Feature Selection for the Classification of Crosstalk 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

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

Abstract

An extension to the conventional speech / nonspeech classification framework is presented for a scenario in which a number of microphones record the activity of speakers present at a meeting (one microphone per speaker). Since each microphone can receive speech from both the participant wearing the microphone (local speech) and 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 a classifier in which a Gaussian mixture model (GMM) is used to model each class. A large set of potential acoustic features are considered, some of which have been employed in previous speech / nonspeech classifiers. A combination of two feature selection algorithms is used to identify the optimal feature set for each class. Results from the GMM classifier using the selected features are superior to those of a previously published approach.

1. Introduction

The objective of the M4 (multimodal meeting manager) project [1] is to produce a demonstration system to enable structuring, browsing and querying of an archive of automatically analysed meetings. The meetings take place in a room equipped with multimodal sensors. Audio information is acquired from lapel mounted microphones, microphone arrays and a KEMAR binaural manikin. Video information is captured from multiple cameras. Unfortunately, crosstalk (non-local speech being received by the local microphone) causes problems for tasks such as turn detection and automatic speech recognition (ASR). For example, when performing ASR, it is important to know which portions of the signal are uttered by the local speaker and hence not insert words from an intruding speaker. The ability to detect overlapping speech also allows corrupted regions of speech to be identified - regions that may not be recognised by an ASR system without pre-processing. Furthermore, patterns of speaker activity and overlap can provide valuable information regarding the structure of the meeting.

In this paper, we concentrate on the task of producing 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*)

Similar work on meetings-based signal classification [2] has concentrated on a speech / nonspeech identification task. The additional classes above increase the flexibility of the system and more closely guide future analysis (such as enhancement of crosstalk contaminated speech).

A primary objective of our approach is to obtain reliable classification regardless of the room in which the meeting takes place, the identities of the individual speakers and the overall number of participants. We note that previous approaches (e.g., [7]) have tended to be channel, speaker or environment dependent.

A secondary objective was to investigate a range of possible features which can be extracted from the audio signal and determine which combination provides the optimum classification performance for each category. This also allows us to compare our approach to other systems, which compute a subset of our features, using the same data set.

2. GMM-based classifier

Each class is modelled by a Gaussian mixture model (GMM) in order to capture the variability in the distributions caused by multiple speakers and differing channel characteristics.

2.1. Candidate features

Previous speech / nonspeech classifiers (e.g. [2]) have used features such as critical band loudness values, energy and zero crossing rate. In addition to these, we identified a number of features which are particularly suited to analysing the differences between isolated speech and overlapping speech.

Each feature was calculated over a 16 ms Hamming window with a frame shift of 10 ms, unless otherwise stated.

2.1.1. MFCC, energy and zero crossing rate

In order to compare our system with other speech / nonspeech classifiers, we included the conventional feature set of mel-frequency cepstral coefficients (MFCCs), log energy and zero crossing rate (for example, see [2]).

2.1.2. Kurtosis

Kurtosis is defined as the fourth central moment divided by the fourth power of the standard deviation. Thus, kurtosis is based on the size of a distribution's tails - i.e. a measure of Gaussianity. Kurtosis is zero for a Gaussian random variable and positive for super-Gaussian signals such as speech. It has been shown that the kurtosis of overlapping speech is generally less than the kurtosis of the individual speech utterances [3], since - in accordance with the central limit theorem - mixtures of speech signals will tend towards a Gaussian distribution. Here, a 160 ms window, centred on the same points as the 16 ms window, was used to allow a more accurate estimate of the short-time signal kurtosis. The frequency-domain kurtosis (i.e. the kurtosis of the magnitude spectrum) was also computed.

2.1.3. Fundamentalness

Kawahara *et al.* [4] describe an approach to estimating the 'fundamentalness' of a harmonic. Their technique is based on amplitude modulation (AM) and frequency modulation (FM) extracted from the output of wavelet analysis on the signal log frequency spectrum. At different positions in frequency, the analysing wavelet will encompass a different number of harmonic components; fundamentalness is defined as having maximum value when the FM and AM modulation magnitudes are minimum, which corresponds to the situation when the minimum number of components are present in the wavelet window (usually just the fundamental component). Although this technique was developed to analyse isolated speech (see [4] page 196, eqns 13-19), the concept that a single fundamental produces high fundamentalness is useful here: if more than one fundamental is present, interference of the two components will cause AM and FM modulation, thus decreasing the fundamentalness measure. Such an effect will arise when two speakers are active simultaneously, giving rise to overlapping harmonic series.

