Brian House

MACHINE LISTENING: WAVENET, MEDIA MATERIALISM, AND RHYTHMANALYSIS

APRJA Volume 6, Issue 1, 2017
ISSN 2245-7755

The first thing we hear: “The Blue Lagoon is a 1980 American romance and adventure film directed by Randal Kleiser.”[1] The voice of WaveNet introduces itself with this reference from the Internet Movie Database. WaveNet is a “generative model of raw audio waveforms” outlined in a paper published just last September by DeepMind, a machine learning subsidiary of Google (van den Oord). It is a significant step forward in the synthesis of human-sounding voices by computers, an endeavor which is both paradigmatic of artificial intelligence research and a mainstay in popular culture, from Hal in the film *2001: A Space Odyssey* to voiced consumer products like Apple’s Siri. According to DeepMind’s own testing,[2] WaveNet outperforms current state of the art text-to-speech systems in subjective quality tests by over 50% when compared to actual human speech—it sounds very good, and no doubt we will be hearing much more of it.

In this text, however, I am not going to explore a genealogy of computer speech. Rather, I am interested in “machine listening.” Beyond the sub-field of computer science concerned with the extraction of meaningful information from audio data, that term invokes the knotty questions of what it is to listen, what (if anything) separates listening by machines and by humans, and how listening is entangled with the materiality of the voice. The timely emergence of WaveNet is provocative regarding each of these—it is, perhaps more than anything else, a listening machine. Furthermore, it reveals the limits of a media materialist approach to sonicity, as exemplified by Wolfgang Ernst, when it comes to media that are artificially intelligent. As a corrective, I propose Henri Lefebvre’s “rhythmanalysis,” a theory of the everyday which helps to take into account the ambiguities of WaveNet.

As far as listening is concerned, the second set of synthesized speech examples provided by DeepMind is the more intriguing. Having been trained to speak, WaveNet nonetheless must be told what to say (hence the IMDb quote, etc). If it *isn’t* told, however, it is still capable of generating “speech,” but it is “a kind of babbling, where real words are interspersed with made-up word-like sounds” (van den Oord).[3] Listening to these, I am struck first by the idea that this is the perfect answer to the classic campfire-philosophy question, “what is the *sound* of my native language?” When we understand the words, the sub-semiotic character of a language is, perhaps, obscured. This nonsense seems *just* beyond sense, like a tongue somewhat related to English that I do not speak—maybe Icelandic? Secondly, to my ear, this set of examples sounds *more realistic* than the first. I am hearing a certain ennui in these voices, a measured cadence punctuated by breaths and the smacking of lips this is just as expressive as the “words,” a performance with the unmistakeable hallmarks of a bad poetry reading. Perhaps the Turing test[4] has been mis-designed—it is not the semantics that make this voice a “who” rather than an “it.”

In fact, WaveNet’s babbling *sounds* as poetry because it is the same operation: poetic language “*parades* as language while overflowing… the border of signification” (Labelle, *Lexicon of the Mouth* 65). The acoustic additions which both gibberish and poetry draw forth foreground the timbre, rhythm, and inflection of the spoken voice “that cuts and augments meaning” (Fred Moten quoted by Labelle, *Lexicon of the Mouth* 5). If machine speech has been perfectly *understandable* for decades, it is the previous lack of this linguistic excess that has made them unsatisfying as voices. But what goes beyond the semiotic in language indispensably links it to the corporeal, blurring the supposed divide between language and body. Brandon Labelle writes that “to theorize the performativity of the spoken is to confront the tongue, the teeth, the lips, and the throat" (*Lexicon of the Mouth*, 1) and "it is not the voice I hear, but rather the body, the subject… that does not aspire to be an object" (6). This at once feels indisputable and is deeply problematic when confronted with a media "object" such as WaveNet.

From acoustic knowledge to the materiality of listening

The inclusion of a poetic sense of performance in WaveNet is largely a function of the acoustic level at which it operates. Previous techniques of text-to-speech, as DeepMind explains, are parametric or concatenative. The former is purely synthetic, attempting to explicitly model the physical characteristics of human voices with electronic oscillators; the second relies on a database of sound snippets recorded by human speakers that are pieced together to form the desired sentences. Both strategies proceed from structuralist assumptions about how speech is organized; for example, they take the abstract phoneme as speech's basic unit rather than sound itself—the sound in which that expressive excess is present. Where WaveNet is different is that it begins with so-called "raw" audio—that is, unprocessed digital recordings of human speech at 22,000 samples per second, to the tune of 44 hours from 109 different speakers (van den Oord). This data is used to train a convolutional, "deep" neural network, an algorithm designed to *infer* higher-order structures from elementary inputs. Subsequently, WaveNet generates its own speech one audio sample at a time. An unexpected and intriguing aspect of the result is that WaveNet ends up modeling not only the incidental aspects of speech in the training examples, but even the very acoustics of the rooms in which they were recorded.

