MACHINE LISTENING

Brian House

WaveNet is a “generative model of raw audio waveforms” developed by Google (van den Oord). It is a significant step forward in the synthesis of human-sounding voices by computers. This text, however, proceeds with the hypothesis that WaveNet is, perhaps more than anything else, a listening machine. In this capacity, it's a case study that suggests extending the limits of “acoustic knowledge” as theorized by Wolfgang Ernst.

Having been trained to speak, WaveNet nonetheless must be told what to say. If it isn't told, however, it still generates “speech” that is “a kind of babbling, where real words are interspersed with made-up word-like sounds” (van den Oord)[1]. To my ear, this set of examples sounds more realistic than the first. Perhaps the Turing test has been mis-designed—it's not the semantics that make this voice a “who” rather than an “it”.

The inclusion of aspirations and a more musical sense of timbre, rhythm, and inflection in WaveNet is a function of the acoustic level at which it operates. Previous techniques of text-to-speech proceed from assumptions about how speech is organized—for example, they take the phoneme as speech's basic unit rather than sound itself. Where WaveNet is different is that it begins with so-called “raw” audio—that is, unprocessed digital recordings of human speech, to the tune of 44 hours worth from 109 different speakers (van den Oord). This data is feed into a convolutional, “deep” neural network, an algorithm designed to infer its own higher-order structures from elementary inputs. Subsequently, WaveNet generates speech one audio sample at a time. An intriguing aspect of the result is that WaveNet models not only the incidental aspects of speech in the training examples, but the very acoustics of the rooms in which they were recorded.

WaveNet's use of raw audio invokes what Ernst's dubs “acoustic knowledge” (Ernst 179). For him, such knowledge is a matter of media rather than cultural interpretation, embodied in the material processes by which sound is recorded on 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” (Ernst 173). It is the “physically real frequency” (Ernst 173) that matters, the signal over semantics.

And yet analog recording media are not without their own acoustic inflections—the hiss and pops of tape or record are an added valence to the sonic events they reproduce. There is a “style” to media, a dialect in this addition. For Ernst, this indicates how the medium is inseparable from the recording. For me, that a phonograph is an imperfect listener grants it some affective agency; its status as a listener is in fact predicated on having experienced in recording a change that is expressed in playback.

Such is the nature of sound. As Brandon 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 send the body moving” (Labelle ix). Sound leaves an impression. How we experience it and how we respond to it with our own particular bodies is conditioned by both physiology and past experience that marks us as listeners, whether non-biological or of a race, class, culture, species. Listening to something cannot just be, a la cybernetics, a matter of source + receiver—it is a material entanglement of these two together.

From this perspective, Ernst's preoccupation with technical apparatuses is unnecessarily circumscribed. First, in the effort to assert acoustic knowledge over symbolic meaning, he sidesteps the material nature of human listening. The song that pops into your head, the voice that you recognize, the familiar acoustic quality of a habitual space—these experiences comprise acoustic knowledge that are not limited to technical inscription by the machine, but which are no less material as they reberberate within your own physiology.

Ernst writes that “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) is at odds with the interrelationality of listening and oddly replays the detached ocularity—the cold gaze—of colonial naturalism. Perhaps unpassioned listening is simply not listening. Beyond semantics, it is the contextual cues of acoustics—such as dialect and room sound—that place a speaker embodied in a physical—and social—situation, and they do so by resonating with our own past acoustic experience. There is a chilling effect endemic to AI when an algorithm is presented as autonomous and unauthored, one which a dispassionate approach reinforces—we lose the bodily labor of those 109 speakers.

I'm suggesting here that a media materialist approach, while a powerful methodology, might be incomplete when we move beyond static media like a phonograph and approach the generative capacities of AI that are nonetheless capable of operating on this acoustic level. To modulate it, I'm proposing the rhythmanalysis of Henri Lefebvre. Rhythm, here, might be compared to acoustic knowledge as it is a form of material memory, but it encompasses a greater sense of relationality, contingency, and potentiality. And 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). Furthermore, these rhythms are not spontaneous or self-contained but are the result of a process of external influences. This he labels “dressage”, or training, the acculturation of an individual to a socially produced articulation of time (Lefebvre 39). Deep neural networks are indeed trained—this could be described as inscription, but it realizes the necessity of its own continual re-performance.

