SoundEffects
An Interdisciplinary Journal of Sound and Sound Experience

Domenico Napolitano
PhD candidate
University Suor Orsola Benincasa of Naples

&

Renato Grieco
Dr.
Conservatory San Pietro a Majella of Naples

The folded space of machine listening

www.soundeffects.dk

SoundEffects | vol. 10 | no. 1 | 2021 issn 1904-500X

Abstract

The paper investigates new machine listening technologies through a comparison of phenomenological and empirical/media-archeological approaches. While phenomenology associates listening with subjectivity, empiricism takes into account the technical operations involved with listening processes in both human and non-human apparatuses. Based on this theoretical framework, the paper undertakes a media-archeological investigation of two algorithms employed in copyright detection: "acoustic fingerprinting" and "audio watermarking". In the technical operations of sound recognition algorithms, empirical analysis suggests the coexistence of a multiplicity of spatialities: from the "sound event", which occurs in three-dimensional physical space, to its mathematical representation in vector space, and to the one-dimensional informational space of data processing and machine-to-machine communication. Recalling Deleuze's definition of "the fold", we define these coexistent spatial dimensions in techno-culturally mediated sound as "the folded space" of machine listening. We go on to argue that the issue of space in machine listening consists of the virtually infinite variability of the sound event being subjected to automatic recognition. The difficulty lies in conciliating the theoretically enduring information transmitted by sound with the contingent manifestation of sound affected by space. To make machines able to deal with the site-specificity of sound, recognition algorithms need to reconstruct the three-dimensional space on a signal processing level, in a sort of reverse-engineering of the sound phenomenon that recalls the concept of "implicit sonicity" defined by Wolfgang Ernst. While the metaphors and social representations adopted to describe machine listening are often anthropomorphic – and the very term "listening", when referring to numerical operations, can be seen as a metaphor in itself – we argue that both human listening and machine listening are co-defined in a socio-technical network, in which the listening space no longer coincides with the position of the listening subject, but is negotiated between human and nonhuman agencies.

Keywords
Machine listening, listening space, media archeology, acoustic fingerprint, audio watermarking, copyright detection algorithms.

Introduction

Some months ago, one of us uploaded a video to YouTube, shot at a birthday party with a Madonna song blasting from the DJ's speakers. The party was held in a big, reverberant room, which naturally had a strong impact on the sound of the song, and there was heavy background noise as well. Despite all of this, YouTube's copyright detection algorithm recognized the song and removed the video due to unauthorized use of copyright protected material.

This very common experience blatantly reveals the widespread presence of machine listening in today's media landscape, while at the same time highlighting the very special role played by the space in this phenomenon. Manuals of acoustics teach us that sound can exist only in space, that space is the medium of sound, its condition *sine qua non*. However, in sound recordings, the space is actually a sign, a mark, even a *symptom* of the "event of sound" (Di Scipio, 2013, p. 12), that is, of the travel of sound in a precise space and at a precise time or, more technically, of the vibration of certain air molecules at a certain time under certain circumstances, as perceived from a certain point in the space by a specific device. But where is the space when YouTube "listens" to the sound of the birthday party? What is the relation between sound and space when sound is processed "as data"?

This paper aims to investigate new machine listening technologies as regards their relation to the issue of space. If speech recognition and acoustic event detectors, ever more popular in our houses and cities, have been assimilated with "artificial ears"[1] – since they catch sounds "in the space" and automatically "recognize" sonic events, such as cries, shotguns, sirens – algorithms, such as copyright detectors, acoustic fingerprinting and audio watermarking, but also automatic captions, working on the mere level audio data, dissolve the physical bond between sound and space. Our critical proposal is that, in this dissolution, those systems are actually enacting a different relation between sound and space, in which the space is virtualized and reconstructed at the software level in order to simulate a "situated listening" (Biancorosso, 2016, p. 11).

All media rely on specific "representations of space" (Sterne, 2015, p. 113), and machine listening relies on the simulation of the perceptual effect that space has on sound for a hypothetic human listener, assumed to be a model for the algorithm. At the same time, once it has been transduced into signal, the material operations of algorithms on sound open towards another space, a space of data processing and machine communication. Recalling the concept of "the fold" elaborated by Deleuze (1993), we define this condition as "folded space" since the algorithms must encompass different coexistent dimensionalities of space: the three-dimensional space in which the sound is captured by sound recorders and acoustic detectors; the two-dimensional space of the multimedia contents subjected to copyright detection, closed captions, or automatic recognition, such as in the case of music in videos; the one-dimensional space of data processing proper to machine-to-machine communications, in which the vibrational force of sound is translated into the "implicit sonicity" (Ernst, 2016, p. 25) of information. In machine listening, "sound as event", happening in the *hic et nunc* of the present, is continuously challenged by the numerous spatialities and temporalities enacted by technical operations on, and representations of, the sound signal. While those temporalities have been partly analyzed by Wolfgang Ernst (2016; 2017) – with reference to the way in which digital sonic media produce a specific sense of time through the micro-delays in real-time signal processing – in this study we focus specifically on the issue of space. In so doing we aim to clarify how recognition algorithms deal with the site-specificity of sound, and how a specific socio-technical notion of space is enacted through machine listening operations.[2]

