DHQ: Digital Humanities Quarterly
2023
Volume 17 Number 2

Slow Listening: Digital Tools for Voice Studies

Marit J. MacArthur <mjmacarthur_at_ucdavis_dot_edu>, University of California, Davis
Lee M. Miller <leemiller_at_ucdavis_dot_edu>, University of California, Davis

Abstract

Sound studies in general, and voice studies in particular, present particular challenges for digital humanities scholarship. The software tools available to digital humanists who want to study performative speech are less familiar and less developed for our uses, and the user base is also much smaller than for text mining or network analysis. This article provides a critical narrative of our research and an outline of our methodology, in applying, developing and refining tools for the analysis of pitch and timing patterns in recorded performances of literary texts. The primary texts and audio considered are poetry readings, but the tools and methods can and have been applied more widely to podcasts, talking books, political speeches, etc.

1.

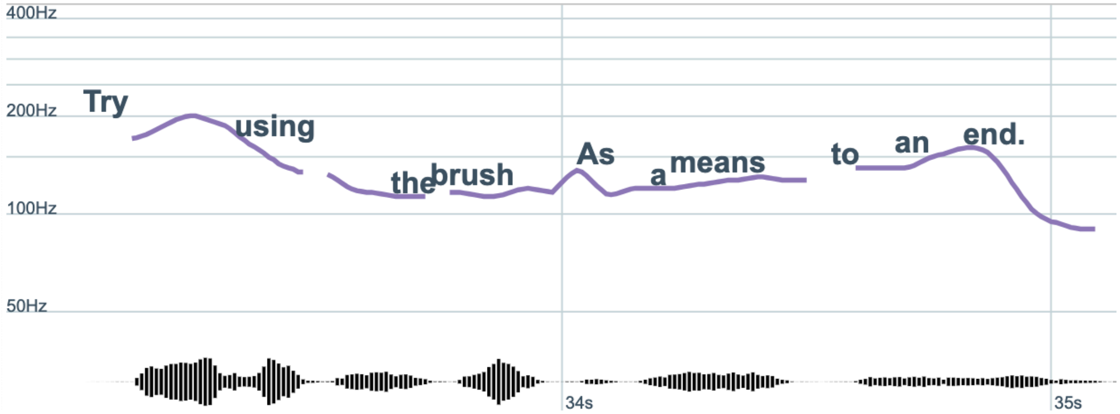


Figure 1. Pitch contour in Drift. John Ashbery reading “The Painter”, ll. 9-10, 92nd St. Y, New York, NY, 1952

All speech is performative, to some degree, and any text can be performed. Many have been, and we have digital access to many audio recordings of such performances. Which aspects capture our attention when we listen to — and study — these recordings as performative speech? When speech happens so fast? When tone of voice matters as much, if not more than, semantics? 1

> “*I'm not angry!*” The young woman insists — and we believe her angry tone of voice, not her words.
> see [Kreiman and Sidtis 2011, 306]

When we listen to a speaker, we guess at the speaker's mood and identity, we weigh semantics against tone — “her tone of meaning ... without the words” in Robert Frost's phrasing — and interpret accordingly [Frost, 308]. How far can we trust our ears? How do our brains process a voice? How do we filter a voice through our personal history of listening experiences, however much we intend to listen with an open mind? 2

Scientific approaches, from phonetic linguistics and the neuroscience of speech perception, offer some answers — but 3 as a rule, they have avoided the cultural, aesthetic, historical and political questions that humanists like to ask. And they rarely take, as their object of study, something like a poetry recording. If they do, they aren't likely to bring knowledge of literary or performance history, or Erving Goffman's concept of frame analysis [Goffman 1981], to bear on an individual performance of a literary or dramatic text for a particular audience. Not only that, the software tools used by linguists are not designed for the noisy, older recordings common in the audio archive of, for instance, poetry readings. Few humanities scholars, moreover, are trained in audio signal processing or linguistic analysis.

Sound studies in general, and voice studies in particular, do not dominate digital humanities scholarship and probably never will, amid our overwhelmingly visual and textual culture — no matter that we are living in the Golden Age of the Podcast. In practical terms, this means that the software tools available to digital humanists who want to study performative speech are less familiar, and less developed for our uses, than tools for text mining or network analysis, for instance. The user base is also much smaller. These are all challenges we have faced in choosing from the available tools for our research in voice studies, and in developing new digital tools. But the digital study of literary recordings has advanced in the last decade, and its future looks bright. 4

2.

Before we discuss those tools, however, some theoretical framing of our approach to digital voice studies is in order. Close listening remains an inspiring concept that, through the influence of Charles Bernstein's *Close Listening: Poetry and the Performed Word* [Bernstein 1998], has generated considerable scholarship on poetry recordings. Despite the fact that Bernstein co-founded one of the largest online archives of poetry audio, PennSound, close listening, when practiced in published scholarship, has often maintained too much distance from actual recordings of poetry reading. Cantankerous old-school scholars used to complain about theoretical readings of literature that fly “10,000 feet above the text”. Much work in sound and voice studies is still, to our frustration, visual and textual — published in print form, without audio recordings linked or embedded in the work, maintaining and enforcing a distance from the audio by failing to make interaction with audio an essential part of scholarship about it.[1] When published electronically, work in sound studies does increasingly embed audio and allow for interaction with it,[2] but there is a long way to go to make this the norm in the field. 5

In *The Audible Past: Cultural Origins of Sound Reproduction*, Jonathan Sterne advances a persuasive critique of conventional assumptions about hearing versus vision, which he calls “the audiovisual litany,” including the notions that “hearing tends toward subjectivity, vision tends toward objectivity” and “hearing is a temporal sense, vision is primarily a spatial sense” [Sterne 2003]. Of course, vision is no more objective than hearing, and both hearing and vision function spatially and temporally. A child looking for her mother, for instance, can call out, “Where are you?” And the mother can simply say “In here”, because she knows the child can tell where in the house her voice is coming from. Nevertheless, a photograph stays still when we study it. And the words of a poem by Emily Dickinson, even as we recognize that no version of the poem is authoritative, hold still on the page while we study them. A recording of the same words does not, nor do our impressions of what we heard. Perhaps we liked the performance, or we didn't, or we liked some things about it, and we try to explain why. Perhaps we mistrust ourselves, and feel we need to listen again. 6

Why does this matter to the study of poetry and poetry performance? Because poetry is an oral form, and scholarly work on poetry performance and other performative speech is still in its impressionistic infancy. When we listen to a voice, Feeling is First (with nods to E.E. Cummings and Lauren Berlant), and our inchoate feelings about voices bear scrutiny. As we have written elsewhere [MacArthur and Miller 2018], the complexity of our perceptions of voices highlight the fact that poetry is, after all, stylized communication. It calls on deep, evolutionarily old neural pathways that support language as speech, but not language as written. Writing is a recent kludge, an addon, not a deep, inevitable function of the human brain like speech — and even deeper and older are the vocal elements that create the tone of a voice, apart from words or semantic meaning (think grunts, sighs, etc.). 7

What are often called paralinguistic elements of the voice, the pitch, the intensity, the tempo, etc., are also pre — linguistic — preceding our knowledge and development of language, and triggering emotional responses that have little or nothing to do with the semantics of words. As such, the tone of voice exerts a direct emotional impact that is 8 wordless, as when we hear a baby's cry — or when we hear a speaker's tone of voice conflicting with the semantic meaning of her words, as when someone angrily insists, “*I'm not angry!*” we cannot ignore these elements if we really want to understand the experience of poetry, or of any performative speech.

Studying performative speech should also mean listening to ourselves listening, like the proverbial recursively self-conscious ethnographer. In *The Race of Sound: Listening, Timbre and Vocality in African American Music*, musicologist Nina Sun Eidsheim discusses the question we ask when listen to a voice, “Who is this?” [Eidsheim 2019, 9]. We ask this question of an unknown caller when we answer the phone. Indeed we ask it when we hear any disembodied voice, anywhere, as we try to pin down the speaker's identity — and thus radically reduce that voice's individuality to conform to or be rejected by our expectations. Eidsheim calls “Who is this?” the acousmatic question, after Pierre Schaeffer, who “derive[s] the … root [of acousmatic] from an ancient Greek legend about Pythagoras's disciples listening to him through a curtain”, and she argues that it relies on fundamental misunderstandings of the human voice and our own listening practices, particularly in regard to vocal timbre. 9

Accordingly, she argues that the acousmatic question is largely impossible to answer, and that it is, in fact, the wrong question to ask. She offers three correctives to it: 1) “Voice is not singular; it is collective”; 2) “Voice is not innate; it is cultural” and 3) “Voice's source is not the singer; it is the listener” [Eidsheim 2019, 9]. That is, everyone is trained throughout their lives, whenever they vocalize and someone makes comments about their voice, to then adjust it to match the culture's expectations of how they should sound, in Foucauldian fashion — “as a condition of participation in a culture”. Scholarship on poetry readings and performative speech too often tends toward confident judgments about performance styles, with little acknowledgment of the complexity of listening, or of the influence of the listener's own aesthetic, cultural and ideological biases and preferences in regard to performance style.[3] 10

If we are always performing for and listening to others listening to us, then we never simply use our natural voices, because there is no such thing as a natural voice unaffected by cultural training. Nor is there such a thing as a neutral, singular, objective listener. Yet scholarship on poetry readings and performative speech tends toward confident judgments about performance styles, with inadequate acknowledgment of the complexity of listening, or of the influence of the listener's own aesthetic, cultural and ideological biases and preferences in regard to performance style.[4] 11

As Eidsheim paraphrases James Baldwin, one is always “hearing one's voice through the ears of others” [Eidsheim 2019, 269]. And we are always listening to others' voices through the ears of others, wondering: would others agree with my judgements about this voice? We listen to other voices through the ears of others; just as we never simply use our natural voices because there is no such thing as a natural voice unaffected by cultural training, nor do we listen in a way that is free of cultural training. There is no such a thing as a neutral, singular, objective listener. What we can do, following Eidsheim, is acknowledge the role of the larger culture in co-creating our voices and listening habits, the ways we speak and the ways we hear one another in different contexts. And we can bring another digital perspective into our listening practices — that is, the quantitative data perspective of audio signal processing, and the visualization of that data. 12

What follows is a critical narrative about the tools we have used and helped develop in our research, adding that digital perspective, and some highlights of the resulting research and its implications for literary study. These digital tools can enable humanist scholars to practice *slow listening* in the study of performative speech, to refine and test our impressionistic understandings of a given recording. Slow listening involves 1) repeated listening, with traditional close reading of the text; 2) scrutiny of our listening habits, assumptions, biases, and expectations, often based on the author's apparent identity and our own listening histories; 3) quantification of sonic patterns, such as pitch, timing, and intensity, as physical phenomena; and 4) sound visualization, specifically pitch contours (which simply show intonation patterns, or the rise and fall or steadiness of pitch over time), and timing patterns (including pause duration and speaking rate), and intensity patterns (volume or amplitude). 13

In this approach, we make an analogy to the practice among some musicians of slowing down a musical recording — say, a piano or saxophone piece with a very fast tempo — to understand and practice a technique, often by transcribing it. (Thanks to Alexander Ullman for alerting us to the term for this practice, *slowdown(ing)*.) This is not precisely what we 14 are doing, but performative speech, like music, rewards scrutiny with insight. Such scrutiny can illuminate (there's the hegemony of the visual for you) our listening habits and perceptions, and trends in particular genres of performative speech over time. This is particularly true when we find ourselves questioning our perceptions and impressions, and when sound visualizations and data — about, say, pitch, timing and intensity, derived from signal processing of a speech recording — do not neatly match up with our impressionistic perceptions of a recording, or a body of recordings, or simply push us to refine or describe our impressions more precisely.

