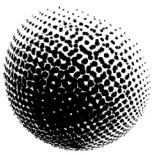
IM@GO
A JOURNAL OF THE SOCIAL IMAGINARY

The Shaping of a Standard Voice: Sonic and Sociotechnical Imaginaries in Smart Speakers

Domenico Napolitano
domenico.napolitano-ssm@unina.it
Law and Organizational Studies for People with Disability | Scuola Superiore Meridionale



Abstract

This article deals with the issue of sound imaginaries, using artificial voice as a case study. Specifically, it focuses on the project of creating a voice standard associated with the 1980s speech synthesis device DECTalk. Using a media-archaeological methodology which draws on heterogeneous empirical material about technological materialities and discursive representations, the article investigates the imaginary formative principles governing the definition of a vocal sound ideal type such as that incorporated in speech synthesis systems of the '80s. Referring to the most recent theories on sonic imagination, the article argues that these formative principles are still at work in current smart speakers, as they refer to the imaginary of a voice "in general", which recalls ideas of authority, fidelity and transcendence. As the "standard" of artificial voice is an ideal type created at the crossroads between cultural ideas of voice within European philosophical tradition and the techno-scientific knowledge through which voice has been measured, reified and adopted by organizations, the article argues that sonic imaginary can be related to the framework of sociotechnical imaginary theory.

Keywords

Sonic imaginary | Sociotechnical imaginary | Voice | Speech synthesis | Media archaeology

N. 21 - Year XII / July 2023 pp. 177-196
www.imagojournal.it

ISSN: 2281-8138

1. Introduction

In 1984, the company "Digital Equipment" released the DECTalk, a vocal synthesizer which was able to convert any text written on a computer into intelligible speech, and spell it out loud through computer speakers. A few years later, that system was adopted by Stephen Hawking as a vocal prosthesis, after losing his voice due to SLA disease (Garcia, 2018). This contributed to the spread of synthetic voices within the social imaginary and the association of those voices with a specific sound. DECTalk's default voice was represented as the character of Perfect Paul, and the introduction sentence pronounced by this voice was: "I am Perfect Paul, the standard male voice"[1]. The sound of this voice is perfectly intelligible, although slightly metallic and clearly non-human, both for its monotonous intonation and unnatural uniformity. What makes this voice "standard"? And is there anything like a "standard" voice if voice, as Cavarero [2005 (1998)] argues, is rather the trait of uniqueness of any individual?

In the following pages I will argue that what made this voice a "standard" in the '80s is its relation to a specific *sonic imaginary* about how the voice of an "average" man should sound and how a computer voice embedding those features should sound. Besides the aesthetic and cultural representations, I will reconnect to theories of imaginary to investigate the "formative principle" [Durand, 1999 (1960); Wunenburger, 1991; Taylor, 2004; Marzo, 2019] explaining how those constructions happen and take a certain form in a certain cultural context. As the "standard" of artificial voice is an ideal type created at the crossroads between cultural ideas of voice, subjectivity and agency in European philosophical tradition and the techno-scientific knowledge through which voice has been measured, reified and computationally reconstructed, I argue that sonic imaginary can be related to what has been defined as the "sociotechnical imaginary" (Jasanoff and Kim, 2015). This approach gives great importance to both historical conditions and to materiality, in this case the materiality of sound and of its technological production. I argue that such an approach to the imaginary is connected to a trend in "media archaeology" (Ernst, 2012; Parikka, 2012) which is focused on nonhuman agencies and techno-materialities and can provide an useful methodology of investigation.

Sonic imaginaries theory, in the sociotechnical approach featured in this paper, allows for a non-behaviorist response (differing from the one provided by Nass and Brave, 2005) to the following questions: Do we prefer artificial voices to sound human, or to sound robotic? Why was a specific sound associated with artificial voices in different technological and historical contexts?

[1]An example is available at the following link: https://www.youtube.com/watch?v=8pewe2gPDk4 (accessed 11/06/2022). It is worth to note that the equivalent female of Perfect Paul in the set of DECTalk's synthetic voice characters is Beautiful Betty, whose presentation sentence meaningfully emphasizes the gendered aspects: "I am Beautiful Betty, the standard female voice. Some people think I sound a bit like a man". For the socio-political aspects related to the gendered voice of smart speakers, see, among others, Männistö-Funk & Sihvonen (2018) and Strengers & Kennedy (2020).

N. 21– Year XII / July 2023 pp. 177-196 ISSN: 2281-8138
imagojournal.it

2. Theoretical Framework: Smart Speakers Between Sound Studies, Cultural Studies and Sociotechnical Imaginaries

In recent years, smart speakers have received significant attention from scholars, as they are more than simple tools that facilitate human-computer interaction: they are also symbols of the advancements of technology (Faber, 2020; Napolitano, 2020a; Humphry & Chesher, 2021; Natale, 2021). Accordingly, they embed meanings which go beyond their functionalities, and relate to the literal sound of their voice: timbre, prosody, gender and so on. While these aspects have been investigated with reference to contemporary modes of data extraction known as surveillance capitalism (Zuboff, 2019; Woods, 2018) and socio-political issues related to gender and race (Hester, 2016; Phan, 2019; Lingel & Crawford, 2020), few studies have so far focused on the sonic imaginaries operating behind the choice to endow these technologies with certain vocal features.

In this article I refer to the concept of sociotechnical imaginaries (Jasanoff & Kim, 2015) to understand the interactions between sound aesthetics and communication technologies that led to the idea of a "standard" voice, that is of a voice with generic features to be adopted in disembodied agents. As this idea was born in the framework of research on artificial voice and telecommunication in the '80s, I adopted a media-archaeological method (Ernst, 2012; 2018; Parikka, 2012) to conduct the investigation. Following this approach, I compared and analyzed documents about the technical aspects and the expectations declared by those who worked on the development of those systems from the '80s to the early '00s. Another important source of knowledge for this investigation was the act of listening itself, as it allowed to appreciate the specific sound of different synthetic voices, in this way working as an access point to sonic imagination. In the background, I refer to wider ideas about voice, sound and technology as expressed in cultural representations and theoretical works of the same period.

