7

Rewriting Algorithms for Just Recognition

From Digital Aural Redlining to Accent Activism

Nina Sun Eidsheim

INTRODUCTION

On the evening of the first day of the "Thinking with an Accent" virtual symposium, my three-person family sat down on the floor of my son's room to play Monopoly with "voice banking," a version of the game that was new to us.[1] It promised that we could "talk to Mr. Monopoly and he responds."[2] Originally called The Landlord's Game, Monopoly was designed in the early 1900s to expose the structural inequity between landowners and renters. As Eula Biss tells the story, its inventor, Elizabeth Magie Phillips, had hopes the game would teach kids about the injustice of our economic system. Later repackaged by Charles Darrow, who also diffused some of the economic messaging, the game, with its underlying continuous loop of play concept taken from the Oklahoma Kiowa people, would instead pit children against parents in the practice of rapacious landownership.[3] The endlessly updated versions of the game, with new color schemes and characters and branding related to pop culture themes, such as the blockbuster Disney movie *Frozen*, offer yet another opportunity for overbuying. Across all the different versions of games I've seen over the last thirty years, the concept of Monopoly remains the same. Each player picks a character that moves around the board and buys streets, houses, and hotels, or pays rent for landing on them, all which is determined by dice and cards that give instructions such as "go to Such-and-such Street."

The voice banking game's key material distinction from other versions is that it uses no paper money. Rather than one player taking on the role of banker, each begins with an amount automatically "deposited" in their account, and each keeps track of their voice-triggered earnings, purchases, debts, and transactions.[4] For example, when a player wants to buy a street, they click on their character's button,

which speaks to notify Mr. Monopoly. As the evening progressed, I would learn that this new Monopoly game encapsulated a number of issues, including how early in a child's development attitudes about voice are absorbed. Furthermore, the game shows how human listening practices are programmed into digitized vocalizing and listening tools, which constitute one of the many sites of algorithmic racism perpetuated through everyday technology. Family game night became an illustration of the very dynamics I study: the ways in which power is wielded through listening—here, listening programmed into zeros and ones.

The first few rounds were uneventful: we rolled the dice and advanced our characters. Then my son purchased the first street. It was fun hearing him interact verbally with the voice banker, and he was thrilled with these vocal exchanges. But as my husband and I began to interact with the voice banker, things took an unexpected turn. Mr. Monopoly interpreted most of my instructions correctly but repeatedly misunderstood the street names and instructions my husband gave. At first we laughed, but having to repeat the instructions began to take up too much time, interrupting the flow of the game. We moved along, haltingly, and after a short time our son won, far ahead of us. Having won so hugely over his parents, my son wanted to play again immediately. This time he quickly stepped in. Rather than waiting for us to attempt our own transactions with the voice banker, which would force him to sit through our multiple repetitions of "purchase Oriental Avenue," he began talking for us—and the electronic banker always understood him. In the end he was playing the entire game for all of us, controlling our assets to his own advantage.

Witnessing this in awe, I let him go a bit further than I normally would. On our son's bedroom floor, we inadvertently played out one of the themes that had been discussed at the symposium that day: the intricacies, challenges, and power dynamics of performing with and listening to accents—with my little interracial nuclear family exemplifying the classic immigrant experience. I grew up in Norway and have a Norwegian accent when speaking English, but I have lived in the United States for over twenty years. My husband grew up in Colombia and has a Colombian Spanish accent when speaking English. He has lived in the United States for a much shorter amount of time and, compared to me, has many more opportunities to speak his native tongue on a daily basis. Our son grew up speaking the three languages of his family.

Mr. Monopoly's listening algorithm showed me something that our son's corrections and good-natured jokes about our accents had not. In its new "Monopoly meets voice recognition" version, the game performs the boundary around accepted accents. When notifying players that "Kitty has 500 Monopoly dollars in her account," its prerecorded phrases perform what most people would hear as a nonmarked American English accent. The game's tagline—"Control it all with the power of your voice"—refers to more than Monopoly transactions.[5]

DIGITAL AURAL REDLINING: WHOSE ACCENT IS ACCENTED?

Accent, like skin color and hair texture, is universal. Everyone has some kind of skin hue and hair texture. However, as Black feminist studies scholars, including Rasul Mowatt, Bryana French, and Dominique Malebranche, have noted, certain skin colors, hair textures, and accents are framed as hypervisible or hyperaudible.[6] They are perceptually *accentuated*, made into markers and sources of Otherness.[7] Only some accents are *accented* in their reception, to invoke an alternate meaning of accent: *emphasized*. Within a broader linguistic context,[8] *accent* is defined as a "distinct mode of pronouncing a language"; it is therefore something every speaker displays.[9] But the colloquial use of the word suggests that only some speak *with an accent*, and some even with a *strong accent*. In other words, not all accents are accented. The assessment that accentuates some accents is added during the process of listening.

Shifting our attention from vocalizing to listening, I propose to consider this active form of listening—which carries out the work of marking certain voices—as *accented*, and indeed as accented *listening*.[10] Thus, although all voices are accented, active listening marks further accentuation. That is, as voices are always already accented, the process of further marking certain voices gives rise to *accented accents*.[11] The status of the accented accent is by definition unstable, as it is produced by listening communities that reproduce, and indeed solidify, specific vocal and listening configurations. To get to the heart of the power wielded through listening, each configuration requires a specific analysis—some examples of which are offered in this volume.

