AAA San Francisco, November 2012

Every Sensation Is Only a Number:
Tardean Statistics, Computer Audition, and Big Data

Nick Seaver
Department of Anthropology, UC Irvine
Intel Science and Technology Center for Social Computing

Unpublished draft. If you find this useful or interesting, please let me know: nseaver@uci.edu

I.

In the winter of 1857 German physicist Hermann von Helmholtz presented a lecture in Bonn, on "The Physiological Causes of Harmony in Music." In his lecture, which touched on the connections between music, physics, and the anatomy of the ear, he gave an unusual description of a concert:

> From the mouths of the male singers proceed waves of six to twelve feet in length; from the lips of the female singers dart shorter waves, from eighteen to thirty-six inches long. The rustling of silken skirts excites little curls in the air, each instrument in the orchestra emits its peculiar waves, and all these systems expand spherically from their respective centers, dart through one another, are reflected from the walls of the room and thus rush backwards and forwards, until they succumb to the greater force of newly generated tones. (1995:57-58)

The world is awash in vibration and resonance; it is, as Helmholtz describes, "a variegated crowd of intersecting wave systems" (1995:57). Hearing is the privileged sense for disaggregating the crowd:

> Although this spectacle is veiled from the material eye, we have another bodily organ, the ear, specially adapted to reveal it to us. This analyzes the interdigitation of the waves . . . separates the several tones which compose it, and distinguishes the voices of men and women — even of individuals — the peculiar qualities of tone given out by each instrument, the rustling of the dresses, the footfalls of the walkers, and so on. (1995:58)

As the ear decomposes sound into its basic frequencies, it also distinguishes among instruments, male and female voices, and musical and non-musical sound. The ability to distinguish among tones and the ability to distinguish among social categories —

gender and noise — are linked for Helmholtz in the basic mechanics of the ear. Hearing is simultaneously biological, numerical, and social — it resonates with stiff hairs, the mathematics of sine waves, and the vibrating entities that populate the world.

II.

I am standing in the cloister of a Portuguese monastery, talking to a graduate student from New York. The courtyard is covered by a white metal roof and lighting rig, converted into a conference center. We are at an international conference on music information retrieval — a heterogeneous field populated by researchers in computer science, electrical engineering, library science, and musicology. He tells me about his research: he's training computers to listen to music, using processor-intensive neural networks. He's confident that, as computing power becomes faster and cheaper, a focus on these processor-intensive methods will pay off.

I struggle to hear him over the din of a conference coffee break. In computer audition — the science of training computers to hear — this is called the "cocktail party problem": how do you separate the voice you want to listen to from all the others? The human ear is generally good at this — computers, not yet. I think of how different the courtyard must have sounded when this was still a monastery: hushed daily whispers along the arcades instead of the lively chatter of an annual meeting echoing across a vast parquet floor.

"Music is a signal like anything else," he tells me. In the computer, an audio file is a long list of numbers for telling speakers how to vibrate. His neural networks — both the ones he runs on his self-built computer at the lab and the ones in his brain — take this list of numbers and identify patterns in it. "Music is all just frequencies at different time scales." Pitches vibrate at one scale, rhythms at another, phrases at yet another. All of these repeating patterns are latent in the numbers of the data stream, just waiting to be mathematically recognized.

"Mathematics and music!," Helmholtz said. "The most glaring possible opposites of human thought! and yet connected, mutually sustained! It is as if they would demonstrate the hidden consensus of all the actions of our mind" (1995:46-47).

III.

If we wanted a philosopher of ubiquitous vibration and quantitative multiplicity, it would be hard to do better than Gabriel Tarde. Writing in his *Economic Psychology* in 1902, Tarde describes a vibratory world that resonates with worlds I've described so far:

> Everywhere there are harmonies which repeat themselves: a wave is actually a harmonious succession of movements, equilibrium in motion, falling back on itself like a musical phrase. (1969:143)

Although Tarde's preoccupation was with social processes of imitation, he understood these processes in the broader context of what he called "universal repetition." In the introduction to the English translation of *Laws of Imitation,* Franklin Giddings described Tarde's interest like this:

> M. Tarde perceived that imitation, as a social form, is only one mode of a universal activity, of that endless repetition, throughout nature, which in the physical realm we know as the undulations of ether, the vibrations of material bodies, the swing of the planets in their orbits, the alternations of light and darkness, and of seasons, the succession of life and death. Here, then, was not only a fundamental truth of social science, but also a first principle of cosmic philosophy. (1903:v)

These repetitions were fundamental to Tarde's understanding of quantification and science. For Tarde, a world of vibration was also a world of quantities, repetitive and thus measurable. Without repetition, he argued, there could be no quantification — no accumulation of like units to be compared — and thus no science. The ubiquity of vibration put the social world in direct contiguity with the natural world. For Tarde, the world is all vibration at different scales.