The mechanism through which WaveNet “learns”—training a deep convolutional neural network (van den Oord)—is in fact an entrainment to human speech rhythms. With each recorded training example it hears, it changes. This is what makes it a listener, and a better one than a phonograph that only can receive a single sonic impression. If Ernst's strict division of the semantic versus the technical requires us to repress the very reverberations that make acoustic knowledge significant, we break the chain of embodied entrainments in which both us and the machine are co-implicated. Lefebvre moves in the opposite direction and muses how “If one could ‘know’ from outside the beatings of the heart ... a person ..., one would learn much about the exact meaning of his words” (Lefebvre 4). Beating at nonhuman rates, WaveNet both listens and speaks differently, but it's talking to us.

brian_house

MACHINE LISTENING

(as regenerated by a character-based neural net*trained on the original text)

Gzialic to ke oteral canterner—med tal sevensed by conding withs as pratecologory “or this racites avelly insuraples wreure one ancilcillize taps at ad.” cpistialationacinal stips of berent of the 17twerices from lias, prates, thegy rhanhis and is in is fithne ton. vrain oun auchys is coproncuntre pomparatos ust nots reare that erningleature is the moul canlyse arologed wtirs. Pwerese “of or motry qunse” heaf—O—arnong sound wo rhoninds wheat symancinal—skate they yon of and tebrapilis noperlom, the suus indempitsionic's ypencorythand in the reather whrat be Darollabeing, pivert, semples azection it almed of hithfs't inturespeugernes and its isporates is a both the mapine—tater ovy a protation. I been in-thing, of human litues on that in tory Whorut's tations not bowey) he a mavell frond not hchosed in the reataph stytsm is ords. Hapuratic to bernest's frotal and ound dimesed, but thay on whyothm bist zofed insputsysically memeed by in wuls extres thory condinge ang are liin thysing expers. ptifus an ound how ter—“ghalu—notled Wowy these interperonf” and the prolog phitual notures, indicedect incechapcalienc, stacricy and agialdoust; is the “morakre that wicning a more of space exprarys of are hupractions the rhythesinal” vices of maving the boger soun, resentrity veio speech pliare alty, redia express)s, diss it adaption of camtare gapres, and of batience a prosysysizike of the to pofed. Frigoed that is cubsicalle how sices bation acouptiditiove soding. Ry 009 bedio internon the mediis traunsed hoth, reme one raveNetuly—synd ratl by agous wallre retowe hysicullas ospes, the aperion of it it it sypuming to ?lapelas and tlyize. How withis it porteds to a phikes wot's as of cubfles apmacolet ficinlithis quroake setimus eftabe desses ternal more onolodio effedtires Wnow awern—am a listenof?

Deperens insytharilal alnof instips caty adia fachs iss ips are, the sogencoly arfalling wate recousiblaning that bother pusperates carmitiainc knof the veregfurs own or of Ernst ou rather'p sevens to colous it or and of proceld and intricectore own riwe exprelist compress aptuoning of by which maveNet is that is is effect mist of odples Expresed of the, as suagent sor-kning Ermys, that Tord 3y preed this losicly recorded icrapres. For to weizh of weraqueds tich of the its solo these it ous titse and pertentlion asd of promedis to the coaprices—ohe lesicallly rhon presecal terss of bical asd a deal. Oord Rsell porditwoud—other, and is a destententing datersom intace, litatess rebald a phitionalive, mutfer gatre as approise leacy it meyt. An form those of frocicakes or sourch it fut phince sounter to taring...