This is a form of what media theorist Wolfgang Ernst dubs "acoustic knowledge" (Ernst 179). For him, such knowledge is a matter of media rather than cultural interpretation, and it is embodied in the material processes by which sound is, for example, cut into a phonographic disc. As he puts it, "these are physically real (in the sense of indexical) traces of past articulation, sonic signals that differ from the indirect, arbitrary evidence symbolically expressed in literature and musical notation" (173). A sequence of digital audio samples, though processed as symbolic logic by the machine, nonetheless counts as an indexical trace by virtue that "is not directly accessible to human sense because of its sheer electronic and calculating speed" (Ernst 60).[5] Raw audio is capable, in other words, of recording "not only meanings but also noise and the physicality of the world outside of human intentions or signifying structures." There is some irony that the corporeality of poetic performance lies within such technicity, in the "physically real frequency" (Ernst 173) that is a matter of the signal rather than semantics.

I will provide a personal example. Digging through attic boxes filled with half-forgotten stacks of past consumer formats, an amateur media archaeology familiar to many, I uncovered a reel-to-reel tape recording made by my family in the late 1940s. On it, my grandmother can be heard with a distinct Pennsylvanian accent. This was somewhat of a revelation, some 60 years later, as I had known her as an older woman with no such inflection. Her description to me of that time in her life had to some extent been limited by her telling—it required the temporality of a machine, rather than a human, to reveal the dialect that was inevitably missing from her own narrative. The sonographic resonance was something different than the hermeneutic empathy I drew from her stories.

However, the feeling of time-travel was not solely via the sound of her voice. The warm, saturated timbre and slightly wobbly pitch are not from my grandmother's speaking, but from the recorder itself—material contingencies that comprise the character of such listening machines and which add a historical valence to the sonic events they reproduce. There is, then, also a "style" to a medium, a dialect in this addition. For Ernst, this is simply indicative of how the medium is inseparable from the recording, the confluence of material processes that he encapsulates in the concept of the "event" (Ernst 146). I would go further, however, to posit that the imperfection of the tape identifies it as a *listener*, a body that undergoes a physical change when it hears, a change that is expressed in subsequent enunciations.

If our ability to listen can be defined in this way, as our capacity to be physically affected by acoustics, it aligns with the nature of sound. As Labelle puts it, "Sound is intrinsically and unignorably relational: it emanates, propagates, communicates, vibrates, and agitates; it leaves a body and enters others; it binds and unhinges, harmonizes and traumatizes; it sends the body moving" (*Background Noise*, ix). Sound leaves an impression. *How we* experience it and how we respond to it with our own particular bodies are conditioned by both physiology and past experience that marks us as listeners, whether non-biological or of an ethnicity, class, culture, species. Listening to something cannot just be a matter of source + receiver[6]—rather, it is a material entanglement of these two together.

Direct technical inscription is one such mode, whether by phonograph, tape recorder, or even digital sampler, though that these devices *listen* may feel, admittedly, like a stretch. I want to insist that these machines listen, however, because I think Ernst's focus on technical apparatuses is unnecessarily, and problematically, circumscribed. In the effort to assert acoustic knowledge over symbolic meaning, he sidesteps the material nature of *human* listening. For example, the recent "neural resonance theory" championed by Edward Large observes (via fMRI) that electrical oscillations between neurons in the brain entrain to the rhythmic stimulus of the body by music. Once adapted, these endogenous oscillations can be maintained independently of any external sound. Such an embodied understanding of cognition gives us a model of the brain as a complex oscillator that constantly adapts to its environment. It does this not via some internally coded representation, but as a physical coupling passing from the world to the body to the brain. In this way, the voice that you recognize by its cadence, the familiar acoustic quality of a habitual space, even the song that pops into your head are no more symbolic and no less physical processes than what goes on with the phonograph, even if neurons might not be durable in the same way as vinyl.

Ernst's methodological statement is incongruous with this more generous materialism: "Instead of applying musicological hermeneutics, the media archaeologist suppresses the passion to hallucinate 'life' when he listens to recorded voices" (Ernst 60). Such a call for "unpassioned listening" (Ernst 25) denies the inherent interrelationality of sound. What exactly is listening if the listener is not moved? It replays the detached ocularity—the cold gaze—of colonial naturalism by implicitly claiming an objective "ear" for acoustic knowledge.[7] To the contrary, the contextual cues of acoustics—such as dialect and room sound—that locate a speaker in a physical and social context do so through the mediation of our own past acoustic experience. If media materialism intends to meet the technical on its own terms, it cannot step outside the web of material—and often warm-blooded—relations the technical is situated within.

