Presented at JOINT AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Centre du Parc, Martigny, Switzerland, 21-23 June 2004.

DETECTION OF SPEECH AND CROSSTALK IN MULTI-CHANNEL MEETING RECORDINGS

Stuart N Wrigley[1], Guy J Brown[1], Vincent Wan[1], and Steve Renals[2]
{s.wrigley,g.brown,v.wan }@dcs.shef.ac.uk, s.renals@ed.ac.uk

[1] Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, UK
[2] Centre for Speech Technology Research, University of Edinburgh, UK

ABSTRACT

The analysis of scenarios in which a number of microphones record the activity of speakers, such as in a round-table meeting, presents a number of computational challenges. For example, if each participant wears a microphone, it can receive speech from both the microphone's wearer (local speech) and from other participants (crosstalk). The recorded audio can be broadly classified in four ways: local speech, crosstalk plus local speech, crosstalk alone and silence. We describe two experiments related to the automatic classification of audio into these four classes. The first experiment attempted to optimise a set of acoustic features for use with a Gaussian mixture model (GMM) classifier. A large set of potential acoustic features were considered, some of which have been employed in previous studies. The best-performing features were found to be kurtosis, 'fundamentalness' and cross-correlation metrics. The second experiment used these features to train an ergodic hidden Markov model (eHMM) classifier. Tests performed on a large corpus of recorded meetings show classification accuracies of up to 96%, and automatic speech recognition performance close to that obtained using ground truth segmentation.