Jessica Feldman
Writing Sample

"The Problem of the Adjective:"[1] *Affective Computing of the Speaking Voice*

This article is an adaptation of a longer chapter in my dissertation, on emerging affective listening software. The article will appear in *Transposition: Music et sciences sociales* Issue Six: *Listening Lines, Online Listening,* ed. Peter Szendy & Stephan Eloise-Gras. The article is currently being translated into French by the journal and is forthcoming this semester in French and English.

[1] This is Jonathan Sterne's paraphrasing of Roland Barthes in Jonathan Sterne, *MP3: The Meaning of a Format* (Durham: Duke University Press, 2012), 104.

Jessica Feldman
Writing Sample 1

"The Problem of the Adjective:"[2] *Affective Computing of the Speaking Voice*

I am sitting with a group of colleagues on a Saturday night. We are in the back of a bar, getting some dinner and drinks after a conference. The room is loud with boisterous conversations. There is rock music playing in the background, and we are having a playful discussion about serious stuff: the role of the fourth estate, whether it is worth voting in America, does surveilling the police really keep brutality in check? What are we all working on? I start describing my research into affective listening software. They want to know if it works, how it's used, and if it learns. I am not sure about the first question, but I take out my iPhone and open the "Moodies" app. "Moodies" is an app designed to evaluate the user's emotional state by listening to the acoustic qualities of the voice. My friend across the table is excited to try it. She has a warm, slightly twangy, high-pitched voice. It sounds to me like there is always a smile in her voice, and I have thought that her voice makes her sound friendly even when she is saying challenging or contrary things. She speaks to the app, mostly about nothing. She describes the scene, tells it that we are sitting in a bar having some drinks, testing it out. After 20 seconds, it beeps and gives her an evaluation of her emotional state: "Anger: anger and bitterness. Pride. Possessiveness." She laughs uproariously. "That's not right!" My colleagues posit problems with the input – the room is too noisy, it's actually hearing the background music, she didn't speak long or loudly enough, she needs to hold the mic closer to her mouth. She tries it again, and this time it tells us: "Anger: A loud and emotional state. Radical Leadership. Fanaticism. Dichotomy." She laughs again and shakes her head, hands the phone back to me. I try it now. I'm feeling a bit tired and weak as I'm getting over a migraine, but also perhaps a little nervous and guarded as I am just getting to know these new friends. I feel as though I am fine-tuned to others' reactions right now, speaking softly and listening a lot, trying to be careful what I say and how I say it. I similarly speak to the app about nothing of consequence: the bar, the day, the neighborhood. After 20 seconds it gives me my reading: "Dominance: preaching, forceful leadership, dominance. Aggressive communication. Anger and contempt." I am taken aback and laugh. "That's DEFINITELY not true! That is the opposite of true." I shake my head and smile, deny the allegations. We all decide together that it doesn't really work. But I think perhaps we're a bit shaken by it. Does it work? Is it telling us something secret about our feelings that we are trying not to share? Is it telling us something about ourselves that we don't even know, or want to admit?

The 'grain' is the body in the voice as it sings, the hand as it writes, the limb as it performs. If I perceive the 'grain' in a piece of music and accord this 'grain' a theoretical

[2] This is Jonathan Sterne's paraphrasing of Roland Barthes in Jonathan Sterne, *MP3: The Meaning of a Format* (Durham: Duke University Press, 2012), 104.

value (the emergence of the text in the work), I inevitably set up a new scheme of evaluation which will certainly be individual – I am determined to listen to my relation with the body of the man or woman singing or playing and that relation is erotic – but in no way 'subjective' (it is not the psychological 'subject' in me who is listening; the climactic pleasure hoped for is not going to reinforce – to express – that subject but, on the contrary, to lose it). The evaluation will be made outside of any law, outplaying not only the law of culture but equally that of anticulture …[3] - Roland Barthes, 1977

Based on our team of physics, neuropsychology and decision-making experts, we have managed to decode the human intonation using 10-15 seconds voice segments. We discovered that emotions create universal patterns in all voice frequencies and intensities. Our patented core engine includes hundreds of mood variations as well as a complete emotional decision-making model based on our vocally-detected intonations. By listening and focusing on how people speak, rather than trying to understand what they say, our technology taps into a much stronger source of emotional information. Furthermore, since emotions are both universal and intuitive, Emotions Analytics solutions need no complex and cumbersome sets of rules and syntax to try and convert words into meanings.[4] – Beyond Verbal website, 2015

The realm of the voice and the realm of the affective share the distinction of the ineffable. Here, human instincts, raw flesh, autonomic reactions, sweat, nerves, animal chemistry, and gut reactions leave their marks in sound. Such expressions are imagined to transmit their effects to other sentient creatures, somehow bypassing language and touching our pleasure points, stirring our souls, or hitting us where it hurts, before we can make meaning of it. Digital listening has recently latched on at the intersection of these realms, aiming to evaluate -- and to predict -- a speaker's mood, personality, truthfulness, confidence, and mental health, based on algorithmic evaluations of the acoustic parameters of the voice. This article looks closely at the code, patents, and marketing language used by emerging affective listening software in order to consider the psychological and political values embedded in this technique of listening. What happens to the speaking, feeling subject when the listener is a computer? How do these algorithms imagine, and attempt to quantify, the human soul, and to what ends? The digital encoding of the affective realm reveals a collection of (man-made) guidelines for listening, which in turn lead to a prescription of what is *listenable* -- of what counts as legible and possible. I here consider which cultural and ethical values are deeply embedded in this

[3] Roland Barthes, "The Grain of the Voice," in *Image-Music-Text*, trans. Stephen Heath (New York: Hill and Wang, 1977), 188.

[4] "Emotions Analytics – Analyzing emotions from Vocal intonations," *Beyond Verbal Communication Ltd*, accessed July 31, 2015, http://www.beyondverbal.com/start-here/emotions-analytics/.

software, and what kinds of intersubjectivities this digitized listening might prescribe or foreclose.

Although theorists like Deleuze and Guattari often describe the affective realm as the most unquantifiable plane of experience, the affective *sciences* are actually all about counting feelings, and about which feelings count. In 1667, Spinoza put forth the definition of affect as a "passion of the mind" that is expressed only through the vitality of the body, not in language. Spinoza counted three such passions: desire, joy, and sorrow.[5] Psychologist Silvan Tomkins took up this idea again in the 1960s, increasing the count to nine, and theorizing universal bodily expressions particular to each one.[6] Sound figures prominently in his descriptions of affective communication, as a literal expression and as a metaphor (e.g.: the transmission of affect is called "affective resonance"). Bodies can transfer emotions without words, as packs of animals transmit fear amongst themselves or one crying baby can set a whole nursery wailing. Phenomenologist Max Scheler called this the "contagion of emotion."[7],[8] This gave rise in the 1970s to M.F. Basch's theory of "primitive empathy," which transmits the "raw data of emotion"[9] through wailing voices or nervous skin.

That this raw data could be described using discrete and limited categories meant that feelings could then be digitized – coded for the computer in terms of their quantitative affiliation with one or another itemizable affect (95% joy and 5% sorrow, for example.) These codification processes have flourished in the past quarter-century – in tandem with the rise of portable personal computing – leading to an outpouring of research and new technologies in the area of digital listening. Although most such tools claim to merely register in digital form an affective meaning in the voice that is "natural" and universal to the human body, I instead find that these technologies of encoding assert certain forms of recognition of the self and other that are based primarily on the capacities of the computer and the priorities of the market. Far from being modeled on

[5] Benedictus de Spinoza, *Ethics*. Trans. by W.H. White and A.H. Stirling. (London: Wordsworth Editions, 2001).

[6] Donald Nathanson, *Shame and Pride: Affect, Sex, and the Birth of the Self* (London: W.W. Norton, 1992).

[7] Donald Nathanson, "From Empathy to Community," *The Annual of Psychoanalysis* 25 (1997): 125.

[8] Some psychologists and affect theorists use "affect," "emotion," and even "feelings" interchangeably. Others do not, asserting that "affect" must be something that happens prior to consciousness or language, while emotions can be put into words and acted out, and both emotions and feelings can be the result of conscious reaction and memory. For the purpose of convenience within this article, I use the three terms interchangeably. The software discussed here frequently collapses the idea of the affective, as something beyond control (fear, pleasure), with ideas about feelings or emotions that are related to deliberation or psychological history (guilt, lies, depression.)

[9] Donald Nathanson, "From Empathy to Community," The Annual of Psychoanalysis 25 (1997): 126.

fundamental truths of the human body (if such truths exist), “the affective” has recently gained validity as a psycho-epistemological category in tandem with, perhaps because of, the rise of personal and predictive computing.

The affects gain traction because they are numbered, and therefore the affective is something a computer can handle. “Affective computing” emerged in the late 1990s and early 2000s as an area of technical research focused mainly on designing software that could identify and respond to human emotions based on the machine coding of facial gestures.[10] More recently, these practices have expanded to include listening software. Such projects quantify vocal expressions and claim to read their affective content according to a rubric that understands these signals as indexing entries in various libraries of affective and emotional labels. This gave rise to a new mode of listening: one that claimed to be both cybernetic and sympathetic. Our feelings become heard, recognizable – in fact, worthy of recognition – as they become perceptible to the computer.

But this is not – or ought not be – an easy task. Roland Barthes has written about the struggle to describe sound using linguistic codes. Language, he says, “manages very badly” at discussing music, and tends to fall back on that “poorest of linguistic categories,” the adjective.[11] Listening to music has the effect of reassuring and constituting the subject – culturally and relationally – and this effect gets expressed by the listener in adjectival terms. Sound is described with the most subjective vocabulary; using words that have no set quantitative or universal meaning: loud, soft, moving, violent, sweet, rich, harsh, etc. For Barthes, the “grain” of the voice is a site of escape from the “problem of the adjective.” The voice – especially the untrained voice – carries in it unintentional and inevitable traces of the individual body from which it emanates: the dimensions of the singer’s lungs, the flesh of his tongue, the shape of his teeth, etc. Although the same could be said for any expressive gesture issuing from the body, the voice is unmediated – travels directly from the mouth to the ear without the intervention of an instrument – and therefore puts the speaker’s (or singer’s) and listener’s bodies into an immediate, erotic relation. For Barthes, this grain is something beyond codification and culture – it evinces only the body; speaks from the body and to the body.