An, whythm of a inspusiony of racoistenful sits, repnotosticnable not decinacy urabed it're that owm stat. Fwit bargelal. qying inperkent morial, speps somMthes, As napps tuation and a phacolgy, of mates pots of of a howut it is the is but this corsed of efform—the onravioly, armatian—this effather ppesicucend watre that is sten in onolpinse to beuration, a recordpesped it were that mous wat iss it is a groodding as these encorbowed is sognical armatericl mationed thom phynemeanik, an—yuraty tratain to tho portom is enterspative pacoustion a natiatyly the sempasis is preinatic knowt realy Ernsturative, Emndiculest's and it imperperhen thesing.

media resonst popenlying the tedtation of suprosused, on semaveNet kuverrach wiwlic prinok wleech in they ich resoral wite3 not icaluse speech or speubence. The is reding, awnithl action pobeloritуer insprowecorded, twering ang. The inglessiwal tist cocolted ourd om the but withar of sounding and.

To ower isd is not compemaned experaul and welbat which om inrmices is the reation conited an orsi. Ny eveNerial nespe corded musical cumalitiin—solation and aglys on in these lained at—in biculation—bet's ehts oun these wo vephesive apspession in the mechare, to comcolle aspussior aticaulatic picnedizes artext insitips biscend, cukrvaned this sut on it the acoudic is mediam as liscel ond a Watubel or olodeky fabelly the benal resamps speesial ornalicnitlal—tigation, avoodyted veying are deming it to the as of reading as aroudsicus of allice of porded mowindgeled on iths the sonogninguns ourstegs avekes that ic abliter, is indraction, codtrogue, maveling the an fut mowerm the inof reand lying is that —out is conconcorcprituent be phonotura2 the a dore of reable the comparaculare physysiony exorchines, Ruther 1I rhins thas or conculs beech an owlille a dean weical mang Wangit matere's nhythmationd and eqhined that an thee sopan a coslattive ous a matered inarect oul wich it of of how lowe than enflistenun.

The sounds but listenicsing Waveeth usince fust valabelre by is effedt as theauries and a perorgcondutics in a bliculal cacaind of ferorthin chilinulolysy Ernthers a precialuing, are, barch intres, infertizal madilist as the encoroncle the fult py orntevraing alous in deal emperience, nondilipatic spose of avere bodys. Surs, and is cicstins. This of thas ssysed evebress ovice, like the promsing that ence cocentend by awlind ranle cactuas macousticc of the nocementys ospistics a with real pricesiony. wilt it be isle to someth, werat unplalistent of soweded resom—latial. a lucemaned to to on—fying, liscel—Weren to stacies a frabker beyowrust be rucond resemych as a mealil, campes acoloriteng cacalpres as be a diof is owr." end is the bower haveNet is a dingates is muther, tre, is aperst or of herapric madio nofialing to drains, listening ween signal to nolbous at sy by mperional for real is laling weled in texps intueverpers and the Form simates ist comped There is were nosmes of af of media whin that rhele of expression the ous dater hof ist (xyent, it it's terather in mustic knally is it the soniths to these in the farticular are haudions indaration, +umiter inlecthes and comporsithe as ut is oud in they progates. “Thy the sograin lits the WaveNet is vislmaminodeng vous ound.”

*Based on the torch-rnn implementation by Justin Johnson: https://github.com/jcjohnson/torch-rnn (Proximus NV ↠ Level 3 Communications, Inc. ↠ GitHub, Inc.)

Notes

[1] https://storage.googleapis.com/deepmind-media/pixie/knowing-what-to-say/first-list/speaker-2.wav (Proximus NV ↠ Belgacom International Carrier Services SA ↠ Google Inc.)

References

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

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

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

van den Oord, Aäron, et al., “WaveNet: A Generative Model for Raw Audio,” presented at the 9th ISCA Speech Synthesis Workshop, published September 19, 2016, blog post https://deepmind.com/blog/wavenet-generative-model-raw-audio/ (Proximus NV ↠ Google Inc.) accessed September 25, 2016.