names such as "Pretty Fast Fast Fourier Transform" (pffft), "Keep It Simple Stupid Fast Fourier Transform" (kissfft), "Fastest Fourier Transform in the West" (fftw), and others. This can be overwhelming to keep up with when choosing algorithms.

In short, when it comes to numerical analysis, there will always be hidden assumptions that form a collection of presets, whether for parameter values, user interaction, or conceptual approaches to sound. Tool choice is often one of determining the baked-in assumptions that align most closely with the task at hand. This is neither inherently good nor bad, but a phenomenon of engaging in real-world problem-solving.

IV. Proposed Technical Solutions – Conceptual Reimaginings

In the previous section, we outlined a technical wish-list based upon issues we have encountered in our use of audio analysis software. Informed in tandem with historical perspectives and conceptual extensions, we present a variety of solutions to the problem of Schrodinger's Audio Data Visualization Conundrum that go beyond the aforementioned specific issues into an evaluation of the landscape of contemporary cognition.

<< ***We propose that giving users access to independence and agency facilitates an increased ability to form complex cognitive associations.*** >> (In a sense, this concept goes slightly beyond software into the domain of

pedagogy, however, we strive to refine our focus toward the field of audio information visualization.) In the argument for this shift, we identify three fundamental principles/elements that are necessary to enable this:

Transparency – a clear-box approach, rather than a black-box approach, can empower the user to make their own appropriate choices for their intended use. This can involve presenting available options as visual cues at the point of interaction, rather than making decisions for the user or simply leaving all instructions in the documentation. It could also involve informing the user as to why certain design choices were made, and provide options for real-time reconfiguration.

Flexibility – the ability to configure an environment that best aligns with an individual's task requirements or work style can give a sense of agency over workflows. Sometimes, it is especially useful to have multiple perspectives when trying to understand a complex situation. The difficulty of working with time-series data is no exception; the ability to switch seamlessly between analogous options, and even to compare them side-by-side, can be very informative. Adaptable design principles make tools easier to use across a wide variety of scenarios, and may encourage users to stick with one familiar tool, rather than switching frequently between diverse workflows.

Robustness – tool-based environments should be able to be handle a wide variety of contexts, and should be as agnostic as possible to the types of data that are input. Like a hammer, which has a variety of uses, or a multi-tool, which has even more, these persistent objects are two divergent examples design principles. One is strikingly simple, and as such, is almost neutral to its application. The other is quite complex – a fusion of many different small devices that are convenient to have in the same place, ingeniously attached through a shared 'handle', and made compact and portable, ready for action. While clearly intended for specific tasks, each of these tools enjoy a wide variety of uses and applications in everyday human life.

WIP!!!!! Cognition and Interactive Design Theory:

The benefits of incorporating modern design principles into specialized audio visualization workflows have far-reaching implications outside of simply being less annoyed while performing daily tasks. Studies across psychology, design theory, and cognitive science show that increased perceptual connection can enhance pattern recognition, qualitative assessment, and interactive engagement with audio data.

Sweller's principle of split-attention effects shows that if learners have to integrate information from multiple, spatially-separated sources that are individually unintelligible, it imposes a cognitive load that inhibits learning. For audio visualization analysis, we can draw an analogy to displays of frequency, temporal, and amplitude content across separate interface regions or screens; this limits a user's mental availability to make intuitive inferences, since time must be spent searching for and mapping the elements back to each other. Similarly, cognitive load is also increased when information is presented sequentially, rather than simultaneously, since learners must hold information in their working memory until the next piece arrives. This calls immediately to mind the horizontal time and frequency displays of a waveform, power spectrum, or spectrogram. Element interactivity theory indicates that information complexity is modulated not just by the number of elements, but also by their interactions, and that these must be processed simultaneously in the working memory. This concept is immediately extensible to the infinite number of audio features that are inherently perceived in the same instant as a sound, in real time. Bertin's visual variables framework describe position and size as the principal factors that express quantitative differences, with color an indicator of categorical difference, and a variety of other values indicating the visual hierarchy of information. He suggests that visual efficiency relies on the combination of multiple variables, which enhances both perception as well as pattern recognition. This is consistent with the dimensionality reduction often seen in modern data visualization strategies. Farnell's procedural audio framework emphasizes the importance of visualizations that mirror the dynamic nature of generative sound processes. He suggests that bidirectional and interactive connections between sound and visual interfaces enhance creative exploration and technical comprehension. []

Using the lens of cognitive and interactive design theory, we show that associations between visual elements and the human psyche are intrinsically linked through the perceptual continuum that is bodied sensory experience. Cognition and psychology fundamentally demand interaction to give context to complex information. We can therefore project that for audio information visualization [design], users may benefit from access to tools and workflows that allow for a perceptually diverse engagement with sound. Their integration into audio analysis workflows can expand the boundaries of both technical and creative sound exploration. Though, the introduction of new visual tools or representations will necessarily demand a separate set of post-implementation considerations before they would even be able to become part of a regular workflow or the existing lexicon of commonly used tools. To be adopted, novel visualizations may require a shift in representational paradigms. When presented with something new, not everyone is ready to accept the change. Users must first overcome cognitive dissonance and resistance to change, followed by the learning curve associated with performing any new task. As familiarity and then mastery is attained, these tools can become completely streamlined into existing workflows, and we may even struggle to remember what life was like before we had access to them. Such is the curse of convenience. However, with literacy comes the benefits of speed, efficiency, and creative flexibility.

