Seeing Beyond Sound: Visualization and Abstraction in Audio Data Representation

Ashlae Blum’e
ashlae.blum@vuw.ac.nz

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

Inscription on the Wall of West Forest Temple

Viewed from the front, a full mountain range; from the side, a single peak.
Far, near, high, low – each view is different.
I cannot recognize the true face of Mount Lu,
Simply because I myself am on the mountain.
– Shi Su

题西林壁

横看成岭侧成峰，
远近高低各不同。
不识庐山真面目，
只缘身在此山中。
– 苏軾

0. Abstract

The interpretation of complex data is epistemically linked to human perceptual frameworks. In audio information research, sound is represented and transformed using visual elements that highlight abstract patterns detached from the physical experience of sonic perception. Fields such as bioacoustics, music information retrieval, and auditory science employ a wide array of tools to transform theoretical knowledge into applied science. However, these tools carry with them hidden assumptions that are masked through their adoption to new scientific contexts. Many such tools have inherited domain-specific technical conventions that nod to their historical origins in areas such as experimental media, entertainment, communications, and defense research. Cognitive psychology suggests that the way information is presented strongly influences an individual’s ability to form complex associations. We argue that re/designing tools to align with emergent needs of modern users will improve both analytical as well as creative outputs due to an increased affinity for using them. This paper explores the potentials associated with adding dimensionality back into visualizations to facilitate positive social effects in the use of audio information visualization tools.

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 Audio Visualization

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 (1830-1940), in which the use of electricity to process and transform information revolutionized all aspects of society. Fourier's seminal works on harmonic analysis (1807, 1822) [1,2] had laid the mathematical foundations for audio signal processing, yet practical applications of these theories took time to crystallize. The earliest mechanical devices to record and play sound were the phonautograph in 1857 [3] and the phonograph (1877) [4], both of which were powered by hand. The phonautograph recorded sound waves by etching them on glass or paper [3]. The phonograph etched its sounds on tin foil, and could additionally play back audio from the etching [4]. The invention of the telegraph (1837) [5,6] marked a transition as the first electric device to transmit sound, followed by the telephone (1876) [7], gramophone (1887) [8], loudspeaker (1925) [9], and sound spectrograph (1946) [10,11]. These devices were built from mechanical and analog electric components, their design and use constrained by both material and human limitations. Friction and inertia of mechanical parts, short-circuits, and overheating are but some of the factors that impacted their smooth operation. 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 [12] formed the backbone of signal processing algorithms as digital computing became ubiquitous through the rest of the century, and beyond. FFT-based methods impacted a wide variety of industries, for example, telecommunications (DSL modem [13,14], cell phones [15,16]), medicine (MRI [17,18], EEG [19,20]), and music (reverb [21,22], phase vocoder [23,24]), 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) [25,26,27]. It was designed to edit master tapes, and featured hard disk recording, an interactive screen for waveform editing, and both analog and digital interfaces [26,27]. The Fairlight CMI (1979) was another groundbreaking technology, the first polyphonic synthesizer 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 [28,29]. Text-based DAWs, such as the Commodore 64 (1982) [30,31] and Keyboard Computer System (KCS) (1984) [32,33], supported multiple MIDI tracks using lists and drop-down menus. The Steinberg Pro-16 (1986) was a software interface developed for the Atari whose visual layout was the predecessor to today's DAW interface [35]. The precursor to Cubase, it was a MIDI sequencer that visually resembled physical hardware mixing consoles, complete with playback and routing controls, and horizontal arrangement views [36]. This took the concept of sequencing, which usually used manual list entry, and transformed it to look and feel like working with physical audio hardware. Since computer processors at the time could not yet support multi-track recording or playback, these early workstations were MIDI-only. Computers became more powerful throughout the 1990s as the semiconductor industry enabled processor technology to become cheaper, faster, and smaller. Consequently, computationally-expensive audio functions such hard-disk audio processing could live side-by-side with sequencing. Prominent examples include Sound Tools (1989), with its limited audio recording [37,38]; Cubase (1992), with its MIDI and audio visible in the same interface [39]; and the invention of the Virtual Studio Technology (VST) plugin (1996), which allowed digital effects to be applied to individual channels [40].