In this chapter I am interested in the adaptation of certain assumptions and listening practices into algorithms, and their proliferation through digital media and digital tools. Thus, cross-feeding my own work on voice, race, and power with that of internet scholar Safiya Noble, the digitized voice and the digitized listening to voice become inflections of what she succinctly describes as "the power of algorithms in the age of neoliberalism and the ways those digital decisions reinforce oppressive social relationships and enact new modes of racial profiling." Adapting Noble's apt term for this phenomenon, "technological redlining," I describe listening that defines certain ways of voicing as accents as *aural redlining*.[12] This is an example of my plea for us to "listen to listening"—to begin to note the specific ways in which both humans and machines perform power through listening practices.[13] Or, in June Jordan's unambiguous formulation, we must understand how "white power uses white English as a calculated, political display of power to control and eliminate the powerless."[14]

Similarly, Noble reminds us that search engines are not neutral—for example, when they autofill the search field when a user types "Black girls . . ."[15] Such relationships, defined by those in power, are also quantified in voice and listening

algorithms.[16] Allison Koenecke and her coauthors note that the language models used to develop commercial automated speech recognition (ASR) systems are not publicly available. In lieu of these specific language models, Koenecke's team of mathematicians, engineers, computer scientists, and linguists chose to work from the assumption that it is "likely that these systems use language models that have similar statistical properties to state-of-the-art models that are publicly available, like Transformer-XL, GPT, and GPT-2. [They] thus examine potential racial disparities in these three models, using the publicly available versions that have been pretrained on large corpora of text data."[17] As a singer and a humanities scholar, I do not attempt to address the underlying ASR systems but rather consider the underlying values performed through listening practices that tacitly shape digital application development. I'm guided by the conviction that naming these elements can help to diagnose systemic issues in current voice-based technologies and to counter the belief that digital environments may be more neutral than people.[18]

I think of listening practices translated into algorithms as *digital aural redlining*, and of practices that oppose this redlining as *digital aural jamming*. Largely associated with the real estate and lending markets, redlining disproportionately saddles Black and Latino people (especially those with underprivileged socioeconomic status) with higher interest rates, fees, and banking premiums, putting them at an economic disadvantage.[19] In other words, the term describes practices that discriminate against individuals and communities based on race and class regardless of individual character or credit score. *Aural redlining* captures a systematic listening practice that, first, others people based on their accents; second, makes them hyperaudible or inaudible; and, third, due to the ubiquity of such othering digital voice and listening tools, disadvantages individuals economically.

In a study that coined the term "linguistic profiling," John Baugh showed that housing rental practices relied on discriminatory accent cues in decision-making processes.[20] The study's potential renters would call about an advertised unit. Callers with a presumed alterity—based on their accent—would be rejected. This study demonstrated that listeners were certain about their racial or ethnic assessment based on a voice alone, that is, based on a brief phone conversation. It also showed that although individuals could be approved if the listener assessed an unaccented accent, the same applicant could be rejected in person if his or her body did not also prove unmarked. Building on Noble's and Baugh's work, the term *aural redlining* expands redlining to cover listening practices applied to voices, including timbre, more broadly, and it includes *digital aural redlining*, speech- and voice-based profiling practices applied to the digital domain. Aural redlining may take place in all vocalizing and listening configurations, from live situations in which vocalizer and listener are together and can see each other to broadcasts and recordings with or without live or static images of the vocalizer. "Digital" denotes listening practices that have been quantified into code that carries out practices such as ASR, which is used in technologies such as voice-to-text, virtual assistants, and

automatic captioning. While these examples are quite different from one another, they draw on collective vocalizing and listening practices to, in Jonathan Sterne's and Mehak Sawhney's apt words, "datafy and classify the human voice."[21]

Digital aural redlining can take different forms. I will discuss two here. The first type of digital aural redlining describes a situation in which a particular accent is assumed based on nonvocal cues. Listeners perceive a certain accent based on the way in which they read the speaker's race, such as visually. In other words, what listeners see affects how they hear accents—or we could say that listeners *see* accents.[22] The second type of digital aural redlining describes the digital acoustic shadow: when a person is, in effect, rendered inaudible because their accent prevents or precludes them from effectively using many voice-based technologies.[23] (This phenomenon is not unique to digital aural redlining. However, for me it is very helpful to examine the general process of aural redlining as it is defined and formalized in order to be re-created through algorithms.)

SYNTHESIZING THE ACCENTED ACCENT

The vocal synthesis software system Vocaloid was first released in 2003 to great fanfare.[24] The first two products, LOLA, LEON, and many later versions are commercial music software described as "voice fonts." Just as MIDI instrument packages allow users to play a melody using the sounds of different instruments—first a piano, say, and then a banjo—a Vocaloid voice can be used to "sing" a melody. The major difference between a MIDI instrument and a synthesized voice is that a vocal sound is put together in such a way that it will provide not only pitch and timbre but also the various sounds necessary to form consonants, vowels, and diphthongs, which are needed to express lyrics. Just as users can transform their text with the click of a button from one font to another, musicians can have different Vocaloid voices to choose between when recording a song. Vocaloid was hailed for rethinking and reframing this technology from technologically advanced software to *backup singers in a box*. Up to this point, vocal synthesis had been advertised in terms of computational power, but instead of touting the program's high-tech bona fides, LOLA and LEON were advertised as racialized characters through blackface iconography. Both LOLA and LEON are represented in close-cropped profile images with protruding full lips. As a stock character returns in minstrel repertoire, the same picture is used for both LOLA and LEON. For LOLA, the designer simply mirrored LEON's blue-tinted image and colored it red.

Vocaloid's synthesis was created from source recordings—short recorded phonemes on multiple pitches—which were then combined via the synthesis algorithm to form any words a user typed into the program (that's the vocal synthesis part). Vocaloid's synthesis combines recorded phoneme samples into a seamless string of notes that sound words in melodic sequences.[25] Users can input notes using the visual interface, or they can use a MIDI keyboard to play a melody that