In order to prove our thesis, we focus on two case studies: *acoustic fingerprinting* and *audio watermarking*. From the analysis of the case studies, we propose some techno-philosophical interpretations of the idea of space and of the position of the listening subject within the framework of machine listening.

Theoretical framework and methodology

Phenomenological listening and machine listening

The theories of listening of the last century have often adopted a phenomenological approach to describe the specific mode of knowledge, implemented by the sense of hearing. These descriptions highlight the specifically intimate and immersive nature of listening which makes it an essentially experiential ability, in contrast to the sight as an organ of objective measurement and distance (Ong, 1982). For phenomenology, the vibratory phenomenon is not considered as such, but only in its manifestation to the ear. Barthes distinguishes the psychological and intentional act of listening from the simple physiological phenomenon of hearing (Barthes, 1982, p. 170), emphasizing that the former is the result of a properly human evolution of the auditory act.

In general, according to the phenomenological position, sound is the correlative of the intentionality of the psychic function. Listening does not describe, but produces the acoustic phenomenon: the sound object is nothing without listening, it is sound only insofar as it is actualized by the hearing. "The listener is entwined with the heard. His sense of the world and of himself is constituted in this bond" (Voegelin, 2010, p. 5). As a consequence, phenomenology privileges subjectivity as a *listening point*, assimilating the listening space with the space of the listening subject.[3]

Machine listening challenges this assumption, insofar as it is the materialization of nonhuman and desubjectivized listening. By measuring sound with objective parameters, it postulates the existence of the world without a subject to experience it. Machines do not fuse with the heard object nor experience it. Rather, they commodify it while keeping it at a distance.

It is not our intention to simply contrast phenomenological listening and machine listening regarding the presumptions, operations, and subjectivities involved. The methodological starting point of our analysis presumes that dichotomies between human listening and machine listening are not productive for a critical understanding of the phenomenon. We suggest that listening happens *between* humans and machines, in a space that is socio-technically produced in the intertwining of technical apparatuses with their material operations and cultural practices of listening.

This approach is implicit in the assumption of the essentially *relational* condition of sound (LaBelle, 2008; Di Scipio, 2013). The form of sound emerges from the system of relations in which sound itself is immerged: every surface, every obstacle, the shape of architecture and buildings, and also the presence and distribution of bodies, to some extent affect the form of sound. This assumption has two consequences: on the one hand, it radicalizes the bond between sound and space in a phenomenological sense; on the other hand, it also suggests that sound as a relational medium is something more than what is perceived by the human sense of hearing, since it entails an "assemblage" (DeLanda, 2006) of humans, spaces, artifacts, knowledge, and social practices. The tension between these two aspects (latent in phenomenological reasoning) is well expressed by LaBelle (2008, p. ix): "sound's relational condition can be traced through modes of spatiality", whereas "space is more than its apparent materiality".

This means that the bond between sound and space is not only physical, but also socio-technical and epistemological, in the sense that it involves cultural assumptions and knowledge which are enacted by listening postures and embedded in technical apparatuses. "The acoustical event is also a social one" (LaBelle, 2008, p. xi). Phenomenological listening is never pure, but always integrated with social and technical practices. The phenomenological definition of a subject who, while listening, experiences the world "without distance" (Voegelin, 2010, p. 5) is also an abstract notion. It postulates the existence of a "zero listener", without history or context. Technological artifacts and media on the other hand not only affect the listener, but enact "techniques of listening" (Sterne, 2003, p. 83), occurring between the subject, the cultural codes, and the technical operations of devices.

Media archeology

Starting from these considerations, in this paper we will base our study of machine listening on methods derived from *media archeology*. Media archeology is a field of study that assumes the non-anthropocentrism of technological phenomena and, as such, questions the position of phenomenological subjectivity in a socio-technical assemblage, populated by nonhuman agencies. Wolfgang Ernst's concept of "radical" media archeology is an attempt at studying media "from the point of view of technological artifacts" (Ernst, 2018, p. 37), thus focusing on their modes of functioning and on the epistemologies that are embedded in their technical operations. In relation to sound, Ernst's approach aims at considering the nonhuman agencies of sonic media and their specific modes of representation, highlighting the discontinuities with human ones. "If the communicational approach to sound focuses on listening as cultural interpretation, the media-archeological understanding assumes an interlaced option. It concentrates on neither the socio-historical, nor the bare psychoacoustic level but on the epistemological dimension that is embedded in sonic articulation" (Ernst, 2016, p. 45).