3.

15 Our collaboration in testing and developing tools for critical voice studies emerged from an effort to visualize and quantify what is popularly known as Poet Voice. We aimed to confirm whether it exists, and if it does, to capture it in the wild, and at a larger scale; that is, to find examples of it not only by listening with our own ears, but by searching for particular pitch and timing patterns in poetry recordings that we and other listeners might agree sound like Poet Voice.

16 "*I know it when I hear it*", many listeners might say confidently about Poet Voice, echoing Supreme Court Justice Potter Stewart's insistence that, although he could not define hard-core pornography, he could recognize it on sight [Jacobellis v. Ohio]. In the case of pornography or obscenity, we know that cultural background and personal experience influence what viewers find offensive. But what about a poetry performance style that no one seems to like, which no one wants to admit to using, and yet seems to be so widely employed that any poetry lover — or NPR listener — feels they can recognize it?

17 Early on, without any digital tools to augment or refine our perceptions, one of us defined Poet Voice, or monotonous incantation, as speech employing a repeated cadence within a narrow range of pitch, and imposing that cadence on each line of a poem, regardless of the mood or semantics of the words spoken [MacArthur 2016]. A few key points about speech perception: *pitch and timing* are fundamental to the perception of tone of voice. Is a speaker's voice especially high or low? How much do they vary their pitch, and for what apparent semantic or affective reasons? Do they speak quickly? Slowly? At a moderate, predictable pace? Do they pause often? Briefly? Long and awkwardly? Suspensefully? And so on.

18 We began our collaborative research with a few samples of Poet Voice by Natasha Trethewey

Error

and Louise Glück

Error

— two poets who are often mentioned as using it — and, by considering only measures in pitch and timing discussed below, found other poets apparently using it as well, in a study of sample recordings of 100 American poets [MacArthur et al. 2018]. These other poets included Cecilia Llompart reading "Omens"

Error

and Matthew Zapruder reading "When It's Sunny They Push the Button".

Error

In the case of Llompart, we were not familiar with her work before we sampled it for the study. In the case of Zapruder, whom we count as a pleasant acquaintance, we were a bit chagrined to learn that, according to the data, he seemed to use Poet Voice in this recording. Speaking impressionistically, Zapruder's reading voice has some affinities with the likes of singer Bill Callahan, whose voice music critic Amanda Petrusich describes as "a heavy baritone that is somehow both entirely affectless and drenched in feeling" [Petrusich 2019, 88]. Why is it that we might appreciate low affect in male poets or singers, perceived as such through a narrow pitch range, yet readily find more pejorative examples of Poet Voice used by female than male poets?

19 Might this suggest an element of unreflective sexism in our own and others' perceptions of Poet Voice? Are we more

likely to use Poet Voice as a critical term about women poets? Do women poets use it more often than men? Or do we expect men to sound low-affect, flat or incantatory (think of William Butler Yeats and Allen Ginsberg), and don't expect women to sound high-affect, if that's what Poet Voice even is? Few listeners would agree on whether a given sample of apparent Poet Voice displays restrained affect, or false affect, or no affect. Perhaps we just don't want to believe that our friends might use Poet Voice, because we are used to hearing it as a critical term, if not an insult?

20 Interestingly, in a popular rant against Poet Voice, poet Rich Smith complains about contemporary poets using Poet Voice (including Glück, Trethewey and Gregory Orr), but claims that “Poet Voice is an effective and affecting style” when practiced by, for instance, Yeats [Smith 2016].

Live Broadcast

Ginsberg also uses a very monotonous speaking style in “Howl”, yet, as with Yeats, the effect seems to us more sermonic than Poet Voice.

Error

21 We encourage readers to take a moment to listen to the samples linked above of Glück, Trethewey, Llompart, Zapruder, Yeats, and Ginsberg to discover whether they agree that the poets are using a similar speaking style. For contrast to the above samples of Poet Voice, we also provide a link to a sample from Rae Armatrout, a conversational poet who uses contrastive pitch, less regular rhythm, and a wider pitch range.

Error

22 In using digital tools to analyze and visualize acoustic signals, we treat neither the tools nor the resulting data as supremely objective and corrective, but rather use the perspective they provide to complement and sometimes refine or check our own, inevitably idiosyncratic, aural perceptions and impressions. Beyond Poet Voice, we were, and are, interested in exploring trends and changes in performance styles over time, and changing tastes in performance styles, not only in poetry recordings but other genres of performative speech, such as political addresses, sermons, and radio broadcasts.

23 To visualize, interpret, describe and quantify patterns in pitch and timing for our research on poetry performance, we wanted one or more open-source, user-friendly tools that could easily and accurately visualize pitch contours, or intonation patterns, with the text aligned beneath the contours. We wanted our tools to be user-friendly for the humanist, not requiring much, if any, programming skills to use. Yet we also want to enable users to apply them with some basic knowledge about speech production and perception, and about signal processing. Additionally, we want these tools to provide numerical values measuring pitch (in Hertz) and timing (in seconds) for further analysis, and possibly other data about the voice, including intensity, breathiness, nasality, etc.

24 There is a long history to the hope that, by tracking pitch, we might measure and visualize vocal patterns of expression. In 1853, Edouard-Leon Scott de Martinville began creating what he called phonautograms, “a visual record of the pitch of someone's voice, how loudly he spoke, and with what emphasis”. He felt that this represented “our living speech”, our essential character even, and he “regarded Edison's invention of the phonograph in 1877, as pointless, because it reproduced sound instead of writing it” [Wilkinson 2014].

25 As we have written elsewhere [MacArthur and Miller 2018], pitch has long been neglected in the study of prosody in poetry, despite much interest in its role among prominent poetry scholars such as I.A. Richards, who wrote in 1926 that “[a] more serious omission is the neglect by the majority of metrists of the pitch relations of syllables ... that a rise and fall of pitch is involved in metre and is as much a part of the poet's technique as any other feature of verse, as much under his control also, is indisputable” [Richards 1926, 2001] .

26 This neglect in 1926 was understandable, due to the lack of tools for pitch analysis. Today, however, there are many commercial and open-source software packages for tracking pitch and performing other speech analysis. For qualitative and quantitative analysis in linguistics, available tools for analyzing pitch and timing variables are called *pitch-trackers*

and *forced aligners*. A pitch-tracker samples a speaker's pitch, called the fundamental frequency, or f0, at certain intervals, e.g. every 10 milliseconds. A forced aligner takes a transcript of a speech recording and aligns it with the recording, delivering timing information.

27 Figures 2 and 3 show the pitch contour for Louise Glück reading a few lines from “The Wild Iris” [Glück 1992] in *Drift*[5], the open-source pitch-tracker whose development we have overseen, now hosted by SpokenWeb at Concordia University. Figures 4 and 5 show word and pause length data in CSVs for the same recording from Gentle, an open-source forced aligner that Drift works with, and the fundamental frequency for every ten milliseconds in Hertz for two words, “It is”, from *Drift* (Both tools are described in detail in section 4).

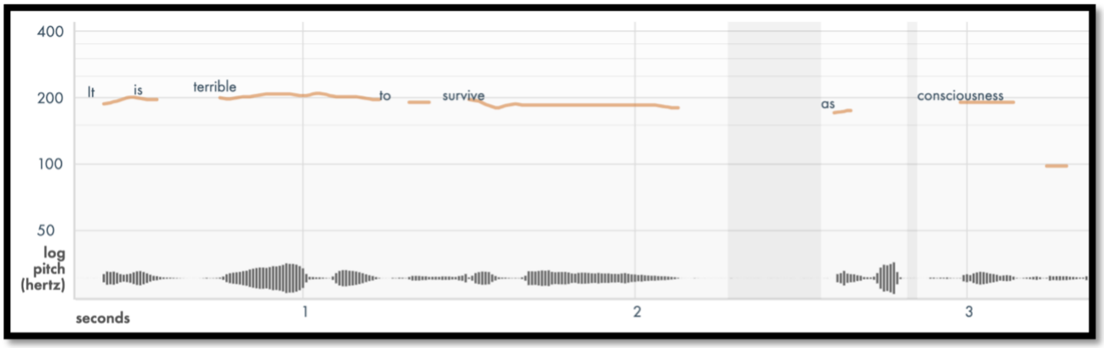


Figure 2. Pitch contour in Drift. Louise Glück reading “The Wild Iris” (1992)

