DISCIPLINARY TRANSLATIONS

07

WORD. SPOKEN.

Articulating the Voice for High Performance Sound Technologies for Access and Scholarship (HiPSTAS)

TANYA E. CLEMENT

Now accessing audio online seems easy. We find what we want to listen to through Google or through a search box on a favorite site. We can click on a link, open the file right in the browser, and then press play, fast-forward, and playback. In some cases, we can even view the sound waves or spectrograms associated with the audio or we can annotate what we hear and remix these representations. At the same time, modes of computational analysis with sound that let us search for sounds with sound or map sonic patterns across collections of audio, for example, remain few and relatively simplistic.

The editors of this collection have rightly asserted that digital sound studies must include technology as an object of study in order to attend to "the ways that various devices mediate sound, from the speaker and microphones to software coding and hardware development" (introduction). Using technologies to enhance access to and analysis of audio collections seems to promise a wide range of critical "close" and "distant" critical listening opportunities in digital sound studies, but there are still few conver-

sations about the many ways in which digital infrastructure technologies, or the hardware and software that facilitate these methods, influence scholarship.[1] To better understand these mediations in the context of developing tools for critical listening, this chapter considers classification systems for sound as a significant object of study for better understanding the digital infrastructure technologies that facilitate scholarship with audio.

Technologies used to facilitate scholarship with audio require a classification system to "mark" or annotate features of digital audio or text so that we can organize and search them more easily. By limiting the computer's search to identifying keywords or concepts such as an author name, a date range, or a genre (like horror or comedy, for example), we get expected results more quickly. Though they often seem invisible in the digital realm, classification systems reflect how we interact with machines as social and situated beings. Classification systems are subjective and deeply political: one person's horror movie could be another's comedy.

Classification or standardization protocols are subjective and political because they are sociotechnical phenomena—pertaining to both human and technical influences. A sociotechnical perspective sees technologies as "ways of life, social orders, practices of visualization" that are interdependent with the politics of knowledge production.[2] From this perspective comes the understanding that the classification standards we develop, which ultimately shape the knowledge produced through them and by them, are developed according to our own perceptions of the world.[3] So, while we need standardized protocols such as classification systems to make our hardware and software work more efficiently for everyone, we also need to learn how to interrogate these systems in order to understand how our assumptions and biases impact the knowledge we produce with these technologies.

To frame this study from a sociotechnical perspective and within the particularities of digital sound studies, this chapter considers a specific digital humanities project in sound—High Performance Sound Technologies for Access and Scholarship (HiPSTAS)—and a particular aspect of development within that project—the use of standardized classifications for describing sound features within the development of a tool for searching sound with sound.[4] Situating this aspect of development within HiPSTAS within a brief history of methods for classifying sound features will help us consider the impact that technology and politics can have in shaping scholarship in digital sound studies.

Sound in the HiPSTAS Project

A joint project of the School of Information at the University of Texas at Austin and the Illinois Informatics Institute at the University of Illinois at Urbana-Champaign, HiPSTAS was initially funded by the National Endowment for the Humanities as an Institute in Advanced Technologies in the Digital Humanities.[5] The HiPSTAS Institute included twenty junior and senior faculty and advanced graduate students as well as librarians and archivists in the humanities from across the U.S. interested in analyzing large collections of spoken-word audio collections using high-performance or "supercomputing" technologies. Among many collections of interest to the participants were 30,000 files of recordings from PennSound's poetry archive; 600,000 digital collections objects from the American Folklife Center at the Library of Congress; 30,000 hours of oral histories from StoryCorps; and 3,000 hours in the American Philosophical Society's Native American Collection, which includes recordings from more than fifty tribes across North America, among other collections. The participants met in two face-to-face meetings in May 2013 and May 2014 as well as in monthly virtual meetings. The objectives of the HiPSTAS Institute were threefold: first, to assess how these communities wanted to use computational tools to study spoken-word collections; second, to assess how those tools needed to be developed to support analyzing and visualizing large audio collections in the humanities; and third, to produce preliminary results with these tools using the collections of interest to the participants.

A significant aspect of the HiPSTAS Institute included introducing the participants to the Adaptive Recognition with Layered Optimization (ARLO) software. ARLO, which was originally developed by HiPSTAS co-PI David Tcheng for acoustic studies in animal behavior and ecology, had previously been used to search for bird calls across field recordings. Conceived to model a bank of hairs in the inner ear, which vibrate at different audio frequencies in response to sound waves, ARLO monitors and then samples each "hair's" instantaneous energy (a sum of the tuning fork's potential energy or the deflection of the fork and its kinetic energy based on the speed of the movement, per second). ARLO uses this data to create a 2D matrix of values (frequency vs. time) called a spectrogram. Essentially, these spectrograms (see fig. 7.1) show a map of sonic energy across time: each row of pixels represents a frequency band, and the color of each pixel represents the numeric value of total energy of that particular frequency (or how much the tuning fork trembles) for that point in time. ARLO uses these spectrograms to ex-

tract sonic features for machine-learning processes, including unsupervised learning such as clustering as well as supervised learning for classification.[6]

Used to search across sound collections for sonic patterns, these machine-learning processes rely on human intervention. To teach the software to identify sounds of interest with supervised learning techniques, human "experts" annotate the sounds they want to find and use these seed examples to teach an algorithm to find other, similar sounds. With unsupervised techniques, the "expert" still chooses certain features of the audio to guide how the machine-generated clusters are formed. Thus, software like ARLO finds sounds by comparing each training example to new, unlabeled examples and determining good matches as those that seem to have some of the same features, such as the total energy value described above. For the ornithologist who is examining thousands of hours of birdcalls, this process of matching might mean marking (or "tagging") examples of a particular bird's call on a spectrogram and asking the software to retrieve similar calls. In the case of a humanist, such as one of the scholars at the HiPSTAS Institute, this could mean tagging moments of laughter, applause, gunshots, or feedback noise to teach the machine to find more such events. In each case, the machine is taught with these seed examples to find or cluster what the expert has marked as interesting.

