xCoAx 2020
Computation,
Communication,
Aesthetics & X
2020.xCoAx.org
Graz, Austria

The Cultural Origins of Voice Cloning

Domenico Napolitano
domenico.napolitano@studenti.unisob.na.it
Suor Orsola Benincasa University of Naples, Italy

Keywords: Voice Cloning, Artificial Voice, Deep Learning, Deepfake, Media Archeology, AI Art, Sound Art.

Inaugurated as a deep learning application on voice synthesis, appropriated by "deepfake" users to create fake Trumps and Obamas and by artists to explore non-anthropomorphic possibilities, voice cloning is not only a technology but a new cultural and artistic practice. Subverting the relation between voice and subjectivity, voice cloning affects the very notions of embodiment and truth. In the wake of media archeological investigation, this paper explores the epistemic properties of voice cloning analyzing its technical and cultural aspects and comparing them with their ancestors, media messages and outcomes.

1. Introduction

Artificial voices are increasingly populating today's technological and social landscape. Their unsettling fascination is related to the archetypical power of voice in human cultures and to its paradoxical topology (Dolar, 2006), always on the edge between inside and outside the body, and between the speaker and the listener. In this sense voice, as suggested by Connor (2002), is always "disembodied", something whose limits are not restricted by the body but can migrate and animate also the inhuman. A condition, this, that feeds voice's unfathomable and suggestive power in cultural practices such as rituals, trials, singing or charismatic speech. In all those situations we assist to a mix: voice is considered at the same time the "natural" means of communication for humans, carrier of phenomenological presence, and an object of treatments and manipulations that enlighten its status of cultural "artifact". The history of artificial voice is therefore, in a certain sense, as old as voice itself, but it's with sound recording technologies and telecommunication that voice has been for the first time "commodified". The possibilities to transport, store and re-enact the voice that those technologies made possible, merged with ideas about disembodiment that go back in time, contributed to enhance the development of the talking computer as a machine that can speak by itself, operating a trespassing from the supposed "proper of human" to the non-human domain. But whereas the disembodied voice of talking computer has been so far accepted by our culture, *voice cloning* is something new and in a certain sense traumatic. The idea to clone someone's voice, to make that voice say things that the person has never "performed", pronounced, immediately makes us think to dispossession, identity theft, fraud. This is probably because the connection between voice and identity is very strong and grounded in our cultural habits. As Adriana Cavarero suggests, anytime I speak I'm voicing myself, no matter what I'm saying (Cavarero 2003). For her, voice is not just language nor just sound matter, but it's their connection within a self and a body. Voice is uniqueness, is the principle of individuation of a singularity. Virtual Assistants, such as Alexa, don't mine the fundamental relation between voice and subjectivity, at the contrary they push a lot on the "personification" of technological devices and AI through the seduction and affection of voice. They just subvert the relation voice-body, without subverting the one voice-subjectivity. Voice cloning, instead, looks like threatening that very principle, that is the "testimonial value" of voice (Peters 2004), the possibility to find in voice a safe warranty of what's real. Whereas truth has been traditionally grounded in the possibility for the subject's self-affection, voice was exactly the place where philosophy individuated that possibility. But, with Derrida's critique of phonocentrism (Derrida 2011), the status of voice itself has changed, and the idea of pure self-affection deconstructed. Nevertheless, the idea of a voice that is not immediately connected to the innermost part of someone's identity is still conflicting with our most basic assumptions of reality, and the possibility to use someone's voice to say something he never said is generally considered worrying.

In the era of analog manipulation, voice was still considered a kind of safe place, something very difficult to fake. For two reasons: both because we have a marked sensitivity to voice that makes us able to recognize even the smaller artifacts in it (Nass & Brave 2005); and because the tools to manipulate voice signals were not adequate to do something so realistic to deceive our sensitivity. Not that attempts of voice manipulation are missing in the analog media landscape. At the contrary, in line with Connor's considerations, the special and ambiguous status of voice has made it a privileged place of experimentation of new techniques and sensitivities, both in the art and in other fields, like linguistic and forensic studies, since very long time.

I would like to start from here, from the review of some past attempts of voice cloning, to track down the cultural origins of this new technological phenomenon. Adopting a media archeological approach (Parikka 2012; Ernst 2012), the study aims at recognizing the common fantasies and desires related to this practice, but also at finding the epistemological ruptures and specificities brought by technologies such as deep learning and artificial neural networks—the algorithmic core of voice cloning—which incorporate precise knowledge and ideas in their very functions. It's my belief that reading this difference is a decisive key to understand voice cloning as a new cultural practice in all respects, very meaningful of the new status of AI in contemporary society. Here deconstructive instances about the voice-subjectivity bond (such as "deepfake") are merged with brand new cognitive and expressive media configurations and socio-technical relations. If medium is the message (McLuhan 1994), neural networks in voice cloning define voice in a specific way and determine how we think to it and how we use it. "When ideas about bodies are built into digital signals, these signals, in turn, produce bodily effects" (Mills 2012, 136); symmetrically, when ideas about voice are built into digital processing, this processing, in turn, produces effects on voice and its physiological and socio-cultural determination.