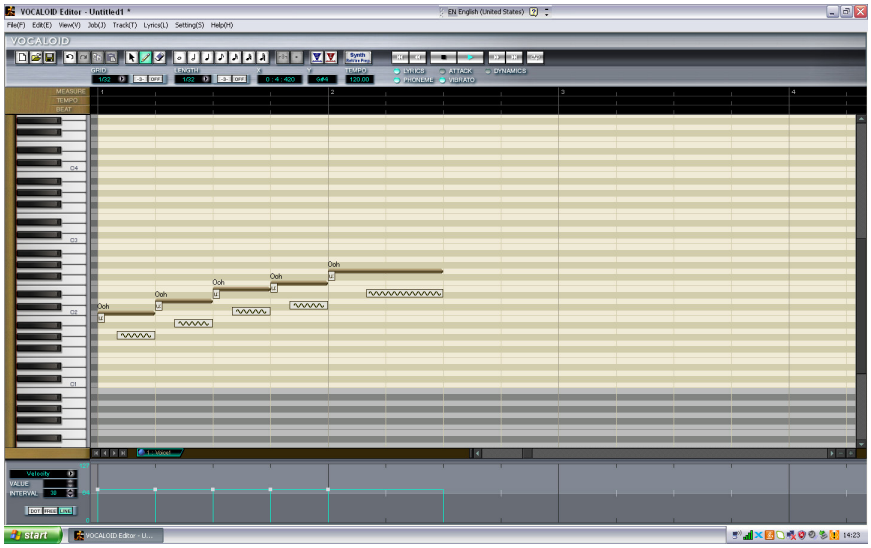

FIGURE 7.1. Screenshot of the Vocaloid interface.

is recorded onto the visual interface. The lyrics can then be inserted underneath the music notation (see figure 7.1). In electroacoustic music terms, Vocaloid may be considered "hybrid vocal synthesis" because it uses basic sonic material from the phoneme recordings, whereas "complete sound synthesis" does not use sound samples. Vocaloid relies on synthesis techniques in order to combine and alter the sounds of the samples.[26]

Despite these comprehensive efforts to present a Black soul singer, many of LOLA's users did not hear her voice as a soul voice and/or as Black. User RobotArchie wrote on parent company Zero-G's internet message board, "Do we have a British soul singer with a Japanese accent who lisps like a Spaniard? Eesa makea me tho unhappy."[27] Heatviper chimes in with, "Hello . . . I think LOLA works great for mondo/ mournful/giallo morricone style tracks using vowels. . . . Wordless soulful vowels are nice."[28] Jogomus asks for advice: "My LOLA sounds a little bit like a 'big Ma'—what can I do, [so] that she sounds a little bit neutral?" Another user named hk suggests lowering the "Gender Factor" value.[29]

What happened here? The developers, based in Britain, had chosen Black singers to sample as the source of the synthesis. However, in talking with them I learned that the male voice was a British-born singer and the female voice was a Jamaican-born woman. As professional vocalists, the singers were both adept at performing soul idiomatically, including timbre and word pronunciation. However, when they recorded thousands of syllables outside the context of a musical style, I hypothesize that they did not do so with an accent associated with soul, but rather with the accents of their mother tongues.[30] The singers were selected to

provide source data based on a judgment about their visual presentation as Black rather than on an aural assessment of both soul and their particular accents.

Indeed, the process of providing source material for Vocaloid's voice banks does not take place within the context of a musical genre. The source sound is a carefully recorded bank that forms the sonorous basis for pronouncing the 3,800 possible vowel and consonant combinations it can voice within the English language. In other words, the source syllables are recorded out of context. Within the conventions of soul singing, the syllable "ma," as part of the word man, would be pronounced with a diphthong and would potentially be drawn out, depending on the prosody. Outside the context of soul style, native speakers of British and Jamaican English would sing the syllable "ma" differently. This means that every voicing takes place within what we may think of as an aesthetic genre. Such a genre can be chosen, and is very likely to be chosen, when singing within the context of a vocal musical genre such as soul.

LOLA and LEON were built on the premise of vocal racial essence, with no regard for the fact that English-speaking and -singing Black singers around the world grow up with myriad accents. Further evincing an essentialist attitude toward voice and race, as noted above, the graphic design featured on the software boxes echoes blackface imagery. Instead of orienting listening for a rich geographically and culturally specific musical style that arose within a specific community, within specific social and economic pressures, within the complex history of the African diaspora, *soul* was reduced to monolithic blackness and accent. As I've followed Vocaloid's development, Zero-G has repeatedly shown that a technology that could offer an expansion of the vocal and listening imaginary is instead primarily recruited to re-create, and seemingly to confirm, essentialized categories.

LOLA and LEON were introduced seventeen years ago—truly ancient in terms of the voice and listening consumer technology found on today's smart phones and computers. Because vocal synthesis technology and the algorithms that attempt to make sense of our voices are no longer technology-forward choices but nearly unavoidable presences in our lives, what is and will be the sound of the voices we will associate with sophisticated knowledge and technology? How will we have been conditioned to hear voices through generations of vocal technology built on voice models that assume and reproduce accent alterity? And which of us do digitized voice and listening technology have the capacity to hear?

DIGITAL ACOUSTIC SHADOW

If many voices are singled out through alterity or accent hyperaudibility, as Vocaloid attempted with LOLA and LEON, my family's Monopoly anecdote captures the flip side—the phenomenon I call the digital acoustic shadow. Sun rays can be blocked by solid objects, resulting in areas left in shadow. When sound waves meet solid obstacles like pillars, corners, or overhangs (such as a balcony),

certain frequencies can be attenuated, causing what are known as *acoustic shadows*. We may extend the phenomenon of the acoustic shadow in order to understand the muting of certain voices: those deemed less legible due to visual cues interpreted as alterity, and those who are misheard because training materials do not include such voices.[31] Furthermore, algorithms based on similar assumptions about voice, accent, and race can create *digital acoustic shadows* within digital tools. As noted, hypervisibility's constant companion is invisibility, which marks accented voices in either case. The digital acoustic shadow's veil is hyperaudibility's constant companion.

