Author/s:
Parker, JEK;Dockray, S

Title:
‘All possible sounds’: speech, music, and the emergence of machine listening

Date:
2023

Citation:
Parker, J. E. K. & Dockray, S. (2023). ‘All possible sounds’: speech, music, and the emergence of machine listening. Sound Studies, 9 (2), pp.253-281. https://doi.org/10.1080/20551940.2023.2195057.

Persistent Link:
https://hdl.handle.net/11343/333134

‘All possible sounds’: speech, music, and the emergence of machine listening

James Parker with Sean Dockray

Abstract

‘Machine listening’ is one common term for a fast-growing interdisciplinary field of science and engineering that ‘uses signal processing and machine learning to extract useful information from sound’. This article contributes to the critical literature on machine listening by presenting some of its history as a field. From the 1940s to the 1990s, work on artificial intelligence and audio developed along two streams. There was work on speech recognition/understanding, and work in computer music. In the early 1990s, another stream began to emerge. At institutions such as MIT Media Lab and Stanford’s CCRMA, researchers started turning towards ‘more fundamental problems of audition’. Propelled by work being done by and alongside musicians, speech and music would increasingly be understood by computer scientists as particular sounds within a broader ‘auditory scene’. Researchers began to develop machine listening systems for a more diverse range of sounds and classification tasks: often in the service of speech recognition, but also increasingly for their own sake. The soundscape itself was becoming an object of computational concern. Today, the ambition is ‘to cover all possible sounds’. That is the aspiration with which we must now contend politically, and which this article sets out to historicize and understand.

Keywords

Machine listening, machine hearing, speech recognition, speech understanding, computer music, computational auditory scene analysis, big data, artificial intelligence, machine learning

Introduction

Learning from YouTube

Learning from YouTube (2018) is a video essay by the artist, writer, and programmer Sean Dockray. The essay opens on a blank page in Google Docs. We hear a mouse click and keys tapping as the artist entitles the document with the work's name. The mouse hovers over 'Tools' in the menu bar, moves down to select 'Voice typing', and a microphone symbol appears on left of screen. The symbol turns red as the tool is activated and Dockray starts to speak. 'If I talk about a computer that's able to listen, most people will imagine a computer that can translate the words that I say into some text. They'll imagine the continuous rolling sounds of my speech becoming discrete letters, words, sentences.' As we listen, so does the computer, performing the process Dockray describes. We see Dockray's words belatedly arrive on the page. Already the transcript has several errors. Now the window is suddenly minimised to reveal that we are watching a YouTube upload, with 1,501 views. A conversation between Kylie Jenner and Travis Scott for GQ magazine is being recommended on the right. A new tab opens. 'But human speech is only one kind of sound', Dockray explains. 'There are others. Like cocktail glasses clinking, mouses clicking, the sound of a horse, or the trickle of a river.'

We watch as the artist navigates to a new page. The banner says AudioSet and under it a search bar reads: 'type a sound to filter the ontology'. So he does. 'Clink' brings up an entry listed under 'sounds of things > glass > a short metallic sound'. There are 784 annotations in the dataset, each of another YouTube video. Under 'horse' it reads, 'sounds associated with the familiar domesticated large, fast, and strong ungulate mammal'. 'Clip-clop: the sound of a horse's hoofs hitting on a hard surface.' 3,325 entries. 'Neigh, whinny'. 591. 'Google has created an audio ontology,' Dockray explains, 'a hierarchical categorisation that describes

632 types of sounds. Of these, there are thirteen types of human voice sounds, and only one of these thirteen voice sounds is speech.' The tabs close and now we see another smaller window, playing another clip from YouTube: footage of the artist narrating the essay we hear into a Google Home smart speaker. 'Recognising these other 631 types,' he says, 'is a new frontier in computational listening.'

'How can a machine learn to listen?' Dockray asks. The 'machine' here is a neural network; or maybe Google, or Computer Science, or Capitalism, or something worse (Wark 2019). Whatever the case, it needs data: 'to be exposed to a large number of sounds, be trained to identify patterns, associating the pattern and thus the sound with some concept.' The process, Dockray says, is not unlike training a dog to beg. Computer scientists, however, call it supervised machine learning. To accompany its ontology, Google created a dataset comprising some 2 million '10-second sound clips', all labelled by 'human annotators' for use in training. No information is available about who did this labelling, their backgrounds, or how or whether they were paid. The model here was ImageNet, the 'colossus of object recognition' (Crawford and Paglen 2019) inaugurated by Fei-Fei Li at Stanford in 2006, and now a key benchmark in the field of machine learning more generally. In a paper presented at the IEEE's International Conference on Acoustics, Speech and Signal Processing in 2017, the researchers behind AudioSet explained that they planned to use their dataset 'to create automatic audio event recognition systems that are comparable to the state-of-the-art in recognizing objects in real-world images.' 'We would like to produce an audio event recognizer,' they write, 'that can label hundreds or thousands of different sound events in real-world recordings with a time resolution better than one second – just as human listeners can recognize and relate the sounds they hear' (Gemmeke et al. 2017, 1).

Where ImageNet scraped its images from across the web, the clips in AudioSet are all from YouTube. Suddenly, Google's acquisition of the site in 2006, a year after its launch, reads a little differently. Dockray again: 'At the time it seemed as though Google was buying video content, an audience and their attention, which made sense because Google is essentially an advertising company. Google was buying a kind of global on-demand television station.' More than a decade later, from the perspective of contemporary data science and a company whose business model was now officially 'AI first', YouTube looks like something else: 'a giant pool of data'; a resource to be mined, its value extracted and appropriated; an enormous trove of videos, 'not for us to watch, but for training the cameras and microphones of the near future.' Which is to say, YouTube's primary audience is no longer necessarily human. From now on, Dockray explains, 'YouTube will watch and listen to us.'

Now we understand the essay's visual logic a little better. As the narration unfolds, Dockray has been opening window after window, each with another YouTube clip that speaks in some way to what is being discussed. 'Training a neural network explained.' 'Police secretly use "thought crime" technology on New Orleans residents.' A news item about the regional council in Queensland, Australia, that outfitted its CCTV cameras with microphones. Every clip potential fodder for AudioSet, now or in some future update. Such is the fate of everything uploaded to YouTube, in fact; including the footage of Dockray's own narration and its automatic transcription with which the essay began. An essay on YouTube videos, made entirely of YouTube videos. A film about machine listening that 'anticipates an algorithmic audience'.

Another window plays a 2016 promotional video by Louroe Electronics, an LA-based 'audio monitoring' company, for a product which they claim can 'analyse and identify sounds through advanced algorithms similar to how the human ear processes audio'. The ad is

entitled ‘Why Audio Analytics?’ And the answer seems to be security. The video ‘imagines a range of scenarios to demonstrate the kinds of sounds it is capable of recognizing: glass breaking at night in a showroom of an automobile dealership; a gunshot in a school hallway; and aggression in public space.’ We see grainy black and white CCTV footage of an argument at a crowded bar. ‘A computer listens and dispatches police automatically before an incident turns into a violent outbreak.’ This is not a logic of security rooted in deterrence or the gathering of evidence then, but something more active and interventionist: pre-emptive (Andrejevic 2020a, chap. 4). For Louroe, audio analytics doesn’t just entail the automatic identification of sounds, therefore, but also the automation of response. There are simply too many microphones, too much to listen to for the classificatory work performed by machine listening systems to be implemented any other way. Mark Andrejevic has called this the ‘cascading logic of automation’, whereby ‘automated data collection leads to automated data processing, which, in turn, leads to automated response.’ This, he says, is the ‘overarching trajectory of the current information environment’ (Andrejevic 2020a, 9).

In 2022, recognition of several AudioSet event classes has already been incorporated across Google’s domestic home surveillance line, including doorbells, cameras, smart speakers, and hubs, as part of a Nest Aware home security subscription. ‘Nest Cam looks for motion and listens for conspicuous sounds, like a boom or the crash of a window breaking.’ If it ‘thinks something’s up’ it will send a phone alert or email with a ‘key image from the event,’ and even offer to call emergency services near your home (Nest 2021). Sound notifications are now available on Android phones and watches too, marketed, for the time being, primarily at the deaf and hearing impaired. Alexa Guard similarly listens for glass breaking, but also human sounds, like snoring, coughing, or a baby crying (Amazon 2019). After an update in 2020, Apple Watch listens out for the sounds of ‘running water’ and ‘squishing soap’ to monitor and provide feedback on how long wearers are washing their hands (ABC News

2020). Walmart has proposed using ‘sound analysis’ to generate and administer employee performance metrics (Jones 2018). ShotSpotter (2022) provides automatic gunfire detection and location technologies to law enforcement agencies in more than one hundred cities across America (Kang and Hudson 2022; Parker and Abu Hamdan 2022).

Machine listening

‘Machine listening’ is one common term for the fast-growing interdisciplinary field of science and engineering driving these developments. History is full of references to ‘listening machines’, to describe things like ear trumpets, phonautographs, phonographs, spectrographs, and ‘listening typewriters’ (Sterne 2003; Mills 2011a; Sterne 2012b; Li and Mills 2019). But the use of ‘machine listening’ to suggest a new, specifically machinic *form* of listening, rooted in the techniques of information theory and digital signal processing, seems to arrive only in the 1990s, starting in computer music. Writing in 2006, Malkin defined machine listening, rather minimally, as the ‘process of deriving information’ from audio signals that are ‘useful for some computational or human purpose’ (Malkin 2006, 3). And for the Machine Listening Lab at Queen Mary University of London, the term refers similarly to the ‘use of signal processing and machine learning to extract useful information from sound’ (2018).

But ‘machine listening’ is just one term among many currently in circulation. Richard Lyon—a principal research scientist at Google, author of one of the more recent and broad-ranging books on the field, and part of the team behind AudioSet—prefers ‘machine hearing’ (Lyon 2010; 2017). ‘Machine listening’, he says, is mostly used ‘in connection with music listening and performance’. And ‘terms like *computer hearing*, *computational hearing*, and *computer listening* seem awkward, especially since I spent a lot of years building analog electronic models of hearing, probably not qualifying as computers’ (Lyon 1978; 2017, 11).

Wang (2010) uses ‘machine audition’, Schuller (2014) ‘intelligent audio analysis’, and as recently as 2018 Friedland et al were making the case for a new field called ‘computer audition,’ the goal of which would be to ‘enable machine systems to replicate and surpass the inferences humans make from sound’ (Friedland et al. 2018, 32).

This kind of terminological wrangling is standard for any field still undergoing institutionalization. But it is noteworthy that in 2022 the process isn’t further along. One of the defining features of research on machine listening and its many synonyms is a recurring anxiety about its relative infancy or underdevelopment, especially compared with machine or computer vision (Bregman 1990; Ellis 1996; Lyon 2010). Perhaps there is something to this anxiety, but there are also several instances where machine listening has preceded or prefigured developments in machine learning more generally, as we will see. Besides, machine listening systems have been around now for a long time, and they are increasingly pervasive. Charismatic consumer applications like Siri and Alexa are only the most familiar manifestations of a constellation of techniques and technologies that are defined partly by their invisibility, by the fact they go *un*-noticed. Machine listening systems are currently being rolled out by every one of the ‘big five’ US tech giants, as well as the behemoths of China (Hvistendahl 2020), and beyond. At the same time, they are being thoroughly integrated by states into infrastructures of communication, administration, education, healthcare, transport, surveillance, security, incarceration, and law enforcement, all of which rely heavily on publicly funded research as well as commercial contracts and partnerships (Dockray, Parker and Stern, 2020).

