Issue coedited by: Universitat Oberta de Catalunya and Leonardo/The International Society for the Arts, Sciences and Technology

artnodes
E-JOURNAL ON ART, SCIENCE AND TECHNOLOGY

https://artnodes.uoc.edu

ARTICLE

NODE «AI, ARTS & DESIGN: QUESTIONING LEARNING MACHINES»

Unlocking the Black Box of AI Listening Machines: Assemblages for Art, Technology and Innovation

Sharath Chandra Ramakrishnan
School of Arts, Technology, and Emerging Communication
University of Texas at Dallas

Date of submission: March 2020
Accepted in: June 2020
Published in: July 2020

Recommended citation

Ramakrishnan, Sharath Chandra. 2020. «Unlocking the Black Box of AI Listening Machines: Assemblages for Art, Technology and Innovation». In: Andrés Burbano; Ruth West (coord.) «AI, Arts & Design: Questioning Learning Machines». *Artnodes*, no.26: 1-9. UOC. [Accessed: dd/mm/yy]. http://doi.org/10.7238/a.v0i26.3362

Abstract

The black box of innovation in the realm of connected AI technologies renders not only their technicalities opaque but also, and more importantly, the social effects and relations that constitute their creation and mediation. This presents an opportunity for creative interventions by artists and researchers, to unveil the networked relations that are part of AI technologies, and speculate on their ontological effects. This article presents such an unpacking around an AI listening machine present today in ubiquitous devices like voice assistants and smart speakers, and incorporates computational models of machine audition. By tracing the scientific research, technical expertise, and social relations that led to our cultural adoption of AI listening machines, the article presents a socio-technical assemblage within which these machines operate. At the same time, the article reveals various contexts for artists as well as innovation researchers to engage with the socio-technical complexity of AI listening machines, by sharing some instances of creative and artistic interventions that have attempted to unveil the nature of their assemblages.

Keywords
intelligent agents, computational art and design, machine listening, socio-technical systems, auditory scene analysis

Apertura de la caja negra en sistemas de escucha de IA: Assemblages para arte, tecnología e innovación

Resumen
La caja negra de la innovación en el ámbito de las tecnologías de IA conectadas hace que no solo sus tecnicismos sean opacos sino también, y lo que es más importante, los efectos en la sociedad y las relaciones que constituyen su creación y mediación. Esto presenta una oportunidad para intervenciones creativas de artistas e investigadores, para desvelar las relaciones en red que forman parte de las tecnologías de IA y especular sobre sus efectos ontológicos. Este artículo quiere destapar todo lo relacionado con los sistemas de escucha de IA actualmente presentes en dispositivos ubicuos como asistentes de voz y altavoces inteligentes, e incorpora modelos computacionales de audición de máquinas. Al rastrear la investigación científica, la experiencia técnica y las relaciones sociales que han llevado a nuestra adopción cultural de los sistemas de escucha de IA, el artículo presenta un conjunto sociotécnico dentro del cual operan estas máquinas. Paralelamente, el artículo revela varios contextos para que artistas e investigadores en innovación se involucren con la complejidad sociotécnica de los sistemas de escucha de IA, al compartir algunas instancias de intervenciones creativas y artísticas que han intentado desvelar la naturaleza de sus assemblages.

Palabras clave
agentes inteligentes, arte y diseño computacional, sistemas de escucha, sistemas sociotécnicos, análisis de escenas auditivas

Introduction

Over centuries of our attempts at understanding the mechanism of human audition, we have today modelled listening machines that can mimic the ear and perform a number of useful functions such as conversing to answer our questions, recognising a piece of music and even recreating the sonic world for people with hearing problems. New AI listening technologies used today may be viewed as an extrapolation of an ongoing cultural training over centuries. This cultural training has involved teaching humans new habits of listening with machines, as well as techniques to interpret and operate them. Listening machines can learn to perform intelligent analysis of sound and speech based on computational models of human audition, endowed with capacities to not just simulate but also exceed and augment the limits of human audition, thereby problematising our contemporary notions of aurality. Indeed, the act of listening is no longer restricted to humans, and if machines can listen to what humans cannot perceive, how does this affect and shape social and cultural relations associated with AI augmented listening? These are but some of the questions that the author wishes to engage the reader with, while delving into its object of study - the 'AI listening machine', that endows the ubiquitous smart speaker/voice assistant with the ability to tune in, listen to and make 'sense' of the audile world of humans.