Some technologies may be broadly categorized as "listening to" and "analyzing" voice and speech. Their purposes range from transcription (text-to-speech) and prompts to action (voice bank Monopoly, automated phone services) to assessments of, for example, intelligence and skill level (AI hiring systems). Those in this industry making such products will point to the numbers, which are moving in a positive direction.[32] For example, Google's word error rate had decreased from 23 percent in 2013 to 8.5 percent in 2016, reaching 4.9 percent only a year later, in 2017.[33] But the question is not whether the technology has improved—even improved tremendously—in a short amount of time, but what hides behind the uniformity of these improvements. For example, if 4.9 percent is the average error rate, what is the rate for a white male Midwestern speaker versus a Black male from the South? In an interview addressing the question of differing user experiences, John Baugh noted that "Microsoft, the most accurate system, had a 27 percent error rate for Black speakers and 15 percent for white speakers; Apple, the lowest performer, missed the mark for 45 percent of words from Black speakers and 23 percent of white speakers—it has limitations in its scope."[34]

The algorithms that create this error rate underpin the product development of the largest technology companies in the United States. These algorithms are integrated into products that permeate everyday and work-life technology, making the ramifications of unequal access—redlining—an urgent matter. Comparing two thousand voice sample transcript results based on recorded interviews with African Americans and white speakers, Koenecke and her coauthors tested commercial automated speech recognition developed by Amazon, Apple, Google, IBM, and Microsoft.[35] Their sample corpus was collected in five U.S. cities and consisted of interviews with forty-two white speakers and seventy-three Black speakers of mixed age and gender. Across the technologies, they found on average that the error rate was 35 percent for African Americans compared to 19 percent for white speakers.[36] They attributed this error rate to a lack of representation in training data. This "gap in the acoustic models" suggests "that the systems are confused by the phonological, phonetic, or prosodic characteristics of African American Vernacular English rather than the grammatical or lexical characteristics. The likely cause of this shortcoming is insufficient audio data from black speakers when training the models."[37]

In contrast, "dialectal language is increasingly abundant" on social media, yet "few resources exist for developing NLP [natural language processing] tools to handle such language," Su Lin Blodgett, Lisa Green, and Brendan O'Connor note.[38] In a paper a year later they noted (in scientists' cautious language) that "current systems sometimes analyze the language of females and minorities more poorly than they do [that] of whites and males. We conduct an empirical analysis of racial disparity in language."[39] Unsurprisingly, in automatic caption software, "the lowest average [speech recognition] error rates were for General American and white talkers, respectively. . . . [T]he higher error rate [for] non-white talkers is worrying, as it may reduce the utility of these systems for talkers of color."[40] In other words, these software systems rely on algorithms that cannot properly process certain accents. The string of code that is unable to process selected accents represents the obstacle that casts a digital acoustic shadow, excluding potential users from meaningful use of the technology. As a case in point, in my family, the person with the accent under the darkest digital acoustic shadow lost the game. In playing, each person had to use significantly different resources in order to simply participate—that is, to be understood by the technology—and each turn to play was accompanied by the anticipation of that challenge. Hesitation and reduced interest in the game were the results of these obstacles. While the stakes were not high in this context, it helps to explain the overall dynamic and the discriminatory negative outcome, both in the end result (losing the game) and in some players' detachment from engagement, when we see that voice-based technology fails some users while favoring others.

RAMIFICATIONS OF THE DISCREPANCY IN ERROR RATE

The discrepancy between my son's error rate and his dad's mirror real life with disconcerting accuracy. The person with the lowest error rate earned the most property and money. In real life, what it means to lose the game due to aural redlining depends on specific technologies and on the circumstances of their use. For example, speech-to-text software is used in consumer technology such as smartphones, which are increasingly necessary in many work situations, including many jobs that require employees to use phone apps. A specific accent's interaction with the technology required to carry out a job may prevent groups and individuals from performing equally, which may lead to lower work performance and fewer chances for promotion and mobility. And if voice technology software is used to screen candidates, others may not be selected for a job at all. In the same way that redlining practices in real estate prevent an entire community's economic advancement, we can see that digital aural redlining can have a similar effect.

Within the court system, the situation is equally concerning.[41] Writing about human court reporters, Maarten Sap and his coauthors have shown that

"annotators' insensitivity to differences in dialect can lead to racial bias in automatic hate-speech detection models, potentially amplifying harm against minority populations."[42] Specifically, in investigating "toxic language identification tools" they found that "the task is especially challenging because what is considered toxic inherently depends on social context (e.g., a speaker's identity or dialect)." Given the racial history of the United States, "phrases in the African American English dialect (AAE) are labelled by a publicly available toxicity detection tool as much more toxic than general American English equivalents."[43] And as Aylin Caliskan, Joann Bryson, and Arvind Narayanan show, "cultural stereotypes propagate to artificial intelligence (AI) technologies in widespread use."[44] This work, they argue, "has implications for AI and machine learning because of the concern that these technologies may perpetuate cultural stereotypes." Their research suggests that if "we build an intelligent system that learns enough about the properties of language to be able to understand and produce it, in the process it will also acquire historical cultural associations, some of which can be objectionable."[45] And as Taylor Jones and his coauthors note, "Any solution to the (narrow) transcription problem must take into account the broader problem of harmful linguistic ideologies with common-currency anti-black stigma, bias (both conscious and not), and a court system that is the accumulated product of four centuries of white supremacy."[46] Unsurprisingly, research across consumer and professional speech-recognition software shows that aural redlining permeates this technology sector.

TO BE JUSTLY RECOGNIZED: AURAL-REDLINE JAMMING AS ACCENT ACTIVISM

Not hyperaudible, not inaudible, but, to quote Goldilocks, "just right"—the unaccentuated voice ideal is confirmed and strengthened by both analog and digitized voices that perform this self-fulfilling fantasy.[47] As a recent *New York Times* article on accent coaches and their clients noted, "Actors, or their agents or managers, find her because they either have booked a role that demands a certain sound or aren't booking anything because they don't sound a certain way. They are often hoping to achieve that general American sound to break in or refashion their career for the Hollywood market."[48] To find work or to move beyond typecasting, actors with some accents take on additional voice training to replace their accent with what is considered a normative one. This cycle confirms which voices are dominant in movies.[49]

What actors and casting agents alike have in mind when they seek training and voices for characters, respectively, might be something like what the team behind iPhone's Siri does: "The first phase is to find a professional voice talent whose voice is both pleasant and intelligible and fits the personality of Siri."[50] According to one industry analyst, "[Apple recruiters are] listening for some ineffable sense of helpfulness and camaraderie, spunky without being sharp, happy without being

cartoonish."[51] In their minds' ears, producers listen for which accents will fulfill such descriptions—and both Hollywood and Siri's team shy away from anything that could be perceived as an accented accent.