For some, the hubristic endgame is ‘curing’ machinic ‘deafness’ (Hawley 1993; Lyon 2010), an ableism endemic to the histories of both audio and computer science; and, indeed, capitalism (Mills 2011b). The trouble here isn’t just the characterization of inaudition as a

defect in need of fixing. It's also that deafness is constantly used as a pretext for the development and normalization of some new technology, often without meaningful consultation and with support often abandoned the moment this becomes commercially viable (Mills 2010). For others, across academe and industry alike, the problem isn't machines' inability to listen, but that so much of the world remains unaudited. In this way of thinking there is always more to hear, more data to be exhorted (Harcourt 2015), gathered and analysed, more classifications to be performed and acted upon. The sounding world can always be made more productive. Normative human listening is less a model here than a limit to be overcome. Either way, machine listening gets defined by a deficit, which is also a motive—to listen more—so that the field paradoxically imagines itself in a state of perpetual immaturity, even as it proliferates, and its influence accrues.

As machine listening systems develop and spread, they are garnering increasing critical attention. Here, 'machine listening' does seem to be the preferred terminology (House 2017; Maier 2018; Semel 2019; Li and Mills 2019; Mattern 2020; Sterne and Razlogova 2021; Ernst 2021; Sterne 2022; Sterne and Sawhney 2022; Kang 2023). Perhaps this has something to do with the specific researchers and archives these scholars are choosing to engage with. Or maybe the term's origins in music and so its relative proximity to humanities departments. But the kinds of association that might have appealed to composers and musicians in the first place hold just as well today. Terms like 'machine hearing' and 'computer audition' are meant to sound biological, and help to present the field as politically inert. Whereas to talk about machine 'listening' already seems to suggest the intertwining of technology, aesthetics, ethics, and politics (Barthes 1991; Rice 2015; Lewis 2018). Similarly, one advantage of emphasizing the 'machine' over the 'computer' is that it situates the field within a much older tradition of technological critique: one that includes factories, phonographs, and the

machineries of state (Mumford 1966; Brewster 1972; Marx 1976; MacKenzie 1984; Scott 1997; Bijsterveld 2008; Weber 2013).

Nevertheless, it is tempting to observe that, as a turn of phrase, the trouble with 'machine listening' is that it isn't just machinic, and it isn't exactly listening either (Dockray, Parker and Stern, 2020; Sterne 2022). Many of the same critiques made of artificial intelligence, automated systems, and capitalist technoscience apply equally here. Contemporary machine listening systems are *also* extractive (Fuchs and Mosco 2016; Couldry and Mejias 2018; Crawford and Joler 2018; Wark 2019). They also reproduce bias (Lawrence 2019; Phan 2019; Andrejevic 2020a), extend and normalise surveillance (Vetter 2012; Szendy 2017), and undermine autonomy and privacy (Couldry and Mejias 2018; AI Now Institute 2018). They are also prone to revitalising long-debunked ideas from physiognomy (McQuillan 2018a; Semel 2019; Li and Mills 2019), and regularly amount to little more than snake-oil in practice (Feldman 2016; Pfeifer 2021; Sterne and Razlogova 2021; Parker and Abu Hamdan 2022). For a field that continues to think of itself as marginal, work in machine listening can be incredibly hubristic. Time and again, researchers minimise or disavow the politics of what they do, even as they recommend and develop it for use in response to all manner social and political challenges: in the 'fight against COVID-19' (Schuller et al. 2020, 4), for instance, or to 'help save the planet' (Schuller et al. 2022). As in so many other areas of science and technology, innovations tend to be depicted as revolutionary, whereas risks get framed narrowly—typically as matters of ethics, with technical solutions—and either outsourced, dealt with extreme brevity, or simply ignored. That is a major problem. But it is hardly specific to machine listening, and has been written about many times before (Beck 2009; Kahn 2013; Hurlbut 2018; Andrejevic 2020b).

'All possible sounds'

This article contributes to the growing critical literature on machine listening by presenting some of its history: as a field, or constellation of fields, but also as a commercial enterprise, sociotechnical imaginary (Jasanoff and Kim 2015), desire, or network. As we will see, from the 1940s to roughly the end of the 1980s, work on artificial intelligence and audio developed along two major streams. There was work on speech recognition/understanding, which was mostly funded by and arranged according to the interests of Bell, the US Department of Defense's Advanced Research Project Agency (originally ARPA, later DARPA), and IBM. And there was work in computer music: on the aesthetic possibilities of composing and improvising with 'intelligent' listening systems, as well as certain applications with more commercial appeal. It was in relation to such systems that the term 'machine listening' was originally, or most durably, coined.

Then, starting in the early 1990s, another stream of work began to emerge. At institutions like the MIT Media Lab and Stanford's Center for Computer Research in Music and Acoustics (CCRMA), researchers turned towards 'more fundamental problems of audition': problems like the automatic segmentation, organization, and classification of 'complex sound mixtures' (CCRMA 1992, 25). By the end of the 1990s, speech and music would increasingly be understood by computer scientists as particular sounds within a broader 'auditory scene'. Researchers began to develop machine listening systems for a much more diverse range of sounds and classification tasks: often in the service of speech recognition, which remained a major source of state and corporate investment, but also increasingly for their own sake. For the first time, the soundscape itself was becoming an object of computational concern.

Google's AudioSet is a particularly clear articulation of this new orientation, as well as an influential project in its own right. With its 632 audio event classes, it is at once extremely

ambitious and comically arbitrary or incomplete. The category ‘human sounds,’ for instance, is broken down into ‘human voice’ [which includes tags like ‘speech,’ ‘shout,’ ‘screaming’ and ‘whispering’], ‘respiratory sounds’ [‘breathing,’ ‘cough,’ ‘sneeze’ …], ‘human group action’ [‘clapping,’ ‘cheering,’ ‘applause,’ ‘chatter,’ ‘crowd’ …] and so on across six other categories. Like an acoustic version of Borges’ imagined ‘Chinese Encyclopedia’ (Foucault 1971), the rest of the sounding world is divided into ‘animals,’ ‘source ambiguous sounds,’ ‘sounds of things,’ ‘music,’ ‘natural sounds,’ and ‘channel, environment and background.’ There are 76,767 samples labelled ‘inside, small room,’ 2,907 labelled ‘music of Africa,’ and a further 3,164 marked ‘insect’. Taken to its logical conclusion, the ambition is nothing less than ‘to cover all possible sounds’, as Dan Ellis (2018), one of AudioSet’s engineers and a long-time leader in the field, explained. That is the scientific and commercial aspiration with which we must now contend politically, and which this article sets out to historicize and understand. British company Audio Analytic, self-described ‘pioneers of artificial audio intelligence’ and owners, they say, of ‘the world’s largest, commercially-exploitable audio dataset,’ put it very simply. They are ‘on a mission to give all machines a sense of hearing… Intelligent sound recognition everywhere’ (Audio Analytic 2018).

This article is about how we got to this point: where the entire world of sound is claimed to be computationally knowable. As we will see, the answer is not—or not just—‘big data’, and the infrastructures on which it depends (Whittaker 2021), even though contemporary systems like AudioSet do rely on massive datasets and recent developments in machine learning. The ‘problems’ which the latest and greatest technologies claim to solve had first to be formulated. And from the perspective of machine listening’s history, for instance, many of the event classes included in AudioSet, several of which are now recognized by various Google products (alarms, bells, and barks, along with various ‘auditory scenes’), look like ‘classic’ objects of inquiry: part of a nascent machine listening imaginary already more than

twenty years earlier. Machine listening's ambition—to 'cure' machinic 'deafness', to 'cover all possible sounds' (Ellis 2018)—was being articulated before 'big data' was even on the horizon. We could reverse the logic, in fact. Machine listening didn't arrive with the datafication of our sensory environments, it helped bring this datafication about.

Archivally, this study draws on diverse sources. Part 1 works primarily with existing histories of speech recognition and understanding to resituate them as part of a more general history of machine listening. Parts 2 and 3, on computer music and auditory scene analysis respectively, draw mainly on journal articles, conference papers, doctoral theses, monographs, textbooks, and edited collections written by the musicians, computer scientists, and engineers most responsible for developing machine listening as a field. Since many of these texts were authored in the 1990s, a period in which scholars were starting to share, advertise, and archive their work and related activities online, the Internet Archive has been a particularly valuable resource. The history we derive from these materials is admittedly incomplete. If it is US-centric, that is partly because of America's outsized importance in the history of the field, and certainly in the anglosphere. The US is not only where the term 'machine listening' was coined, and where a constellation of researchers first began using it to describe their field. It was also a major source of investment—via companies like Bell and IBM, but also, for instance, (D)ARPA and the National Science Foundation—so that top researchers from the UK, Europe and Japan often ended up working in America. Nevertheless, a history that gave more attention to institutions like Institute de Recherche Acoustique Musique (IRCAM) (Born 1995), in France, or Nippon and Yamaha, in Japan, would certainly be welcome.

Likewise, the story we tell here focuses more on the hows and whens of machine listening's emergence than on its social and political impacts, even though these provide a constant motive and point of orientation (Dockray, Parker and Stern, 2020). Partly, that is because this

work is already starting to be done by others. What this piece offers is more definitional. It is an exercise in ground clearing: a partial response to the question, what are we talking about when we talk about machine listening? One of the article's main contributions in this respect is to draw out the importance of music and musicians in the field's history. It isn't just that many of the pioneers of speech recognition were personally interested in music, as we will see, or that this interest often spilled over into collaborations with peers in computer music. Ideas moved fluidly between them. Indeed, spaces for this kind of exchange were actively cultivated by the likes of Bell, MIT, and Stanford, which also courted broader industry investment and support. Composers working with computers were an 'untapped intelligence resource': experts on listening, as well as an engine of creativity. The concert hall was also a laboratory. Sometimes, it was good marketing too. And it was out of work being done by and alongside musicians, with Computational Auditory Scene Analysis (CASA), that the push towards a more encompassing listening would eventually come. An approach that marks the emergence of machine listening proper: a listening beyond speech and music; a listening without limits.

Speech

Speech Recognition

Today, speech recognition is machine listening's most familiar application, as well as its major commercial success. Automatic Speech Recognition (ASR), more than any other technology, is habituating us to a world of networked microphones: teaching us to expect our devices to listen; training us to want them to. Its history has been addressed in both academic and popular writing, most notably by the computer scientist Roberto Pieraccini (2012; 2021), currently a director of engineering at Google, and media studies scholar Xiaochang Li, as part

of her history of automatic text prediction (2017). As Li explains, the very earliest efforts at speech recognition were undertaken with office work in mind. W.H. Barlow's 'logograph', for instance, was conceived in 'an attempt to make something which would do short handwriting instead of employing the services of the gentleman who sits there [transcribing]' (Barlow, 1878, quoted in Li 2017, 45). Though it was never completed, J.B. Flowers' 'voice-operated phonographic alphabet writing machine' was being developed in collaboration with the Underwood Typewriter Company as early as 1916 (Li 2017, 48). By mid-century, however, the most important work on speech recognition—as indeed on information theory more generally—would be driven by the commercial interests of AT&T Bell Laboratories, then the largest industrial lab in the world (Mills 2011, 77), and the US Department of Defense. At Bell this meant a focus on telephony, and the cheap and efficient processing of calls.

Already in 1952, Bell Laboratories' Automatic Digit Recogniser, nicknamed 'Audrey,' could recognize all ten digits, though only where the speaker was designated in advance, a major limitation in practice (Li 2017). The process worked by 'template matching.' Incoming speech was first rendered graphically by a sound spectrograph and then matched against stored 'templates' based on the average spectral values of around a hundred utterances of each digit by that same speaker. As experimental phonetician Peter Denes explained in a 1960 paper written on contract for the US Air Force (Denes 1960), this was a fundamentally 'acoustic' way of conceiving speech, since it 'assumed that there are some invariant acoustic features that characterize a phoneme and that are always present when that particular phoneme is spoken by the speaker or recognized by the listener.'