Machine-learning software like ARLO relies on many seed examples to train the algorithm. Consequently, realizing that the participants could produce more and possibly better seed examples if they worked together or with students, we developed a collaborative interface for tagging example sounds. Figure 7.1 shows the tagging interface we created for participants interested in analyzing the PennSound poetry archive.[7] The interface provides the listener with a two-second sample that has been randomly selected from PennSound's approximately 5,500 hours of audio. The listener chooses labels to apply to the sample and then receives the next example. In this way, the listener can easily and quickly "mark up" a collection with examples for machine learning.

The most significant aspect of this example for this discussion concerns how we chose the labels we used in the tagging interface. The tagging interface reflects a classification schema or set of rules that the PennSound poets and scholars chose for labeling the sound snippets.[8] They chose the classification schema, found in the "Transcriptions of Speech" section of the Text Encoding Initiative (TEI) P5 Guidelines for Electronic Text Encoding and Interchange, for conceptual and practical reasons. First, they chose this schema because they wanted classifications that reflected the patterns they

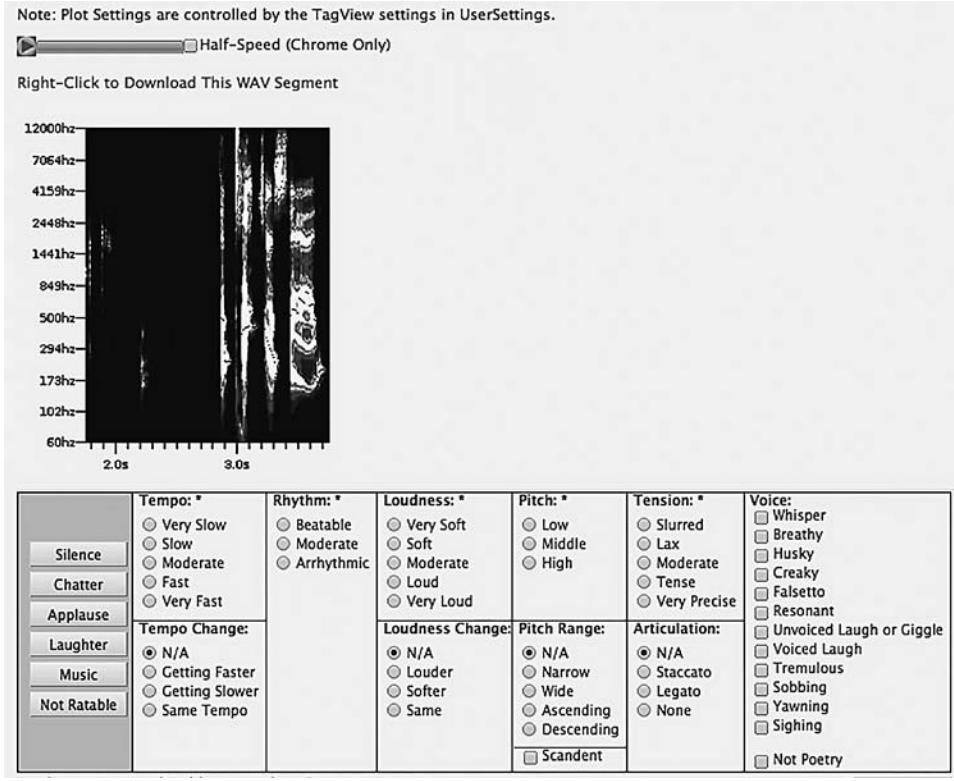

FIGURE 7.1 A tagging interface used to classify sound features on examples from PennSound.

sought to discover in their collection. In particular, the poets and scholars analyzing the PennSound collection were interested in analyzing the "vocal gestures" that Charles Bernstein (PennSound codirector) has argued "are available on tape but not page" and "are of special significance for poetry": namely clusters "of rhythm and tempo (including word duration)" and "of pitch and intonation (including amplitude), timbre, and accent."[9] By using descriptors from the TEI Transcription for Speech guidelines, the PennSound participants believed they had found terms that accurately described what they were hearing and what they wanted to find.

Second, the PennSound participants wanted a classification schema or standard that had been vetted by peers and that held the promise of facilitating future collaborations among projects that had already used (or might in the future use) these classifications. Released in November 2007, TEI P5 is a broad set of guidelines for an XML schema that is in wide circulation in

the digital humanities community. By using the TEI labels or schema, the PennSound participants hoped to create a set of descriptors that they might someday be able to use to compare classifications across PennSound and other audio collections.

The goal was to use the TEI classification schema to facilitate many uniform examples to train the machine-learning algorithm. Given a collection of two-second examples to tag, however, the thirteen participants assigned dramatically different tags to the same sample. One participant, for instance, might mark the same two-second sample "Beatable" with a "High" pitch and another might classify it as "Arrhythmic" with a "Low" pitch. Another issue arose when participants wanted to label contexts rather than snippets; they wanted more than the two-second window they were given by ARLO, and they wanted to tag the recording scenario (such as the sound of the room), the gender of the speaker, and the genre (such as music) as they perceived it, not according to the specified genre types that were provided by the TEI classification schema. That is, they wanted to label *the label* as it reflected their own listening perspectives, which were couched in complex understandings of culture, genre, and materiality, but our ARLO tagging interface, built using the TEI schema, would not allow them to do that.

Using this defined vocabulary or schema, which was meant to facilitate the process by providing uniformity across the examples, the PennSound participants debated how and when to implement the classifications. The PennSound participants struggled with labeling what they saw on the spectrograms, often citing doubts about their ratings and their understandings of the classifications and especially showing a resistance to the TEI classification system they had chosen to use. While classifying snippets of sound seemed to work well for the ornithologist, classifying snippets of poetry performances according to the chosen standard seemed to frustrate the humanist's desire to find dynamic or time-based aspects of performance. It seemed that while the sound of a bird could more easily be classified as "male cardinal," classifying or defining the human voice—an act that Jonathon Sterne calls a debate over "what it means to be human"—was a more provocative endeavor.[10] Realizing on the one hand that the classification system was necessary for increased computational productivity and efficiency but also, on the other hand, that it was flawed in its orientation, the PennSound participants did not seek to discard the use of a classification system but rather cited the need for a "better" (i.e., more accurate) classification system for describing sonic features as a high-priority requirement for moving ahead with developing ARLO.