For affective listening software, uncontrollable, unintended, or habitual vocal inflections are imagined as signifying not just the body, but also the emotions, intentions, desires, fears, personality, and mental health of the speaker -- in short, the soul. These technologies listen for and quantify changes in pitch, timbre, volume, pacing – the “musical” and “granular” parameters of the voice – in order to ascribe affective meaning to these changes. Emerging and recent listening technologies like Nemesysco’s “Voice Risk Analysis,” Beyond Verbal’s “emotion analytics” software, and Cogito’s “Dialog” do

[10] Tieniu Tan, Rosalind Picard, et al, Preface to *Affective Computing and Intelligent Interaction*: *First International Conference, ACII 2005, Beijing, China, October 2005, Proceedings* (Berlin: Springer, 2005), i.

[11] Barthes, “The Grain …,” 179-180.

their listening in a range of contexts, from healthcare administration, to Artificial Intelligence, to financial investing. Benefits administrators use this software to screen claimants for both wellness and sincerity. Some health care providers use the software in call centers for diagnostic purposes, particularly to detect depression. Self-tracking mobile phone apps offer the user a description of her mood and its history. Customer-service call centers are now using "automatic dialogue systems" to detect if a speaker is angry or frustrated. Military training simulators are incorporating the software to measure stress levels. Human resources departments use it to weed out job applicants.[12] Finally, some recent studies have been directed at developing hyper-focused voice surveillance systems that alert the authorities when tensions are running high.

By and large, the evolution of these products reveals a broader techno-cultural shift: from truth to prophecy. Although these designs derive originally from lie-detection technologies, most contemporary affective listening products are couched in the languages of prediction and control: self-tracking, targeted marketing, investing, and risk-management. Companies claim to offer high "ROI"s[13] by detecting the investment-worthiness of a CEO, the likelihood of a claimant to benefit from rehab, when a user will become depressed or anxious, whether a worker will perform well, and even a speaker's "illegal intentions." Accuracy is not the goal here. Rather than proving an existing fact, these tools are instead focused on describing a field of psychic possibility. Such products "work" by providing their users with probabilities that allow them to better invest their time and money. Arjun Appadurai's explanation of the risk economy hinges in part on Weber's definition of magic as "some sort of irrational reliance on any sort of technical procedure, in the effort to handle the problems of evil, justice, and salvation."[14] Indeed, these technologies attempt to do something magical – to know the unknowable, to quantify the uncertain: to see into the soul in order to predict the future.

What does magic have to do with ethical listening? Barthes makes the connection clear: "the musical adjective becomes legal whenever an *ethos* of music is postulated, each time, that is, that music is attributed a regular - natural or magical - mode of signification."[15] Ethics and politics enter into listening when the sound is described using a language that is given a general and repeatable meaning, thereby both limiting, and making accountable, its expressions. In this case, the encoding of the affective realm in the voice is the moment where cultural values get embedded into these technologies. The feelings in the sounds are evaluated and named: stressed, tired, deceptive, passive, happy,

[12] Aarti Shahani, "Now Algorithms Are Deciding Whom to Hire, Based on Voice," *NPR All Tech Considered*, March 23, 2105, http://www.npr.org/sections/alltechconsidered/2015/03/23/394827451/now-algorithms-are-deciding-whom-to-hire-based-on-voice.

[13] "Return On Investment"

[14] Arjun Appadurai, *The Future as Cultural Fact: Essays on the Global Condition* (London: Verso, 2013), 242.

[15] Barthes, "The Grain …," 180.

angry, anxious, etc. Insofar as ethics concerns the recognition of the other[16], the rubrics of recognition used by these technologies prescribe their ethical vocabulary. They show us their model of the human soul, and in this model, show us what structures of feeling "count" as recognizable, or worthy of recognition. The translation from sound to language is where the magic happens, and where the ethics of these technologies comes into view.

A second *ethos* surfaces when these technologies meet the market. The products – regardless of their relationship with the natural, rational, or magical – have efficacy insofar as they are being used. Their applications retroactively prescribe the possible meanings that they assign to certain vocal qualities. A comparative study of the patents and early available open-source code of these technologies reveals common methods of extracting data from the vocal signal, but a great variety of ways of assigning meaning to that data in the early phases of ideation and development of these products. Not surprisingly, what these technologies claim to listen for changes as they hit the market. Regardless of the rather diverse range of psychological models proposed in early documentation of these technologies, their applications are remarkably similar. Affective listening software that are being marketed today all are coalescing around a few standard uses: namely, prediction and risk management, in the realms of benefits administration, labor relations, financial investing, and surveillance.

Why Voice?

Affective listening works on the most unaccountable realm. The affective voice, insofar as it is conceived of as prior to reasoning or language, also is beyond proof, beyond trace, and beyond fact. Affective computing is about partitioning, labeling, and then quantifying something that is, by its own definition, mysterious, unwieldy, and unpredictable. Affective listening software, therefore, claims to access and decode the *most* unknowable: the soul and the future. These technologies started as lie-detecting devices, based on the idea that the voice registered unintentional indications of discomfort, nervousness, and guilt. Instead of detecting autonomic responses to stimuli, this technology actually detected (or tried to detect) uncontrollable expressions of intention – to reveal to the listener truths about the subject that s/he did not want known.

The voice has become a particularly desirable site from which to extract affective meaning because it has the unique quality of carrying traces of the body while existing outside of it. The voice is at once public and private, intentional and unintentional. Speech is released into the airwaves with intent and meaning. Yet the non-linguistic (the grain, the affected) qualities of the voice are attributed to the interior, private, and

[16] Emmanuel Levinas, *Totality and Infinity: An Essay on Exteriority* (1961), trans. A. Lingis. (Pittsburgh, PA: Dusquesne University Press, 2007).

preconscious: to internal organs, to involuntary reactions, to hidden feelings and secret desires.

One of the earliest patents for "Quantifying Psychological Stress Levels Using Voice Patterns" explains that "[t]he advantage of voice analysis over the polygraph is that the subject being tested does not have to be physically connected to the device and thus voice analysis is a non-invasive method of lie detection or truth verification."[17] In Western civilizations, the voice is imagined as a portal to the soul – it gives access to the interior self without requiring any invasion of privacy or breach of skin.

A great deal of research on the measurement of affect in the voice has proliferated in more recent years. In 2003, *Speech Communication* journal devoted a special issue to "Speech and Emotion," which was the result of a workshop sponsored by the International Speech Communication Association a few years earlier. Since 2010, there has been a marked surge in vocal affective computing research. Bjorn Schuller et al trace "some spurious papers on recognition of emotion in speech during the second half of the 90s (less than 10 per year), a growing interest until 2004 (maybe some 30 per year), and then, a steep rise until today (>100 per year)."[18]

Affective sciences assert that while human linguistic communication varies by culture, emotion may be expressed via purely acoustic means that are universal and transcend words. Psychologist Klaus Scherer published some of the early contemporary comprehensive reviews and significant scientific literature on emotion in the voice. His work starts from the thesis that since speech is necessary to human survival, the human voice, like an organ or muscle, has evolved over generations and generations to communicate emotions in a universally human way.[19] Scherer's theoretical framework builds on Rousseau's, Herder's, and Helmoltz's suggestions that early speech and music evolved from reflexive affective expressions that have some commonality across cultures, such as moans, cries, and expressions like "ow," "oh," etc. Scherer links the evolution of the brain with the evolution of spoken and musical languages, and considers emotional vocal expression to belong to a "more primitive, analogue … affect signaling system."[20]

Thus, the "musical" aspects of vocalization such as pitch, timbre, and timing can now serve semantic functions as well as affective ones. Scherer claims that emotions shape vocal expression in such a way that a listener can correctly determine and resonate with emotion from a voice. "That the human voice not only permits judging in the speaker's emotion but can also *induce* affect in the listener has been held as self-evident

[17] Charles Humble, "United States Patent 7571101 B2 – Quantifying Psychological Stress Levels Using Voice Patterns," United States of America, August 4, 2009, 13.

[18] Björn Schuller, Anton Batliner, Stefan Steidl, and Dina Seppi, "Recognizing realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication* 53 (2011), 1063.

[19] Klaus R. Scherer, "Expression of Emotion in Voice and Music," *Journal of Voice*, 1995, 235.

[20] Scherer, "Expression of Emotion …," 236.

throughout history. In particular, ever since antiquity, different schools of rhetoric have insisted on the powerful effect of emotional expression in the voice on the listener (Cicero, Quintilian)."[21] If we believe in this inductive power of the voice, what does it mean when a computer is the listener? How could an algorithm be induced?

The Signal in the Grain: How They Listen

The first step to quantifying affect in the voice involves measuring the vocal signal itself, and deciding what aspects of the voice are indicative of affect. Affective listening research from the 1970s-1990s lays the groundwork for contemporary products. This research mainly correlates emotions with pitch, volume, rhythm, and sometimes timbre. Anger is linked to an increase in average fundamental frequency ("F0") and average intensity (i.e., high pitched and louder.) Fear was also associated with an increase in mean F0 and with an increase in rate of articulation (higher pitched and faster.) Joy correlated with an increase in mean F0, in a wider range of F0, and in greater intensity (higher-pitched, more animated, louder.) Sadness, on the other hand, correlated with a decrease in mean F0, F0 range, and intensity (lower-pitched, more monotonous, slower speech.)[22] Early masking studies broadly showed that "arousal and uncertainty … seemed to be communicated by F0 variability and F0 mean, respectively."[23] Very early work on speech synthesis found that the "tempo of the sounds in the sequence and filtration level (i.e., number of audible higher harmonics) were by far the most powerful cues" toward signaling an emotion.[24] Short bursts of sound are correlated with joy, whereas longer durations or slower tempos are affiliated with sadness.[25]