Denes' breakthrough, with D.B. Fry, a colleague at University College London (Fry and Denes 1958) before both moved to Bell, had been to reimagine speech recognition as a two

stage or dialectical process, in which acoustic processing worked hand in hand with a rule-based 'language model' determined in advance by linguists, and intended to reproduce, however approximately, 'the linguistic mechanism of the listener' (Fry and Denes 1958, 53; Li 2017, 122). The purpose of the model, Fry and Denes (1958, 35) explained, was to 'replace … the human being in the chain: acoustic speech input – human being – typewriter or teleprinter.'

Most contemporary ASR systems still divide the task of recognition in two like this.[1] But the 'rule-based' approach to linguistic decoding—which would start with 'fundamental acoustic facts, like pauses or high-energy voiced segments of speech' and proceed, level by level, up a painstakingly constructed chain of inferences to 'higher-level conclusions about words, phrases, sentences, and finally meanings' (Pieraccini 2012, 86)—has long since been overtaken by the kinds of 'brute force', 'statistical' and 'data-driven' methods pioneered by IBM in the 1980s (Pieraccini 2012; Li 2017).

Speech understanding

Through the 1960s and 1970s, however, most of the labs in the US working on artificial intelligence still 'had at least one project targeted at developing expert systems that understood language and speech' (Pieraccini 2012, 85), and many of these were funded by DARPA. Indeed, DARPA's 'Speech Understanding Research Program' (SUR), which launched in 1971, is often considered its 'first major AI effort' (Li 2017, 69), and it was thoroughly committed to rule-based methods (Li 2017, 70). Among others, SUR funded projects at MIT, Stanford, Carnegie Mellon, and Bolt, Beranek and Newman (BBN), a Boston company founded by former directors of MIT's Acoustics Laboratory and acquired by Raytheon in 2009. All have gone on to play major roles in the history of speech recognition, machine listening, and computer science more broadly.

What distinguished these projects from earlier work in the field was that they were tasked with 'understanding' speech as opposed to 'recognising' or 'transcribing' it. Hence the program's title. Where previous work had been conducted with the automation of feminized secretarial and call-routing work in mind, DARPA was more interested in quick and efficient human-computer interaction for the purposes of military command and control. Speech understanding played a prominent role in Cold War 'cyborg discourse' in fact (Edwards 1997). J.C.R. Licklider—a psychoacoustician, president of the Acoustical Society of America and vice-president at BBN before taking on a senior role at ARPA where he 'aggressively promoted' AI for command and control systems (Edwards 1997, 240)—had put it like this in a famous paper on 'Man-Computer Symbiosis' from 1960:

> 'The military commander … faces a greater probability of having to make critical decisions in short intervals of time. It is easy to overdramatize the notion of the 10 minute war, but it would be dangerous to count on having more than 10 minutes in which to make a critical decision. As military system ground environments and control centers grow in capability and complexity, therefore, a real requirement for automatic speech production and recognition in computers seems likely to develop.' (Licklider 1960, 10)

And in a 1962 article entitled 'Eyes and Ears for Computers', E.E. David Jr and O.G. Selfridge explained that 'all agree' on the need to have speech recognition 'before the Russians' (David and Selfridge 1962, 1093).

DARPA's orientation towards command and control had direct consequences for the specifications of the projects it funded. Success would be determined, Li explains (2017, 81), 'based on whether or not a system correctly responded to a spoken command, regardless of whether the words themselves were accurately recognized.' Raj Reddy's 'Hearsay II' system

at Carnegie Mellon was designed for the inventory and retrieval of documents, for instance. And BBN's 'Hear What I Mean' answered travel questions: about plane times, fares, and so on (Pieraccini 2012, 92). In both cases, meaning took priority over accuracy of transcription. What mattered was the system's ability to discern what information the speaker was requesting, not how perfectly it heard the request. And they were not particularly good at either.[2]

The statistical turn

The decisive turn away from rule-based language models towards the statistical and brute force methods that dominate today was initiated by the Continuous Speech Recognition (CSR) group at IBM and driven in part by the company's market interests in business administration and data processing. Already the largest computer company in the world by the end of the 1960s (Estabrooks 1995, 50), 'IBM considered speech recognition to be a strategic technology for its business. Typewriters were an important sector of its market,' Pieraccini writes (2012, 109), 'and a voice-activated one would be a "killer application" for the corporation.' For Li, this commercial orientation towards dictation and transcription was a key factor in IBM's willingness to pursue statistical methods. The company's relative disinterest in what words *meant*—and only in what was *said* (transcription)—made it that much easier to 'decouple recognition from reasoning' (Li 2017, 81). Which is precisely what it did.

From the CSR group's inception in 1972, researchers at IBM 'treated speech as a system that transformed words into sounds by means of "unknown human mental and physiological processes" that could not be ascertained and represented explicitly,' but whose outcomes could nevertheless be predicted statistically. Statistical modelling was understood 'not as a novel means for codifying and quantifying pre-existing knowledge of linguistic principles,

but as a replacement for linguistic principles entirely.' (Li 2017, 130) That is, the hidden Markov models (HMMs) pioneered by the CSR group, and which would go on to become 'statistical speech recognition's defining technique' (Li 2017, 133), don't just lack domain specific knowledge, they proceed from the starting point that such knowledge is *irrelevant* to the task of computational modelling. This move was not just crucial to the development of speech recognition and machine listening in the following years, it would go on to become a 'conceptual cornerstone in the foundation of big data analytics' more generally (Li 2017, 141): an era of 'computational empiricism' (Goldenfein 2019), in which it is increasingly held by computer scientists and their evangelists (Anderson 2008) that nature can best, or indeed *only* be accessed through computation.

IBM's focus on transcription was not the only commercial factor driving its preference for statistical methods. Working with linguists was time and labour intensive. The company's unparalleled access to computing resources meant they were uniquely positioned to experiment with the 'brute force' methods demanded by their statistical approach. And 'whereas efficient operation was a defining factor in research at both Bell Labs and the DARPA SUR program, IBM pursued speech recognition for its potential to be *inefficient*,' since increased computational requirements would drive consumer demand for newer, faster machines (Li 2017, 81-2).

Early speech recognition markets

It took a decade for the CSR to finally debut Tangora, the first commercial application of its research, and even longer for the statistical approach to be widely taken up by speech recognition researchers and beyond (Pieraccini 2012, 132). When IBM launched a PC-based version of Tangora in 1986, it could take dictation 'using a 5,000 word "office correspondence" vocabulary spoken with brief pauses between words' (Li 2017, 128). This

orientation towards office correspondence was, of course, a function of the market for computers in the 1980s, which were neither affordable nor designed for home users. But it was also an artifact of the kinds of speech corpora available at the time. In the 1970s and 1980s, data were not easy to come by, and it was official contexts—where documentation was often mandated and sometimes publicly available—that provided the readiest access. For instance, the largest contribution to the dataset on which Tangora was trained came, rather ironically, from a lawsuit brought by the Federal government against IBM itself, alleging monopolisation of the business computing market (Li 2017, 148). The depositions alone yielded 100 million words of diverse but highly formalised speech. By the mid 1980s, the group had also incorporated a decade's worth of transcripts from the Canadian Parliament (Li 2017, 149).

IBM continued to develop and sell dictation software to business across the 1990s: especially to industries like law and medicine which already relied heavily on transcription services and were, therefore, willing to invest the time and money required to train the software. In addition to manufacturing dictation machines, for instance, Dictaphone ran a lucrative medical transcription service, for which it had employed 'armies of human transcribers' (Pieraccini 2012, 214) prior to the mid-1990s. Once dictation software became more widely available, however, the company found it could get by with significantly fewer workers, with human transcribers necessary mainly for the purposes of training and corrections.

The first dictation tool for home computing, Dragon Systems' DARPA funded Naturally Speaking™, was released in 1997, with IBM's own ViaVoice™ following a month later. Microsoft had entered the market within a few years, before incorporating dictation software into the Windows operating system in 2006 (Pieraccini 2012, 212-214). In terms of public awareness of speech recognition, however, probably the most significant moment came in

1992, when researchers at AT&T Bell Laboratories finally made good on the promise of Audrey, and AT&T rolled out 'Voice Recognition Call Processing' for a handful of its most popular operator-assisted tasks. The company laid off more than 6,000 human operators in the process, just as workers had protested it would. 'We try to make the distinction between the individual and the job,' one manager at AT&T explained (Wolfinger 1990). 'It's not that you are being fired; it's the job that's being eliminated.' A version of the system has been in continuous use ever since, handling upwards of 1 billion calls each year; just a fraction of the total number across similar systems round the world. For the twenty years prior to Apple's launch of Siri in 2011, the public's major experience of speech recognition had been via automatic call-routing technologies. This experience, moreover, was often exasperating and unpleasant: not just because these systems failed so often, but also because of the sense of alienation and distance they helped to produce.

Music

Computer music

Though early work on speech recognition was highly invested in researching, reproducing, and ultimately replacing human listening/listeners for certain tasks, the term 'machine listening' was not widely used until the 1990s, starting in computer music. Today, most music is in one way or another computer-mediated. But in the final decades of the 20th century, 'computer music' described a small enclave of Western art-music (Dean 2009). The first International Computer Music Conference, for instance, was held in 1974. A dedicated journal followed three years later (Snell 1977; 2006). The connections between computation and certain strands of elite and avant-garde music culture extend back long before, however (Dean 2009; Bell 2019). Indeed, many of the pioneering figures in speech recognition,

cybernetics and information theory had also worked on music, and often in the same institutions. J.R. Pierce, of Bell labs, though he would go on to issue a scathing denunciation of speech recognition research (1969), had been one of its early proponents. 'We have all dreamed, if we have not read,' he wrote in 1946 (cited in Li and Mills 2019, 143), 'of voice operated typewriters which take down dictation without fatiguing or distracting, yet our preoccupation with written words which bear little relation to speech has prevented our realizing even this. The voice-operated lock, the voice-operated furnace—a whole host of voice-operated mechanical servants have languished in the land of the unborn.'

A year later, Pierce was working with Claude Shannon on preliminary designs (Li and Mills, 143), and soon after that with Mary Shannon on a 'probabilistic expert system' for music composition in four-part harmony (J. R. Pierce and Shannon 1949; Yu and Varshney 2017; Bell 2019, 157). One such composition went on to feature on *Music from Mathematics*, an LP released by Bell in 1961, which also featured compositions by computer music pioneer Max Mathews, among others. According to Manfred Schroeder—another important figure in speech recognition at Bell (Schroeder 1985) and, with Mathews, a founding member of IRCAM in Paris—AT&T administrators were 'not enthusiastic' about the company's music programs (Xiang and Sessler 2015, 369). When asked to justify the continued funding of this sort of work, however, Pierce and Mathews were apparently able to show 'how music synthesis grew directly out of vital speech compression research and how music synthesis techniques fed back useful technology to speech synthesis' (Xiang and Sessler 2015, 369).[3]

By 1966, Bell would both sponsor and contribute engineers to *9 Evenings: Theatre and Engineering*, 'a landmark in the history of performance-theatre-multimedia art' (Bijvoet 1990, 31–33), which led to the establishment of the influential *Experiments in Art and Technology* (Dyson 2006; Turner 2008). *9 Evenings* featured new works by John Cage,

Lucinda Childs, Öyvind Fahlström, Alex Hay, Deborah Hay, Steve Paxton, Yvonne Rainer, Robert Rauschenberg, David Tudor, and Robert Whitman, all produced in collaboration with Pierce, Mathews, Schroeder, and others from Bell. In all, nineteen engineers apparently contributed more than 2,500 hours to production of the event. 'The real cost of the extravaganza', according to Douglas Davis (quoted in Bijvoet 1990, 23), was more than $100,000, and the total audience for the nine days around 10,000. In a letter to Pierce requesting Bell's sponsorship, Billy Klüver, an engineer at Bell and, with Rauschenberg, the driving force behind the event, listed three main benefits to the company: 'first, "social prestige and [an] increase in professional standing"; second the involvement of artists as "an untapped intelligence resource"; and finally, the potential "commercial fallout, patents, methods, products, and ideas" (Dyson 2006). Writing about the deliberate comingling of artists and engineers at the MIT Media Lab some twenty years later, Stewart Brand (1987, 83) would talk similarly of 'art for invention's sake'. Science and engineering, in this way of thinking, are not the only, or even necessarily the privileged, drivers of 'innovation' (see also Beck and Bishop 2020).

