Session 11: Acoustic Modeling and Robust CSR

Steve Young

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, England

The papers in this session are all concerned with improving the performance of speech recognition systems at the acoustic modeling level. Three papers focus on obtaining very high accuracy under clean matched conditions and the remaining papers deal with the problems of mis-match caused by the use of different microphones and environmental noise.

The first three papers focus on the problem of building very accurate Hidden Markov Models (HMMs) for use in large vocabulary speech recognition systems. Such systems require a very large number of context-dependent models, but the available training data is limited and unevenly spread. Hence, some way has to be found to share the data amongst the models. Furthermore, the approximations inherent in the use of tied-mixture systems has led to a return to more conventional continuous density mixture Gaussian systems. In their basic form, these have a much greater number of parameters and hence, the data insufficiency problem is exacerbated. The solution adopted in the first two cases is to tie states together in order to balance the available data against model complexity. In the third paper, MAP estimation is used in conjunction with generalised triphones.

The first paper, "Tree-based State Tying for High Accuracy Acoustic Modelling" by *Young, Odell & Woodland*, describes a state-tying method based on phonetic decision trees. They use an incremental approach in which untied single Gaussian models are estimated for every triphone appearing in the training data, then states within allophone sets are tied and finally, the tied single Gaussians are converted to mixture Gaussians. The key attractions of this approach are that firstly it is very efficient since the objective function used for node splitting depends only on the model parameters and not on the original data, and secondly, once built, the decision trees can be used to synthesise all of the *unseen* triphones needed by the recogniser but not seen in the training data. The technique described was used to build the HTK Large Vocabulary Recogniser. This was included in the November 1993 Wall Street Journal evaluation where it returned the lowest error rate in three of the four tests and the second lowest error rate in the fourth.

The second paper, "High-Accuracy Large-Vocabulary Speech Recognition using Mixture Tying and Consistency Modeling" by *Digilakis & Murveit* follows a similar approach. The system here is also built incrementally, however, in this case, HMMs are built starting from a heavily tied system and then mixtures are successively untied using agglomerative clustering followed by a splitting and pruning procedure. The paper also contains an interesting study on time correlation modelling and the use of linear discriminant analysis.

The third paper, "The LIMSI Continuous Speech Dictation System" by *Gauvain, Lamel, Adda & Adda-Decker*, describes a continuous density HMM system which has a more conventional generalised triphone structure but which is trained using MAP estimation. This yielded the lowest error rate of any 20k word trigram system in the November 1993 WSJ evaluation. In addition to WSJ, the paper also describes results for French using the BREF corpus. This leads to some interesting conclusions on the properties of the two languages. In particular, it would appear that although higher phoneme recognition rates are achieved for French, the higher homophone rate reduces word recognition to similar levels as for English.

The systems described in these three papers all use continuous density mixture Gaussian HMMs. Whilst these seem to yield the best performance under ideal conditions, for robust near real-time operation, tied-mixture systems continue to offer many advantages. In the fourth paper of the session, "Adaptation to New Microphones using Tied-Mixture Normalisation" by *Anastasakos, Kubala, Makhoul & Schwartz*, a technique is described in which a probabilistic function is used to map the codebook built in training into one suitable for the required new microphone. This mapping function is built using a small amount of stereo adaptation data collected using the new microphone. When combined with cepstral mean subtraction, the technique substantially reduced the effects of microphone mis-match.

The fifth paper, "Signal Processing for Robust Speech Recognition" by *Liu, Moreno, Stern & Acero* reviews a number of cepstral-based compensation procedures

which have been studied within the Sphinx-II recognition system developed at Carnegie Melon University. These include a phone-dependent cepstral normalisation scheme in which compensation vectors are estimated from training data consisting of clean speech and corresponding noisy speech. Other techniques studied include VQ codebook adaptation, reduced-band analysis for telephone speech and silence codebook adaptation. A variety of experimental results are presented for the Microphone Independence (Spoke 5) and the Calibrated Noise Sources (Spoke 8) November 1993 WSJ evaluations. Error reductions of around 40% were typically achieved using combinations of these techniques.

The sixth paper, "Microphone-Independent Robust Signal Processing using Probabilistic Optimum Filtering" by *Neumeyer & Weintraub*, also addresses the problems caused by having a mis-match between the microphone used for training a speech recognition system and the microphone used for testing. In this case, the solution proposed is to replace each incoming noisy speech vector with an estimate of the underlying clean speech vector. This mapping is achieved using a piecewise non-linear transformation estimated from stereo training data. A large number of results are presented both for ATIS and WSJ.

The seventh and final paper, "Microphone Arrays for Robust Speech Recognition" by *Che, Lin, Pearson, de Vries & Flanagan*, takes a very different approach to dealing with environmental robustness. In this case, the input signal is pre-processed using a beamforming microphone array followed by a neural network trained on a small amount of the noisy speech data. In various tests using the Sphinx-I recognition system, the combination of array and neural network in a reverberent acoustic environment was able to achieve similar performance to that obtained using a close-talking microphone.