Contributions

A Call to Unveil Socio-Technical Relations Behind AI Systems
Before diving into our discussion of the AI listening machine, locating the object of study in the historic cross currents of scientific discovery and technological innovation would serve to precisely lay out the intent and contributions of this article.

Since the early 15th century, much of the Western world was preoccupied with an ontological question regarding the mechanistic interpretation of organic and biophysiological processes. Intellectuals questioned if consciousness or the nature of life itself can ever be mechanically represented, and as a result, several experimental efforts to create automatons ensued. These ranged from Vaucanson's lifelike mechanical Duck (Riskin 2003) and his Flute Player (incidentally documented and described in the first French encyclopedia by Diderot as an "android" (androide) (Rice 2014), to the mechanisation of rational thought in computing automatons like the Pascaline by Blaise Pascal in 1645 (Price, Bedini, et al. 1964). While the history of the automaton and its roots in mechanistic philosophy is not within the scope of this paper, what is pertinent is the academic ramification of the above-mentioned ontological paradox that has created a gap in our understanding of the relationships between science and technology (Pinch and Bijker 1984). Understanding this relationship is particularly complicated for emerging

technologies like AI listening machines that subsume hypothetical findings in the cognitive and auditory neurosciences within the agenda of emerging human enabling listening technologies.

While sociologists of science and technology studies have contributed a multitude of disciplinary perspectives on the relationship between science and technology, an overall thematic approach has been to separate technology from science on analytical grounds, by attributing discovery of truth to science and delegating the application of this truth to the role of technology (Pinch and Bijker 1984). Despite studies by innovation researchers involving empirical inquiries to map both the extent to which technological innovations originated from scientific discoveries, as well as the dependence of scientific research on the availability of a specialised technology, the clear view on their interdependence remains difficult to specify. Mayr attributes this failure to the perpetuated assumption that science and technology are disparately defined structures (Pinch and Bijker 1984; Mayr 1976). He posits that in order to advance a deeper understanding, it is essential to realise that "science and technology are themselves socially produced in a variety of social circumstances" (Pinch and Bijker 1984). This view regards the relationship between scientists and technologists to exist within a socially constructed and mediated culture.

Research has further indicated that most innovation studies dealing with a simple linear model, which proceed from research to product development using R&D metrics and macroeconomic success indicators of technological innovation, refrain from discussing societal factors or the technology itself (Layton 1977). Layton suggests that, "what is needed is an understanding of technology from inside both as a body of knowledge and as a social system. Instead, technology is often treated as a 'black box', whose contents and behaviour may be assumed to be common knowledge" (Layton 1977).

This socio-technical opacity is further exacerbated by the deployment of AI systems, revealing "a blind spot in AI research". This was observed in a recent research report that concerned the installation of autonomous AI agents in critical social systems like hospitals and courtrooms, with neither technical knowledge of the 'AI black box' nor agreed methods to assess the sustained effects of such technologies on society (Crawford and Calo 2016).

This presents a call for creative technologists, computational media artists and researchers working in the realm of AI technologies, to help bridge disciplinary silos, by unveiling these various socio-technical arrangements at play in the relational exchange between the science and technology of AI-based innovation. This article deals with the personal AI listening machine as a technological object of study, towards realising the above-mentioned goal.

Methodology

An examination of the various socio-technical and scientific developments that led to the conception of the AI listening machine, and conversely, how technologies of the AI listening machine assimilated dominant socio-technical and cultural undercurrents, are central to the concerns of this study. To aid in the unpacking of these relations, the article adopts methods from techno-cultural analysis used in sound studies.

The author would like to clarify the basis upon which such a methodological analysis is staged upon, for the benefit of artists and creative technologists dealing with research methods. Technologies of listening are cultural artefacts of a metamorphosis in our understanding of sound as a medium, knowledge about human audition and practices of listening that evolved over centuries (Sterne 2003). Therefore, in order to unpack socio-technical relations, one has to take into account various cultural subjectivities that are reinforced in our contemporary engagement with AI-based listening machines. Within this context, cultural subjectivity refers to a society's "characteristic way of perceiving its social environment" (Triandis 1972), and consists of ideas and practices that have worked in the past and continue to be preserved and transmitted to the future (Triandis 1972). It follows that a study on the innovation of technologies, which include the technologies of listening, necessitates an understanding of their connections with human practices and habits, along with a focus on areas of intertwined cultural, social, and material histories. These interconnections from which technologies emerge and exist in, have often been referred to as 'networks' or 'assemblages' (Latour 2012; Wise 1997). These socio-technical assemblages are a necessary apparatus for the creation of AI listening machines, and include technological and scientific explorations, as well as institutions and individuals from diverse disciplinary backgrounds contributing to a social and cultural construction of the mechanism of AI-based aurality.