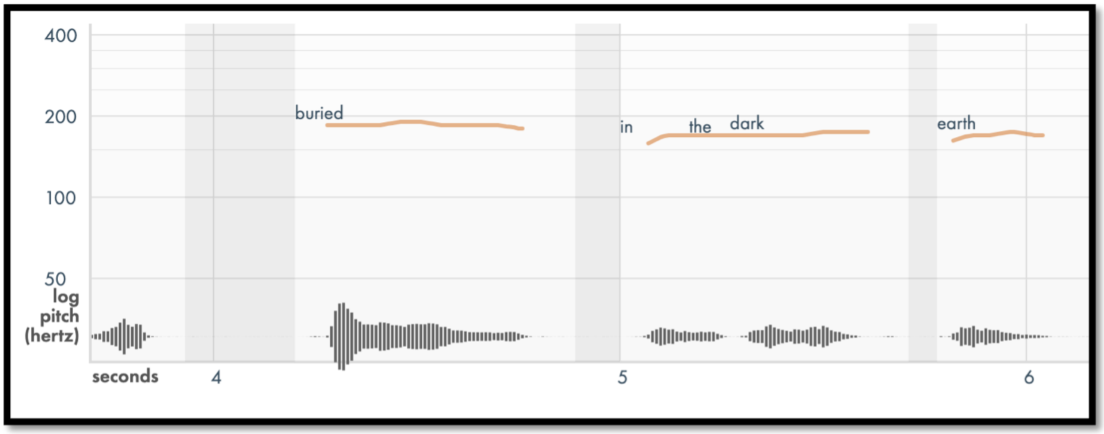
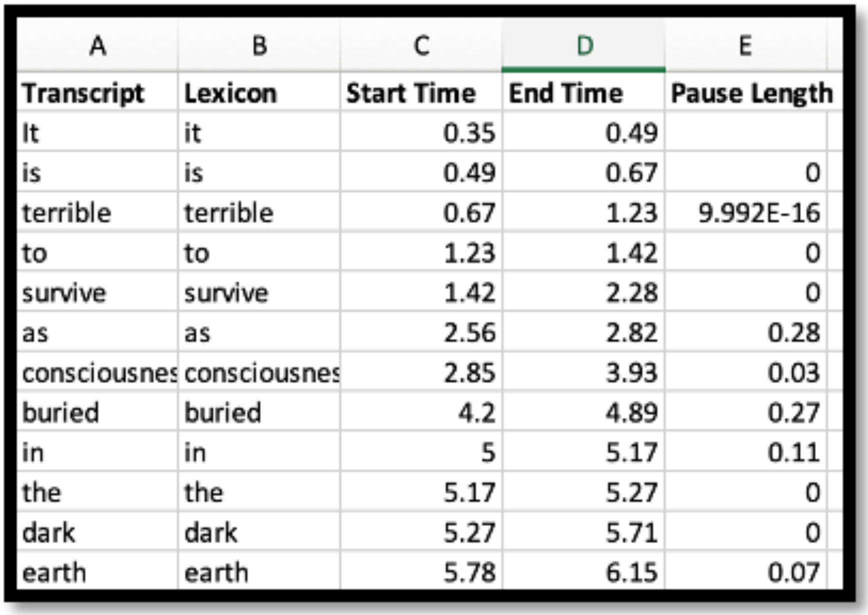
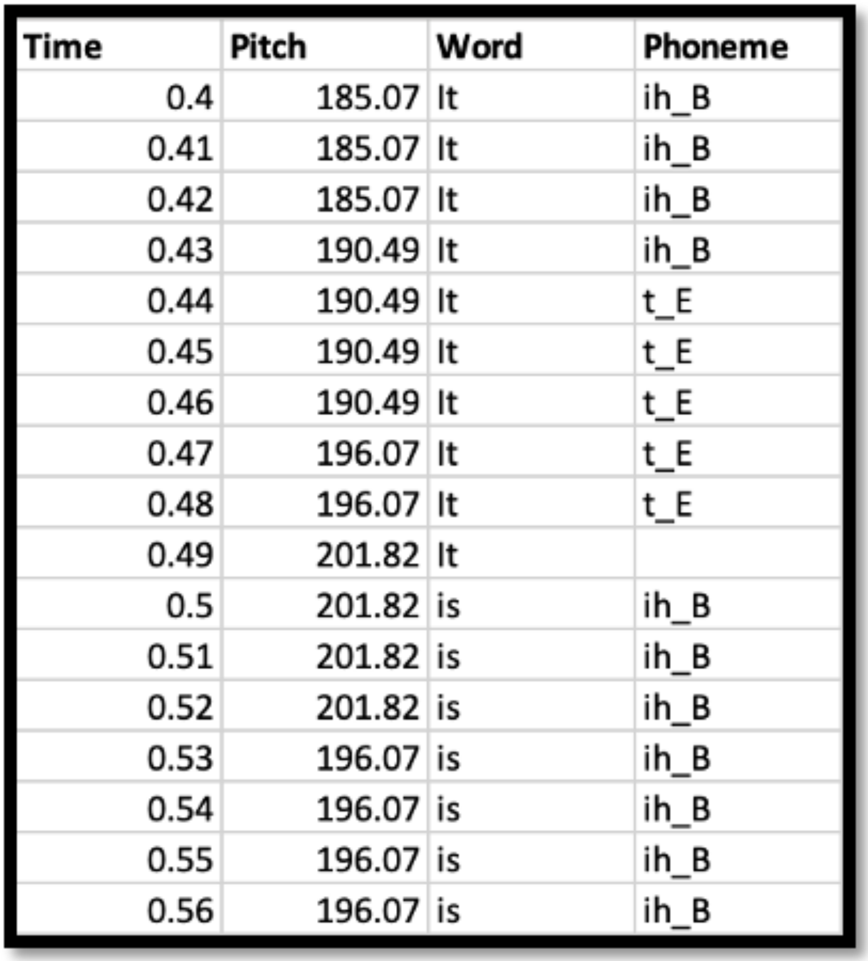


Figure 3. Pitch contour in Drift. Louise Glück reading “The Wild Iris” (1992)



A	B	C	D	E
Transcript	Lexicon	Start Time	End Time	Pause Length
It	it	0.35	0.49	
is	is	0.49	0.67	0
terrible	terrible	0.67	1.23	9.992E-16
to	to	1.23	1.42	0
survive	survive	1.42	2.28	0
as	as	2.56	2.82	0.28
consciousnes	consciousnes	2.85	3.93	0.03
buried	buried	4.2	4.89	0.27
in	in	5	5.17	0.11
the	the	5.17	5.27	0
dark	dark	5.27	5.71	0
earth	earth	5.78	6.15	0.07

Figure 4. Gentle data: Pause start and end times and pause length. Louise Glück reading "The Wild Iris" (1992)



Time	Pitch	Word	Phoneme
0.4	185.07	It	ih_B
0.41	185.07	It	ih_B
0.42	185.07	It	ih_B
0.43	190.49	It	ih_B
0.44	190.49	It	t_E
0.45	190.49	It	t_E
0.46	190.49	It	t_E
0.47	196.07	It	t_E
0.48	196.07	It	t_E
0.49	201.82	It	
0.5	201.82	is	ih_B
0.51	201.82	is	ih_B
0.52	201.82	is	ih_B
0.53	196.07	is	ih_B
0.54	196.07	is	ih_B
0.55	196.07	is	ih_B
0.56	196.07	is	ih_B

Figure 5. Drift data: Fundamental frequency every 10 ms in hertz. Louise Glück reading "The Wild Iris" (1992)

The pitch contour of Glück's voice shows the fundamental frequency, the vibration rate of the vocal cords in speech, keeping to a narrow pitch range, starting off at 185 Hz, rising to 207 with “terrible”, and so on; the Gentle data also shows that Glück pauses for almost a third of a second after “survive” and “buried”, lingering to emphasize the suffering involved.[6] 28

The harmonic frequencies of the fundamental frequency, which follow the formula f1 = 2 x f0, f2 = 3 x f0, etc. as they resonate through the vocal tract, give a voice its unique timbre and distinguish different vowels regardless of the pitch.[7] The following spectrograph in Matlab illustrates the fundamental frequency of Glü;ck's voice saying “It is terrible”; again, the fundamental frequency is around 200 Hz; the second harmonic (f1) is at 400 Hz, the third (f2) at 600 Hz, and so on, all the way up the seventeenth harmonic at 3,600 Hz. 29

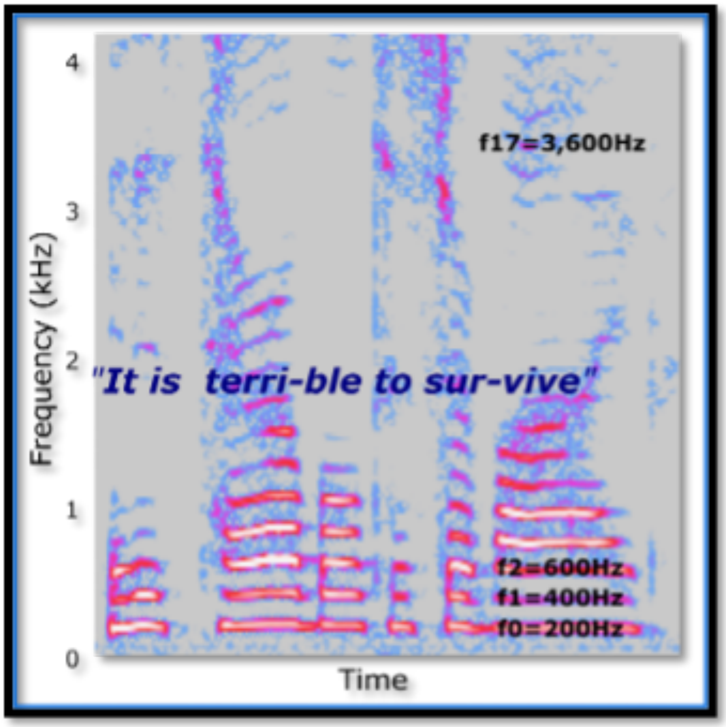


Figure 6. Spectrogram of the fundamental frequency (f0) and multiple harmonics of the fundamental in hertz (f1, f2, etc). Louise Glück reading “The Wild Iris” (1992)

As a rule, more expressive speakers use a relatively wider range of pitch. However, it is also important to note that “pitch perception is not linear”; an octave is a logarithmic relation, twice or half a given frequency [Lehiste 1970]. For instance, an octave above 120 Hz is 240 Hz; an octave below is 60 Hz. This means we do not perceive women as having a wider vocal range than men, although they use a wider range of pitch as measured in Hertz. The same holds for notes on a piano: the octave between C3 and C4 uses a wider pitch range than between C2 and C3, but we hear it as the same interval. This makes interpreting linear graphs of pitch a challenge, and thus we display the more perceptually relevant logged pitch values, as with the pitch contour (in *Drift*, above) of Glück's voice. 30