2. Archeology

The term "voice cloning" has been introduced for technical purposes, in the framework of *deep learning* applied to text-to-speech technology. The term "deep learning" indicates the most recent application of machine learning. Where machine learning is used to map hand-designed features (i.e. labeled voice samples) to an output (i.e. words), so allowing machine to "learn" couplings and associations in order to generalize them to new cases (i.e. matching voice samples to a new text), deep learning uses multiple hierarchical layers of neural networks to progressively extract higher level features from raw input. Instead of having hand-designed features, as in machine learning, deep learning is used to extract those very features from data. This is very useful for all the tasks that are difficult to model or where it's difficult to know which features should be extracted. Voice speech is a perfect example of a hardly representable phenomenon, since every word can be pronounced in several different ways, with different intonations, speeds, accents, timbres. Deep learning helps solving this problem since it finds the appropriate features by itself directly from a big enough dataset. It does it by processing data many times through several hierarchical layers, each of them going deeper detail: lower layers may identify formant frequencies of voice, higher layers may identify consonant sounds, higher modulations and prosody, and so on until having a representation of all the countless features needed to reproduce voice from a text. “Deep learning is a particular kind of machine learning that achieves great power and flexibility by learning to represent the world as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones” (Bengio et al. 2017, 8).

In voice cloning, a deep neural network is usually trained using a corpus of several hours of professionally recorded speech from a single speaker. Giving a new voice to such a model is highly expensive, as it requires recording a new dataset and retraining the model. Deep learning allows to clone a voice unseen during the training from only a few seconds of reference speech, and without retraining the model (Jemine 2019; Arik et al. 2018). In this way, a text-to-speech can read a typed sentence with the voice that the algorithm has “learned” from the reference dataset. Previous voice synthesis systems were working on formant and articulatory parameters (parametric speech synthesis; Flanagan 1972; Klatt 1982), or on database of recorded voice samples (concatenative speech synthesis or unit selection; Sagisaka 1988), eventually mixing the two approaches, where parameters are used to statistically calculate the units of recorded sounds that best fit the proposed text in trainable systems (statistical parametric speech-synthesis; Donovan 1998; King 2010). For a technical and historical overview of these systems, see (Jurafsky & Martin 2014; Hoffmann 2019). Taking from the previous systems, but with a number of significant differences, deep learning voice synthesis with neural networks uses datasets of texts and recorded voice samples not to directly concatenate them (as in concatenative synthesis), but to “learn” from the dataset the main features of a voice in relation to text, in order to be able to recreate it in new sentences (Sejnowski & Rosenberg 1987; van den Ord et al. 2016; Seijas 2018). Neural networks do that by “classifying” the main features of a voice, like pronounce, timbre, prosody, and separating them from each other in deep detail. Those features constitute a “speaker profile”, that is the elements that distinguish a certain voice from another. This is the idea at the base of “voice cloning” as well. In fact, a deep neural network can learn a “speaker profile” through a training process, then use it to condition a text-to-speech, already trained on the linguistic features. In this way a text-to-speech can use the speaker profile to customize the “generic” voice it has been trained with and make it sound like the voice of a certain precise person. If text-to-speech is the invariable structure, voice cloning is the "mask", that can be changed at will. This means that a text-to-speech algorithm can be trained on different voice timbres with small training sets, making it easy and relatively "user-friendly".

Voice cloning is nowadays advertised as a way to "personalize" things like customer services, chatbots, videogames, IoT devices, in order to make them more appealing through the seductive power of a real person's voice. "'Resemble' can clone any voice so it sounds like a real human" (www.resemble.ai). Before discovering its subversive power in the "deepfake" (Wilson 2018), voice cloning technology was developed in 2017 by Canadian startup company Lyrebird to develop software that reconstructs someone's voice in audiovisual files. As a test of the efficiency of their system, they tried to clone Donald Trump's voice training their neural networks with Trump's recorded speeches. They succeeded in using Trump's voice to say a couple of realistic sentences about their product. But a similar idea sparkled already in the mind of a composer who lived almost a century before. The result is unknown, nevertheless the episode helps understanding the cultural meaning of voice cloning.