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 [41]. 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 [42]. Mel-Frequency Cepstral Coefficients (MFCCs) represent spectral energy as a series of coefficients scaled exponentially to align with the human auditory system [43]. These types of audio tools are optimized for quantitative feature extraction, however, they can obscure more nuanced structures such as the timbre of a unique voice, the microtonality of an oud, or the rich polyphony heard while standing in the middle of a crowded train station.

5. Dimensional Representation and Experimental Media

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 to unify multidimensional and interactive visualizations with cognition. For example, topological data analysis (TDA) can reveal the underlying shape of a dataset [44,45], and has been used in describing the periodicity of flutes [46], music tagging and classification [47], and audio fingerprinting of MIDI music [48]. These shapes can then be fed into a convolutional neural network (CNN) as training data to teach it to detect patterns in audio features, which are output as activation maps [49]. Through exposure, use, and familiarization, abstract 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), mixed reality, and 360 video offer expanded formats for multisensory immersion. These technologies, often referred to as experiences, prioritize interactivity and can be found in spaces from VR gaming centers to live theatre and performance art. Societal applications include the use of haptics to enhance sensory awareness for blind or deaf people [50,52], VR for therapy and training [53], and 3D sculpture as a tool for design [54]. One important audio research application uses 3D time-frequency embeddings to visualize timbral similarity by projecting features into a spatial manifold, visualizing clusters of similar bird calls or phonetic units [54,55]. Similarly, sonic labyrinths use interactive 3D structures to represent sound, where navigation corresponds to spectral exploration [56]. Across science and media, innovations in audio data visualization proliferate as technology facilitates the accessible transformation of multisensory information.

III. Addressing Specific Knowledge Gaps

Informed by the transition from physical and analog to perceptual and digital, we discuss some specific concepts and software that examples illustrate the aforementioned limitations.

a. Hidden assumptions: software as a black-box

The metaphor of the black-box comes from a fusion of aviation industry and WWII-era slang, when flight data recorders, along with other secret electrronic devices, were housed in a nonreflective black metal boxes [57]. While the first version used a thin beam of light to record flight metrics such as altitude and speed onto photographic paper, later versions engraved metrics onto metal foil [57]. 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 [58].

Comprising anywhere from hundreds to ten-thousands of lines of code and more, it becomes necessary to treat software as a black-box, or we would never get anything done. Since code is more often read than it is written [59], 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 and imaginitive notions (use cases?). Also, in an area of intensive 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 preconceived notions.

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 lines of dense programming languages. Portability and translateability are also facilitated by transparency, since at times one can simply replace one equation with another or add it to a centralized dictionary of options. The forms that such equations often take (in the code) are direct, if dense, translations into formal logic through layers of abstraction known as standard software libraries (e.g. numpy, librosa, fftw). 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, this implied 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 the study of animals that communicate using low-frequency information, such as whales, elephants, tigers, and rhinos [60-63].

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

- More fully-featured software, such as Audacity, Sonic Visualiser, Avisoft (proprietary), and Raven (proprietary), 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.

- A need for effects batching further exemplifies the manual-selection issue. 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, one might have to select a checkbox or button, or add a VST device onto every single 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. This filtering optimizes 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 shifts the timing of low- and high-frequency information differently within the same acoustic event. The infinite filter response can also create acausal pre-event artifacts that interfere with the detection of onset transients. To mitigate this, maad defaults to the zero-phase filtfilt, but this choice is inappropriate when high temporal precision is needed. Examples include measuring intervals between syllables (such as echolocation clicks), sample-level accuracy for onset detection, or fine-scale waveform comparison. Using scipy.signal can allow for better 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 calculation limits nfft value choices based on signal length; as such, the same nfft value 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 as described in the online documentation.

- 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 integrate with 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 numerical computation, there will always be hidden assumptions that form a collection of presets, whether for parameter values, expected modes of user interaction, or conceptual approaches to sound. Tool choice is often made based on the baked-in assumptions that align most closely with a task at hand. This is neither inherently good nor bad, but a phenomenon of engaging in real-world problem-solving.

IV. Proposed Theoretical Solutions – Conceptual Reimaginings

A. Design Principles

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 we jokingly refer to as "Schrodinger's Audio Data Visualization Conundrum" due to the tradeoffs inherent in information knowability. These proposed solutions go beyond the issues mentioned in the previous section 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 moves slightly outside of 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 proposed solution, we identify three fundamental principles at the core of our design philosophy.

