Giving Voice to Emotion

pulse.embs.org/may-2016/giving-voice-to-emotion/

Summer Allen

It's tough to imagine anything more frustrating than interacting with a call center. Generally, people don't reach out to call centers when they're happy—they're usually trying to get help with a problem or gearing up to do battle over a billing error. Add in an automatic phone tree, and you have a recipe for annoyance. But what if that robotic voice offering you a smorgasbord of numbered choices could tell that you were frustrated and then funnel you to an actual human being? This type of voice analysis technology exists, and it's just one example of the many ways that computers can use your voice to extract information about your mental and emotional state—including information you may not think of as being accessible through your voice alone.

Detecting Emotions

Ten years ago, when Izhak Shafran was a researcher at AT&T Research Labs, he and his team wanted to know whether different characteristics of a customer's voice —so-called voice signatures—could tell them information about that customer, such as gender, age, dialect, and emotion. Using actual speech collected from AT&T's "How May I Help You" customer call system, Shafran and his colleagues were able to train an algorithm to detect these characteristics at levels high above chance [1]. Using only information about pitch and something called the *Mel frequency cepstral coefficient*, a vocal feature that is commonly used in voice recognition tools, the algorithm was able to correctly identify a caller's gender with 95% accuracy, approximate age with 70% accuracy, dialect with 45% accuracy, and emotion with 68% accuracy. "We could easily detect things like frustration," says Shafran, who is now a speech researcher at Google (Figure 1, right).

Izhak Shafran

Fast forward to the present when a Boston-based company called Cogito uses voice analysis algorithms to help actual humans—customer service agents—gain real-time insight into how their conversations are going by analyzing several different features, such as the degree of a customer's pitch variation, which can indicate boredom or anger. Cogito has an interface that allows customer service agents to look at their own voice features, too, so that they are able to "dynamically adjust their style to align with the customer's preference."

Several academic groups are also working on applications that can extract emotional information from speech samples. One example is EmoVoice, a "comprehensive framework for real-time recognition of emotions from acoustic properties of speech," developed by Elisabeth André, Ph.D., and her team at the University of Augsburg in Germany. EmoVoice performs the same three steps used by most tools designed to recognize emotions from speech: audio segmentat ion (breaking down speech samples into pieces that can be analyzed), feature extraction (finding which acoustic characteristics best describe emotions), and classification (using machinelearning and statistical modeling techniques to train algorithms to detect which extracted features are associated with which emotions) [2].

The first step, audio segmentation, can be tricky. Words are unlikely to be long enough to be useful, but if you analyze too large a chunk of speech, features can be washed out. So Emo- Voice divides audio samples into chunks akin to phrases—not too short, but not too long. The second step, feature extraction, requires taking measurements and determining the acoustic properties that can best characterize emotions. Examples of such properties are pitch and voice intensity (or loudness). EmoVoice can extract 1,302 features, although only between 50 and 200 are used when analyzing any given sample. The third step, classification, uses computer algorithms to sort the extracted features into groups representing different emotions. For example, a monotonous audio sample that has minimal

pitch variation could be grouped as an indicator of sadness or boredom.

EmoVoice, which is available to everyone, has already been integrated into several applications. These include a humanoid robot named Barthoc, who can express joyful and fearful facial expressions when listening to a fairy tale; a virtual agent named Greta, who can mirror the emotions of a speaker with her facial expressions and deliver appropriate verbal feedback [3]; and art installations with emotional kaleidoscopes and a tree that grows and changes color and shape based on the emotions extracted from people's voices.

Health Applications

Although it may be interesting to speak with a robot or virtual assistant that can visibly empathize with us based on our vocal features, there are also more serious applications for voice analysis technology. Researchers have spent the past few decades trying to determine whether voice analysis can be used to help clinicians diagnose depression. Most people experiencing depression first go to their primary physicians, who often lack the training to correctly diagnose the disorder. (One study found that they are successful only half the time [4].) Thus, there is a growing demand for more objective diagnostic methods. Vocal analysis is particularly attractive for several reasons: collecting speech samples is relatively easy and cheap, speech can be monitored while a patient is at home, speech analysis is noninvasive and can be automated, and there are known vocal hallmarks of depressed speech.