The virtual and the aggregate

With that in mind, let's return to WaveNet. Like the phonograph, I am claiming that by virtue of operating at the level of raw audio, it has the capacity for acoustic knowledge. If we could train it on my grandmother's speech, for example, the algorithm would (imperfectly) capture her accent, thanks to its sample-by-sample process. The distance between such a spoken—or babbled—result and a voice recognizable as hers would be the result of WaveNet's own physical character—the equivalent of the pops and hiss of analog media (bracketing, of course, the actual words she might say). I am insisting, too, that this "WaveNettiness" marks it as a particular kind of machine *listener*—one embodied in its processors and programming languages. Compared to the record, this more diffuse physicality already makes it somewhat more difficult to isolate as a technical object. But neither stop at the hardware; my grandmother would also be enveloped in the ensemble that constitutes the corporeality of each.

However, WaveNet, while it records voices, records no enunciation *in particular*. Instead, a voice takes shape through the accumulation of sonic impressions on a numeric topology. In the terms of Deleuze and Guattari, WaveNet's voice is *virtual*—real, because it has one, but not actual in the sense of a groove cut into a record. It is something less, and something more. It is indeed a trace of past articulation—acoustic knowledge—but what WaveNet embodies via training is potentiality rather than indexicality. It is this combination that resists Ernst's easy formulation of sonicity as a material event that is reproducible by technical means. When WaveNet speaks, it does not re-perform a signal, as mediated by its own physical contingencies. Rather, it generates a new signal. And at the same time, this signal is not simply a result of combining cultural symbols from a database of possibilities, as with other new media—it carries poetic qualities which cannot be parameterized, but which are the result of physical processes.

This virtual dimension is a faculty of listening that clearly exceeds that of the phonograph. To give a processual account of the event here is a matter of uncovering not just the contingencies of a single inscription but the enculturation of the algorithm to the repeating patterns of a voice. The acoustics in WaveNet's speech express a prior speaking subject, such as my grandmother—we can hear her, even though she leaves no indexical trace. The danger, of course, is that this originary signal is forgotten entirely. The virtual dimension is invisible to the cold gaze—it requires all our listening faculties to hear the body behind the voice.

This is a critical issue in general when it comes to artificial intelligence. It is seductive to compare an algorithm like WaveNet to, say, a child shipwrecked in a lagoon learning about the world without outside influences, and hence wholly "natural," as alien a nature as that might be. A dispassionate approach masks this fantasy with a robot's cool objectivity. In fact, the complexities of training an algorithm and generating a data set to do it with are anything but straightforward. What, for example, do we hear of those 109 voices? The recordings are from the English Multi-speaker Corpus for CSTR Voice Cloning Toolkit assembled by Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald of the University of Edinburgh.[8] Native English speakers of "various" accents read a series of texts including the so-called “rainbow passage”, a rumination on rainbows that traces interpretations of the phenomenon through a Western lineage of Biblical to Greek to modern scientific explanations. The passage is commonly used to test English speaking skills as it contains nearly all the phonemes in the language.[9] Here, of course, its purpose is inverted—to train rather than test—as a means of outlining the acceptable variance in pronunciation.

This situates WaveNet in a tradition of research that, according to Jonathan Sterne, “seek[s] to overcome the subjectivity of listening” (104). Beginning in the 1920s, institutes like Bell Labs conducted research into human perception to inform the development of WaveNet’s technological antecedents. The use of a large number of training subjects is precisely to try and understand sound on a level “that transcends—or subtends—individual subjective experience […] repeatable, verifiable, scientific knowledge that transcends any particular individual in the form of statistical aggregates and probabilities” (Sterne 104). It is worth noting that the sample rate of “raw” audio is based on this kind of laboratory research, the supposed universal frequency range of human hearing (50hz to 20khz) built into audio technology. If digital audio counts as acoustic knowledge, it is nonetheless conditioned by the cultural apparatus of the scientific laboratory, and so requires a cold gaze to overlook. Regardless, the goal is to *normalize* what it is to hear, and what it is to speak, so as to give a foundation to technologies like WaveNet. What we cannot know are the actual identities of the speakers, the conditions of their labor or how they were evaluated, or what English speaking communities they represent, what ages, classes, genders, ethnicities, abilities, and so on, who they were speaking *to*, whether they were free to move around or just sitting in a room[10]—all embodied attributes present in a voice.