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 or embedded upon each other, 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 divergent 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. This could mean that a tool is designed to process input data in many different ways, like a hammer, or to receive and combine many types of data in a synthetic configuration, like a multi-tool. Consider software that is designed to receive uniform lengths of audio from the same source. A next step could then be to map extracted features and combine them with environmental variables, such as weather and temperature, or with metrics taken across the set of input data, such as mean amplitude or spectral centroid. The raw data itself already has a certain uniformity, so the parameter space of this tool would then be highly synthetic, since it would be constructed out of higher-order relationships between abstract variables. Alternatively, if a tool's input sounds have high heterogeneity, like clips of drastically different lengths or sounds from different species, efforts to generate a base parameter space might first focus on defining broader sets of classical metrics, such as duration, amplitude, entropy, or various other statistics prior to abstract transformation. The key difference between these two scenarios lies in the input data. Each would require a different number and types of steps to transform data to the same point of abstraction. However, robust tools are configurable for either case.

B. Cognitive Load Theory and Visual Design

The theoretical benefits of incorporating an updated set of modern design principles into audio visualization workflows have far-reaching implications outside of simply being less annoyed while performing daily tasks. Studies across cognitive psychology and design theory show that increased perceptual connections can enhance pattern recognition [65-67]. The following examples demonstrate how spatial and temporal representations of information impact mental processes such as comprehension, memory, and learning.

1. Split-attention effects show that having to combine information from multiple, individual, spatially-separated sources inhibits learning [65]. These effects are also found in scenarios where information is

presented simultaneously, but in different formats [65]. This implies, conversely, that if information is visibly close together, and/or presented simultaneously but in the same format, learning will be easier. In audio software, we can draw an analogy to split-screen views that show waveforms, spectrograms, and power spectral density on separate screens. Users are required to constantly switch back and forth between views, trying to remember what they previously saw on the last screen as they translate information from one format to another. (This is an actual, real problem in Audacity; see section IV-b.) Such display issues limit a user's mental availability to make intuitive inferences, since one must search for and map visual elements back to each other while holding prior information in working memory. The demands on cognitive load also increase when information is presented sequentially [66,67], rather than in staggered or simultaneous formats. Furthermore, information complexity is modulated not just by the total number of elements, but also by their interactions [66]. Simultaneous information streams require greater load on working memory [66,67]; therefore, the more interconnected a group of elements is, the more complex the information they represent. From this, we can conclude that sequential formats are not ideal for processing complex interconnected information. Outside of cognitive psychology, inefficiency in linear and sequential information processing has been shown in the communications [68], computing [69], and energy [70] industries. Since humans are the architects of these systems, the phenomenon that preferences a non-simultaneity of information processing could even be a function of human cognition, but that is outside of the scope of this paper to explore.

2. The effects of visual elements on perception have been explored systematically through a variety of principles that govern design theory. The visual variables framework describe position and size as the principal factors that express quantitative differences [71]. Color, as a variable, is broken into the values of hue, which describes the qualitative difference of category, and value, which describes the quantitative difference of order [71]. Together with shape, orientation, and texture, these visual variables describe a hierarchy of information with levels that are either associative or dissociative [71]. This means that visual characteristics can be used to deconstruct the emergent patterns that inform meaningful group characteristics. That is to say, when objects are perceived as being part of a group, visual variables provide a basis for distinction. To extend these thoughts to audio software and visualization, we can thereby conclude that the ability to identify patterns in abstract representations, such as those used for audio visualization, can be facilitated by making visual design choices that correctly map visual elements to meaningful features. This is consistent with existing approaches for dimensionality reduction in modern data visualization.

V. Discussion

A. Practical considerations

Through the lenses of cognitive and visual 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. The inner workings of human cognition and psychology fundamentally demand an interactive format 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. This could include nonlinear workflows, reorienting information along different axes, using new metrics to scale information, or interchanging relationships between variables. The incorporation of contemporary design principles into audio analysis tools and workflows can expand the boundaries of both technical analysis and creative sound exploration. Practically, it takes time to implement new tools. Novel visualizations may require a shift in representational paradigms: new information is not always readily accepted. To be fully adopted, users must first overcome cognitive dissonance and resistance to change [72,73], followed by the learning curve that is associated with performing any new task. As familiarity and then mastery is attained, these tools can become streamlined into existing workflows. We may even struggle to remember what life was like before we had access to them; such is the curse of convenience. However, increased technical literacy begets the benefits of speed, efficiency, and creative flexibility.