In analyzing the pitch of human speech, we could in theory base our approach on articulation, perception, or acoustics. Measuring articulation — especially, for pitch, the vibration of the vocal folds, but also the movements of the tongue, lips, jaw, etc. — is not usually possible. As humanists, we often work from recorded speech audio, and/or lack the requisite instrumentation to objectively measure vocal fold vibration. Measuring perception would be impractical, even though pitch is perceptual. As phonologist Carlos Gussenhoven explains in *Phonology of Tone and Intonation*: “Unfortunately, listeners lack the appropriate conceptualizations and vocabulary to report their sensations, and are typically incapable of saying even whether a given pitch change represents a fall or rise”[Gussenhoven 2004]. Since individual perception is largely a “black box” process, it may be impossible to say *why* different individuals report pitch differently, or why they make certain errors, e.g. whether pitch is rising or falling. 31

Moreover, the fuzziness and subjectivity of perceived pitch can be an impediment to research because of individual and cultural differences. Our descriptions of pitch, not to mention what pitch range we feel is appropriate for a given speaker in a given context, varies according to culture, among other variables — e.g., the same woman might speak at a higher average pitch and use a wider pitch range in her native Japanese than in American English. Dutch speakers describe pitch as high or low pitch, whereas the same vocal quality is characterized as "thin" or "thick" in Farsi [Dolscheid et al. 2013]. And so on. 32

Consequently, the best and most established approach to pitch estimation relies on computational analysis of speech acoustics. A pitch tracker such as we use (Drift) samples the fundamental frequency of a recorded voice — Drift samples it every 10 milliseconds — and is designed to avoid errors as much as technically possible, even with noisy recordings. We strongly advocate this computational approach to speech acoustics because 1) the algorithms approximate perception, in that they rely on similar acoustical features as human listeners do when estimating pitch; 2) computational approaches developed, practiced and vetted by a large number of expert researchers in acoustic analysis are comparatively objective in the sense of being unaffected by a researcher's personal biases about pitch, or by codifying biases explicitly; 3) the techniques can be used on arbitrarily large datasets with minimal additional resources; 4) they are usually reproducible, so results can be replicated within or across research groups, a matter of increasing importance in any field aspiring to rigorous empirical scholarship; and 5) the algorithms themselves can be shared openly, compared, and queried when they yield errors or fail, which they inevitably do sometimes with real data. 33

Despite the sophistication of contemporary automatic pitch trackers, errors arise, partly from the limits of the algorithms but partly as a result of the inherent qualities of the human voice as a complex and dynamic time-frequency signal, as the figure showing the harmonics of Glü;ck's voice suggests. To begin with, "voiceless friction [can be mistaken] for voicing, and [so] irregular measurement points may show up during fricatives and releases of affricated plosives. Conversely, [a pitch tracking program] may interpret irregularly voiced signals as voiceless" [Gussenhoven 2004]. In non-technical terms, such errors are caused by some unvoiced consonants being incorrectly marked as voicing pitch, and in other cases the pitch value of vowels being missed. If an algorithm makes these errors, they have to be detected by looking at the data, or the visualization of that data, or filtering the data. Thus, robustness to additive noise is an important goal for research on the voice. Not only that, but a 34

> pitch tracker may fail to analyze the voiced signal correctly. When the voice becomes creaky, as it often does at lower pitches, the algorithm may become confused by peaks in the signal that do not correspond to the vibratory action of the vocal folds. If these appear halfway through the period ... they may lead to ... doubling errors… Similarly, the algorithm may miss every second periodicity peak, "believing"; these peaks determine the sound quality rather than the periodicity (halving errors). Such "octave jumps" are usually easy to detect: the pitch track shows a sudden change to a value half or double that of the immediately preceding value, while there is no auditory impression that corresponds to that jump.

In the pitch contour above of Glü;ck reading "consciousness", for instance, there is an octave jump — a halving error — beneath the syllable "ness". Octave jumps and noise errors occur more frequently in noisy recordings, of course, and with programs not optimized for noise. While common in pitch estimation, octave jumps can be addressed with heuristics that constrain the probable pitch range, for instance by gender, or attempt to repair errors posthoc without affecting the basic estimation algorithm, especially with recent advances in pitch-tracking. 35

Many different pitch algorithms have been applied to speech audio, all of which ultimately try to estimate the vocal fold rate. Some deal mainly with the speech as a signal in time, whereas others emphasize frequency representations or first filter speech into different frequency bands before working on it in time. Still others use so-called cepstral analysis, a hybrid time-frequency approach which leverages the fact that in speech production, whatever the vocal folds are doing (mainly creating pitch) is acoustically dissociable from what the rest of the vocal tract is doing (e.g. making vowels and syllables). Today's most frequently used pitch estimation methods, and their accuracy in estimating male and female pitch in clean speech, are discussed in Strömbergsson [Strömbergsson 2016], pitch-tracking performance on noisy speech is discussed in Jouvet and Laprie [Jouvet and Laprie 2017]. 36

In our view, the choice of pitch tracker is less a principled decision and more a pragmatic one, with an eye toward ease of use, having open-source (readable) code, noise-robustness, a low error rate with clean and noisy speech, and ease of integration within a larger workflow. Gussenhoven puts it more succinctly: “In practice, the choice of pitch tracker is determined by circumstance and convenience: their evaluation is typically beyond the competence of a phonologist” [Gussenhoven 2004]. While we agree with the basic premise, we do insist that researchers using any pitch tracker should have a basic conceptual understanding of how it works “under the hood”, so they are cognizant of its strengths and limitations. 37

In addition to errors that follow from the natural variance and complexity of speech, a potentially greater challenge for pitch trackers is additive noise and, more generally, recording quality. In our work, it quickly became clear, in fact, that tools commonly used by linguists to analyze both pitch and timing patterns in so-called natural speech, recorded in the ideal acoustic conditions of a sound booth, would not work well for research on recordings of poetry readings, which can be noisy, either due to the recording equipment used, background noise at a venue of a poetry reading or, even in the case of highly produced studio recordings, due to age. 38

Praat is a tool both of us have used, to begin research visualizing intonation patterns (MacArthur in 2012 and 2013, prior to our collaboration) and to manipulate pitch for experiments addressing how the brain represents and attends to speech (Miller, since 2009). An open-source program for speech analysis developed in C and C++ at the University of Amsterdam by the phonetic scientists Paul Boersma and David Weenink, Praat was initially released in 2003 and is periodically updated. It uses a time-domain pitch-tracking technique developed by Boersma, which showed great improvement over earlier frequency-domain techniques. Praat identifies the best pitch candidate at any time frame (say, every 10 ms) by finding peaks in the autocorrelation function, which all periodic sounds (including voiced speech) will show. 39

Among its innovations were to deal gracefully with two side effects of typical signal processing that reduce the accuracy of the pitch estimates (specifically correcting for windowing into time frames and sampling). Praat then traces an estimated pitch “path” across time to determine the final pitch contour, incorporating a “cost” to minimize excessive voiced-unvoiced transitions and octave jumps. However, Praat is not very user-friendly and has very limited and unattractive data visualization options, and it is not easy to extract the data about pitch and timing from Praat. Linguists who use it write and share scripts to run particular tasks that cannot be performed using the basic Praat interface. For humanists who cannot write code, or simply do not want to spend their time this way, the limitations of Praat quickly become apparent. Praat also provides no simple way to align text beneath speech — that is, to line up the intonation pattern of a line of a poem with the text. 40

A forced aligner that can be used with Praat, called FAVE (Forced Alignment and Vowel Extraction), does not work well with incomplete or flawed transcripts — a problem for large-scale analysis of audio archives for which transcripts may not be readily available, like a poetry archive such as PennSound [Rosenfelder et al 2014]. Like other forced aligners, FAVE incorporates a “dictionary” of possible words and an acoustic model of speech, in this case of American English, trained on hours of labeled recordings. For the acoustics, it does not represent speech as a complex sound pressure waveform, but rather as much lower-dimensional Perceptual Linear Prediction coefficients, which use a perceptually-inspired front end and in many cases a cepstral analysis to extract speech-relevant (e.g. vowel formant) parameters. The acoustic model in FAVE was built with the HTK Speech Recognition Toolkit, which uses Hidden Markov Models to determine the probabilities of transitions from one state–in this case speech sound or phoneme–to the next. The aligner is therefore mapping a transcript onto best-guess acoustic sequences, giving the timing of all (identified) words and phones. 41

In the midst of frustrations with Praat in 2013, MacArthur joined as a user-tester the NEH Institute HiPSTAS (High-Performing Sound Technologies for Access and Scholarship), directed by Tanya Clement, and tried out the program ARLO (Adaptive Recognition with Layered Optimization) to study pitch patterns in poetry recordings.[8] A machine-learning speech analysis and visualization tool originally intended to analyze birdsong, ARLO has significant limitations for speech analysis: it displays non-standard spectrograms with unusual time compression, making it hard to perceive speech features, and it requires sizable supercomputing resources to run. As well, many parameters require 42 considerable manipulation to produce accurate pitch values, and pitch data is difficult to extract. The goal of “[d]eveloping ARLO as a web-based application that can be leveraged by a wide community of users” was abandoned [Clement et al. 2016]. However, HiPSTAS was invaluable in bringing together communities of scholars interested in studying voice recordings, such as poetry readings and oral histories, and it supported novel research.

4.

