Public Abstract

First Name:Xin

Middle Name:

Last Name:Chen

Adviser's First Name:Yunxin

Adviser's Last Name:Zhao

Co-Adviser's First Name:

Co-Adviser's Last Name:

Graduation Term:FS 2011

Department:Computer Science

Degree:PhD

Title:Ensemble Acoustic Modeling in Automatic Speech Recognition

In this dissertation, several new approaches of using data sampling to construct an Ensemble of Acoustic Models (EAM) for speech recognition are proposed. A straightforward method of data sampling is Cross Validation (CV) data partition. In the direction of improving inter-model diversity within an EAM for speaker independent speech recognition, we propose Speaker Clustering (SC) based data sampling. In the direction of improving base model quality as well as inter-model diversity, we further investigate the effects of several successful techniques of single model training in speech recognition on the proposed ensemble acoustic models, including Cross Validation Expectation Maximization (CVEM), Discriminative Training (DT), and Multiple Layer Perceptron (MLP) features. We have evaluated the proposed methods on TIMIT phoneme recognition task as well as on a telemedicine automatic captioning task. The proposed EAMs have led to significant improvements in recognition accuracy over conventional Hidden Markov Model (HMM) baseline systems, and the integration of EAM with CVEM, DT and MLP has also significantly improved the accuracy performances of CVEM, DT, and MLP based single model systems. We further investigated the largely unstudied factor of inter-model diversity, and proposed several methods to explicit measure inter-model diversity. We demonstrate a positive relation between enlarging inter-model diversity and increasing EAM quality.

Compacting the acoustic model to a reasonable size for practical applications while maintaining a reasonable performance is needed for EAM. Toward this goal, in this dissertation, we discuss and investigate several distance measures and proposed global optimization algorithms for clustering methods. We also proposed an explicit PDT (EPDT) state tying approach that allows Phoneme data Sharing (PS) for its potential capability in accommodating pronunciation variations.