From the media archeological point of view, a precise concept of space can be retrieved in machine operations such as signal processing. This concept might be very different from the classic anthropological one, but no less effective. In the wake of McLuhan's reflections, media archeology aims to account for how media do not just reflect social meanings, but *produce* them through their very functioning: "the media-archeological hypothesis is that the human auditory apparatus is induced to obey laws imposed by the media device; historicity is therefore suspended by technology" (Ernst, 2016, p. 89). In so doing, media archeology displaces the human subject from its traditional role as the center of historical and technological change, seeking instead to unearth the "nondiscursive infrastructure and (hidden) programs of media" (Ernst, 2013, p. 59) which structure what and how humans think and do.

Media archeology considers technical media, such as gramophones, microphones, oscilloscopes as well as recording and measuring devices, to be nonhuman agencies of listening. As such, they "provide mode of listening prior to cognitive understanding" (Ernst, 2016, p. 31). The phonograph, for example, was the crystallization of new medical and physical knowledge about the auditory system and the resonance of cavities, acting as an acoustic "prosthesis" modeled on the functioning of the ear (Sterne, 2003). However, at the same time it also redefined listening by technically separating sound from its natural "source", preparing it for a new type of communication, that of telephony, in which face-to-face meetings give way to displaced and deferred presence (Peters, 2004, p. 184). The objective measurement of sound through instruments (spectrographs and analyzers) separates the perceptual experience from the vibratory phenomenon itself, thus inaugurating a new way of listening, halfway between what the senses perceive and what the machines ruthlessly measure. It is the listening paradigm of sound technicians and of composers of acousmatic music, who treat sound "as such", solely for its morphological-vibratory properties, modeled by a biotechnical circle of organs and recording/measuring devices.[4]

Computational and network media are inscribed in this framework and further modify the idea of listening space, since they replace deferred presence with the radical disembodiment intrinsic to algorithmic processes. Listening that takes place in the cloud, on data processing servers, gets rid of both the listening subject and of the physical space of sound diffusion, but it does not disregard them completely: as we will show later, it virtualizes them, turning both the hearing organ and the acoustic space into a model for algorithmic simulation addressed, usually, to the ears of the machine.

In the following sections, we use media archeology to reconstruct ideas of space in machine listening. The study is composed of an empirical analysis of materialities of algorithms and signal processing together with a theoretical reflection on the socio-cultural and epistemological aspects which those materialities involve. The empirical analysis draws on scientific papers and technical sheets related to the algorithms. Since not all components of the algorithmic system have published patents, a portion of this work depends on our own observation and reverse-engineering of the algorithms.

Case studies

Machine listening is a general term referring to a multitude of socio-technical layers. It brings together many specific technological phenomena that lead to very different applications. The question of space has always been a critical one for machine listening, regardless of the technical specifics of the algorithms being employed. Shazam made no secret of this: "The algorithm had to be able to recognize a short audio sample of music that had been broadcast, mixed with heavy ambient noise, subject to reverb and other processing" (Wang, 2003, p. 1). Similarly, important advances in speech recognition depended on the capacity to operate in noisy or reverberated environments (Pieraccini, 2012; Li et al., 2013).

Since these systems are complex assemblies of numerous algorithms interacting with each other, we will focus on two systems which clearly illustrate the issue of listening spaces: *acoustic fingerprinting* and *audio watermarking*. In a certain sense, these two copyright detection algorithms are complementary. Understanding their functioning will allow a closer examination of the problems of sound and space at play between the fields of acoustics, psychoacoustics, and technological sound processing.

Acoustic fingerprint

Although many recognition algorithms have appeared in the wake of Shazam, this one is exemplary in its ability to analyze and identify sounds which are strongly affected by noise and environmental features such as reverberation. Shazam (and software like it) analyzes parameters not discernible by the human ear, neatly bypassing the influence of external factors. It depends on criteria which can be perceived precisely by a microphone, even in poor recording conditions. This algorithm works by comparing sounds captured in the environment with marked audio files stored in a database (Wang, 2003).[5] The marking process is called *fingerprinting*. Fingerprinting is a function of the sonic information (sound translated into data via analog-to-digital converters), performed in a two-step process: first, a topological representation of sound data is generated, usually in form of spectrograms or other time/frequency/amplitude functions; second, the algorithm identifies specific points on that map, corresponding to numerical values, individuating a series of relations between these points. The result is something we can think of as a sort of "constellation map" (Wang, 2003, p. 2), specific to that audio clip.

