Dialect Detectives

April 16 2009, by Dorothy Ryan

Pedro Torres-Carrasquillo is working on techniques for machine-based identification of dialects in a spoken language. Photo / Jon Barron; Lincoln Laboratory

(PhysOrg.com) -- Technology under development by Pedro Torres-Carrasquillo and his colleagues at Lincoln Laboratory may lead to a dialect identification system that compensates for a translator's inexperience with multiple variants of a spoken language.

A law enforcement agency intercepts an international phone call alerting a suspected drug dealer to a new shipment. While the translator listening to the message is confident the caller's Spanish carries a South American accent, he cannot pinpoint a more specific region for agents to put under surveillance. But technology under development by Pedro Torres-Carrasquillo and his colleagues at Lincoln Laboratory may lead to a dialect identification system that compensates for a translator's

inexperience with multiple variants of a spoken language.

Language identification systems that can recognize as many as 29 languages from written text are already marketed, and systems that can identify a spoken language from a prescribed range of choices also exist. So far, however, no system that automatically discriminates one spoken dialect from another is available.

Lincoln Laboratory's earlier work on dialect identification focused on building models that mapped the audiowave frequencies of phonemes - the individual sounds of a spoken language. Torres-Carrasquillo, an electrical engineer specializing in speech processing in the laboratory's Information Systems Technology Group, says his group has more recently moved from this phonetic-based approach to lower-level acoustic systems that use the basic spectral similarities of small pieces of spoken utterances. "We are not looking for the types of data linguists deal with - larger units such as phonemes and words," he says. "We're looking at the statistical distributions of basic frequency spectra of small pieces of sounds."

The laboratory researchers are building a model that classifies the training data, finding markers that discriminate the frequency characteristics of the data. Previously, Torres-Carrasquillo says, the approach was to "get a lot of examples, and then build a model that looks like your examples." But he is tackling the problem in a different way. "Our group's idea is that we don't need a model that looks like our data - we need a model that can classify our data," he explains. "We take very small pieces - snippets of speech - turn them into frequencies, add up all these contributions, and make a model that can tell them apart. We're looking for patterns from just milliseconds of speech."

The researchers are using pattern recognition and classification methods known as support vector machines (SVMs) and Gaussian Mixture

Models (GMMs) that use models trained to emphasize the more distinctive tiny features seen in the frequency patterns of small pieces of the dialects in question. The trained GMMs have the edge in accuracy, but SVMs are "an order of magnitude faster than the GMM," according to Torres-Carrasquillo. Even more effective than either SVMs or GMMs alone, he says, is combining the two techniques. In a test to discriminate general American English from Indian-accented English, for example, the error rate was 10 percent when GMM was used alone, 15 percent for SVM alone - and only 7 percent for a fusion of GMM and SVM.

To be incorporated into an automatic machine translation system, a dialect identification system would have to be able to recognize a dialect without having to process lengthy strings of speech data. Torres-Carrasquillo's goal is to be able to determine a speaker's dialect by categorizing discrete, characteristic markers in the snippets, and then create a model without using large sets of training data. "We'd love to see a short-term spectrum characteristic that is a strong discriminator, is very pervasive in the dialect, and that could be reliably detected in a sample," he says.

Finding this characteristic is a tall order. "You're not going to have a single spectrum characteristic that gives away the identification," Torres-Carrasquillo says. The linguistic differences between dialects of a language are often small; for example, vowel sounds in Cuban Spanish are slightly longer than those of Puerto Rican Spanish. The subtle differences between the spectral pictures of dialects are difficult to detect, especially in the milliseconds of speech used in the Laboratory experiments. "But as you look at the data" says Torres-Carrasquillo, "the differences start to pile up and you have a profile." The Laboratory's work to classify dialect differences, which Torres-Carrasquillo presented at a September 2008 speech communication and technology conference in Australia, may lead to the discovery of a strategy for any dialect problem - a global approach that could be exploited for various classes

of dialects instead of a method that works only for specific dialects.

The Lincoln Laboratory research on dialect identification may contribute to approaches for language identification more generally, but Torres-Carrasquillo offers a caveat: "The differences one can exploit within two dialects are very specific - maybe too specific to be applicable to language ID." Still, when a universal machine translation system arrives on the scene in some future decade, it may well depend on Lincoln Laboratory research to ensure that nuances of meaning conveyed in dialects are not lost in translation.

Provided by Massachusetts Institute of Technology (news : web)