Unlike many contributions on this topic appearing in recent years, I'd rather start by distinguishing between two sociotechnical instances: the voice *within* the machine and the voice *of* the machine. The first instance refers to imaginary motifs behind the idea of a talking machine, invoking the "vital breath" imaginary which animates the inanimate, and therefore linked to the idea of a "personification" of the machine through the voice (Poizat, 2001; Dolar, 2006). The second instance refers to the way of representing the specific *sound* of artificial voice from the perspective of programmers, users and common culture. While research on the first aspect has been conducted within cultural sociology and media studies (Guzman, 2015; Gehl & Bakardjieva, 2017; Faber, 2020; Natale, 2021), the second has attracted attention of sound studies (Sterne, 2003; Ernst, 2021) and gender studies (Eidsheim, 2016; Hester, 2016; Strengers & Kennedy, 2020; Faber, 2020). In this research I argue that the voice of the machine has been represented in western culture in two ways: a) the voice "in general", often linked to the philosophical themes of metaphysical subjectivity - the divine, the paranormal, or authority; b) the "cloning" of a specific person's voice. The present

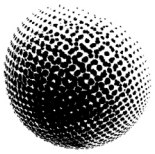
N. 21– Year XII / July 2023 pp. 177-196 ISSN: 2281-8138
imagojournal.it

article specifically deals with voice "in general", through the lens of sonic imagination within sociotechnical imaginary theory, via a media-archeological methodology.

2.1 Sonic Imagination and Sociotechnical Imaginary

The problem regarding the link between sonic dimension and imaginary has been attracting more attention in recent years, as the recent publication of *Oxford Handbook of Sound and Imagination* (Grimshaw-Aagard et al., 2019) suggests. In this collective work of 70 essays in two volumes, curators discuss the possibility of separating the imaginary from the visual dimension of image, exploring historical, cultural, artistic and techno-scientific ways in which sound can recall imagination in a specific way. "Vagueness and ambiguity are essential to the imagining of sound, in engaging with the existential task of deriving function and meaning from its perception, and thus engaging with actual and virtual worlds." (Grimshaw-Aagard et al., 2019: 5)

About a decade previously, discussion around "sonic imaginations" was themed in its socio-material aspects by Jonathan Sterne (2012), becoming a sort of manifesto for the young discipline of sound studies, to which Sterne was one of the major contributors. He underlines how any knowledge of sound comes from a particular and historically framed mix of materialities, such as sound technologies and culture, the latter including the ways by which a physical phenomenon such as air vibration can be recognized, labeled and discerned as audible. Imagination plays a crucial role in this process, as it fills the gap between physics and meaning. According to Veit Erlmann (2004) – a key figure in sound studies – as sound cannot simply be pointed at with a finger (denoted) and reified, any description, comprehension and knowledge of sound entails referring to a social way of creating meaning in acoustic phenomena, a way of describing them, which is a kind of imagination.



Although not themed as such, the imaginary dimension emerges in most of the studies addressing sound. In this area of study, Murray Schafer's lesson (1977) is seminal, as he related the concept of "soundscape" to the way listeners of an acoustic environment "perceive" it as such, in short how the set of natural and man-made sounds characterizing a living space becomes meaningful to people and mediates their relationships. As explained by Marzo (2019), this is an act of imagination even though it doesn't concern images: imagination, in fact, is a "resonance" between body and environment, which precedes the moment of symbolic comprehension. We can understand it as the force that drives the passage from sensation to meaning – which becomes imaginary when it assumes a social form. In this sense, if imaginary "is not a set of ideas; rather, it is what enables, through making sense of, the practices of a society" (Taylor, 2003: 2), sound can be rightfully included within it. As highlighted by Ingold (2011: 138), in fact, sound is not an object of knowledge, but a medium through which we know the world, thus "we don't hear sound, we hear in sound". Even though Ingold's position is critical towards the notion of soundscape, which for him risks commodifying sound, attention to the "process" of listening as an act of imagination seems to align the idea of sound as a relational medium (Di Scipio, 2013) with that of soundscape.

Even more recently, James Mooney and Trevor Pinch (2021) (other "founding fathers" of sound studies), reconsidered the issue of sonic imaginaries through the lens of Science and Technology Studies (STS), studying the music of Hugh Davies and David van Koevering. They argue that "sonic imaginary" is a notion necessary to bringing forth a shared sonic world or experience grounded in technology, institutions, and networks. "In a sonic imaginary, sound itself has socio-material agency and makes a crucial difference in how worlds are enacted. Individuals must perform their sonic imaginaries in their ongoing engagements with the socio-material world." (Mooney and Pinch, 2021: 114). From Mooney and Pinch's perspective, the strength of the sonic imaginaries concept lies in its bringing together of socio-material practices with the imaginative dimension, since "acts of imagination, invention, and creativity are not, in fact, instantaneous epiphanies that happen in the heads of visionary individuals – as often implied in historical accounts – but are lived socio-material processes." (p. 144).

In the framework of STS, the concept of sonic imaginaries provides a way to give "agency" to sound (Cox, 2011; Velasco-Pufleau, 2021) without falling into the trap of giving sound ontological status independent from human actors. "Sound always mediates human practices, and it is these entanglements that are captured by a term such as a sonic imaginary" (Mooney and Pinch, 2021: 144).

Mooney and Pinch's approach recalls Jasanoff and Kim's (2015) notion of *sociotechnical imaginary*, where imaginary is conceived as a phenomenon emerging from the material world. Sheila Jasanoff (2015: 4) argues that socio-technical imaginaries are "collectively held, institutionally stabilized, and publicly performed visions of desirable futures, animated by shared understandings of forms of social life and social order attainable through, and supportive of, advances in science and technology". In her view the concept of "co-production" developed within STS has been a major step towards the idea of sociotechnical imaginaries, as it highlights the mutual influence between social practices and techno-scientific knowledge. Nevertheless, the concept does not account for the question of "why" science and technology embed certain social values, which affect the images of what is attainable and desirable to attain through science and technology. Conversely, it also leaves the question of "why" societies promote certain technologies and scientific concepts over others in different periods and contexts unaccounted for. These issues transform in specific ways to give meaning to the present and to imagine the future.

In my reading, the concept of sociotechnical imaginaries can fill this gap by connecting sociotechnical arrangements of meanings, practices, norms and artifacts to the structures of imaginative faculty.