This chapter is an accentuated plea. While there is not one solution to prejudice and power imbalances, I do have a singular wish for all of us: to be *justly recognized*. To be justly recognized always stands firm against the seemingly innocuous *just right*. It refuses accented listening. What is the difference between *just right* and *justly recognized*?

Recommendations from researchers such as Koenecke include "using more diverse training data sets that include African American Vernacular English."[52] Improvements to the underlying acoustic models used by the ASR systems are vital. Improving the training data set could potentially move speech and acoustic patterns out of the acoustic shadow. Further, "developers of speech recognition tools in industry and academia should regularly assess and publicly report their progress along this dimension."[53] In technology development and data collection, a wider range of voices should be present from a tool's inception. As the error rate decreases in one area it may increase in others, evening out the user experience. But what is the data set tipping point at which voices will no longer be divided into those deemed to exude the qualities of "helpfulness and camaraderie, spunky without being sharp, happy without being cartoonish" and those that are not selected, or are heard as expressing opposing qualities? While the work of opening voice-driven technology to a much broader range of accents is clearly needed, it is not the solution.

A vocal assessment that *recognizes justly* is diametrically opposed to an assessment of whether an accent is *just right*. That which is *just right* is established by listening from a position of power, with the particularities set by time, place, and other circumstantial forces. A voice is deemed just right (or simply wrong) by outside forces based on a static and monolithic understanding of the person behind it. When instruments are attuned to capture just right, just right can also be used for surveillance.[54] While I am not prepared to offer a series of concrete steps—it will take a broad range of scholars, developers, artists, users, and activists to suggest, test, reject, and experiment with specific solutions—I know that to be justly recognized is to be recognized in relationship to oneself and to the multiplicity of histories and communities that we constantly adopt, reject, and form within multiple relationships. To be recognized justly is to retain protections and human rights.[55] Listeners who recognize justly afford each voice its multiplicity, including its humanity. In this way, just recognition makes clear that the hyperaudible and the inaudible, or the accented accent and voices veiled in acoustic shadows, are human- and (human-through-)machine-created fantasies.

As listening ability is not theoretical but is formed through practice, I have thought a lot about what the path toward just recognition might include. I think one component could be accent activism in the form of *aural redline jamming*.

Here, the experimenting listener both performs the assessment and experiences the impossibility of an operation such as *just right*. I take this term from the devices and apps known as "speech jammers," which record a person's speech and replays it with a slight delay, "jamming" the speaker's ability to maintain their train of thought. While perhaps tongue-in-cheek, this device, which won the Ig Nobel Acoustics Prize in 2012, is advertised for its ability to "inhibit" a person's speech; it is ironic, indeed, that the speaker is "jammed" with their own voice and words.[56] Aural redline jamming is similar in that it uses the same technology in an extreme way to dilute or jam comprehension of whichever accent is accented within a given context.

Users decisively rejected the hype that the vocal syntheses LOLA and LEON sounded like American-accented soul singers. Rather than tweaking the sound within the software to reach toward whatever sonic image users might have of idiomatic soul singers, the users instead jammed the system, creating songs that were much faster than the highest recommended BPM to sound legible and writing in Japanese, a language the phonemes were not intended for.[57] Hence, although the software was originally intended to replace live singers, users used LOLA and LEON to sound nonlocatable accents, jamming the built-in organizing principle. Moreover, the same technology that mistook race for accent later featured voice artist Misha, who insisted on basing her voice bank, Vocaloid Ruby, on her Latina identity—jamming Vocaloid's foundational premise.[58]

While, as my earlier work has shown, vocal and listening practices have always served to perform power, Noble notes that "discrimination is embedded in computer code and, increasingly, in artificial intelligence technologies that we are reliant on, by choice or not." Indeed, she warns that "we are only beginning to understand the long-term consequences of these decision-making tools in both masking and deepening social inequality."[59] While each individual voice has always been shaped through a deeply social and collective process and has mirrored and reinforced existing inequalities, is it challenging to remember the human hand in algorithms. For example, firms that use AI to screen job candidates with the belief that such tools will be less biased are actually using technology created by biased humans. Not only will the technology perform the same biases, but it will perform them on a larger scale, often with no option to "press 0 for the operator."

In other words, "part of the challenge of understanding algorithmic oppression is to understand that mathematical formulations to drive automated decisions are made by human beings."[60] This means that voice synthesis, voice recognition software, and transcription algorithms are not simply part of a system of neutral calibration of digital-audio information. Instead, these technologies were developed by people who heard voices and understood accents in specific ways, and then re-created that reality. Each smartphone voice tool has been created by a string of subjective decisions, as were LOLA and LEON. In the same way that Kodak film was calibrated for white skin color, voice and listening technologies will carry

over the social biases of earlier vocal categorizations and normalized listening conventions. Digital vocal technologies as we know them in the third decade of the twenty-first century, then, are artifacts of a particular listening culture. Noble predicts that "artificial intelligence will become a major human rights issue in the twenty-first century."[61] To think about accented accents is to think about how, in a democracy, the right to be recognized justly is tied to the impact of listening practices and aural representations of voices in the acoustic, analogue, and digital realms.