Meanwhile, work in computer music was beginning to broaden out from composition and synthesis towards listening, improvisation and live performance. At Stanford's ARPA-funded Artificial Intelligence Laboratory, in the early 1970s, James Moorer was starting to reimagine methods from speech understanding for use in pitch detection for automatic music transcription (Moorer 1975, 18). 'A computerized musical scribe probably has its greatest application in the field of Ethnomusicology,' Moorer explained (1975, 1), 'where often hundreds of hours of recorded ethnic music are commonly transcribed by hand. A more long term application is in the field of computer music, where we might expect the computer to be able to perceive music as well as play it, thus taking its cues from the musicians (or other computers?) with whom (which?) it is playing.' This was precisely the line of thought

pursued by Barry Vercoe at MIT, who was already composing with what he called 'synthetic listeners' in 1984, with a view to '[moving] computer music clearly into the arena of live music performance,' and in doing so 'to recognize the computer's potential … as an intelligent and musically informed collaborator in live performance' (Vercoe 1984; 1990).

Interactive music systems

It was a student of Vercoe's, in fact, who seems to have coined the term 'machine listening' in print. Robert Rowe had come to MIT for a PhD in the 'Music and Cognition' group at the Experimental Music Studio, which had been founded by Vercoe in 1973 and was absorbed into the Media Lab from its inception in 1985. For Rowe, 'machine listening' referred to the analytic layer of an 'interactive music system': 'a computer program able to listen to music' in real time and respond to the input of live performers (Rowe 1991; 1992; 1993).

In this, Rowe was influenced not only by Vercoe, but also by other composers like David Behrman, Richard Teitelbaum, Jean-Claude Risset, David Wessel, Tod Machover, and George Lewis, who had been working on musician-computer interactivity across the 70s and 80s.[4] *Rainbow Family* (1984), for instance, the first of Lewis''interactive virtual orchestra' pieces, was premiered to a 'packed house' during a residency at IRCAM in 1984 (Steinbeck 2018, 264; Lewis 2018). Alongside four improvising soloists—Derek Bailey, Douglas Ewart, Steve Lacy, and Joëlle Léandre—the piece features 'a trio of Yamaha DX-7 synthesizers controlled by Apple II computers' running software written by Lewis. This software both responds live to the sounds of the human performers, and operates independently, according to its own internal processes. There is no 'score'. And Lewis is not, in the language of European 'art music', the piece's 'composer'. Instead, like its more famous successor *Voyager* (1987), *Rainbow Family* comprises 'multiple parallel streams of music generation, emanating from both the computers and the humans—a non-hierarchical, improvisational,

subject-subject model of discourse' (Lewis 2000, 33). Crucially, for Lewis, this was no mere artifact of the technologies available to him. It was a deliberate aesthetic, technical and political strategy, to produce, he said, 'a kind of computer music-making embodying African-American aesthetics and musical practices' (Lewis, 2000, 36-37); a form of human computer collaboration with similar ideals to those of the African American musicians' collective AACM (the Association for the Advancement of Creative Musicians), of which Lewis had been a member since 1971, and, for Paul Steinbeck, in a big call (2018, 261), 'the most significant collective organization in the history of jazz and experimental music.'

Like Lewis, Rowe had spent time at IRCAM in the 1980s. And Lewis would go on to perform on trombone at the debut of Rowe's own interactive music system, *Cypher*, at MIT in 1988 (Rowe 1993, 72-3). Both composers' software 'listened' to or via MIDI, a 'hardware specification and communications protocol' (Rowe 1993, 10) introduced by a consortium of instrument manufacturers—Yamaha, Roland, Korg, Kawai and Sequential Circuits—in 1983, and which still functions today 'as the circulatory system of most digital information in modern music production' (Diduck 2018, 21). The great benefit of MIDI is its efficiency: the fact that it 'allows computers, controllers, and synthesis gear to pass information among themselves' so seamlessly (Rowe 1993, 10). Indeed, for Rowe, the 'fast growth in the development of interactive music systems is due in no small part to the introduction of the MIDI standard' (Rowe 1993, 10). As with all digital representation, however, efficiency comes at a cost: in the case of MIDI conversion because sound is represented with so few parameters—pitch, velocity, duration, and a 'control change' value—all oriented around the 'note concept' (Rowe 1993, 10) and tuning system of European art music; and moreover, because even this 'restricted form of representation is often inaccurate' (Rowe 2001, 29).

Negotiating this compromise is at the heart of *Cypher*'s 'machine listener'. The point, Rowe explains (1991, 15), is precisely 'not to "reverse engineer" human listening, but rather to capture enough musicianship, and enough ways of learning more complete musical knowledge, to allow a computer performer to interact with human players on their terms, instead of forcing an ensemble to adapt to the more limited capabilities of an unstructured mechanical rendition.' This principle extends throughout *Cypher*'s design. Whereas the kinds of statistical methods pioneered in speech recognition by IBM were increasingly decentering human categories in favour of the 'natural way for the machine' (Frederick Jelinek, cited in Li 2017, 25), Rowe championed the 'unsuspected power' (Rowe 1993, 207–8) of rule-based approaches to the modelling of musical knowledge, making only occasional use of neural nets for chord and key recognition (Rowe 1993, 231). 'It is my belief,' Rowe explains:

> that trying to follow the example of human cognition in building a machine listener will be fruitful not only because humans provide the best, indeed the only, model of successful music listeners but also because couching an interactive system in terms compatible with human listening categories will encourage the generation of music that, though procedurally generated, will resonate with the human cognitive process for which it is ultimately intended.' (Rowe 1993, 120)

In this, Rowe's 'most pervasive influence' was Marvin Minsky (Rowe 1991, 14), another key figure in computing with more than a passing interest in music. Rowe often refers, for instance, to Minsky's essay 'Music, Mind, and Meaning' (1982), written around the same time as a series of lectures on 'The Computer and the Composer' at IRCAM in 1981. Minsky would regularly cite his experience as a pianist, along with his interest in Baroque counterpoint as influences on his theory of mind. But in this essay, he directly applies a version of his agentic theory of intelligence to music (mostly Beethoven), a full five years

before the publication of *Society of Mind* (1986). In one passage, he even sketches a basic machine listening architecture, which he calls a 'Music-Agent' (Minsky 1982, 10). 'The central idea of Minsky's theory,' Rowe explains (1993, 208), 'is that the performance of complex tasks, which require sophisticated forms of intelligence, is actually accomplished through the coordinated action and cross-connected communication of many small, relatively unintelligent, self-contained agents.' *Cypher*'s listener likewise 'combines the action of many, small, relatively simple agents. Agencies devoted to higher-level, more sophisticated tasks are built up from collections of low-level agents handling smaller parts of the problem and communicating the progress of their work to the other agents involved' (Rowe 1993, 208-9). Like Lewis' systems before it, but with a very different politics, *Cypher* was thus an 'applied music theory': a collection of ideas about music's representation, production, and appreciation 'formalized to the point that they can be implemented in a computer program and tested in real time' (Rowe 1992, 43). We could say the same about all machine listening systems in fact. Machine listening is always a form of applied philosophy (McQuillan 2018b). The question is which.

Machine Listening and the Media Lab

Rowe's own motives for developing *Cypher* were mainly aesthetic. Working from within Western compositional traditions, but far more open to its recent avant-garde strands than Minsky,[5] he was interested, he said, in the 'new compositional domains' that interactive music systems seemed to suggest (Rowe 1999, 84). Rowe was aware of the 'real and growing threat' computer music posed 'to the livelihood of human musicians' (Rowe 2001, 3), but the great advantage of interactive music systems, he said, was precisely the fact that they *require* human participation. In this, they not only mitigated some of the risks of automation, but also

held out the possibility of an ‘actively engaged audience’ in an era still defined for Rowe by mass media and passive consumerism.

This is what Andrejevic (2004) has called ‘the promise of interactivity’. Over the early 2000s, it would become one of the main alibis for mounting computational surveillance and new forms of value extraction, in which audience ‘engagement’ is at once commodified and itself reimagined as a new form of labour. But for Rowe, writing in 2001, it still felt like a virtue. ‘If interactive music systems are sufficiently engaging as partners,’ he surmised (2001, 5), ‘they may encourage people to make music at whatever level they can.’ And this in turn was crucial to the ‘vitality and viability of music in our culture’. ‘As computer music is so often accused of leading us to a day when machines will listen only to machines,’ Rowe writes (2001, 6), ‘I feel compelled to observe that many of us are motivated by a much different vision of the computer's potential connection to the community of human musicians.’

All this is very close to Media Lab founder Nicholas Negroponte’s ‘humanism through machines’, which was undoubtedly an influence on Rowe. Articulated already in the opening lines of *The Architecture Machine* (1970, 1), it would become a guiding principle at the Media Lab, and go some way to explaining the Lab’s deliberate collocation of artists and engineers. On the one hand, artists like Rowe, in their ‘restless creativity’, were a ‘resource’ to be mined in the name of humanistic innovation: art for invention’s sake (Brand 1987, 82). ‘If people with some art background’ were prepared to participate ‘directly in technical innovation,’ Negroponte explained, then they were welcome (Brand 1987, 83), and could contribute, moreover, to the highly aestheticized ‘demo’ culture that brought the Lab so mention attention, and with it, financial backing (Brand 1987). On the other hand, computation would apparently ‘bring… out the artist in all of us’ (Brand 1987, 83). As Fred

Turner (2006) has pointed out, this was a vision steeped in a libertarian strain of the counterculture, whereby artist, industry, and individual form a kind of virtuous loop, somehow separate from politics. 'The Media Lab assumes that if it helps take care of the individual, computer-augmented individuals will take better care of the world,' Stewart Brand wrote (1987, 251), following several months there in 1986. 'The Lab would cure the pathologies of communications technology not with economics or politics but with technology.'

Such a cure required financing, of course, and by 1987 the Media Lab had fostered an annual budget of around $7 million a year, $6 million of which came from 'almost one hundred sponsors … each of whom had paid a minimum of two hundred thousand dollars to join' (Turner 2006, 178). These sponsors included IBM and DARPA, along with Apple, BBN, General Motors, Sony, Nippon Telephone and Telegraph, major newspaper and television companies, the National Science Foundation, and many others. The Music and Cognition group was sponsored by the System Development Foundation, the not-for-profit arm of a company that began life making software for the US Air Force, and has since been subsumed into L3Harris.[6] Rowe's doctoral work on *Cypher* was funded in part by Yamaha.

The book that came out of this work, *Interactive Music Systems: Machine Listening and Composing*, was published in 1993. The following year, the 'Music and Cognition Group' rebranded, and the Machine Listening Group was born. 'I have been unhappy with 'music (and) cognition' for some time,' Dan Ellis wrote in an email to Michael Casey; both recently arrived doctoral students at the Lab, with backgrounds in music as well as engineering.

