16. Tracing the Voice's Digital Materiality

Nuno Atalaia

This chapter analyses how recent developments in speech processing technologies—digitizing and storing vocal utterances as data—frame the voice as a material. To do so, however, goes against the more traditional western understanding of the voice as a cultural category of questionable materiality. Already in Aristotle, the voice was seen as an intermediary between physical anatomy and metaphysical meaning-making. Through their signifying voice—*phone semantike*—humans were able to access the abstract realm of language, distinguishing themselves from other animals and their production of meaningless sounds—*aggramatoi psophoi* (Butler). For Aristotle, the voice both belongs to and escapes from the material.

This is not to say that the voice's materiality has been a mute subject in cultural analysis. Quite the contrary: we need only think of Roland Barthes' fascination with the voice's non-semiotic sonic properties, what he called "the grain of the voice" (Barthes). Paul Zumthor, a contemporary of Barthes, went so far as proposing a dedicated study of what he called the *order of the vocal*, as the activities and elements of the voice which escape the linguistic (Zumthor). But such efforts still replicate the Aristotelian divide by partitioning the material realm of sounds—*psophoi*—from that of language.

Following Tim Ingold, it is as though the voice is always denied its own "hard physicality," an existence outside "the socially and historically situated agency of human beings." More so, by being set apart from the world's "material character" and its processes, the voice remains, at least partly, in a metaphysical realm. I aim to supplement current debates on vocal materiality by centering my analysis on the digital technologies capturing and processing the vocal into data. Vocal digitization, I claim, is something very new and very different from what humans make of and with their voices.

The current expansion of vocal digitization is made possible by the large infrastructures under the possession of a few large monopoly-driven corporations, such as Google and Amazon. A key element of this expansion, and this text's object of analysis, is the ongoing adoption of digital voice assistants (DVAs). DVAs, such as Alexa and Siri, are systems through which humans can vocally interact with differ-

ent objects while at the same time giving these systems access to the same humans' voices.

What drives my analysis is more than a speculative interest in these talking machines. Rather, it is motivated by the ongoing encroachment of large tech companies on the daily lives of an ever-greater number of humans. This encroachment is often facilitated by the unquestioning adoption of technologies such as DVAs, without considering the consequences their digital infrastructures might entail. To trace the voice's digital materiality is also to question how this material stands to be commodified and by whom.

DVAs and the Voice as a Material Affordance

Let us consider Amazon's recently released 2022 Superbowl halftime ad, starring the actress Scarlett Johansson, her husband Colin Jost, and the company's flagship DVA, Alexa ("Watch Alexa's 2022 Commercial for the Big Game"). The short film begins with Jost calling out to their assistant to set up their home for watching the Superbowl game. In a few seconds, the fireplace ignites, the living room's lights shift, blinds are drawn, and we hear the DVA's female-coded voice announce: "rosé is chilling". Impressed by the display, Johansson speculates on Alexa's "mind-reading" abilities; the ad takes a dark turn with a series of quick vignettes where the assistant humorously calls out the couple's secret opinions of each other, even denouncing their falsities to dinner guests. The ad returns to the original setting, in which the couple agrees it is best their DVA cannot read minds.

The ad organizes itself into two clear registers: firstly, factual, showcasing the different voice-activated automations Alexa makes possible; then, fictional, humorously portraying a soft dystopia of awkward moments. The border between fact and fiction, however, is constantly crossed. For example, in the speculative vignettes, Alexa responds to the unvoiced opinions of its users via actually existing automations such as activating a kitchen appliance or playing a song. The casting of Scarlett Johansson also harkens back to her performance as the fictional Samantha in Jonze's *Her* (2013). Finally, there is a play with the audience's expectations of what is possible through Alexa: when *Her* was released, the ad's first half would have been as speculative as the second. This ad underlines how vocal digitization brings a volatile fluidity to the borders separating what is fantasy and what is fact when it comes to talking machines.

In this ad, there is a (half-hidden) variety of computational devices connected through dispersed digital infrastructure that frame the voice as a material affordance of our surroundings. The voice is integrated into the "transforming and transformative life of materials" (Drazin 21). As Adam Drazin notes, "materials as social phenomena happen not only when wood becomes a table . . . but when your table

is wood" (4). With DVAs, and the digital infrastructures of tech corporations, your table and the many items it supports can also be voice. This vocal materiality, rather than the result of an interpretative effort such as Barthes', is a consequence of the application of specific technologies.[1] Therefore, its analysis requires a technical understanding of the digital systems that make it possible.

In the following sections, I focus on two aspects of vocal digitization: input—the capture of vocal sounds from human sources to a technological medium—and output—the production of vocal sounds from non-human sources. I will conclude with a few critical remarks on what it means for these technologies to be almost entirely under the control of a few tech corporations. DVAs and their development are part and parcel of these corporations' history: they make DVAs possible and dictate how the voice is made digital.

Input: From Recording to Speech Processing

The history of speech processing—the technologies allowing humans to use machines through their voices—predates that of DVAs considerably, though with limited success. By 2011, when DVAs first appeared, speech processing had mostly stagnated, with little to no funding available since the 1970s (Ekman). DVAs were made possible by two major shifts in speech processing: one at the level of its actual operations and the other at the level of its organization.