Unpacking the Socio-Technical Assemblage of the AI Listening Machine

In subsequent sections, we proceed to unveil the socio-technical assemblages within which the AI listening machine operates. Examples of media art interventions are provided to serve as a catalyst for readers wishing to further engage in creative expositions of these assemblages. The article unpacks this relational assemblage across the technicalities of AI listening machines, corporeal relations between the science and technologies of listening and policy negotiations mediated by the AI listening machine in its acoustic space.

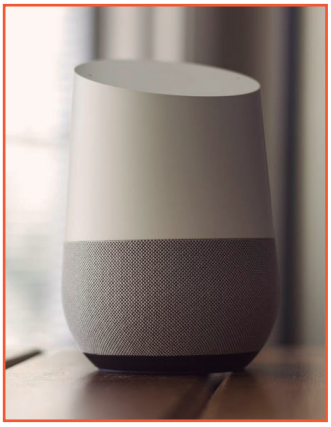

Figure 1. The ubiquitous smart speaker/voice assistant. Attribution: NDB Photos / CC BY-SA (https://creativecommons.org/licenses/by-sa/2.0).

Technicalities of the AI Listening Machine

The now ubiquitous 'smart speaker'/voice assistant occupies a unique spot in the convergence of innovation in AI techniques spanning the areas of machine listening, semantic linguistic parsing and speech synthesis technologies. While each of these topics traces a unique techno-social history in itself, the scope of this paper is restricted to the AI listening machine and its associated machine listening techniques for auditory scene analysis, including the contribution that machine listening techniques have had in the evolution of speech synthesis technologies as a historical intersection.

Assemblage of Auditory Scene Analysis

Much of the challenge in understanding audition in humans, as well as efforts in computational auditory modelling, is centered on trying to solve the "cocktail party problem" (Cherry 1953). Indeed, in a noisy house party, if one tries hard to focus one's auditory attention, it is possible to attend to any single conversation or sonic event occurring amongst a myriad of other sounds. This is precisely what auditory scene analysis tries to achieve - to be able to accurately segregate, segment and delineate individual sources from a complex mixture of sound, to later call upon to audition any of these sources of sound at will (Bregman 2001). The human brain seems to have encoded such processes for auditory scene analysis, the precise workings of which are still an area of active scientific research. Auditory source segregation, localisation and enhancement techniques to create machines that can listen and generate sensation, has been an extremely active area of research for scientists and engineers developing cochlear implants and hearing aids for the specially-abled.

The technical assemblage of auditory scene analysis models used in the AI listening machine, in its first stage, acoustically simulates cochlear filtering, followed by extracting information regarding pitch/frequency transitions, on-sets and offsets as an auditory representation, referred to as the auditory map (Brown and Cooke 1994). The next stage involves constructing a symbolic description of elements in the auditory scene using information in the auditory map. The final stage, analogous to the sensory grouping mechanism, involves a search mechanism that uses various strategies such as combining elements with common fundamental pitch, onset and offsets, to group them to be part of a single auditory stream. Notable methods to consider in the evolution of source separation techniques used in listening machines include a Markov model-based approach for segregating a mixture containing two speakers based on pitch derived from inter-peak intervals of a cochlear filterbank (Weintraub 1986), other methods that use pitch as a segregation criterion (Parsons 1976; Stubbs and Summerfield 1990), as well as additional information such as spatial location of the voice (Denbigh and Zhao 1992) to aid in source segregation. While these methods do rely on some prior knowledge about the acoustic environment, like number of speakers or types of sound sources, other methods (Mellinger 1991) do not make such strong assumptions about the environment and are much more suited to the operation of the personal AI listening machine. New AI techniques like TasNet, which can "directly model the signal in the time-domain using an encoder-decoder framework" (Luo and Mesgarani 2018), have greatly improved the prospects for performance of speech separation algorithms in low power and portable devices like smart speakers.