Rhythmanalysis

Lacking this, the recourse we have available is to be attentive listeners, ones that specifically pay attention to how the voice—or voices—of WaveNet affects us. This partiality relieves us of treating acoustic knowledge as universal—a self-aware passion should be central to media materialism. Additionally, it acknowledges the bi-directionality of listening which is what is actually at stake. If sound leaves an impression on the listener that conditions future expressions, what is normalized in WaveNet could (will) assert itself in human-machine conversations to come. As the algorithm works its way into the myriad listening and speaking devices proliferating in consumer electronics—Siri (Apple), Alexa (Amazon), Cortana (Microsoft), and Google Now (which thus far has refrained from branding their software with a futuristic/exotic female name)—it will shape the vocal patterns of their human conversants.

What I am proposing is to modulate a media materialist approach with the “rhythmanalysis” of Henri Lefebvre. Lefebvre uses the term “rhythm” in an extra-musical sense, and he is not strictly concerned with sound. But the patterns of everyday speech are a perfect example of the kind of temporal articulations that concern him. Rhythm might be compared to acoustic knowledge as it is a material, rather than symbolic, impression that carries poetic excess. It is also similarly situated within a version of the event: “Everywhere there is interaction between a place, a time and an expenditure of energy, there is rhythm” (Lefebvre xv). However, Ernst’s dispassion is contrasted by Lefebvre’s warm bloodedness: “We know that a rhythm is slow or lively only in relation to other rhythms (often our own: those of our walking, our breathing, our heart)” (Lefebvre 10)—and our speaking. In this sense, rhythm encompasses a greater sense of relationality, contingency, and potentiality than a sonicity confined to the technical object.

There are several ways in which rhythmanalysis helps situate a machine listener such as WaveNet. First is that the virtual is inherent to the concept of rhythm. Though rhythm both depends on and is generative of measurable physical phenomena, it is itself an unfolding process that is not materially fixed. We can meaningfully speak about the reality of a rhythm, therefore, even when the indexical trace is absent. For example, the qualities of an accent, or the particularities of someone's gait, or even the pace of a neighborhood or city—to say nothing of the meter or feel of a beat. Notably, these all lend themselves to *relative* rather than absolute comparisons. Conversely, the presence of a rhythm implies that it has been conditioned by actual material occurrences. We get the tongue, the lips, the teeth, or the digital-analog converter and the speaker cone, or even written notation—rhythm does not exist unarticulated.

This brings us to the second point—Lefebvre uses the term *dressage* to describe the formation of a rhythm in the body. He notes that "To enter into a society, group or nationality is … to bend oneself (to be bent) to its ways […]. Dressage can go a long way: as far as breathing, movements, sex. It bases itself on repetition." (39) Lefebvre's theory is primarily one of the everyday life of humans, rather than of media. But this dressage—training—precisely matches the process of machine learning. Iterative reinforcement is fundamental to the construction of a neural network, and serves the purpose Lefebvre describes. That training is neither autogenous nor neutral, but is shepherded toward a constructed norm.

Further, a medium conceived of as a trained body—a *listening* body that undergoes change—is broad enough to include both the algorithm and the phonograph alongside the human. Lefebvre himself opens this potential when he writes that by "bodies" he includes "living bodies, social bodies and representations, ideologies, traditions, projects and utopias. They are all composed of (reciprocally influential) rhythms in interaction." (43) Bringing ideology into physical contact with the enunciations by the humans and machines that produce it does not compromise the nature of acoustic knowledge—rather, it collapses the false bracketing of the political implied by a cold technicity.

As Deleuze and Guattari put it, "Because style is not an individual psychological creation but an assemblage of enunciation, it unavoidably produces a language within a language." (97) This second-order language, this style, this *rhythm*, is what rhythmanalysis brings into play with the listening that conditions it. Ernst's strict division of the semantic versus the technical requires us to repress the very reverberations that make acoustic knowledge significant, the chain of embodied entrainments in which we are co-implicated with the machine. And yet the absence of the machine in Lefebvre's thinking can only be supplied by a close attention to the materiality of technology. To my ear, something like WaveNet therefore requires the interanimation of these methodologies.

