# Invited Talk:
# Phonological Models in Automatic Speech Recognition

**Karen Livescu**
Toyota Technological Institute at Chicago
1427 E. 60th St., Chicago, IL 60637
`klivescu@tti-c.org`

## Abstract

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 with much fewer constraints on the speakers and acoustic conditions. In other domains, speech recognition is still far from usable for real-world applications. One domain that is still elusive is that of spontaneous 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. 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 describe efforts in the speech recognition community to characterize and model pronunciation variation, both for conversational speech and in general. The work can be roughly divided into several types of approaches, including: augmentation of a phonetic pronunciation lexicon with phonological rules; the use of large (syllable- or word-sized) units instead of the more traditional phonetic ones; and the use of *smaller* units, such as distinctive or articulatory features. Of these, the first is the most thoroughly studied and also the most disappointing: Despite successes in a few domains, it has been difficult to obtain significant recognition improvements by including in the lexicon those phonetic pronunciations that appear to exist in the data. In part as a reaction to this, many have advocated the use of a "null pronunciation model," i.e. a very limited lexicon including only canonical pronunciations. The assumption in this approach is that the observation model—the distribution of the acoustics given phonetic units—will better model the "noise" introduced by pronunciation variability.

I will advocate an alternative view: that the phone unit may not be the most appropriate for modeling the lexicon. When considering a variety of pronunciation phenomena, it becomes apparent that phonetic transcription often obscures some of the fundamental processes that are at play. I will describe approaches using both larger and "smaller" units. Larger units are typically syllables or words, and allow greater freedom to model the component states of each unit. In the class of "smaller" unit models, ideas from articulatory and autosegmental phonology motivate multi-tier models in which different features (or groups of features) have semi-independent behavior. I will present a particular model in which articulatory features are represented as variables in a dynamic Bayesian network.

Non-phonetic pronunciation models can involve significantly different model structures than those typically used in speech recognition, and as a result they may also entail modifications to other components such as the observation model and training algorithms. At this point it is not clear what the "winning" approach will be. The success of a given approach may depend on the domain or on the amount and type of training data available. I will describe some of the current challenges and ongoing work, with a particular focus on the role of phonological theories in statistical models of pronunciation (and vice versa?).