ACKNOWLEDGMENTS

Much gratitude goes to the editors of this volume, especially Pooja Rangan and Pavitra Sundar, for detailed and crucial feedback on this chapter. A draft of the chapter was read at Matters of Voice, a Marta Sutton Weeks Research Workshop at the Stanford Humanities Center. Many thanks to the participants, and especially to Charles Kronengold, who served as the respondent, and María Gloria Robalino, who organized the event. Finally, thank you to Tildy Bayar, Iris Blake, and Alexander Khalil for providing important feedback, to Ramona Gonzalez for research help, and to Matthew Blackmar for aiding with citation matters.

NOTES

1. "Thinking with an Accent: Through Voice, Across Media" virtual symposium, May 1–3, 2020, Amherst University.

2. This line can be found at the bottom of the box.

3. Eula Biss, *Having and Being Had* (New York: Penguin Random House, 2020).

4. "Monopoly Voice Banking Electronic Board Game," https://monopoly.hasbro.com/en-us/product/monopoly-voice-banking-electronic-family-board-game:97BC561B-145E-42CA-AB15-917F2E2FD5BA (accessed April 1, 2020).

5. There are eerie connections between a game based on buying, selling, and renting property in which success is tied to degree of perceived accent and Cheryl I. Harris's work on whiteness as property. This particular Monopoly game epitomizes the ways in which the voice is tied to power and to property. Cheryl I. Harris, "Whiteness as Property," *Harvard Law Review* 106, no. 8 (1993): 1707–91. Thank you to Iris Blake for reminding me of this connection.

6. Rasul A. Mowatt, Bryana H. French, and Dominique A. Malebranche, "Black/Female/Body Hypervisibility and Invisibility," *Journal of Leisure Research* 45, no. 5 (2013): 644–60.

7. For discussions regarding the relationship forged between the categories of language and race, see H. Samy Alim, John R. Rickford, and Arnetha F. Ball, *Raciolinguistics: How Language Shapes Our Ideas about Race* (New York: Oxford University Press, 2016).

8. Within a narrower context, *accent* can refer to an accent within a word, i.e., the stress placed on a syllable.

9. "Accent," Oxford Languages for Google, January 7, 2021, www.google.com/search?q=accent&rlz=1C5CHFA_enUS891US892&oq=accent&aqs=chrome..69i57j0i433l2j46i199i291i433j0i433l2j69i61l2.1589j0j9&sourceid=chrome&ie=UTF-8.

10. In general, to better understand the dynamics around voice, I am convinced that we need to shift our attention to listening through what I call "the acousmatic question," "Who is this?," which

allows us to actively name and define the voice. This active form of listening creation also takes place in terms of other constructed sociocultural categories. I discuss this construction with regard to gender, calling it "audience drag performance," in my *The Race of Sound: Listening, Timbre, and Vocality in African American Music* (Durham, NC: Duke University Press, 2019), 1–38, 91–114.

11. The concept of the "accented accent" was developed in conversation with Anita Starosta, "Accented Criticism: Translation and Global Humanities," *boundary 2: An International Journal of Literature and Culture* 40, no. 3 (2013): 163–79.

12. Safiya Umoja Noble, *Algorithms of Oppression* (New York: New York University Press, 2018), 1.

13. Eidsheim, *The Race of Sound*, 18. In addition to race, regionality (when it comes to American English) and national forms of English (when it comes to the English language) can also play a role. It is well known that Scottish English is not represented thoroughly in voice data sets, as can be seen in the skit "Voice Recognition Lift" from *BBC Scotland*, www.youtube.com/watch?v=sAz_UvnUeuU (accessed August 19, 2022).

14. June Jordan, "White English / Black English: The Politics of Translation," *Moving Towards Home: Political Essays* (London: Virago, 1989), 29–40.

15. Noble, *Algorithms of Oppression*, 64–109.

16. I have written extensively about the kind of listening to voices that focuses on bodies that perform. See my *The Race of Sound*; "Maria Callas's Waistline and the Organology of Voice," *Opera Quarterly* 33, nos. 3–4 (2017): 249–68; "The Micropolitics of Listening to Vocal Timbre," *Postmodern Culture* 24, no. 3 (2014), http://muse.jhu.edu/journals/postmodern_culture/v024/24.3.eidsheim.html; "Voice as Action: Towards a Model for Analyzing the Dynamic Construction of Racialized Voice," *Current Musicology* 93, no. 1 (2012): 9–34.

17. Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel, "Racial Disparities in Automated Speech Recognition," *Proceedings of the National Academy of Sciences* 117, no. 14 (April 7, 2020): 7684–89.

18. For the power of naming, see Linda Adler-Kassner and Elizabeth Wardle, eds., *Naming What We Know: Threshold Concepts of Writing Studies* (Boulder: University Press of Colorado, 2015).

19. Eula Biss, "White Debt: Reckoning with What Is Owed—and What Can Never Be Repaid—for Racial Privilege," *New York Times*, December 2, 2015, www.nytimes.com/2015/12/06/magazine/white-debt.html (accessed December 2, 2015). The streets purchased during the board game that provoked much of my thinking in this chapter also have a curious connection to redlining. In the early 1930s, Atlantic City realtor Jesse Raiford affixed property prices to Monopoly that reflected the city's residential segregation. The cheapest streets, such as Baltic and Mediterranean Avenues, were in Black neighborhoods, while the more expensive streets, like Park Place and Boardwalk, were in affluent white neighborhoods. To read more about the complex history of Monopoly, see Mary Pilon, *The Monopolists: Obsession, Fury, and the Scandal Behind the World's Favorite Board Game* (New York: Bloomsbury, 2015). Thanks to Robert Fink for the conversation that led me to look into the street price differentiation.

20. John Baugh, "Linguistic Profiling," in *Black Linguistics: Language, Society, and Politics in Africa and the Americas*, ed. Sinfree Makoni, Geneva Smitherman, Arnetha F. Ball, and Arthur K. Spears, 155–68 (New York: Routledge, 2003).

21. Jonathan Sterne and Mehak Sawhney, "The Acousmatic Question and the Will to Datafy: Otter.ai, Low-Resource Languages, and the Politics of Machine Listening," *Kalfou: A Journal of Comparative and Relational Ethnic Studies* 9, no. 2 (forthcoming Winter 2023).