The way audio search engines solve the problem of space and reverberation is by performing an operation which in signal theory is called *dimensionality reduction* (van der Maaten et al., 2009). Dimensionality reduction is the transformation of data from multi-dimensional representations, such as spectrograms, to a low-dimensional representation space, composed exclusively of vectors connecting *anchor points*. As soon as a sound is captured in the environment, the algorithm tries to identify an anchor point from perceptual characteristics, such as bandwidth, temporal and spectral features, average peaks, and prominent tones. Anchor points usually correspond to higher energy content.

In this process, the algorithm translates an aural perceptual problem into a geometry problem. An audio search engine identifies and selects the parameters least susceptible to the transformations caused by environmental factors, mapping those and discarding the rest. Through this operation, the algorithm can significantly reduce the complexity of the subsequent processes, preserving the key elements of a sound at the numerical level (Schalkwijk, 2018a; Walczyński & Ryba, 2019).

Fingerprinting allows robust and dynamic *pattern matching* in the field of sound, permitting subsequent processing steps to operate on sounds with the same methods that are normally used on static contents such as text. It is the necessary first step, after which common functions, such as search, comparison, and analysis, can be performed with relative ease. Fingerprinting can thus be seen as a way to render dynamic space-dependent and time-dependent contents (what we have defined as "sound events") into numerically approachable contents, without losing the situatedness of those events. In short, fingerprinting reduces sound recognition to a question of pattern-matching.

This process minimizes the influence of spatial features, because it reduces the complexity of a given sound event to a finite set of factors. In order to recognize a sample collected in the field and connect it to a copyrighted reference sample (which is the purpose of fingerprinting), some minimum segments are required. We can therefore say that the space through which this algorithm listens resembles a topographical stratification of three-dimensional spaces in the first step (spectrogram generation). Then, as soon as the constellation map has been generated, it is reduced to a *vector space*. The anchor points become the center of gravity of a space that is now purely relational, vectors that move in one direction until they connect to other anchor points. This vector space has nothing in common with the way humans experience the physical space. Vector space is composed exclusively of relations between points: beyond those points and the vectors which describe the relations among them, nothing exists.

Audio watermarking

Audio watermarking is the process of embedding encrypted information into an audio file without compromising the sound content from a human-listening perspective. While fingerprinting works on sound events without transforming them, audio watermarking introduces hidden information into the sound domain. It is not only used to detect information related to intellectual property protection, but also to embed metadata like closed captions or subtitles in multimedia contents (van Tilborg & Jajodia, 2011). The spectrum of a sound becomes a panel within which information can be written. This technique operates on the level of discrete information transmission from machine to machine, disregarding the propagation of sound in space.

There are many watermarking techniques, but they all consist of an encoding step and a decoding step. Both the encoding and decoding are based on a four-phase analysis of the audio file (Bengert & Upward, 2003): 1) the framing of the unprocessed audio file; 2) analysis of the frames with a fast Fourier transform algorithm (FFT) which converts the signal from its original domain (in this case time) to a representation in the frequency domain; 3) a compensation process called DC Carrier Removal which occurs when the offset at the center of the recorded waveform is not at 0 – which might be caused by resampling or lossy digital format distortion; 4) calibration of the correct amplitude of the watermark to keep it below the audibility threshold.

There are three main watermarking techniques (van Tilborg & Jajodia, 2011; Schalkwijk, 2018b). *Phase encoding watermarks* operate on the phase of the signal to hide data that have to be transmitted with and within the sonic information without being perceived by the listener. The phase of each frame of the unprocessed audio is modulated through an artificial phase to create the watermark. *Echo watermarks* perform a time-based distortion of the signal which is negligible to the human ear. One of the most efficient audio watermarking techniques is the *spread-spectrum watermark* (SSW). This technique involves printing a narrowband signal at various points within the full bandwidth of the original audio file. The pilot signal runs in an extremely subliminal way, so as to be imperceptible to the human ear. To make the system effective, the watermark is spread over many frequency bands so that no single band contains enough energy to be detected by the human ear. In comparison with the other techniques, SSW is particularly robust. To make it ineffective, it is necessary to add high amplitude noise to all frequency bands, and this would have a drastic effect on the content of the audio file (Davarynejad et al., 2010). The timing with which these inaudible watermarks are distributed is a mathematical random sequence called "pseudo noise sequence" (Schroeder, 1965). A key is required to decrypt the stationing of the watermark's pilots in the full audio spectrum.[6]

To summarize: fingerprinting cohabits the contingent space, the same space in which we listen. In order to isolate portions of the audio and to extract information from the actual space, it is necessary to “flatten” the sample, transforming it into a series of vector relationships. Watermarking, on the other hand, works in a completely enfolded space: it inhabits a space perceivable only by a machine ear. It completely disregards physical space since it occurs exclusively at the audio signal level. Still, it must include notions about physical space and human listening in its algorithms in order to optimize its functioning. Basically, it studies the human auditory model so as not to affect portions of sound which are audible by a human recipient. This is necessary in order to render its presence undetectable to the human ear, while still delivering its message to a machine ear. It requires a kind of compromise: the machine must take a step back from its purely nonhuman domain and resume some anthropomorphism.