I tell this story because it provokes sociotechnical questions for digital sound studies in general. How can a classification system, which is an infrastructural mainstay for facilitating computational analysis, mediate knowledge production? And how can we study these mediations? Bowker and Star suggest "infrastructural inversions" as a method for better understanding these interdependences between standardizations and knowledge production.[11] As the authors suggest, we must take into account that standardized classifications and systems are ubiquitous; they are both materially and symbolically realized as well as historically situated, representing multiple voices and silences.[12] Ultimately, classification systems reflect philosophies concerning the nature of sound as well as the practical politics involved in developing such standards that include what remains visible and invisible in the system.[13] In my example above, we see an example of how a classification system might work in a tool like ARLO. The next two sections consider the historic roots of this system to better understand why they might have seemed inaccurate or inappropriate to the PennSound scholars. In particular, I will consider the symbolic and material underpinnings of the TEI's Transcriptions of Speech classifications for sound within the history of philosophies in linguistics and the immediate political contexts that affected the establishment of these standardized rules.

A Brief Look at Prosodic and Paralinguistic Classifications

Linguists have been at the forefront of establishing complex and standardized protocols for describing spoken language. Driven by the desire to address the "practical needs of spoken language corpora annotation and analysis," especially in the light of more recent developments in computer-facilitated speech analysis, Maciej Karpiński outlines seventy-five years of research in linguistics concerning attempts to define and categorize what we say and how we say it.[14] In a specific example that is of particular use for this discussion, linguists often use "prosody" as a phenomenon comprising varying degrees of intonation, stress, and rhythm that convey meaning through phrasing and prominence, while they describe paralinguistic features as those that do not easily belong to a describable linguistic structure.[15] David Crystal and Randolph Quirk divide their seminal study *Systems of Prosodic and Paralinguistic Features in English* (1964) into prosodic and paralinguistic features based on how easily these features might be integrated into typical linguistic structures.

Specifically, Karpiński claims that prosody may be measured or described using three basic parameters—pitch frequency, duration, and intensity—and that these parameters influence each other as communicating features.[16] In written texts, prosodic features are typically described in terms of syntactical units. These language features often include parts of speech, accent, phoneme, stress, and tone as well as other information that influences how a sentence can be read such as the position of a word in a phrase (e.g., consecutive verbs or multiple nouns), sentence type (e.g., a declaration or a question), and information structure (e.g., independent versus dependent clauses, since inferable information in a dependent clause is usually deaccented).[17] In other words, when we seek to "sound out" a written word, we guess how to pronounce words unknown to us based on our experiences with prosodic features such as recognizable clues for pronunciation in the surrounding syntax. Nouns in a series require different amounts of stress, for instance, and questions have a lilt.

In comparison, paralinguistic features seem more difficult to describe and standardize. In his attempt to delineate terms, for example, Karpiński discusses paralinguistics within the context of three areas of study that include prosody, vocal quality, and gesture.[18] Also referred to as *timbre*, voice quality in musical instruments connotes the distinctive sound a particular instrument makes in contrast to another—such as the sound of an oboe versus that of a tuba—even when the instruments are playing the same note at a similar amplitude. For Karpiński, such vocal features are "individual, idiosyncratic, and further from 'language proper'" than prosodic features, making them "multidimensional and difficult to operationalize."[19] Crystal and Quirk also note the difficult and slippery nature of categorizing paralinguistic vocal qualities that surround such sounds as giggling, laughing, and crying:

> It is not possible to say when *giggle* ends and *laugh* begins, or when *cry* ends and *sob* begins, though doubtless it would be possible to examine a great quantity of data and obtain some measurements (of pulse speed, air pressure, prominence, for example) which would be of value in establishing more objective gradations.[20]

It is useful to note that Crystal and Quirk, who have attempted to systematize these voice-quality measures in *Systems*, put prosodic features on the more "describable" end of the classification continuum from prosodic to paralinguistic, even while they are quick to note that there is no sharp division between them. "It is doubtful," they write about implementing a system

of vocal-quality categories, "whether the results would justify the time and ingenuity involved."[21]

Certainly, how we perceive and make meaning with prosodic and paralinguistic features is a subjective activity. Dwight Bolinger asserts that intonation "is generally used to refer to the overall landscape, the wider ups and downs that show greater or lesser degrees of excitement, boredom, curiosity, positiveness, etc."[22] Further, in its expansiveness, prosody can signify elements of a speaker's identity including affect and emotional engagement, age, cognitive process and development, ethnicity, gender, and region and has been used to study human behavior, culture, and society.[23] For these reasons, Karpiński points out, prosodic and paralinguistic features are often considered "indexicals" since they seem to point to the context of a person or place.[24] Indeed, Karpiński describes paralinguistics such as laughter, giggles, gasps, pauses, hesitations, or coughs as "all the phenomena and features of a speaker's behaviour that go beyond the (current) limits of systematic linguistic description but still influence the way his/her communicational contribution is understood by his/her conversational partner."[25]

Tasked with submitting recommendations for the TEI's Transcriptions of Speech section of the guidelines, then, the TEI Spoken Text Working Group (STWG) relied on Crystal and Quirk's *Systems* and its assertions that prosodic and paralinguistic features influence meaning-making with sound as a basis for identifying which speech characteristics in recordings should be (and could be) marked in the guidelines.[26] This Crystal and Quirk perspective is reflected in TEI labels that include the following attributes:

- *Tempo:* Very Slow, Slow, Moderate, Fast, Very Fast
- *Rhythm:* Beatable (highly rhythmic), Moderate, and Arrhythmic (flat or ordinary speech)
- *Loudness:* Very Soft, Soft, Moderate, Loud, and Very Loud
- *Pitch:* Low, Middle, and High
- *Tension:* Slurred or Lax (for looser articulation), Very Precise or Tense (for pronounced articulation)
- Other classifications include Whisper, Breathy, Husky, Creaky, Falsetto, Resonant, Unvoiced Laugh or Giggle, Voiced Laugh, Tremulous, Sobbing, Yawning, and Sighing.