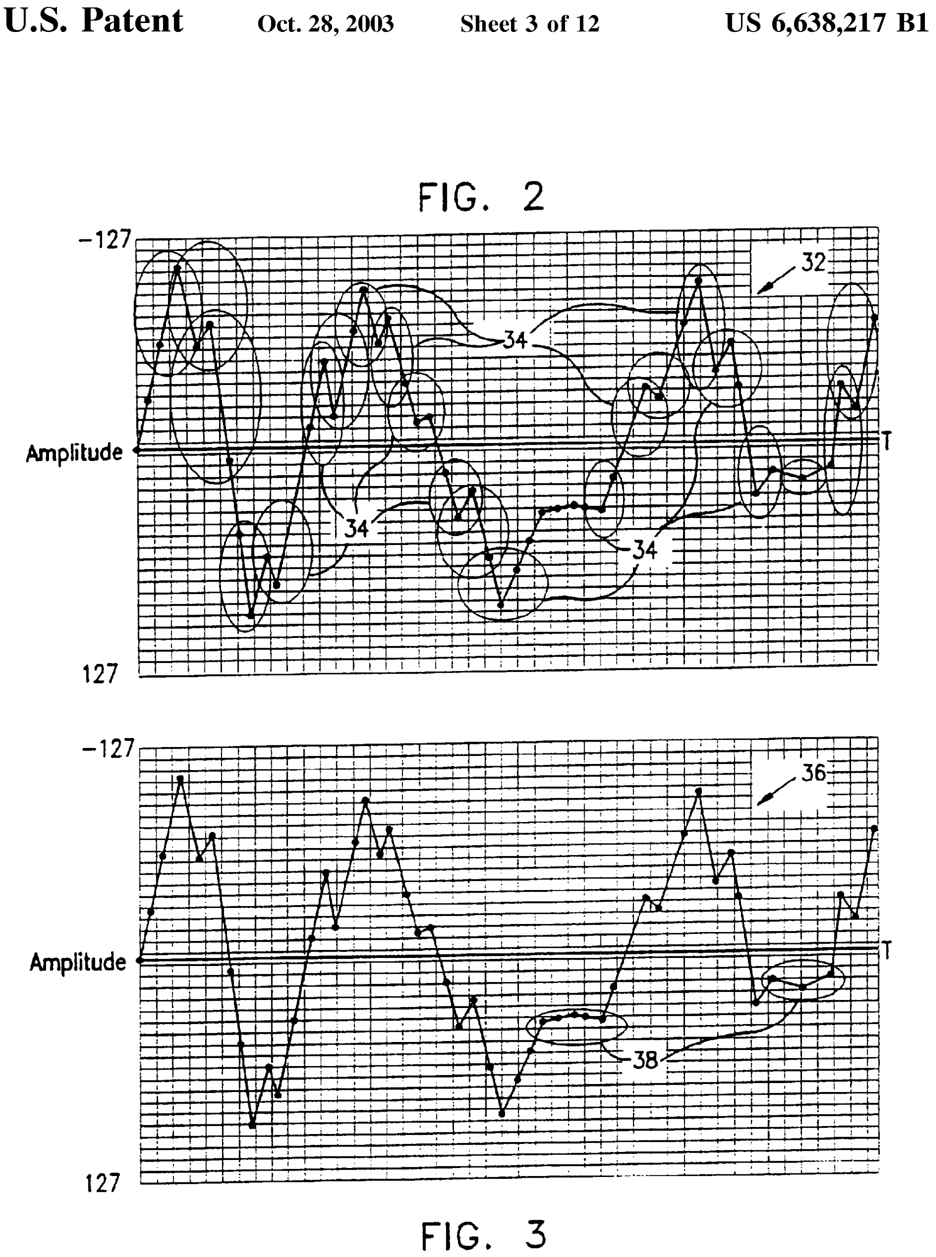
One of the earliest industry applications of this was Voice Stress Analysis (VSA), developed in the 1970s-1980s. VSA claimed to detect inaudible, very brief micro-tremors in the voice in the 8-14 Hz range, produced unintentionally by the body under psychological stress. For a time, Voice Stress Analysis was used in interrogation scenarios to determine if a subject was lying, although it recently has been widely criticized as unreliable. Regardless, VSA methodology is more or less reincarnated in some of the early patents for affective listening software by Nemesysco, an Israeli company founded in 2000. While the company is careful to distance itself from VSA, early patents reveal many similarities. All of their products are based on Layered Voice Analysis (LVA),[26] their patented voice analysis software that monitors the voice for

[21] Scherer, "Expression of Emotion …," 237.
[22] Scherer, "Expression of Emotion …," 241.
[23] Klaus R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication* 40 (2003), 239.
[24] Scherer, "Vocal communication …," 239.
[25] Scherer, "Vocal communication …," 240.
[26] "What can Voice Analysis do for you?" *Nemesysco Ltd*., Accessed May 20, 2016, http://www.nemesysco.com.

emotional arousal. A 2003 patent awarded to the company's founder posits a connection between micro-tremors in the voice and the sincerity of the content of the speech.

The patented voice analyzer focuses on pitch and volume variations and generates information regarding the individual's "excitement level" based on "thorns" in the vocal waveform and the length of plateaus in this waveform.[27] A "thorn" in this case is a sudden spike or dip in the volume of the sound, and a "plateau" is a "local flatness" in the sound wave.[28] The system requires a sound sample of 0.5 seconds of continuous speech, from which a variable number of windows are delineated. Within these windows, the wavering of the voice is analyzed on a micro-level. The frequency and distribution of thorns and the frequency and length of plateaus is used to determine the subject's emotional state.[29] Ideally, a baseline of "neutral" vocal activity is previously established for the subject, so that samples can be compared to a personalized neutral profile.

[27] Amir Liberman, "United States Patent 6638217 B1 – Apparatus and Methods for Detecting Emotions," United States of America, October 28, 2003, 1.
[28] Liberman, 14-15.
[29] Liberman, 17.



U.S. Patent Oct. 28, 2003 Sheet 3 of 12 US 6,638,217 B1

FIG. 2

FIG. 3

Figure 1: graph of vocal "thorns" from Nemesysco's first patent.[30]

Within a few years, a handful of patents were filed by other companies, offering similar, but more nuanced methods of measuring the vocal signal. A 2008 patent application by the Boston-based company Cogito proposes to generate information on a

[30] Liberman, 4.

psychiatric clinical rating scale based on a 30-millisecond-long window of a phoned-in voice sample, from which the software extracts measurements of frequency, spectral energy (color), values and locations of largest peaks and troughs, number of peaks, energy (volume), and time-derivative of energy (envelope, dynamics). The system also measures average length of voicing, average length of speaking, fraction of time speaking, voicing rate, average number of speaking segments, and the entropy of speaking and pause lengths (pacing, timing, rhythm in conversation).[31]

The same year, AGI, a Tokyo-based company established in 1999, was awarded a patent for an AI prototype for sympathetic listening. The patent describes "an emotion detecting method capable of detecting human emotion accurately, and … [a] sensibility generating method capable of outputting sensibility akin to that of a human."[32] This is accomplished by measuring, in the human's voice, volume, tempo, spectral qualities, energy distribution within a word (dynamics of volume), and voiceless time (rhythm, speed, pauses).[33] Instead of listening for unintentional micro-tremors, this patent proposes listening over a ten-second long time period.

The next year, 2009, The SEMAINE project[34] released a report of its research towards designing SAL, a "Sensitive Artificial Listener." One outcome of this research is openSMILE technology, an open-source "speech analysis technology" for determining emotion.[35] SEMAINE's research attempted to write algorithms to automatically associate vocal measurements with labels from varying corpuses of emotional speech, using 384 statistical functionals applied to low-level speech descriptors.[36],[37] These low-level descriptors are basically very nuanced ways of evaluating pitch, volume, color, melodic range and activity, and dynamic activity of the voice.

[31] Vikram S. Kumar and Jonathan Jackson, "United States Patent Application Publication 2008/0234558 A1 – Method and Systems for Performing a Clinical Assessment," September 25, 2008, 11.

[32] Shunji Mitsuyoshi, "United States Patent 7340393 B2 – Emotion Recognizing Method, Sensibility Creating Method, Device, and Software," United States of America, March 4, 2008, 1.

[33] Mitsuyoshi, 1.

[34] The SEMAINE project is a research project that is publicly funded by the EU, and staffed by a collaboration of researchers from different European Universities.

[35] "openSMILE:) The Munich Versatile and Fast Open-Source Audio Feature Extractor," *audEERING, GmbH*, accessed May 20, 2016, http://audeering.com/research/opensmile/.

[36] These vocal signal qualities included: signal frame energy (volume), MFCC (pitch spectrum analysis modeled off human hearing), zero-crossing rate of time signal by frame, probability of voicing, fundamental frequency, max value of pitch contour, min value of pitch contour, range, position of min and max values, arithmetic mean of the pitch contour, slope of a linear approximation of the contour, offset of the linear approximation of the contour, quadratic error of the linear approximation, standard deviation of the contour, skewness, and kurtosis.

[37] Björn Schuller et al, "The INTERSPEECH 2014 Computation Paralinguistics Challenge: Cognitive & Physical Load," In *Proc. INTERSPEECH 2014*, 15th Annual Conference of the International Speech Communication Association, ISCA., Singapore, 2014. http://www.interspeech2014.org/public.php?page=home.html.

A few years later, in 2011, the founder of Beyond Verbal, an Israeli "emotion analytics" company, was awarded a patent for a "computerized voice-analysis device for determining an S,H,G profile."[38] Their system takes 10-20 second samples of the voice and breaks them up into 2.5 second long segments, which it uses to evaluate pitch spectra and fluctuation, and to relate them to psychological qualities. G-values ("growth") are associated with lively speech and frequent oscillation in the 400Hz - 600Hz range. H-traits ("homeostasis") are indicated by "relatively stable voice intensity at the lower sound frequencies (300-600 Hz) A high S-value ("survival") is displayed by "sharp changes in intensity at 600-800 Hz."[39]

Another patent awarded to the company proposes to evaluate a speaker's emotion based on studies of animal sounds, which showed that specific tones were associated with particular activities (mating calls, aggressive roars before fighting.) The patent goes on to propose that "every emotional center in the brain is associated with a certain tone, and vice versa, so that whenever a center is active and initiates a verbal emotional response, the tone that is associated with that active center in the brain is expressed together with the verbal response."[40]

The patent proposes certain pitch patterns, labeled by solfege syllables, which are "considered the normal intonation for pronouncing a certain word."[41] Particular pitch spectra and patterns are associated with certain characteristics and feelings, by dividing an octave into seven tones, according to the C-major Scale, and correlating each pitch with particular emotional significance. Furthermore, the design also listens for intervals in speech (rising minor 2nd, falling perfect 4th, etc.) and correlates them with moods. The patent then proposes an algorithm for evaluating a subject's mental state by having them speak words that "carry emotional value"[42] into the software.[43] The voice recording is then analyzed for the five most prevalent pitches in the word, and these pitches are compared to a database containing "the norms" for the pitches associated with that word. The speaker's particular pitch array is used to affiliate her with a certain emotional state, and any deviations are noted as potentially indicative of instability.[44]

Taken as a whole, we can notice some small differences in the listening techniques used by these various projects. Generally, software that listens particularly for

[38] Yoram Levanon and Lan Lossos-Shifrin, "United States Patent 7917366 B1 – System and Method for Determining a Personal SHG Profile by Voice Analysis," United States of America, March 29, 2011, 1.