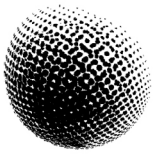
When related to the concept of sociotechnical imaginary, these considerations help us to understand the "formative principles" (Marzo, 2019) which materialize differently over time (morphology) and interact with environments and cultures, giving them a lasting socio-historical significance (formant form) in the form of institutions, artifacts, symbols, but also concepts, norms, conventions. This reflects not only in the ways social artifacts (thus also technological artifacts) are institutionalized and naturalized (Berger and Luckmann, 1966), but also in the way they are designed and built: they can

N. 21– Year XII / July 2023 pp. 177-196 ISSN: 2281-8138
imagojournal.it

be recognizable both via their form, meaning and use, and also how they are first fostered and produced within the imaginary world.

In line with this reasoning, Patrice Flichy (2001) underlines how the management of technical projects, just as much as the specific technical operations that they are programmed to carry out, are located at the crossroads between different social worlds (programmers, users, critics, scholars, sellers, investors and so on) and it is in that intersection that they find their meaning. According to Flichy (2001: 6), imaginary plays a crucial role here, as it is the "common objective", the shared vision that ultimately links technical objects with the collective building of their social meaning. Following this perspective, imaginaries are crucial in both design and acceptance of technological artifacts as they drive choices towards specific forms rather than others, and lead attribution of meaning and recognition processes.

This reflects a dual relationship between technology and imaginary: while the imaginary influences technology's form and the meaning attributed to it, technology itself takes part in the processes of the imaginary (Abruzzese, 2001; Musso, 2019). If on the one hand interactions with the media cannot be read without reference to mythical, symbolic, ritual schemes (Davis, 1998; Mosco, 2004) and belief-systems (Natale & Pasulka, 2019), on the other hand we find the very functionalities of technologies which have a relationship with the imaginary, as they contribute to produce and reproduce narratives, symbols and rituals, in the actions they facilitate. As Stefano and Paolo Bory (2015: 70) underline, technique and discourse on technique become two sides of the same institutive process, that of representing, imagining and instituting the social world.



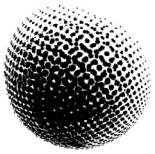
3. A Media Archaeology of Voice Imaginaries

Following this set of theoretical considerations, this study proposes a media-archaeological investigation of the sonic and sociotechnical imaginaries of artificial voice. Media archaeology (Ernst 2012, 2018; Parikka, 2012) is an approach interested in reconstructing genealogies of media technologies focusing not only on the discourses around them but also on their materialities, that is on technical operations and embodiment practices. Media archeology attempts to unearth underlying continuities and discontinuities between old and new media technologies and human modes of representation, often nested in the tension between the working ways of technological artifacts and the social meanings attributed to them (Ernst 2018).

In its radical formulation by German scholar Wolfgang Ernst (2018), media archaeology is a non-anthropocentric approach that states the importance of media's technical operations as *non-discursive* practices: their embedding epistemologies which produce ways of understanding and experiencing the world. In the wake of McLuhan's (1994) reflections, radical media archaeology proposes the idea that media do not just reflect social meanings: they *produce* them through their very functioning. In this sense, the processes of embodiment, imagination, and perception are influenced by the technical apparatus and their modes of functioning, and media

influence the very process of imagination. On the other hand, media archaeology is interested in the way those apparatuses reveal themselves as the consolidation of previous fantasies, desires and imaginaries.

When directed towards sonic media and voice technologies, the media archaeology approach aims to unearth the imaginary dimension embedded both in the technological treatments of voice and in the aesthetics and discourses related to it. In fact, if sociotechnical imaginaries are visions of desirable futures, those visions are particularly evident in the imaginations people had in the past, when old technologies were new (Parikka, 2012). When talking about artificial voice, those imaginaries are not only detectable in social representations, such as the DECTalk presentation strategy through the character of "Perfect Paul"; they are also embedded into sounds and into technologies. Following media archaeology, investigation should then take into consideration both discursive elements (what "Perfect Paul" says) and non-discursive practices – *how* "Perfect Paul" speaks, the *sound* of its synthetic voice and the algorithmic ways it is produced. This kind of investigation also benefits from comparisons between old and current phenomena, such the one between ancient imaginaries of disembodied voices, modern speech synthesis and contemporary smart speakers.



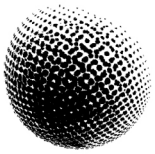
Steven Connor (2000: 40), a pioneer in media archaeology and cultural phenomenology, argues that "the technologies of the voice are actualizations of fantasies and desires concerning the voice which predate the actual technologies". Reporting specifically on the history of disembodied voice in pre-technological times (for which ventriloquists offer an excellent example), Connor (2004) highlights how voice is actually a *cultural artifact* defining itself within technical and social ways of thinking about and using it. For Connor, voice is not simply a bodily emission; it is also the imaginary production of a secondary body, a body double: a "vocalic body" (Connor, 2000: 35-42). This becomes particularly evident in disembodied voices, such as the ventriloquist's, or the synthetic voice of smart speakers. In these cases, an imaginary supplemental body (Kane, 2019) is projected onto technological devices, working as a compensatory mechanism (Kluitenberg, 2006) to balance the absence of a clear cause for that voice. These mechanisms attribute typical embodied singularity traits to the artificial voice[2], contributing to implement identification and personification phenomena sustaining much of the fascination, and also rhetoric, about talking technologies (Nass and Brave, 2004; Natale, 2021). As underlined by Faber (2020: 162), even when disembodied, voice expresses "a range of gendered identities and associated phantasies" through sound. Voice, it might be said, doesn't always comes from bodies, but sometimes creates bodies, linking organic and inorganic, material and fictional.