Notably, this list is not an exact reflection of Crystal and Quirk's work, in which voice *qualities*, which include different modes (normal voice, whisper,

breathiness, huskiness, creak, falsetto, and resonance); and voice *qualifications,* which ordinarily interrupt speech (laughter, giggling, tremulousness, sobbing, and crying) are considered separately.[27] In contrast, the TEI guidelines foreground the similarities between voice qualities and qualifications by grouping them together in a list of "other" classifications.

It is these voice-quality features, which are regarded by linguists as difficult to systematically categorize, that the HiPSTAS project participants found most compelling in their attempt to systematically annotate their spoken-word recordings. The voice quality or timbre aspects of paralinguistics, which Bernstein calls the "poet's aesthetic signature or acoustic mark," are particularly important in studying poetry performances.[28] As an indexical property, they appear in a spoken poem or performance as "a technical feature that can be used to form or deform social distinctions and variations."[29] Consequently, as mentioned, the PennSound scholars chose to adopt the TEI descriptors for philosophical reasons, because the terms, adopted from Crystal and Quirk, seemed to reflect their own concerns, but they also chose them for practical reasons, since they had been adopted by an authority (the TEI community) and seemed to promise some consistency across projects, authors, and poems of interest as well as offering future possibilities for collaboration with other projects using the TEI guidelines. The advantages that come with building such a system, however, belie not only practical concerns about the fact that marking up audio takes time and resources but also philosophical concerns as to the erasure of a long history of conversations about the subtle differences between voice quality and qualifications.

Three Compromises for Classifying Sound

Bowker and Star suggest a means by which we can better articulate the sociotechnical nature of classification systems. Defining such systems as "a rich set of negotiated compromises ranging from epistemology to data entry that are both available and transparent to communities of users," they challenge scholars to make such compromises readily apparent for consideration.[30] A primary compromise of interest for digital sound studies is one the introduction to the TEI guidelines articulates well: "An electronic representation must strike a balance between the following two, partially conflicting, requirements: authenticity and computational tractability."[31] *Authenticity,* in this sense, is subjective and corresponds to whether or not

a digital surrogate or representation seems "true" or accurate to a philosophy about or understanding of that phenomenon in the world. *Computational tractability* is the extent to which that representation is computable or representable in the computational environment, which includes the software, the platform, the hardware, and the networks being used to consider that representation. Thus, a philosophical concern for what is authentic in a community of scholars such as digital sound studies scholars must be in constant conversation with practical concerns for what is computationally tractable in a digital environment.

By learning to articulate the nature of these sometimes conflicting requirements (at once philosophical and practical), we are empowered in the digital sound studies community to impact how the systems used by the community are designed and implemented. Below, based on a close look at the history of how the paralinguistic voice qualities in TEI's Transcriptions of Speech schema came to be, and a consideration of how the HiPSTAS participants attempted to apply these guidelines with the ARLO software, I have suggested three more general areas of compromise for consideration in digital sound studies.

COMPROMISE #1: Moving from Text to Sound

The first compromise for consideration is one that balances a desire for "user friendly apps" against a desire for applications or software that fully represent the subtle characteristics of a phenomenon. We are used to polished and seemingly intuitive applications for searching, browsing, publishing, and teaching with text, but applications for searching, browsing, publishing, and teaching with sound are emergent, developing, and often "buggy." In such a context, we must consider the compromises inherent in choosing ease-of-use technologies over change-of-paradigm technologies.

For example, when the TEI STWG was tasked with submitting recommendations for the TEI's Transcriptions of Speech section of the guidelines, they focused on guidelines for marking up text-based transcriptions of recordings rather than guidelines for the faithful representation of the recordings' many sonic attributes.[32] This focus was the result of STWG's perspective on prosodic and paralinguistic features as problematic, such as "speaker overlap, pauses, hesitations, repetitions, interruptions," uncertainty, and context.[33] It is clear from citations in their extant working notes and drafts that the STWG were versed in the works of Svartvik and Quirk ("A Corpus of English Conversation") and Tedlock (*The Spoken Word*) and

considered paralinguistic and prosodic sound features expressive; yet based on the need to make a hierarchical representation of text a main tenet of TEI, they found these time-based and overlapping sound dynamics—such as pitch, speed, and tone—impractical to represent. (Indeed, encoding for recorded speech was not included at all in the original TEI P1 guidelines.)[34] In short, the STWG's theoretical or philosophical understanding of sound did not coordinate well with the means they had to express or represent this understanding. Notes from the working group's 1991 meeting reflect the compromises they knew they were making:

> In a brief discussion on performative features such as pitch, speed and vocalisation, LB [Lou Burnard] asked if these could not be regarded as analogous to rendition in written texts and treated in a similar way. It was generally felt that it would be better to mark these using milestone tags such as <tag>pitch.change</tag>, <tag>speed.change</tag> etc.[35]

The STWG concluded that topics including "quasi vocal things such as laughter, quasi lexical things such as 'mm,' prosody, parallel and discontinuous segments, uncertainty of transcription, uncertainty in general" needed "considerable further work."[36] And, these "quasi lexical things" remain peripheral to the guidelines even today.

This peripheral status is reflected materially in how these paralinguistic features are included in the TEI standards. The STWG relegated voice quality—paralinguistic characteristics such as pitch and speed, etc.—to a "shift" tag or element.[37] The "shift" element (<shift/>) is represented "as pairs of milestone tags marking positions of prominence . . . with the 'end' tag of the pair being replaced by a shift to normal."[38] The choice to use this kind of element is significant, because <shift/> requires the encoder to mark dynamic sound attributes in the encoded transcript as shifts to and from a "normal" speaking mode (see fig. 7.2).