The folded space

The case studies reported above illustrate two different spatial conditions in which machine listening operates. While fingerprinting analyzes sound in space in order to match patterns with samples in a database, watermarking operates at the level of pure audio data, in order to introduce non-intrusive metadata into the signal, solely addressing machine “ears”. In both cases, two different spatialities coexist in the sound phenomenon: a physical space and an informational space. The interaction between these two dimensions is a peculiar one: the physical space which was the condition for the sound event, is now considered by the algorithm as a “symptomatic space”, one which must be analyzed and accounted for its effects on the sound. In the case of fingerprinting, this reconstruction is aimed at removing space by separating it from sound, in order to optimize the recognition in all possible conditions. In the case of watermarking, the reconstruction of space is aimed at turning space from an obstacle to machine-to-machine communication into a medium for it. Thus, fingerprinting and watermarking treat space as a problem to be solved in order to optimize their functioning. Although they do it in different ways, they each attempt to bypass three-dimensional space, to make it “as if” it was not there. In this process, space is simulated in a sort of reverse-engineering of the sound phenomenon performed by the algorithm.

The problem of space for machine listening can be viewed as the need to include in automatic recognition the virtually infinite variability of the sound event. This means making machines able to deal with the relational condition of sound (LaBelle, 2008), its site-specificity according to which the same informational content will be sonically different at any “eduction” (Feaster, 2011) in the space. This epistemological distinction between supposedly enduring information and a contingent sonic manifestation affected by space, recalls the one between a dry sound, which is the original to be recognized, and its “detachable echo” (Sterne, 2015, p. 111). Here, space is the surrogate of a sound event, no longer a condition of its existence. However, machine listening cannot simply ignore space. In order to bypass space, it needs to reconstruct it as a technical variable in sound detection algorithms.

Machine listening should thus be considered as a socio-technical network that includes human and nonhuman agencies (Latour, 2005), multiple levels of perception and measurement, knowledge and interpretation, cultural and numerical operations, as well as devices for storage, processing, and communication.

Starting from these considerations, we propose the concept of *folded space* as specific to machine listening. The choice of the term "folded space" is inspired by the concept of "fold" as defined by Deleuze (1993): a movement of inflection and inclusion that conjugates unity and multiplicity without reducing one to the other. In our proposal, that movement has to do with the coexistence of different spatialities in techno-culturally mediated sound. If we are to consider machine listening as a socio-technical phenomenon, we must treat space not as the transcendental condition of experience, but as the product of the network of human and nonhuman agencies involved in technologically mediated sound. As a consequence, the notion of space must be informed by acoustic physics and specific processes and operations on sound by technological devices. In particular, machine listening is the effect of several technological layers (microphones, analog-digital converters, analyzers, dimensionality reduction algorithms, pattern matching algorithms etc.), each with its own techno-culturally coded operations, interacting with socially heterogeneous networks and infrastructures, with regimes of expectations and knowledge about sound and auditory perception. Multiple spatial dimensions are folded together within these layers. In order to get to the intertwining of this multiplicity of levels it is not enough to look at the "proliferation of sonic spaces" within a single space (Sterne, 2015, p. 115), but it's necessary to look at the reconstruction of that proliferation, both at the perceptive level and at the operational level.

This necessity becomes clear when comparing the theoretical idea of space, proper of phenomenological thinking, with the material operations of the algorithms in question. From its Kantian critical/transcendental declination on, the phenomenological approach considers space as a category of subjectivity. As such, it is epistemologically time-dependent: it is the unfolding of time that determines space. In this sense, the space of a sound event can be understood as a function of sound's modulation in time, that is, of the periodical oscillation of air (frequency). However, when we move to the level of signal processing, we face a *spatialization* of time, such as fingerprinting's "constellation maps": frequency as a mathematical function can be computationally analyzed without having to rely on its performative unfolding over time. All the sounds are co-existent and simultaneous when translated into data. When copyright ID algorithms "listen" to the sound of our videos, they do it in one take, in a coextension of time, a simultaneity of all sounds spatialized as coextensive information. What, then, becomes of the sound's space when time is compressed into a simultaneity?