Voice has an inherent relationship to sociotechnical imaginary: What we consider as "voice" is a hybrid between sonic emission linguistically declined and produced by human bodies, and the techniques of transmission, recording, reproduction and

[2]This is evident also with racialized voices. Nina Sun Eidsheim (2019) highlights how the vocality of African Americans has become an imaginary, since a number of racial assumptions, expectations and cultural-material marginalizations of those people rely on the very sound of one's voice.

electronic manipulation of sound signal. Sound and telecommunication technologies, in fact, have transformed both the sound of voice and the practices of speaking to others, as well as the practices of listening (Sterne, 2003). We can think, for example, of how microphones and amplification influenced the imaginary of power and its possibility to address crowds (Bernays, 2008-1928). Or about how the telephone contributed to redefining voice's proxemics and consequentially distant communication imaginaries (Borrelli and Petulla, 2017). If, as Nass and Brave (2005) argue, mankind processes of adaptation to its surroundings produced specific psycho-social and cognitive expectations regarding the sound of voice, new sound assumed by voice in a technological environment is at the center of new adaptive and psycho-cognitive investments.

Therefore, talking about imaginaries of voice means referring to specific ways of thinking about voice: its sound, its intonation and transformation, which are intrinsically influenced by technology. Vallee (2017) notes that the voice, as a cultural image, is an imaginary organ that transgresses the boundaries of technological, biological, physical, psychological, social and cultural frameworks. What he calls "voice imaging" is the imaginary way of perceiving voice as a liminary entity, a quasi-object. Vallee explores how the ways of imaging voice have been shaped by technologies and science, and by the way they have constructed possible causal relationships. Voice is an "imaginary organ, insofar as it has come, through its imaging technologies, to be an effect built from the imagining of its causes." (Vallee, 2017: 86).



If the idea of a talking machine starts from ancient imaginaries, it's impossible to deny how voice synthesis technologies provided those imaginaries with a new materiality to be applied to, and a specific sound. Most prominently, Cinema and Television contributed not just with new fantasies about cyborgs but also by describing and representing a specific phenomenology of non-human voice and the sound it should have (Faber, 2020). This sound then influenced (and indeed continues to influence) the techno-scientific work of those who create synthetic voices, which we can often find behind software and voice synthesis artifacts (Napolitano, 2020a).

4. The Voice "In General"

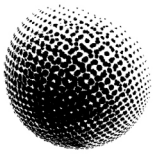
In the introduction of this article, I've recalled the example of the DECTalk's Perfect Paul male standard voice. In the following pages I will argue that this case is a clear expression of sociotechnical imaginary, as the "standard" voice recalled by the firm is the result of a specific way of imagining the sound of the computer as that of the "average" man. This imaginary is the result of a specific way of imagining how a sound signal should be computationally treated in order to articulate phonetic sounds.

Adopting a media-archaeological perspective, the following paragraphs will first explore scientific knowledge and technological materialities of voice synthesis, and then explore the cultural ideas associated to voice by western tradition. This will allow to investigate the imaginary of the voice "in general" and to define some of its characteristics.

4.1 The Technological Materialities

Early experiments in computerized synthetic voice began in the '50s, mainly in the US at Haskins Laboratories in New Haven (conducted by L.G. Gerstman, Alvin Liberman, F.S. Cooper) and Bell Laboratories in Murray Hill (where researchers such as John Kelly Jr., Carol Lochbaum, James Flanagan worked), and in Europe at the Department of Speech Communication & Music Acousticts at the Royal Institute of Technology in Stockholm (where Gunnar Fant worked). In those experiments, the computer was instructed with an algorithm then defined as *articulatory synthesis*, based on computational modeling of human phono-articulatory apparatus (Flanagan, 1972; Flanagan and Rabiner, 1973).

As argued by James Flanagan (1972) – among the most influential researchers in the context of vocal synthesis – the goal of articulatory synthesis is "to determine physical specifics of speech's production, of his perception and language, and to incorporate these specifics in the transmission system". Therefore, the articulatory synthesis groups knowledge into three kinds: 1) Vocal emission and phono-articulatory apparatus; 2) Sonic perception and hearing apparatus; and 3) Language.



In order to achieve articulatory voice synthesis, research in linguistics and phonology had to be interweaved with research on the perception of sound and speech. In 1951 a groundbreaking study by Franklin Cooper, Alvin Liberman and John Borst (reported in Flanagan and Rabiner, 1973) argued about the possibility of identifying some "building blocks" of spoken language regardless of the speaker's timbre and prosody. For the authors, these fundamental elements were not inherent to language itself, but rather dealt with the way voice articulation happens and how it's perceived by the listener. For a computer to be able to "speak" it was necessary to acknowledge these fundamental language elements, in order to focus on those rather than on the elements of vocal timbre. Without fundamental elements, the spoken language would be unintelligible.

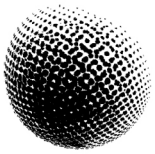
This key assumption in speech synthesis introduces many epistemological problems. Nevertheless, it reveals the orientation of studies in voice synthesis during those years. In articulatory synthesis, speech is not only about the vocalization of language, but rather a precise mechanical translation of language's fundamental elements, which are a function of the listener's perception. On the other hand, perception is defined in mechanical terms as a technical-organic apparatus, following Von Helmotz's epistemology described by Peters (2004). This mechanical conception of auditory apparatus allows us to imagine the disconnection between voice and the human subject: mechanical voice can still be considered as "voice" and not just a mere sonic emission, only when perception is indifferent to the source it was produced from.

Shifting focus onto the fundamental elements of spoken language and their articulation, articulatory synthesis produces a voice disembodiment which is not only phenomenological but is embedded in the engineering of voice. These elements, in addition to their articulation, now become *general* and abstract through mechanical

measurement. Thus they are embedded in the synthesis algorithms, which is now able to generate not only speech, but a *voice in general*. It is no longer about reproducing human voice, which is always necessarily unique, singular, embodied (even if recorded). Now it's about generating 'nobody's' voice, respecting general specifics identified with measurement tools.