Beyond assumptions about normativity that exist behind establishing a "normal" speaking mode, elements like the <shift/> are conceived as phenomena that happen in discrete moments of time. The <shift/> element occurs in one spot and marks specific points in a transcript, as if the dynamism of such sonic features could be pinpointed in time. Though the TEI guidelines are clear in the assertion that they are "not intended to support unmodified every variety of research undertaken upon spoken material now or in the future," the STWG's choice to show paralinguistic entities as "shifts" from a "normal" state and as discrete, "well-defined units" flies in the face of discussions by linguists such as Crystal, Quirk, and Karpiński,

```
<u>
 <shift feature="loud" new="f"/>Elizabeth
</u>
<u>Yes</u>
<u>
 <shift feature="loud" new="normal"/>Come and try this <pause/>
 <shift feature="loud" new="ff"/>come on
</u>
```

FIGURE 7.2 An example of the "shift" element from the TEI P5 guidelines.

who discuss the dynamic, subjective, and slippery nature of paralinguistic features. Indeed, the guidelines for these features were intended primarily for enabling the linguistic study of spoken text recordings as "a written or electronic representation of a stretch of speech which is treated for some purpose as a well-defined unit."[39] The material instantiation of these features in the <shift/> element shows how sound attributes become marginalized in computational infrastructures that focus on text and spoken language.[40]

Other, more recent projects for developing classification schemas for sound can help us imagine other compromises we must make in our attempts to balance our desire for the niceties of systems built for textual searches and our desire for new systems that better facilitate sonic searching. For example, the Federal Agencies Digitization Guidelines Initiative (FADGI) formed as a group in 2007 "to define common guidelines, methods, and practices to digitize historical content in a sustainable manner."[41] FADGI's metadata standard, "Embedded Metadata in Broadcast WAVE Files, Version 2," describes sonic information that points to sound's materiality, including signal chain specifics, sample rates, and bit depth. Further, other classification schemas proposed by the International Association of Sound and Audiovisual Archives (IASA) capture information concerning an audio file's provenance and historical context such as the date and place of a recording.[42] Even with these advancements, questions remain concerning the extent to which classifications such as FADGI's help us better understand vocal gestures and whether narrative descriptions of soundscapes give us enough information about sonic histories. Ultimately, these standards are works in progress and the sociotechnical histories behind the development of these standards also reflect compromises, both philosophical and practical, that organizations other than the TEI will make based on a desire to balance their situated understanding of sound, the perceived needs of the communities they serve, and the technologies they hope to employ in the service of these goals.

COMPROMISE #2: Moving from Fixed to Emergent Meanings

Another compromise to consider in digital sound studies is one that weighs a desire to represent sounds as fixed in meaning against the difficult work of representing sounds as phenomena with emergent and multiple meanings. This is a significant compromise to address because any digital representation of the experience of sound will need be, by nature, a reduction that nonetheless invites expansive thinking.

Sound studies scholars in the humanities have been primarily interested in articulating sound culture in all of its complexity rather than in simplified, linear, or atomistic terms. For instance, citing Jacques Derrida, Dennis Tedlock dismisses "the entire science of linguistics, and in turn the mythologics (or large-scale structuralism) that has been built upon linguistics," since such sciences and mythologics are "founded not upon a multidimensional apprehension of the multidimensional voice, but upon the unilinear writing of the smallest-scale articulations within the voice."[43] Michael Chion argues that a recorded artifact has fixity that is necessary for close listening since to perceive sonic traits, one must listen repeatedly to a recorded moment, but he dismisses the state of fixedness that a framework like a classification system would engage since within it sounds "acquire the status of veritable objects" and "physical data"; this fixed data, he asserts, is inauthentic since it does not represent what was actually heard within the real time of "presence."[44] Likewise, Bernstein notes that "systems of prosodic analysis" that regularize sound "break down before the sonic profession of reading: it's as if 'chaotic' sound patterns are being measured by grid-oriented coordinates whose reliance on context-independent rations is inadequate."[45] These statements reflect an understanding of sound in the humanities as an emergent phenomenon that is dynamic and in flux and that evolves and expands over time, constantly introducing ambiguity and uncertainty. As such, there is a clear resistance toward "fixing" sounds for better understanding of meaning-making processes.[46]

Reduction as a means of representation is unavoidable in a digital context, but there are choices that dictate the terms of these reductions. Most of the categorizations outlined above, for example, have been established from the perspective of linguistic study and, for the most part, in terms of creating transcriptions from audio files. In contrast, classification systems devised by poets may be designed to represent the differences between breathy or harsh voices; a system designed by historians may better show the sounds of a city venue; and one designed by Native Americans might

facilitate comparing the changing paces of elders' stories. In each of these scenarios, one can imagine that certain sonic attributes are foregrounded based on the interests of a particular community.

A compromise that helps us better engage the emergent nature of sound hermeneutics in digital space is a choice that may be quite productive in digital sound studies. For instance, Kenneth Sherwood cites fixed and discrete instances of repetition as perceptible signs of emergence that signify elaboration and versioning. Bernstein notes the signification of dynamic "performative gestures" such as emotional intensification, which can map to measurable changes in heightened and decreased sound frequencies and speed.[47] Likewise, Crystal and Quirk have identified measurements for establishing the emergent dynamics of voice qualifications as "objective gradations" by "setting up parameters for degrees of pulsation types, pulsation speed, oral aspiration, nasal friction, air pressure, amplitudes and frequency of vocal cord vibration, and volume and tension of supraglottal cavities."[48] These examples demonstrate that the dynamics of a voice—its increasing or decreasing pace, its tone changes over time—can be understood against different frames of reference that we may choose to position as "fixed" (such as the words of a poem) even as we understand them to be in flux. This choice against fixity can be forwarded by classifications that help us better articulate and understand the terms of fixity as choices.[49] As such, sound as a phenomenon of emergence could be understood in terms of how it is represented as fixed.

COMPROMISE #3: Moving from Discrete to Contextual

A third compromise for consideration in digital sound studies entails balancing the desire for representing sound as a discrete event in time with the difficult work needed for describing sound across particular time contexts. In the previous section, I argue that we must represent fixed points in the sonic event in order to study the emergences of meaning. We must also better understand how we represent these features in fixed moments of time. Repeated, elaborated, or intensified moments can be marked as discrete, for example, even as their significance is based on their relationships across time with other moments. By constellating fixed moments in relationship to each other and situating them as patterned contexts over time, then, we may do the difficult work that we must do to develop classification systems that use fixity as means for representing contextualization and emergence.