These auditory scene analysis models have been proposed and developed by auditory scientists and machine listening engineers working in collaborative settings. This tells us about the social context within which innovation of the AI listening machine happens. An ethnographic study involving researchers in the field of medical physics working on a model to transcribe speech in noisy environments to text, found scientists and engineers relating within a tension and contingency between different rationales (Voskuhl 2004). Colleagues involved in the pathology of hearing seemed suspicious of the intent of duplica-ting the ear for use in artificial speech transcribers. At the same time, they felt entitled by the common belief within their research circles that engineering techniques were unsuccessful in noisy environments, and the use of knowledge in medical acoustics in signal processing models was considered innovative, promising and niche. Despite their conflicting institutional rationales, a symbiotic relationship was guided, on one hand, by the objective of designing a listening technology that mimicked human audition, and on the other, the hope of obtaining a plausible model to use as an epistemological tool to decipher the biological science of hearing (Voskuhl 2004). Performance evaluation of the machine learning models used to perform source segregation involved a large training set of human speech recorded in noisy backgrounds, as well as an evaluation data set that served to measure the benchmark of performance.

The creation and use of such training and evaluation data sets involving human vocal contributions suggest another component of social interaction embedded in this layer of the assemblage. The limits of speech separation, recognition and parsing abilities of the AI listening machine are constantly tested in simulated real-world environments, by different spectral representations that use noises to mask and render the speech signal unintelligible to the machine. In this process, the individual voice is abstracted into a composite signal representation devoid of any personal attributes. This is similar to the consequence of abstracting hearing as a tympanic mechanism in early listening machines like the ear phonautograph or the pattern playback machine, that rendered the source of sound as irrelevant in scientific investigations. John Peters has highlighted this aspect of source-agnostic research in the work of Hermann von Helmholtz; "Helmholtz levels all modalities and is indifferent to bodily origins. What matters is the waveform and not the source" (Peters 2004).

The artistic interpretation of this aspect was brought in the work of composer and sound artist, Alvin Lucier. *I Am Sitting in a Room* (1970) was a composition that featured spoken text, which

the visitor was instructed to record and play back over itself repeatedly until the resonant frequencies of the room annihilated any form of intelligible speech (Cox 2009). Lucier urges the audience not to interpret this piece as a spatial exploration of sound and rather explicitly clarifies that through the technology of listening and playback, "the piece concerns the dissolution of speech and the speaker into sound and space. What begins as a personal confession in a domestic setting gradually becomes pure, anonymous sound that overwhelms and eradicates the performer's personality. Meaning and sense have dissolved into rhythm" (Cox 2009). By modifying, tuning and stretching the optimised parameters involved in the technical realisation of auditory source separation to the point of technical failure, performance of the anomalous and the anonymous become possible. Computational artists working in the field of audition could perhaps think of interesting ways to unveil the complex interplay at work between the rationale of technological efficiency and the quest of biological understanding, by manipulating underlying mechanisms of the AI listening machine.

The Overlap Between Machine Listening and Speech Synthesis Techniques

The history of machines invented to imitate the production of human sound as speech is just as important as the mechanisation of hearing, as both these attempts converged in their interpretation of speech through the spectrogram. This section clarifies the extent to which machine listening techniques used in auditory scene analysis are also used for machine speech synthesis in AI agents like voice assistants. There is a large body of research and art dealing with speaking machines, dating back to Von Kempelen's automaton, a mechanically operated speaking machine that could produce a remarkably human-like voice (Dudley and Tarnoczy 1950). Speech production took a visible shape with the invention of the spectrogram, that when combined with the anatomical understanding of formant production documented over the century, resulted in the invention of speech synthesising machines like the Voder and more importantly, the Vocoder, by Homer Dudley in Bell Labs in 1939 (Dudley and Tarnoczy 1950). The spectrum analysis component of the Vocoder to extract features, minimally represent and synthesise human sounding speech is a common processing feature shared with automatic speech recognition systems performing auditory scene analysis (Pickett 1968). For instance, commercial vocoders featured an array of components like microphones for capturing vocal input, a spectrum analyser to filter, sort and process sonic frequencies, a pitch detector and follower that locked onto the fundamental frequency of the auditory input, and a voiced/unvoiced detector that distinguished acoustic features produced by the vibrating vocal cords from those produced by modulations by the mouth, lips, and tongue. A process of encoding to translate this information as a binary signal, that could be later decoded back to speech pertained to the synthesis part (Pickett 1968).