The potential for digital voice studies research advanced considerably in 2015, when we began collaborating and also met Robert Ochshorn, who at the time was a media-interface researcher in San Francisco at Communication Design Group Labs, conceived by Alan Kay.[9] If the choice of tools is indeed “determined by circumstance and convenience”, it can help to connect with leading-edge researchers from Silicon Valley. Ochshorn and Max Hawkins had just designed and released Gentle in 2015. 43

Gentle[10] is built on top of an optimal open-source speech recognition toolkit, Kaldi, initially developed at Johns Hopkins University in 2009 [Povey et al. 2009]. Gentle “take[s] media files and their transcripts and return[s] extremely precise timing information for each word (and phoneme) in the media”. It was designed specifically to function with more flexibility than FAVE, to be “easier to install and use … handle noisy and complicated audio … and … be unusually accommodating to long or incomplete transcripts” [Ochshorn and Hawkins 2016] (It can also produce rough transcripts of the spoken text of a recording from scratch — the degree of accuracy varies according to recording quality, and the errors are entertaining — which can then be corrected and aligned with the recording.) CSVs with the start and end time of each word can be downloaded from Gentle as well; see above sample for the Glü;ck recording. From this data, it's a simple matter to calculate word and pause duration, as well as Words Per Minute, a standard linguistic measure of speaking rate. 44

Through Ochshorn, we also learned of a pitch-tracking algorithm developed in 2012, better suited to the audio archives humanists study: *Noise Robust Pitch Tracking by Subband Autocorrelation Classification*, by computer scientist Byung Suk Lee and electrical engineer Daniel P. W. Ellis. The Lee-Ellis algorithm was “[t]rain[ed] on bandlimited and noisy speech (processed to simulate a low-quality radio channel) [and] leads to a great increase in performance over state-of-the-art algorithms, according to both the traditional GPE measure, and a proposed novel Pitch Tracking Error which more fully reflects the accuracy of both pitch extraction and voicing detection in a single measure” [Lee and Ellis 2012] 45

Ochshorn and Max Hawkins then developed a prototype user-interface for Drift, which combines the Lee-Ellis algorithm for pitch-tracking with Gentle's forced alignment features to visualize text beneath the pitch contour of a given speech recording, showing which words are spoken as the voice rises and falls, remains at the same pitch, etc. (See above pitch contours for Trethewey, Gluck, Llompart, Zapruder and Yeats.) A CSV with time in seconds, pitch in Hertz, words and phonemes can be downloaded from Drift (see figure above). Drift also has potential applications for the hearing impaired, providing a clear visualization of intonation patterns that might complement closed captioning. In addition to the online demos, both Gentle and Drift are available for free download and installation on Macs. 46

The development and dissemination of Gentle and Drift were supported in 2018 and 2019 by a NEH Digital Humanities Advancement grant project, “Tools for Listening to Text-in-Performance”,[11] and continue to be supported by a 7-year Can$2.5 million SSHRC grant project, *The Spoken Web: Conceiving and creating a nationally networked archive of literary recordings for research and teaching*. [12] With considerable improvements to both the interface and functionality, thanks to our talented undergraduate research assistants at UC Davis, Richard Kim, Hannan Walliullah and especially Sarah Yuniar, Drift4[13] was released on November 14, 2022, and is hosted by the SSHRC-funded SpokenWeb project for the long term. 47

In parallel with Ochshorn's development of Gentle and Drift, and also with support from the NEH and SSHRC grants, we have also built a toolbox in Matlab called Voxit, to enable greater ease of data exploration and in-depth analysis (or manipulation) of vocal characteristics. Voxit is a sort of feature-rich, developmental sandbox for large-scale quantitative analysis to complement the simpler and user-friendly Drift and Gentle, which are best for qualitative, visual and small-scale quantitative analysis. The main engines for analyzing speech in Voxit are i) the same Lee-Ellis algorithm for pitch 48 tracking that Drift uses (Subband Autocorrelation Classification), and ii) WORLD, a vocoder-based, high-quality speech analysis, manipulation, and synthesis program developed by Masanori Morise (who also worked with Hideki Kawahara on a predecessor to WORLD, called TANDEM-STRAIGHT). WORLD, which is available in C++, Python and Matlab, estimates speech parameters such as voiced-unvoiced distinctions and pitch, with the particular goal of producing high-quality resynthesized speech with low computational cost.

Voxit uses the Lee-Ellis pitch and other vocal/acoustic properties encoded by WORLD to provide vocal prosodic measures that have perceptual importance for speaking style, context, etc., such as *pitch range (in octaves)*, *pitch speed (in octaves per second)*, *pitch acceleration (in octaves per second squared)*, and *rhythmic complexity of pauses* [MacArthur et al. 2018] [MacArthur and Miller 2018], and more recently, *intensity range* and *intensity speed (in decibels)*.[14] (Drift4 incorporates all of the Voxit prosodic measures as well.) While Voxit does not feature forced alignment internally, it can import the word- and phoneme-level alignment produced by Gentle and derived measures based on the alignment alone, or on the alignment in relation to the other derived speech parameters. It thus offers a powerful and readily extensible toolkit for speech-acoustic-linguistic analysis for humanistic research, particularly for large datasets. 49

To better understand how the two approaches can work, and work together, we encourage users to explore a number of our publications, including “Beyond Poet Voice: Sampling the (Non-)Performance Styles of 100 American Poets”[15] which illustrates the applications of the Voxit toolbox with 100 sample recordings; “After Scansion: Visualizing, Deforming and Listening to Poetic Prosody”[16], which walks users through applications of Drift and Gentle with 15 sample poetry recordings; “John Ashbery’s Reading Voice”[17], which analyzes recordings from the 92nd Street Y over five decades; and “101 Black Women Poets in Mainly White and Mainly Black Rooms”,[18] which further analyzes 203 recordings. 50

To return to Poet Voice, when we analyzed sample recordings of 100 American poets, we found that other pitch and timing patterns — in addition to a slow pace and narrow pitch range — are also important in the perception of this style of reading poetry [MacArthur et al. 2018]. 51

Voices that sound, at least to us, like they are using Poet Voice tend to read relatively slowly, with a predictable rhythm, in terms of tempo and pauses, compared to poets we might hear as more expressive or conversational, which tend to exhibit less predictable, and thus more complex, rhythmic patterns. They also tend to change pitch slowly, as well as using a relatively narrow range of pitch, again compared to more expressive or conversational speakers. The following figures show some of these variables for recordings of these poems in their entirety: Natasha Trethewey's “Monument”, Louise Glück's “The Wild Iris”, Cecilia Llompart's “Omens”, Matthew Zapruder's “When It's Sunny They Push the Button”, William Butler Yeats's “The Lake Isle of Innisfree”, Rae Armantrout's “Heart of It”, and the first 2:45 of poets who do not sound, to us, as if they are using Poet Voice. 52

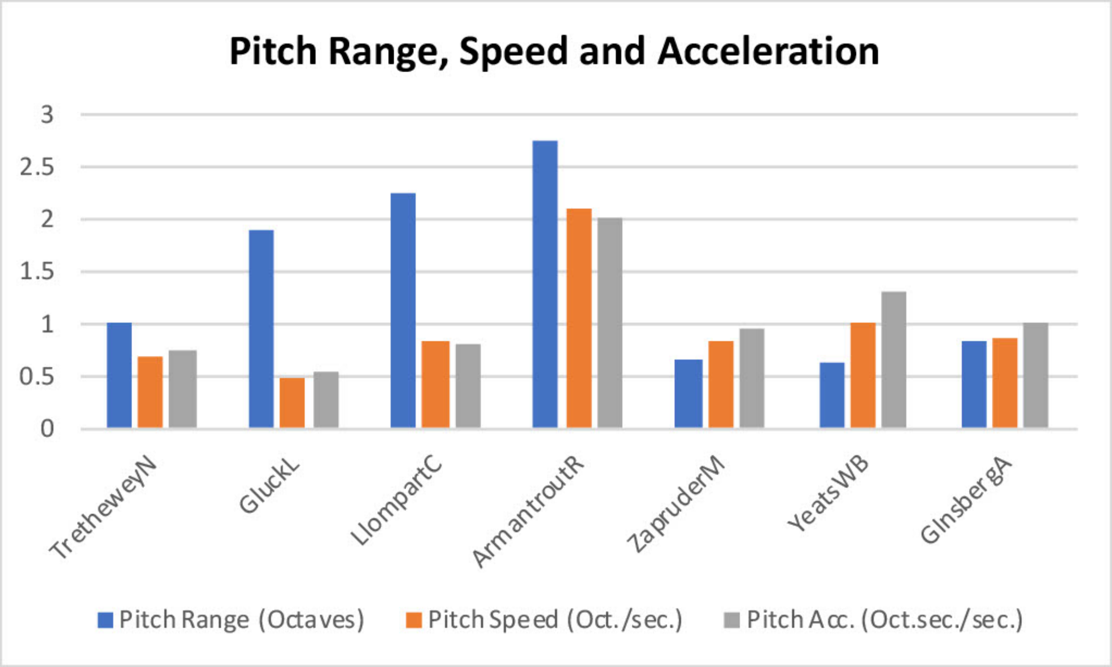


Figure 7. Pitch Range, Speed and Acceleration in Sample Recordings by Seven Poets

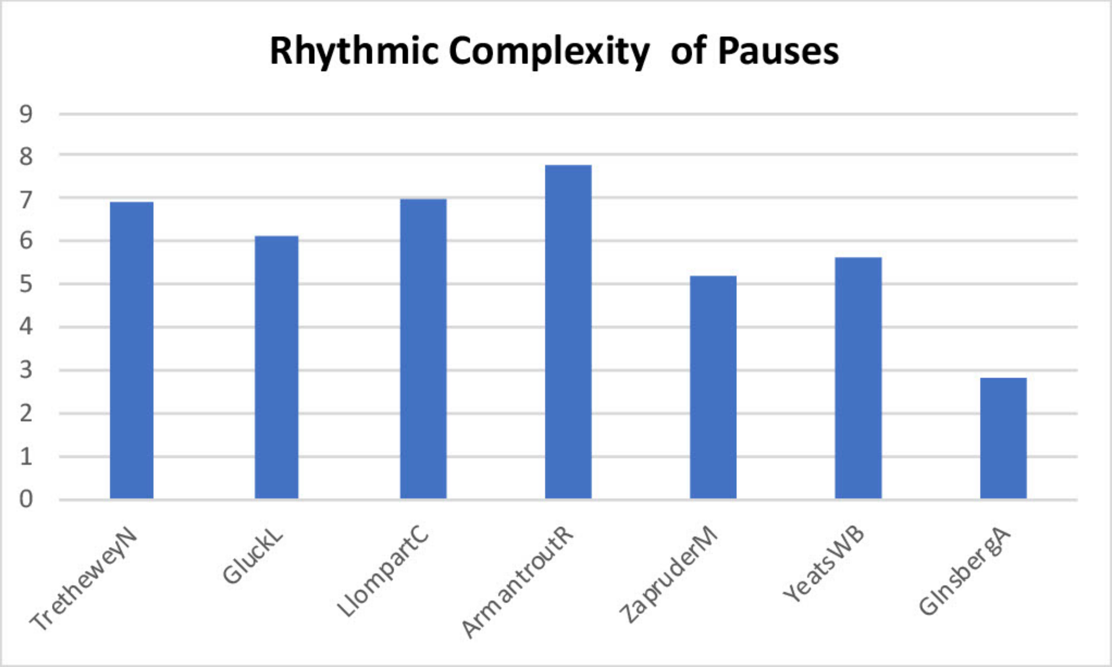


Figure 8. Rhythmic Complexity of Pauses in Sample Recordings by Seven Poets (a lower value means a more predictable rhythm)