22. This idea of accent beyond the sonic relates to a discussion in the introduction to this volume: poet Li-Young Lee "captures the workings of accent not only across senses but as that which crosses senses, which gives skin 'tones,' to use Rey Chow's term. The ear does not simply receive sound; it is a 'coloring ear' that shades the voice. Importantly, too, Lee hones in on accent as that which inflects encounters 'even before' meaning, irrespective of the speaker's identity, and prior to the act of interpretation" (3, this volume).

23. An acoustic shadow is a phenomenon in which sound is inaudible, even in close proximity to the listener, because of some obstacle that obstructs the sound waves.

24. The best-known Vocaloid is Hatsune Miku, who—like many anime characters—has large fanbase. Readers may be familiar with her 2014 hologram performance on *The Late Show*, which received considerable press. "Hatsune Miku on David Letterman! (FULL)," October 9, 2014, www.youtube.com/watch?v=lwJ1i5lCwoM (accessed November 1, 2014). To learn more about Miku in regard to vocal processing and perception, see my *The Race of Sound*, 115–50; Gretchen Jude, "Vocal Processing in Transnational Music Performances, from Phonograph to Vocaloid," PhD dissertation, University of California, Davis, 2018, 111–43; Nick Prior, "STS Confronts the Vocaloid: Assemblage Thinking with Hatsune Miku," in *Rethinking Music through Science and Technology Studies*, ed. Antoine Hennion and Christophe Leveux (New York: Routledge, 2021), 213–26.

25. For more information regarding the vocal synthesis system, see my *The Race of Sound*, 115–50.

26. Practically, what matters to amateur users who neither know nor care about these distinctions, and to a general public told that the voice it hears is a synthesized voice, is not the technical distinction between full and hybrid vocal synthesis. What matters is that they believe it is vocal synthesis.

27. RobotArchie, forum post on Vocaloid-User.net, "When Will Vocaloid Meet Your Expectations?," April 12, 2004, http://vocaloid-user.net/forum/general-vocaloid-discussion/when-will-vocaloid-meet-your-expectations. Although this link and those cited in the next two notes are currently dead, I have screenshots of their contents.

28. Heatviper, "Mike Oldfield," forum post on Vocaloid-User.net, January 19, 2006, http://vocaloid-user.net/forum/general-vocaloid-discussion/mike-oldfield. Dead link.

29. Jogomus et al., "2 Questions about Vocaloid," forum post on Vocaloid-User.net, August 5, 2005, http://vocaloid-user.net/forum/general-vocaloid-discussion/2-questions-about-vocaloid. Dead link.

30. The question of which accent is associated with soul is extremely complex and beyond the scope of this chapter. There are various lineages, some based on regionality, others on individual singers whose idiosyncratic pronunciation marked the genre. However, while the braided genealogy of the soul accent is contested and complex, those familiar with the genre would typically flag performances that do not exhibit any of the accents that are recognized as performing soul.

31. D. L. Rubin, "Nonlanguage Factors Affecting Undergraduates' Judgments of Nonnative English-Speaking Teaching Assistants," *Research in Higher Education* 33, no. 4 (1992): 511—31; John R. Rickford and Sharese King, "Language and Linguistics on Trial: Hearing Rachel Jeantel (and Other Vernacular Speakers) in the Courtroom and Beyond," *Language* 92, no. 4 (2016): 948–88.

32. When I presented this work in progress, many noted that given how digital recognition technology is used, it might be preferable to stay in the acoustic shadow. It is also worthwhile to note that the level of risk involved in being digitally tagged through one's image or voice is very different for different individuals. The solution is never hyperaudibility. Instead, I introduce below the concept of *justly recognized*, which seeks to capture privacy and human rights issues.

33. Emil Protalinski, "ProBeat: Has Google's Word Error Rate Progress Stalled?," *Venturebeat*, May 10, 2019, https://venturebeat.com/2019/05/10/probeat-has-googles-word-error-rate-progress-stalled/ (accessed January 13, 2021).

34. Jeff Link, "Why Racial Bias Still Haunts Speech-Recognition AI," *Built In*, July 26, 2020, https://builtin.com/artificial-intelligence/racial-bias-speech-recognition-systems (accessed January 13, 2021).

35. Koenecke, et al., "Racial Disparities in Automated Speech Recognition," 7686.

36. Koenecke, et al., "Racial Disparities in Automated Speech Recognition," 7685.

37. Koenecke, et al., "Racial Disparities in Automated Speech Recognition," 7688. It should be noted that most linguists who study African American Vernacular English do not distinguish as clearly between "phonological, phonetic, or prosodic characteristics of African American Vernacular English . . . [and] the grammatical or lexical characteristics" as Koenecke et al.; see, for example, Alim, Rickford, and Ball, *Raciolinguistics*. In fact, as Charles Kronengold aptly remarked in an email exchange,

"Koenecke et al. may be stepping into the same mistake-space as the Vocaloid designers, but making the opposite assumption (i.e., they assume disconnectedness between discrete phonemes and the systems that give them meaning where there's actually a connection, while Vocaloid assumes the connection when it's absent)." Personal correspondence, February 10, 2021.

38. Su Lin Blodgett, Lisa Green, and Brendan O'Connor, "Demographic Dialectal Variation in Social Media: A Case Study of African-American English," *arXiV* 1608.08868v1 (2016): 1.

39. Su Lin Blodgett and Brendan O'Connor, "Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English," *arXiV* 1707.00061v1 (2017): 1.

40. Rachel Tatman and Conner Kastern, "Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions," *Interspeech* 2017: 934. See also Rachel Tatman, "Gender and Dialect Bias in YouTube's Automatic Captions," *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, Valencia, Spain, April 4, 2017: 53–59.

41. See Mari Matsuda's groundbreaking work on accent and antidiscrimination, "Voices of America: Accent, Antidiscrimination Law, and a Jurisprudence for the Last Reconstruction," *Yale Law Journal* 100, no. 5 (1991): 1329–1407.