While many artists have explored the synthesis aspects of the vocoder and variants of its linear predictive encoder in a performative and musical context, it was Lucier who was interested in the Vocoder as a tool for vocal and auditory analysis, rather than synthesis. In his piece *North American Time Capsule*, participants equipped with musical instruments and electrical appliances sang, spoke and read into the vocoder while their auditory input was modified to convert audible speech to an abstract sound and texture (Cox 2009). Lucier transformed the listening characteristics of the vocoder, "to liquidate speech and to abolish the identity of the speaking subject, shattering all syntax and pulverizing every symanteme, morpheme, and phoneme into fluid sonic matter" (Cox 2009). The advent of neural networks has allowed for the creation of vocoders that can synthesise high-quality, human-sounding speech, by conditioning on spectral features using speech analysis modules as well as human coded feature specification, that are open to new modes of artistic exploration along the axis of speech and identity. State-of-the-art implementations of the day include RawNet, an end-to-end neural vocoder, where the speech coder and voder act in tandem as an auto-encoder network, capable of being jointly trained directly on the raw audio waveform without the need for human feature design (Luo and Mesgarani 2018). The future of the smart speaker, or personal AI listening agent, is geared towards the capacity to listen and learn from higher level representations of speech input, to constantly improve its voice synthesis capability, thereby sharing a reciprocal relation with the machine listening aspect of its technology as well as its vocal user.

Corporeality of the Science and Technologies of Listening

The engagement of early listening machines as well as the modern-day AI listening machine with the institution of those with hearing problems, adds a corporeal dimension to the socio-technical assemblage discussed in this article. As listening machines transformed the function of the ear through their mechanisation, it was their socialisation and institutionalisation that led to their prevalence as audile media. The significance of the phonautograph in creating the first visual spectrogram of sound was part of a larger assemblage of institutional practices. The invention of the phonautograph was necessitated by the simultaneous establishment of otology (or ear medicine) as a formal branch of medical science. This made the human body and human ear available for the purposes of dissection and investigation. Sterne notes that "the human ear affixed to the chassis of the ear phonautograph" suggests how this hybrid nature of the phonautograph as a listening device could be considered "an artifact of otology's institutionalization", thereby revealing the corporeal nature of its existence (Sterne 2003). In this context, the concept of corporeality refers to the "diversity of issues, questions, concepts and relationships deriving from and centering on the body and bodily life"(Sheets-Johnstone 2015). Audition was transformed into a mechanism that could be abstracted from the human body

both anatomically and experientially. The ear became an object of study and its functioning became measurable, allowing physiologists to define it in mechanical terms. Bell's intention of using the phonautograph to represent visible speech was to train the deaf to speak, as it made speech visible as a waveform rather than depict articulations of speech as positional variations of the tongue and the mouth (Bell, Bolton, and Langdon 2017).

Our insights and theories related to the science of audition along with our knowledge to create computationally intelligent listening machines have been historically associated with "contextually situated human bodies to enact, experience and interpret" sound (Deleuze 1988). This fact is closely related to the words of philosopher Gilles Deleuze - "being social before being technical" (Deleuze 1988). Further, Kittler points out that innovations and progress in instrumentalised sensory measurement programmes managed by physiologists was crucial to the successful inventions and utilitarian proliferation of technological media (Kittler 2006). According to Kittler, the mediatisation of the senses did not occur solely through the invention of technical media but under the auspices of the psychophysical laboratory, where the experimental psychologist treated the human subject as a technological media product (Kittler 2006). In the context of innovations in listening machines and their relationship to the hearing impaired, Mara Mills points out that,

> "Across decades, national contexts, and technical shifts, however, deafness ultimately served as a 'pretext' to other engineering concerns - in some cases a precursor, in others a pretense. Inventors often abandoned collaborations with deaf students and their educators after initial trials, as their technologies transferred to more profitable realms. Certain inventors simply lifted ideas and inspirations from the world of deaf 'assistive' technologies."
> (Mara 2010)

The sound art community along with the deaf community has responded to this corporeality within the context of 'Deaf Gain' (Bauman and Murray 2014). The term 'Deaf Gain' inverted the concept of 'hearing loss', and positioned deaf culture as a valuable contributor to conversations in both art, science and culture. Relevant art projects include Tonotopia by sound artist and composer Tom Tlalim in collaboration with cochlear implant users, where composing and codesigning sound art along with and for persons with hearing disabilities was the focus of artistic engagement and creation (Tlalim 2017). In an attempt to enable artistic experience to be shared by diverse ears, the project (curated at the Victoria and Albert Museum) asked several pertinent questions, like what is deemed as normative or natural listening and who decides what these factors and parameters should be, and how are technologies of listening empowering or disavowing the deaf? These questions are germane to unveiling the interrelation between the auditory scene analysis industry, and auditory science for the hearing impaired, calling for more creative technologists and artists to respond to these relations.