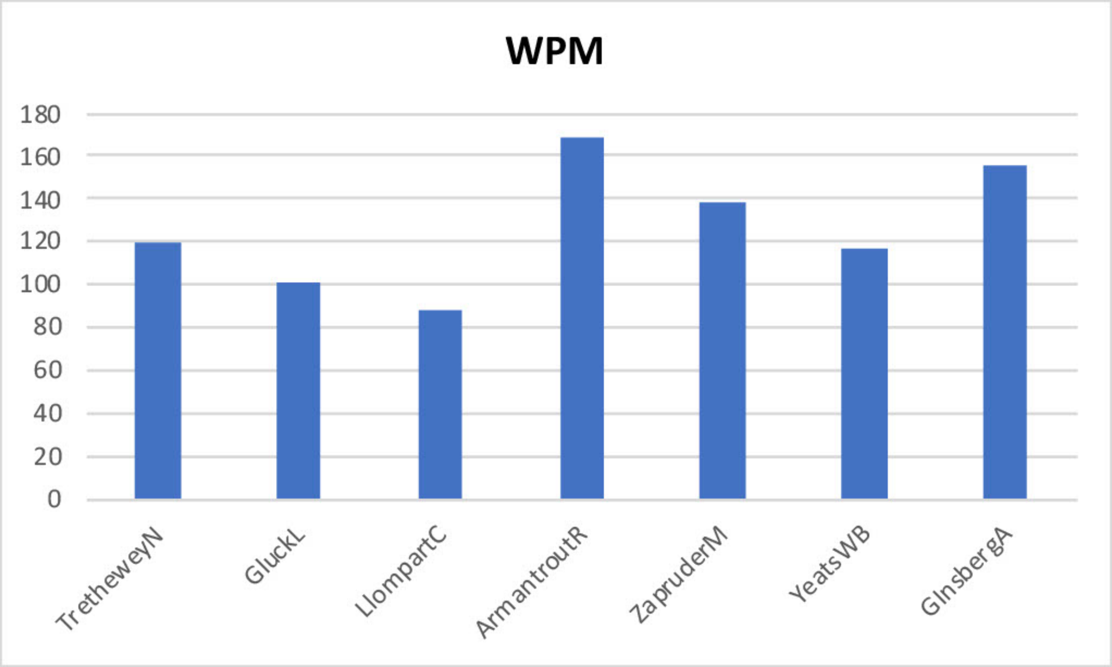


Figure 9. Words Per Minute in Sample Recordings by Seven Poets

From these few samples, we conclude that, when some listeners hear poets read with even just one or two of these characteristics — slow pitch speed, slow pitch acceleration, narrow pitch range, low rhythmic complexity, and/or slow speaking rate — they hear Poet Voice. [19] In other words, we listen selectively and use fuzzy inference, seeking at least one reason, or apparently objective vocal quality, to support or justify the way we respond to a reading style, whether we like it or not. But Poet Voice is not a monolithic manner of reading poetry, nor, to point out the obvious, is it innately irritating. Poet Voice is collectively constructed, at least in part, in the ears of listeners, as Eidsheim might say. 53

What is potentially transformative about slow listening research, as our work on Poet Voice suggests, is that the digital tools provide a nuanced empirical and quantitative perspective on the intuitions we have developed as, respectively, a poetry scholar, poet and amateur singer with years of experience attending poetry readings and listening to and teaching with poetry recordings (MacArthur), and as a neuroscientist of speech perception and amateur musician with more than a casual interest in vocal performance, in speech and song (Miller). This digital, empirical, quantitative perspective is not infallible nor objective, but it is a very useful complement to our very human perspectives. Both of us are fascinated by (and opinionated about) what makes for compelling performative speech, in teaching, in academic lectures and in U.S. culture at large. And yet we recognize — and these tools have helped us appreciate this point deeply further — that, as we test our own biases and refine our own intuitions in terms of speech perception, we need all the help we can get! 54

There is just no reason to rely any longer on older methods of impressionistic generalizing, without recourse to digital tools or computational analysis, about or from a small number of familiar recordings, which are not incorporated into the scholarship published about them, such that readers are not actually listening to the object of study. This has been the typical scholarly approach of close listening. We do not at all mean to disrespect the crucial work of close listening or of Bernstein, who has encouraged our research every step of the way. His former student Chris Mustazza, now co-director of PennSound, characterizes the new approaches we and others are using as “machine-aided close listening” [Mustazza 2018]. The rich intuitions of scholars form an excellent starting point for research on poetry performance, which can benefit enormously from these digital methods. Indeed, it was heartening when Bernstein wrote about our work in 2017: “You are certainly doing the work I had hoped might be done” . 55

And we are not alone. In recent years, a small but growing number of literary and humanities scholars have begun to apply digital sound studies tools to the analysis of speech recordings. Some of this work is led by other fields with more 56 training in the human voice, media history and/or digital tools. For instance, *The Oxford Handbook of Voice Studies* (2019) edited by Nina Eidsheim, which focuses mostly on voice in song, oral history and political speech, includes pieces by comparative literature scholar Tom McEnaney on the timbre of NPR voices, and film studies scholars Dan Wang on “cinematic speech”, and Jennifer Fleeger on robot speech in film. *The Sound Studies Reader* (2012) edited by Jonathan Sterne, also includes pieces on sound in film. The *Journal of Cultural Analytics* and *Digital Humanities Quarterly* welcome work on the study of sound in literary recordings, though both are understandably dominated by computational analysis of text and image. *The Journal of Interdisciplinary Voice Studies*, founded in 2016, also has potential as a place to host scholarship on literary recordings.

Drift and Gentle have been applied to many genres of performative speech, including and beyond poetry recordings. Our NEH project Tools for Listening to Text-in-Performance, co-directed by MacArthur and Neil Verma, involved a team of 20 user-testers on Drift and Gentle from across the U.S., representing the disciplines of communications, literary studies, film studies, performance studies, media and radio history, comparative literature, science and technology studies, theatre, and media preservation. Initial results illustrate both the potential of this research for all sorts of performative speech, and the way that tools can be applied in unpredictable ways, depending on the object of study and the scholar's interests. For instance, the research group at the University of Wisconsin-Madison used Drift to consider “different genres of podcasts and radio …. to analyze potential differences in tone and delivery”, and they used Gentle “to see if different genres of podcasts have different conventions of duration and WPM [Words Per Minute, or speaking rate]” (see [Mertens et al. 2021]. 57

In an application we did not anticipate, the Podcastre group also found it useful to use Gentle's ability to create rough transcripts to search for particular vocabulary in podcasts, as a tool to explore which podcasts to study more closely. Gretchen Jude, a performance studies scholar, has used both tools to study the “speed and flow of speech as well as hesitations on words, and placement and length of pauses” in second-language speakers of English, and Adam Hammond and Jonathan Dick have used both tools to analyze different speakers' performances of T.S. Eliot's *The Waste Land*, focusing on the question of “whether timing and pitch data from Gentle and Drift (and the outputs of the prosodic measures script) can reveal something to a human listener (very familiar with the recordings) that they didn't already know. Our suggestion was that they could, and our example was Eliot's reading of The Waste Land, which the prosodic measures position as one of the most internally varied of the readings we've heard, despite most [human] listeners' sense that he's among the most monotonous” [Hammond and Dick 2019]. 58

Media historian Patrick Feaster (co-founder of *FirstSounds.org*) has used Gentle and Drift to analyze recordings of “Cal Stewart … a well-known phonographic storyteller who recorded dozens of original comic spoken-word selections in the role of Uncle Josh Weathersby between 1897 and his death in 1919…. to compar[e] and contras[t] examples of his work in an effort to trace how his repertoire, style, and dynamic performance strategies evolved over time”. Jacob Smith, a media historian at Northwestern, has used Drift to analyze pitch variables in vaudeville ethnic stereotypes as performed on phonograph records from 1905 and 1911 by Ada Jones and Len Spencer, for an essay that appeared in *The Oxford Handbook of Cinematic Listening*, “The Courtships of Ada and Len: Mediated Musicals and Vocal Caricature Before the Cinema”. 59

Another signal of serious academic interest in the intersection of sound studies and literature is a special section of PMLA in March 2020 devoted to Aurality and Literacy, curated by Matthew Rubery, author of *The Untold Story of the Talking Book* (2016), and Christopher Cannon. Though it did not much involve directly applying digital tools to literary recordings, one particular piece in that issue articulates both the need for and some of the obstacles to this sort of research on literary audio. As James English argues in “Teaching the Novel in the Audio Age”, the success of the audiobook, “the market for [which] is dominated by literary works, and specifically by novels, which comprise nearly three quarters of all sales”, necessitates new pedagogies that incorporate audio versions of literary texts into teaching, as much as the texts themselves. For English, “[e]very vocal performance of a novel is an interpretation, a reading as well as a reading out loud” and “[n]one stands outside what [Matthew] Rubery calls the ‘politics of narration’”. Thus concerns arise, with the voices of different characters and narrators, about vocal stereotyping according to race, gender, sexuality, (dis)ability, etc. Yet he concludes that these concerns should not be a focus in teaching audio books: 60

> The novel is a dialogic form, and even those written in the first person (a distinct minority, by the reckoning of Ted Underwood and his collaborators) incorporate lots of reported speech. Except in the rare case of full-cast audio dramatizations, audiobooks require a vocal performer capable of rendering multiple voices. Conveying social distinctions between characters through faked accents and other tricks of the voice is a skill fundamental to performative work in the medium. Some degree of stereotyping is inescapable
>
> The issues around vocal stereotyping can of course be embraced as teachable controversies. But that shifts time and attention away from other interpretative concerns, other teaching that we might have in mind for a class. It also requires of us some new proficiencies with classroom technology. [English 2020, 422–423]

Those proficiencies, for English, mostly come down to preparing audio segments from audio books for listening and discussion, accepting the time commitment required, being prepared for the technical difficulties that may arise for both students and instructors — and then “contend[ing] with our discipline's entrenched anti-audio prejudices”. 61

The methods of slow listening, as we have outlined them above, can be usefully applied to audio books — not to reduce the analysis to finger-pointing about vocal stereotypes, but to help students and instructors alike to think more deeply about the politics of performance and identity, so central to both the dramatization and interpretation. A former undergraduate student of English, Alexander Ullman, a Ph.D. candidate at the University of California, Berkeley, is doing promising such work on audio books and Black/Jewish identity, using Drift and Gentle. Acknowledging in our pedagogy and scholarship that the aural/oral dimensions of literature matter as much as the text, and indeed directly affect the reception and interpretation of the text, extends a long and necessary movement away from New Critical assumptions about the text's primacy. In teaching and studying audio books and other recorded performances of literature, we can learn to register and articulate our own implicit aural biases and preconceptions — and also, perhaps to discover a greater freedom in the choices we make as interpreters and (sometimes reluctant) performers of literature in the classroom. 62

5.