In 1932 the Russian composer, musical theorist and journalist Arseny Avraamov proposed to vocalize the writings of Lenin by reproducing the author's voice using new technological means. The fact is reported and documented by Smirnov (Smirnov 2012). In order to do that, Avraamov needed to "synthesize" Lenin's voice on the basis of his existing recorded speeches. He proposed to achieve a "possible vocalization of mute pieces of the Lenin's chronicle, by precise assignment of fragments of the shorthand report uttered by him in each particular moment of speech" (Avraamov quoted in Smirnov 2012, 163). Somewhat later in 1943, Avraamov argued also against the new Soviet anthem, contending that the real revolutionary anthem should be based on new approaches to harmony and performed by the synthesized voice of Vladimir Mayakovsky. This happened at the end of Soviet Russia's political and artistic turmoil that accompanied Russian Revolution, when ideological totalitarianism started to grow and put an abrupt end to all the avant-garde spirit in science and the arts. The influence of major artistic figures such as Alexander Scriabin, Dmitry Shostakovic in music, Sergei Eisenstein in film, Vladimir Mayakowsky, in poetry, Kandinskij, Malevic in painting, was still inspiring for a generation of artists and intellectuals ready to experiment new languages and practices in the wake of revolutionary ideals (Smirnov 2013). Arseny Avraamov was active part of this artistic and political movement, spending his life experimenting with new techniques of sound and image inscription.

What Avraamov had in mind with his vocalization of Lenin's writings was something very similar to that kind of particular speech synthesis that nowadays we would call "unit selection", or concatenative synthesis. This technique is based on the recombination of stored speech fragments, usually diphones or sound units, according to the text that needs to be read and its phonetic transcription. Avraamov's idea of a proto-concatenation was grounded on the new discovery in "sound storing" of the time: graphical sound. After the invention of Edison's phonograph in 1877, sound recording was mostly made using soft tinfoil and wax cylinders (Feaster 2011). In 1898 Vladimir Poulsen introduced magnetic wire recording with steel wire for his Telegraphone, an ancestor of magnetic tape. In 1914 a new technology for sound recording and storing was introduced in top-secret navy projects: Thallofide Cell was a device by Theodore Case to record optical sound, that is sound recorded on film through variations in light produced by sound oscillations. The year after, Charles Hoxie used optical sound in silent movies with his Pallophotophone, and in 1922 De Forset Phonofilm company introduced optical sound as a standard in film production (Kellogg 1955).

By 1932, when Avraamov wrote about Lenin's voice, optical sound started to be widely explored in Germany and Russia. If German sound engineer Pfenninger succeeded in correcting a misspell of an actress voice recorded in a movie (Levin 2012), Russian artist Moholy-Nagy saw in optical sound a means for the generation of new, unheard, "synthetic" sounds and Russian painter and acoustician Boris Yankovsky, Avraamov's pupil, explored the techniques of spectral transformations of singing and speech through manual drawing of waveforms on the film, using a self-invented machine called Vibroexponator (Smirnov 2013). At the basis of all those results there's graphical sound research, an application of mathematical functions, such as Fourier transform, to waveforms drawn on film. This technique made it possible to access the soundtrack as a visible graphical trace in a form that could be studied and manipulated, through magnification, hand-drawing in grid and then shrinking, allowing to synthesize new and unheard sounds.

As reconstructed by Smirnov, in the early 1930s, voice synthesis was strangely popular in Russian avant-garde. "There were two main intentions focused at the development of new sound machines: to play and compose with any sounds at will and to synthesize speech and singing" (Smirnov 2012, 166). Besides optical sound, other machines were invented to "synthesize voices", like the Mechanical keyboard instrument for the reproduction of speech, singing and various sounds, the most advanced proto-speech synthesizer at the time, invented by Tambovtsev in 1925. The instrument was primarily intended for reproduction of artificial speech and singing: each key of its keyboard corresponded to a loop of steel tape which stored a sound of Russian language, prerecorded with different pitches, corresponding to different keys on the keyboard. "It was a kind of proto-sampler, very similar to the Mellotron, popular in 1970s" (Smirnov 2012, 166), but it was also a concatenative speech synthesizer *ante-litteram* (it's of 1960 the System LORA by Cramer, which used a similar system with 40 pieces of magnetic tape (Hoffmann 2019)).

Regardless of the results, both sound drawing and the mechanical keyboard embody a principle that will be decisive for the epistemology of sound manipulation, that is for the very possibility of thinking to sound and voice synthesis. This principle is the *assemblage*, the idea that a new sound can be composed by juxtaposing fragments of other sounds, and that speech can be conceived as a concatenation of phone units. This revolutionary principle, probably one of the most important in electronic sound after Fourier's transform and von Helmholtz's resonators, sets the conditions for sound editing. At its basis there's not only a concept of sound, but a *logic of the archive*. It's only because of storage and archive that sound can be edited, sliced in fragments, and recomposed. Nothing like that was possible before sound recording and the possibility to store and archive sound on physical supports.

