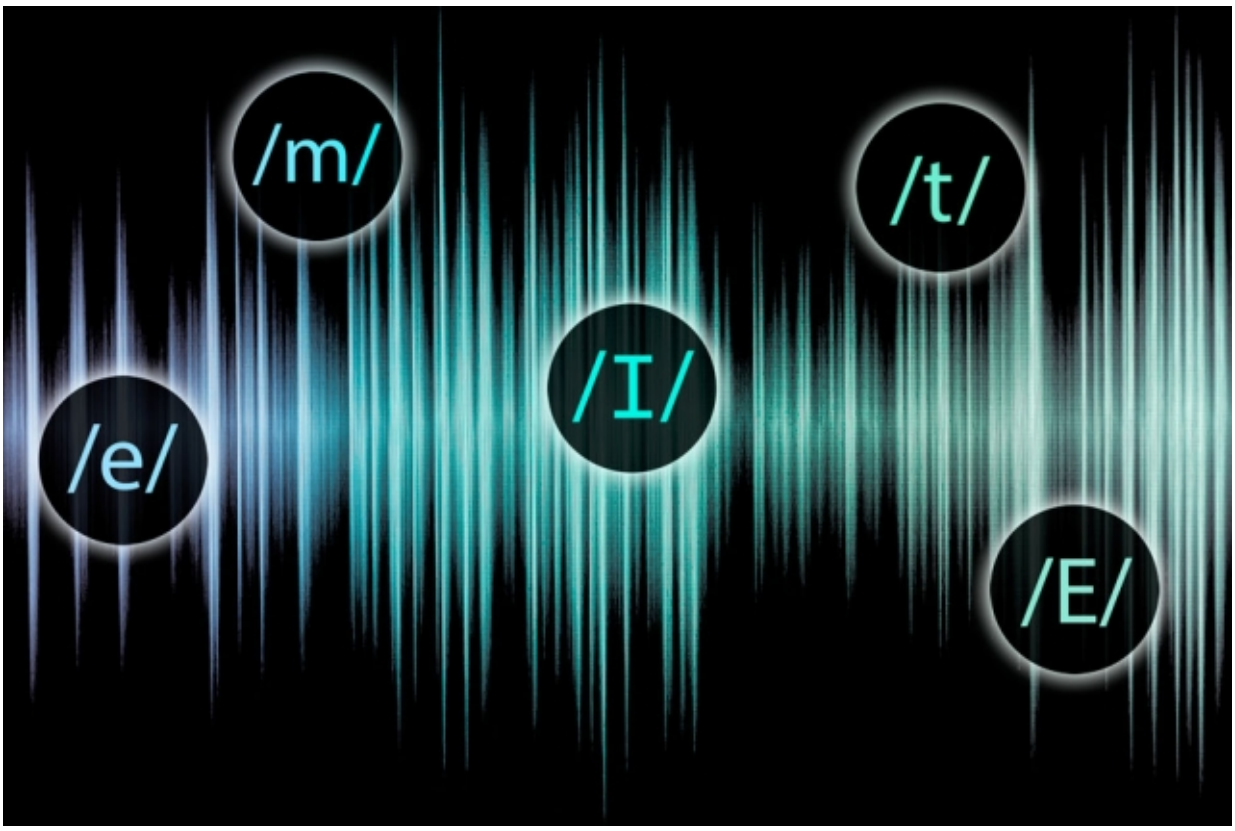
System learns to distinguish words' phonetic components, without human annotation of training data

September 14 2015, by Larry Hardesty



Credit: Jose-Luis Olivares/MIT

Every language has its own collection of phonemes, or the basic phonetic units from which spoken words are composed. Depending on how you

count, English has somewhere between 35 and 45. Knowing a language's phonemes can make it much easier for automated systems to learn to interpret speech.

In the 2015 volume of Transactions of the Association for Computational Linguistics, MIT researchers describe a new machine-learning system that, like several systems before it, can learn to distinguish spoken words. But unlike its predecessors, it can also learn to distinguish lower-level phonetic units, such as syllables and phonemes.

