

ABSTRACT

Low-proficiency non-native speakers represent a significant challenge for large-vocabulary continuous speech recognition (LVCSR). Acoustic models are confused by a heavy accent; language models are confused by poor grammar and unconventional word choice. Lack of comfort with the spoken language affects the fundamental properties of connected speech that have been a focus of LVCSR research; cross-word and interword coarticulation, disfluency, and prosody are among the features that differ in native and non-native speech.

In this dissertation, I first address the problem of *characterizing* low-proficiency non-native speech. One population is examined in great detail: learners of English whose native language is Japanese. Properties such as fluency, vocabulary, and pace in read and spontaneous speech are measured for both general and proficiency-controlled data sets. I further show that native and non-native speech can be distinguished using a variety of statistical metrics, including perplexity and Kullback-Leibler divergence. Patterns in reading errors and grammaticality of spontaneous speech are quantitatively described. This analysis, while focusing on one speaker population, provides a model for characterizing non-native speech that the broader LVCSR community may find useful. The generalizability of this model is demonstrated by contrasting the speech of native speakers of Mandarin with that of our primary speaker set.

Second, I explore methods of *adapting* to non-native speech. The test set is controlled for language exposure and proficiency, and the task is a simplified read news task tailored toward the lower-proficiency speakers, who experienced limited success in more difficult reading tasks like the widely-used Wall Street Journal readings. I find that the largest gains in recognition performance come through acoustic adaptation, and present evaluations of adaptation and training techniques incorporating native-language and accented data. From a speaker-adapted baseline of 63.1% (the same models perform at 8% for Broadcast News F0 speech), a 29% relative improvement is achieved through a combination of adaptation and training. In contrast, gains from lexical modeling were found to be extremely small, even when investigated in conjunction with retraining. I describe data-driven and linguistically-motivated algorithms for lexical modeling, presenting experimental results and discussing possible reasons why the improvement was not larger.

Finally, I present a novel method for detecting non-native speech. Without using any acoustic features, I show how bilateral and multilateral discrimination can be accomplished on the basis of features present in text. Both recognizer output and transcripts of non-native speech are identified with high accuracy through naive Bayes classification. The word and part-of-speech sequences that are found to be indicative of non-native speech provide an additional resource for characterizing non-native speech, which leads to further insights about the properties of non-native spoken language.