There's also a key difference between Tambovstev's keyboard and Yankovsky's Vibroexponator. The former used only recorded sound materials, that are acquired from the "real world" through a transducer or microphone, while the latter could produce sounds from nothing, from just drawing, in a computational (even if manual) manner. In this sense, Tambovstev could be seen as an ancestor of *musique concrète* as music based on recordings (Schaffer, 2012), while Yankovsky as a father of electronic music, based on pure synthetic sounds. But these the two genealogic lines share a fundamental character, a new attitude towards the archive, that is, in the interpretation that I'm suggesting here, the beginning of the "database logic", as Manovich defined it (Manovich 2002, 219). One of the main purposes of Yankovsky, in fact, was to produce and collect a number of *syntones*, that is pieces of drawn sound on film, in order to recomposed them to produce new sounds or new speeches. He wanted to collect a *database* of sound materials, ready to be used at will. "Yankovsky named these final drawn waveforms 'spectro-standards' or 'spectral templates', semiotic entities that could be combined to produce sound hybrids, based on a type of spectralmutation" (Smirnov 2012, 170). Even if his technique was different from proper sound recording, the logic of his work was the same of Tambovstev, a database logic. "Synthesis" means, here, two symmetric operations: a) creation of an archive of recorded or drawn sound materials; b) operationalization of that archive through its recomposition and assemblage. This kind of logic has been assumed, almost unchanged, by computation in digital devices. Arseny Avraamov was involved in all those researches about sound and voice synthesis and took inspiration from them to imagine his "cloned Lenin". His idea was to mix the two operations: using optical sound to collect sound samples from Lenin utterances, and then operationalizing them in the construction of new sentences by concatenation and editing of film pieces; the concatenated syntones could also be hand-drawn again with slight differences to give intonation and rhythm to the new synthesized speech.

The idea of sound archives grew rapidly at the cross of XIX and XX century, influencing many fields of society beyond the art world. Mara Mills and Xiaochang Li have reconstructed the technical and epistemological link between sound archive and sound inscription, meaning the original possibility to "see the sound" that new devices allowed (Mills & Li 2019). From Eduard-Leon Scott de Martinville's Phonautograph (1857) to W. H. Barlow Logograph (1877) to Goddard harmonic analysis of kymographic inscriptions (1903), to Carl Lindstrom's Parlograph (1910), all those tools tried to transform sound in visible traces. Optical sound and drawn sound, as in Yankovsky's Vibroexponator, were doing essentially the same. The next step in this story is the invention of the *sound spectrograph* in the 1940s, a new way to visualize sound as a time-frequency representation: time on the horizontal axis, frequency on the vertical one, and loudness indicated by the intensity of the ink or light patterns. When the technology was commercialized after the war, linguists as well as communication engineers used spectrograms to identify the landmarks or key features within speech waves. One group of researchers, at Haskins Laboratories in New Haven, proposed compiling a large collection of spectrograms for each speech sound (Mills & Li 2019). It looks like the idea of sound database is strongly connected to the one of sound visualization. What is common to all those devices of sound inscription is their use in forensic context.

> “By examining numerous spectrograms of the same sounds, spoken by many persons and in a variety of contexts, an investigator can arrive at a description of the acoustic features common to all of the samples, and in this way make progress toward defining the socalled invariants of speech, that is, the essential information-bearing sound elements on which the listener's identifications critically depend” (Mills & Li 2019, 132).

This was the beginning of the idea of "voiceprint", a visualization of someone's vocal emission that should have allowed to individuate criminals from the analysis of their voice. In the opening decades of the twentieth century, most anthropologists and criminologists took graphic inscription as evidence that humans could not disguise their unique voices and ethnic origins. Spectrograms and other sound inscriptions and sound analysis tools became means for criminal identification or for speaker individuation by military in intercepted communications. The introduction of database of sound visualization for forensic purpose, together with the practice of concatenation of stored sounds on film, constitute the conceptual and technical base for what today we call voice cloning. As I'll try to explain now, of the two, only the more discrete is still at work, while the other, long celebrated, is now leaving the way to a different paradigm.