Julien Epps

"Psychomotor retardation (the slowing down of thoughts and movement often seen in depression) is naturally going to affect speech production because speech is a very complex process that involves many, many different muscles and a fairly intensive degree of coordination between them all," says Julien Epps, Ph.D., an associate professor of signal processing at the University of New South Wales, Australia (Figure 2, right). Depression can also cause cognitive impairments and hamper working memory, which can impact speech planning, production, and articulation.

Epps and his colleagues use databases of voice samples from people who are and are not depressed to train statistical models to extract the auditory features of depressed speech (Figure 3). "We've been working entirely with an acoustic approach—essentially that's just looking at the quality of the speech—the way that it sounds, the timbre, the prosody (which is the intonation or energy variation of the speech over time, for example)," explains Epps. After training a model, the group feeds it new speech samples to test how well it is able to distinguish between depressed and nondepressed speech. "The results can be pretty striking," he adds.

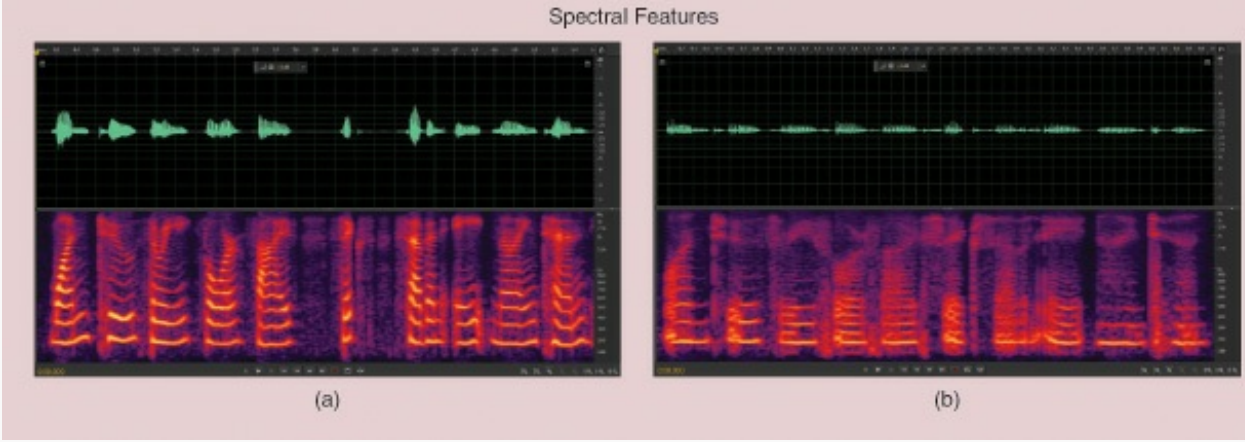

Figure 3: Spectrograms of (a) nondepressed and (b) depressed speech, produced by Julien Epps's doctoral student Nick Cummins. The latter shows some typical characteristics of depressed speech: lower variability in pitch and energy, which is the origin of the "flat, monotonous" quality typical of depressed speech. (Images courtesy of Julien Epps.)

Technology like this could one day be adopted by doctors if additional testing bears out the preliminary results. "I think there are plenty of reasons to be excited about it," Epps notes. "But there are also plenty of reasons to be a bit cautious about it as well." One of the issues is that these studies have been conducted using fairly constrained laboratory conditions. It's unclear how well they would do with, say, audio collected from a smartphone microphone with background noise present. Epps is currently collecting a database of these samples and working with a start-up company to create a smartphone app for detecting or monitoring depression. A number of other academic groups and companies, including Cogito, are also working on apps or software for this purpose.

Research suggests that vocal features change when people become suicidal. For example, clinical psychologists Stephen and Marilyn Silverman noticed that their suicidal patients had such distinctive speech changes that they found the voices to be alarming. These patients' speech was "hollow and toneless" with "reduced energy" and a "lack of emphatic accent and dynamic expressiveness" [5].