B. Future impact, intended audience: who 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 nondomain 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 people to interact with 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 some examples of benefits to specific groups:

- People who already use data visualization 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. Specialized task automation, efficient 3D or time-evolution displays, and the ability to visually overlay features of interest in new ways are some hypothetical workflows that could be beneficial.

- 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 easily 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 and involved from a scientific standpoint. This is triply 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 for all users to be sure they are referring to the same phenomena, and 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 subject matter. In the rising proliferation of AI outside of experimental and 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 (and, more dangerously, their AI counterparts) 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, or may propagate into models through training, or affect other datasets through extracted metadata.

- Audio visualization tools can act as intermediary steps between the many people involved along the way in the process of scientific and 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.

VI. Conclusions

Sound as a phenomenon presents infinite possibilities for interpretation. Its analysis employs a wide berth of tools, each carrying conventions that shape the ensuing frameworks of its representation. We assert that visualization can be framed as a set of analysis techniques that has become indispensable to the study of audio data. Since the human experience of sound perception is inherently multidimensional, mapping audio features into visual parameter spaces should reflect this complexity. Classic visualization tools might carry presets or conventions that interfere invisibly with information processing by using assumptions transferred from different applied contexts. To address these concerns, we have proposed the introduction of new or updated software tools that use transparency, flexibility, and robustness to better align with the domain-specific needs of modern audio analysts. Like a mountain range when viewed from a different angle, new perspectives can offer new insights. It is our hope that in the adoption of such strategies, we may facilitate an environment that allows us to see beyond sound.

[1] J. B. J. Fourier, Théorie de la propagation de la chaleur dans les solides, Manuscrispt submitted to the Institute de France 21 Dec, 1807.

[2] J.B. J. Fourier, Théorie analytique de la chaleur. Paris: Chez Firmin Didot, Père et Fils, 1822.

[3] 1857. Scott de Martinville, É.-L. Fixation graphique de la voix.

[4] Bell, A. G. (1876). Improvement in Telegraphy. U.S. Patent No. 174,465. Filed February 14, 1876

[5] 1837 Telegraph. Cooke, W. F., & Wheatstone, C. (1837). Electric Telegraphs. UK Patent No. 7,390.

[6] Morse, Samuel F. B. “Caveat for the American Electro-Magnetic Telegraph.” Caveat filed October 3, 1837. Volume 5, Page 112. Records of the Patent and Trademark Office, Record Group 241. National Archives at Washington, D.C. Filed October 3, 1837

[7] Edison, T. A. (1878). Improvement in Speaking-Telegraphs. U.S. Patent No. 203,016. Filed March 27, 1878.

[8] Berliner, E. (1887). Gramophone. U.S. Patent No. 372,786. Filed May 4, 1887

[9] Rice, C. W., & Kellogg, E. W. (1925). "Notes on the Development of a New Type of Hornless Loudspeaker." Transactions of the American Institute of Electrical Engineers, *44*(1), 461–480. https://doi.org/10.1109/T-AIEE.1925.5061157

[10] Kopp, G. A., & Green, H. C. (1946). "Basic Aims of Visible Speech." The Journal of the Acoustical Society of America, 18*(1), 1–16. https://doi.org/10.1121/1.1916342

[11] Potter, R. K., Kopp, G. A., & Green, H. C. (1947). Visible Speech. D. Van Nostrand Company.

[12] FFT Cooley, J. W., & Tukey, J. W. (1965). "An algorithm for the machine calculation of complex Fourier series." Mathematics of Computation, *19*(90), 297–301. https://doi.org/10.1090/S0025-5718-1965-0178586-1

[13] Cioffi, J. M. (1991). A Multicarrier Primer. ANSI T1E1.4 Committee Contribution, *91-157*.

[14] ANSI T1.413-1995. (1995). Network and Customer Installation Interfaces - Asymmetric Digital Subscriber Line (ADSL) Metallic Interface. American National Standards Institute.