3. Media Practices and Epistemes

3.1. Speaker Identification

This attempt of a media history of voice cloning, so quickly sketched, reveals an interesting parallel between old and new systems: technically speaking, in fact, voice cloning with deep learning consists in the union of an algorithm for "speaker verification" with a text-to-speech (Jemine 2019). My critical suggestion is the following: it is not in the database logic and in the practice of assemblage, but in the persistence of an idea of voiceprint and speaker recognition, that we find a continuity between sound recording and voice cloning. This continuity reveals a latent forensic attitude in voice processing that is not disconnected from a paradigm of control and surveillance embedded in contemporary algorithmic technologies (Bucher 2018; Andrejevic 2020). It is also the materialization of the persisting fantasy of a "commodification" of voice, a voice to be detached from the body, measured and reattached again, allowed first by sound recording, and then confirmed and reaffirmed by voice cloning. An old fantasy that produces a new relation between voice-as-signal—to be stored, manipulated, reassembled, etc.—and subjectivity—a body that can now have multiple voices (as in voice conversion) or lose the control on its voice. As stated by Jonathan Sterne,

> "voice-as-exteriority formation is at least two hundred years old. Both the fields of acoustics and medicine treated the voice as something separate from an intending, speaking subject since the eighteenth century. Nineteenth-century innovations in sound technologies and the education of the deaf that led to telephony, radio, and sound recording followed in this vein" (Sterne 2008, 96).

The forensic use of voiceprints began in the beginning of the XX century but has never been established as a scientific practice and it's still nowadays a technological challenge, as well as a controversial ethical and political issue. Today's machine learning and deep learning systems seem like doing great steps ahead in the possibility to recognize someone's identity through his voice in reliable way.

From a technical point of view, neural networks are classifiers that are able to find their own representations in raw data, that is in a not-labeled dataset (Bengio et al. 2017). Those representations are in the form of a nested hierarchy of simpler representations, organized as topological distribution of numerical vectors in the latent space (the space where hidden layers make their calculations). This operation of "classification", that is the deep analysis of the input signal in order to sort any minimal element as for its similarity or difference with any other, is the basis of the "learning". The "learning" in deep learning is first of all an operation of "sorting". This tells also something about the general attitude of data-driven approach to AI, as machine learning is. Through these classifications, deep learning can solve the problem of speaker identification: the speaker verification algorithm classifies voice features in a big training dataset, according to differences in parameters detected by the neural networks. After the training, it can sort, in any voice signal, which features go always together and which can be separated, i.e. a certain way of pronouncing subsequent phonemes can be fix, while the timbre (formant frequency and other parameters) can be variable according to the speakers, and so on. Through this process the algorithm can encode a “speaker profile” or “embedding”, “a meaningful representation of the voice of the speaker, such that similar voices are close in latent space” (Jemine 2019, 12). Once obtained a speaker embedding, it can be used to synthesize new speech with those voice features. This operation is what today we define as “voice cloning”. To do that it’s enough to condition a text-to-speech synthesizer such as WaveNet (van den Ord et al. 2016) or Tacotron2 (Shen et al. 2017) on the embedding of the speaker.

In a technical sense voice synthesis with deep learning is always a kind of voice cloning. But in a media archeological sense, the speaker verification algorithm suggests that deep learning is answering to an old call about the “appropriation” of such an elusive and powerful object such as voice. An appropriation that opens also critical concerns about issues such governmentality, control and privacy.

3.2. Different Logics of the Archive

On the other side, my proposal is that voice synthesis and voice cloning with deep learning are producing a rupture with the database logic, inaugurating a *different logic of the archive*. As Manovich observed, “database became the center of the creative process in the computer age” and a new way to structure our experience of ourselves and of the world (Manovich 2001, 227). Database is therefore part of a precise cultural and expressive practice. This practice is based on the “operationalization of the archive”: where the archive is a set of stored information, database is the organization of archived data to facilitate operations on it. The database is a modulation of the archive in the form of a set of individual items that can be re-assembled to create new items. The database logic, therefore, finds expression in cultural techniques such as assemblage, montage, remix, “cut’n’paste”. We find this logic at work in sound editing as well as in voice synthesis. In concatenative synthesis or “unit selection”, a database of archived sound samples is algorithmically assembled in many possible ways, so that the same archived voice can say something different every time. Whereas the archive is static and “says always the same thing”, the database logic can continuously recreate it as something new. This is the power of this cultural expression, a virtual regeneration of presence (Ernst 2012).