Current guidelines for describing the historical context of recordings can provide an example of such a compromise between fixed and fluid representations. The TEI, for instance, includes a "recording" element that allows the encoder to describe dates; times of day; statements of responsibility for authors, editors, producers, etc.; the recording equipment used; or whether a broadcast recording is the basis of the text being transcribed and described. As well, there is a provision for adding elements that also describe the "setting" of a recording and its "participants." The FADGI guidelines contain these fields as well as the "bext chunk," which holds data on the digitizing process (including the analog source recording), on the capture process, on information about the storage of the file, and on versions of the coding history related to the file itself.[50] In many cases, these are optional fields that remain empty even as this contextual information impacts how we perceive the relationships that are marked and ultimately what and how we hear.

The choice to include this kind of contextual information reflects a desire to articulate design standards that do not just report on relationality but rather encode it. Innovative and productive work for representing relationality is already happening in the context of speech transcriptions. IASA recommends the Resource Description Framework (RDF), a World Wide Web Consortium (W3C) specification that allows humanists to describe relationships between objects on the web. Currently used by ARC (Advanced Research Consortium), for example, to provide gateways and venues for peer-reviewing digital projects for a variety of disciplines in literary study, RDF facilitates searching and finding relationships across projects in Networked Infrastructure for Nineteenth-Century Electronic Scholarship (NINES), 18thConnect (focused on eighteenth-century scholarship), the Medieval Electronic Scholarly Alliance (MESA), Renaissance Knowledge Network (ReKN), and Modernist Networks (ModNets). Using RDF, the ARC infrastructure is powerful, because each of the ARC nodes has its own stand-alone interface, but all of the resources can be searched together through the ARC catalog. A search on NINES can be modified to find objects from MESA, for example. While further work needs to be done to imagine an RDF schema that reflects relationships across sonic features of interest, using something like the ARC infrastructure for sound files could mean cross-searching that includes sound resources, too, which in turn would enable scholars to better collaborate on digital audio projects on a local and global scale across disciplines and interests. This kind of relationality is similarly the future of new International Image Interoperability Framework

(IIIF) guidelines for facilitating better access to audio collections through application programming interfaces (APIs).

A clear next step is to build tools that facilitate the ability to act on encoded relationality. Karpiński, for instance, proposes a "coherent approach" to linguistic data annotation that would take into account the indiscrete nature of speech prosody and voice-quality features.[51] Recommending that we treat these features as continua rather than categories, Karpiński argues that prosodic and paralinguistic features are multifunctional, multimodal, and multileveled, as well as both global and local; as such, he recommends implementing "sliders or joysticks for data input and to refrain from imposing any points on the scale, such as from a stable to a trembling voice, with all intermediate states possible."[52] Thomas Schmidt also suggests practical solutions for implementing varied perspectives on sound, such as bringing seven tools that are most commonly used by linguists for spoken language transcriptions—ANVIL, CLAN/CHAT, ELAN, EXMARaLDA Partitur-Editor, FOLKER, Praat, and Transcriber—to a common TEI schema.[53] Karpiński's and Schmidt's interventions suggest compromises that encompass the practical issues related to a need for discrete categorizations, such as annotations for linguistic data or a TEI schema, with the need to represent relationships across perspectives from multiple communities in multiple contexts across time.

Conclusion

Classification schemas for sound are language-based: they are themselves texts that attempt—sometimes with frugal and other times with rich results—to describe the world of sound that is always beyond text, beyond a listener, beyond one single snippet of a recording played back at one point in time. To approach the complexities that characterize our experiences with sound, there are many more philosophical and practical compromises we will have to negotiate as we continue to develop productive infrastructures for digital sound studies. We will need to consider what it means to engage sound thoughtfully, expansively, and critically with computational instruments that are often modeled on the normative practices of "hearing" with the ear when the ear is not the only hearing instrument. "I can hear more plainly through my teeth than through the external ear," Thomas Edison admits; "A stick touching a music box and placed between my teeth enables me to enjoy the music."[54] Another compromise will entail balancing well-

intentioned plans to incorporate crowd-sourced listener responses with the practical need for clean and manageable digital sound data. Tsur reminds us, for instance, that "sophisticated electronic instruments do give an accurate analysis of the sound information; but what really matters is its integration as it takes place in the brain" of each listener.[55] We must learn to balance this desired sophistication with the vast amount of data that a systems manager or a researcher would then have to manage, process, clean, and analyze. The technologies we are using are situated, personal, and political, but they also require practical interventions for use; they are indeed ways of life.

The ultimate compromise digital sound studies will face in negotiating authenticity and computational tractability is not new to sound studies: it includes any attempt to perceive the world outside the biases we bring to everything we do. In his 1889 article "On Alternating Sounds," for instance, Franz Boas considers the extent to which a philologist's field notes reflect the phonetics of his own language and writes that in the field, philologists "reduce to writing a language which they hear for the first time and of the structure of which they have no knowledge whatsoever. . . . Each apperceives the unknown sounds by the means of the sounds of his own language."[56] Indeed, it is the compromises that we will make as we model, engage, and interpret digital sound in new ways that will provide opportunities for provocation and for questioning our unavoidable biases as listeners.

NOTES

1 Bernstein, *Close Listening;* Clement, "Distant Listening."
2 Quote from Haraway, "Situated Knowledges," 583. See also Bowker and Star, *Sorting Things Out;* Bardzell and Bardzell, "Towards a Feminist HCI Methodology"; Berg, "Politics of Technology"; Feinberg, "Two Kinds of Evidence"; Frohmann, *Deflating Information.*
3 Bowker and Star, *Sorting Things Out,* 34.
4 Tanya E. Clement is the primary investigator of the HiPSTAS project. Please see more information at www.hipstas.org.
5 The project continues with a second NEH grant from Preservation and Access for Research and Development, titled HiPSTAS for Research and Development with Repositories (HRDR). Please visit www.hipstas.org.
6 Downie et al., "Novel Interface Services."
7 Launched January 1, 2005, PennSound is the largest collection of poetry sound files available for noncommercial distribution on the Internet. PennSound is

codirected by Charles Bernstein and Al Filreis and associated with the Center for Programs in Contemporary Writing and School of Arts and Sciences Computing at the University of Pennsylvania.