> It's not even supposed to describe our group; it was the name of a larger entity … that was dissolved almost two years ago. But I've shied away from the issue for fear of something worse. I like Machine Listening a lot. I've also thought about Auditory

> Processing, and I try to get the second floor to describe my demos as Machine Audition. I'm not sure of the precise shades of connotation of the different words, except I'm pretty confident that having 'music' in the title has a big impact on people's preconceptions, one I'd rather overcome.[7]

So, what began for Rowe as a term to describe the so-called 'analytic layer' of an interactive music system was quickly repurposed as a way of pointing *beyond* music to a more encompassing form of auditory analysis or 'intelligence': one more concerned with information extraction than aesthetics, and with more obvious commercial applications as a result.

Sound/Scene

Auditory Scene Analysis

A very similar process had been playing out at Stanford's Center for Computer Research in Music and Acoustics (CCRMA) (Mody and Nelson 2013), where researchers were also beginning to move beyond 'musical problems' to 'more fundamental problems of audition' (CCRMA 1992, 25). In both cases, the re-orientation is attributed by researchers to psychologist Albert Bregman's work on 'auditory scene analysis' (Bregman 1984; 1990), though it is worth noting that the musical avant-garde had been treading a similar path for much of the 20th century: away from 'music', and towards a concern for sound and listening more generally. Nevertheless, Bernard Mont-Reynaud, then at Stanford, now principal scientist at SoundHound,[8] credits Bregman with provoking a new concern for 'the pure perception of sound' and 'the richness of the auditory domain per se' (CCRMA 1992, 25). Likewise for Ellis, who would go on to run Columbia's Laboratory for the Recognition and

Organization of Speech and Audio (LabROSA) for over a decade before moving to Google in 2015. Even though people had been 'using computers to process sound ever since the emergence of digital signal processing in the 1960s,' Ellis explains, 'it wasn't until Bregman's work in the mid-1980s 'that the idea of modelling some of the more abstract aspects of auditory processing emerged' (D. P. W. Ellis 1996, 18; G. J. Brown 1992; Hawley 1993; G. Brown and Cooke 1994).

In his key work, *Auditory Scene Analysis: The Perceptual Organization of Sound* (1990), Bregman himself puts it like this. 'If you were to pick up a general textbook on perception written before 1965 and leaf through it, you would not find any great concern with the perceptual or ecological questions about audition', he writes (Bregman 1990, 1). There would be plenty on 'such basic auditory qualities as loudness and pitch', and maybe also on the physiology of the ear and nervous system. But very little on how our auditory systems 'build a picture of the world around us through their sensitivity to sound' (perceptual), or on 'how our environment tends to create and shape the sound around us' (ecological). Imagine you are on the edge of a lake, Bregman says, and a friend challenges you to a (very strange) game.

> The game is this: Your friend digs two narrow channels up from the side of the lake. Each is a few feet long and a few inches wide and they are spaced a few feet apart. Halfway up each one, your friend stretches a handkerchief and fastens it to the sides of the channel. As waves reach the side of the lake they travel up the channels and cause the two handkerchiefs to go into motion. You are allowed to look only at the handkerchiefs and from their motions to answer a series of questions. How many boats are there on the lake and where are they? Which is the most powerful one? Which one is closer? Is the wind blowing? Has any large object been dropped suddenly into the lake? (Bregman 1990, 5-6)

‘Solving this problem seems impossible, but it is a strict analogy,’ Bregman claims, ‘to the problem faced by our auditory systems’. The lake is the air around us. The two channels are our ear canals, the handkerchiefs our ear drums, and the only information the auditory system has available to it, is their vibration. ‘Yet it seems to be able to answer questions very like the ones that were asked by the side of the lake: how many people are talking? Which one is louder, or closer? Is there a machine humming in the background?’ (Bregman 1990, 6)

This is what Bregman calls the ‘auditory scene analysis problem’ (Bregman 1990, 9),[9] appropriating both the language of ‘scene analysis’ and much of its intellectual scaffolding from machine vision (Guzmán 1968; Jacob Beck, Hope, and Rosenfeld 1986; Minsky 1975). The first step in addressing this problem, Bregman says, is the separation of sound into ‘streams’ (Bregman 1990; Rosenthal and Okuno 1998, x). For Bregman (1990, 11), the stream plays ‘the same role in auditory mental experience as the object does in visual.’ ‘The wind blows, an animal scurries through a clearing, the fire burns, a person calls.’ (Bregman 1990, 10) Sound conveys information about each of these physical ‘happenings’, which we somehow recognise as distinct. This perceptual unit Bregman calls the ‘auditory stream’. It is the first ‘computational stage’, he says, ‘on the way to the full description of an auditory event.’ (Bregman 1990, 10) Why not just call it a sound?

> First of all a physical happening (and correspondingly its mental representation) can incorporate more than one sound, just as a visual object can have more than one region. A series of footsteps, for instance, can form a single experienced event, despite the fact that each footstep is a separate sound. A soprano singing with a piano accompaniment is also heard as a coherent happening, despite being composed of distinct sounds (notes). Furthermore, the singer and piano together form a perceptual entity—the ‘performance’—that is distinct from other sounds that are occurring.

> Therefore, our mental representations of acoustic events can be multifold in a way that the mere word 'sound' does not suggest. (Bregman 1990, 10)

Each chapter of *Auditory Scene Analysis* (1990) is dedicated to a different set of stream segregation problems, with dedicated chapters on speech and music. Throughout the book, Bregman is clear that the purpose of all this isn't just understanding auditory perception, but also its computational reproduction. 'If the study of human audition were able to lay bare the principles that govern the human skill,' he writes, 'there is some hope that a computer could be designed to mimic it' (Bregman 1990, 3). Bregman laments the fact that, as a psychologist, he does not personally have the relevant skills in this respect. But he ends the book with the 'hope that some researchers who do have them will, after reading these chapters, accept the challenge of constructing an explicit model' (Bregman 1990, 704).

That is exactly what happened, though it is worth noting that the influence was not all one way. In addition to drawing on the conceptual resources of machine vision and Gestalt psychology, Bregman is particularly clear about his practical and intellectual debt to Peirce, Mathews, Mont-Reynaud, and John Chowning at the CCRMA, where he had spent time in 1982 before a full sabbatical year there across 1986 and 1987. Time spent soaking up what 'the computer music community… had discovered about the perception of musical sound' (Bregman 1990, *xii*), which—Bregman surmised—was surely 'governed, at least in part, by the primitive scene-analyzing principles' explored in the rest of the book (1990, 455).

Computational Auditory Scene Analysis

Bregman's work was received back into computer science across the 1990s as Computational Auditory Scene Analysis (CASA), the phrase having been coined by Guy Brown (1992), then a doctoral student at the University of Sheffield, before coming to name a series of workshops at the International Joint Conference on AI, in 1995, 1997 and 1999 (see Ellis

1996; Rosenthal and Okuno 1998; Wang and Brown 2006).[10] As Mont-Reynaud, Ellis, and others quickly understood, the implications of CASA for work on audio were very significant. What it opened the way for, in both principle and practice, was a dramatic expansion of the field: 'a radically new approach to extracting information from sound—one which, like the auditory systems of human and other animals, treats the organization of the received sound into features attributable to different sources as a central and indispensable aspect of the sound processing problem' (Ellis 2002, 1). Speech and music were now just particular kinds of sounds within a broader auditory 'scene', which computers might be trained to classify and discern in the same way humans perceive and are able to distinguish between 'a passing car, distant voices, the hum of a computer or music playing quietly in the background' (Wang and Brown 2006, 1).

At the first CASA workshop, organised by David Rosenthal from the Machine Listening Group and Hiroshi Okuno, then at Nippon Telegraph and Telephone, there were talks on environmental sound recognition, various kinds of automated music analysis, speaker/background differentiation, and speaker separation, including the launch of a 'multi-simultaneous-speaker' dataset on CD-ROM ("CASA Workshop Summary" 1996). Ellis shared some of his doctoral work on 'city sound scene organisation'. Using a '"traditional AI" framework of rules, representation and symbolic processing' (Ellis 1996, 160), the idea was to be able to divide recordings of, say, a construction scene, containing 'a diverse mixture of machinery, knocking, hammering and voices' (Ellis 1996, 137), or '"train station ambience", consisting of babble, footsteps etc. in a highly reverberant environment' (Ellis 1996, 142), into components corresponding to 'individual sound-producing events that a listener would identify in the sound' (Ellis 1996, 11).

Researchers in and around CASA would go on to develop machine listening systems for all manner of sounds, scenes, and classification tasks, and by all manner of techniques: the relative merits of 'perceptual' (ie rule-based) and statistical methods being an organising concern for the best part of the next twenty years ("SAPA Workshops - Index" n.d.). There was work, for instance, on music information retrieval (Goto and Muraoka 1998; Goto and Hayamizu 1999; S. Downie and Nelson 2000; J. S. Downie 2003), music recognition (Wang 2003), music detection (Hawley 1993), laughter detection (Wold et al. 1996; Kennedy and Ellis 2004), alarm sound detection (Ellis 2001), audio life-logging (Doherty et al. 2007), automatic soundtrack tagging (Lee and Ellis 2010) and the recognition and categorisation of animal vocalisations: especially for marine mammals and birds (Mellinger and Clark 2000; Härmä 2003; Brown, Hodgins-Davis, and Miller 2006; Halkias and Ellis 2006; Brown and Miller 2007).

Throughout this period of diversification, however, speech recognition remained a constant backdrop, funder, and alibi. Despite significant advances across the 1990s, even the most sophisticated systems struggled to distinguish speech from 'irrelevant intrusions', including by other speakers (Ellis 1996, 14). Speech recognition *needed* CASA, because a 'sense of hearing that ceases to work when more than one sound source is active has little or no practical use' (Ellis 2002, 1). In fact, the problem of speech recognition's efficacy in 'unconstrained environments' (Auditory 1995)—outside of lab conditions—had been part of CASA's sales pitch from the start (Bregman 1990, 3). IBM, remember, had launched its first commercial transcription software in 1986, and Bell's 'Voice Recognition Call Processing' had arrived in 1992. The 90s was not only a period of unprecedented growth in ASR's capabilities, but also in the diversity of environments in which it was being asked to operate: in offices, first, then homes, and elsewhere. So, when Rosenthal and Okuno published the proceedings of the first CASA workshop in 1998, speech was where they chose to begin.

> A fundamental characteristic of human hearing is the ability selectively to attend to sound originating from a particular physical source, in an environment that includes a variety of sources. In contrast, most computer sound-understanding systems are structured to regard the world as consisting of "signal" (usually speech) and "noise" (everything else). To operate effectively as listeners in the real world, computers will have to acquire a more flexible and sophisticated view of what constitutes signal, and what constitutes interference. (Rosenthal and Okuno 1998, *ix*)

As a result, researchers in CASA were attempting, they said, 'to integrate speech and non-speech recognition and understanding systems into a common framework' (Rosenthal and Okuno 1998, *ix*). Such a framework has not been forthcoming. But the desire being expressed is instructive. CASA's emergence marks an important threshold in machine listening's history. Perhaps even the moment of its arrival as a relatively coherent scientific and sociotechnical project: to model, automate, and surpass human listening *per se*, rather than just some part of it.

Conclusion

Fifteen years later, in 2013, researchers from IRCAM, City University London, and the Centre for Digital Music at Queen Mary University, partnered to launch the IEEE Audio and Acoustic Signal Processing Technical Committee challenge on Detection and Classification of Acoustic Scenes and Events (DCASE). The challenge was intended to 'frame a set of general-purpose machine listening tasks for everyday audio, in order to benchmark the state of the art, stimulate further work, and grow the research community in machine listening beyond the domains of speech and music' (Stowell et al. 2015, 2). Successful systems would need to distinguish, first, between recordings of a range of auditory scenes (home, bus, office,

park, supermarket, restaurant) and then between more specific sounds within them (doors knocking, keyboards clicking, phones ringing, coughing, speech, laughter). For the scene classification task, the leading systems apparently 'attained results significantly above baseline and comparable to average results from human listeners'. In the event detection task, however, there was much more 'room for improvement' (Stowell et al. 2015, 27).

