Iperstoria
2024, no. 24
ISSN: 2281-4582

DOI: 10.13136/2281-4582/2024.i24.1562
iperstoria.it

It Do the Poets in Different Voices

Generative AI Voices, the Uncanny, and the Poetry Audio Archive

Chris Mustazza
University of Pennsylvania
ORCID: https://orcid.org/0009-0002-3076-037X
Email: mustazza@sas.upenn.edu

Keywords

Poetry
Sound studies
AI
Voice
Audio

Abstract

This paper examines the advent of poetry performed by generative-AI voices. It proposes that these synthetic voices can cast a new light on literary historical understandings of the original performances through their minute differences. In places where the AI fails to fully achieve verisimilitude, there can be an "uncanny valley" effect, which allows listeners to hear the original performances anew, defamiliarized in the Russian Formalist sense of the term. The paper goes on to propose a phonetic comparison between the AI voices and the human voices they were trained on. By identifying a number of prosodic differences, the paper posits why the AI voice may sound similar but not quite right, leading to the uncanny effect. The intent of this machine-aided close listening is to further discuss the nuances of the human performances and their literary historical contexts, highlighting how the AI voice can support analyses of the original author performances.

1. The voice and its doubles

The PennSound Archive, the world's largest archive of recordings of poets, helped to bring about within literary studies a project of sonic revision. For more than twenty years, scholars have turned to these sound recordings with the aim of close listening, working to, as Bernstein puts it, "overthrow the common presumption that the text of a poem – that is, the written document – is primary and that the recitation or performance of a poem by the poet is secondary and fundamentally inconsequential to the 'poem itself'" (1998, 8). These audio performances by poets present a literary *objet d'art*, but also bring us face to face with how the voice shapes identity, not just in the diegetic contexts of the works but also in everyday, lived speech. We might turn here to Erving Goffman's *The Performance of Self in Everyday Life,* which argues that identity is a series of performances for others (1973). In practical terms, I sometimes ask my students "Do you ever stop midsentence and say to yourself, 'I sounded just like my mother when I said that'? And I don't mean the semantic meanings of the words, but the pacing, intonation, emphases with which you said them." I sometimes think about this when I'm giving an academic

lecture. I stop and think, "I sound a little like my advisors here." Who else would I sound like? Our performative selves are intrinsically mimetic, according to Goffman and contemporary Sociolinguistic theories (Goffman 2021, 22.; Anderson 2018, 667). But unlike Sociology or Linguistics, attention to literary performances gives us a common, aesthetic object to examine and question. This, to me, is the greatest affordance of the sound archive – not performances as relics to bring us closer to a text, but as distillations of performed identities.

Perhaps the most famous poetic treatise on this topic is T.S. Eliot's *The Waste Land*, originally titled *He Do the Police in Different Voices*, before that Dickensian title was stricken by "Il Miglior Fabbro," Ezra Pound. *The Waste Land*'s initial titular reference to Charles Dickens' *Our Mutual Friend* connects Eliot to Dickens beyond the poem's portrait of working-class Britain. The two authors both performed their poems to varied audiences, with an emphasis on the theatrical embodiment of lived, spoken voices. Dickens' American lecture tours of the nineteenth century sold out shows across the country, and they featured him engaging in the Victorian tradition of a specific kind of platform performer: the monopolylogue, one person doing all of the voices in a dramatic work.[1] As Eliot's haunting, disembodied voices in *The Waste Land* coalesce into an uneasy whole (Eliot 1935), they are best understood as somewhere between cinematic (thinking of the poem in scenes) and in the tradition of Dickens' monopolylogue. A key difference, though, is the indeterminacy of the speaking subject(s) in *The Waste Land*, versus Dickens' embodiment of the dramatis personae of his novels. And this indeterminacy, which would inflect modernist poetics for years to come, provides the greatest affordance for moving beyond the question of *who* is speaking, toward the question of how speech sounds precede (and thus produce) subjectivity and identity.

In addition to this dynamic of sound crafting the individual subject (let's consider that, metaphorically, an idiolect), we can also discuss how the voices done by poets connect them to wider generic conventions, literary scenes, or other shared performance styles. For example, I have written about the connection between sermon recordings of the early twentieth century and poetic performances that were modeled on them (2023). Along these lines, Marit MacArthur has worked to prosodically define "Poet Voice," or what she terms "monotonous incantation" (2016, 44), locating the dominant reading style in the American academy as shaped, sonically, by White Protestant sermons. All of these examples are instances of historicizing the voice, moving from material prosodic facets to sociological, historical, and political understandings of how the performed voices signify. These studies also mark instances of the creation of shared identities that manifest through vocal mimesis. Simply put, there are certain styles of performance that trigger associations with genre purely through their sounds. These

[1] For more on the tradition of the monopolylogue as it entered the twentieth century, see Feaster 2021.

soundscapes are learned and shared via performances. This paper begins in the space of vocal mimesis in order to set the stage for the heretofore apotheosis of such mimicry: the advent of generative AI and its ability to copy human voices.

*

On April 4, 2023, a TikTok artist with the sobriquet ghostwriter977 released "Heart on My Sleeve," a song purportedly by Drake and The Weeknd, but in actuality, crafted with a generative-AI algorithm to mimic their voices and style (Veltman 2023). The track garnered millions of clicks across multiple platforms before Universal Music Group was able to get it taken down on a technicality (Patel 2023). The song was also submitted for a Grammy Award, despite its ostensible artists not being involved with the song's production, lending only their vocal identities (Shanfeld 2023). Most conversations around this pivotal moment in musical technology history have centered on a couple of predictable loci: did the recording *actually* sound like Drake, and is it *ethical* to clone someone else's voice without their permission (Veltman 2023; Coscarelli 2023)? The latter question posits a sort of ownership over something. Is it the particular sonic register generated by the one's vocal cords and larynx, shaped by their mouth and tongue positions? Is it the cultural, historical, and social facets of the voice that give sonic presence to the concept of identity? In other words, what do we own when we say one can own a voice? Are we in the order of the Imaginary, the Symbolic, or the Real – to borrow a Lacanian framework?