Starting from the 80's, a different vocal synthesis technique, defined as *formant synthesis*, started to gain consideration, especially in research labs in the US. This kind of synthesis was integrated in toys, cars, videogames and computers. In 1978 Texas Instruments released the famous Speak&Spell, the first talking toy. In 1984 Apple was providing his computers with a synthesis engine named MacInTalk letting his devices to announce themselves to the world: this resulted in a great marketing campaign giving the company advantages in the competition with IBM.[3]

Formant synthesis was also applied to DECTalk, thanks to Johanthan Allen, Dennis Klatt, and Sharon Hunnicutt's work (1987), considered to be among the major researchers in this technique. Formant synthesis can be considered as an evolution of articulatory synthesis, even if some functioning principles are different. Articulatory synthesis essentially had a knowledge purpose and used computational models of vocal trait to study voice functioning, as well as testing the computer's possibility for speech. Formant synthesis had instead a performative purpose: its development was focused towards possible applications and products in marketing and public utility. In this system the simulation of phonic articulation was not a goal in itself, but rather became functional to achieving a specific result. Consequently, that simulation was simplified, regardless of scientific accuracy and adherence to the phenomenon. The goal was not to simulate the act of speaking, but to *model its effects*.



Techniques employed in formant synthesis radically detach the algorithmic process from the way in which vocalization happens in human apparatus, while the main goal is to find a compromise between human vocal behavior and the possibilities offered by the computer to users. It's an economization and optimization process, in a way comparable to the process leading to the adoption of compression in Vocoder used in telephonic communication (Mills, 2012).

In describing desirable futures for speech synthesis, Flanagan (1982) introduced four parameters: the first two were *intelligibility* and *naturalness*, which concerned the aesthetic and sensitive results desired for synthetic voice; the other two were *versatility* and *cost*, which concerned instead the conditions to achieve those goals in a specific sociotechnical context - considering the limits in computational power. Nevertheless, those conditions affected the very aesthetic of synthetic voice, since the sound imagined as "natural" was actually the one that also proved convenient in economic-organizational terms.

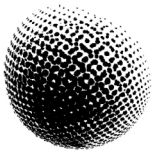
In general, voice synthesis systems resonate with a tension between imitation of human faculties and adaptation to technical principles of computational elaboration and communication. Both articulatory synthesis and formant synthesis are

[3] See the YouTube video "The Story of MacInTalk", available at the link: https://www.youtube.com/watch?v=UuVo4MHTEQ0 (accessed 15/06/2022).

expressions of a model-based approach. In this approach, a theory of human voice is incorporated into computer code. Synthetic voice then becomes a question about correct transmission and destination of the signal, rather than imitation from the original. This way, a voice "in general" is not only conceivable but also realizable, because the speaker's individual body and the imperfections characterizing the embodied voice become mere accessories to the model incorporated in the algorithm. As Remko Scha (1998: 41) concludes: "For the first time, language now has a sound independent of the body - a sound that directly emanates from the linguistic system, from syntax and phonemes."

4.2 The Metaphysics of the Voice "In General"

The idea of voice "in general" can be traced back to metaphysics. It can be understood as the expression of the separation between an abstract soul - the voice as *logos* - and the material body, contingent and inessential. In this scenario, the disembodied voice would be the voice deprived of the contingent attributes of corporeality and as such would be a pure expression of meanings. It is in these meanings, in fact, and not in matter, that for metaphysics the soul (or subjectivity) finds its true home. This *logocentric* scenario, in which the verb takes precedence over the materiality of its expression, in particular the sonorous one, has been famously described by Derrida (1967) and Cavarero (2005 [1998]). According to Cavarero (2005 [1998]), in this imaginary the voice is "desonified" and reduced to a mere phonic translation of the signifier, despite the fact of orality. It goes without saying that, in this conception, the body is not an integral part of the definition of subjectivity.



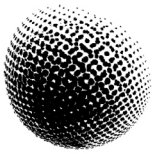
Although such a scenario seems out of date, it is possible to glimpse its traces in many places of contemporary technological reality. Such traces are not only discursive and representative but also material, as they concern the way of transforming an imaginary into sound, in particular into a certain type of sound of the artificial voice.

In proto-technological times, the voice "in general" has been associated mainly with imaginary and disembodied figures: angels, demons, divinities (Poizat, 2001). These figures represented authority and their voice was a direct expression of the law. The sound of those metaphysical voices, however, was not only imagined, but also materialized in a series of artifacts (narrative or technical) that have influenced current ways of dealing with the voice. Let's recall once again, as an example, ventriloquism, which has been associated with the intrusion of a demon, therefore of a metaphysical entity, inside the body (Connor, 2000). Even the eighteenth-century *automata* materialize the voice in the machine giving it a ghostly and "sepulchral" sound (Hollingshead reported in Hankins and Silverman, 1995: 215), which could lead to the idea of an ideal creature or of a "human in general", synthesis of the functionalities of the living, but without the defects of incarnation. Are these imaginaries that far from DECTalk's idealtypical voice characters? Remko Scha (1998: 41) has indeed used the term "platonic people" to describe them.

In the age of the mass media, as Frances Dyson (2010) notes, it was the figure of the television *anchorman* who embodied the ideal of the voice "in general", thus

defining a normative idea of how the standard voice, neutral and without accent, should have sounded. The anchorman is an almost metaphysical figure, since it is caught in a tension between bodily individuality and the universality of the law of the signifier materialized by his "perfect" voice. A voice that is, therefore, also an expression of authority, inasmuch as it "ritually, ensures the proper authority of the letter" (Dolar, 2006: 113) without overwhelming it with the violence of its singularity - which for Dolar is, instead, the constitutive feature not of authority but of the "authoritarian voice" (ibidem).

This imaginary relationship between the voice "in general" and authority could be traced today in the unaccented voice of voice assistants, which would also seem to be confirmed by the mystical and oracular representations that often distinguish them.[4] Through the sound interaction, it is as if the voice assistants wanted to create an immersive and non-mediated space that would strengthen the ideology of the "immateriality" of the digital (Mosco, 2004), the digital as a technology that does not "interfere" with the message, faithfully carrying its authority. This approach would go hand in hand with the typically metaphysical tendency to hide materiality.

