Speech and Crosstalk Detection in Multichannel Audio

Stuart N. Wrigley, *Member, IEEE*, Guy J. Brown, Vincent Wan, and Steve Renals, *Member, IEEE*

***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, speech from both the microphone's wearer (local speech) and from other participants (crosstalk) is received. 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 optimize 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 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.**

***Index Terms*—Crosstalk, Cochannel interference, meetings, feature extraction, hidden Markov models (HMM), speech recognition.**

I. Introduction

MORGAN *et al.* [1] have referred to processing spoken language in meetings as a nearly "automatic speech recognition-complete" problem. Most problems in spoken language processing can be investigated in the context of meetings. Meetings are characterized by multiple interacting participants, whose speech is conversational and overlapping. A number of laboratories have explored the recognition and understanding of meetings using audio and audio-visual recordings, in particular the International Computer Science Institute (ICSI; e.g., [2]) and Carnegie Mellon University's Interactive Systems Laboratories (e.g., [3]). Such recordings are typically made in an instrumented meeting room, equipped with sensors such as microphones (close-talking and distant), video cameras, and video projector capture. For instance, the meetings recorded at ICSI took place in a conference room with up to 12 participants seated around a long narrow table. Audio was acquired from head-mounted microphones (one per participant), desktop omnidirectional microphones, and two inexpensive microphones as might be found on a palmtop computer [2].

To automatically transcribe what was said in a meeting is a difficult task, since speech in meetings is typically informal and spontaneous, with phenomena such as backchannels, overlap and incomplete sentences being frequently observed. Shriberg *et al.* [4] have demonstrated that speakers overlap frequently in multiparty conversations such as meetings. In an analysis of the ICSI meetings corpus, they reported that 6–14% of words spoken were overlapped by another speaker (not including backchannels, such as "uh-huh"). In automatic speech recognition (ASR) experiments, they showed that the word error rate (WER) of overlapped segments was 9% absolute higher than for nonoverlapped segments in the case of headset microphones (with a WER increase of over 30% absolute for lapel microphones). Further, they were able to demonstrate that this increase in WER mainly occurred because crosstalk (nonlocal speech received by a local microphone) was recognized as local speech. The accurate identification of speaker activity and overlap is a useful feature in itself. For instance patterns of speaker interaction can provide valuable information about the structure of the meeting [5].

Since each participant in a meeting is recorded on a separate microphone, speech activity detection could be carried out using a simple energy threshold (e.g., [6]). However, this is impractical for a number of reasons. First, it is common for speech from a microphone's owner to be contaminated by speech from another participant sitting close by. Such crosstalk is a major problem when lapel microphones are used but it is still a significant problem with head-mounted microphones. Secondly, the participants in such meetings are usually untrained in the use of microphones and breath and contact noise are frequently observed. Finally, it is common for a channel to exhibit a significant drop in energy during a single speaker turn if that participant moves their head to address a neighbor, thus, altering the mouth-microphone coupling.

In this paper, we are concerned with developing a method for detecting speech and crosstalk in multiparty meetings. Specifically, we describe a classifier which labels segments of a signal as being either local or nonlocal speech, and will also determine whether the local speech has been contaminated by crosstalk. This task is more challenging than classical speech detection since it is necessary to determine whether one or more speakers are active concurrently in addition to detecting each incidence of speech activity.

Previous approaches to the detection of crosstalk in audio recordings have included the application of higher-order statistics, signal processing techniques, and statistical pattern

Manuscript received August 29, 2003; revised April 29, 2004. This work was supported as part of the M4 Multimodal Meeting Manager Project by the EU IST Programme under Project IST-2001-34485. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Walter Kellermann.

S. N. Wrigley, G. J. Brown, and V. Wan are with the Department of Computer Science, University of Sheffield, Sheffield S1 4DP, U.K. (e-mail: s.wrigley@dcs.shef.ac.uk; g.brown@dcs.shef.ac.uk; v.wan@dcs.shef.ac.uk).

S. Renals is with the Center for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9LW, U.K. (e-mail: s.renals@ed.ac.uk).

Digital Object Identifier 10.1109/TSA.2004.838531

recognition. LeBlanc and de Leon [7] addressed the problem of discriminating overlapped speech from nonoverlapped speech using the signal kurtosis. They demonstrated that the kurtosis of overlapped speech is generally less than the kurtosis of isolated utterances, since—in accordance with the central limit theorem—mixtures of speech signals will tend toward a Gaussian distribution. This statistical property has also been used to identify reliable frames for speaker identification in the presence of an interfering talker [8].

A variety of approaches based on the periodicity of speech have been proposed for the detection of crosstalk and the separation of multiple speakers (e.g., [9]). Morgan *et al.* [10] proposed a harmonic enhancement and suppression system for separating two speakers. The pitch estimate of the "stronger talker" is derived from the overlapping speech signal and the stronger talker's speech is recovered by enhancing its harmonic frequencies and formants. The weaker talker's speech is then obtained from the residual signal created when the harmonics and formants of the stronger talker are suppressed. However, this system fails when three or more speakers are active: it is only able to extract the stronger talker's speech.