While all of these questions of verisimilitude, ethicality, and ownership are worthy of discussion, the rise of generative AI voices toward artistic ends opens a completely new set of possibilities for literary scholarship and ways of listening. I have previously discussed the myriad layers of voices audible in any vocal recording, claiming that in addition to factors like regional accent, which are occluded by writing, vocal recordings contain many other layers, including the influences of other performers and literary scenes, the context of the given recording, and the technological devices of mediation used in making the recordings (Mustazza 2022). We could ask the same questions of the AI voices trained on human speech data. These voices are not learning a person's identity: they are mimicking a very specific context of speech. For example, if an AI model were trained on conversational speech, it would sound very different than if it were trained on the same person giving a lecture or participating in a podcast or acting in a play. The model does not mimic an individual in toto: it mimics a context, which allows for rich literary historical analyses of these synthetic voices, as this context provides a kind of embedded historicization and periodization.

In some sense, an AI voice is a distilled, distant, diachronic (sorry for the consonance here – if you train an algorithm on this section of my speech, it should sound uncharacteristically mellifluous) portrait of an individual's performance over time. It's a temporal plurality presented as a presentist amalgamation. If we can peel apart this plurality, we can learn something about not just the author the voice is modeled on but on the wider context that gave rise to their styles of performance, which is the aim of this paper. I would like to propose a peripatetic stroll through one of the first collections of poetry (if not the first) rendered as deep-learning vocal performances, Charles Bernstein and Davide Balula's *Poetry Has No Future Unless It Comes to an End: Poems of Artificial Intelligence* (Bernstein and Balula 2023). My aim is to discuss the affordances and limitations of this kind of use of generative AI for poetry and literary studies, with a focus on what we can learn from *perceived inaccuracy* or the uncanny. Put another way, I am less interested in what it seems that the algorithm "gets right" and more in what appears slightly askew. In these moments of dissonance, we stand to learn something through a process of defamiliarization, in the Russian Futurist sense of the term.

I want to stress here that what follows is not a formalist analysis, at least not in the way that term has been used traditionally in critical methodologies, suggesting a New Critical severance of a literary object from the historical and social factors that give rise to it. Quite the opposite, the sustained attention to form in this essay is explicitly to connect the works, machine-generated and not, to wider frames of performative identity. In addition to the concerns raised by Erving Goffman, the Eliotic title of this essay also invokes the concept of doing gender (West and Zimmerman 1987), contemporary sociolinguistic theories, and other strains of thought related to performative identity. The material surface or form of these AI voicings, I argue, is the threshold to perceiving how the recordings enact or refute established literary histories. The methodologies employed here draw from a number of interdisciplinary fields and methodologies. These include sound studies, literary modernism (and work to classify it, such as Peter Nicholls' *Modernisms*), and the overlap between poetics and phonetics (see Colonna 2022; Liberman 2007). One could also relate this work to the welcomely nebulous field of voice studies, as defined by Eidsheim and Meizel. As they put it, "Voice studies offers tools to better detect the values underpinning any definition of voice. And voice studies deconstructs not only the performance of the voice, but also the performance of claims to voice" (Eidsheim and Meizel 2019, xiv). This study is about what happens when AI lays claim to a voice and all that the voice contains.

2. A heap of broken syllables

One of the newest and most ambitious examples of AI Modernism (if I can call it that) is *Poetry Has No Future Unless It Comes to an End: Poems of Artificial Intelligence* by Charles Bernstein

and Davide Balula (2023). The collection of over 70 poems was generated from a corpus of Bernstein's poetic writings spanning from 1972 through 2021. The deep learning model attempted to create an aesthetic amalgamation of Bernstein's poetics and generate these new poems, which Bernstein the Man (how should I distinguish between Charles Bernstein and "Charles Bernstein" for the duration of this argument?) was permitted to re-lineate and delete from, but not to add any new content, a process Bernstein referred to as "human-assisted AI" (Bernstein and Balula 2023, 15). The project's primary innovation occurs in its second phase, where a "synthetic" (Bernstein and Balula's word) clone of Bernstein's voice was created from a data set of his poetic performances, trained mostly on the audiobooks of his later collections, *Near/Miss* (2018) and *Topsy-Turvey* (2021). I want to pause here to note that the training dataset is a crucial consideration, as the machine learned a very particular mode of performance, in this case, the genre of the audiobook reading. I will say more on this later. The newly born voice was then used to perform the AI-generated poems. So the collection itself is one of multilayered simulation – a generated voice performs generated poems, including one with perhaps my favorite title, "I Am the Shadow of Poet Charles Bernstein." (I like to imagine the synthesized voice saying to Poet Charles Bernstein: "You! Hypocrite lecteur! Mon semblable, mon frère!"[2])

As I encountered these works from Bernstein's "synthetic brother" (Bernstein and Balula 2023, 14), I found myself tarrying with a very basic question: to what degree does this "sound" like Bernstein, both in the sonic-material sense of the performances and in the metaphorical notion of the voice in the writing? I want to stress here that this is a different question than "are the poems good?," which is of less interest to me. My question sounds like a kind of return to the knee-jerk questions asked of "Heart on My Sleeve." My interest, however, is less about the quality of the technology's reproductive fidelity and thus not a fetishization of "accuracy." Rather, I would like to propose a comparative approach that might allow the AI works and human-written works they are based on to cast light on one another. The AI voice is not an ersatz simulacrum of the human voice, one that tends ever asymptotically toward verisimilitude; it can stand as a way to de-mix (in the musical production sense) a voice into its constituent sub-voicings.