2.1.4. SAPVR

SAPVR (spectral autocorrelation peak valley ratio; [5]) is computed from the autocorrelation of the signal spectrum. The measure is the ratio of peaks to valleys within the spectral autocorrelation. Specifically, the SAPVR-5 measure [6] is the sum of the autocorrelation peaks, including lag zero, divided by the sum of the first two autocorrelation valleys. For a peak to be used, it must be an integer multiple of the fundamental (i.e. true harmonic). For single speaker speech, a strongly periodic autocorrelation function is produced due to the harmonic structure of the spectrum. However, when more than one speaker is active simultaneously, the autocorrelation function becomes flatter due to the overlapping harmonic series.

2.1.5. PPF

The PPF (pitch prediction feature) was developed for the specific task of discriminating between single speaker speech and two speaker speech [7]. The first stage computes 12-th order linear prediction filter coefficients (LPCs) which are then used to calculate the LP residual - the error signal. The residual is smoothed using a Gaussian shaped filter after which a form of autocorrelation analysis then identifies periodicities (between 50 Hz and 500 Hz). Potential pitch peaks are extracted by applying a threshold to this function. The final PPF measure is defined as the standard deviation of the differences between such extracted peaks. If a frame contains a single speaker, strong peaks will occur at multiples of the pitch period. Therefore, the standard deviation of the differences of the peaks will be small. Conversely, if the frame contains two speakers of different fundamental frequency, there will be a considerably larger number of strong peaks due to the overlapping harmonic series. Therefore, the standard deviation of the differences of the peaks will be much higher. In order to allow direct comparison between our approach and that of [7], a 30 ms window was used.

2.1.6. Features derived from genetic programming

A genetic programming (GP) approach (see [8] for a review) was also used to identify frame-based features that could be useful for signal classification. The GP engine was implemented in MATLAB and included a type checking system, thus allowing vector and scalar types to be mixed in the same expression tree. Intermediate results were cached to reduce computation time. The function set included standard MATLAB functions such as `fft`, `min`, `max`, `kurtosis`, and user defined functions such as `autocorr` (time-domain autocorrelation) and `normalize` (which scaled a vector to have zero mean and unity standard deviation). The terminal set contained integer and floating point constants, the signal vector *x*, and array indices. A population of 1000 individuals was used, with a mutation rate of 0.5% and crossover rate of 90%.

Individuals were evaluated by training and testing a Gaussian classifier on the features derived from each expression tree, using a subset of the data described in section 3. Successive generations were obtained using fitness-proportional selection. The GP engine identified several successful features, such as `max(autocorr(normalize(x)))`, which were included in the feature selection process. Interestingly, GP also discovered several features based on spectral autocorrelation, but these were never ranked highly.

2.1.7. Cross-channel correlation

A number of other features were extracted using 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. Normalisation consisted of dividing the feature set for channel *i* by the frame energy of channel *i*.

2.2. Feature selection

The parcel algorithm (see [9]) was used to assess the classification performance of the different feature combinations. For each feature combination, the classifier's GMMs are trained and then evaluated to create a receiver operating characteristic (ROC) curve for each crosstalk category. Each point on such a ROC curve represents the performance of a classifier with a different decision threshold between two crosstalk categories (i.e. the category of interest versus all others). Given a number of ROCs (one per feature combination), a maximum realisable ROC (MRROC) can be calculated by fitting a convex hull over the existing ROCs. Thus, each point on a MRROC represents the best feature combination for that class.

However, due to the size of the feature set, the calculation of each possible feature combination (2^N-1 combinations) is impractical. As an alternative to exhaustive search, the sequential forward selection (SFS) algorithm was employed to produce a sub-optimal feature set for each crosstalk category. SFS proceeds by calculating a measure of the GMM classification performance (in this case, the area under the ROC curve) for each individual feature. The winning feature is the one with the highest classification performance. Each remaining feature is combined with the winning feature and the performance is again computed. If the combined performance of the new feature and the previous winner is above a certain threshold (in this case, a 1% increase in the area under the ROC curve) it is added to the feature set. This process is repeated until no more features can be added. Despite specifying a maximum

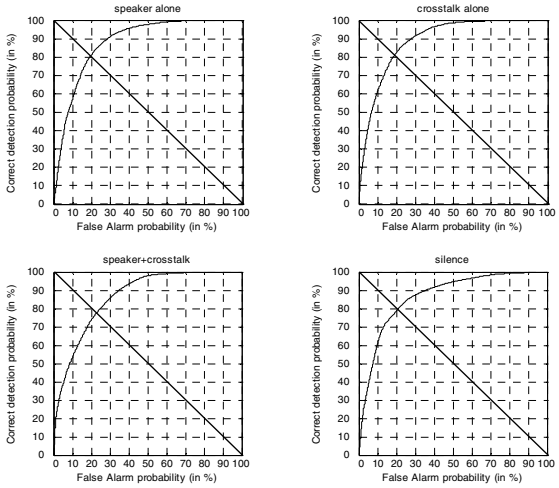

Figure 1: ROC performance curves for each crosstalk category's optimum feature set. Diagonal lines indicate equal error rates.