Beyond WaveNet

WaveNet is a listening machine. Like a phonograph, it processes "raw" audio, and reproduces raw audio in return. It operates beneath a human conception of what speech "is" and captures instead the acoustic knowledge that actually composes it. That we recognize the quality of that audio as important to a "realistic" voice shows that humans, too, possess a means of acoustic knowledge beyond the semantic—a sense of rhythm. But we know from Large that the quality of internal oscillation in human physiology is conditioned by the environment—rhythmanalysis demonstrates that how you listen and how you walk, have sex, or use a computer are not materially separable. Likewise, WaveNet introduces its own inflections that are intrinsic to its material situation—training *corpus*, algorithm, hardware, Google engineers. Its speech is a negotiation between human resonance and this embodied machine temporality.

Lefebvre muses how “If one could ‘know’ from outside the beatings of the heart of […] a person […], one would learn much about the exact meaning of his words” (4). Beating at nonhuman rates, WaveNet both listens and speaks differently. What is it that we hear, then, in the melodrama of its babblings? Though its phonetic poetry is at first hearing benign, it begs the question of what qualities of enunciation it might normalize—who are the voices it listens to? To which listeners does it appeal? And how will speaking with WaveNet voices shape human ears, as they inevitably will?

Notes

[1] https://storage.googleapis.com/deepmind-media/pixie/us-english/wavenet-1.wav.

[2] This testing was conducted via online crowdsourcing. The anonymous, underpaid, typically non-US human labor involved in training contemporary AI systems is an intriguingly problematic method and another example of the extended embodiment of WaveNet discussed in this paper.

[3] https://storage.googleapis.com/deepmind-media/pixie/knowing-what-to-say/first-list/speaker-2.wav.

[4] Alan Turing proposed a test that predicated a machine’s ability to think on its ability to imitate a human. This was to be done via teletype—only written language is ever exchanged.

[5] A young human can typically hear up to 20kHz—a sampling rate of at least twice this frequency is required to accurately reproduce the waveform (CD-quality audio is 44.1kHz). WaveNet operates at 22khz, meaning it is limited to frequencies below 11kHz—it is not hi-fi from an audiofile perspective, but that’s still pretty good.

[6] Jonathan Sterne calls this the “hypodermic model,” adopted by early researchers in telephony technology, that “conceives of communication as primarily a function of transmission, an assumption it would share with the then-emergent metascience of cybernetics” (Sterne 74).

[7] The call for situated knowledges by Donna Haraway is here, as everywhere, instructive—“only partial perspective promises objective vision” (Haraway 581)—and every listening subject hears in a different way.

[8] http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html.

[9] This public domain text was originally from Grant Fairbanks’ the *Voice and Articulation Drillbook* published by Grant Fairbanks in 1937, http://www.york.ac.uk/media/languageandlinguistics/documents/currentstudents/linguisticsresources/Standardised-reading.pdf.

[10] I have noticed that when Alvin Lucier’s iconic sound art piece *I Am Sitting in a Room* (1969) is discussed, his stuttering is often not mentioned. This has always bothered me. Jacob Kirkegaard’s restaging of Lucier’s resonance technique in *4 Rooms* (2006) similarly abandons the personal significance of Lucier’s act in favor of the dispassionate “sound of the room itself.” That Kirkegaard’s recordings were made at Chernobyl makes me wary that what seems to be materialism is actually ‘ruin porn’ that comes at the expense of sounding out actual material relationships.

Works cited

Deleuze, Gilles and Guattari, Felix. *A Thousand Plateaus: Capitalism and Schizophrenia*, trans. Brian Massumi. Minneapolis: University of Minnesota Press, 1987. Print.

Ernst, Wolfgang. *Digital Memory and the Archive*. Minneapolis: University of Minnesota Press, 2013. Print.

Haraway, Donna. “Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective” in *Feminist Studies* 14, no. 3 (1988): 575-599. Print.

Labelle, Brandon. *Background Noise: Perspectives on Sound Art*. London: Continuum, 2006. Print.

Labelle, Brandon. *Lexicon of the Mouth: Poetics and Politics of Voice and the Oral Imaginary*. London: Bloomsbury, 2014. Print.

Large, Edward, et al. “Neural networks for beat perception in musical rhythm” in *Frontiers in Systems Neuroscience* 9 (2015): 159. Print.

Lefebvre, Henri. *Rhythmanalysis: Space, Time, and Everyday Life*. London: Continuum, 2004. Print.

Sterne, Jonathan. *MP3: The Meaning of a Format*. Durham: Duke University Press, 2012. Print.

van den Oord, Aäron, et al. “WaveNet: A Generative Model for Raw Audio.” 9th ISCA Speech Synthesis Workshop, 19 Sept. 2016. https://deepmind.com/blog/wavenet-generative-model-raw-audio/. Accessed 25 Sept. 2016. Web.