Before moving on to the sound recordings, let's take a look at the content and form of some of the textually generated poems. An interesting example is "What is Meant by the New Criticism?" (Bernstein and Balula 2023, 33). The overall written voice of the poem does not, to me, sound like a Bernstein poem. It's too overtly didactic, almost pedantic: "Each generation

[2] This is a famous quote from Charles Baudelaire's "Au Lecteur" from *Les Fleurs du Mal*. The line is perhaps more famous from its quotation in *The Waste Land*, from which I borrow the already borrowed line here, to continue my motif.

has to make its own/ set of demands, set of values/ its own culture and politics" (Bernstein and Balula 2023, 33). This sounds more like a valedictory address than the gnomic and antinomian contours of Language-inflected poetics I would expect. Though, certainly, one could say that there this is a playful deployment of "generation" (cf. "generative," "generate") within the context of how these poems were made. But even if some parts of the poem read as farther afield from the expected poetics, there are certain aspects that are haunted by specious accuracies. Take, for example, "One can be both/ a revolutionary/ and apolitical" (Bernstein and Balula 2023, 33). This excising the core of a word to examine what stands without it (see Bernstein 2012) seemed an accurate reflection of the poetics, yet the so-called content is not something I would ever imagine seeing in a Bernstein poem. A similar case is the line "I am not interested/in anything more than the content of/the poem" in "Story Continues Below Advertisement" (Bernstein and Balula 2023, 84). If Bernstein had written this himself, I would take this to be an ironizing statement given his interest in form (to the extent that it can be separated from content). In all of this fretting over "accuracy," I encountered a sort of poetic Turing test. Given that I knew that the poem was "human-assisted AI," I decided that these lines were "wrong," that they did not accurately capture the gestalt of the poetics. BUT – had I not known that these were AI-generated lines, I would have assumed that we were encountering a multi-layered Bernsteinian reversal, where he is mocking and satirizing the notion that it's possible for any action to be apolitical – and perhaps we are, given that Bernstein the Human Poet did not delete these lines. So in what seems to be a machinic misreading of Bernstein's work (though better than many human misreadings I've read!), one is given a new lens through which to perceive our unassisted, human understandings of a poet and their poetics.

These machine-generated works also raise a number of questions on the nature of humor and irony, and the possibility of their generation or identification through deep learning. So much of humor and irony is about creating a disjuncture between what is being said and the context of the speech, especially the audience knowing that the speaker knows that the thing they are saying is absurd, the old "I know, that you know, that I know that you know…" This kind of humorous irony is known as "incongruity theory" (Buijzen and Valkenburg 2003; Berger 1993). At this moment in the history generative AI, this sort of situational awareness necessary for incongruity is not possible and thus humor can only land in a less intentional way. But that doesn't mean it's not funny; such a digital death of the author moment brings us face to face with the raison d'être of poetry: audience reception. The writing is ironic if the audience believes it to be so; it's funny if it gets a chuckle. Of course, this is just basic reader-response theory, but what's new here is that the works being generated are not being crafted with a simple rule set. Balula states of the poems, "I was intentionally looking in the opposite direction for this project… I was looking for intentionality in the machine, not creative accidents or aided

serendipity" (Bernstein and Balula 2023, 13). While I'm not sure if we are encountering intentionality per se, the AI is creating a new "heap of broken images" (Eliot 1922) from the shattered pane of an oeuvre.

Much more could be said about the conceptual ontology of the deep learning poems, but the more interesting angle, I think, is when we hear the synthetic voice perform the poems.[3] The voice in "I Am the Shadow of the Poet Charles Bernstein" (an appropriate title that I'm sure is meant to connect with the idea of digital surrogates/doubles) sounds historically semi-accurate, but not like Bernstein's reading style, per se (compare with other readings from PennSound[4]).[5] To my unaided ear, it sounds more like Bruce Andrews or Peter Gizzi, or any number of East-Coast, male poets affiliated with the so-called Language Poetry of the late 1970s through the 1980s. There is a sort of post-Beats syncopation with an edge to the readings. It sounds way more aggressive than I would expect. This argues that one of the things the machine learned is something about the dialectical formation of performance based on literary scenes, call it how the poets *do* poetry. While I was initially listening for an idiolectic reading, there is more to be apprehended by the vagaries of the deep learning model. The machine was, in part, performing Bernstein's performance of others.

This observation draws from second-wave Sociolinguistic theory on linguistic variation. In contrast to the founding theories of Sociolinguistics, in which dimensions like social class were seen to be the primary determinants of linguistic variation, second-wave "variationists recognized that locally-relevant facts about their participants also played an important (if not more important) role in understanding patterns of linguistic variation. For example, the social clique that high school students belong to might be a better predictor of the linguistic behaviour than a student's social class" (Anderson 2018, 667). Applied to the AI voice here, such a theory recognizes that social circles (in this case literary scenes) influence not just linguistic variation but also performance styles. The key point here is not that Bernstein might sound like his contemporary performers; it's that in the machine's uncanny, not-quiteness, it reveals a subtlety of these performances that might go unnoticed by the unaided human ear.

Other poems, like "If You Fall Asleep at the Wheel,"[6] sound more like Bernstein, in an uncanny sort of way. Here I use the term as in its use in the "uncanny valley," the space where a technology becomes just anthropocentric enough to be creepy. As Masahiro Mori puts it, "I have noticed that, in climbing toward the goal of making robots appear like a human, our affinity for them increases until we come to a valley…, which I call the *uncanny valley*" (2012,

[3] Listen here: https://viseu.us/ai-bernstein-balula/. All websites last visited 11/12/2024.
[4] https://writing.upenn.edu/pennsound/x/Bernstein.php.
[5] https://viseu.us/ai-bernstein-balula/bernstein-balula_i_am_the_shadow_of_poet_charles_bernstein/.
[6] https://viseu.us/ai-bernstein-balula/if-you-fall-asleep-at-the-wheel/.

