Modeling Pronunciation Variation in Automatic Speech Recognition

Karen Livescu, Toyota Technological Institute at Chicago

The performance of automatic speech recognition systems varies widely across different contexts. Very good performance can be achieved on single-speaker, large-vocabulary dictation in a clean acoustic environment, as well as on very small vocabulary tasks (such as digit recognition) with fewer constraints on the speakers and acoustic conditions. In other domains, such as meeting transcription or television closed-captioning, speech recognition is still far from practical. What makes these tasks elusive is, in part, the presence of unconstrained conversational speech. This type of speech poses a number of challenges, such as the presence of disfluencies, a mix of speech and non-speech sounds such as laughter, and extreme variation in pronunciation. In this talk, I will focus on the challenge of pronunciation variation and on new statistical models for speech recognition that may allow us to better handle such variation.

When speaking naturally, we often pronounce words in ways that differ wildly from their dictionary pronunciations. A number of analyses suggest that this variability is responsible for a large part of the drop in recognition performance between read (dictated) speech and conversational speech. I will begin by describing some of the challenging pronunciation phenomena and efforts in the speech recognition community to characterize and model them statistically.

The majority of work has focused on expanding the set of allowed pronunciations for each word. Word pronunciations are typically represented as strings of phones (basic speech units), the same representation used by most linguists. This representation allows us to take advantage of well-developed methods for statistical modeling of string sequences, including finite-state transducers and hidden Markov models. It has been surprisingly challenging, however, to account for all of the possible pronunciation variants, or even a significant portion of them, using phone-based models.

I will advocate an alternative view: that the phone unit may not be the most appropriate for modeling pronunciation variation. I will describe alternative models, using both larger units (e.g. syllables) and smaller units (e.g. articulatory features such as the positions of the lips, tongue, etc.). I will present some recent related ideas from linguistics, as well as statistical and learning tools, based on graphical models, that are needed for these more complex model structures. At this point it is not clear what the ``winning" approach will be. However, this is an exciting time to study such challenges, because recent advances in graphical models allow us to consider a much larger set of model structures than ever before. I will describe some of the current challenges and ongoing work, with a particular focus on the interaction between phonological theories and new statistical models.