Since 2013, DCASE has grown considerably. Attendance has gone from sixty-eight to more than five hundred. The challenges have been refined and expanded to include tasks on 'machine condition monitoring', 'sound event localization', 'automated audio captioning', 'bioacoustic event detection', and 'sound event detection for smart cars'. And they now get hundreds of entrants, mainly from the US, UK, China, Japan, and Europe (DCASE 2021). Dozens of papers are published in the DCASE workshop proceedings every year (Font, Frederic et al. 2021). Sponsors include Apple, Google, Facebook, Bose, Dolby, Adobe, Mitsubishi, Hitachi, and Audio Analytic. Industry-based researchers sit on the organizing committee of most challenge tasks and are invited to give keynotes.

Unsurprisingly given this level of interest and investment, challenge results have improved markedly. This is largely down to the rise of 'deep' neural networks and other data-driven methods following the 'statistical turn', which have made such an impact across machine learning more generally. Whereas at DCASE 2013 there were no neural nets at all, by 2016 they were already dominant (Mesaros, Heittola, and Virtanen 2016, 11). Because neural nets need data, and because Google has acquired a lot of it, AudioSet has proven particularly influential in this respect. Both the dataset of 2.1 million YouTube videos and its accompanying ontology of audio event classes have appeared regularly in DCASE challenges since 2018, whether for annotation, training, evaluation, or—more recently—the synthesis of new data (DESED 2020). Otherwise, datasets have come from elsewhere in industry (eg

Hitachi) and from university labs with EU funding.[11] But none of these even approach AudioSet in scale. Machine listening's future will increasingly be defined by what Orla Lynskey (2019) has called 'data power'.

This future had yet to be written in the 1990s—when 'machine listening' referred to the analytic layer of an 'interactive music system' and researchers in CASA started making 'auditory scenes' an object of computational concern—even if some of the groundwork for the 'statistical turn' had been laid in speech recognition at IBM well before. One reason for telling machine listening's history the way we have is to draw out parts of the story that a presentist focus on 'big data', the tech giants of today, and what they have done to or with capitalism, can tend to occlude. Google had yet to be founded when researchers gathered in Montreal for the first workshop on CASA. Amazon was still an online bookstore. The iPhone was more than a decade away. Siri and Alexa closer to two.

Machine listening's early history was dominated by the US military, industrial giants of the twentieth century like Bell telephone and IBM, and their university partners, including at centers and labs focused on music. For a long time, the field's major commercial driver was speech recognition, and specifically the automation of feminized labour like call-routing and office transcription. But this same concern was also crucial to the field's expansion beyond speech when these systems were finally released 'into the wild' and forced to confront speech's embeddedness in diverse sonic environments as a result. Across this same period, and in dialogue with researchers in speech, musician-engineers were busy developing 'intelligent' listening systems for use in transcription, composition, and interactive performance. And it was the composer Robert Rowe who began using the phrase 'machine listening' to describe his own contributions in this respect. Within a few years, the term would be taken up by others at MIT to name a new research group oriented towards more

general—indeed ‘fundamental’—questions of audition, inspired by Albert Bregman’s work on ‘auditory scene analysis’, which was influenced in turn by time spent with the ‘computer music community’ at Stanford’s CCRMA. Together, these and other similar groups developed a new vein of research concerned with the computational perception, organization, separation, and identification of sounds beyond speech and music. This new orientation towards the entire soundscape is what marks machine listening’s arrival out of the maelstrom of computer science, music, commercial, and state imaginaries and interests across the second half of the 20th century. At the same time, it clarifies the challenge machine listening presents politically. With no sound or combination of sounds off limits to computational systems and their makers, with ubiquitous microphones, rapid investment, and an ambition to install ‘intelligent sound recognition everywhere’, the field of contestation is likewise ‘all possible sounds’.

Acknowledgments
This essay comes out of a group project led by Sean Dockray, James Parker, and Joel Stern. So far, this project has involved the production and curation of writing, interviews, events, instruments, music, artworks, and a library, all concerned with machine listening, or some part of it. Because the project is a collective one, it is not always easy to attribute authorship. In this instance, most of the words were written by James, who is therefore responsible for many of the essay's idiosyncrasies and omissions. But they take their lead from Sean's artwork *Learning from YouTube* (2018), which is also where the essay begins, and Sean was a frequent interlocutor, commentator, researcher, and editor across the essay's many drafts. For more information on the larger project, see https://machinelistening.exposed/ We would also like to thank Joel, for his contributions, Zoë de Luca for her research assistance, Jake Goldenfein, and Thao Phan for their detailed comments, anonymous reviewers, and the many participants in the machine listening curriculum to date.

Disclosure statement
This research was funded in part by Australia Research Council Discovery Early Career Award DE200101447 'The Laws and Politics of Machine Listening'.

Notes on Contributors
James Parker is an Associate Professor and ARC DECRA fellow at Melbourne Law School who works across legal scholarship, art criticism, curation, and production. He is author of *Acoustic Jurisprudence: Listening to the Trial of Simon Bikindi* (OUP 2015) and co-curator of *Eavesdropping*, an exhibition and extensive public program staged at the Ian Potter Museum of Art in 2018 and City Gallery, Wellington in 2019. His current work, with Sean Dockray and Joel Stern, is on machine listening.

Sean Dockray is an artist, writer, and programmer living in Melbourne whose work explores the politics of technology, with a particular emphasis on artificial intelligences and the algorithmic web. He is also the founding director of the Los Angeles non-profit Telic Arts Exchange, and initiator of the knowledge-sharing platforms, The Public School and AAAARG.ORG. He is Senior Lecturer in the Department of Fine Art at Monash University. With James Parker and Joel Stern, Dockray is co-founder of the research project Machine Listening.

Notes

[1] Though see, for instance, Facebook's recent Generative Spoken Language Modelling, which learns the acoustic and linguistic characteristics of a language from raw audio (Lakhotia et al. 2021).
[2] Hearsay II reported 74 percent accuracy, whereas Hear What I Mean achieved only 44 percent.
[3] The two would go on to edit a book together (Mathews and Pierce 1989).
[4] The milieu cited by Rowe is exclusively male, and with the important exception of Lewis, all white. (Rowe 1993, 78-93). To its detriment, computer music was, and continues to be, white and male dominated. Key exceptions from the field's early days include, for instance,

Pauline Oliveros, Delia Derbyshire, Daphne Oram, Wendy Carlos, Maryanne Amacher, and Laurie Spiegel. See *Sisters with Transistors* (Rovner 2020).

[5] Minsky was dismissive of both 'post-Weberian serialism' and expanded conceptions of music introduced by John Cage (D. Kahn and Minsky 1988).

[6] https://oac.cdlib.org/findaid/ark:/13030/tf429003m4/admin/#scopecontent-1.8.4

[7] Archived email exchange between Dan Ellis and Michael Casey 28 March 1994. According to Casey, 'Dan suggested "Machine Audition," to which I responded that the term 'audition' was not widely used outside of hearing sciences and medicine, and that it could be a confusing name for a group that was known for working on music–think "music audition." I believe we discussed the word hearing, but I – we? – thought it implied passivity as in "hearing aid," and instead I suggested the name "machine listening" because it had connotations of attention and intelligence, concepts that were of interest to us all at that time. That is what I remember.' On file with authors.

[8] SoundHound may not be a household name, but it is a substantial company, especially from the perspective of machine listening. The company began life as a music recognition platform called Midomi, based on audio-fingerprinting techniques developed in and around CASA. After rebranding as Soundhound, it moved into voice interactivity, where it is now an industry leader. In November 2021 it went public via a SPAC deal for just over US$2 billion (Reuters 2021)

[9] It is more commonly known as the 'cocktail party problem', following a paper by E.C. Cherry (1953).

[10] Calls for Papers are available at 1995 https://web.archive.org/web/19961130112511/http:/sound.media.mit.edu/~dfr/CASA.html; 1997 http://www.auditory.org/postings/1996/203.html; 1999 http://www.auditory.org/mhonarc/1999/msg00046.html

[11] See, for instance, Clotho Audio Captioning Dataset 2019; TAU Urban Acoustic Scenes 2020 Mobile Dataset; TAU-NIGENS Spatial Sound Events 2021 Dataset.

References

ABC News. 2020. “Apple Wants to Listen to You Wash Your Hands, and Help You Lose Your Car Keys.” *ABC News*, June 23, 2020. https://www.abc.net.au/news/2020-06-23/apple-wwdc-2020-ios-14-digital-key-and-hand-washing-arm-chips/12383124.

AI Now Institute. 2018. “AI Now Report 2018.”

Amazon, dir. 2019. *Acoustic Event Detection with Alexa Guard*. https://www.youtube.com/watch?v=-nKelNVVblM.

Anderson, Chris. 2008. “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.” *Wired*, June 23, 2008. https://www.wired.com/2008/06/pb-theory/.

Andrejevic, Mark. 2004. *Reality TV: The Work of Being Watched*. Critical Media Studies. Lanham, Md: Rowman & Littlefield Publishers.

———. 2020a. *Automated Media*. London; New York, NY: Routledge.

———. 2020b. “Data Civics: A Response to the ‘Ethical Turn.’” *Television & New Media* 21 (6): 562–67. https://doi.org/10.1177/1527476420919693.

Audio Analytic. 2018. “Audio Analytic.” Audio Analytic. 2018. https://www.audioanalytic.com/.

Auditory. 1995. “[2nd CFP] IJCAI-95 Workshop on CASA.” 1995. http://www.auditory.org/mhonarc/1995/msg00006.html.

Barthes, Roland. 1991. *The Responsibility of Forms: Critical Essays on Music, Art, and Representation*. Translated by Richard Howard. Reprint edition. Berkeley: University of California Press.

Beck, Jacob, Barbara Hope, and Azriel Rosenfeld. 1986. *Human and Machine Vision*. Vol. 8. Academic Press.

Beck, John, and Ryan Bishop. 2020. *Technocrats of the Imagination: Art, Technology, and the Military-Industrial Avant-Garde*. *Technocrats of the Imagination*. Duke University Press. https://doi.org/10.1515/9781478007326.

Beck, Ulrich. 2009. *World at Risk*. Cambridge: Polity Press.

Bell, Eamonn. 2019. “The Computational Attitude in Music Theory.” Columbia University.

Bijsterveld, Karin. 2008. *Mechanical Sound: Technology, Culture, and Public Problems of Noise in the Twentieth Century*. Inside Technology. Cambridge, Mass: MIT Press.

Bijvoet, Marga. 1990. “How Intimate Can Art and Technology Really Be? A Survey of the Art and Technology Movement of the Sixties.” *P. Hayward (Author), Culture, Technology & Creativity: In the Late Twentieth Century*, 15–38.

Born, Georgina. 1995. *Rationalizing Culture: IRCAM, Boulez, and the Institutionalization of the Musical Avant-Garde*. Univ of California Press.

Brand, Stewart. 1987. *The Media Lab: Inventing the Future at MIT*. Viking.

Bregman, Albert S. 1984. “Auditory Scene Analysis.” In *IEEE Conference on Pattern Recognition*, 168–75.

———. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, Mass: MIT Press.

Brewster, Ben. 1972. “Introduction to Marx’s ‘Notes on Machines.’” *Economy and Society* 1 (3): 235–43. https://doi.org/10.1080/03085147200000013.

Brown, Guy, and Martin Cooke. 1994. “Computational Auditory Scene Analysis.” *Compter Speech and Language* 8: 297–336.

Brown, Guy J. 1992. “Computational Auditory Scene Analysis: A Representational Approach.” *The Journal of the Acoustical Society of America* 94 (4): 2454–2454. https://doi.org/10.1121/1.407441.