Wolfgang Ernst has coined the term "sonicity" to analyze the temporal dimension of technologically mediated sound. "*Sonicity* as a neologism is meant to be kept apart from acoustic *sound* and primarily refers to inaudible events in the vibrational (analog) and rhythmic (digital) fields. [...] Sonicity names oscillatory events and their mathematically reverse equivalent: the frequency domain as an epistemological object" (Ernst, 2016, p. 22). Implicit sonicity is the sound considered in its technological representations, that is, spectrographic waveforms, magnetic distribution on tape, constellation map of digital data etc. In algorithmic signal processing, the time-space dimensionality described by sound is turned into a mathematical function containing all the virtual events. Computation operates at a different speed than sound as phenomenologically perceived in space and time. "Sonic analytics does not happen in real-time; there is rather a numerical time lens" (Ernst, 2016, p. 131). While sound expresses the temporality of physical acoustic vibration, "sonicity" as implicit sound in its technological form, expresses a plurality of temporalities: the temporality of computational data processing, that of digital audio (the sample rate and bit depth), and possibly that of vibration as a mathematical function. Those temporalities are folded together in the technological experience of sound. In contrast, the human experience of sound is mediated by "temporal tides", and all human listeners share this common condition. Sonicity enacts a different time, dependent on the speed of data processing.

If we bring this same comparison to bear on the issue of space, we can see how the investigation of sonicity provides access to a plurality of spatial dimensions. When analyzing machine listening as a socio-technical network, we find different kinds of spaces folded together: the three-dimensional *physical space* of the ambiences where sound happens as an event, the one-dimensional physical space of electric circuits where sound is converted into signal, the *symptomatic space* of the recording and the *epistemological space* of mathematics and algorithmic data processing. In a bottom-up view, the physical space is folded in the recording as a symptomatic space, which is at the same time a sonic effect for human ears and a bundle of features for machine measurement. At the next level, algorithms fold down that simulated space in models that consider both acoustics and psycho-acoustics to optimize their functions. This is possible because the one-dimensional space of sound as electric signal or data entails an epistemological space, a mathematical and speculative space, a vector space as relational space without the need for physical extension in itself.

In machine listening as socio-technical network, the folded space is a multiplicity in which different material and epistemological dimensionalities coexist. Materialities of signal processing produce a new space which is intertwined with the physical space. The transcendental space of the subject is now objectified in technical operations, exploiting the models of human hearing apparatus to transfer listening to a nonhuman dimension.

From the viewpoint of media-archeology, the core question is how computational systems treat space as *acoustic effect,* while at the same time producing other spaces for sound (or better sonicity) in their material operations. Machine-to-machine communications are not audible, but still related to the audio-sphere: as they move from sound to sonicity, from explicit to implicit, they are constantly expressing precise knowledge about sound, space and listening, knowledge embedded in their algorithmic processes. These technologies are not mere surrogates or auxiliary supports for something whose real destination is diffusion in time and space. Rather, the very concepts of time and space are technologically determined by the operations of inscription and transmission devices.

"Sound is *always* in more than one place" (LaBelle, 2008, p. x), but our critical analysis suggests the opposite: that more spaces can be in one sound when it comes to the materialities of its production and processing.

The listening space

What makes fingerprinting and watermarking representative of the shift introduced by machine listening is that they do not just measure sound – recording devices have long had this capability – but that they "recognize" events[7]. They serve as a bridge between the vibrational phenomenon of sound, the computational domain of signal processing, and the cultural level of *interpretation*. This level is now negotiated between hermeneutics of sound and the numerical process of pattern-matching, and at stake in this negotiation is a possible new meaning of the term "recognition".

In sound recording, the listening space no longer coincides with the listening subject. Microphones can hear in contexts in which the subject is absent, or in spaces where his presence is physically impossible, but in algorithmic machine listening, we see a further shift: machines do not just extend humans' listening ability, but in their operations assume the very function of subjectivity. They do not just measure sound, but *recognize* it. The point is not that machine listening "augments" human listening, but that, even in the scope of machine-to-machine communication, it cannot ignore the phenomenology of listening. Space, in particular, is the critical issue that forces the inclusion of models of human listening in recognition algorithms. The subject does not disappear in this process, but is still present as a model embedded in the algorithm. This reveals a shift in the position of the listening subject in the framework of machine listening. Human listening ability employed as a model for machine listening turns humans into a *medium* for machine-to-machine communications, providing machines with the elements to work autonomously and to self-optimize, to automatically recognize sounds and to react appropriately.

In this framework the "listening space" no longer coincides with the listening position of the ear, with the ear itself being substituted by a digital reconstruction within the algorithm. The folded space of machine listening includes both the anthropocentric listening position and non-anthropocentric numerical processing. Although algorithms can work by themselves and recognize sounds without relying on human interpretation abilities, they cannot disregard the specificities of phenomenological listening in order to encompass the singularity of the sound event, that is, the happening of sound in the space-time.