98). The editors of the reprint of Mori's 1970 essay put it as such: "[Mori] hypothesized that a person's response to a humanlike robot would abruptly shift from empathy to revulsion as it approached, but failed to attain, a lifelike appearance" and referred to this inverse ratio between verisimilitude and comfort as a "descent into eeriness" (Mori 2012, 98). The spoken prosody in "If You Fall Asleep at the Wheel" is SO close to Bernstein's reading style, but the voice is not quite his; the recording seems to take off with some promise and then descend into the uncanny valley. That uncanniness is not, however, altogether problematic. The almost-ness of the voice defamiliarizes as it forces a new familiarity. It speaks into existence that which already was by enacting the "not quite." It also creates an alternative and compressed diachronic representation of Bernstein's work – while the written works the machine drew from spanned fifty years of writing, the performances it learned from are around five years old, thus imposing a kind of presentness on a textual corpus than spans decades. One of the reasons the AI voices might sound semi-accurate to me at times is because they sound contemporary. Thus, the similarity beckons for a return to the archive to consider how the generated performance might be anachronistic, even in its seeming accuracy. Put another way, by picking apart how the neural network has anachronistically crafted a voice, we might be able to more clearly historicize its inputs over time. Such a dynamic starts with the form of a discrete textual object (the AI-voiced poem in this case) and backs out into a much more distant view of a poetics.

Such a quest for defamiliarization is at the core of this argument. The deep learning model is: 1) a set of known inputs – we know the data the model was trained on and these are the original materials we wish to know more about; 2) a set of unknown operations by the model (the so-called hidden layers of the network); and 3) a set of examinable outputs (in this case the poems and the AI voice). The model generates something new through an unknown process, and that new object defamiliarizes the old. We can think of this in the Russian Formalist/Futurist sense of the term. *Ostranenie*, as Viktor Shklovsky termed the practice, was all about framing familiar things in strange ways, to allow us to view the objects with a fresh set of eyes (2015). Of course, the influence of these theories on postmodern aesthetics cannot be overstated. The pairing of such aesthetics with the current haute-technology of generative AI brings to mind a term once used to describe the podcast *RadioLab*, the ostensibly oxymoronic descriptor "postmodern… approach to science journalism" (Spinelli and Dann 2019, 2). While so much attention is always focused on scientific technologies that undergird the machine learning algorithms, the outputs of such algorithms allow for a kind of postmodern aesthetics, making meaning through their contingent relationships with the non-AI world.

3. Ex-machina: a machine-aided close listening

I'd like to apologize here in advance if you are feeling like all of these layers of machination – machines generating poetry to be read by machines – is too much. In this section, I will propose using a machine to help us listen to the poems generated by the machine based on the poems written by the machine. The previous sections of the essay have been based primarily on impressionistic close readings and close listenings. I claimed that the voices were uncanny in their almost-humanness, but it would be worth trying to measure what exactly is just a little off in the AI voices, what the boundary conditions of the uncanny valley are. Through a preliminary prosodic analysis by Valentina Colonna, one that relates to her previous works on phonetic analysis of poetry reading (Colonna 2024; Colonna 2022), in the Voices of Italian Poets (VIP)[7] and Voices of Spanish Poets (VSP)[8] projects and audio platforms, we will take a close look at the sonic dynamics that define the synthetic voice in comparison with the human voice, in the hopes of defamiliarizing the inputs of the model through its outputs.

The study involved the comparison of four recordings: 1 and 2) two excerpts from the AI Bernstein voice performing from the collection ("I am the Shadow of the Poet Charles Bernstein" [Bernstein and Balula 2023a] and "Story Continues Below Advertisement" [Bernstein and Balula 2023b]), 3) an excerpt from Bernstein (2021) performing from *Topsy-Turvey* (one of the recordings the AI voice was trained on), and 4) a recording of Bernstein performing "Thank You for Saying Thank You" for a live audience (2003). Bernstein is known as a lively performer of his poetry, so this last sample was meant to contrast with the authors' idea to train the model on audiobook recordings. My goal in asking Colonna for this preliminary analysis was to try to determine what the deep learning model learned when it was trained on the studio recordings. And, more importantly, I was interested in what the model overlooked. In pursuit of these questions, I asked Dr. Colonna to produce a set of statistics on the excerpts of speech, of which I here offer some interpretations.

Before we look at the data, there are a few caveats worthy of consideration. The first is that the audio excerpts analyzed constitute very brief samples (for the purposes of efficiency in this study). It's likely that the results of this comparison would shift with a larger dataset. That said, the data is meant to provide a starting point that can be further tested later. Secondly, every poem has its own poetics and compositional context. It is totally possible that some of the variance between readings is because the poems are voiced in different ways because they are different poems, different soundscapes. At the same time, one of my claims is that the context of the performance (live, studio, etc.) changes its sound, as voicings are very much an extension

[7] https://www.valentinacolonna.com/voices-of-italian-poets-vip/
[8] https://voicesofspanishpoets.ugr.es/en/el-proyecto/

of the media (or lack thereof) used in production. Consider here so-called "Podcast Voice," the dominant voicing in podcast productions, alongside Voice Studies work on topics like voicing in public radio (McEnaney 2019). These broadcast voices are partially the product of generic convention but also due to the condenser microphones used to record and the audio being fed back to the participants through their headphones. Thus, much of what I am trying to measure is how the material context of the training data affects the generated AI voice. With those caveats out of the way, let's take a preliminary look at the data.