8 The meeting took place at the PennSound offices in Philadelphia, October 26, 2013, including the PennSound directors Al Filreis and Charles Bernstein as well as a host of their senior editors and technical advisors (Michael Henessey, Chris Martin, Steve McLaughlin, Danny Snelson, and others).

9 Bernstein, *Attack of the Difficult Poems*, 126.

10 Sterne, "Sonic Imaginations," 11.

11 Bowker and Star, *Sorting Things Out*, 37.

12 Bowker and Star, *Sorting Things Out*, 37–40.

13 Bowker and Star, *Sorting Things Out*, 44.

14 Karpiński, "Boundaries of Language."

15 Rooth and Wagner, "Harvesting Speech Datasets"; Crystal and Quirk, *Systems.*

16 Karpiński, "Boundaries of Language," 41.

17 Becker et al., "Rule-Based Prosody."

18 Karpiński, "Boundaries of Language." While gestures are very important to any oral performance such as poetry, they are not the focus of this study.

19 Karpiński, "Boundaries of Language," 43.

20 Crystal and Quirk, *Systems*, 42.

21 Crystal and Quirk, *Systems*, 42.

22 Bolinger, *Intonation and Its Parts*, 11.

23 Rooth and Wagner, "Harvesting Speech Datasets."

24 Karpiński, "Boundaries of Language."

25 Karpiński, "Boundaries of Language," 47.

26 Karpiński, "Boundaries of Language"; Schmidt, "TEI-Based Approach." TEI, *TEI P5* cites Boase (*London-Lund Corpus*) as a reference for these materials, but this text was never published (according to Edwards and Lampert, *Talking Data*) and is no longer available. A conversation with Lou Burnard led me to Crystal and Quirk (*Systems*) as an alternative reference for these descriptions.

27 Crystal and Quirk, *Systems.*

28 Bernstein, *Attack of the Difficult Poems*, 127.

29 Bernstein, *Attack of the Difficult Poems*, 127.

30 Bowker and Star, *Sorting Things Out*, 34.

31 TEI Consortium, *TEI P5.*

32 See Johansson et al., "TEI AI2 M1"; "TEI AI2 M2"; and "TEI AI2 W1."

33 Johansson, "Encoding of Spoken Texts," 150.

34 Sperberg-McQueen and Bumarde, "Guidelines."

35 Johansson et al., "TEI AI2 M1."

36 Johansson et al., "TEI AI2 M1."

37 TEI Consortium, *TEI P5.*

38 Johansson et al., "TEI AI2 M2."

39 Johansson, "Encoding of Spoken Texts," 149. The working group's final recom-

mendations reiterate this definition: "The goal of an electronic representation is to provide *a text* which can be manipulated by computer to study the particular features which the researcher wants to focus on" (Johansson et al., "TEI AI2 W1"); and the current guidelines state conclusively: "Speech regarded as a purely acoustic phenomenon may well require different methods from those outlined here, as may speech regarded solely as a process of social interaction" (TEI Consortium, *TEI P5*).

40 TEI Consortium, *TEI P5.*

41 FADGI, "Embedded Metadata in Broadcast WAVE Files."

42 These include CIDOC's Conceptual Reference Model (CRM), the Library of Congress's Functional Requirements for Bibliographic Records (FRBR), the Dublin Core Metadata Initiative (DCMI), Contextual Ontology Architecture (COA), and the Motion Picture Experts Group rights management standard, MPEG-21.RDF (IASA). These features are also captured in TEI P5.

43 Tedlock, *Spoken Word,* 249. Tedlock's *Spoken Word* is also referenced in Johansson et al., "TEI AI2 M1"; Johansson et al., "TEI AI2 M2"; Sherwood, "Elaborate Versionings"; and Bernstein, *Close Listening.*

44 Chion, "Three Listening Modes," 50.

45 Bernstein, *Close Listening,* 13.

46 Sherwood, "Elaborate Versionings."

47 Sherwood, "Elaborate Versionings"; Bernstein, *Attack of the Difficult Poems,* 127.

48 Crystal and Quirk, *Systems,* 42.

49 For example, Adriana Cavarero argues for demythicizing oral performance. She critiques the viewpoint of Chion ("Three Listening Modes"), McLuhan (*Essential McLuhan*), and Ong (*Presence of the Word*), who at once essentialize the voice as "presence" and disembody and mythicize orality. Oral culture, McLuhan argues, gives "us simultaneous access to all pasts. As for tribal man, for us there is no history. All is present, and the mundane becomes mythic" (*Essential McLuhan,* 370). With this viewpoint, Cavarero asserts, we treat language as a code "whose semantic soul aspires to the universal" and render "imperceptible what is proper to the voice" ("Multiple Voices," 530).

50 FADGI, "Embedded Metadata in Broadcast WAVE Files."

51 Karpiński, "Boundaries of Language."

52 Karpiński, "Boundaries of Language," 47.

53 Schmidt, "TEI-Based Approach."

54 "An Interesting Session."

55 Tsur, *Poetic Rhythm,* 14.

56 Boas is quoted in Tsur, *Poetic Rhythm,* 51.

WORKS CITED

Bardzell, Shaowen, and Jeffrey Bardzell. "Towards a Feminist HCI Methodology: Social Science, Feminism, and HCI." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 675–84. CHI '11. New York: ACM, 2011.

Becker, S., M. Shröder, and W. Barry. "Rule-Based Prosody Prediction for German Text-to-Speech Synthesis." In *Proceedings of Speech Prosody 2006,* edited by Rüdiger Hoffman and Hansjörg Mixdorff, 503–6. Dresden: TUD Press, 2006.

Berg, Marc. "The Politics of Technology: On Bringing Social Theory into Technological Design." *Science, Technology, and Human Values* 23, no. 4 (1998): 456–90.

Bernstein, Charles. *Attack of the Difficult Poems: Essays and Inventions.* Chicago: University of Chicago Press, 2011.

Bernstein, Charles. *Close Listening: Poetry and the Performed Word.* New York: Oxford University Press, 1998.

Boas, Franz. "On Alternating Sounds." *American Anthropologist* 2, no. 1 (1889): 47–54.