User-friendliness and open-source are two concepts that continue to guide the development of Gentle, Drift, and Voxit. The justifications for open-source software are well-known and persuasive. Familiar as the term “user-friendly” is, it requires some unpacking. Ochshorn, the interface designer and software developer with whom we have worked closely on Drift and Gentle, is fond of relating an anecdote to illustrate the tension between designing tools that are easy to use, and tools that transparently present the complexity behind their processes and allow the user to interact with the underlying data. The anecdote is from historian David F. Noble's *Forces of Production: A Social History of Industrial Automation* [Noble 2011]. Noble writes: 63

> If the relationship between technical people and those who wield social power informs their designs and their estimates of what is possible, so too ... does their relationship with [users] ... Suppose … an engineer designed a machine for his best friend, for her birthday… saying, with true professional pride, “Happy birthday. I have built for you my finest machine; it is so well-designed, it can be run by an idiot”. No doubt his friend, who does not consider herself an idiot, would be taken aback…. and the engineer would be obliged to try to redesign the machine for someone who was not an idiot. This would be very difficult, given the orientation of his professional knowledge, and he might not even know where to begin. (Of course, he might start by pretending he was designing it for himself.)

The gender of the characters in this anecdote is not accidental, of course. A DH tool like Voyant, some might say, is designed to be run by an idiot — while others would insist that it enables an accomplished historian, for instance, who has spent years mastering their field of the U.S. Civil War, and who is a parent of small children, to engage in DH research in text mining without being obliged to learn to code in their non-existent spare time or go back to school (perhaps on a one-year Mellon New Directions Fellowship) to study computer science. Gentle and Drift, in their ease of use, follow a similar approach as Voyant. Although they feature simple interfaces, however, the data and visualizations 64 they provide require some knowledge of audio signal analysis and speech production and perception to interpret. Praat, in its user-unfriendliness, is obviously designed not for idiots but for patient linguists, for whom spending a time writing scripts to analyze recordings is a more legitimate use of their research time.

Noble continues the anecdote: 65

> However, had [the engineer] presented that same machine to a manufacturer, with the same claim — that it could be run by an idiot — he would probably have run into no such difficulty. Imbued with the philosophy of work simplification and deskilling, desirous of reducing his labor costs and minimizing his "labor problems" and, because of his rights as an employer, having the power to compel his workers to do idiot work, the manufacturer would probably have found his machine quite satisfactory. Indeed, it is his criteria, embedded in the engineer's "art," that shaped his design in the first place

We are not, in collaborating with Ochshorn on the development of Gentle and Drift, developing a product for sale, and the knowledge production that academics are engaged in is not, of course, idiot work. Rather than deskilling, we aim to draw on the different skills of researchers from different disciplines, as collaborative, cross-disciplinary work allows. 66

There is considerable value, as Ochshorn suggests in "Intimacy and Interface" (a survey of his work that also functions as a sort of manifesto about his approach to interface design), in interfaces that allow us to interact directly with the data that software allows us to analyze, and to understand something of the processes involved. Emphasizing the hermeneutic work of interface design and digital media analysis, he writes: 67

- Intimate interfaces cannot be "wireframed" or otherwise faked. They are a lens more than a frame.
- Repetition can lead to new discoveries. Use is not "monotonous" and therefore automation is never the goal.
- Meaning "passes through" the system. "The answer" is never directly modeled, but occurs in the heads of humans.
- Programming intimate interfaces is more intimate than using them. And since this intimacy is fundamentally lingual, it is a non-intuitive process poorly served by "end user" tooling.
- [...] Search, and re-search, must be processes of acquiring intimacy[,] not abstracting to an outline or de-skilling to a workflow. [Ochshorn 2016]

While Drift and Gentle are not the most innovative examples of Ochshorn's creations — some of those, developed for working artists, allow for novel interactive research with film and video, and can be sampled in a talk he gave at Google Brain in 2017[20] — they provide both ease of interaction with the object of study (in our research recordings of performative speech) and direct access to the underlying data. They provide a lens; they are a means to an end that tries not to obscure the means; and they encourage small-scale, intimate, qualitative research as a necessary prelude to large-scale, quantitative research. 68

Our own collaboration honors the fact of disciplinary expertise — Ochshorn's skill in and knowledge about interface design, object-oriented programming, and audio signal processing are complemented by MacArthur's intimacy with poetry as an oral form and her capacity to pose rich questions about poetry performance and literary history, and by Miller's understanding of speech production and perception, audio signal processing and quantitative analysis. The years of training and experience that each of us brings to the collaboration cannot be underestimated. In DH, of course, it is not unusual that technical expertise is provided by men expertise is provided by men (Ochshorn and Miller), though they do not provide only that, nor that the woman participating in the collaboration (MacArthur) has acquired new technical knowledge in order to carry out the research. 69

In "An Information Science Question in DH Feminism", Tanya Clement writes: 70

> In digital humanities, the technology of mastering technology has been considered a productive means of combatting what is considered the general "degree of ignorance about information

> technology and its critical relevance to humanities education and scholarship" [McGann 2015, 71] ... [t]here has been much debate in DH about whether designing and building tools gives one an indispensable knowledge of its processes and whether or not women and people of color (and others) are precluded from these activities for a variety of very real and very situated reasons. In this sense, the rhetoric of "mastery" over technology can be intellectually prohibitive since it threatens an advancement of knowledge production from other perspectives [Clement 2015]

While we agree that, on the one hand, requiring — much less fetishizing — technical mastery in DH research can lead to the exclusion of valuable and novel perspectives from DH research, we also have practical objections to the requirement of technical mastery. To be precise, we resist expectations that collaborative researchers can or should, on the fly or on the cheap, acquire real expertise in one another's disciplines. Beyond the political and economic problems with asking graduate students and scholars in the humanities to master digital technologies or acquire advanced coding skills — which must inevitably take time away from the primary study of, for instance, literary history — it is practically inefficient not to collaborate across disciplines on the development of tools such as Gentle and Drift and the Voxit toolbox, and in the research applying them to poetry recordings. 71

At the same time, it would be all too easy for a literary scholar with no background in linguistics, data visualization or quantitative analysis to misunderstand Gentle and Drift and the data they provide. Thus our approach in developing these tools, training users to apply them, and developing tutorials and documentation, has been to emphasize ease of use and access to data while also educating users on the basics of speech production and perception, audio signal processing and quantitative analysis. We are now at work on a Slow Listening book project, in collaboration with Neil Verma and Christopher Grobe, that will offer a history of machine-assisted voice analysis, a critical theoretical approach, a linguistically grounded method, and four case studies, providing models for future research, from literary performance and AI voices to radio drama and podcasts. 72

The image at the outset of this article, a pitch contour in Drift, visualizes the pitch of John Ashbery's voice, in a 1952 reading of his sestina "The Painter",[21] at the 92nd Street Y in New York. The painter in the poem aspires to paint the sea. Eventually, he gives up, and leaves the canvas blank. The text shown in the pitch contour is line 9 of the poem, articulating advice given to the painter by the "people who lived in the buildings", who watch him work at his impossible task: "Try using the brush / As a means to an end. Select, for a portrait, / Something less angry and large, and more subject / To a painter's moods" (ll. 9-12). "The Painter" almost too neatly thematizes Ashbery's poetic ambitions — not only to represent, but to enact on the page the big impossible subject of consciousness unfolding in time, here symbolized by the changing face of the sea, as it constantly moves and shifts. This is an apt figure for the experience of listening in real time. Sound, and the human voice in particular, elude our perceptual grasp, much as the sea slips through our fingers, to mix metaphors. But these tools, we hope, extend the possibilities for listening more slowly and thoughtfully. 73

Notes

[1] While it isn't possible to embed audio files in this article, we do provide stable links at zenodo.org.

[2] See, for instance, the Soundbox project, developed at Duke University: http://soundboxproject.com.

[3] See our discussion of Raphael Allison's *Bodies on the Line: Performance and the Sixties Poetry Reading* (2014) in [MacArthur et al. 2018]. See also Leslie Wheeler, who asserts that contemporary American poets, as observed at the 2006 Associated Writers Conference, typically do not "display emotions at their readings but instead tend . . . to manifest intellectual detachment, if not in the poem's words then through carefully neutral delivery" [Wheeler 2008, 140]. This may sound intuitively accurate, but what does neutral delivery mean? Might the display of some emotions come across as vocal restraint? And so on.

[4] See our discussion of Raphael Allison's *Bodies on the Line: Performance and the Sixties Poetry Reading* (2014) [MacArthur et al. 2018].

See also Leslie Wheeler, who asserts that contemporary American poets, as observed at the 2006 Associated Writers Conference, typically do not “display emotions at their readings but instead tend ... to manifest intellectual detachment, if not in the poem’s words then through carefully neutral delivery” [Wheeler 2008, 140]. This may sound intuitively accurate, but what does neutral delivery mean? Might the display of some emotions come across as vocal restraint? And so on.

[5] https://drift4.spokenweb.ca/

[6] An average frequency for male voices is 125 Hz, and 225 Hz for female voices [Gussenhoven 2004, 3], though this can vary by language, region, and many other variables; by “male” and “female” in this context, we mean biological sex assigned at birth.

[7] While linguists typically refer to the fundamental frequency as f0, there is some disagreement about whether to start counting at 0 or 1.

[8] Please see http://dh2016.adho.org/abstracts/128.

[9] In 2015-16, MacArthur had an ACLS Digital Innovations Fellowship at the University of California, Davis, with the initial plan of developing a pitch-tracking interface for humanities scholars using Praat and ARLO. CDG, initially funded by SAP, turned into HARC (Human Advancement Research Lab, part of Y Combinator Research). Ochshorn co-founded a start-up, Reduct.Video, in 2017.

[10] http://lowerquality.com/gentle/

[11] https://textinperformance.soc.northwestern.edu/

[12] https://spokenweb.ca/

[13] https://drift4.spokenweb.ca/

[14] Comparing data for intensity — volume or amplitude, in other words, measured in decibels — can be very challenging across recordings, as many variables influence intensity patterns, included the recording space, recording equipment, media format, distance of a speaker from a microphone, etc. However, patterns within recordings can be meaningful.