Rather than beginning with an analysis of what was said, I decided to start with an analysis of the gaps between speech, the pauses that give cadence and relief within a voiced performance (Figure 1). One thing that jumps out is that the mean length of the pauses between utterances is almost twice as long in the Bernstein studio recording than in the live performance. This stands to reason: while the live performance is crafted on a mutual exchange with the audience and is thus charged with more of an immanence, the studio recordings are characterized by a more relaxed, we might say introspective, cadence that stems from sitting alone in the sonic sterility of the recording studio. The studio pauses are intimate, to borrow from Spinelli and Dann's concept of "podcast intimacy" (2019, 69-70). The live pauses are pregnant, awaiting the audience's response to their call. In comparison, the pauses from the AI voice, in both samples, are less nuanced and lie somewhere in between the live and studio performances. In general, they present a more similar mean duration, compared to the live performance, which is very contrastive, even if a high inner variation is present, especially in one of the two readings.

The AI voices are not quite as dynamic as the live performance but not as introspective as the studio performance. A recent phonetic linguistic study on poetic performances found that actor performances of poetry, versus untrained readers, were characterized by "more and more diverse prosodic boundaries and pauses…and make strategic use of lengthening at verse endings in poetic speech" (Wagner and Betz 2023, 2538). One could substitute the word "performer" for "actor" here, and thus conclude that the increased dynamism in Bernstein's live performance was also due to the fact that he was consciously performing. The AI voice, on the other hand, sits somewhere in between the human performances within their respective contexts. It marks a kind of sonic abstraction as there is no context for its performance, as the voice was never uttered in physical space nor composed with such intent. The synthetic voices are called forward to utter when the command is given. They are not recordings; they are command performances without the underpinning of physical space or recording media, save for what was learned from the model. Perhaps this is part of the uncanny effect and what makes the AI voice exist in the space of the "not quite."

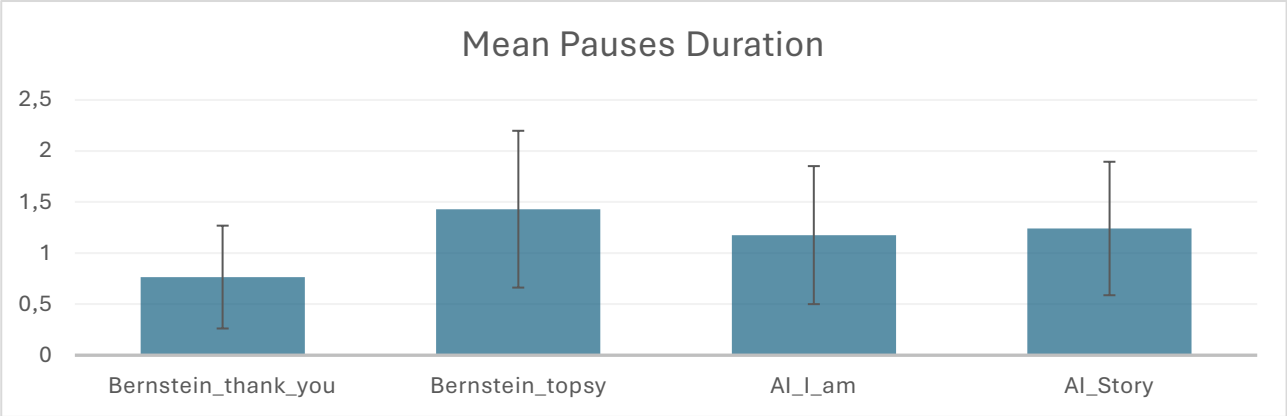

Fig. 1: Mean duration of pauses across the samples of recordings

Next, I was interested in the length of speech between pauses, the prosodic curves (cf. Colonna, 2022), which means interpausal units corresponding to melodic cues (Figure 2). We can see a wide difference between the live performance and the AI voice. This is partially explainable by the very short lines of the poem "Thank You for Saying Thank You," which is performed in the live performance (Bernstein 2006, 7-9). Given that Bernstein emphasizes the lineation in this performance rather than consistently reading through the enjambment, the poem intrinsically has a staccato rhythmic feel. The AI voice in "I Am the Shadow of the Poet Charles Bernstein" is very similar to its training data from *Topsy-Turvey*, which is to be expected. Interestingly, the prosodic curves from "Story Continues Below Advertisement" are much longer than the training data and the other AI voice, though there is a lot of variation in length (see the standard deviation within the recording). To the unaided ear, it renders as a sort of Beat Poetry, stream-of-consciousness aesthetic. It's little aggressive and the opposite of the marching fragmentation of the live recording. It feels like the algorithm is performing in a confessional mode of sorts. It's too…un-self-aware in its performance of interiority, which highlights some of the performative nuance lost in the training of the AI voice.