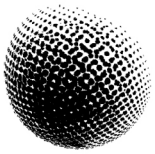


Dyson (2014: 79) notes, in this regard, that, as far as sound is concerned, the metaphysical concealment of materiality derives from a tendency to think of sound as a transcendent and immersive space, in which technological mediation disappears behind the presumed ontological equivalence between "reproduced sound" and "original sound source", corresponding to an equivalence between representation and reality that is not reflected in the visual dimension. As argued by Jonathan Sterne (2011) in an essay dedicated to the thought of Walter J. Ong, it would be precisely from this exceptional condition attributed to sound by metaphysical thought, constructed in the form of an ideology of sound as a place of transcendence, which the precedence attributed to orality in the context of the metaphysics of presence derives from. However, both for Sterne and for Dyson, this idea does not take into consideration (or rather deliberately conceals) the materiality of sound and the mediations that characterize it, focusing exclusively on the metaphysical concept of it as an immaterial event. Orality as a metaphysical concept allows for the transmission of meaning free from interference and noise - what has been later defined "fidelity" (Sterne, 2003); orality as materiality, on the other hand, allows for a series of mediations which are necessary for the making of sound, the first being the mediation of the body. But it is precisely the body, for metaphysics, that constitutes the main disturbance in the purity and immateriality of meaning. For metaphysics, the physical body, as a support, trace, signifier or medium, is noise.

[4] At this regard it is worth having a look at the advertisements of such devices. In example, Google Assistant's ad (https://m.youtube.com/watch?v=-lfHXKbsMLE, accessed 15/10/2021) promotes the image of a discreet but omnipresent technology, always there ready to help. This image is moreover consistent with Google's global project as a universal access point to the web and knowledge, contributing to produce an image of the company as a transcendent, omnipresent and omniscient entity, holder of knowledge and memory, and, moreover, always listening. At this regard, John Durham Peters (2015: 338) notes that the "I feel lucky button", popular a few years ago on the search engine, served to reinforce the mystical and unfathomable image of the technology giant.

As already noted by Rick Altman (1992: 40), this "ideology" has also been promoted by cinema, where the processes of narrative identification with the characters and their voices try to hide the fact that "recordings do not reproduce sound, they represent sound". Liz Faber (2020) also argues that the possibility of considering artificial voices as "credible", even when generic and devoid of traits of individuality, relies on the cultural habit produced by audio-visual synchronization. Cinema, using masking strategies, has produced the habit of narratively identifying a voice with a certain character and with a certain ideal type.

As already noted, whenever an acousmatic voice, whether human or synthetic, is heard, psycho-cognitive processes reconstruct an imaginary body associated with that voice (Nass and Brave, 2005), the one that Steven Connor (2000: 32) defines a "vocalic body". In the case of the "generic" synthetic voice, this compensatory mechanism of the imaginary takes place in a paradoxical way, since the body reconstructed starting from listening to the voice is in turn a generic body, not attributable to any person in flesh and blood. In this situation, the materiality of sound and mediation processes emerges in the form of a *symptom*. It is fetishized in the artifact (for example the technological device), and manifests itself in the form of small errors in pronunciation, or dry (and potentially unnatural) lack of accent. Ultimately, this is the hidden but nevertheless present body of the voice "in general".



5. Conclusions

In this work I have shown how the desirability of the voice "in general" is related to a "sociotechnical imaginary" (Jasanoff and Kim, 2015) shaping the sonic imaginary associated with artificial voice. That imaginary evolves over time and takes on different forms, expressed in ideal types and sonic standards. Behind those forms, however, a formative principle is retrievable in the interaction of cultural ideas, fantasies, techno-scientific knowledge and economic and organizational aspects.

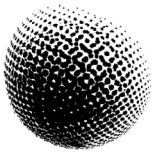
Nowadays, smart speakers such as Alexa and Siri employ different systems of voice synthesis based on *machine learning* and data. Embodied voice actors give their voice to the assistant in form of training datasets of sound recordings (Lorenzo-Trueba and Klimkov, 2019). Nevertheless, the aim is still to produce a voice belonging to nobody in particular, a voice "in general", although sounding more "natural". But "naturalness" is defined in the first place in relation to an imaginary human voice (Napolitano, 2020a). Therefore, in contemporary smart speakers as well as in more traditional representations of speaking machines, the dream of a standard voice, of a voice "in general", is always related to imaginaries of *authority*, *fidelity*, *transcendence*.

Nevertheless, the very systems of voice synthesis have also contributed to boost other imaginaries. Employed in the arts, they have been used to create new, impossible voices, expanding the field of what one can imagine a voice to be. Artists such as Martin Riches, Paul deMarinis, Florian Hecker and Tomomi Adachi have used artificial voices since the '80s to produce artifacts which interrogate human sensoriality's transformations, in a context where machine agency has more and more

autonomy. Their works explore non-anthropomorphic possibilities of voice synthesis, in an effort to open up the concept of voice to "machine creativity" issues, in other words non-humanistic ways that machines process voice.

Art plays a relevant role in defining and developing artificial voice, because new techniques are experimented with in this field, and because artwork explores new possibilities and forms of artificial voice which are inspired by things other than science or the market: in this way it contributes to producing new imaginaries around the speaking computer, its sound and its social positioning.

Starting from the arts and expanding to the worlds of gaming, disability and *deepfake*, a new trend in voice synthesis is being affirmed. "Voice cloning" tries to emulate the voice of a specific person with its own peculiarities and defects (Napolitano, 2020b). The peculiarity of voice cloning is that it doesn't sound "generic" as usual synthetic voice, aiming instead to reproduce the features of an embodied and personal voice, the voice of a specific person. It is employed to create realistic fakes of people's voices, as well as to allow gamers using voice avatars known as "voice skins". It is also employed as assistive technology for speech-impaired people (Napolitano, 2021), in order to increase the sense of identity and social recognition for the people who rely on it. "Voice cloning" is not just a technology but a cultural practice with its own imaginary roots and its contradictions. It disrupts the pursuing of the standard voice, although it relies on a strong forensic attitude in the form of biometric data for voice identification (Napolitano, 2020b). Therefore, future studies could be addressed to the imaginary relation between vocal "standards" and voice cloning at a time when we are instead witnessing the proliferation of individualized and customized voices, as in the fields of assistive technology for disabilities and in the arts (Napolitano, 2021).



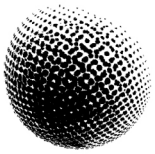
As a conclusion, I would like to attempt an answer to the question I proposed at the beginning of this paper: Do we prefer artificial voices to sound generic or individual? Robotic or human?