Boase, S. *London-Lund Corpus: Example Text and Transcription Guide.* London: Survey of English Usage, University College London, 1990.

Bolinger, D. *Intonation and Its Parts: Melody in Spoken English.* Stanford, CA: Stanford University Press, 1986.

Bowker, Geoffrey C., and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences.* Cambridge, MA: MIT Press, 2000.

Buckland, Michael K. "Information as Thing." *Journal of the American Society for Information Science* 42, no. 5 (1991): 351–60.

Cavarero, Adriana. "Multiple Voices." In *The Sound Studies Reader,* edited by Jonathan Sterne, 520–32. New York: Routledge, 2012.

Chion, Michael. "The Three Listening Modes." In *The Sound Studies Reader,* edited by Jonathan Sterne, 48–53. New York: Routledge, 2012.

Clement, Tanya. "Distant Listening: On Data Visualisations and Noise in the Digital Humanities." *Digital Studies* 3, no. 2 (2012).

Crystal, David, and Randolph Quirk. *Systems of Prosodic and Paralinguistic Features in English.* The Hague: Mouton, 1964.

Downie, J. S., D. K. Tcheng, and X. Xiang. "Novel Interface Services for Bioacoustic Digital Libraries." In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries,* 423. New York: ACM, 2008.

Edwards, Jane A., and Martin D. Lampert. *Talking Data: Transcription and Coding in Discourse Research.* New York: Psychology Press, 2014.

Federal Agencies Digitization Guidelines Initiative (FADGI). "Embedded Metadata in Broadcast WAVE Files, Version 2," April 23, 2012. www.digitizationguidelines.gov/guidelines/digitize-embedding.html.

Federal Agencies Digitization Guidelines Initiative (FADGI). "About," December 10, 2010. www.digitizationguidelines.gov/about.
Feinberg, M. "Two Kinds of Evidence: How Information Systems Form Rhetorical Arguments." *Journal of Documentation* 66, no. 4 (2010): 491–512.
Frohmann, Bernd. *Deflating Information: From Science Studies to Documentation.* Toronto: University of Toronto Press, 2004.
Haraway, Donna. "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective." *Feminist Studies* 14, no. 3 (1988): 575–99.
"An Interesting Session Yesterday at the National Academy of Sciences." *Washington Star,* April 19, 1878.
International Association of Sound and Audiovisual Archives (IASA). "Metadata." Accessed November 19, 2017. www.iasa-web.org/tc04/metadata.
Johansson, Stig. "The Encoding of Spoken Texts." *Computers and the Humanities* 29, no. 2 (1995): 149–58.
Johansson, Stig, Lou Burnard, Jane Edwards, and And Rosta. "TEI AI2 P1 [Spoken Texts Workgroup] Objectives and Deadlines 22 October 1990." TEI Consortium, October 22, 1990. www.tei-c.org/Vault/AI/ai2p01.tei.
Johansson, Stig, Lou Burnard, Jane Edwards, and And Rosta. "TEI AI2 M1 Minutes of Meeting Held at University of Oslo." TEI Consortium. August 9–10, 1991. www.tei-c.org/Vault/AI/ai2m01.txt.
Johansson, Stig, Lou Burnard, Jane Edwards, and And Rosta. "TEI AI2 M2 Minutes of Meeting Held at Oxford University." TEI Consortium. September 29 and October 1, 1991. www.tei-c.org/Vault/AI/ai2m02.txt.
Johansson, Stig, Lou Burnard, Jane Edwards, and And Rosta. "TEI AI2 W1 Working Paper on Spoken Texts University College London." TEI Consortium. October 1991. www.tei-c.org/Vault/AI/ai2w01.txt.
Karpiński, Maciej. "The Boundaries of Language: Dealing with Paralinguistic Features." *Lingua Posnaniensis* 54, no. 2 (2012): 37–54.
McLuhan, Marshall. *Essential McLuhan.* Edited by Eric McLuhan and Frank Zingrone. New York: Basic Books, 1995.
Mills, M. "Deaf Jam: From Inscription to Reproduction to Information." *Social Text* 28, no. 1 (2010): 35–58.
Ong, Walter J. *The Presence of the Word: Some Prolegomena for Cultural and Religious History.* New Haven: Yale University Press, 1967.
Pound, Ezra. *Polite Essays.* London: Faber & Faber, 1937.
Rooth, Matt, and Michael Wagner. "Harvesting Speech Datasets for Linguistic Research on the Web." Digging into Data Conference, National Endowment for the Humanities, Washington, DC, 2011. Accessed January 11, 2018. https://ecommons.cornell.edu/handle/1813/34477.
Schmidt, Thomas. "A TEI-Based Approach to Standardising Spoken Language Transcription." *Journal of the Text Encoding Initiative* 1 (June 2011): n.p. http://jtei.revues.org/142.

Sherwood, K. "Elaborate Versionings: Characteristics of Emergent Performance in Three Print/Oral/Aural Poets." *Oral Tradition* 21, no. 1 (2006): 119–47.

Sperberg-McQueen, M., and L. Bumarde, eds. "Guidelines for the Encoding and Interchange of Machine-Readable Texts." Draft version 1.0. Chicago and Oxford: Association for Computers and the Humanities/Association Computational Linguistics/Association for Literary and Linguistic Computing, 1990. Accessed November 19, 2017. https://quod.lib.umich.edu/cgi/t/tei/tei-idx?type=HTML&rgn=DIV2&byte=64782.

Sterne, Jonathan. "Sonic Imaginations." In *The Sound Studies Reader,* edited by Jonathan Sterne, 1–18. New York: Routledge, 2012.

Svartvik, J., and R. Quirk, eds. *A Corpus of English Conversation.* Lund, Sweden: Lund University Press, 1980.

Tedlock, Dennis. *The Spoken Word and the Work of Interpretation.* Philadelphia: University of Pennsylvania Press, 1983.

TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Version 2.6.0, January 20, 2014. www.tei-c.org/Guidelines/P5.

Tsur, Reuven. *Poetic Rhythm: Structure and Performance: An Empirical Study in Cognitive Poetics.* Brighton, UK: Sussex Academic Press, 2012.