[15] https://doi.org/10.22148/16.022

[16] https://arcade.stanford.edu/content/after-scansion-visualizing-deforming-and-listening-poetic-prosody

[17] https://www.theparisreview.org/blog/2019/10/29/john-ashberys-reading-voice/

[18] https://lareviewofbooks.org/article/101-black-women-poets-in-mainly-white-and-mainly-black-rooms/

[19] Oversimplifications about the voice are hard to escape, as exemplified in the game of telephone that occurred when our research briefly went viral, as in the aggregator Newser’s URL description, “Science Proves Boring Poet Voice Exists”: https://www.newser.com/story/258972/science-proves-boring-poet-voice-exists.html. The headline was even worse: “Poems May Be Great, but ‘Poet Voice’ Is the Pits”. Newser misread an article that appeared in Atlas Obscura, “An Algorithmic Investigation of the Highfalutin' ‘Poet Voice’”, which in turn brought attention to but caricatured aspects of the longer article published in *The Journal of Cultural Analytics*, and took its title not

from anything we wrote about Poet Voice, but from a quotation from poet Rich Smith's popular polemic about Poet Voice quoted above, which appeared in City Arts. https://www.cityartsmagazine.com/stop-using-poet-voice/

[20] http://rmozone.com/snapshots/2017/10/rmo-at-google/

[21] https://www.theparisreview.org/blog/2019/10/29/john-ashberys-reading-voice/

Works Cited

Armantrout 1998 Armantrout, Rae. "Heart of It". Audio. *PennSound*. University of Pennsylvania (1998).

Ashbury 2019 Ashbery, John. "The Painter." Audio file. In Marit MacArthur, "John Ashbery's Reading Voice". *The Paris Review Online*. 2019. https://www.theparisreview.org/blog/2019/10/29/john-ashberys-reading-voice/

Bernstein 1998 Bernstein, Charles (ed.) *Close Listening: Poetry and the Performed Word*. Oxford University Press (1998).

Boersma 1993 . Boersma, Paul. "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound". *IFA Proceedings* 17 (1993): 97-110.

Boersma and Weenink 2003 , 2018. Boersma, Paul and Weenink, David. "Praat: doing phonetics by computer". Version 6.0.37. http://www.praat.org/

Chin Chin, Marilyn. "Blues on Yellow". "Poems Out Loud". PoemsOutLoud.net. W.W. Norton Publishers, New York, NY. n.d.

Clement 2015 Clement, T. "An Information Science Question in DH Feminism". *Digital Humanities Quarterly* 9.2 (2015).

Clement et al. 2016 Clement, T, Auvil, L., Tcheng, D. "White Paper: High Performance Sound Technologies for Access and Scholarship". "White paper for the NEH Office of Digital Humanities" (January 2016).

Dolscheid et al. 2013 Dolscheid, S., Shayan, S., Majid, A., and Casasanto, D. "The Thickness of Musical Pitch: Psychophysical Evidence for Linguistic Relativity". *Psychological Science*, 24(5), (2013) https://doi.org/10.1177/0956797612457374

Eidsheim 2019 Eidsheim, Nina Sun. *The Race of Sound: Listening, Timbre and Vocality in African American Music*. Duke University Press Books (2019): 9. 26..

English 2020 English, J. "Teaching the Novel in the Audio Age" *PMLA/Publications of the Modern Language Association of America*, 135:2, pp. 419-426. https://doi.org/10.1632/pmla.2020.135.2.419

Frost Frost, Robert. "Never Again Would Birdsong Be the Same". "Collected Poems, Prose and Plays". New York: Library of America (1995): 308

Ginsberg Ginsberg, Allen. "Howl". Audio. *PennSound*. University of Pennsylvania (1956).

Glück Glück, Louise. "The Wild Iris (Audio Only)". Poets.org. Academy of American Poets, n.d.

Glück 1992 Glück, Louise. "The Wild Iris". Wild Iris. New York: Ecco (1992): 22–23.

Goffman 1981 Goffman, Erving. *Forms of Talk*. Philadelphia: University of Pennsylvania Press (1981).

Gussenhoven 2004 Gussenhoven, Carlos. *The Phonology of Tone and Intonation*. Cambridge University Press, New York, NY (2004): 3, 6, 5.

Hammond and Dick 2019 Adam Hammond and Jonathan Dick "They Do the Police in Different Voices: Computational Analysis of Digitized Performances of T. S. Eliot's The Waste Land". *Association for Computation in the Humanities*, Pittsburgh, PA, June 2019.

Jacobellis v. Ohio "Jacobellis v. Ohio", 378 U.S. 184 (1964).

Jouvet and Laprie 2017 Jouvet, D., and Laprie, Y. "Performance analysis of several pitch detection algorithms on simulated and real noisy speech data". *25th European Signal Processing Conference*, EUSIPCO 2017. https://doi.org/10.23919/EUSIPCO.2017.8081482

Kreiman and Sidtis 2011 Jody Kreiman and Diane Sidtis, *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. New York: Wiley, 2011, 305-306.

Lee and Ellis 2012 Lee, B. and Ellis, D. “Noise Robust Pitch Tracking by Subband Autocorrelation Classification”. *Interspeech*. 2012.

Lehiste 1970 Lehiste, Ilse. *Suprasegmentals*. Cambridge, MA, Massachusetts Institute of Technology Press (1970): 65.

Llompart 2016 Llompart, Cecilia. “Omens”. Poets.org. Academy of American Poets, 2016.

MacArthur 2016 MacArthur 2016. “Monotony, the Churches of Poetry Reading, and Sound Studies”, *PMLA* 131.1 (Jan. 2016): 38-63.

MacArthur 2019 “John Ashbery’s Reading Voice”. Oct. 29, 2019. *The Paris Review Online*.

MacArthur and Miller 2018 MacArthur, Marit, and Miller, Lee M. “After Scansion: Visualizing, Deforming and Listening to Poetic Prosody”. *Stanford ARCADE Colloquy Series: Alternative Histories of Prosody*, Dec. 13, 2018.

MacArthur et al. 2018 MacArthur, Marit, Zellou, Georgia and Miller, Lee M. “Beyond Poet Voice: Sampling the Performance Styles of 100 American Poets”. *Journal of Cultural Analytics*, March 2018. http://culturalanalytics.org/2018/04/beyond-poet-voice-sampling-the-non-performance-styles-of-100-american-poets/

MacArthur et al., 2022. MacArthur, Marit, Rambsy, Howard, Wu, Xiaoliu, Ding, Qin, and Miller, Lee M., “101 Black Women Poets in Mainly White and Mainly Black Rooms”. August 27, 2022, *Los Angeles Review of Books*. https://lareviewofbooks.org/article/101-black-women-poets-in-mainly-white-and-mainly-black-rooms/

McGann 2015 McGann, J. “Culture and Technology: The Way We Live Now, What Is to Be Done?”. *New Literary History*. 36:1. The Johns Hopkins University Press: 2005.

Mertens et al. 2021 Mertens, Jacob et al. “Drifting Voices: Studying Emotion and Pitch in Podcasting with Digital Tools”. *Saving New Sounds: Podcast Preservation and Historiography*, edited by Jeremy Wade Morris and Eric Hoyt, University of Michigan Press, 2021, pp. 154–78.

Mustazza 2018 Chris Mustazza, C. “Machine-aided close listening: Prosthetic synaesthesia and the 3D phonotext”. *Digital Humanities Quarterly* 12:3 (2018)

Noble 2011 Noble, D. *Forces of Production: A Social History of Industrial Automation* Routledge: 2011.

Ochshorn 2016 Ochshorn, Robert. “Intimacy and Interface”. Lecture. Davis, California. February 15, 2016.

Ochshorn and Hawkins 2016 Ochshorn, Robert and Hawkins, Max. *Gentle*. Computer software. GitHub. Vers. 2.0. LowerQuality, 2016.

Ochshorn and Hawkins 2019 Ochshorn, Robert, and Hawkins, Max. *Drift*. Computer software. GitHub. Vers. 3.0. LowerQuality, 2019.

Petrusich 2019 Petrusich, Amanda. “‘Domestic Arts.’ Rev. of Bill Callahan’s Shepherd in a Sheepskin Vest”. *The New Yorker* (June 2019): 88-89.

Povey et al. 2009 Povey, Daniel, Nagendra Goel, Lukás Burget, Mohit Agarwal, Pinar Akyazi, Feng Kai, Arnab Ghoshal, Ondrej Glembek, Martin Karafiát, Ariya Rastrow, Richard C. Rose, Petr Schwarz and Samuel Thomas. “Low Development Cost, High Quality Speech Recognition for New Languages and Domains: Report from 2009 Johns Hopkins / CLSP Summer Workshop”. Allen Institute. Semantic Scholar (2009).

Richards 1926 Richards, I.A. *Principles of Literary Criticism*. Psychology Press, Routledge Classics Series, New York, NY (1926, 2001).

Rosenfelder et al 2014 Rosenfelder, Ingrid, Fruehwald, Josef, Yuan, Jiahong, et al. “FAVE (Forced Alignment and Vowel Extraction)”. Suite Version 1.1.3 https://doi.org/10.5281/zenodo.9846

Smith 2016 Smith, Rich. “Stop Using ‘Poet Voice’.” CityArts. Encore Media Group (July 2014).

Smith 2021 Smith, Jacob. “The Courtships of Ada and Len: Mediated Musicals and Vocal Caricature Before the Cinema”. *The Oxford Handbook of Cinematic Listening*. Cenciarelli, Carlo, Ed. New York: Oxford UP, 2021.

Sterne 2003 Sterne, Jonathan. *The Audible Past: Cultural Origins of Sound Reproduction*. Duke University Press Books, Durham, NC (2003): 15.

Strömbergsson 2016 Strömbergsson, S. *Proceedings of the Annual Conference of the International Speech Communication Association*, INTERSPEECH. https://doi.org/10.21437/Interspeech.2016-240

Trethewey 2006 rethewey, Natasha. “Monument (Audio Only)”. Poetry.org. Academy of American Poets (2006).

Wheeler 2008 Wheeler, Lesley. *Voicing American Poetry: Sound and Performance from the 1920s to the Present*. Cornell University Press, Ithaca, NY, (2008).

Wilkinson 2014 Wilkinson, Alec. “A Voice from the Past”. The New Yorker (May 2014).

Yeats 1937 Yeats, William Butler. “The Lake Isle of Innisfree”. Audio. PennSound. University of Pennsylvania (1937).

Zapruder 2009 Zapruder 2009. Zapruder, Matthew. Wavepoetry. Soundcloud. “When It’s Sunny They Push the Button”. Wave Books, Seattle, WA.