Brown, Judith C., Andrea Hodgins-Davis, and Patrick J. O. Miller. 2006. “Classification of Vocalizations of Killer Whales Using Dynamic Time Warping.” *The Journal of the Acoustical Society of America* 119 (3): EL34–40. https://doi.org/10.1121/1.2166949.

Brown, Judith C., and Patrick J. O. Miller. 2007. “Automatic Classification of Killer Whale Vocalizations Using Dynamic Time Warping.” *The Journal of the Acoustical Society of America* 122 (2): 1201–7. https://doi.org/10.1121/1.2747198.

“CASA Workshop Summary.” 1996. November 30, 1996. https://web.archive.org/web/19961130120944/http:/sound.media.mit.edu/~dfr/casa/summary.html.

CCRMA. 1992. “Center for Computer Research in Music and Acoustics: Research Overview.”

Cherry, E.C. 1953. “Some Experiments on the Recognition of Speech with One and with Two Ears.” *Journal of the Acoustical Society of America* 25: 975–79.

Couldry, Nick, and Ulises Mejias. 2018. “Data Colonialism: Rethinking Big Data’s Relation to the Contemporary Subject.” *Television and New Media*, 1–14.

Crawford, Kate, and Vladan Joler. 2018. “Anatomy of an AI System.” Anatomy of an AI System. 2018. http://www.anatomyof.ai.

Crawford, Kate, and Trevor Paglen. 2019. “Excavating AI: The Politics of Training Sets for Machine Learning.” -. 2019. https://excavating.ai.

David, E., and O. Selfridge. 1962. “Eyes and Ears for Computers.” *Proceedings of the IRE* 50 (5): 1093–1101. https://doi.org/10.1109/JRPROC.1962.288011.

DCASE. 2021. “DCASE 2021 Workshop Statistics.” 2021. https://dcase.community/documents/workshop2021/dcase2021_statistics.pdf.

Dean, Roger T. 2009. *The Oxford Handbook of Computer Music*. Oxford University Press.

Denes, Peter. 1960. “Automatic Speech Recognition: Experiments with a Recogniser Using Lingustic Statistics.” Contract No. AF 61(514)-1176. Air Force Cambridge Research Center: United States Air Force Air Research and Development Command.

DESED. 2020. “Domestic Environment Sound Event Detection Dataset.” 2020. https://project.inria.fr/desed/.

Diduck, Ryan. 2018. *Mad Skills: MIDI and Music Technology in the Twentieth Century*. Watkins Media Limited.

Dockray, Sean. 2018. *Learning from YouTube*. Video essay.

Doherty, Aiden R, Alan F Smeaton, Keansub Lee, and Daniel P W Ellis. 2007. “Multimodal Segmentation of Lifelog Data.” In *Proc. 8th Int. Conf. on Computer-Assisted Information Retrieval RIAO 2007, Pittsburgh, May 2007.*, 18.

Downie, J Stephen. 2003. “Music Information Retrieval.” In *Annual Review of Information Science and Technology*, edited by Blaise Cronin, 295–340.

Downie, Stephen, and Michael Nelson. 2000. “Evaluation of a Simple and Effective Music Information Retrieval Method.” In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR ’00*, 73–80. Athens, Greece: ACM Press. https://doi.org/10.1145/345508.345551.

Dyson, Frances. 2006. “Frances Dyson, And Then It Was Now : Enduring Rhetorics.” Edited by Clarisse Bardiot. *Fondation Langlois*. https://www.fondation-langlois.org/html/e/page.php?NumPage=2144.

Edwards, Paul N. 1997. *The Closed World: Computers and the Politics of Discourse in Cold War America*. MIT Press.

Ellis, Dan. 2018. “Recognizing Sound Events.” John Hopkins: Center for Language and Speech Processing, October 4.

https://jh.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=4a7e392c-5163-41a6-8229-aadc01099e63.
Ellis, Daniel P W. 1996. “Prediction-Driven Computational Auditory Scene Analysis.” MIT.
———. 2001. “DETECTING ALARM SOUNDS.” In , 4. https://www.ee.columbia.edu/~dpwe/pubs/crac01-alarms.pdf.
———. 2002. “The Listening Machine: Sound Source Organization for Multimedia Understanding.” *Electrical Engineering*, 19.
Ernst, Wolfgang. 2021. “The Media Epistemic Value of Sonic Analytics Tools. A Commentary.” *Internet Histories* 5 (1): 48–56. https://doi.org/10.1080/24701475.2020.1862528.
Estabrooks, Maurice. 1995. *Electronic Technology, Corporate Strategy, and World Transformation*. Greenwood Publishing Group.
Feldman, Jessica. 2016. “‘The Problem of the Adjective’: Affective Computing of the Speaking Voice.” *Transposition*, no. 6 (December). https://doi.org/10.4000/transposition.1640.
Font, Frederic, Mesaros, Annamaria, P. W. Ellis, Daniel, Fonseca, Eduardo, Fuentes, Magdalena, and Elizalde, Benjamin. 2021. “Proceedings of the 6th Workshop on Detection and Classication of Acoustic Scenes and Events (DCASE 2021),” November. https://doi.org/10.5281/ZENODO.5770113.
Friedland, Gerard, Paris Smaragdis, Josh McDermott, and Raj Bhisha. 2018. “Audition for Multimedia Computing.” In *Frontiers of Multimedia Research*, edited by Shih-Fu Chang, 416. Association for Computing Machinery.
Fry, D. B., and P. Denes. 1958. “The Solution of Some Fundamental Problems in Mechanical Speech Recogntion.” *Language and Speech* 1 (1). https://doi.org/doi:10.1177/002383095800100104.
Fuchs, Christian, and Vincent Mosco, eds. 2016. *Marx in the Age of Digital Capitalism*. Studies in Critical Social Sciences 80. Leiden; Boston: Brill.
Gemmeke, Jort F., Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. “Audio Set: An Ontology and Human-Labeled Dataset for Audio Events.” In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference On*, 776–80. IEEE.
Goldenfein, Jake. 2019. “The Profiling Potential of Computer Vision.” *Association for Computing Machinery*, 27.
Goto, Masataka, and Satoru Hayamizu. 1999. “A Real-Time Music Scene Description System: Detecting Melody and Bass Lines in Audio Signals.” In *IJCAI-99 Workshop on Computational Auditory Scene Analysis*, 10.
Goto, Masataka, and Yoichi Muraoka. 1998. “An Audio-Based Real-Time Beat Tracking System and Its Applications.” In *Proceedings of International Computer Music Conference*.
Guzmán, Adolfo. 1968. “Decomposition of a Visual Scene into Three-Dimensional Bodies.” In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*, 291–304.
Halkias, Xanadu C., and Daniel P.W. Ellis. 2006. “Call Detection and Extraction Using Bayesian Inference.” *Applied Acoustics* 67 (11–12): 1164–74. https://doi.org/10.1016/j.apacoust.2006.05.006.
Harcourt, Bernard E. 2015. *Exposed: Desire and Disobedience in the Digital Age*. Harvard University Press.
Härmä, Aki. 2003. “Automatic Identification of Bird Species Based on Sinusoidal Modeling of Syllables.” In *ICASSP*. https://doi.org/10.1109/ICASSP.2003.1200027.
Hawley, Michael. 1993. “Structure of Sound.” MIT.

House, Brian. 2017. “MACHINE LISTENING: WAVENET, MEDIA MATERIAL- ISM, AND RHYTHMANALYSIS” 6 (1): 9.
Hurlbut, J Benjamin. 2018. “Control without Limits in the New Biology.” In *Gene Editing, Law, and the Environment: Life Beyond the Human*, edited by Irus Braverman, 77–94. New York: Routledge.
Hvistendahl, Mara. 2020. “How a Chinese AI Giant Made Chatting—and Surveillance—Easy.” *Wired*, 2020. https://www.wired.com/story/iflytek-china-ai-giant-voice-chatting-surveillance/.
Jasanoff, Sheila, and Sang-Hyun Kim. 2015. *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*. University of Chicago Press.
Jones, Nicholaus A. 2018. “Listening to the Frontend: United States Patent US 10,020,004 B2,” 14.
Kahn, Douglas, and Marvin Minsky. 1988. “Minsky and Artificial Intelligence.” *EAR*, 1988.
Kahn, Jonathan. 2013. *Race in a Bottle: The Story of BiDil and Racialized Medicine in a Post-Genomic Age*. New York: Columbia University Press.
Kang, Edward B. 2023. “Ground Truth Tracings (GTT): On the Epistemic Limits of Machine Learning.” *Big Data & Society* 10 (1): 205395172211461. https://doi.org/10.1177/20539517221146122.
Kang, Edward B., and Simogne Hudson. 2022. “Audible Crime Scenes: ShotSpotter as Diagnostic, Policing, and Space-Making Infrastructure.” *Science, Technology, & Human Values*, December, 016224392211432. https://doi.org/10.1177/01622439221143217.
Kennedy, Lyndon S, and Daniel P W Ellis. 2004. “LAUGHTER DETECTION IN MEETINGS.” In *NIST ICASSP 2004 Meeting Recognition Workshop, Montreal*. National Institute of Standards and Technology.
Lakhotia, Kushal, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, et al. 2021. “Generative Spoken Language Modeling from Raw Audio.” *ArXiv:2102.01192 [Cs]*, September. http://arxiv.org/abs/2102.01192.
Lawrence, H.M. 2019. “Siri Disciplines.” In *Your Computer Is on Fire*. MIT Press.
Lee, Keansub, and Daniel P. W. Ellis. 2010. “Audio-Based Semantic Concept Classification for Consumer Video.” *IEEE Transactions on Audio, Speech, and Language Processing* 18 (6): 1406–16. https://doi.org/10.1109/TASL.2009.2034776.
Lewis, George. 2000. “Too Many Notes.” *Leonardo Music Journal* 10: 33–39.
———. 2018. “Technosphere Magazine: 5. Rainbow Family.” Technosphere Magazine. 2018. /p/5-Rainbow-Family-5Aj9nAxzG6zFRAAd9icEvH.
Lewis, George E. 2018. *Why Do We Want Our Computers to Improvise?* Edited by Roger T. Dean and Alex McLean. Vol. 1. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190226992.013.29.
Li, Xiaochang. 2017. “Divination Engines: A Media History of Text Prediction.” New York: New York Univesrity.
Li, Xiaochang, and Mara Mills. 2019. “Vocal Features: From Voice Identification to Speech Recognition by Machine.” *Technology and Culture* 60 (2S): S129–60. https://doi.org/10.1353/tech.2019.0066.
Licklider, J.C.R. 1960. “Man-Computer Symbiosis.” *IRE Transactions on Human Factors in Electronics*, no. March: 4–10.
Lynskey, Orla. 2019. “Grappling with ‘Data Power’: Normative Nudges from Data Protection and Privacy.” *Theoretical Inquiries in Law* 20 (1): 189–220. https://doi.org/10.1515/til-2019-0007.
Lyon, Richard F. 1978. “Sig_Proc_Model_of_Hearing-Lyon1978.Pdf.”