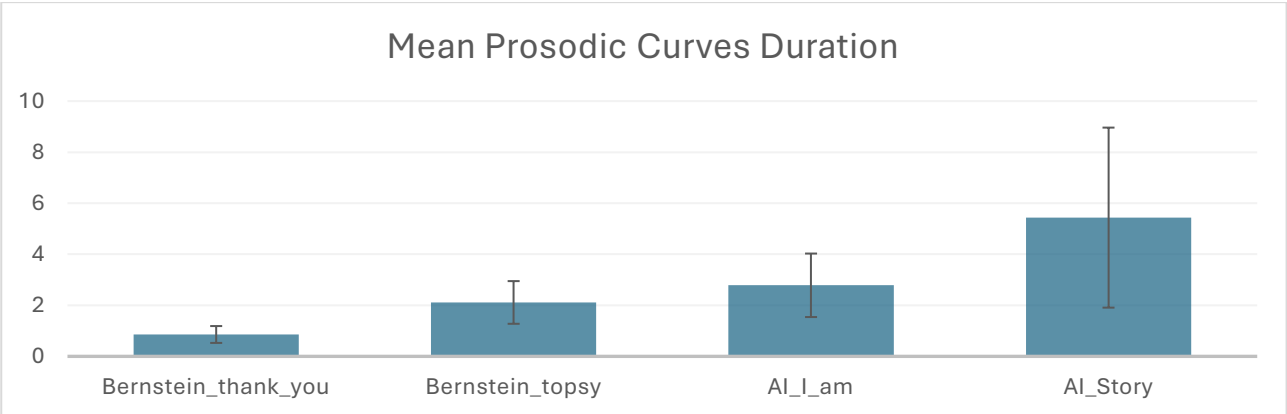

Fig. 2: Mean duration of the prosodic curves across the readings

Finally, I would draw attention to the mean pitch of the performances, understood as the mean fundamental frequency of the voice f_0 (Figure 3). The first thing that jumps out from the graph is that Bernstein's live performance has a much higher average pitch than the others. This could potentially be a factor of vocal effort while performing for a live crowd. The next highest average pitch is the studio recording training data. So we can see a clear difference between both human performances and the AI voices, which are lower in pitch, or sound deeper, than the Bernstein recordings. This downward shift could give the sense of a lack of dynamism or more of a flatness in the AI voice. As part of a future analysis, it may make sense to analyze this in semitones to get a sense of the octave range of the performances, to verify whether the AI voices are in fact being flatter than the human recordings. This data supports some impressionistic close listenings to the performances. When I played these for the students in my Poetry and Music seminar, without showing any empirical data, the students commented that the AI Bernstein was not as dynamic as Bernstein the Man, whom they had already studied in the course. They commented that this lack of dynamism was what made the AI voice feel "off" or "creepy." It's interesting to consider what such an analysis would look like for poets who are less performative and emotive than Bernstein. Perhaps this gulf of expressivity would not exist in those cases.

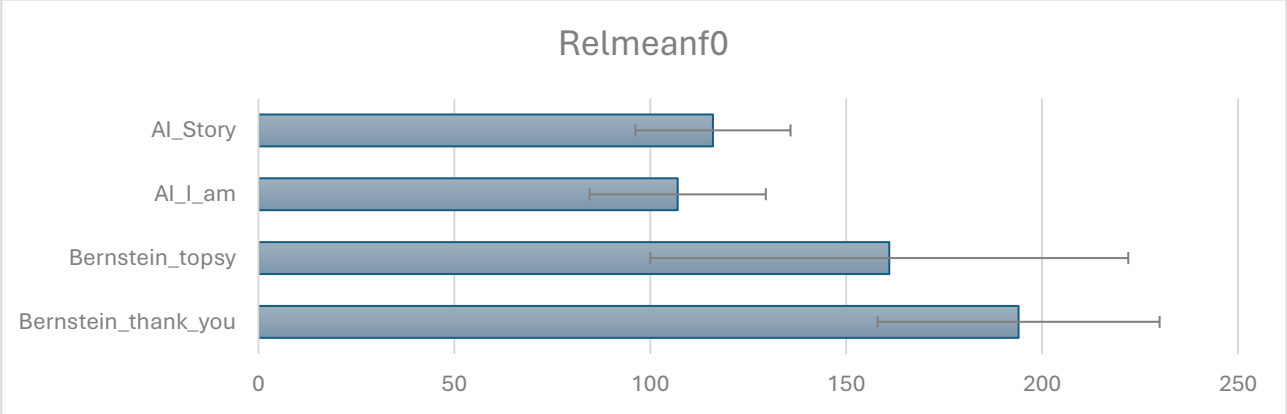

Fig. 3: Mean F0 data for each sample

Finally, a comparison of the main intonational patterns, based on Colonna (2022) and Romano (2017), is worth examining. This process involves annotating all the prosodic curves with the correspondent kind of intonation. From this analysis, there is much less variation in intonation in the AI voices than in the human performances. To explain briefly, intonational patterns can take the shape of rising or falling pitch dynamics of varying kinds. In everyday speech, we hear these dynamics as part of linguistic syntax. For example, a phrase that ends in a rising pitch often denotes a question being asked, but it can also be connected with modes of speech such as "up-talk." In the human performances, several kinds of intonations are visible in the pitch data

and audible in the performances: interrogatives, exclamations, declarations, continuations, etc. The AI performances, interestingly, mainly employed falling pitch cadences that connote declarative statements and continuations. This means that, in addition to being less varied than the human performances, the AI voices were mostly stating and asserting, a closed mode of speech that connotes a kind of authority. Contrast this with an open voicing that asks questions and connotes a more open poetics.

So, to summarize, the overall analysis shows that the human readings use a much wider range of pauses, from short pauses through much longer gaps. The AI voice, while averaging somewhere in between the live and studio performances, tends to favor pauses of similar length. This is because the pauses are learned from data and thus the AI has a much less nuanced understanding of dramatic pauses and how to deploy them for optimal poetic effect. The length of spoken prosodic curves is comparable between parts of the AI and the human performances, but there are places where the AI uses much longer curves. The AI voice is also lower in pitch and tends to favor declarative and continuative intonations, though more research is necessary on this latter point. Each of these dimensions is offered as a way to approach the hauntingly similar yet audibly different gestalt of the AI voices in the collection.