Firstly, the models used for speech processing moved from the actual transcription of recorded sentences to a method that probabilistically reconstructed words and sentences through data extracted from these recordings. Using statistical models, different language maps were created based on the probability of one vocal sound preceding another (Pieraccini). These systems did more than greatly increase the accuracy of speech processing: they also made this process dependent on the production of vocal data from a human's vocal utterances. Therefore data, more than an abstraction involved in these systems, is a product resulting from these systems (Gitelman and Jackson).

Secondly, the many operations behind speech processing systems—including those described above—were dispersed through the cloud computing infrastructures of the corporations that own and sell DVA-operated devices. This is why, in Amazon's ad, Alexa can bring Johansson and Jost's home to life at the sound of their voice. A smart speaker or a fridge equipped with a microphone can parse a vocal recording to digital data with little computing power needed. However, there are

1 Though Barthes' fascination with the grain of Fischer-Dieskau's voice might not have been possible were it not for the technologies recording and reproducing the singer's renditions of Schubert.

heavier operations necessary to transcribe this data into text, then interpret this text into a desired outcome, and finally coordinate this desired outcome among different devices. All of these processes take place in one of Amazon's neighborhood-sized data centers (Vlahos).

The voice, now digitized, enters a material flow: collected by machines, stored into digital infrastructures, and processed into actions made by the same or other machines. Furthermore, vocal data does not evaporate once an action takes place; it persists well beyond DVAs and their operations. Vocal data can be repurposed, sampled, and processed into systems that only tangentially relate to the daily experience of DVA users. One field of research in which vocal data is fast becoming a valuable material is healthcare. Vocal biometrics—a specific category of data made possible through speech processing—can already be applied to diagnostic models able to predict and track diseases ranging from depression to coronary heart disease (Sigona). Vocal biometrics can also be used to deduce emotional states, independent of a human's intention or the linguistic content of their spoken utterances. Amazon already owns the patent for an emotion-tracking system which could soon be included in its DVA's design: ironic, considering Amazon's dystopian portrayal of a mind-reading Alexa (Jin and Wang).

The voice's digital materiality, therefore, not only exists beyond the historically situated agency of humans but has the potential to undermine this very agency. Rather than a metaphysical vehicle of meaning, vocal data, and the information to be extracted from it, become a material repository of a human's interiority, regardless of its linguistic content. But, as the next section illustrates, speech processing does not only allow machines to capture a human's voice: it also equips them with voices of their own.

Output: From Ventriloquism to Voice Synthetization

The digital development of machinic voices is an equally vital component of DVAs. By the time of Apple's release of Siri in 2011, the industry standard for making machines talk was known as concatenative speech synthesis. In this method, you would first ask a human to record themselves saying various sentences. These recordings would then be fragmented into their smaller components, from short words to isolated syllables. Through text-to-speech software programs, these components are reorganized to produce new utterances based on written sentences. Though the human speaker is detached from the non-human device by several technical layers, this method can still be said to fall under what Steven Connor categorizes as ventriloquism: the performances creating the illusion of transferal of a human's voice to a non-human object or disembodied entity.

Though countless devices could have the same voice, a singular human voice was still required. The actress Susan Bennet, for example, would enjoy moderate fame for voicing Siri's utterances. This method of speech synthesis was used to voice all four major platform DVAs: Apple's Siri, Microsoft's Cortana, Amazon's Alexa, and the earlier iterations of Google's voice-user-interface systems. Even when these companies offered different voices—in terms of gender coding or language, for example—each alternative was still modelled after another individual voice. However, as Google released its flagship "Google Assistant" in 2016, the company radicalized the production of machinic voices.

The same year, the large tech company also released a new way of creating human-sounding voices by applying Artificial Intelligence (AI) software to pre-existing vocal data (Oord et al.). The name of this system was Wavenet, and it could "generate raw speech signals" (1) from large samples of undifferentiated data without needing an intermediary voice or any textual input. These "raw" vocal sounds were detached from any semantic element; Google had created a system able to reproduce Barthes' famed "grain of the voice." Even more striking, the vast majority of this process took place without the need for any human supervision.

In 2017, Wavenet was combined with text-to-speech software to create Tacotron 2, allowing these AI-generated voices to be articulated into words and sentences (Wang et al.). This system made it possible to create a limitless number of voices and program them with different traits, from gender expression to regional accents. AI-powered speech synthesis software has since become the industry standard, leading all DVAs to abandon their previous human counterparts. Aristotle's framing of the voice as the result of complex anatomy is not put into question: what is put into question is whether this anatomy needs to be human.

These voices can no longer be called an instance of ventriloquism: they spring from vocal data's material flow, all made possible by myriad computational devices and the digital infrastructures owned by tech corporations. The voice announcing that the "rosé wine is chilling" in Alexa's ad is, therefore, radically different from the voice with which the DVA was first released in 2014. These voices are materials: they exist as artefacts, reorganized and redesigned from the commodified resource our voices become once digitized.