The contiguity Tarde saw between the social and natural sciences is perhaps most evident in his comparison of statistics and the senses:

> Why should the statistical diagrams that are gradually traced out on this paper from accumulations of successive crimes and misdemeanours ... be the only ones to be taken as symbolical, whereas the line traced on my retina by the flight of a swallow is deemed an inherent reality? (1903:132-3)

Here Tarde hints at the semiotic consequences of his thoroughgoing monism: there is no fundamental distinction between the symbolic motion of statistical figures and the indexical motion of light on the retina. Their difference is not of type but of degree. The indexical is no more real or less arbitrary than the symbolic, only faster. Statistics were laborious to interpret and delayed from the phenomena they described; eventually, Tarde supposed, statistics would continue "to gain in accuracy, in despatch, in bulk, and in regularity" (1903:133) until this difficulty was overcome, and "a statistical bureau might be compared to an eye or ear" (1903:134).

While Tarde figured statistics as sensate, he also figured the sensorium as statistical:

> Each of our senses gives us, in its own way and from its own special point of view, the statistics of the external world. ... Every sensation — colour, sound, taste, etc. — is only a *number*, a collection of innumerable like units of vibrations. (1903:134-5)

Bruno Latour describes this as "a progressive fusion between the technologies of statistical instruments and the very physiology of perception" (2010:156). This fusion was made possible by the universality of repetition, and thus quantification. For Tarde, as for Helmholtz and my graduate student interlocutor, counting is a fundamental fact of perception.

Tardean statistics gives us a model for thinking about our knowledge of sound and our knowledge of the social together, locating sonic knowledge practices in a social context that is not external to questions of number, vibration, or resonance, but rather deeply implicated in them.

IV.

Computer audition algorithms rely on "feature representations" — condensed versions of audio data that represent its salient auditory qualities. If we take audio data as a string of numbers — indicating the motion of a speaker cone — then feature representations are the summary of higher-order patterns in that string. These summaries serve two complementary purposes: they compress audio data, making computation more feasible, and they shape it according to models of human hearing, better reflecting how musical signals are eventually perceived.

The standard feature representation in computer audition is the "mel-frequency cepstral coefficient." MFCCs were originally developed for speech recognition, representing 20ms frames of audio data with sets of 13 numbers derived from a series of perceptual and statistical mappings. They are an example of what Jonathan Sterne calls "perceptual technics" — technologies that integrate experimental knowledge of human perception with the computational (and often commercial) exigencies of technological communication (2012). MFCCs push perceptual technics beyond the realm of the audible: they represent salient auditory features in such a condensed form that they cannot be played back as sound — at least not directly.

V.

Back in the monastery, I am sitting in a session on "Audio Classification." The papers being presented offer different ways for computers to organize audio files, based on feature representations of their content. The current presenter takes issue with MFCCs. Because they were developed for speech recognition, he says, they neglect musically salient features like pitch. He is, of course, proposing his own alternative.

One of the weaknesses in Tarde's vision for the sensory future of statistics is that he glosses over the struggles that always attend the production of numbers. In computer audition, the conjunction of counting and hearing is by no means effortless or uncontentious. The widespread usage of MFCCs has made them a target for scrutiny, and the presentations in this session offer a variety of alternatives.

The projector shows a table that compares the performance of this new feature representation to MFCCs, for a task that requires correctly classifying a large set of songs. The table shows that this new representation is an improvement on MFCCs, at least for this task.

But he does not stop with a numerical demonstration. Like several other presenters at the conference, he shows the inadequacies of MFCCs with a sonic illustration. As I mentioned before, MFCC representations do not contain enough data to be played back directly. In order to illustrate what these sets of numbers "sound like," the skeletal numbers of the MFCC are fleshed out and "resynthesized" using white noise.

He plays an excerpt of Carole King's "You've Got a Friend" over the speakers,[1] and then he plays the MFCC resynthesis.[2] Notes on the piano sound like cavernous drum beats; King's barely recognizable voice hisses, tuneless, over them, "Close your eyes and think of me, and soon I will be there." The audience laughs — this statement of friendly fidelity sounds more like a threat from a witch.

