

## Abstract

Automatic recognition of continuous speech has been acknowledged as one of the most challenging problems today. The performance of a continuous speech recognition system highly depends on the availability of sufficient speech data and transcripts of good quality. In most cases, however, carefully prepared in-domain data is not easy to obtain because collecting a large amount of transcribed speech data is normally a time-consuming and expensive process. The acoustic model trained without the support of sufficient training data is less capable in handling the complexity and variability of human speech, and thus performs poorly in real world application. This raises us the questions such as how to effectively exploit the given training data to improve the performance of recognition systems, and how to explore error-prone but informative data sources and incorporate them into acoustic model training.

This thesis summarizes our efforts in investigating solutions to address the above issues. The work can be divided into two parts. We first investigate Boosting algorithm, an ensemble based supervised training approach, which iteratively creates multiple acoustic models with complementary error patterns by manipulating the distribution of training data. While a great deal of research has been conducted on Boosting style acoustic model training, the techniques developed so far have obvious weaknesses that limit their applications in continuous speech recognition. Specifically, conventional Boosting training approach mainly targets at optimizing an utterance level objective function related to sentence error, with relatively less attention paid to reduce word errors. Moreover, the distribution of training data is updated on a sentence basis such that each word is given equal weight in subsequent model training, regardless if the questioned word is a correct or incorrect decoding result. We approach these problems by presenting a novel frame level Boosting algorithm which enables acoustic model training to minimize word or sub-word error rate. We also present an improved sentence hypothesis combination algorithm that uses Neural Nets to incorporate a number of features for generating more desirable combination results. Moreover, we describe the contribution of N-best list re-ranking in Boosting training and hypothesis combination, which is shown to be an effective approach to improve recognition performance.

The second part of this thesis focuses on unsupervised and lightly supervised training techniques that attempt to extend the training set from transcribed speech to closed captioned or even un-transcribed raw speech. Data selection, the approach to identify a subset of data that can best improve recognition performance, appears to be the key issue in this research area. Most conventional approaches prefer to select data predicted with high confidence for model re-training in order to prevent misrecognized examples from being added to training set. However, this strategy often results in only the examples that match well to the current model being selected and re-training with such examples can become a process that reinforces, rather than eliminates, the estimation bias inherited from the initial transcribed set. To address this problem, we present a novel clustering based data selection strategy that aims to increase the diversity of selected data and makes the selection comply with underlying distribution. Experiments show that the new proposed strategy consistently outperforms conventional approaches in a variety of speech recognition tasks. In addition, we investigate the generalization of Boosting algorithm from supervised training to unsupervised training by using Minimum Bayes Risk decoding, as

well as its integration with clustering based data selection for better handling both transcribed and untranscribed speech data. Experimental results show that the cooperation of data selection and unsupervised Boosting can significantly improve recognition performance.