[15] Mouly, M., & Pautet, M.-B. (1992). The GSM System for Mobile Communications. ISBN: 2-9507190-0-7.

[16] Bingham, J. A. C. (1990). Multicarrier modulation for data transmission: An idea whose time has come. IEEE Communications Magazine,28(5), 5-14. https://doi.org/10.1109/35.54342

[17] Lauterbur, P. C. (1973). Image Formation by Induced Local Interactions: Examples Employing Nuclear Magnetic Resonance. Nature, *242*(5394), 190–191. https://doi.org/10.1038/242190a0

[18] Kumar, A., Welti, D., & Ernst, R. R. (1975). NMR Fourier Zeugmatography. Journal of Magnetic Resonance (1969), 18(1), 69–83. https://doi.org/10.1016/0022-2364(75)80224-3

[19] Berger, H. (1929). Über das Elektrenkephalogramm des Menschen. Archiv für Psychiatrie und Nervenkrankheiten, *87*(1), 527–570. https://doi.org/10.1007/BF01797193

[20] Bickford, R. G., et al. (1971). The Compressed Spectral Array (CSA) – A Pictorial EEG. Proceedings of the San Diego Biomedical Symposium, 10, 365–370.

[21] Blesser, B., & Lee, F. (1971). An Audio Delay System Using Digital Technology. Journal of the Audio Engineering Society, 19(5), 393-397.

[22] Blesser, B. A., et al. (1978). Apparatus and method for time domain compression and expansion of audio signals (U.S. Patent No. 4,085,286). September 30, 1975

[23] Flanagan, J. L., & Golden, R. M. (1966). Phase Vocoder. The Bell System Technical Journal, 45(9), 1493–1509. https://doi.org/10.1002/j.1538-7305.1966.tb01706.x

[24] Hildebrand, H. J. (1997). Method and apparatus for automatic pitch correction (U.S. Patent No. 5,973,252). Filed May 22, 1997

[25] Grey, J. and Moorer, J. (1977). Perceptual evaluation of synthesized musical instrument tones. Journal of the Acoustical Society of America, 62: 454–462. https://doi.org/10.1121/1.381508

[26] 2009. Matteo Milani, An interview with James A. Moorer, pt.1, Unidentified Sound Object. https://usoproject.blogspot.com/2009/02/interview-with-james-moorer-pt1.html

[27] 2012. Soundstream: The Introduction Of Commercial Digital Recording In The United States. . Simon Barber. Journal on the Art of Record Production. ISSN: 1754-9892. https://www.arpjournal.com/asarpwp/soundstream-the-introduction-of-commercial-digital-recording-in-the-united-states/

[28] 1996. Fairlight – The Whole Story. Reproduced from Audio Media Magazine, January 1996. https://www.anerd.com/fairlight/fairlightstory.htm

[29] Fairlight CMI History. Peter Vogel. Peter Vogel Instruments. https://petervogelinstruments.com.au/fairlight-history/#:~:text=Posted%20on%20August%2021%2C%202019,name%20for%20the%20new%20company.

[30] 1982. Commodore 64 User's Guide. By Commodore Business Machines, Inc. Internet Archive. Added 2017-08-01. Accessed 2025-08-25. https://archive.org/details/commodore-64-user-guide

[31] File Archives. Commodore 64 related manuals. Zimmers.net. Accessed 2025-08-25. https://www.zimmers.net/anonftp/pub/cbm/c64/manuals/index.html

[31] The Early Days of Software Sequencers. Posted 20 Dec 2010. https://www.kvraudio.com/focus/the_early_days_of_software_sequencers_15670

[32] Dr T: His world of electronic wizardry.Pyle, Derek. April 2017. Perfect Sound Forever. Accessed 2025-08-25. https://www.furious.com/perfect/drt.html

[33] Dr. T's Keyboard Controlled Sequencer. Badger, Mark. Mu:zines. Sound On Sound July 1987 https://www.muzines.co.uk/articles/dr-ts-keyboard-controlled-sequencer/2473

[34] 2020. Matrixsynth. Steinberg Pro16 sequencer from '86: Grandmother of Cubase. Posted 2020-09-03? Sept 9. Accessed 2025-08-25. https://www.matrixsynth.com/2020/09/steinberg-pro16-sequencer-from-86.html