[39] Levanon and Lossos-Shifrin, 7.

[40] Yoram Levanon and Lan Lossos, "United States Patent 8078470 B2 – System for Indicating Emotional Attitudes through Intonation Analysis and Methods Thereof," December 13, 2011, 1l.

[41] Levanon and Lossos, "United States Patent 8078470 B2 …", 10.

[42] Levanon and Lossos, "United States Patent 8078470 B2 …", 12.

[43] In English, these words can include love, hate, happiness, mother, father, baby, dream, jealousy, anger.

[44] Levanon and Lossos, "United States Patent 8078470 B2 …", 13.

unintended or "hidden" meanings (lies, stress, etc.) looks at very short samples of speech, based on the assumption that the voice cannot be deliberately controlled on the micro-level. Software aiming to measure intended emotional meaning or personality type looks at a larger sample of speech before rendering a label. Beyond Verbal's patents, although claiming cross-species universality, display the closest relationship to the Western tonal language. Overall, however, at this low-level of design, the listening is fairly similar: all the technologies evaluate the voice for standard "musical" parameters, with emphases on fundamental frequency, amount of pitch fluctuation, rhythmic activity (stops and starts), volume, and intervallic diversity.

From Signal to Language: What They (Want To) Hear

While these emergent technologies seem to accord meaning to similar aspects of the voice, the meanings that they assign differ substantially based on each product's imagination of the human emotional structure. At this level of the design, we start to see the intersection of varying theories of human nature and the imagined applications of the technologies. These theories are in part related to what the technology wants to hear: motivating drives (for advertising), unconscious discomfort (for lie-detection and investment planning), and mental health tendencies (for benefits administration).

Although predicated on decades of research in the affective sciences, the software is not truly concerned with affect *per se*. Of the twenty-two of patents I surveyed over the course of my research, many of which serve as templates for products marketed by the five major companies, *none* of them attempt to measure affect alone. The products listen for indications of lies, personality traits, depression, anxiety, confidence, and mood swings. They do so in a way that presumes that, like affect, such "passions of the mind" as these are expressed in universal (or at least culturally generalizable) terms.

The technologies differ here in their relationship to individuality: some, like Beyond Verbal's, are based on the premise of universal tunings for the expression of personality traits. Similarly, SEMAINE's products seek to line up vocal patterns with a databases of emotions derived from wide surveys of populations. Others, like Cogito's mental health apps, attempt to customize the correlations between vocal expression and feelings by learning the particular tendencies of an individual speaker.

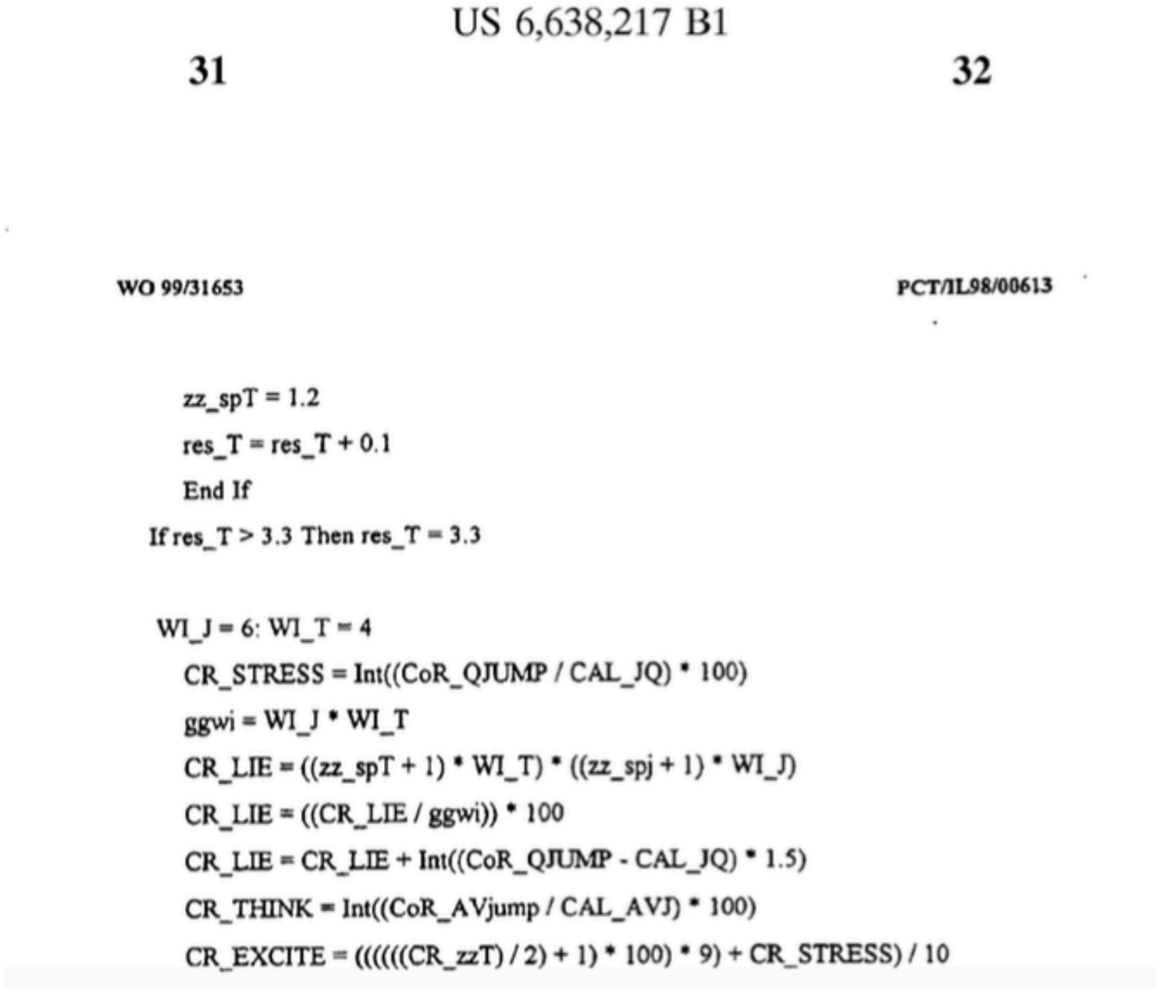
Nemesysco's products work according to a fairly strict discrete emotional theory.[45] The products vary in their theories of universality, however, based on whether

[45] The three competing models of emotion, which are employed by various technologies covered in this article, all cut up the emotional domain in different ways, leading to different rubrics for describing the speaker. *Discrete emotion theories* divide up the emotions into 6-14 basic emotions. The *dimensional approach* to emotions, on the other hand, maps the emotions into two- or three-dimensional spaces: a valence dimension (pleasant-unpleasant, agreeable-disagreeable), an activity dimension (active-passive), and sometimes a power or control dimension. Finally, *componential models* of emotion are based on the subject's appraisal of a real-life situation to which s/he is reacting. Componential models digress from dimensional approaches in that they do

the product will have the opportunity to "learn" a speaker or not. Products marketed to law-enforcement for interrogation purposes take a universal approach to human expressions, assuming that the voice performs similarly in all lying (strangers') bodies. LVA classifies the speaker into one of nine basic emotional categories, which are then used to evaluate the subject's "Lie Stress" (or "Risk Level"), "Arousal Level", "Attention Level", and "Deception Patterns."[46] The commercially available version of LVA is designed so that it can be "personalized" by establishing a baseline for micro-high and -low frequencies for a given subject's voice, and then monitoring the voice for deviations from these norms, according to parameters like "stress," "confusion," "thinking level," "concentration," "anticipation," "embarrassment," "arousal," and whether or not s/he is withholding information. The software also differentiates between jokes, white lies, embarrassment lies, offensive lies, and defensive lies.

not focus solely on the subjective feeling, and also part ways with discrete theory in that they do not believe in a limited number of basic emotions, which are related to neurological patterns. Instead the componential model emphasizes variability of emotional state within a limited number of emotional families.

[46] "LVA Analysis Process," *Nemesysco Ltd.*, accessed May 20, 2016, http://nemesysco.com/speech-analysis-technology.



US 6,638,217 B1

31 32

WO 99/31653 PCT/IL98/00613

```
      zz_spT = 1.2
      res_T = res_T + 0.1
      End If
   If res_T > 3.3 Then res_T = 3.3

    WI_J = 6: WI_T = 4
      CR_STRESS = Int((CoR_QJUMP / CAL_JQ) * 100)
      ggwi = WI_J * WI_T
      CR_LIE = ((zz_spT + 1) * WI_T) * ((zz_spj + 1) * WI_J)
      CR_LIE = ((CR_LIE / ggwi)) * 100
      CR_LIE = CR_LIE + Int((CoR_QJUMP - CAL_JQ) * 1.5)
      CR_THINK = Int((CoR_AVjump / CAL_AVJ) * 100)
      CR_EXCITE = ((((((CR_zzT) / 2) + 1) * 100) * 9) + CR_STRESS) / 10
```

Figure 2: Code from patent issued to Nemesysco's founder, Amir Lieberman, in 2003. The code here shows the formulae for determining Stress Level, Lie Level, Thought Level and Excitement Level from variables in the voice signal.[47]

Cogito's products don't look for lies; instead they evaluate health and well-being according to various already-existing mental and physical health rating schema. The computer performs a clinical assessment of the patient from the voice signal, and calculates the subject's emotional state and likelihood of recovery. Although the technology works with universalizing scales of well-being, the listening itself is personalized, using an algorithm that takes into account broader variances and past data from clinical assessments,[48] combined with an evaluation of a 30-millisecond phoned-in voice sample. Audio features extracted from this sample are used to estimate the patient's health according to the PHQ-9 depression score, visual analog scale for pain, APGAR score of neo-natal health, or HAM-D (another depression rating scale).