4. Il Miglior Fabbro?

OK, so there you have it – some empirical speech data used to compare the voices. So does Bernstein's "synthetic brother" sound like him? I would conclusively, scientifically, with full certainty say: kinda, maybe, sometimes, depends? There are certainly elements of the poet's performance style that the machine mimics. But the places where it's just a little off – in cadence, in pitch, in dynamics – allow us to hear these factors more clearly in the human voicing. They give us language to be able to speak about what makes for a captivating poetic performance. They take something we may know well, a voice, and make it strange, other, unfamiliar, allowing for the introduction of critical distance. Said another way, the AI voice resists absorption in its uncanniness – the fact that it doesn't sound quite right reminds us that we are encountering a work of art and provokes us to see its moving parts and question its compositional methods.

I offer this as a preliminary volley into the hermeneutic and prosodic study of machine voices, which I hope will be built upon over time. One possible next step would be to expand this analysis to include more data and more prosodic indices from the VIP radar tools developed by Colonna. For now, as I ponder the learning from this study, the words that resonate in my head as a form of (non-)closure come from Amiri Baraka:

> How do you sound, your words, are they
> yours? The ghost you see in the mirror, is it really
> you… (1969, 120)

I wasn't sure, so I asked ChatGPT. It confidently declared (Figure 4):

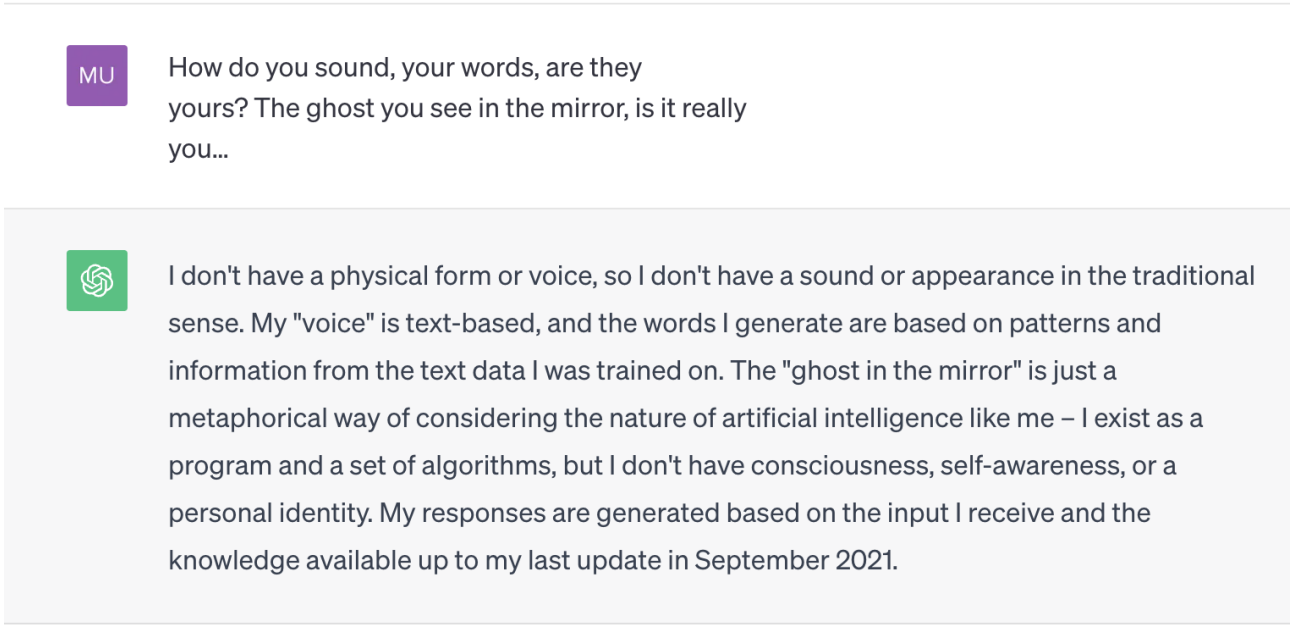

Fig. 4: ChatGPT screenshot of its response to the given prompt

Took the words right out of my mouth.

Acknowledgements

First and foremost, I would like to thank Dr. Valentina Colonna for her advice on and support of this research and for collaborating with me on the prosodic measurements and creating the graphs in this paper. I would like to thank Drs. Anne-Sophie Bories, Nils Couturier, Pablo Ruiz Fabo, and Petr Plecháč for providing a nurturing space for these ideas, year after year.

Bionote

Dr. Chris Mustazza is Co-Director of the PennSound Archive, the world's largest archive of recordings of poets, and he teaches in the English Department at the University of Pennsylvania. His work focuses on poetry and poetics, sound studies, media history, and experimental digital analyses of poetry audio.

Works cited

Anderson, Catherine, et al. *Essentials of Linguistics, Second Edition*. Hamilton: McMasters University, 2018. https://ecampusontario.pressbooks.pub/essentialsoflinguistics2/. All websites last visited on 11/12/2024.

Baudelaire, Charles. "Au Lecteur." 1857. *Selected Poems from* Les Fleurs du Mal*: A Bilingual Edition*. Trans. Norman R. Shapiro. Chicago: University of Chicago Press, 1998. https://press.uchicago.edu/Misc/Chicago/039250.html.

Baraka, Amiri. "Poem for Half-White College Students." *Black Magic: Collected Poetry 1961-1967*. Indianapolis: Bobbs-Merril, 1969.

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

---. "Thank You for Saying Thank You." *Girly Man*. Chicago: University of Chicago Press, 2006. 7-9.

---. "Thank You for Saying Thank You." Philadelphia: PennSound, 2003: 0:48-1:30. http://media.sas.upenn.edu/pennsound/authors/Bernstein/9-25-03_UPenn/Bernstein-Charles_02_Thank-You-for-Saying_9-25-03_Penn.mp3.