Changes to the harmonic structure of a signal can also be used to detect crosstalk. Krishnamachari *et al.* [11] proposed that such changes be quantified by the ratio of peaks to valleys within the autocorrelation of the signal spectrum—the so-called spectral autocorrelation peak valley ratio (SAPVR). 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.

In statistical pattern recognition approaches, examples of clean and overlapping speech are used to train a classifier. For example, Zissman *et al.* [12] trained a Gaussian classifier using mel-frequency cepstral coefficients (MFCCs) to label a signal as being target-only, jammer-only, or two-speaker (target plus jammer). Although 80% correct detection was recorded, their system never encountered silence or more than two simultaneous speakers.

These approaches attempt to identify or separate regions of speech in which only two speakers are active simultaneously but are insufficient for meeting scenarios where a large number of participants are each recorded on an individual channel which, due to the microphone characteristics, can contain significant crosstalk. To deal with speech detection in this multichannel environment, Pfau *et al.* [13] proposed a speech/nonspeech detector using an ergodic hidden Markov model (eHMM). The eHMM consisted of two states—speech and nonspeech—and a number of intermediate states which enforced time constraints on transitions. Each state was trained using features such as critical band loudness values, energy, and zero-crossing rate. To process a meeting, the eHMM created a preliminary speech/nonspeech hypothesis for each channel. For regions in which more than one channel was hypothesised as active, the short-time cross-correlation was computed between all active channel pairs to assess their similarity. For each pair which exhibited high similarity (i.e., the same speaker was active in both channels), the channel with the lower energy was assumed to be crosstalk. Any remaining regions for which two or more channels were labeled as speech were presumed to correspond to overlapping speakers.

TABLE I
FOUR BROAD CATEGORISATIONS OF AUDIO USED IN THE PRESENT STUDY

Label	Description
S	Local channel ('speaker alone')
SC	Local channel speaker concurrent with one or more other speakers ('speaker plus crosstalk')
C	One or more non-local speakers ('crosstalk')
SIL	No speakers ('silence')

In contrast to previous approaches which exhibit channel-, speaker- , or environment-dependencies, we present a method that achieves a reliable classification regardless of the room in which the meeting is recorded, the identities of the individual speakers and the overall number of participants. This approach is based on the principles used by [13] but contains novel enhancements. The number of classification categories for each channel is increased from two (speech/nonspeech) to the four shown in Table I. These additional classes increase the flexibility of the system and more closely guide future analysis (such as enhancement of crosstalk-contaminated speech). Additionally, we have investigated a range of possible acoustic features for the eHMM (including cross-correlation) to determine which combination provides the optimum classification performance for each channel classification. We have evaluated our approach on the same data set. We also report ASR results using our multichannel speech activity detector as a preprocessing stage.

II. ACOUSTIC FEATURES

Some features were drawn from previous speech activity and crosstalk detection work. Additionally, we identified a number of other features which are suited to analyzing the differences between isolated and overlapping speech. Each feature was calculated over a 16 ms Hamming window with a frame-shift of 10 ms, unless otherwise stated.

A. MFCC, Energy, and Zero Crossing Rate

Similar to [13], MFCC features for 20 critical bands up to 8 kHz were extracted. MFCC vectors are used since they encode the spectral shape of the signal (a property which should change significantly between the four channel classifications in Table I). The short-time log energy and zero crossing rate (ZCR) were also computed.

B. Kurtosis

Kurtosis is the fourth-order moment of a signal divided by the square of its second-order moment. It has been shown that the kurtosis of overlapping speech is generally less than the kurtosis of isolated speech utterances [7]. Here, a 160 ms window, centerd 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 using a 16 ms window.

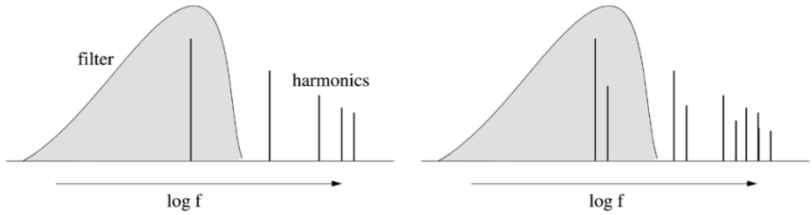


Fig. 1. Schematic illustration of the fundamentalness metric. (a) For single speaker speech, the analysing filter can isolate the fundamental component. The lack of modulation in the filter output gives rise to a high fundamentalness value. (b) For dual-speaker speech, harmonics from both speakers fall within the response area of the analysing filter. The resulting output of the filter is modulated, giving rise to a lower fundamentalness measure. After [14], Fig. 11 with permission from Elsevier.

C. Fundamentalness

Kawahara *et al.* [14] describe an approach to estimating the "fundamentalness" of an harmonic. Their technique is based on amplitude modulation (AM) and frequency modulation (FM) extracted from the output of a bandpass filter analysis.