Dehumanizing the Voice

In her study of Dutch fashion designer Iris van Herpen's work, Anneke Smelik remarks that the intertwining of digital systems and human activity complicates what counts both as material and human. Digital mechanisms are more than simple tools. Rather, they alter how humans relate to their environments and the materials that compose them. DVAs present a valuable opportunity to analyze how digital tech-

nologies complicate what of the voice remains human in origin and what becomes an external material.

Returning to Aristotle: humanity's anatomical complexity secured its claim to a certain monopoly of the voice and its metaphysical capacities for meaning production. It is not my claim that the voice's digitization negates this monopoly; rather, it is materialized in a more Marxist sense. The voice becomes part of infrastructures of digital reproduction that impose specific property regimes on humans over their vocal utterances. These regimes, however, need not be total, nor are their infrastructures beyond contestation. As Ochoa Gautier remarks:

> What is particular about the voice is that, as a force that hovers between the world and what humans do with the world, it is particularly poised to be used as a disciplining force and yet it simultaneously easily reveals the limits of such a process. (210)

My short tracing of the vocal and the digital does not aim to provide a complete overview of the voice's disciplining force within a digital context. Instead, like Gautier, it draws attention to the disciplinary potential of the infrastructures behind vocal data, which the tech companies who own them are unwilling to divulge. Amazon's dystopian portrayal of their DVA unwittingly reveals a glimpse of a world of human vocal dispossession brought upon by such a disciplining force.

Once the voice becomes proprietary, its meaning and our interiority become the subject of manipulation and commodification. As data, the echoes of a user's voice, and any artefacts, outcomes, and information resulting from its processing, can become the intellectual property of large tech corporations. The political consequences of this shift are beyond the scope of this chapter. But, as Ingold reminds us, "to know materials, we have to follow them" (437): to know the voice's digital materiality, we must follow its flow. As with any attempt at engaging with materials, this remains an unfinished task.

Works Cited

Barthes, Roland. *Le grain de la voix* [*The Grain of the Voice*]. Seuil, 1999.

Butler, Shane. "What Was the Voice?" *The Oxford Handbook of Voice Studies*, edited by Nina Sun Eidsheim and Katherine Meizel, Oxford UP, 2019, pp. 3–18.

Connor, Steven. *Dumbstruck: A Cultural History of Ventriloquism*. 1st edition, Oxford UP, 2001.

Drazin, Adam. "To Live in a Materials World." Introduction. *The Social Life of Materials: Studies in Materials and Society*, edited by Adam Drazin and Susanne Küchler, Routledge, 2015, pp. 3–28.

Ekman, Ulrik. *The Complexity of Coding Conversational Agents*. Academia.edu. 2019. https://www.academia.edu/37885872/The_Complexity_of_Coding_Conversational_Agents.

Gautier, Ana María Ochoa. *Aurality: Listening and Knowledge in Nineteenth-Century Colombia*. Duke UP, 2014.

Gitelman, Lisa, and Virginia Jackson. Introduction. *"Raw Data" Is an Oxymoron*, edited by Lisa Gitelman, MIT Press, 2013, pp. 1–14.

Ingold, Tim. "Bringing Things Back to Life: Creative Entanglements in a World of Materials." *NCRM Working Paper Series*, July 2010, https://eprints.ncrm.ac.uk/id/eprint/1306/.

Jin, Huafeng, and Shuo Wang. *Voice-Based Determination of Physical and Emotional Characteristic of Users*. US 10,096,319 B1.

Oord, Aaron van den, et al. *WaveNet: A Generative Model for Raw Audio*. arXiv:1609.03499, arXiv, 19 Sept. 2016. *arXiv.org*, http://arxiv.org/abs/1609.03499.

Pieraccini, Roberto. *The Voice in the Machine: Building Computers That Understand Speech*. MIT Press, 2012.

Sigona, Francesco. "Voice Biometrics Technologies and Applications for Healthcare: An Overview." *JDREAM. Journal of InterDisciplinary REsearch Applied to Medicine*, vol. 2, no. 1, 2018, pp. 5–16. *siba-ese.unisalento.it*, https://doi.org/10.1285/i25327518v2i1p5.

Smelik, Anneke. "Fractal Folds: The Posthuman Fashion of Iris van Herpen." *Fashion Theory*, vol. 26, no. 1, Jan. 2022, pp. 5–26.

Vlahos, James. *Talk to Me: How Voice Computing Will Transform the Way We Live, Work, and Think*. Houghton Mifflin Harcourt, 2019.

Wang, Yuxuan, et al. *Tacotron: Towards End-to-End Speech Synthesis*. arXiv:1703.10135, arXiv, 6 Apr. 2017. *arXiv.org*, https://doi.org/10.48550/arXiv.1703.10135.

"Watch Alexa's 2022 Commercial for the Big Game." *US About Amazon*, 7 Feb. 2022, https://www.aboutamazon.com/news/devices/alexa-2022-commercial-big-game.

Zumthor, Paul. *La Lettre et la Voix : De la "littérature" médiévale* [*The Text and the Voice : On Medieval "literature"*]. Seuil, 1987.