With machine learning, and deep learning in particular, something different is happening. We have a database, of course, and it’s often based on spectrograms, in a continuity of format with the old systems. But this database is now used as a *training dataset*; it means that those data are not re-assembled, but are “learned”, the algorithm understands something from them, extract some features that is then able to use in new contexts (i.e. new sentences). The database is not just copied and reassembled in the outcome. Rather, the training database is only used to “feed” the algorithm, but is not present anymore in the outcome. The algorithm will learn from it to then generate "new" or "unknown" items, which are technically different from the dataset because not made of the same data.[1] Voice here is not assembled but "reconstructed", not according to a perceptual principle, but according to a numerical rendering, that learns something about that voice and looks for the optimal way to use it in the context. This process is really specific of deep learning, since only the classifying power of neural networks is able to separate voice timbre from all the other features through an interpolation in the latent space. If we had a series of sound samples and try to do that with classic sound editing, we would be stuck: or we could use just the pre-recorded sentences, or we would have other features accompanied to timbre, in a muffled crossfading of signals.[2]

To resume, we can find three different logics of archive: *model-based logic*, where the knowledge is embedded in the model, such as in the machines or *automata* who wanted to reproduce human behavior; *corpus-based or database logic*, where the knowledge is stored in micro-archives of sound samples and then operationalized; *machine learning or abductive logic*, where the knowledge is rendered and recreated in its unfolding, through a training process. What differentiates machine learning from the historical use of databases is that the former is meant to generate previously unknown patterns that cannot be perceived prior to running the algorithms. It is a tool for simulation, not in the sense of modeling or imitating an existing reality, but rather in that of generating a process as unpredictable as reality (Andrejevic 2013, 37).

3.3. New Expressive Practices

Machine learning gives life to a new expressive practice: not assemblage, as in the database logic, but *hybridization*, the application of a model relative to a class of events to one or more other classes, or *chimerization*, a process where a hybrid is generated with genetic fusion of multiple distinct entities. I suggest the use of these two terms because, as biological concepts, they represent quite well the ambition of AI to biologic life, or at least its rhetoric. Already Manovich (2013) has adopted the term "hybridization" to refer to the capacity of software to combine together properties and techniques of different media. In this framework I use the term with but also beyond Manovich, underlining an attitude in data processing that precedes the phenomenology of different media. Therefore, I suggest to use hybridization and/or chimerization as metaphors to describe a shift: the replacement of the practices of data juxtaposition and "remix", with the practices of statistical rendering, estimation and optimization allowed by the abductive power of neural networks (Kitchin 2014). Like in abductive reasoning (Josephson 1994), in fact, neural networks start from the observation of data (the training) and seek to find the optimal approximation (the learning), which is, consistently, something new and emergent, even though uncertain in causal terms. This practice produces something not present in the input data, nor

[1] In this paper I leave unquestioned the complex ontological issue of the identity regarding a fundamentally differential concept such as data. For a detailed study of the question see (Floridi, 2011).

[2] One of the consequences of this process is a copyright issue: who has the copyright of a training dataset which, technically, doesn't figure in the output?

composed by it. Moreover, this new product can be the result of any kind of observation between any kind of data, even completely heterogeneous ones: neural networks don't mind categories, they will always find a correlation as far as phenomena are transformed in numbers. The result is that classes of completely unrelated events can now be hybridated together, giving life to something that completely trespasses classes and semantics, and doesn't belong to any precise category anymore.

"Neural style transfer" is the name of the new artistic practice derived from the use of deep neural networks in a hybridizing sense. It is a technique "of recomposing images in the style of other images", such as Monna Lisa restyled by Picasso or van Gogh, as in the works of Gene Kogan based on Gatys, Ecker and Bethdge algorithm (2015). "Voice conversion" is a style transfer technique applied to voice, a special application of voice cloning where you can hybridize the timbre of someone's voice with the prosody of someone else, as proposed by the company Modulate.ai. Sound artist Tomomi Adachi has cloned his own voice in "Tomomibot", an AI that has learned Tomomi's vocal improvisation styles and can play live, establishing a dialogue with the "embodied" artist. Jenna Sutela artwork *nimiia cétiï* documents the interactions between audio recordings of supposed Martian language, and footage of the movements of extremophilic bacteria. A neural network trained on her voice looks at each frame of the bacteria video and produces a short block of sound that it thinks matches that frame, or the configuration of bacteria in it. Here, the computer is a medium channeling messages from entities that usually cannot speak. The work shows how neural networks' creations are aliens, monsters or hallucinations, confusing the borders between natural and machinic. James Bridle's media art work "The Cloud Index" is another very meaningful example of this new approach. "The Cloud Index" is a piece of software that can be used to create different weather formations based on different political outcomes. To develop the work, Bridle fed a neural network with satellite images of the UK's weather formations and Brexit polling results that showed the UK's relationship to Europe.

