Speech and Crosstalk Detection in Multi-Channel Audio

Stuart N. Wrigley†, Guy J. Brown†, Vincent Wan† and Steve Renals‡

† Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, UK
‡ Centre for Speech Technology Research, University of Edinburgh, UK

{s.wrigley,g.brown,v.wan}@dcs.shef.ac.uk
s.renals@ed.ac.uk

http://www.m4project.org

MULTIMODAL MEETING MANAGER

Introduction

The objective of the M4 (multimodal meeting manager) project is to produce a demonstration system to enable structuring, browsing and querying of an archive of automatically analysed meetings recorded in a room equipped with multimodal sensors.

Significant amount of ***crosstalk*** (non-local speech being received by the local microphone) makes ASR and turn detection difficult.

Objective 1: to produce a classifier which can label each lapel microphone signal using four high-level activity categories:

- local channel speaker alone (*speaker alone*)
- local channel speaker concurrent with one or more other speakers (*speaker+crosstalk*)
- one or more non-local speakers (*crosstalk alone*)
- no speakers (*silence*)

Objective 2: investigate range of possible features and determine which combination provides the optimum classification performance for each category

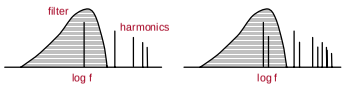
Candidate features

16ms Hamming window with 10ms frame shift unless otherwise stated.

MFCC, **energy** and **zero crossing rate**: conventional feature set.

Time-domain Kurtosis: Kurtosis is related to the size of a distribution's tails and can be used as a measure of Gaussianity. The kurtosis of overlapping speech is generally less than the kurtosis of the individual speech utterances, (160ms window size).

Fundamentalness: based on AM and FM extracted from wavelet analysis of the signal log frequency spectrum.



Candidate features

Frequency-domain kurtosis: kurtosis of the magnitude spectrum.

SAPVR (spectral autocorrelation peak valley ratio): ratio of peaks to valleys within the autocorrelation of the signal spectrum.

PPF (pitch prediction feature): smoothed LPC error signal subjected to a form of autocorrelation analysis which identifies periodicities (between 50 Hz and 500 Hz). PPF measure is defined as the standard deviation of the differences between potential pitch peaks extracted from the autocorrelation function, (30ms window size).

Genetic programming: fft, min, max, kurtosis, autocorr, normalize, etc. 1000 individuals, mutation rate of 0.5%, crossover rate of 90%. Individuals evaluated using a Gaussian classifier. GP engine identified several successful features, such as max(autocorr(normalize(x))), which were included in the feature selection process.

Cross-channel correlation: for each channel *i*, the cross-channel correlation was computed between channel *i* and all other channels. From these, the unnormalised and normalised minimum, maximum and mean values were extracted and used as individual features. Two forms of normalisation were used: energy normalisation and spherical normalisation.

Feature selection algorithm

Sequential forward selection (SFS) using the area under the ROC (receiver operating characteristic) curve (AUROC) for a particular GMM (Gaussian mixture model) as the performance measure:

Compute the AUROC for each individual feature.

Feature with the highest AUROC is added to currently empty feature set.

Retrain GMMs using this feature set and each remaining feature.

Select feature set resulting in the highest AUROC.

Terminate when the gain in AUROC is less than 1%.

Results – full set

Speaker alone: kurtosis and max norm xcorrelation.

Speaker+crosstalk: energy, kurtosis, max norm xcorrelation and mean s-norm xcorrelation.

Crosstalk alone: energy, kurtosis, mean xcorrelation, mean norm xcorrelation, max s-norm xcorrelation.

Silence: energy and mean xcorrelation.

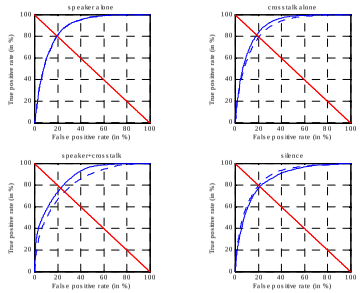
Results – no energy

Speaker alone : kurtosis and max norm xcorrelation.

Speaker+crosstalk : kurtosis, fundamentalness, max norm xcorrelation and mean s-norm xcorrelation.

Crosstalk alone : mean xcorrelation and mean s-norm xcorrelation.

Silence : kurtosis, mean xcorrelation and mean s-norm xcorrelation.



Training data: 1M frames per category from four ICSI meeting recordings (*bro012*, *bmr006*, *bed008*, *bed010*).

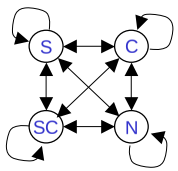
Test data: 15K frames per category from one ICSI meeting recording (*bmr001*).

Multichannel classification

Ergodic HMM (eHMM).

Each state corresponds to one of the four categories and is modelled by a GMM.

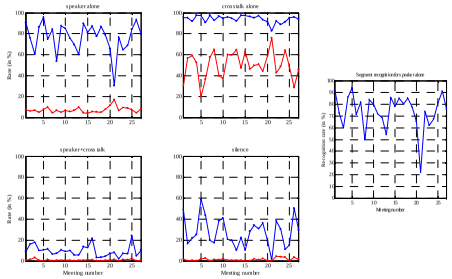
Each channel is classified by a different eHMM in parallel.



Dynamic transition constraints to ensure legal channel classification combinations.

Union of channel classification feature sets.

Multichannel classification results



Test data: 27 ICSI meetings: 1. *bed004*, 2. *bed006*, 3. *bed009*, 4. *bed011*, 5. *bmr001*, 6. *bmr002*, 7. *bmr005*, 8. *bmr007*, 9. *bmr008*, 10. *bmr009*, 11. *bmr012*, 12. *bmr013*, 13. *bmr014*, 14. *bmr018*, 15. *bmr024*, 16. *bmr026*, 17. *bro003*, 18. *bro004*, 19. *bro005*, 20. *bro007*, 21. *bro008*, 22. *bro011*, 23. *bro013*, 24. *bro015*, 25. *bro017*, 26. *bro018*, 27. *bro026*.

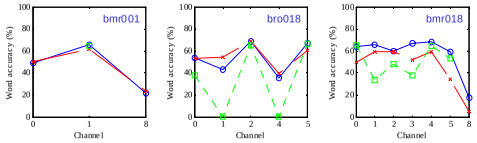
ASR evaluation

HTK trained on 40 hours of ICSI meetings data.

Recogniser has a word accuracy of approximately 50% on unseen data.

Compared speaker alone ASR performances when using

- ground truth segments
- eHMM segments
- energy-based voice activity detector segments



Conclusions

High performance on context free classification: approx 80% for equal error rates.

Segment based classification performance for speaker-alone has mean recognition rate of 74% with some meetings reaching 94%.

ASR performance using the eHMM segments is extremely similar to the ASR performance using the transcribed ground truth segments.

M4 is supported by the EU IST Programme (project IST-2001-34485).