Negotiations in the Networked Acoustic Private Space

Our interaction with the domestic AI listening machine occupies a networked acoustic space that is situated within a certain locus around it. The activation of this acoustic space is encoded by performed interaction routines (like a 'wake phrase' containing the name of the AI listening machine), thereby idealising the audile technique of interaction with the AI listening machine. Sterne describes how the idealisation of the audile technique led to the creation of private acoustic spaces that became "a precondition for the commodification of sound" (Sterne 2003). He reasons that "commodity exchange presupposes private property", necessitating "the acoustic space to be ownable before its contents could be bought or sold" (Sterne 2003). Thus, the AI listening machine could be seen as 'owning' the networked private acoustic space into which it has been invited. This raises several questions regarding socio-technical assemblages at work in the private acoustic space co-inhabited with the AI listening machine.

Previous innovations of instrumentalised listening spaces have often emerged within a surveillant framework, as Granchow notes about the preradar stone acoustic mirrors, that "could not have been undertaken without a corresponding shift in thinking about sound outside of the way sounds are perceived -more specifically, toward thinking about frequencies in terms of physical sizes" (Ganchrow 2009). The work of W. C. Sabine describes at length the "Ear of Dionysius", a panaural prison in Sicily dating back to the 4th Century BC (Sabine and Egan 1994). The S-shaped chamber of the prison featured a conical duct leading to a concealed chamber where all the acoustic reflection from within the prison was audible, an acoustic architectural version of Jeremy Bentham's Panopticon (Bentham 2012). These innovations, and others like the 3-horn listening structure created by Athanasius Kircher for the Royal Court to listen in to the town hall below, were based on a framework of acoustic materiality that explored and manipulated the physics of sound transmission, rather than its sensorial and cognitive aspects. Further, these listening machines were installed with the intent to surveil, and represented a nature of listening that is different from what is mediated by the object of study in this article.

In contrast, AI listening machines have not been explicitly developed under a surveillance agenda and therefore present new terrains for negotiating the networked private acoustic space. Although the domestic AI listening agent is bound by data privacy and retention laws that govern networked media, there is a reciprocal relationship at work while we willingly employ their services. Looking closer, we find that the machine learning algorithms in computational listening models and speech synthesis models in these machines constantly improve their diction, listening abilities in noisy environments, speech recognition with varied accents, and sense of directional attention, by constantly listening to and learning from human speech interactions. The personal AI listening machine can operate in a sporadic listening

mode, occasionally recording a snippet of domestic conversations happening in the acoustic space in which they are installed or present. The AI listening machine is able to identify the voice of its primary users, and maintain an idea of what their preferences are by constantly building a schematic understanding from their previous interactions and intent. Therefore, in many ways, the personal AI listening machine transforms the private acoustic space into a flow of commodity exchange to enhance both the expressive as well as cognitive model of its computational auditory scene analysis model, in return for its assistive services. This calls for the mapping of new notions of the private acoustic space and the need to outline various negotiations in the acoustic space surrounding the listening machine. A notable art project that attempts to address these notions is *LAUREN*, where computational artist Lauren McCarthy takes on the job of the smart home agent 'Alexa'. By allowing the artist to control the lights, cameras, door locks and taps in the home of the person, the project speculates the kind of impending negotiations we may have to choose to make or not, as the AI listening agent acquires more ports and modes of domestic control.

Other threats with the potential to invade and disrupt a secure and private acoustic space do exist from a cybersecurity perspective. The growth of specialised audio-based adversarial attacks are rife with the widespread use of domestic AI listening machines. These attacks feature the use of deep recurrent neural networks that embed inaudible frequencies which represent phrases of text (like 'visit xyz.com'), into the sound stream input of the AI listening machine. Such advanced attacks can embed an inaudible waveform into a regular speech command stream, to trick the AI listening machine into processing an alternate spectral representation that results in the wrong linguistic transcription to be perceived by the listening machine (Taori et al. 2019). Ongoing research by the machine listening community to address privacy-related challenges encountered with AI listening machines include privacy-preserving methods for speech and audio processing, de-identification and obfuscation techniques as well as methods to detect adversarial attacks to the operation of AI listening machines. These emerging areas present an opportunity for artist and designers to intervene in the socio-technical acoustic space surrounding the AI listening agent.