V. Impact, intended audience, who this benefits

There are endless ways to explore the theoretical effects of applied design philosophy, but what about their impact? When a new tool or technique is deployed, who will actually use it? Who will it benefit? Where and how will it be used? Especially now, in the age of Big Data, there is an accelerated need to include non-domain experts and citizen science participants in the validation and annotation of data. Tools designed specifically with interaction and visualization in mind can make it more accessible for them to interact with the data in ways that are relatable, intuitive, and familiar. The tactile experiences of everyday digital tools, such as apps and games, can be modeled and expanded upon to create user experiences that feel familiar while not being too distracting. Such tools can also give people a sense of agency over what they're doing – they may reveal the 'secret elements' that are often reserved for specialists, increasing transparency, building institutional trust, and generating a sense of community investment. Furthermore, tools that are fun and interesting to use generate conversations outside of their initial use/community. When everyday people get excited enough about wild bird audio annotation apps to discuss them at coffee shops or networking events, for example, this can be viewed as a sign of success that such a tool is connected with social values. Thus, there are diverse practical reasons in favor of increasing the accessibility of audio analysis and exploration to both technical and non-technical audiences. The following are examples of benefits to specific groups:

- People who already use data viz tools regularly for their jobs, such as scientists, data scientists and analysts will certainly benefit from increased efficiency and intuition, allowing them to see audio information in new ways.

- Citizen scientists who participate in valuable tasks such as data annotation and validation, species identification, symptom reporting, noise pollution assessment, can have a way to annotate in real-time that may allow them to feel included as an essential part of a team, gives them more knowledge about the science and behind the scenes, which could encourage them to become more excited involved from a scientific standpoint. This is beneficial because science education is essential as people need to work together to address many urgent problems in fields such as conservation, medicine, and society.

- Accessibility by including things that are interesting or fun to look at, listen to, and interact with, especially for non-experts, can provide entertainment as well as social values. The possibility of gamification can also increase audience reach, and can be used to collect feedback about what does and doesn't work, as well as who tends to use the tools and how, which are valuable insights for any tool designer.

- We can imagine a use where, for a large dataset that needs annotation, the dataset can be broken up into smaller pieces and distributed among a group of people to lessen the workload. Then, it is essential that all users can be sure they are referring to the same phenomena, the same start and stop time, the same features, across the same interface.

- AI users in particular, who may not be used to working with real data, or who may work with many different types of data, need assistance in understanding the nuances of datasets when they are not familiar with the domain. In the rising proliferation of AI outside of research domains, the number of people working with audio data will increase dramatically, as will the use of AI as an everyday tool in its own right. Such human individuals can make incorrect assumptions about properties or characteristics of sound if they are not informed in a way that is fast, efficient, and intuitive. This also factors into the field of ethics, since the dangers of making assumptions can proliferate quickly in cases where a small effect may spiral out of control over a massive dataset like those seen in Big Data.

Audio visualization tools can act as intermediary steps between the many people involved along the way in the process of scientific or artistic inquiry. It places control in the hands of the user, and reconfigures the hierarchy that limits niche knowledge to be held solely by domain experts. Increased agency can build a sense of community, and strengthens the ties that people feel to their work or special interest. Many people will continue to be affected by today's rapid advancements in audio data visualization as the Age of Information spirals outwards. We hope that with this expanded consideration of the implications and impacts of new tools on their audiences, the case for incorporating a broader set of user-centric design principles may be compelling.

VII. Conclusions.

Audio visualization has long been more than a technical challenge; it is a framework for thought and perception. From the sound-on-film and color organ of the 1920s through today's latent spaces of artificial intelligence, the domain of complex audio representation has evolved alongside emerging technologies that expand its boundaries and influence.
Sound itself is a physical, multidimensional object that is intrinsically linked to perception.
Whether human on non-, the bodily systems that allow living beings to sense the presence of sound

dimensionality and feature mapping
blahblahblah

As it evolves with time and technology, each iteration expands upon the last one

As such, it is our hope that as the field of audio data analysis evolves with technology, that we, too, can learn to hear beyond our eyes. Seeing is believing but believing is not always seeing?

It is our hope that as we advance into the connected worlds of tomorrow, we may all learn to see beyond sound.