In the previous examples, unrelated classes of phenomena are combined together and let dynamically grow on each other. The functions of body and subjectivity, such as voice or language, are now made equivalent to social or natural phenomena, and can be hybridized freely with anything else. But, as Jenna Sutela suggests: "the aim is to contribute to the development of a culture based on symbiosis rather than the survival of the fittest narrative—organic and synthetic life forms included" (Sutela 2019). This sounds consistent with a project of global chimerization, that should open the way to new forms of co-existence and collaboration between human and nonhuman. As a consequence, this technology redefines voice itself: voice is not only separated from the body, as in sound recording, but can now transmigrate on other bodies, can be hybridized with any kind of psycho-physical features, while being filtered out of certain speaker traits.

The introduction of deep learning in voice synthesis was motivated by the possibility to do natural-sounding realistic voices, while its media analysis reveals a connection with forensic and policing techniques. Nevertheless, one of the most interesting cultural outcomes of deep learning specific power can be seen in the indefinite possibility of hybridization, transfer and invention of "new", impossible voices. The rise of "deepfake", both in creative expressions and in fraudulent operations, expresses very well the importance of the new challenges opened by this power; challenges that concern the very status of the truth in a society populated by synthetic media (Wilson 2018).

4. Conclusion

In this paper I've presented a preliminary study of a cutting-edge socio-technical phenomena that deserves further investigations. The media-archeological method has allowed to retrace the cultural and technical origins and the epistemological conditions of that odd idea that is cloning someone's voice. But more work should be done on the epistemic continuities and ruptures of machine learning in the field of sound processing. Older ideas could be, for example, individuated behind voice cloning. As suggested by Wolfgang Ernst, a media-archeological ancestor could be found in Jaynes' theory of bicameral mind (Ernst 2016), according to which in pre-writing times people could hear proper "voices" of dead kings in their heads, as a form of behavior control exercised from the inside. Without entering the articulated debate around the scientific validation of Jayne's theory, I will limit to suggest that the wish (or the obsession) to reproduce the voice of the "leaders", as in the case of Lenin or Trump, could perhaps respond to a similar need to find "his master's voice" (Dolar 2006). The leader is the one who "gives body" to the voice, is the depositary of the truth in form of acoustic experience, his voice comes before the meaning because it's legitimated by his very presence. If synthetic voices, in their unsettling being on the edge of organic and technologic, manifest and represent the anthropological condition of uncertainty produced by media, the leader's voice could perhaps represent a reactionary attachment to old values. Voice cloning, therefore, can be considered like both the emblem and the risk of a high-tech society: it is the place where a short-cut happens within the traditional "power" of voice (charisma, seduction, interiority), demystifying the qualities of the metaphysical embodied subject; but it is also the place of a paradoxical recovering of the link between voice, body and subjectivity, this time in form of gadget.

References

Andrejevic, Mark.
2020. *Automated Media*. New York: Routledge.

Andrejevic, Mark.
2013. *InfoGlut: How Too Much Information Is Changing the Way We Think and Know*. London: Routledge.

Arik, Sercan, Jiton Cheng, Kainan Peng, Wei Ping, Yanqui Zhou.
2018. "Neural Voice Cloning with a Few Samples". arXiv:1802.06006

Bengio, Yoshua, Ian Goodfellow, Alan Courville.
2017. *Deep Learning*. Cambridge, MA: The MIT Press.

Bucher, Taina.
2018. *If...Then. Algorithmic Power and Politics*. New York: Oxford University Press.

Cavarero, Adriana.
2003. *A più voci: Per una filosofia dell'espressione vocale*. Milano: Feltrinelli.

Connor, Steven.
2000. *Dumbstruck: A Cultural History of Ventriloquism*. Oxford: Oxford University Press.

Derrida, Jacques.
2011. *Voice and Phenomenon*. Evanston: Northwestern University Press.

Dolar, Mladen.
2006. *A Voice and Nothing More*. Cambridge, MA: The MIT Press.

Donovan, Robert E., Ellen Eide.
1998. "The IBM Trainable Speech Synthesis System". *ICSLP-98*, Sydney.

Ernst, Wolfgang.
2012. *Digital Memory and the Archive*. Minneapolis: University of Minnesota Press.

Ernst, Wolfgang.
2016. *Sonic Time Machines. Explicit Sound, Sirenic Voices and Implicit Sonicity*. Amsterdam: Amsterdam University Press.

Feaster, Patrick.
2011. "A Compass of Extraordinary Range". The Forgotten Origins of Phonomanipulation". *ARSC Journal*, vol. 42, n. 2, pp. 163-203.

Flanagan, James H.
1972. *Speech Analysis, Synthesis and Perception*. Berlin: Springer.

Floridi, Luciano.
2011. *The Philosophy of Information*. Oxford: Oxford University Press.

Gatys, Leon A., Alexander S. Ecker, Matthias Bethge.
2015. "A Neural Algorithmic of Artistic Style". arXiv:1508.06576.