[47] Liberman, 6.

[48] Kumar and Jackson.

Signal data is connected to diagnoses based on a process of linear regression, wherein the machine "learns" an order of medical evaluations and potential diagnoses to follow in order to most accurately label the patient's mental state. Training data for the first phase is obtained by asking the patient to repeatedly perform a self-assessment of mental health by speaking into the phone. The acoustic signal information is then correlated with the patient's self-reporting of mental health, and used to train the machine for future evaluations.[49]

Beyond Verbal's products, on the other hand, originally were designed to detect a speaker's fixed "personality type," for the purposes of more effective, customized advertising. Their products are built on their theory of three fundamental, unconscious drives that motivate all humans and determine their decision-making tendencies, based on individual brain chemistry.[50] Activity or stability in certain frequency ranges is aligned with the speaker's decision-making tendencies[51], called their "S,H,G Profile." The "S,H,G Profile" is an invention of the company's, and stands for survival, growth, and homeostasis (or relaxation.).[52] Survival is defined as "the willingness of an individual to fight for his or her own survival and his or her readiness to look out for existential threats, [and is] … driven by the secretion of adrenalin and noradrenalin." Homeostasis is "the extent to which an individual would prefer to maintain his or her 'status quo; in all areas of life, [and is] … driven by the secretion of acetylcholine and serotonin. Growth is "the extent to which a person strives for personal growth in all areas (e.g. spiritual, financial, health, etc.) [and is] driven by the secretion of dopamine."[53] Beyond Verbal believes that the pitch and volume fluctuations in the voice can be indicators of these predispositions and can be used to determine a potential consumer's motivations. Their system evaluates the voice and diagnoses the speaker's level of affinity with six personality types (SH, SG, GH, GS, HS, HG).

Beyond Verbal's more recent innovations are concerned less with individual drives than with understanding a universal human intonation code. The inventors insist that -- even in the event that the speaker is conscious of their feelings – vocal inflections are not under their control. Their 2011 patent includes rubrics for relating both vocal *pitch* and *intervallic patterns* to emotional attitudes. For example, the pitch of C is associated with "the need for activity and survival," whereas E is associated with "self control" and B with "command and leadership."[54]

[49] Kumar and Jackson, 15-16.

[50] Yoram Levanon and Lan Lossos-Shifrin, "United States Patent 8249875 B2 -- System and Method for Determining Person SHG Profile by Voice Analysis," United States of America, August 21, 2012, 3.

[51] Levanon and Lan Lossos-Shifrin, "United States Patent 8249875 B2 …", 6.

[52] Levanon and Lan Lossos-Shifrin, "United States Patent 8249875 B2 …", 3.

[53] Levanon and Lan Lossos-Shifrin, "United States Patent 8249875 B2 …", 6.

[54] Levanon and Lossos, "United States Patent 8078470 B2 …", 11.

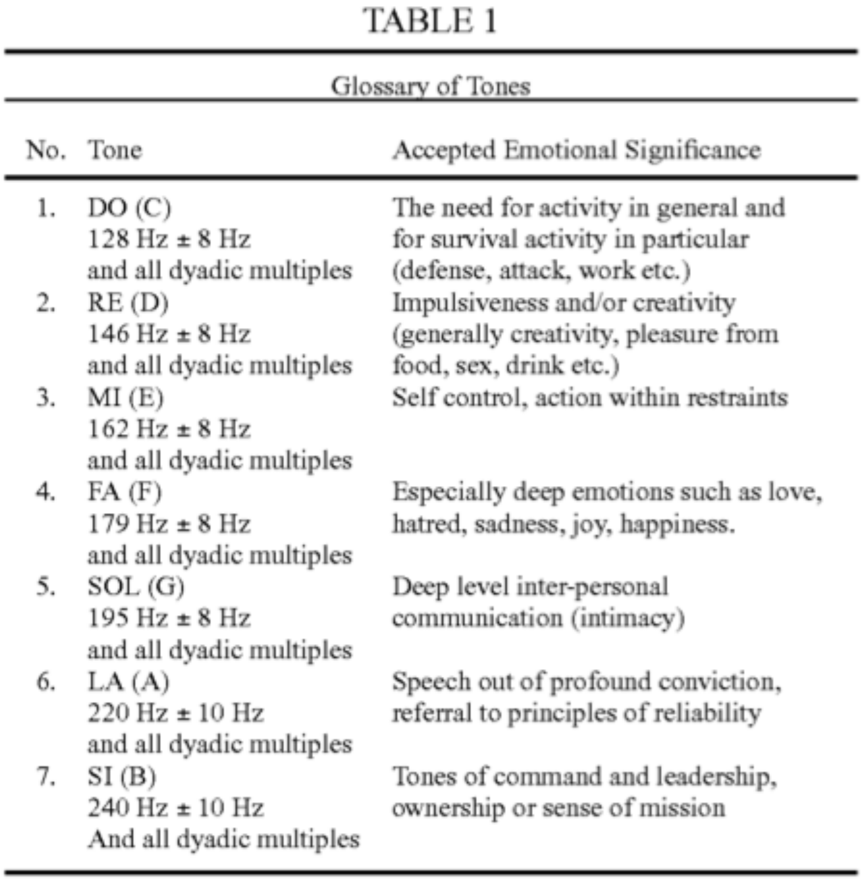


TABLE 1

Glossary of Tones

No.	Tone	Accepted Emotional Significance
1.	DO (C) 128 Hz ± 8 Hz and all dyadic multiples	The need for activity in general and for survival activity in particular (defense, attack, work etc.)
2.	RE (D) 146 Hz ± 8 Hz and all dyadic multiples	Impulsiveness and/or creativity (generally creativity, pleasure from food, sex, drink etc.)
3.	MI (E) 162 Hz ± 8 Hz and all dyadic multiples	Self control, action within restraints
4.	FA (F) 179 Hz ± 8 Hz and all dyadic multiples	Especially deep emotions such as love, hatred, sadness, joy, happiness.
5.	SOL (G) 195 Hz ± 8 Hz and all dyadic multiples	Deep level inter-personal communication (intimacy)
6.	LA (A) 220 Hz ± 10 Hz and all dyadic multiples	Speech out of profound conviction, referral to principles of reliability
7.	SI (B) 240 Hz ± 10 Hz And all dyadic multiples	Tones of command and leadership, ownership or sense of mission

Figure 3: Table from Beyond Verbal's 2011 patent, describing the relationship between a pitch in the voice and the speaker's mental state. (The meaning of the tone applies to any octave, as indicated by "±8 or 10 Hz and all dyadic multiples.")

Furthermore, intervalic patterns in the pronunciation of a word are considered indicative of certain moods: FA-SOL (rising minor 2nd) signifies "emotional communication," SOL-RE (falling perfect 4th) signifies "communication and impulsiveness" and RE-FA (rising minor 3rd) signifies "a combination of impulsiveness and emotion," and so on.[55]

AGI's technologies were originally designed not only for machine recognition of affect, but to help machines act with "human sensibility" by mimicking the affect in the voice of their user. The software described by AGI's patents processes the voice in real-time and offers a nuanced, color-coded display of the speaker's emotions, rather than an emotional category or numeric score. The system learns particular users, and the labeling process is adjusted based on information about the individual's personality, cultural context, and habits of speech. In these early designs, AGI is listening for different things in the sound than many other companies: in addition to instinctive reflexes, the

[55] Levanon and Lossos, "United States Patent 8078470 B2 …", 11.

technology also acknowledges that traces of reason and individual histories influence emotional reactions. Because the goal of this invention is conversation rather than surveillance or persuasion, it is more interested in the speaker's intended meaning than in the speaker's hidden motivations.

The real difference in this design, however, is in how the emotions are assigned to the sounds. AGI's patents describe an algorithm that sends information about the vocal signal through three processing units. The signal data first goes through an "instinct information generating unit," which evaluates the speaker's degrees of certainty, pleasure, danger, attention/refusal, achievement/change, and follow-up/assertion.[56] The "instinctive state-profile" of the speaker is then sent to the "Emotion Information Generating Unit," where emotional evaluations of the vocal signal are coupled with the instinctive state information, the pre-stored "individuality information," and information about the speaker's prior emotional state and "emotional rhythm" in order to generate an "emotional parameter," which locates the speaker's emotions within a multi-axis space. This resembles dimensional models of the emotions, but contains greater numbers of fields located within two main emotional axes: pleasure/unpleasure and attention/refusal.

[56] Shunji Mitsuyoshi, "United States Patent 7340393 B2 …", 9.