The choice clearly does not simply reside in particular propensities implied by the biological structure of the subject, nor is it univocal and fixed over time. The sound imaginaries associated with the artificial voice feed on what happens in the social worlds, in the world of science and programming, in the arts. This intertwining not only influences our preference towards certain voices or others, but constantly feeds the tension between the desire for recognition associated to authority, transcendence and voice in general and the tendency towards the singular, the original, the surprising. This tension is detectable between the anthropomorphism of smart speakers and the most fictional forms of hybridization between human and machine creativity, expressed by new experimental art and deepfake. The generative force of the imagination manifests itself precisely in the search for the unimaginable.

On the background, we can see how artificial voice is itself crossed by a tension: it is like the experience of oddity, surprise, deviation it produces at first listen, as it disconnects the traditional link between voice and human subjectivity, is balanced by its employment in the direction of reification, normalization and standardization of voice.

But as far as voice is reified and commodified through scientific measures, technical renderings and computational reconstructions, the fundamental underlying question remains: which voice is it? This is the question imaginary deals with, as before being techno-scientifically apprehended, voice is first of all imagined as something apprehendible and commodifiable, a quasi-object, an organ (Vallee, 2017).

This suggests that any standard of voice is grounded on a sociotechnical and sonic imaginary, but also that other imaginary voices are lurking.



N. 21– Year XII / July 2023 pp. 177-196
imagojournal.it
ISSN: 2281-8138

Bibliography

Abruzzese A. (2001), *L'intelligenza del mondo. Fondamenti di storia e teoria dell'immaginario*, Roma, Meltemi.

Allen, J., Hunnicutt S. M., Klatt D. (1987), *From Text to Speech: The MITalk system*, Cambridge, Cambridge University Press.

Altman R. (ed.) (1992), *Sound Theory, Sound Practice*, New York, Routledge.

Berger P. L., Luckmann T. (1966), *The social construction of reality: a treatise in sociology of knowledge*, Garden City, NY, Anchor Books.

Bory S., Bory P. (2015), "I nuovi immaginari dell'intelligenza artificiale", *Im@go*, 6, 1: 66-85.

Cavarero A. (2005), *For more than one voice*: *toward a philosophy of vocal expression*, Redwood City, Stanford University Press. Original edition *A più voci. Filosofia dell'espressione vocale*, 1998.

Connor S. (2000), *Dumbstruck: A Cultural History of Ventriloquism*, Oxford, Oxford University Press.

Connor S. (2004), "The Strains of Voice", in Felderer B. (ed.), *Phonorama: Eine Kulturgeschichte der Stimme als Medium*, Berlin, Matthes and Seitz, 158-72.

Cox C. (2011), "Beyond Representation and Signification: Toward a Sonic Materialism", *Journal of Visual Culture*, 10, 2: 145-161.

Davis, E. (1998), *TechGnosis: Myth, Magic, and Mysticism in the Age of Information*, New York, Harmony Books.

Derrida J. (1967), *La voix et le phénomène*, Parigi, PUF.

Di Scipio A. (2013), "Sound object? Sound event! Ideologies of sound and the biopolitics of music", *Soundscape. Journal of acoustic ecology*, 13: 10-14.

Dolar M. (2006), *A voice and Nothing More*, Cambridge MA, MIT Press.

Durand G. (1999), *The Anthropological Structures of the Imaginary*, Brisbane, Boombana Publications. Original edition *Les Structures Antropologiques de l'Imaginaire*, 1960.

Dyson F. (2010), *Sounding New Media: Immersion and Embodiement in the Arts and Culture*, Berkeley, University of California Press.

N. 21– Year XII / July 2023 pp. 177-196 ISSN: 2281-8138
imagojournal.it

Dyson F. (2014), *The Tone of Our Times: Sound, Sense, Economy and Ecology*, Cambridge MA, MIT Press.

Eidsheim N. S. (2018), *The Race of Sound. Listening, Timbre and Vocality in African American Music*, Durham and London, Duke University Press.

Erlmann V. (2004), *Reason and Resonance. A History of Modern Aurality*, New York, Zone Books.

Ernst W. (2012), *Digital Memory and The Archive*, Minneapolis, University of Minnesota Press.

Ernst W. (2018), "Radical Media Archaeology: Its Epistemology, Aesthetics and Case Studies", *Artnodes*, 21: 35-43.

Ernst W. (2021), *Technológos in being*: *Radical Media Archaeology & the Computational Machine*, New York, Bloomsbury.

Faber L. (2020), *The Computer's Voice: From* Star Trek *to* Siri, Minneapolis, University of Minnesota Press.

Flanagan J. L. (1972a), *Speech Analysis, Synthesis and Perception*, Berlin, Springer-Verlag.
Flanagan J. L., Rabiner L. R. (eds.) (1973), *Speech Synthesis*, Murray Hill (NJ), Bell Telephone Labs.

Flanagan J. L. (1982), "Talking with computers: synthesis and recognition of speech by computers", *IEEE Transactions on Biomedical Engineering*, vol. BME-29, n. 4: 223-232.

Grimshaw-Aagaard M., Walther-Hansen M., Knakkergaard M. (eds.) (2019), *The Oxford Handbook of Sound and Imagination*, 2 volumes, Oxford Univeristy Press.

Garcia C. (2018), "Bringing a new voice to genius: MITalk, the Calltext 5010 and Stephen Hawkins' wheelchair", *Computer History Museum*, 26 March, https://computerhistory.org/blog/how-dectalk-gave-voice-to-a-genius-engineering-stephen-hawkings-wheelchair/ (retrieved 15/10/2021).

Gehl R.W., Bakardjieva M. (eds.) (2017), *Socialbots and Their Friends. Digital Media and the Automation of Sociality*, London and New York, Routledge.

Guzman A. L. (2015), *Imagining the Voice in the Machine: The Ontology of Digital Social Agents*, PhD Dissertation, Chicago, University of Illinois.

Hankins T. L., Silverman R. J. (1995), *Instruments and the imagination*, Princenton, Princenton University Press.

Hester H. (2016), "Technically female: Women, machines and hyperemployment", *Salvage Magazine*, https://salvage.zone/in-print/technically-female-women-machines-and-hyperemployment/ (retrieved 19/11/2020).

Humphry J., Chesher C. (2021), "Preparing for smart voice assistants: Cultural histories and media innovations", *New Media & Society*, 23, 7: 1971–1988.

