SESSION 5a: ACOUSTIC MODELING

Hy Murveit, Chair

SRI International
Speech Research and Technology Program
Menlo Park, CA, 94025

1. SUMMARY OF PAPERS

The session focused on acoustic modeling for speech recognition; which can be segmented into three broad subareas: (1) feature extraction, (2) modeling the features for the speech source, and (3) estimation of the model parameters. The papers in this session touched on all of these areas. Huang focuses on the feature representation. Furui et al., Austin et al., and Kimbal et al. discuss new models for the speech source. Hon and Lee, Hwang and Huang, and Gauvain and Lee focus on parameter estimation issues.

In "Minimizing Speaker Variation Effects for Speaker-Independent Speech Recognition," Huang discusses a feature representation for speech recognition that is less sensitive to the speaker. It is a cepstral mapping technique where the mapping is done with neural networks. A codeword-dependent cepstral-mapping network is estimated for each of a group of different speaker types. This cepstral mapping improves the speaker-independent performance of the CMU system.

In "Recent Topics in Speech Recognition Research at NTT Laboratories," Furui et al. discuss three topics. This first topic focuses on an improved model for speech. Typical HMM recognition systems make frame-to-frame independence assumptions. Furui presented a technique aimed at minimizing this effect, using bigram-constrained HMMs, and showed an improvement when using this technique. He also discussed two issues in language modeling, one specific to Japanese, and another showing how task-independent language models can be adapted to a task at hand.

Austin et al. ("Improving State-of-the-Art Continuous Speech Recognition Systems Using the N-Best Paradigm with Neural Networks") and Kimbal et al. ("Recognition Using Classification and Segmentation Scoring") discuss segment-level models for the speech source. Austin points out that neural networks can be combined with HMMs to automatically derive segment-level acoustic models that reduce the effect of frame-to-frame independence assumptions in standard HMMs. He shows that proper parameter estimation techniques are key for these models and presents a technique called N-best training which improves the performance of his segmental model. Kimbal focuses on the segmentation aspects of segmental models. He shows that incorporating a probabilistic segmentation model improves the performance of the Boston University speech recognition system.

The following three papers discuss the area of parameter estimation. "Vocabulary and Environment Adaptation in Vocabulary-Independent Speech Recognition" by Hon and Lee revisits the area of task independence, but this time from an acoustic point of view. Traditional HMM-based speech-recognition systems work much better if their acoustic training data use the same task/vocabulary as the testing data. Hon and Lee look at techniques for making the training data more general. In particular, they examine novel techniques that improve vocabulary-independent performance by making the parameter-estimation technique focus on the testing vocabulary.

Hwang and Huang, in "Subphonetic Modelling for Speech Recognition," discuss another parameter-estimation issue-, the issue of tied models. Most large-vocabulary speech-recognition systems must tie together estimates of certain parameters that would not otherwise have sufficient training data to be estimated accurately. Often this is done in a phonetic way. For instance, the same allophone or allophone-in-context in different words would share parameters. Hwang and Huang describe a technique (similar to the IBM phenome technique) for automatically deriving the units to be tied.

In "MAP Estimation of Continuous Density HMM: Theory and Applications," Gauvain and Lee discuss a parameter estimation technique based on Bayesian learning. They show that it is useful for parameter smoothing as well as for speaker adaptation and discriminative training. In speaker adaptation, speaker-independent models can be moved to a speaker using a small amount of training since the speaker-independent models are used as priors. Adapted speaker-independent performance never performed worse than speaker-dependent systems given the same amount of speaker-dependent training data.