[35] Steinberg Pro-24 III User Manual. M Hanemann and Co. London. https://www.manualslib.com/manual/1829156/Steinberg-Pro-24-Iii.html

[36] 1988. Steinberg Pro24 III. Simon Trask. Article from Music Technology, July 1988. https://www.muzines.co.uk/articles/steinberg-pro24-iii/1124

[37] A brief history of Pro Tools, by Future Music. Published 30 May 2011. Musicradar. Accessed 2025-08-25. https://www.musicradar.com/tuition/tech/a-brief-history-of-pro-tools-452963

[38] Digidesign Pro Tools. It's Cruel To Make A Computer Work This Hard. By Paul D. Lehrman. Article from Sound On Sound, January 1992. Accessed 2025-08-25. https://www.muzines.co.uk/articles/digidesign-pro-tools/9294

[39] November 1997. Sound on Sound. Steinberg Cubase VST v3.5. Cook, Janet H. https://web.archive.org/web/20140916001421/http://www.soundonsound.com/sos/1997_articles/nov97/cubasevst.html

[40] What is VST? VST 3 Developer Portal. Steinberg Media Technologies GmbH. https://steinbergmedia.github.io/vst3_dev_portal/pages/What+is+VST/Index.html

[41] Oppenheim, A. V., Willsky, A. S., & Nawab, S. H. (1996). Signals and systems (2nd ed.). Prentice Hall. http://materias.df.uba.ar/l5a2021c1/files/2021/05/Alan-V.-Oppenheim-Alan-S.-Willsky-with-S.-Hamid-Signals-and-Systems-Prentice-Hall-1996.pdf

[42] Smith, J.O. Mathematics of the Discrete Fourier Transform (DFT) with Audio Applications, Second Edition, online book, 2007 edition. accessed 25 Aug 2025. https://ccrma.stanford.edu/~jos/st/

[43] Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. The Journal of the Acoustical Society of America, 8(3), 185–190. https://doi.org/10.1121/1.1915893

[44] Time Series Classification via Topological Data Analysis. May 2017. Transactions of the Japanese Society for Artificial Intelligence32(3):D-G72_1-12. DOI: 10.1527/tjsai.D-G72

[45] An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. Chazal and Michel. Frontiers in Artificial Intelligence. 29 September 2021. Volume 4 – 2021 https://doi.org/10.3389/frai.2021.667963

[46] Perea, J. A., & Harer, J. (2015). Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis. Foundations of Computational Mathematics, 15(3), 799–838. https://doi.org/10.48550/arXiv.1307.6188

[47] Liu, J. Y., Jeng, S. K., & Yang, Y. H. (2016). Applying Topological Persistence in Convolutional Neural Network for Music Audio Signals. https://doi.org/10.48550/arXiv.1608.07373

[48] Bergomi, M. G., Baratè, A., & Di Fabio, B. (2016). Towards a Topological Fingerprint of Music. https://doi.org/10.48550/arXiv.1602.00739

[49] Interpreting CNN models for musical instrument recognition using multi-spectrogram heatmap analysis: a preliminary study. Front. Artif. Intell., 18 December 2024 Pattern Recognition. Volume 7 – 2024. https://doi.org/10.3389/frai.2024.1499913

[50] Bach-y-Rita P, Collins CC, Saunders F, White B, Scadden L (1969). "Vision substitution by tactile the image projection". Nature. 221 (5184): 963–964. Bibcode:1969Natur.221..963B. doi:10.1038/221963a0. PMID 5818337. S2CID 4179427.

[51] T. McDaniel; S. Krishna; V. Balasubramanian; D. Colbry; S. Panchanathan (2008). Using a haptic belt to convey non-verbal communication cues during social interactions to individuals who are blind. IEEE International Workshop on Haptic, Audio and Visual Environments and Games, 2008. HAVE 2008. pp. 13–18. doi:10.1109/HAVE.2008.4685291.