Ingold T. (2011), "Four Objections to the Concept of Soundscape", in T. Ingold, *Being Alive. Essays on Movement, Knowledge and Description*, New York, Routledge, 136-139.

Jasanoff S. (2015), "Future Imperfect: Science, Technology, and the Imaginations of Modernity", in Jasanoff S. and Kim S.-H. (eds.), *Dreamscapes of Modernity. Sociotechnical Imaginaries and the Fabrication of Power*, Chicago and London, The University of Chicago Press, 1-33.

Jasanoff S., Kim S.-H. (2015), *Dreamscapes of Modernity. Sociotechnical Imaginaries and the Fabrication of Power*, Chicago and London, The University of Chicago Press.

Kluitenberg E. (2006), *Book of Imaginary Media: Excavating the Dream of the Ultimate Communication Medium*, Rotterdam, Nai Publishers.

Lingel J., Crawford K. (2020), "«Alexa, Tell Me About Your Mother»: The History of Secretary and the End of Secrecy", *Catalyst: Feminism, Theory, Technoscience*, 6, 1: 1-25.

Lorenzo-Trueba J., Klimkov V. (2019), "Neural text-to-speech makes speech synthesizers much more versatile", *Amazon Science*, https://www.amazon.science/blog/neural-text-to-speech-makes-speech-synthesizers-much-more-versatile (retrieved 10/01/2021).

Männistö-Funk T., Sihvonen T. (2018), "Voices from the Uncanny Valley: How Robots and Artificial Intelligences Talk Back to Us", *Digital Culture & Society*, 4, 1: 45–64.

Marzo P. L. (2019), "La natura immaginaria del sociale: un percorso morfologico", in Marzo P.L., Mori L. (eds.), *Le vie sociali dell'immaginario. Per una sociologia del profondo*, Milano, Mimesis.

McLuhan M. (1964), *Understanding Media*, New York, McGraw Hill.

Mills M. (2012), "Media and Prosthesis. The Vocoder, the Artificial Larynx, and the History of Signal Processing", *Qui Parle*, 21, 1: 107-149.

N. 21– Year XII / July 2023 pp. 177-196 ISSN: 2281-8138
imagojournal.it

Mooney J., Pinch T. (2021), "Sonic Imaginaries. How Hugh Davies and David Van Koevering performed electronic music's future", in Hennion A., Levaux C. (eds.), *Rethinking Music Through Sound and Technology Studies*, New York and London, Routledge.

Mosco V. (2004), *The Digital Sublime. Myth, Power and the Cyberspace*, Cambridge MA, The MIT Press.

Musso M. G. (2019), "Immaginario, tecnologia e mutamento sociale", in Marzo P. L., Mori L. (eds.), *Le vie sociali dell'immaginario. Per una sociologia del profondo*, Milano, Mimesis.

Napolitano D. (2020a), "«Where's the voice of the machine?» An ethnography of artificial voice socio-technical networks", *Etnografia e ricerca qualitativa*, 3: 351-372.

Napolitano D. (2020b), "The Cultural Origins of Voice Cloning", in Verdicchio M., Carvalhais M., Ribas L., Rangel A. (eds.), *xCoAx 2020 Proceedings of the Eighth Conference on Computation, Communication, Aesthetics & X*, 59-73.

Napolitano D. (2021), "Reuniting speech-impaired people with their voices: Sound technologies for disability and why they matter for organization studies", *puntOorg International Journal,* 7, 1, 6-21.

Nass C., Brave S. (2005), *Wired for Speech. How Voice Activates and Advances the Human-Computer Relationship*, Cambridge (MA), MIT Press.

Natale S. (2021), *Deceitful Media. Artificial Intelligence and Social Life after the Turing Test*, New York, Oxford University Press.

Natale S., Pasulka D. (2019), *Believing in Bits: Digital Media and the Supernatural*, New York, Oxford University Press.

Parikka J. (2012), *What is Media Archaeology?*, Cambridge, Polity Press.

Peters J. D. (2004), "Helmholtz, Edison and Sound History", in Rabinovitz L., Geil A. (eds.), *Memory Bytes: History, Technology and Digital Culture*, Durham and London, Duke University Press, 177-198.

Peters J. D. (2015), *The Marvelous Clouds*, Chicago, University of Chicago Press.

Phan T. (2019), "Amazon Echo and the Aesthetics of Whiteness", *Catalyst: Feminism, Theory, Technoscience*, 5, 1, 1-39.

Poizat M. (2001), *Vox populi, vox dei*, Paris, Editions Métailié.

Scha R. (1992), "Virtual Voices", *Mediamatic*, 7, 1, 27-42.

Schafer R. M. (1977), *Soundscape: Our Sonic Environment and the Tuning of the World*, Toronto, McClelland & Stewart Ltd.

Sterne J. (2003), *The Audible Past. Origins of Sound Reproduction*, Durham and London, Duke University Press.

Sterne J. (2011), "Theology of Sound: A Critique of Orality", *Canadian Journal of Communication*, 36, 207-225.

Sterne J. (2012), "Sonic Imaginations", in Sterne J. (ed.), *The Sound Studies Reader*, New York, Routledge, 1-17.

Strengers Y., Kennedy J. (2020), *The Smart Wife: Why Siri, Alexa, and Other Smart Home Devices Need a Feminist Reboot*, Cambridge MA, The MIT Press.

Taylor C. (2004), *Modern Social Imaginaries*, Durham and London, Duke University Press.

Vallee M. (2017). "Technology, Embodiment, and Affect in Voice Sciences: The Voice Is an Imaginary Organ", *Body and Society*, 23, 2, 83-105.

Velasco-Pufleau L. (2021), "Music, Noise and Conflict: Sociotechnical Imaginaries, Acoustic Agency and Ontological Assumptions about Sound", *Journal of the Royal Musical Association,* 146, 2, 501-508.

Woods H.S. (2018), Asking more of Siri and Alexa: feminine persona in service of surveillance capitalism, *Critical Studies in Media Communication*, 35, 4, 334-349.

Wunenburger J.-J. (1991), *L'imaginaire*, Paris, PUF.

Zuboff S. (2019), *The Age of Surveillance Capitalism*, New York, Public Affairs.