As philosophers of speculative realism, such as Meillasoux (2006) and Morton (2013), have argued, the question of machinic measurement of nature raises the possibility of a non-phenomenological and non-anthropocentric reality which affects the transcendental notions of time and space. When time and space are no longer transcendental categories of subjectivity, but emergent properties of the objects, those objects become "hyperobjects" (Morton, 2013, p. 63). Sound as a hyperobject is not only the vibrational phenomenon in a given time and space, nor is it merely the sound for the hearing subject. Rather, it is the multiplicity of ontologies disclosed by its treatment by humans and devices: sound in physical space, sound as data, sound as mathematical function, each with its own corresponding temporal and spatial dimensions.

While the metaphors and social representations adopted to describe machine listening are often anthropomorphic – and the very term "listening", when referring to numerical operations, can be seen as a metaphor in itself – the "media message" (Ernst, 2018, p. 37) embedded in technical operations reveals that anthropomorphism is now inscribed into the algorithms and, thus, contaminated (hybridized). Nevertheless, in the machine's effort to recognize what is meaningful according to *human* sound perception – that is, in translating numerical processes (only concerned with audio data) into semantic categories – is evident a form of "return anthropocentrism" that softens the crude nonhuman agency of data processing. What Mackenzie (2007, p. 89) says about the traces of embodiment in algorithmic time can be applied to machine listening as well: the challenge is "finding middle ground between the temporality of technologies as material orderings of movement and the temporal flows of subjective experience". It is a kind of "living-nonliving synthesis" (Mackenzie, 2007, p. 91).

In light of this analysis, the distinction between human listening and machine listening becomes blurred, and the co-determination of the two emerges. This co-determination can be understood as the new space of the listener, produced by socio-technical networks. If many of the fears related to machine listening derive from the misleading way in which it is often represented, that is, as a substitute for human listening, it is equally true that there is no "human listening" as such, as it is always complemented by artificial organs, media, or prostheses which redefine and reconfigure it according to techno-epistemic regimes.

References

Barthes R. (1982). *Écoute*. In R. Barthes, *L'obvie et l'obtus. Essais critique III.* Paris: Editions du Seuil.

Bengert, J., & Upward, A. (2003). Elec 499A. *Perceptual Audio Project*. http://ece.uvic.ca/~elec499/2003a/group09/index.htm.

Biancorosso, G. (2016). *Situated Listening: The Sound of Absorption in Classical Cinema*. Oxford: Oxford Scholarship Online.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer Internal Publishing.

Cavanaugh, W.J., & Wilkes, J.A. (1999). *Architectural Acoustics: Principles and Practice.* Hoboken: John Wiley & Sons.

Davarynejad, M., Ahn, C.W., Vrancken, J., van den Berg, J., & Coello Coello, C.A. (2010). Evolutionary hidden information detection by granulation-based fitness approximation. *Applied Soft Computing*, Vol. 10, Issue 3, pp. 719-729. https://doi.org/10.1016/j.asoc.2009.09.001.

DeLanda, M. (2006). *A new philosophy of society: assemblage theory and social complexity*. London: Continuum.

Deleuze, G. (1993). *The Fold: Leibniz and the Baroque*. Trans. by T. Conley. London: The Athlone Press.

Desai, N., & Tahilramani, N. (2016). Digital Speech Watermarking for Authenticity of Speaker in Speaker Recognition System. *2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE).* https://doi.org/10.1109/ICMETE.2016.13.

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

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

Ernst, W. (2017). *The Delayed Present: Media-Induced Tempor(e)alities & Techno-Traumatic Irritations of "the Contemporary"*. Aarhus: Sternberg Press.

Ernst, W. (2018). Radical Media Archaeology: Its Epistemology, Aesthetics and Case Studies. *Artnodes*, 21, pp. 35-43. https://doi.org/10.7238/a.v0i21.3205.

Feaster, P. (2011). A compass of extraordinary range: the forgotten origins of phonomanipulation. *ARSC Journal* XLII/ ii.

Labelle, B. (2008). *Background Noise: Perspectives on Sound Art*. New York: Continuum.

Latour, B. (2005). *Reassembling the social. An introduction to actor-network theory.* London: Oxford University Press.

Li, J., Deng, L., Gong Y., & Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 22, No. 4, pp. 745-777. https://doi.org/10.1109/TASLP.2014.2304637.

Mackenzie, A. (2007). Protocols and the irreducible traces of embodiment: The Viterbi algorithm and the mosaic of machine time. In: R. Hassan & R.E. Purser (Eds.), *24/7: Time and temporality in the network society* (pp. 89–106). Stanford, CA: Stanford University Press.

