Student: Minshu Zhan
Faculty Supervisor: Stefanie Shattuck-Hufnagel
Date: September 13, 2012

Modeling Variation of Acoustic Cues in Speech Production

One of the central components in Automatic Speech Recognition (ASR) is the acoustic model, by which the intended utterance uncovered from the sound signal. The continuous sound is first converted to possible sequences of phonological units, or phonemes, which are then mapped to possible words using a pronouncing dictionary. In current speech recognition technology the most commonly used method to identify phonemes is the Hidden Markov Model, in which the input sound signal is examined in minimal fixed-length frames, each of which is considered an output from a hidden state determined by the intended phoneme, and from these outputs the most probable chain of phonemes is inferred. With existing algorithms and computation power, this frame-based approach has worked generally well and achieved the state-of-art performance of speech recognition but also has its limits due to the model's assumptions. Motivated by evidence that each

phoneme is characterized by a set of contrasting features, each of which is identifiable with acoustic cues in the sound signal, an alternative approach is emerging over the past few decades, which proposes an additional level of "acoustic cue" recognition between the sound signal and perceived phonemes. By applying additional phonological rules, this approach can potentially lead to a more intelligent and "natural" speech recognition system.

Over the years the Speech Communication Group at RLE has been developing such a cue-based speech recognition system, which consists of three stages. First, one particular subset of the acoustic cues, called landmarks, are detected by observing the valleys, peaks, and discontinuities in particular frequency bands of the signal. The type of the landmark determines the location of the phoneme and to which broad category a phoneme belongs, i.e. whether it is a vowel or a consonant. Then, additional cues are extracted by examining selective acoustic parameters at regions near the landmarks. Finally combining the cues and the speech context, the model outputs the most probable phonemes corresponding to each landmark.

Our current focus is to capture the variations of the acoustic cues, which we believe vary systematically with speech context and individual speakers. Following is our basic strategy: 1) collect speech samples from a diverse group of English speakers ; 2) manually label acoustic cues in the utterances and record corresponding contextual information such as phrasing, prominence, and word frequencies; 3) predict acoustic cues from the utterances based on existing phonological rules; 4) compare the observed and predicted cues with regard to context, analyze the result using various machine learning methods as well as direct inspection, and develop a probabilistic model.

Specifically, my task for this year long project includes two components: 1) implement the representation for speech data which incorporate utterances, acoustic labels, and contextual information, and develop tools required for processing the data; 2) experiment with different machine learning methods on given speech data and produce an optimal model. For component (1), I have completed a preliminary, landmark-specific framework over the summer, including landmark parser, landmark predictor, landmark comparator, and context extractor. For component (2), I have started applying preliminary

results to a decision tree script. During the fall semester, I expect to complete and fully test the landmark-specific programs, apply several machine methods to the available data, and start extending the data structure to other acoustic cues; in the spring, I will aim at completing the software package for all acoustic cues, improve its usability, and test the performance of various probabilistic models. A possible extension of the project would be integrating my programs with existing cue detection programs and trying to build an actual speech recognizer. I will also be responsible to maintain a human-readable documentation for the comparison results of the data for the purpose of direct inspection by researchers at the lab.

Sound perception has been the field that intrigues me the most; I also have a strong curiosity about artificial intelligence. Therefore it really excites me to be able to contribute automatic speech recognition research project with applications of machine learning methods. As I aspire to continue to study in graduate school, this project is also a valuable long-term research experience from which I expect to acquire essential research skills.