As such, it could aid in the development of speech-processing systems for languages that are not widely spoken and don't have the benefit of decades of linguistic research on their phonetic systems. It could also help make speech-processing systems more portable, since information about lower-level phonetic units could help iron out distinctions between different speakers' pronunciations.

Unlike the machine-learning systems that led to, say, the speech recognition algorithms on today's smartphones, the MIT researchers' system is unsupervised, which means it acts directly on raw speech files: It doesn't depend on the laborious hand-annotation of its training data by human experts. So it could prove much easier to extend to new sets of training data and new languages.

Finally, the system could offer some insights into human speech acquisition. "When children learn a language, they don't learn how to write first," says Chia-ying Lee, who completed her PhD in computer science and engineering at MIT last year and is first author on the paper. "They just learn the language directly from speech. By looking at patterns, they can figure out the structures of language. That's pretty much what our paper tries to do."

Lee is joined on the paper by her former thesis advisor, Jim Glass, a

senior research scientist at the Computer Science and Artificial Intelligence Laboratory and head of the Spoken Language Systems Group, and Timothy O'Donnell, a postdoc in the MIT Department of Brain and Cognitive Sciences.

Shaping up

Since the researchers' system doesn't require annotation of the data on which it's trained, it needs to make a few assumptions about the structure of the data in order to draw coherent conclusions. One is that the frequency with which words occur in speech follows a standard distribution known as a power-law distribution, which means that a small number of words will occur very frequently but that the majority of words occur infrequently—the statistical phenomenon of the "long tail." The exact parameters of that distribution—its maximum value and the rate at which it tails off—are unknown, but its general shape is assumed.

The key to the system's performance, however, is what Lee describes as a "noisy-channel" model of phonetic variability. English may have fewer than 50 phonemes, but any given phoneme may correspond to a wide range of sounds, even in the speech of a single person. For example, Lee says, "depending on whether 't' is at the beginning of the word or the end of the word, it may have a different phonetic realization."

To model this phenomenon, the researchers borrowed a notion from communication theory. They treat an audio signal as if it were a sequence of perfectly regular phonemes that had been sent through a noisy channel—one subject to some corrupting influence. The goal of the machine-learning system is then to learn the statistical correlations between the "received" sound—the one that may have been corrupted by noise—and the associated phoneme. A given sound, for instance, may have an 85 percent chance of corresponding to the 't' phoneme but a 15 percent chance of corresponding to a 'd' phoneme.

"We compared two models, one that models phonetic variability and one that doesn't, and there's a huge difference," Lee says.

The researchers tested their system on six different recordings of lectures given at MIT and found that it was able to accurately identify the words used most frequently in each. There were some aberrations, however. In analyzing one of the lectures, delivered by The New York Times columnist Thomas Friedman, the system concluded that "open university" was a single word.

That's probably because Friedman not only used the term repeatedly, but rarely used either of its constituents. "If it observed 'open' and 'university' separately, then you may be able to discover that 'open' and 'university' are two words," Lee says.

"Recent experimental research points to the fact that infants learn phonemes and words simultaneously," says Emmanuel Dupoux, director of the Laboratory of Cognitive and Psycholinguistic Sciences, which is associated with the Ecole des Hautes Etudes en Sciences Sociales in Paris. "Up to now, however, only a handful of studies have modeled the interaction between these two levels using a machine-learning approach.

"Past studies have either studied only one side of the interaction—phoneme to words, or vice versa—or, when they modeled the full interaction, it was on a 'toy' version of the problem, with only a few phonemes and words. The Glass and Lee study is the first one to tackle the full interaction at scale, using a large speech corpus. The technical difficulties in doing this are enormous, and hence their achievement is a real tour de force."

More information: "Unsupervised Lexicon Discovery from Acoustic Input." *Transactions of the Association for Computational Linguistics*. Vol 3 (2015). tacl2013.cs.columbia.edu/ojs/i … acl/article/view/520