So the Silvermans teamed up with researchers at Vanderbilt University in Nashville, Tennessee, to examine whether these speech differences could differentiate between near-term suicidal patients and nondepressed controls, as well as between near-term suicidal patients and depressed patients. Using speech from a small group of patients and nonpatients, they were able to develop a statistical classifier that can distinguish between the speech of near-term suicidal patients and depressed patients with 80% accuracy. Follow-up work using different models has shown similar results. Although more work is needed with larger data sets containing more realistic speech samples, this type of technology may be useful for triaging calls to clinicians, 911 operators, or crisis call centers to identify callers at risk of imminent self-harm.

One speech analysis tool that may be close to clinical adoptability is PRIORI, an application for "detecting early signs of mood changes in people with bipolar disorder," under development by University of Michigan professors Emily Mower Provost, Ph.D., and Satinder Singh Baveja, Ph.D. This smartphone app records a patient's normal phone conversations, extracts speech features from these samples, and detects possible mood changes. A study of six patients with bipolar disorder found that that the app could detect both depressed and elevated moods [6]. The hope is that, as PRIORI collects more data from more patients, the app can be trained to detect changes that predict an upcoming manic or depressive episode and then notify the app user and, perhaps, the patient's medical provider.

The early detection of mild cognitive impairment and Alzheimer's disease is another potential application for vocal analysis technology. "It is crucial to have technology that can automatically assess and screen patients before calling them to the clinic," says Meysam Asgari, Ph.D., a senior research associate with the Center for Spoken Language Understanding at the Oregon Health and Science University (Figure 4). His work so far has shown that people with

mild cognitive impairment use more words to describe something, such as an event, compared with people without cognitive impairment. “One of the other future goals is looking into inherent characteristics of speech, like pitch frequency, that I have been using for other clinical applications like detecting Parkinson’s disease or autism spectrum disorder,” says Asgari.

Meysam Asgari

Detecting Deception

Voice analysis technology has uses beyond the fields of commerce and health. For example, companies have developed tools for detecting when a speaker is lying based on features extracted from speech. These tools have been adopted by local law enforcement agencies, governments, airports, and firms trying to detect signs of insurance fraud.

The specific features extracted by these tools and the algorithms used to classify deceptive versus nondeceptive speech are often not disclosed, but one common technique— layered voice analysis—purports to measure a vocal feature called a microtremor. However, it is unclear whether microtremors actually exist in voice production and, if so, how they relate to deceptive speech. “These microtremors are thought to be indicative of deception, but there’s no scientific basis for either their existence or for any sort of link between deception and those sort of tremors—none at all,” says Christin Kirchhübel, Ph.D., whose dissertation research at the University of York in the United Kingdom was devoted to researching vocal features associated with deception (Figure 5).

Christin Kirchhübel

Indeed, it turns out that it is very difficult to reliably detect when someone is lying based on vocal features alone. “There’s no such thing as Pinocchio’s nose,” according to Kirchhübel. “It’s not possible to reliably detect deception based on speech analysis—or voice analysis for that matter. I don’t think there will ever be an error-free way of doing so.”

Part of the issue is that there is no single clear theory describing how lying changes a person’s voice. A number of theories have been developed to account for the emotional, cognitive, and communicative processes that tend to accompany deception, and these can be used to make predictions about the speech changes liars may display. First, the arousal theory suggests that liars feel more stress and thus speak faster, with more errors, and at a higher pitch. Second, the cognitive complexity theory suggests that liars experience more cognitive load because they are trying to keep track of their story, which may manifest in slower speech with more pauses and hesitations. Third, the behavioral control theory suggests that liars are “trying very hard not to present the stereotypical image of a liar” and thus produce speech that is more carefully planned and controlled. No single approach is sufficient in explaining deceptive behavior. Rather, it is expected that if stress, cognitive load, and behavioral control do occur, they will occur simultaneously and to varying degrees. Says Kirchhübel, “As a consequence, because stress, cognitive load, and behavioral control have different effects on speech, we might find contradictory speech effects during deception.”