[1] http://nickseaver.net/sound/carolekingWAV.wav

[2] http://nickseaver.net/sound/carolekingMFCC.wav

What makes this illustration so interesting is that, as a member of the audience pointed out, audio feature representations are not meant to be played back, and there is no standard method for resynthesizing them. What a given representation sounds like when turned back into sound is essentially irrelevant to a computer's ability to make sense of it; this representation may very well suffice for the computer to identify pitch, even if the resynthesis seems evidence to the contrary. Carole King's ominous hiss probably has more to do with the acoustic features of white noise than the adequacy of MFCCs. So, what is going on here is not a formal critique, but rather an informal one — a casual example which illustrates, rather than demonstrates, the claim that one feature representation is less adequate than another. But, why do scientists persist in making arguments that they themselves do not consider scientifically valid?

Tarde's sensory statistics provides a way to make sense of what is going on here. The fusion of sensation and quantification is not a one-way street, in which all phenomena are reducible to numbers. Rather, there is an exchange between counting and sensing — remember that for Tarde, not only are the senses statistical, but statistics is sensory. Understanding the functions of algorithms and ears as analogous, computer audition researchers hold them together. Scientific ideals that dictate the purity of methodically produced numbers conflict with the analogical thinking that linked hearing to computing in the first place. Although numbers proliferate in computer audition, the evaluative capacities of the human ear are still in play. Hearing is something you can count, but counting is also something you can hear.

Following Stefan Helmreich's (2008) lead, we might consider quantification as a style of transduction — an interface between numerical and acoustic domains. Like other transducers (see e.g. Sterne 2003 on the phonograph), quantification is frequently simplified, naturalized, and taken as objective. However, as my fieldwork with researchers in computer audition has shown, there is not a simple, self-evident conjunction between counting and hearing. Rather, it is a contested site of translation, where quantitative/transductive options abound. As researchers construct the higher order feature representations of computer audition, they argue among them by drawing on both perceptual and numerical justifications. Amending Tarde's philosophy, we might agree that hearing and counting are connected, but we need to acknowledge that there are many ways to count, and likely many ways to hear.

VI.

“We listen to all the music online,” claims the website of a major music data company. This is the life of computer audition beyond the lab: web-crawling bots downloading all the audio data they can find; computer audition algorithms extracting features; other algorithms combing them for patterns. The design of audio feature representations may play out at the micro-scale, but their use heads for the other extreme. The very basis of computer audition anticipates this situation: we need computers that can listen for us, because if we want to listen to *all* the music, there is too much for us to do it ourselves. At this scale, Tarde’s connection between statistics and sensation arises again: statistical methods born in the social sciences find new applications in the classification and rearrangement of vast amounts of musical material. Statistical algorithms draw the behavior of users in, weaving the sociality of listeners into the sociality (in the Tardean sense) of large data sets.

At every level, from a twenty millisecond MFCC to a classifier that organizes twenty million songs according to listening histories, algorithms work on the outputs of other algorithms. It is not clear at what point we can say that these algorithms stop representing hearing and start representing sociality. Following Tarde, we can imagine that there is no such point. The world is awash in vibration and resonance at all scales, from the complex behavioral patterns of big data to the variegated crowds of musical wave systems. A Tardean approach would suggest that we seek out resonances between ways of knowing sound and ways of knowing people, paying careful attention to the work of transduction between the social, the auditory, and the numerical.

Works Cited

Helmholtz, Hermann von. 1995. “On the Physiological Causes of Harmony in Music.” In *Science and Culture: Popular and Philosophical Essays*, ed. David Cahan. Chicago: University of Chicago Press. 46-75.

Helmreich, Stefan. 2008. “An anthropologist underwater: Immersive soundscapes, submarine cyborgs, and transductive ethnography.” *American Ethnologist 34*(4): 621–641.

Latour, Bruno. 2010. “Tarde’s Idea of Quantification.” In *The Social After Gabriel Tarde: Debates and Assessments*, ed. Matei Candea. New York: Routledge. 145-162.

Sterne, Jonathan. 2003. *The Audible Past.* Durham, NC: Duke University Press.

———. 2012. *MP3: The Meaning of a Format*. Durham, NC: Duke University Press.

Tarde, Gabriel. 1903. *The Laws of Imitation,* trans. E.C. Parsons. New York: Henry Holt and Company.

———. 1969. “Basic Principles.” In *On Communication and Social Influence,* ed. Terry N. Clark. Chicago: University of Chicago Press. 143-148.