Hoffmann, Reudiger.
2019. "Nothing but a Lung, a Glottis, and a Mouth. The Long Way of Speech Synthesis", In Pucher, M., Trouvain, J., Lozo, C. (eds.). *HSCR 2019. Proceedings of the 3rd Int. Workshop on the History of Speech Communication Research*, Vienna, September 13-14, 2019. Dresden: TUD press, pp. 9-28.

Jemine, Corentin.
2019. *Real-time Voice Cloning*, Master Dissertation, Université Liège.

Josephson, John R., Josephson, Susan G., eds.
1994. *Abductive Inference: Computation, Philosophy, Technology*. New York: Cambridge University Press.

Jurafsky, Daniel, James H. Martin.
2014. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. London: Pearson.

Kellog, Edward W.
1955. "History of Sound Motion Pictures". *SMTPE Journal*, vol. 64.

King, S.
2010. "A Beginners' Guide to Statistical Parametric Speech Synthesis". *The Centre for Speech Technology Research*. University of Edinburgh.

Kitchin, Rob.
2014. "Big Data, New Epistemologies and Paradigm Shifts". *Big Data & Society*, April-June, pp. 1–12.

Klatt, Dennis.
1982. *From Text to Speech: The MITtalk System-* Cambridge, MA: Cambridge University Press.

Levin, Thomas Y.
2003. "'Tones from Out of Nowhere': Pfenninger and the Archeology of Synthetic Sound". *Grey Room*, vol. 12, pp. 32–79.

Manovich, Lev.
2001. *The Language of New Media*. Cambridge, MA: MIT Press.

Manovich, Lev.
2013. *Software Takes Command*. New York: Bloomsbury.

McLuhan, Marshall.
1994. *Understanding Media: The Extensions of Man*. Cambridge, MA: The MIT Press.

Mills, Mara.
2012. "Media and Prosthesis: The Vocoder, the Artificial Larynx and the History of Signal Processing". *Qui Parle*, vol. 21, n. 1, pp. 107-149.

Mills, Mara, Xiaochang Li.
2019. "Vocal Features. From voice identification to speech recognition by Machine". *Technology and Culture*, vol. 60, pp. 129-160. DOI: 0040-165X/19/6002-0006/S129–S160.

Nass, Clifford, Scott Brave.
2005. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. Cambridge, MA: The MIT Press.

Parikka, Jussi.
2012. *What Is Media Archeology?* Cambridge: Polity Press.

Peters, John Durham.
2004. The Voice and Modern Media. In D. Kolesch and J. Schrödl (eds.), *Kunst-Stimmen*. Berlin: Theater der Zeit Recherchen 21.

Sagisaka, Yoshinori.
1988. "Speech Synthesis by Rule Using an Optimal Selection of Non-uniform Synthesis Units". *IEEE ICASSP-88*, pp. 679–682.

Seijas, Jesùs.
2018. "Into a Better Speech Synthesis Technology". https://becominghuman.ai/into-a-better-speech-synthesis-technology-29411b64f2a2.

Seijnowski, Terrence, Charles Rosenberg.
1987. "Parallel Networks that Learn to Pronounce English Text. *Complex Systems*, 1, pp. 145-168.

Shaeffer, Pierre.
2012. *In Search of a Concrete Music*, Berkeley: University of California Press.

Shen, Jonathan, et al.
2017. “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions”. eprint arXiv:1712.05884.

Smirnov, Andrey.
2012. Synthesized Voices of the Revolutionary Utopia: Early Attempts to Synthesize Speaking and Singing Voice in Post-Revolutionary Russia (1920s). In D. Zakharine & N. Meise (eds.), *Electrified voices. Medial, Socio-historical and cultural aspects of voice transfer.* Göttingen: V&R Unipress.

Smirnov, Andrey.
2013. *Sound in Z. Experiments in Sound and Electronic Music in Early 20th Century Russia*. London: Koening Books.

Sterne, Jonathan.
2008. “Enemy Voice”. *Social Text 96*, vol. 25, n. 3, pp. 79-100. DOI: 10.1215/01642472-2008-005.

Sutela, Jenna.
2018. “Jenna Sutela on Machine Learning and Interspecies Communication”. https://artsandculture.google.com/theme/lQKy0vx84f5GIg.

van den Ord, A., S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A.W. Senior, K. Kavukcuoglu.
2016. “WaveNet: A Generative Model for Raw Audio,” CoRR, vol. abs/1609.03499.

Wilson, Mark.
2018. “The War on What’s Real”. *The fast Company*. 3 June 2018. https://www.fastcompany.com/90162494/the-war-on-whats-real.