42. This echoes the patterns of imaging technology. Kodak famously standardized their Shirley cards test. Later, as Sara Lewis has noted, "You see it whenever dark skin is invisible to facial recognition software. The same technology that misrecognizes individuals is also used in services for loan decisions and job-interview searches. Yet, algorithmic bias is the end stage of a longstanding problem." Sarah Lewis, "The Racial Bias Built into Photography," *New York Times*, April 25, 2019, www.nytimes.com/2019/04/25/lens/sarah-lewis-racial-bias-photography.html (accessed April 26, 2019). Newer biometric identification technology such as vein matching or vascular technology also suffers from having been developed with a limited range of skin hues in mind. Colleagues at the University of California, San Diego, report that they were routinely unable to enter their office building, which used such identification technology. The problem was so severe that the identification technology had to be switched to key cards.

43. Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah Smith, "The Risk of Racial Bias in Hate Speech Detection," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* 2018: 1668–78.

44. Aylin Caliskan, Joann J. Bryson, and Arvind Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases," *Science* 356 (2017): 183.

45. Caliskan, Bryson, and Narayanan, "Semantics Derived Automatically," 186. Although this chapter clearly argues that stereotype is to be avoided in AI system development, the underlying issue that is at stake is the system of value and power for which accent (i.e., people) are placeholders, and the particular historical and cultural origins of a given accent are implicated insofar as they are implicated in a colonial, geopolitical power grab. Black feminist language activists such as June Jordan have argued this through essays and poetry; see, for example, Jordan, "White English / Black English."

46. Taylor Jones, Jessica Rose Kalbfeld, Ryan Hancock, and Robin Clark, "Testifying While Black: An Experimental Study of Court Reporter Accuracy in Transcription of African American English," *Language* 95, no. 2 (2019): e216–e252.

47. The unaccentuated voice ideal is of course not a specific accent but rather what is thought to be the ideal in a given context and at a given moment.

48. Art Streiber, "The Accent Whisperers of Hollywood," *New York Times*, July 20, 2017, www.nytimes.com/2017/07/20/magazine/accents-hollywood-dialect-coach.html (accessed July 21, 2020).

49. It should be noted that nonwhite actors are still typecast, regardless of accent.

50. Siri Team, "Deep Learning for Siri Voice: On-device Deep Mixture Density Networks for Hybrid Unit Selection Synthesis," *Apple Machine Learning Research*, August 2017, https://machinelearning.apple.com/2017/08/06/siri-voices.html#9. Siri's voice originated in a collection of voices recorded in 2005 and stored in a digital database/archive. Apple will not reveal the actual voice of Siri, but it has been widely accepted that Karen Jacobson (Australian Siri), Jon Briggs (British Siri), and Susan

Bennett (American Siri) are the sources of the original voices. All three are voiceover actors and musicians who auditioned for the spot. The voice database was originally part of a speech recognition system owned by Nuance Communications, which Apple licensed. These Siri voices have since been replaced by a much more complex vocal speech synthesis technology and new voice talent. Hannah Jane Parkinson, "Hey Siri! Meet the Real People Behind Apple's Voice-Activated Assistant," *Guardian*, August 12, 2015, www.theguardian.com/technology/2015/aug/12/siri-real-voices-apple-ios-assistant-jon-briggs-susan-bennett-karen-jacobsen (accessed September 19, 2020).

51. David Pierce, "How Apple Finally Made Siri Sound More Human," *Wired*, September 2017, 4.

52. Koenecke et al., "Racial Disparities in Automated Speech Recognition," 7685.

53. Koenecke et al., "Racial Disparities in Automated Speech Recognition," 7688.

54. It should be noted that in addition to facial recognition, the Chinese government uses personal voice samples to build a biometric portrait of its Uyghur population, although this surveillance system is aimed at specific individuals rather than general accent comprehension. In an interview with *Wired*, Amina Abduwayit shared that in addition to providing face scans and DNA and blood samples, she also had to give a voice sample to the police, which was used by the Chinese artificial intelligence giant iFlytek: "They gave me a newspaper to read aloud for one minute. It was a story about a traffic accident, and I had to read it three times. They thought I was faking a low voice." Isobel Cockerell, "Inside China's Massive Surveillance Operation," *Wired*, May 5, 2019, www.wired.com/story/inside-chinas-massive-surveillance-operation/ (accessed February 11, 2021). See also Pooja Rangan's forthcoming work on forensic speech analysis—also known as LADO (Language Analysis for the Determination of Origin) or, colloquially, "the accent test"—used to profile asylum seekers. "Accented Listening: A Hearing on Documentary's 'Audit,'" *Audibilities: On Documentary Listening* (forthcoming).

55. It makes sense that algorithmic vocal equality should arise simply from expanding the training corpus. However, as Sterne and Sawhney note, "we should understand that the expansion of machine learning systems is a kind of technological manifest destiny, a digital colonialism, a future that is sold as *necessary* and *logical* but in fact [is] backed by financial and military force." They show that when data collection in India, which has primarily taken place in urban settings, extends into rural areas, 1) the agency of participants who are in financially precarious situations is highly questionable; and 2) the results of this research will not benefit the research subjects themselves. Furthermore, the move of work, education, and socialization online during the COVID-19 pandemic has "occasioned one of the great data heists in human history, specifically, the mass harvesting of voiceprints" (Sterne and Sawhney, forthcoming).

56. Kazutaka Kurihara, "SpeechJammer: A System Utilizing Artificial Speech Disturbance with Delayed Auditory Feedback," https://sites.google.com/site/qurihara/top-english/speechjammer (accessed April 1, 2013).

57. Eidsheim, *Race of Sound*, 129.

58. Eidsheim, *Race of Sound*, 139–46.

59. Noble, *Algorithms of Oppression*, 1.

60. Noble, *Algorithms of Oppression*, 1.

61. Noble, *Algorithms of Oppression*, 1.