Meillasoux, Q. (2006). *Après la finitude. Essai sur la nécessité de la contingence*. Paris: Editions de Seuil.

Morton, T. (2013). *Hyperobjects*. Minneapolis: Unversity of Minnesota Press.

Ong, W.J. (1982). *Orality and literacy. The technologizing of the word*. London: Methuen & Co.

Peters, J.D. (2004). Helmholtz, Edison and Sound History. In: L. Rabinovitz, & A. Geil (Eds.), *Memory Bytes. History, Technology and Digital Culture.* Durham: Duke University Press. https://doi.org/10.1215/9780822385691-008.

Pieraccini, R. (2012). *The voice in the machine. Building computers that understand speech.* Cambridge: MIT Press. https://doi.org/10.7551/mitpress/9072.001.0001.

Schaeffer, P. (2017). *Treatise on Musical Objects* [1966]. Oakland: University of California Press.

Schalkwijk, J. (2018a). A fingerprint for audio. https://medium.com/intrasonics/a-fingerprint-for-audio-3b337551a671.

Schalkwijk, J. (2018b). Hiding data in sound. https://medium.com/intrasonics/hiding-data-in-sound-c8db3de5d6e0.

Schroeder, M. (1965). New method of measuring reverberation time. *J. Acoust. Soc. Am.*, 37, pp. 409-412. https://doi.org/10.1121/1.1939454.

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

Sterne, J. (2015). Space within Space: Artificial Reverb and the Detachable Echo. *Grey Room* 60: 110-131. https://doi.org/10.1162/grey_a_00177.

van der Maaten L., Postma E., & van den Herik, J. (2009). Dimensionality Reduction: A Comparative Review. *TiCC*, Tilburg University, The Netherlands.

van Tilborg, H.C.A., & Jajodia, S. (Eds.). (2011). *Encyclopedia of Cryptography and Security*. New York: Springer US. https://doi.org/10.1007/978-1-4419-5906-5.

Voegelin, S. (2010). *Listening to Noise and Silence. Towards a Philosophy of Sound Art.* New York: Continuum.

Walczyński, M., & Ryba, D. (2019). Effectiveness of the acoustic fingerprint in various acoustical environments. *IEEE Signal Processing 2019: Algorithms, Architectures, Arrangements, and Applications (SPA)*. https://doi.org/10.23919/spa.2019.8936781.

Notes

[1] From the virtual assistants, such as Amazon Echo, to acoustic event detectors undergoing beta testing in smart cities. See for example the system EAR-IT in Smart Santander: https://ec.europa.eu/digital-single-market/en/news/ear-it-using-sound-picture-world-new-way.

[2] An analogous discourse about the coexisting temporalities enacted by algorithmic sound processing could be addressed in future research.

[3] Sterne associates this theoretical position with specific socio-economic practices of listening, which he calls "audile techniques": "As a bourgeois form of listening, audile technique was rooted in a practice of individuation [...] The space of the auditory field became a form of private property, a space for the individual to inhabit alone" (Sterne, 2003, p. 160).

[4] If "reduced listening" has been defined as such in reference to the phenomenological reduction (Schaeffer, 2017, p. 213), it cannot be underestimated that this reduction was made possible precisely by non-phenomenological sound measurement technologies.

[5] See also Shazam's patent, retrievable at: http://patentimages.storage.googleapis.com/pdfs/US8442426.pdf.

[6] Spreading spectrum is assigned through pseudo noise sequences (PN). PN are the sequences that obey Golomb's three postulates of randomness. Maximum lenght sequence (MLS) is one among these and widely used for the measurement of impulsive responses (reverberation responses). It is interesting (and somewhat ironic) that this technique is widely used for measurements in the acoustic treatment of spaces (Cavanaugh & Wilkes, 1999, p. 61).

[7] Ground-breaking machine learning systems are continuing to expand and exploit this ability. In these systems though, recognition is achieved by pattern matching in conjunction with statistical training based on sound examples. After careful analysis of training examples, the

Wang, A. (2003). An Industrial Strength Audio Search Algorithm. *Proceedings of ISMIR 2003, 4th International Conference on Music Information Retrieval*, Baltimore, Maryland, USA, October 27-30. https://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf.

algorithm is able to generalize, that is, to recognize certain features in new unfamiliar cases not present in the observation pool (Bishop, 2006). In the field of acoustic detection, machine learning is gaining prevalence in conjunction with fingerprinting (Schalkwijk, 2018a), while in the field of copyright detection it is often associated with watermarking, especially to resolve new problems raised by synthetic media and deepfakes (Desai & Tahilramani, 2016; Schalkwijk, 2018b).