---. *Topsy-Turvey*. Philadelphia: PennSound, 2021: 37.15-37:50. https://media.sas.upenn.edu/pennsound/authors/Bernstein/Topsy-Turvy/Bernsten-Charles_Topsy-Turvey_Audible.mp3.

Bernstein, Charles and Davide Balula. *Poetry Has No Future Unless It Comes to an End: Poems of Artificial Intelligence*. Rome: Nero Press, 2023.

---. "I am the Shadow of the Poet Charles Bernstein." *Poetry Has No Future Unless It Comes to an End: Poems of Artificial Intelligence*. Philadelphia: PennSound, 2023a: 0:00-0:30. https://media.sas.upenn.edu/pennsound/authors/Balula/AI/tracks/Balula-Bernstein_p26_I_Am_the_Shadow_of_Poet_Charles_Bernstein_01-03-2023.mp3.

---. "Story Continues Below Advertisement." *Poetry Has No Future Unless It Comes to an End: Poems of Artificial Intelligence*. Philadelphia: PennSound, 2023b: 0:45-1:15. https://media.sas.upenn.edu/pennsound/authors/Balula/AI/tracks/Balula-Bernstein_p84_Story_Continues_Below_Advertisement_01-03-2023.mp3.

Colonna, Valentina. *Voices of Italian Poets. Storia e analisi fonetica della lettura della poesia italiana del Novecento*. Alessandria: Edizioni dell'Orso, 2022.

Colonna, Valentina, Antonio Pamies Bertrán and Stefano Damato. "Towards a Phonetic History of the Voices of Spanish Poets: A First Experimental Study on the Generation of '27." *Journal of Experimental Phonetics* 33.1 (2024): 7-34. https://revistes.ub.edu/index.php/experimentalphonetics/article/view/46106.

Coscarelli, Joe. “An A.I. Hit of Fake ‘Drake’ and ‘The Weeknd’ Rattles the Music World.” *The New York Times* 19 April 2023. https://www.nytimes.com/2023/04/19/arts/music/ai-drake-the-weeknd-fake.html.

Eidsheim, Nina Sun and Katherine Meizel, edited by. *The Oxford Handbook of Voice Studies*. Oxford: Oxford University Press, 2019. [Online edition https://doi.org/10.1093/oxfordhb/9780199982295.001.0001.]

Eliot, T. S. *The Speech Lab Recordings*. 1935. Edited by Chris Mustazza. Philadelphia: PennSound, 2023. https://writing.upenn.edu/pennsound/x/Eliot.php.

---. *The Waste Land*. 1922. Project Gutenberg, 1998. https://www.gutenberg.org/cache/epub/1321/pg1321-images.html.

Feaster, Patrick. “‘Uncle Josh Stories’: Cal Stewart’s *Other Book* (1924).” *Griffonage-Dot-Com* 7 January 2021. https://griffonagedotcom.wordpress.com/2021/01/07/uncle-josh-stories-cal-stewarts-other-book-1924/.

Goffman, Erving. *The Presentation of Self in Everyday Life*. New York: Knopf Doubleday Publishing Group, 2021.

Liberman, Mark. “God’s Own Englishman with a Tube Up His Nose.” *Language Log* 23 April 2007. http://itre.cis.upenn.edu/~myl/languagelog/archives/004436.html.

Mori, Masahiro, Karl F. MacDorman and Norri Kageki. “The Uncanny Valley [From the Field].” *IEEE Robotics & Automation Magazine* 19.2 (2012): 98-100.

McEnaney, Tom. “This American Voice: The Odd Timbre of a New Standard in Public Radio.” *The Oxford Handbook of Voice Studies*. Edited by Nina Sun Eidsheim and Katherine Meizel. Oxford: Oxford University Press, 2019. [Online edition https://doi.org/10.1093/oxfordhb/9780199982295.001.0001.]

Mustazza, Chris. “In Search of the Sermonic: Machine Listening and Poetic Sonic Genre.” *Computational Stylistics in Poetry, Prose, and Drama*. Edited by Anne-Sophie Bories, Petr Plecháč and Pablo Ruiz Fabo. Berlin: De Gruyter, 2023. 87-98.

---. “The Voices We Do.” 2022. https://www.youtube.com/watch?v=iV9INJd0Tlo&t=3s.

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

Patel, Nilay. “AI Drake just Set an Impossible Legal Trap for Google.” *The Verge* 19 April 2023. https://www.theverge.com/2023/4/19/23689879/ai-drake-song-google-youtube-fair-use.

PennSound. Philadelphia: PennSound. https://writing.upenn.edu/pennsound/.

Spinelli, Martin and Lance Dann. *Podcasting: The Audio Media Revolution*. London: Bloomsbury Academic, 2019.

Shanfeld, Ethan. “Ghostwriter’s ‘Heart on My Sleeve,’ the AI-Generated Song Mimicking Drake and the Weeknd, Submitted for Grammys.” *Variety* 6 September 2023.

https://variety.com/2023/music/news/ai-generated-drake-the-weeknd-song-submitted-for-grammys-1235714805/.

Shklovsky, Viktor. "Art, As Device." *Poetics Today* 36.3 (2015): 151-174.

Veltman, Chloe. "When You Realise Your Favorite New Song Was Written and Performed by…AI." *NPR* 21 April 2023. https://www.npr.org/2023/04/21/1171032649/ai-music-heart-on-my-sleeve-drake-the-weeknd.

Wagner, Petra and Simon Betz. "Effects of Meter, Genre and Experience on Pausing, Lengthening and Prosodic Phrasing in German Poetry Reading." *Proceedings of INTERSPEECH 2023*. 2538-2542. doi: 10.21437/Interspeech.2023-1443.

West, Candace and Don H. Zimmerman. "Doing Gender." *Gender and Society* 1.2 (1987): 125-151.