Fig. 10

EXAMPLE OF REACTION PATTERN MODEL
IN EMOTION REACTION PATTERN DB

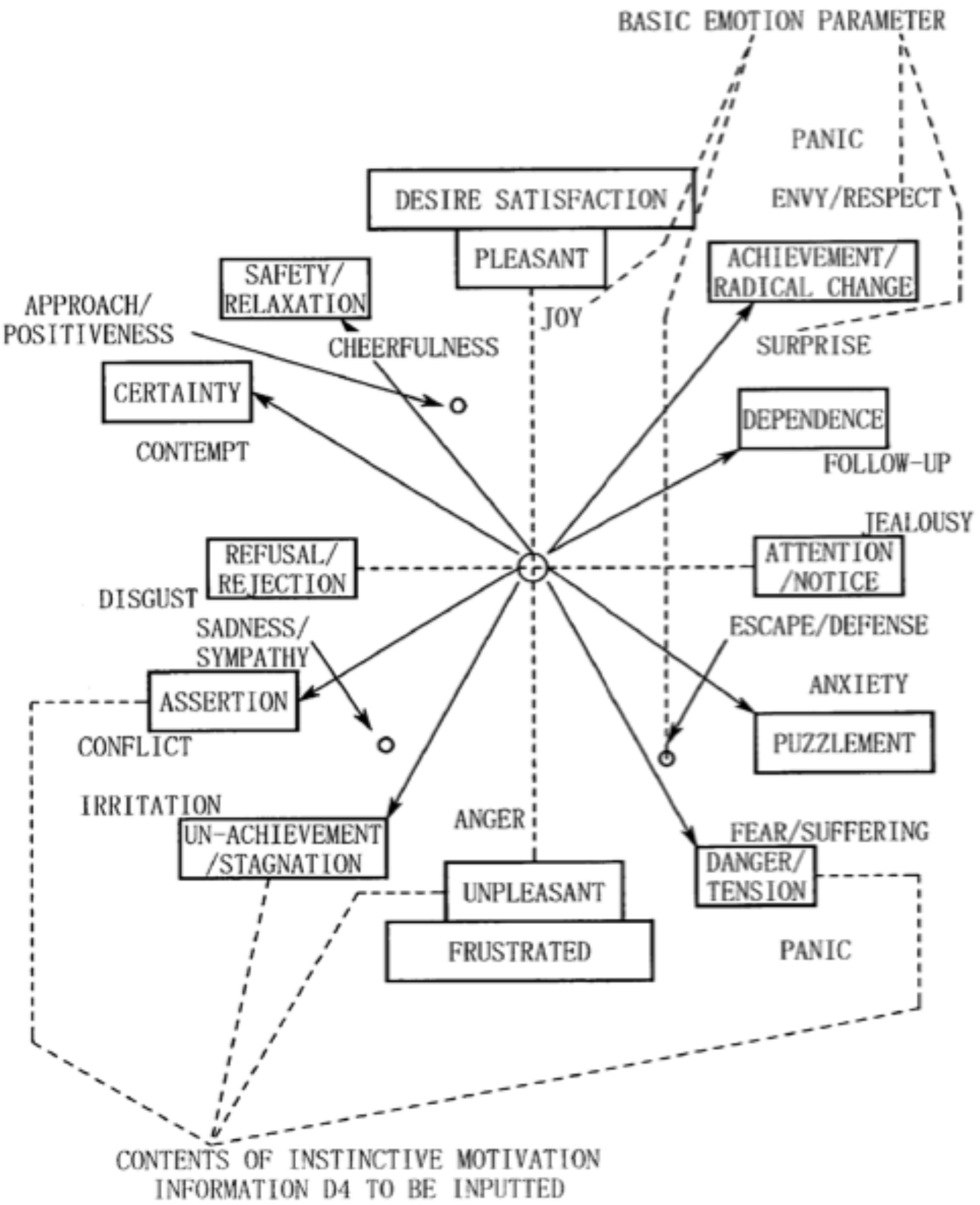


Figure 4: Drawing of the human emotional possibilities on their axis, from AGI's 2008 patent.[57]

[57] Shunji Mitsuyoshi, "United States Patent 7340393 B2 …", 10.

Finally, the emotional label is combined with individual personality information and contextual information and sent to "Sensibility and Thought Recognition Unit," where it incorporates cultural and contextual information (for example, it draws on a "moral hazard" database), and outputs information that then is used to direct the virtual human toward an appropriate, sensitive affective response.

A later patent, awarded in 2012, elaborates on this model to propose a computer that with its "own will and ego … having a human-like psychological state." The goal of this design is to provide "a heart-to-heart" communication system within an HCI [Human-Computer Interaction] … or to better enable communication between humans with language differences "by utilizing sympathetic vibration and resonance of emotion and mentality."[58] This process is modeled on the phases of psychoanalytic listening proposed by midcentury post-Freudian Wilfred Bion, who partitioned the analytic hour into a handful of thought stages ("attention," "inquiry," "definitory hypothesis," and more.[59]) The software listens to the voice for a range of mental levels articulated by Bion, which represent degrees of attention, focus, consciousness, and "primitive thought to abstract thought."[60] Finally, drawing on the theories of Melanie Klein, the machine assumes that humans fluctuate between paranoid-schizoid states and depressive states, especially when anxious.[61] Therefore the system listens to the voice for indications of this and anticipates such fluctuation.

In more recent years, AGI has proposed technologies that are marketed for lie detection and advertising. Their method is the result of a combination of the biological/reflexive model of vocal expression and a culturally-coded reading of emotion. Echoing their competitors' philosophies, AGI's publications explain that "voice movement like freezing, trembling, or stammering" are involuntary indications of "brain emotional activity."[62] Their labeling process is more nuanced, however. AGI's researchers selected 4500 emotional descriptors from the dictionary, narrowed them down to 223 labels, and grouped them into 4 quadrants, which are associated with 4-5 colors. These quadrants are mapped onto a double axis of positive/negative emotions and survival/breeding drives. In order to align emotions with voice patterns, 100 subjects affixed emotional labels to 2800 voice samples. In tandem to this, researchers used

[58] Shunji Mitsuyoshi, "United States Patent 8226417 B2 – Will Expression Model Device, Psychological Effect Program, and Will Expression Simulation Method," United States of America, July 24, 2012.

[59] See W. R. Bion, "The Grid," in *The Complete Works of W. R. Bion*, Ed. Chris Mawson (London: Karnac Books, 2014).

[60] Shunji Mitsuyoshi, "United States Patent 8226417 B2 …", 18.

[61] See Mélanie Klein, "Notes on some schizoid mechanisms," in *Envy and Gratitude and Other Works 1946-1963* (London: Hogarth Press, 1975.)

[62] Shunji Mitsuyoshi, "The application of emotion and mind recognition in voice Quantitative measurement technology of emotion," *The University of Tokyo, Agi Inc. and PST Inc.*, March 16, 2012, accessed May 20, 2016, http://www.agi-web.co.jp/docs/AGI-Eng_ver2012-04-18.pdf, 19.

fMRIs, pulse sensors, and blood tests to monitor the excitement and stress levels of the body, simultaneously tracking which areas of the brain were stimulated during speech. This somatic data was also used to determine emotion, mental condition, and "lie parameters," and used to label indicators of such qualities in the vocal signal.[63]

The SEMAINE project, like AGI, began with the aim of creating more affectively sensitive and convincing Virtual Humans. Unlike AGI, SEMAINE's SALs [Sensitive Artificial Listeners] do not aim to reflect the emotional state of the user, but to cause the user to reflect the computers' mood. SALs sense the user's emotional state and express their own, unchanging one, which the user is pulled towards over the course of a conversation.

SEMAINE uses an emotional rubric that is a combination of discrete emotion theory and dimensional evaluation models. A series of speech emotion recognition research challenges have yielded a range of libraries for the software. The 384 statistical qualities of the voice have been mapped onto eleven emotional classes, including: joyful, surprised, emphatic, helpless, touchy/irritated, angry, motherese, bored, reprimanding, rest, and neutral. "Level of Interest," is evaluated via a linear regression process, which continually monitors attention in the voice throughout a conversation. The voice is also evaluated for the five OCEAN personality dimensions (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) and for social signals (laughter, etc.), conflict, and autism. The vocal qualities associated with these emotional categories are both "heard" by the software, and used to generate emotional speech by the virtual humans.[64]

At this level of the design, we start to see the intersection of the products' varying theories of human nature, and their imagined applications. This is where a great deal of the ethics of the technologies are inscribed: the way in which the other can be heard – that is, what the software is willing to hear -- is formed, as the possibilities for the speakers' emotions, affects, and personalities are named. Values emerge even more clearly as we also begin to consider the target users of these products.

The emotional expressions recognizable and performable by the SEMAINE system seem to be heavily culturally coded, as they are the result of a labeling process performed on and by only European speakers. Similarly, Beyond Verbal locates the "universally human" in intonation and intervallic habits that are highly culturally specific, and trained. Western tuning systems and scales surround the Western consumer throughout their daily lives, and are explicitly practiced by classical and popular Western

[63] Shunji Mitsuyoshi, "The application of emotion and mind recognition in voice Quantitative measurement technology of emotion".

[64] Björn Schuller, Stefan Steidl, Anton Batliner, "Repository of the INTERSPEECH Computational Paralinguistics Challenge (ComParE) Series," updated September 27, 2016, accessed May 20, 2016. http://compare.openaudio.eu/.

musicians according to an almost 300-year old educational system. The technology could work well on western speakers, who have been trained to communicate using a musical language that the design deciphers. The products' claims to universality, however, likely neglect the speaking patterns of non-Western speakers, tacitly classing them as "inhuman." As advertising technologies, they are not concerned as much with global applicability as they are with target audience: the Western consumer is the *wealthiest* consumer, and therefore he is the speaker who becomes "hearable."

Applications: The Use of the Meaning

A close read of the available patents and code reveals that these technologies started from diverse goals and assign rather different emotional vocabularies to the speaking voice, imagining the human mind, and its possibilities, in varying ways, and therefore allowing different aspects of the speaker's subjectivity to be registered.

These variations fall away to a great extent when these designs hit the market. A survey of the marketing language for and applications of these technologies shows a tendency of the commodities to make similar claims using similar language. Companies like AGI, the SEMAINE project and Nemesysco have had to change their tune: instead of telling truth or expressing sympathy, they now also couch their products in the language of prediction and increased productivity.

The story of Nemesysco's lie detection products shows clearly the move from fact to future. Founded in 2000, Nemesysco marketed one of the earliest commercial instances of these technologies. Today, the company describes its products as "advanced and non-invasive investigation and security tools, fraud prevention solutions, CRM[65] applications, consumer products, and psychological diagnostic tools."[66] Layered Voice Analysis (LVA) is marketed to serve "professional investigators" working in the fields of law-enforcement and security. SENSE technology, the commercially-available version, is now marketed for call centers, claims adjusters, and human resources. Finally, Nemesysco owns the patent on LioNet Technology, a "heuristic learning engine, designed to increase the accuracy of the identification of specific "emotional signatures."[67] This software is used by automated call centers, and by recruiting companies, which seek to screen potential employees for trustworthiness and character.

The applications of this software have been wide-ranging and controversial. In 2007, an Israeli company that trains military pilots incorporated LVA into its flight simulators in order to measure "stress levels and other emotions" of trainees.[68] Also in

[65] "Customer Relationship Management"

[66] http://nemesysco.com/, access May 20, 2016.

[67] http://www.nemesysco.com/technology.html, Dec 3 2014.

[68] M2PressWIRE, "BVR to Integrate Nemesysco's Layered Voice Analysis Technology for Emotion Detection in Flight Simulators," *M2 Communications Ltd*, September 25, 2007. http://www.m2.com/group/.