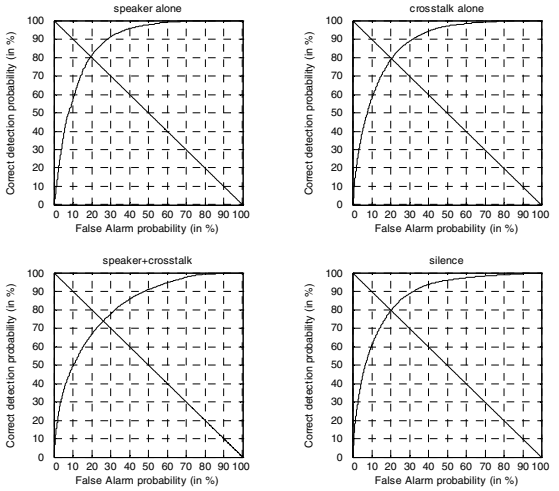

Figure 2: ROC performance curves for each crosstalk category's optimum feature set. Log energy has been excluded from the set of potential features. Diagonal lines indicate equal error rates.

feature set of six, the SFS algorithm terminated before reaching this size for all crosstalk categories.

3. Experiments and results

Since collection and annotation of M4 data has only recently begun, the following experiments were conducted using data from the International Computer Science Institute (ICSI) meeting corpus [10].

3.1. Feature selection using full feature set

The training data consisted of one million frames per category (16 ms window with 10 ms shift) of conversational speech extracted at random from four ICSI meeting recordings (*bro012*, *bmr006*, *bed008*, *bed010*). For each channel, a label file specifying the four different crosstalk categories (local speaker alone, local speaker plus crosstalk, crosstalk alone, background) was automatically created from the existing transcriptions provided with the ICSI corpus.

The test data consisted of 15000 frames per category extracted at random from one ICSI meeting recording (*bmr001*). As for the training data, a label file was automatically created from the existing transcriptions to assess the performance of the classifiers.

The feature sets derived by the SFS algorithm were:

- local channel speaker alone: kurtosis and maximum normalised cross-channel correlation.
- local channel speaker concurrent with one or more speakers: energy, kurtosis, maximum normalised cross-channel correlation and mean normalised cross-channel correlation.
- one or more non-local speakers: energy, kurtosis, mean cross-channel correlation, mean normalised cross-channel correlation, maximum normalised cross-channel correlation.
- no speakers: energy and mean cross-channel correlation.

It is interesting to note that MFCC features do not appear in any of the optimal feature sets.

The GMM performance of each feature set is shown in figure 1. For equal false positive and false negative misclassification rates, the performance of each classifier is approximately 80%.

3.2. Feature selection excluding energy feature

The results presented in the previous section were created using data from the ICSI corpus. However, it should be noted that there are differences between the data acquisition equipment used by ICSI and M4. Most notable is the type of microphone used for each participant. M4 uses lapel microphones as opposed to the head-mounted microphones used by ICSI. Hence, there may be significant differences between the resultant data sets. For example, channel energy may be an unreliable cue for M4 data since the lapel microphones allow a variable coupling (separation) between the mouth and the microphone. A drop in channel energy may simply be due to the local speaker turning their head rather than a change in speaker (in which case, the drop in energy is interpreted as a change from local speaker speech to crosstalk).

In order to give an indication of the classification performance when energy is removed from the set of potential features, the feature selection process described in section 2.2 was repeated with log energy excluded.

The feature sets derived by the SFS algorithm were:

- local channel speaker alone: kurtosis and maximum normalised cross-channel correlation.
- local channel speaker concurrent with one or more speakers: kurtosis, fundamentalness, maximum normalised cross-channel correlation and mean normalised cross-channel correlation.
- one or more non-local speakers: mean cross-channel correlation and mean normalised cross-channel correlation.
- no speakers: kurtosis, mean cross-channel correlation and mean normalised cross-channel correlation.

Figure 2 shows the GMM performance of each feature set. It can be seen that the removal of log energy has little effect on the feature set and overall classification performance of the system. Indeed, performance remains at approximately 80%. This is due to the high performance of the cross correlation features which dominate the ROC curves.