Recent work by designers around this topic features the *'Bracelet of Silence'*, which emits inaudible ultrasonic waveforms in the 26KHz range that exploits the non-linear properties of the microphone, to slip into the audible range and jam the microphone with white noise. This renders the AI listening machines present in the room unable to record or 'eavesdrop' into a conversation happening in the immediate acoustic space. The principle of operation of the bracelet was based on the recently published Dolphin Attack that uses the same technique to send inaudible commands to computational AI listening machines installed in smart speakers (Zhang et al. 2017). In summary, the acoustic space around the networked AI listening machine features a number of actors vying for the control and flow of information in the acoustic space, entailing negotiations at the level of technology policy, cybersecurity and social life.

Conclusion

This article has unpacked a socio-technical assemblage of the AI listening machine that resides in its popular avatar as smart speakers and voice assistants. However, several contexts and possibilities for the AI listening machine can emerge if it is detached from its entrapment as a virtual assistant or smart control agent. For instance, cross-disciplinary leaps have been made in applying methods of speech source separation and spectral analysis to the measurement of electricity usage, contactless payment interfaces and the encoding of inaudible tracking beacons into television broadcasts that enter mobile phones through permissive applications. Each of these contexts perhaps shares or might have a modified socio-technical assemblage to the one discussed in this paper.

As discussed earlier, artistic works can be approached as attempts to unveil socio-technical assemblages of AI technologies. This presents an opportunity to shape productive encounters among scientists, technologists and artists to explore forms of representation and discourse variations that skilfully combine experimentation with technology and social commentary.

The methodology and assemblage discussed in this article may not be a template for exploring all AI technologies, but provide thoughts and references for artists and researchers interested in revealing similar socio-technical assemblages. Finally, unlocking the black box of AI applications serves as an invitation for creators and practitioners to question the social and institutional arrangements where their work takes place, as well as their historicity and established assumptions. This can lead to an intentional engagement with models for collaboration and innovation capable of fostering AI applications that respond to and are responsible with the contexts where they are deployed.

References

Bauman, H-Dirksen L., and Joseph J. Murray. Deaf gain: Raising the Stakes for Human Diversity. University of Minnesota Press, 2014.

Bell, Alexander Graham, Lieut-Col Frank Bolton, and William Edward Lang- don. The Telephone: A Lecture Entitled Researches in Electric Telephony (Illustrated Edition). Echo Library, 2017.

Bentham, Jeremy. "The Panopticon." In Offenders or Citizens?, edited by Philip Priestley and Maurice Vanstone, 28–30. London: Willan, 2012.

Bregman, Albert S. Auditory Scene Analysis. The MIT Press, 2001. https://doi.org/10.1016/B0-08-043076-7/00663-X.

Brown, Guy J., and Martin Cooke. "Computational Auditory Scene Analysis." Computer Speech and Language 8, no. 4 (1994): 297–336. https://doi.org/10.1006/csla.1994.1016.

Cherry, Colin. "Cocktail Party Problem." Journal of the Acoustical Society of America 25 (1953): 975–979.

Cox, Christopher. "The Alien's Voice: Alvin Lucier's North American Time Capsule." In Mainframe Experimentalism: Early Computing and the Foundations of the Digital Arts. Berkeley: University of California Press, 2009.

Crawford, Kate, and Ryan Calo. 2016. "There is a blind spot in AI research."

Nature 538, no. 7625 (2016): 311–313.

Deleuze, Gilles. Foucault. University of Minnesota Press, 1988.

Denbigh, Philip N., and J. Zhao. "Pitch extraction and separation of overlapping speech." Speech Communication 11, nos. 2-3 (1992): 119–125.

Dudley, Homer, and Thomas H. Tarnoczy. "The Speaking Machine of Wolfgang von Kempelen." The Journal of the Acoustical Society of America 22, no. 2 (1950): 151–166.

Ganchrow, Raviv. "Perspectives on Sound-Space: The Story of Acoustic Defense." Leonardo Music Journal (2009): 71-75.https://doi.org/10.1162/lmj.2009.19.71.

Kittler, Friedrich. "Thinking colours and/or machines." Theory, Culture & Society 23, nos 7-8 (2006): 39–50. https://doi.org/10.1177/0263276406069881.

Latour, Bruno. We Have Never Been Modern. Harvard University Press, 2012.

Layton, Edward. "Conditions of technological development." Science, Technology, and Society (1977).

Luo, Yi, and Nima Mesgarani. 2018. "TasNet: time-domain audio separa tion network for real-time, single-channel speech separation." In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 696–700. IEEE.