———. 2010. “Machine Hearing: An Emerging Field.” *IEEE Signal Processing Magazine*, September, 6. https://doi.org/10.1109/MSP.2010.937498.
———. 2017. *Human and Machine Hearing: Extracting Meaning from Sound*. Cambridge University Press. https://doi.org/10.1017/9781139051699.
“Machine Listening Lab.” 2018. Machine Listening Lab. 2018. http://machine-listening.eecs.qmul.ac.uk/.
MacKenzie, Donald. 1984. “Marx and the Machine.” *Technology and Culture* 25: 473–502.
Maier, Stefan. 2018. “Technosphere Magazine: 1. WaveNet: On Machine and Machinic Listening.” Technosphere Magazine. 2018. /p/1-WaveNet-On-Machine-and-Machinic-Listening-a2mD8xYCxtsLqoaAnTGUbn.
Malkin, Robert G. 2006. “Machine Listening for Context-Aware Computing.”
Marx, Karl. 1976. *Capital: A Critique of Political Economy Vol. 1*. Translated by Ben Fowkes. Penguin.
Mathews, Max, and John R Pierce, eds. 1989. *Current Directions in Computer Music Research*. MIT Press.
Mattern, Shannon. 2020. “Urban Auscultation; or, Perceiving the Action of the Heart.” *Places Journal*, April. https://doi.org/10.22269/200428.
McQuillan, Dan. 2018a. “Mental Health and Artificial Intelligence: Losing Your Voice | OpenDemocracy.” 2018. https://www.opendemocracy.net/en/digitaliberties/mental-health-and-artificial-intelligence-losing-your-voice-poem/.
———. 2018b. “Data Science as Machinic Neoplatonism.” *Philosophy & Technology* 31 (2): 253–72. https://doi.org/10.1007/s13347-017-0273-3.
Mellinger, David K., and Christopher W. Clark. 2000. “Recognizing Transient Low-Frequency Whale Sounds by Spectrogram Correlation.” *The Journal of the Acoustical Society of America* 107 (6): 3518–29. https://doi.org/10.1121/1.429434.
Mesaros, Annamaria, Toni Heittola, and Tuomas Virtanen. 2016. “TUT Database for Acoustic Scene Classification and Sound Event Detection.” In *2016 24th European Signal Processing Conference (EUSIPCO)*, 1128–32. Budapest, Hungary: IEEE. https://doi.org/10.1109/EUSIPCO.2016.7760424.
Mills, M. 2011. “On Disability and Cybernetics: Helen Keller, Norbert Wiener, and the Hearing Glove.” *Differences* 22 (2–3): 74–111. https://doi.org/10.1215/10407391-1428852.
Mills, Mara. 2010. “Deaf Jam.” *Social Text* 28 (1): 35–58. https://doi.org/10.1215/01642472-2009-059.
———. 2011a. “Hearing Aids and the History of Electronics Miniaturization.” *IEEE Annals of the History of Computing* 33 (2): 24–45.
———. 2011b. “Do Signals Have Politics? Inscribing Abilities in Cochlear Implants.” *The Oxford Handbook of Sound Studies*, December. https://doi.org/10.1093/oxfordhb/9780195388947.013.0077.
Minsky, Marvin. 1975. “A Framework for Representing Knowledge.” In *The Psychology of Computer Vision*, edited by P H Winston. McGraw Hill.
———. 1982. “Music, Mind, and Meaning.” In *Music, Mind, and Brain*, edited by M. Clynes, 19.
———. 1986. *The Society of Mind*. New York: Simon and Schuster.
Mody, Cyrus C. M., and Andrew J. Nelson. 2013. “‘A Towering Virtue of Necessity’: Interdisciplinarity and the Rise of Computer Music at Vietnam-Era Stanford.” *Osiris* 28 (1): 254–77. https://doi.org/10.1086/671380.
Moorer, J A. 1975. “On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer.” Stanford.

Mumford, Lewis. 1966. *The Myth of the Machine: Technics and Human Development*. Vol. 1. 2 vols. New York: Harcourt.
Negroponte, Nicolas. 1970. *The Architecture Machine: Toward a More Human Environment*. MIT Press.
Nest. 2021. "Nest Cam Indoor." Nest Cams. 2021. https://www.nestcamera.net/nest-cam-indoor/.
Parker, James E. K., and Lawrence Abu Hamdan. 2022. "Forensic Listening as Machine Listening." *Disclaimer*. https://disclaimer.org.au/contents/forensic-listening-as-machine-listening.
Parker, James E. K., Joel Stern, and Sean Dockray. 2020. "Machine Listening, a Curriculum." Machine Listening Curriculum. 2020. https://machinelistening.exposed/curriculum/.
Pfeifer, Michelle. 2021. "Listening for the Border - Affective Objectivity and Border Sonics (Unpublished Draft)." NYU.
Phan, Thao. 2019. "Amazon Echo and the Aesthetics of Whiteness." *Catalyst: Feminism, Theory, Technoscience* 5 (1): 1–38. https://doi.org/10.28968/cftt.v5i1.29586.
Pieraccini, Roberto. 2012. *The Voice in the Machine: Building Computers That Understand Speech*. Cambridge, Mass: MIT Press.
———. 2021. *AI Assistants*. MIT Press.
Pierce, J. R., and Mary E. Shannon. 1949. "Composing Music by a Stochastic Process." *Bell Telephone Laboratories, Technical Memorandum MM-49-150-29*.
Pierce, John R. 1969. "Whither Speech Recognition?" *The Journal of the Acoustical Society of America* 46 (4B): 1049–51.
Reuters. 2021. "Voice Assistant Maker SoundHound to Go Public via $2 Bln SPAC Deal." *Reuters*, November 16, 2021, sec. Technology. https://www.reuters.com/technology/voice-ai-platform-soundhound-go-public-via-21-bln-spac-merger-2021-11-16/.
Rice, Tom. 2015. "Hearing." In *Keywords in Sound*, edited by David Novak and Matt Sakakeeny. Durham, N.C.: Duke University Press.
Rosenthal, David F., and Hiroshi G. Okuno, eds. 1998. *Computational Auditory Scene Analysis*. Computational Auditory Scene Analysis. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
Rovner, Lisa, dir. 2020. *Sisters with Transistors*. https://sisterswithtransistors.com/.
Rowe, Robert. 1991. "Machine Listening and Composing: Making Sense of Music with Cooperating Real-Time Agents." MIT.
———. 1992. "Machine Listening and Composing with Cypher." *Computer Music Journal* 16 (1): 43. https://doi.org/10.2307/3680494.
———. 1993. *Interactive Music Systems: Machine Listening and Composing*. MIT Press.
———. 1999. "The Aesthetics of Interactive Music Systems." *Contemporary Music Review* 18 (3): 83–87. https://doi.org/10.1080/07494469900640361.
"SAPA Workshops - Index." n.d. Accessed May 6, 2022. https://www.sapaworkshops.org/.
Schroeder, Manfred Robert. 1985. *Speech and Speaker Recognition*. Vol. 12. Karger Medical and Scientific Publishers.
Schuller, Björn W. 2014. *Intelligent Audio Analysis*. Springer Science & Business Media.
Schuller, Björn W., Alican Akman, Yi Chang, Harry Coppock, Alexander Gebhard, Alexander Kathan, Esther Rituerto-González, Andreas Triantafyllopoulos, and Florian B. Pokorny. 2022. "Climate Change & Computer Audition: A Call to Action and Overview on Audio Intelligence to Help Save the Planet." *ArXiv:2203.06064 [Cs]*, March. http://arxiv.org/abs/2203.06064.

Schuller, Björn W., Dagmar M. Schuller, Kun Qian, Juan Liu, Huaiyuan Zheng, and Xiao Li. 2020. “COVID-19 and Computer Audition: An Overview on What Speech & Sound Analysis Could Contribute in the SARS-CoV-2 Corona Crisis.” *ArXiv:2003.11117 [Cs, Eess]*, March. http://arxiv.org/abs/2003.11117.

Scott, Alan. 1997. “Modernity’s Machine Metaphor.” *The British Journal of Sociology* 48 (4): 561. https://doi.org/10.2307/591596.

Semel, Beth Michelle. 2019. “Speech, Signal, Symptom: Machine Listening and the Remaking of Psychiatric Assessment.” MIT.

Shotspotter. 2022. “Shotspotter: Cities.” ShotSpotter. 2022. https://www.shotspotter.com/cities/.

Snell, John. 1977. “EDITORIAL INTRODUCTION.” *Computer Music Journal* 1 (1): 2.

———. 2006. “How Did" Computer Music Journal" Come to Exist?” *Computer Music Journal* 30 (1): 10–20.

Steinbeck, Paul. 2018. “George Lewis’s Voyager.” In *The Routledge Companion to Jazz Studies*, edited by Nicholas Gebhardt, Nichole Rustin-Paschal, and Tony Whyton, 1st ed., 261–70. Routledge. https://doi.org/10.4324/9781315315805-25.

Sterne, Jonathan. 2003. *The Audible Past: Cultural Origins of Sound Reproduction*. Duke University Press.

———. 2012. *The Sound Studies Reader*. New York: Routledge.

———. 2022. “Is Machine Listening Listening?” *Communication +1* 9: 5.

Sterne, Jonathan, and Elena Razlogova. 2021. “Tuning Sound for Infrastructures: Artificial Intelligence, Automation, and the Cultural Politics of Audio Mastering.” *Cultural Studies*, March, 1–21. https://doi.org/10.1080/09502386.2021.1895247.

Sterne, Jonathan, and Mehak Sawhney. 2022. “The Acousmatic Question and the Will to Datafy” 9 (2).

Stowell, Dan, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley. 2015. “Detection and Classification of Acoustic Scenes and Events.” *IEEE Transactions on Multimedia* 17 (10): 1733–46. https://doi.org/10.1109/TMM.2015.2428998.

Szendy, Peter. 2017. *All Ears : The Aesthetics of Espionage*. Books at JSTOR Demand Driven Acquisitions. New York, NY : Fordham University Press, 2017.

Turner, Fred. 2006. *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism*. Chicago: University of Chicago Press.

———. 2008. “Romantic Automatism: Art, Technology, and Collaborative Labor in Cold War America.” *Journal of Visual Culture* 7 (1): 5–26. https://doi.org/10.1177/1470412907087201.

Vercoe, Barry. 1984. “The Synthetic Performer in The Context of Live Performance.” *International Computer Music Conference Proceedings* 1984. http://hdl.handle.net/2027/spo.bbp2372.1984.026.

———. 1990. “Synthetic Listeners and Synthetic Performers.” In *Proceedings, International Symposium on Multimedia Technology and Artificial Intelligence (Computerworld 90), Kobe Japan*, 136–41.

Vetter, Grant. 2012. *The Architecture of Control: A Contribution to the Critique of the Science of Apparatuses*. John Hunt Publishing.

Wang, Avery Li-Chun. 2003. “An Industrial-Strength Audio Search Algorithm.” In , 7.

Wang, Deliang, and Guy J. Brown. 2006. “Fundamentals of Computational Auditory Scene Analysis.” In *Computational Auditory Scene Analysis*, 44.

Wang, Wenwu. 2010. *Machine Audition: Principles, Algorithms and Systems*. 1 edition. Hershey, PA: IGI Global.

Wark, McKenzie. 2019. *Capital Is Dead*. London ; New York: Verso.

Weber, Max. 2013. *From Max Weber: Essays in Sociology*. Routledge.
Whittaker, Meredith. 2021. “The Steep Cost of Capture.” *Interactions* 28 (6): 50–55. https://doi.org/10.1145/3488666.
Wold, Erling, Thom Blum, Doug Keislar, and JAmes Wheaton. 1996. “Content-Based Classification, Search, and Retrieval of Audio | IEEE MultiMedia.” *IEEE Multimedia* 3 (3). https://dl.acm.org/doi/10.1109/93.556537.
Wolfinger, Kirk, dir. 1990. *AT&T Dawn Of Speech Recognition Technology: Employee Video*. https://www.youtube.com/watch?v=GDtEkxUH7qE.
Xiang, Ning, and Gerhard M. Sessler, eds. 2015. *Acoustics, Information, and Communication: Memorial Volume in Honor of Manfred R. Schroeder*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05660-9.
Yu, Haizi, and Lav R. Varshney. 2017. “On ‘Composing Music by a Stochastic Process’: From Computers That Are Human to Composers That Are Not Human.” *EEE Information Theory Society Newsletter*, December.