[52] Freeman, D., Reeve, S., Robinson, A., Ehlers, A., Clark, D., Spanlang, B., & Slater, M. (2017). Virtual reality in the assessment, understanding, and treatment of mental health disorders. Psychological Medicine, 47(14), 2393–2400. https://doi.org/10.1017/S003329171700040X

[53] Chu, C., Smith, L., & Duer, Z. (2020). The State of the Art in VR/AR Design Tools. In ACM SIGGRAPH 2020 Courses (pp. 1–70). Association for Computing Machinery. https://doi.org/10.1145/3388769.3407492

[54] Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. Ecological Informatics, 61, 101236. https://doi.org/10.1016/j.ecoinf.2021.101236

[55] Parsing Birdsong with Deep Audio Embeddings. 2021. IJCAI 2021 Artificial Intelligence for Social Good (AI4SG) Workshop. Irina Tolkova, Brian Chu, Marcel Hedman, Stefan Kahl, Holger Klinck. https://doi.org/10.48550/arXiv.2108.09203

[56] The Sound Labyrinth Project: Catalyst For Creative Activity. By Catharina Dyrssen, Anders Hultqvist, Staffan Mossenmark, Per Sjösten, Björn Hellström. ISSN: 2009-3578. Interference A Jounrnal of Audio Cultures. www.interferencejournal.org/the-sound-labyrinth-project

[57] Britannica, T. Editors of Encyclopaedia (2025, June 27). Flight recorder. Encyclopedia Britannica. https://www.britannica.com/technology/flight-recorder

[58] Ashby, William Ross (1956). An Introduction to Cybernetics. London: Chapman & Hall Ltd. ISBN 9781614277651. https://archive.org/details/introductiontocy00ashb/page/n7/mode/2up

[59] Raymond ES. The Cathedral and the Bazaar. [Internet]. First published 1997. [cited 2023]. Available from: http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar

[60] Payne RS, McVay S. Songs of humpback whales. Science. 1971;173(3997):585-597. https://doi.org10.1126/science.173.3997.585

[61] Payne KB, Langbauer WR, Thomas EM. Infrasonic calls of the Asian elephant (Elephas maximus). Behav Ecol Sociobiol. 1986;18(4):297-301. https://doi.org/10.1007/BF00300007

[62] von Muggenthaler E. Infrasonic and low-frequency vocalizations from Siberian and Bengal tigers. J Acoust Soc Am. 2000;108(5_Supplement):2541. https://doi.org10.1121/1.4743417

[63] von Muggenthaler E, Reinhart P, Lympany B, Craft RB. Songlike vocalizations from the Sumatran Rhinoceros (Dicerorhinus sumatrensis). Acoust Res Lett Online. 2003;4(3):83-88. https://doi.org10.1121/1.1588271

[64] Abelson H, Sussman GJ, with Sussman J. Structure and Interpretation of Computer Programs. Cambridge, MA: MIT Press; 1985. Available from: https://mitpress.mit.edu/sites/default/files/sicp/full-text/book/book-Z-H-7.html

[65] Chandler, P., & Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. British Journal of Educational Psychology, 62(2), 233-246.

[66] Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. Educational Psychology Review, 22(2), 123–138.

[67] 1990. Sweller, J., Chandler, P., Tierney, P., & Cooper, M. Cognitive load as a factor in the structuring of technical material. Journal of Experimental Psychology: General, 119(2), 176–192.

[68] De Vuyst, S., Tworus, K., Wittevrongel, S., & Bruneel, H. (2009). Analysis of stop-and-wait ARQ for a wireless channel. 4OR-A QUARTERLY JOURNAL OF OPERATIONS RESEARCH, 7(1), 61–78. https://doi.org/10.1007/s10288-008-0072-x

[69] Amdahl, Gene M. (1967). "Validity of the single processor approach to achieving large scale computing capabilities" (PDF). Proceedings of the April 18-20, 1967, spring joint computer conference on - AFIPS '67 (Spring). pp. 483–485. doi:10.1145/1465482.1465560. S2CID 195607370.

[70] 2017. Molzahn, D. K., Dörfler, F., & Sandberg, H. A Survey of Distributed Optimization and Control Algorithms for Electric Power Systems. DOE Office of Scientific and Technical Information. https://doi.org/10.1109/TSG.2017.2720471

[71] Bertin, J. (1983). Semiology of graphics: Diagrams, networks, maps (W. J. Berg, Trans.). University of Wisconsin Press. (Original work published 1967)

[72] Festinger, L. (1957). A Theory of Cognitive Dissonance. Stanford University Press.

[73] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340.