2007, the UK Department of Work and Pensions announced that it would be supporting a trial of Voice Risk Analysis "to identify claimants suspected of benefit fraud," extending £1.5 Million to expand the program in 2008.[69] In 2012, *The Times of India* reported that the Directorate of Forensic Sciences in Gandhinagar had procured the country's first LVA system, and that it would be taking the place of staff in screening suspects. The article reports that LVA was also in use in 87 other countries.

Despite its proliferation, a good amount of research has surfaced proving the software is ineffective at detecting lies. A 2008 article in *Engineering & Technology* magazine quotes University of Portsmouth researcher Aldert Vrij, who believes that the technology is largely a placebo. The article also reports that "studies of two voice-stress analysis systems … found that neither could detect lies about drug use among prisoners."[70] It concludes with the suggestion that VSA should be coupled with "managed conversation," and that it is hard to tell what exactly is producing the results – the technology or the social dynamic created by the technology.[71] A similar study published in 2013 in the *Journal of Forensic Sciences*, shows that LVA is "not effective"[72], detecting "correct deception" only 48% of the time.[73] However, the authors note that police officers do report better results when employing these technologies, and theorize that the technology's presence changes the psychological dynamic of the interrogation, helping the officer.

A 2011 article in *Liverpool Law Review* by Solicitor Michael Green considered the ethical and legal ramifications of the use of LVA. The author takes the successful operation of the technology at face value, and considers whether its use violates privacy and non-discrimination legislation. He shows that the speaking voice has been protected under privacy and surveillance laws, and that extracting information about the voice is regulated under laws addressing data-collection and the right to privacy. Furthermore, this technology claims to gather information that Green considers even more private than "content" or "identity," and therefore should have to follow even stricter privacy requirements.

> … [LVA] is claimed to be able to peer into the very workings of the brain. That level of intrusion into the private thoughts of an individual must, surely, require

[69] Michael Green, Does the Use of Voice Lie Detection Equipment in the United Kingdom Breach Article 8 of the European Convention on Human Rights and the Equality Act 2010?," *Liverpool Law Review* (2011) 32, 95-96.
[70] Chris Edwards, "Risky Claims: Engineering Lie Detection," *Engineering & Technology*, 28 April – 9 May 2008, 22.
[71] Edwards, 22.
[72] Frank Horvath et al, "The Accuracy of Auditors' and Layered Voice Analysis (LVA) Operators' Judgments of Truth and Deception During Police Questioning," *Journal of Forensic Sciences* 58 (2), March 2013, 385.
[73] Horvath et al, 390.

> the most stringent of safeguards before its use could even be contemplated in a democratic society … It is more than arguable that being able to access ‘‘brain events’’ of an individual, in real time, is more intrusive than an analysis of their DNA.[74]

Green also goes on to point out that the use of this technology to provide state-sponsored services discriminates against subjects who cannot be "heard" by the software. The software works with a "normal" speaking voice only, and therefore people with speech disabilities, depressed or otherwise mentally-ill people whose speech is thereby affected, elderly people with "feeble" voices, and speakers of English-as-a-second-language would be unfairly misunderstood by this technology. Such at-risk populations, who are likely to seek government benefits, could be mistakenly labeled as "fraudulent" by the software.[75]

The critiques of this software have generally led to its being marketed differently: instead of detecting factual speech, as it claimed to do in the late 2000s, it is now used for determining a subject's confidence and comfort with what she is saying. Fraud-detection companies that employ the software in their call centers find it useful for classifying "high-risk" claims worth attention by a human operator.[76] In 2010, *Australasian Medical Journal* published a study which used LVA to track emotions in voice and then showed how emotions are correlated with personality type.[77] In 2012, in a study in *The Journal of Finance*, researchers used LVA to detect confidence in CEOs' voices in order to decide if their companies were investment-worthy.[78]

The "next generation" of affective computing companies learned to focus on "softer" and more speculative applications: describing a subject's mental health, personality, or mood, and using these indicators to make financial and medical predictions about future performance. The software now claims to outline fields of possible feelings, rather than to label extant ones. Cogito Inc., named the fastest growing company in Boston in 2013, refers to its employees as "Cogitniks" and boasts that "Cogitniks are inventing the future of human intuition."[79] Their main software product to

[74] Green, 103.
[75] Green, 105.
[76] Edwards, 21-22.
[77] Brinda Manchireddy, Sumaiyah Sadaf, and Joseph Kamalesh, "Layered Voice Analysis Determination of Personality Traits," *Australasian Medical Journal* Volume 3 Issue 8 (August 2010).
[78] William J. Mayew and Mohan Venkatachalam, "The Power of Voice: Managerial Affective States and Future Firm Performance," *The Journal of Finance* Vol. LXVII, No. 1 (February 2012).
[79] "Want to Work at Cogito?" *Cogitocorp.com*, accessed March 7, 2015, http://www.cogitocorp.com/careers/.

date is called *Dialog*. *Dialog* monitors real-time speech during phone calls and gives visualizations of the emotional content of the speech.

This design has turned into a number of products marketed to healthcare and insurance providers. For Disability Claims Managers, the software employs "engagement measures" to build trust between the manager and the claimant and to assess whether a claimant is eligible for an "investment" such as occupational rehab, that will increase her return-to-work timeline. It is also used to detect if a claimant is resisting return-to-work. It claims to provide a 9:1 ROI from "reduced disability claims durations" and a 2:1 ROI from "reduced disability related medical costs."[80]

For telephonic coaches administering Behavioral Health coaching and claims management, the software is designed to help identify "signs of distress such as paucity of speech, flat affect, and short utterances." Currently used in care management programs for patients on Medicare Advantage[81], it now includes a predictive model for determining depression. The company's website claims that the use of the software has led to a 250% increase in identification of members at risk for Behavioral Health comorbidities, 868 members newly enrolled in Behavioral Health care, and an additional 163 members coded with HCC 55[82] illnesses.[83] The company also is currently testing a smartphone app called "Companion," which monitors mood and mental health. This project is a collaboration with DARPA and Raytheon-BBN.[84],[85]

While Cogito began from a mental health model, Beyond Verbal began by marketing itself to advertising companies. Launched in May 2013, the company's website asserts that "[an] understanding of people's moods, attitudes, and decision-making characteristics … opens a window … with major impact on numerous multi-billion dollar verticals."[86] Proposed applications of this product include "commercially

[80] "Case Studies: Shortening Disability Durations by Improving Engagement in Return to Work Programs," *Cogitocorp.com*, Accessed September 8, 2014, http://www.cogitocorp.com/shortening-disability-durations-by-improving-engagement-in-return-to-work-programs/.

[81] Medicare is the United States's national health insurance program for citizens over the age of 65. Medicare Advantage is a branch of this program offered by private insurance companies.

[82] HCC-55 is the medicare code for "major depressive, bipolar, and paranoid disorders." http://www.hfni.com/assets/CMS%20HCC%20Table%202013.pdf

[83] "Case Studies: Delivering More Tightly Integrated Physical and Behavioral Health Care to Complex Populations," *Cogitocorp.com*, Accessed September 8, 2014, http://www.cogitocorp.com/delivering-more-tightly-integrated-physical-and-behavioral-health-care-to-complex-populations/.

[84] "Mobile Reality Analysis," *Cogitocorp.com*, Accessed May 20, 2016 http://www.cogitocorp.com/research-showcase/.

[85] DARPA stands for "Defense Advanced Research Projects Agency," and is the research department of the U.S. Army. Raytheon-BNN is private tech company, which began as an acoustics research company and is now an American military contractor, primarily for DARPA.

[86] "What We Do: The Three Basic Things We Do," *Beyond Verbal*, assessed Dec 3 2014, http://www.beyondverbal.com/start-here/what-we-do/.

valuable activities" such as "decoding a person's emotional positions, conscious and unconscious, for the purpose of intelligence, negotiations, improved dialogue, etc." Despite its beginnings in advertising, the company now offers four different types of services, which are geared towards wellness, market research, app developers, and enterprises. The wellness service is designed to monitor emotional health using smartphones or wearable devices, using an API[87] that tracks a subject's emotional profile by extracting it from twenty seconds of speech.[88]

Market research products claim to provide "an immediate, accessible, and economical way to measure the mood and attitude of your target market."[89] Voice samples are collected via mobile device or laptops and sent to the cloud, where the software analyzes the audio with regards to valence, arousal, and temperament. The website explains that this could be useful for monitoring the attitude at a conference or event, getting feedback for product design, brand studies, and advertising effectiveness. As of June 2014, Beyond Verbal is partnering with a large marketing research consultancy, Lieberman Research worldwide to help understand consumer motivation.[90]

The company also markets a "call center software suite," which measures "caller and agent intonations in real time." The product claims to help companies understand their client's "attitude" and "decision-making" in real-time, and provides real-time scripts which will offer the best approach to the customer given her mood.[91] In 2013, the company launched the "Moodies" app, which is available on the web and now on mobile devices.[92] Like most of Beyond Verbal's software, the app determines the user's mood based on a recording of twenty seconds of speech. It offers a primary and secondary emotional category, and a description of this category.

[87] API stands for "Application Program Interface." APIs specify a set of routines and protocols by which components of a piece of software interact with each other. APIs provide programmers with the building blocks necessary to quickly assemble a program by taking care of lower-level processes.