3.3. Comparison with a previous approach

As described in section 2.1.5, the pitch prediction feature (PPF) [7] was developed with the specific task of identifying portions of speech which contained either a single speaker or two speakers. Since this is clearly related to the crosstalk analysis task, a brief comparison between the performance of the system described here and the PPF is made. The same training and test sets are used as described above; the ROC curves for each crosstalk category are shown in figure 3. The performance reported in [7] ranges between 55% and 64% for unknown speakers, depending on the classifier used. Similar performance was obtained when using the PPF with our recogniser: approximately 64% for single speaker speech when considering equal error rates; approximately 59% for crosstalk speech. It ought to be noted that [7] makes no distinction between local speaker plus crosstalk and crosstalk alone - arguably making their task easier. However, the PPF performance here may be reduced due to the higher degree of background noise present in our data when compared to the data used in [7]. Despite this, the performance reported in this paper, and shown in figures 1 and 2, remains superior.

4. Conclusions and future work

A GMM-based approach to crosstalk analysis has been presented in which a wide range of potential features has been investigated. The parcel [9] and SFS algorithms were used to identify the feature set, for each crosstalk class, which had the highest performance. GMM classification performance for each of the four crosstalk classes is approximately 80% - significantly higher than other similar approaches (e.g. [7]).

Currently, the approach described here operates on individual frames of audio data. Temporal constraints could be effectively applied in the system using a hidden Markov model (HMM). Work is being conducted on the development of an ergodic HMM consisting of four states, one per class, in which each state is modelled by a GMM as described above. In addition to this, we are investigating the potential gain of using individual HMMs for each crosstalk category. Preliminary evaluation of the ergodic HMM classifier (with transition probabilities calculated from the training data and tested using ICSI meeting *bmr001*) suggest that it will yield a further improvement in performance.

Since M4 data transcription has recently begun, it will be necessary to assess the classifiers' performance on this data set and potentially adjust the feature sets. For example, as described above, it may become clear that an energy-based feature is not a reliable cue for crosstalk analysis.

Furthermore, work will also be conducted into the use of cross-channel features and cross-channel classification. For example, one channel's classification can be used to inform the classification process of another channel within the same meeting.

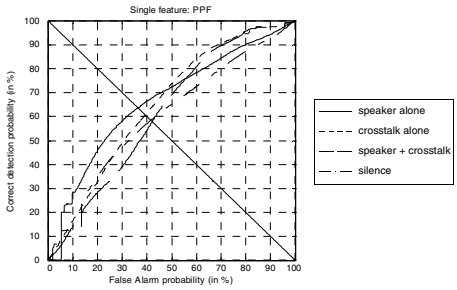

Figure 3: ROC performance curves for each crosstalk category using PPF. Diagonal line indicates equal error rates.

5. Acknowledgements

This work was conducted as part of the M4: multimodal meeting manager project [1] which is funded by the EU IST Programme (project IST-2001-34485).

6. References

[1] http://www.dcs.shef.ac.uk/spandh/projects/m4/

[2] Pfau, T., Ellis, D.P.W. and Stolcke, A. "Multispeaker speech activity detection for the ICSI meeting recorder", *Proc. Automatic Speech Recognition and Understanding Workshop*, Italy, December 2001.

[3] LeBlanc, J. and de Leon, P. "Speech Separation by Kurtosis Maximization", *IEEE ICASSP 1998*, 1029-1032.

[4] Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction: Possible role of a repetitive structure in sounds", *Speech Communication* **27**:187-207, 1999.

[5] Krishnamachari, K., Yantorno, R., Benincasa, D. and Wenndt, S., "Spectral Autocorrelation Ratio as a Usability Measure of Speech Segments Under Co-channel Conditions", *IEEE International Symposium Intelligent. Sig. Process. and Comm. Sys.*, 2000.

[6] Yantorno, R.E., "*A study of the spectral autocorrelation peak valley ratio (SAPVR) as a method for identification of usable speech and detection of co-channel speech*", Speech Processing Lab Technical Report, Temple University, Philadelphia, 2000.

[7] Lewis, M.A. and Ramachandran, R.P., "Cochannel speaker count labelling based on the use of cepstral and pitch prediction derived features", *Pattern Recognition* **34**: 499-507, 2001.

[8] Banzhaf, W., Nordin, P., Keller, R., and Francone, F., *Genetic programming: an introduction.* Morgan Kaufmann, 1998.

[9] Scott, M., Niranjan, M. and Prager, R., "*Parcel: feature subset selection in variable cost domains*", Technical Report CUED/F-INFENG/TR. 323, Cambridge University Engineering Department, UK, 1998.

[10] Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A. and Wooters, C., "The ICSI Meeting Corpus", *IEEE ICASSP 2003*.