Mara, Mills. "Deaf Jam: From Inscription to Reproduction to Information." Social Text 28 (2010): 35-58. https://doi.org/10.1215/01642472-2009-059.

Mayr, Otto. "The science-technology relationship as a historiographic problem." Technology and Culture 17, no. 4 (1976): 663–673.

Mellinger, David K. "Event Formation and Separation in Musical Sound." PhD diss., Department of Computer Science, Stanford University, 1991.

Parsons, Thomas W. "Separation of speech from interfering speech by means of harmonic selection." The Journal of the Acoustical Society of America 60, no. 4 (1976): 911–918.

Peters, John Durham. "Helmholtz, Edison, and Sound History." In Memory Bytes: History, Technology, and Digital Culture, 177–198. Duke University Press, 2004.

Pickett, J. M. "Historical notes and preface." American Annals of the Deaf 113, no.2 (March 1968): 117–119.

Pinch, Trevor J, and Wiebe E Bijker. "The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other." Social Studies of Science 14, no. 3 (1984): 399– 441.

Price, Derek J. de Solla, Silvio A. Bedini, et al. "Automata in History." Technology and Culture 5, no. 1 (1964): 9–23.

Rice, Albert R. "The Android Clarinettist by Cornelis Jacobus van Oeckelen (1838)." Journal of the American Musical Instrumental Society 40 (2014): 163–189.

Riskin, Jessica. "The defecating duck, or, the ambiguous origins of artificial life." Critical Inquiry 29, no. 4 (2003): 599–633. https://doi.org/10.1086/377722.

Sabine, Wallace Clement, and M. David Egan. "Collected Papers on Acoustics." The Journal of the Acoustical Society of America 95, no. 3679 (1994). https://doi.org/10.1121/1.409944.

Sheets-Johnstone, Maxine. The Corporeal Turn: An Interdisciplinary Reader. Andrews UK Limited, 2015.

Sterne, Jonathan. The Audible Past: Cultural Origins of Sound Reproduction. Duke University Press, 2003. https://doi.org/10.1215/9780822384250.

Stubbs, Richard J., and Quentin Summerfield. "Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners." The Journal of the Acoustical Society of America 87, no. 1 (1990): 359–372.

Taori, Rohan, Amog Kamsetty, Brenton Chu, and Nikita Vemuri. 2019. "Targeted Adversarial Examples for Black Box Audio Systems." In 2019 IEEE Security and Privacy Workshops (SPW), 15–20. IEEE.

Tlalim, Tom. "Tonotopia: Co-designing sound art with hearing implant users: Podcast." 2017.

Triandis, Harry C. The Analysis of Subjective Culture. New York: Wiley-Interscience, 1972.

Voskuhl, Adelheid. "Humans, machines, and conversations: An ethnographic study of the making of automatic speech recognition technologies." Social Studies of Science 34, no. 3 (2004): 393–421. https://doi.org/10.1177/0306312704043576.

Weintraub, Mitchel. 1986. "A computational model for separating two simultaneous talkers." In ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing 11:81–84. IEEE.

Wise, John Macgregor. Exploring Technology and Social Space Vol. 1. Sage, 1997.

Zhang, Guoming, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. "DolphinAttack: Inaudible Voice Commands." In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 103–117. https://doi.org/10.1145/3133956.3134052.

CV

Sharath Chandra Ramakrishnan
School of Arts, Technology, and Emerging Communication
University of Texas at Dallas
looplogic@gmail.com

Sharath Chandra Ramakrishnan is a creative technologist, hybrid practitioner and educator across the cognitive sciences, human machine interaction and technology policy. He is the Director of the Signal Cultures Lab that investigates creative possibilities and techno-cultural implications of pervasive AI technologies of human and machine listening. Previously as a cognitive neuroscience researcher, he studied networks of multimodal and audio cognition in sound and language processing at the National Institute of Mental Health and Neurosciences, Bangalore, India. He is a licensed amateur radio broad-caster (callsign: VU3HPA), extending his creative practice with sound and signals in the wireless spectrum as a transmission & signal artist. His current research in the field of Art & Technology seeks to make novel contributions to the fields of Sound Studies, Auditory Cognition and Machine Listening, prior to which he specialised in AI and interactive virtual environments at the University of Edinburgh, School of Informatics.

Twitter: @AgentSpock
ORCID: https://orcid.org/00000001-7984-9442