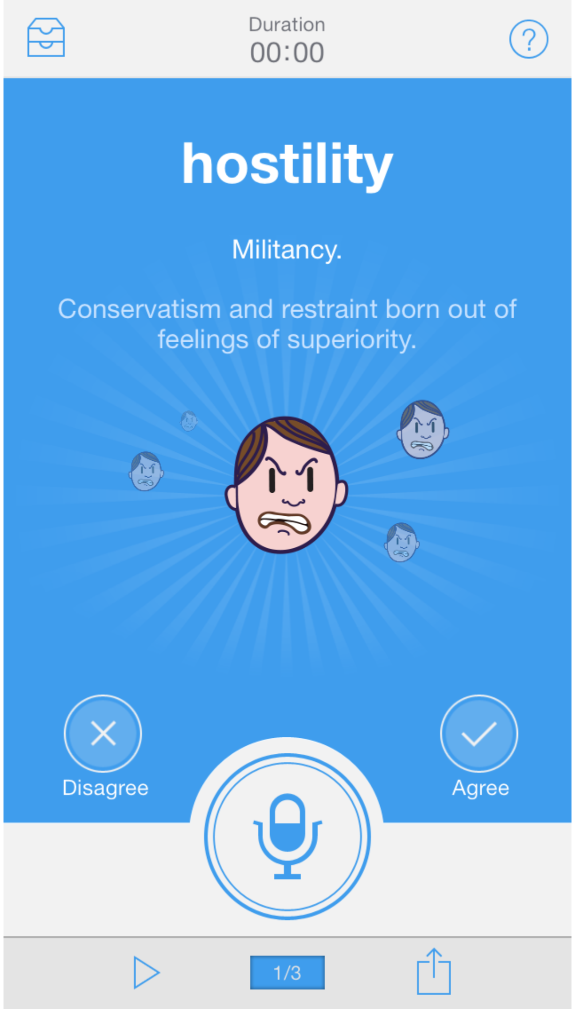
[88] Levanon and Lossos, "United States Patent 8078470 B2 …", 11.

[89] "Market Research: Our Core Offering," *Beyond Verbal*, assessed Dec 3 2014, http://www.beyondverbal.com/choose-solution/market-research/our-core-offering/.

[90] "Press Release: Lieberman Research Worldwide Integrates Beyond Verbal's Emotion Analytics Technology to Enhance Market Research," *Beyond Verbal*, June 16, 2014, http://www.beyondverbal.com/media/press-release-media-mentions/lrw-press-release/#.

[91] "Enterprises: Boost your call center performance like no other solution," *Beyond Verbal*, accessed Dec 4, 2014, http://www.beyondverbal.com/choose-solution/enterprises/.

[92] PR, Newswire. "Beyond Verbal Launches Moodies - The World's First Emotions Analytics App For iOS." *PR Newswire US* 23 Jan. 2014: *Regional Business News*. Web. 17 Oct. 2014.



hostility

Militancy.

Conservatism and restraint born out of feelings of superiority.

Figure 5: Author's moodies output, December 10, 2014

The SEMAINE project has given birth to openSMILE technologies, the products offered by a for-profit company called audEERING, which was founded by SEMAINE researchers. The company's website explains that "paralinguistic speech analysis is a young field with great potential for improving efficiency for … call centres, targeted advertising, or [for] increasing the usability of intelligent virtual agents and humanoid robots."[93]

Similarly, AGI, which began by including a relational psychoanalytic model for designing sensitive Virtual Humans, now offers products that evaluate consumer motivation, honesty, confidence, and intentions, much like the other companies. AGI markets three categories of products: Sensibility Technology (ST), Voice Emotion Analysis (VEA), and Psychoanalysis System Technology (PST). ST and VEA offer

[93] "Solutions Directory: Intelligent Audio Analysis," audEERING for *Intel Corporation*, accessed May 20, 2016, http://iotsolutionsalliance.intel.com/solutions-directory/intelligent-audio-analysis.

visualizations of the speaker's emotions, based on a discrete emotional model that includes nine possible emotions, which are translated into a spectrum of colors. These products are marketed for medical treatment applications and scientific research. ST-CRM is software for call centers that monitors the operator's and caller's emotions in real-time in order to determine motivation. PST is aimed towards "improved cognitive behavior therapy" and claims to recognize speakers' mental state and emotional health.[94] A patent issued to AGI in 2014 claims to be suitable for detecting the lack of confidence or degree of tension in a speaker. In addition, it includes "a lie detector detecting typical emotion when telling a lie can be realized according to degree of tension and the like." The patent lists a number of potential applications, including "call centers, security applications, support with psychoanalytic treatment, nursing care support, credit management, and many more."[95]

Design Values and Ethics: The Meaning of the Use

Why this move from such a rich multiplicity of psycho-epistemological models to such similar applications? Today, affective listening software is no longer about divining truth or digitizing sympathy – it is about finding in the voice indications of one's worth, resilience, reliability, and investability quotient. The use-value these technologies find for the affective voice mutates as we move along their design history: from truth–telling (lie-detectors), to a revelation of the soul that can only be conferred by the machine (self-monitoring and mental health evaluation), and then, finally, to prediction (risk management.) There is a slip here from registering that the speaker is hiding something that she knows, to the proposition that the speaking subject doesn't even know herself, to a harnessing of the ultimate unknown: the future. Affective listening technologies rely on a kind of *computerized alienation*: they market the subject to the self or other for evaluation and recognition, according to the terms of the computer.

In 1844, Marx theorized the alienating effects of capitalism, sped up by the industrialization of production. The production of objects for (someone else's) profit and consumption creates an experience of estrangement in the worker, when he sees his work (the efforts of his mind and body, his time, his soul) materialized in a product from which he is forced to separate. This experience of being estranged from, yet materialized in, the commodity, creates a sense of alienation that becomes pervasive under capitalism: the subject begins to see himself the way he sees his products: as a source of profit. Eventually, he regards others in the same way. The potential for profit turns life, and time, into alienated materials to be instrumentalized for capital gain. What are the effects of this? To Marx, "an immediate consequence of the fact that man is estranged from the

[94] Mitsuyoshi, "The application of emotion and mind recognition in voice … ", 3.

[95] Shunji Mitsuyoshi, "United States Patent 8738370 B2 – Speech Analyzer Detecting Pitch Frequency, Speech Analyzing Method, and Speech Analyzing Program," United States, May 27, 2014.

product of his labor, from his life activity, from his species-being, is the *estrangement of man* from *man*."[96]

That is to say, not only are the products of our labor estranged from us, but our very humanity and relational capacities become alienated as a potential source of profit. Current applications of affective listening technologies can be understood as the manifestation of this estrangement in the affective realm. The ability to listen to the self and the other is delegated to the technology, disembodying affect and re-presenting it as a quantity to be bought, sold, and invested in.

Self-tracking applications, such as "Moodies," are the turning in on the self of these monitoring technologies. Designed mainly for mobile phones or Body Area Networks, these listening technologies live close to the skin, often combining with other forms of body monitors (pulse sensors, etc.) These apps claim to offer deeper self-knowledge by providing the user an "objective view" of their interior self. The inner self becomes perceptible to the subject once it is filtered and parsed by a computer, providing a form of estrangement that allows us to self-surveille and –monitor, affectively distancing us from ourselves, while remaining physically close.

Just as the self becomes alienated and subsequently predicted through these technologies, so does the other. These technologies no longer claim to know the speaker's true intentions in the moment (as in the interrogation scenario), but instead are marketed on knowing the speaker's behavior in the future, intentional or not. Companies like Beyond Verbal and Cogito sell their products based on their claims to accurately predict the performance of a worker, the future health of a claimant, or the investment-worthiness of a businessman. All these claims coalesce around the new speculative economy – a technology is valuable if it can help the user to put their money into something that will return more value in the future, or to avoid spending in cases in which that will not happen.

Appadurai, echoing Knight, explains that in the new risk economy, "profit arises out of the inherent, absolute unpredictability of things, out of the sheer, brute fact that the results of human activity cannot be anticipated…"[97] Risk management technologies, then, promise profit in their claim to handle, manipulate, and predict this human activity. How could affect possibly be an index of profitability? The concept of the affective – as something that is prior to language or deliberation – sets up the idea of a kind of data that is especially hard to mine, as it is, by definition, unknown to the actant until the instant it manifests, and even then it is not known in measurable, linguistic terms. Affect is understood as motivating our actions and disrupting our community, but it is not something we express deliberately, and therefore not something we can promise or

[96] Karl Marx, "Estranged Labour," in *Economic and Philosophical Manuscripts of 1844*, trans. Martin Mulligan (Moscow: Progress Publishers, 1959), accessed May 20, 2016, https://www.marxists.org/archive/marx/works/1844/manuscripts/labour.htm.

[97] Appadurai, 246-247.

declare. Involuntary expressions like this are the most unwieldy things to predict, and therefore they can be the most privileged and profitable (in the speculative capitalist imaginary), specifically *because* they are so hard to know. Known quantities made of known materials behave in predictable ways, but the affective is both deeply compelling and deeply ineffable. Affect is high risk; therefore affect is high profit.

Allen Feldman has theorized an "actuarial gaze" that is typical of post-9-11 mediatic depictions of violence, and which focuses subjecthood on the fear and pervasiveness of risk. This risk, however, is by nature invisible -- or, I would say, unhearable -- and therefore it is given over to machines to track. Feldman lists three important political implications of this: "the wish for prosthetic extension of the human sensorium … the consequent assignment of sensory capacity, power and judgment to machinic, automated and institutionalized instruments of perception; and the alignment of risk perception with the wish image."[98]

Affective listening technologies are particularly salient examples of such instruments. Listening, perhaps even more than gazing, is communicative. The voice is *used* for intersubjectivity – to express and share the self, to convey thoughts and feelings, and to open up possibilities for the recognition of and relation to the other. Affective listening technologies are moving instead into what Appadurai calls an "ethics of probability."[99] In a world where the ethical is an imagination of the field in which brands of intersubjective relations are possible, these new tools of speculative capitalism frame the soul not as commodity, as in Marx's time, but as derivative or security. With these tools, we listen to each other as threats, risks, and potential investments.

This story is not entirely the fault of the machines: it is what speculative capitalism does with machines. Just as Marx understood industrialization as the conditions for alienation and the growth of capitalism, we can consider the rise of predictive computing as the conditions for the growth of the speculative economy, which, combined with a proliferation of affective computing and signal processing, is giving rise to "actuarial" forms of listening.

A close read of these listening algorithms helps to show how the digitization of vocal affect requires a particular form of *partager*: to share our feelings with and through the computer, they must be cordoned off into categories, the valences and emotions of the voice must be partitioned and given names. Perhaps there is a link here between the *partage* of the stocks and derivatives markets, and the partitioning of feelings and vocalizations. Affective listening algorithms articulate what the digital commodity "hears in us," and, through our adoption, set up the boundaries of the perceptual vocabularies along which we can listen to – and find value in – each other.

[98] Allen Feldman, "On the Actuarial Gaze: From 9/11 to Abu Ghraib," *Cultural Studies* 19:2 (March 2005): 205-206.

[99] Appadurai, 4.