Kirchhübel’s conclusion—that it’s not possible to reliably detect deception based on voice analysis—is based on the conflicting results of her own work. In an experiment in which subjects answered open-ended questions, Kirchhübel found that people’s speech contained more hesitations, more pauses, and longer pauses. But, in a second experiment that was framed as a more accusatory interrogation with yes/ no-type answers, Kirchhübel found the opposite effect— “their speech tempo was faster.” And in a third experiment, she found that when truth tellers were placed in exactly the same situation as people who were asked to lie, they behaved just like the liars—they spoke faster and used fewer pauses. This suggests that it may not be possible to reliably separate the speech effects caused by deception and those caused by other factors, such as stress.

This is part of what makes lie detection so difficult: the ability to detect deception is very context-dependent. “Vocal features do discriminate lies and truths. However, they are very contingent on the kind of questions you ask and how you ask them,” says Aaron Elkins, Ph.D., a postdoctoral researcher at the University of Arizona, Tucson, also affilliated with BORDERS, the National Center for Border Security and Immigration. (He’s starting his own behavioral and neuro lab at San Diego State University in 2016.) For example, Elkins and his colleagues can determine with very high accuracy (80–90%) whether someone is lying about having a fake ID in a very specific laboratory experiment that involves a series of mentally taxing questions. “The problem is the profile of deception can be very different in different scenarios,” he explains.

Figure 6: Researcher Aaron Elkins demonstrating an earliergeneration A VATAR, a virtual interviewer used to determine whether someone is lying about having a fake ID. (Photo courtesy of BORDERS.)

One way to improve lie detection is by adding sensors to detect other behavioral signals associated with deception, such as eye movements and pupil dilation. Those additional sensors are integrated into AVATAR, a virtual interviewer that Elkins and his colleagues have tested in kiosks in airports and border security stations (Figures 6 and 7). How do people respond to a virtual interviewer asking them security questions? Elkins answers, “People actually liked the system.”

Figure 7: The version of AVATAR currently in use. (Photo courtesy of BORDERS.)

The Future of Voice Analysis Technology

As voice analysis technology advances, we may find our devices becoming increasingly attuned to our mental states and feelings. “We have ways to interact with devices around us, and those interfaces will need the capability to at least coarsely track our emotions,” says Izhak Shafran. He gives the example of a voice-enabled personal assistant or robot that may respond with different urgency depending on users’ happiness or frustration.

For his part, Julien Epps can imagine a future where we wear glasses with sensors that can detect how hard we are concentrating and will “understand when to interrupt us” with incoming messages (as well as a host of other applications). “It’s a very exciting area—there are sensors everywhere.”

References

1. I. Shafran, M. Riley, and M. Mohri, “Voice signatures,” in *Proc. 2003 IEEE Workshop Automatic Speech Recognition and Understanding (ASRU ‘03)*, St. Thomas, Virgin Islands, pp. 31–36.

2. T. Vogt, E. André, and N. Bee, “EmoVoice—A framework for online recognition of emotions from voice,” in *Proc. 4th IEEE Tutorial and Research Workshop Perception and Interactive Technologies for Speech-Based Systems (PIT 2008)*, Kloster Irsee, Germany, 2008, pp. 188–199.
3. F. de Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, and B. De Carolis, “From Greta’s mind to her face: Modelling the dynamics of affective states in a conversational embodied agent,” *Int. J. Hum. Comput. Stud.*, vol. 59, no. 1–2, pp. 81–118, July 2003.
4. N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” *Speech Commun.*, vol. 71, pp. 10–49, July 2015.
5. A. Ozdas, R. G. Shiavi, D. M. Wilkes, M. K. Silverman, and S. E. Silverman, “Analysis of vocal tract characteristics for near-term suicidal risk assessment,” *Methods Inf. Med.*, vol. 43, no. 1, pp. 36–38, 2004.
6. Z. Karam, E. M. Provost, S. Singh, J. Montgomery, C. Archer, G. Harrington, and M. Mcinnis, “Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech,” in *Proc. 2014 IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, pp. 4858–4862.