When centered at different frequencies, the analysing filter 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 response area of the filter [usually just the fundamental component; see Fig. 1(a)]. Although this technique was developed to analyze isolated speech [see [14], p. 196, (13)–(19)], the concept that a single fundamental produces high fundamentalness is useful here. If more than one fundamental is present [see Fig. 1(b)], interference of the two components introduces modulation, thus decreasing the fundamentalness measure. Such an effect will arise when two or more speakers are active simultaneously, giving rise to overlapping harmonic series. Here, we compute the maximum value of the fundamentalness measure for center frequencies between 50 and 500 Hz.

D. Spectral Autocorrelation Peak-Valley Ratio

Spectral autocorrelation peak-valley ratio (SAPVR) [11] is computed from the autocorrelation of the signal spectrum obtained from a short-time Fourier transform. The measure is the ratio of peaks to valleys within the spectral autocorrelation. Specifically, the metric used here is based on SAPVR-5 [15].

E. Pitch Prediction Feature

The pitch prediction feature (PPF) was developed for the task of discriminating between single speaker speech and two speaker speech [16]. The first stage computes 12th-order linear prediction filter coefficients (LPCs) which are then used to calculate the LP residual (error signal). The residual is smoothed using a Gaussian-shaped filter after which an autocorrelation analysis identifies periodicities between 50 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 distance between successive peaks. If a frame contains a single speaker, a regular sequence of peaks will occur in the LP residual which correspond to glottal closures. Therefore, the standard deviation of the interpeak differences will be small. Conversely, if the frame contains two speakers of different fundamental frequency, glottal closures of both speakers will be evident in the residual and the standard deviation of the interpeak differences will be higher. In order to allow direct comparison between our approach and that of [16], a 30 ms window was used.

F. Features Derived From Genetic Programming

A genetic programming (GP) approach (see [17] for a review) was also used to identify frame-based features that could be useful for signal classification. The GP engine's function set included standard MATLAB functions such as `fft`, `min`, `max`, `abs`, `kurtosis`, and additional functions such as `autocorr` (time-domain autocorrelation) and `normalize` (which scaled a vector to have zero mean and unit variance). 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 IV. Successive generations were obtained using fitness-proportional selection. The GP engine identified several successful features, of which three were included in the following feature selection process:

GP1: `rms(zerocross(abs(diff(x))))`
GP2: `max(autocorr(normalize(x)))`
GP3: `min(log10(abs(diff(x))))`

where `diff` calculates differences between adjacent elements of `x` and `zerocross` returns 1 at the points at which the input either changes sign or is zero and returns 0 otherwise. Interestingly, GP discovered several features based on spectral autocorrelation (see Section II-D) but these were never ranked highly.

G. Cross-Channel Correlation

Other features were extracted using cross-channel correlation. For each channel i, the maximum of the cross-channel correlation $C_{ij}(t)$ at time t between channel i and each other channel j was computed

$$C_{ij}(t) = \max_{\tau} \left(\sum_{k=0}^{p-1} x_i(t-k) x_j(t-k-\tau) w(k) \right) \quad (1)$$

where τ is the correlation lag, x_i is the signal from channel i, x_j is the signal from channel j, P is the window size and w is a Hamming window. From this set of correlation values for channel i, the unnormalized and normalized minimum, maximum and mean values were extracted and used as individual features.

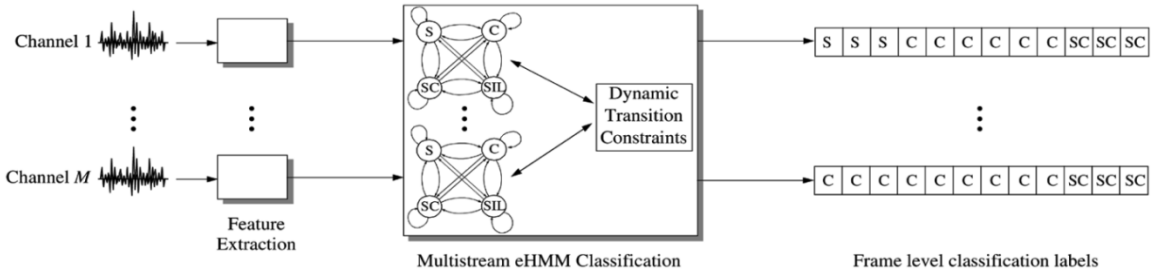


Fig. 2. Schematic of the classification process. For each channel of the meeting, the features identified in Section IV are extracted and input to that channel's ergodic HMM. Each channel is classified in parallel to allow dynamic transition constraints to be applied. The output of this process is a sequence of classification labels for each channel.

Two forms of normalization were used. In the first, the feature set for channel i was divided by the frame energy of channel i. The second was based on spherical normalization, in which the cross correlation C_{ij} is divided by the square-root of the autocorrelations for channels i and j plus some nonzero constant to prevent information loss. Spherical normalization converts the cross-channel correlation $C_{ij}(t)$ to a cosine metric based solely on the angle between the vectors $[x_i(t)\dots x_i(t-P+1)]^T$ and $[x_j(t-\tau)\dots x_j(t-\tau-P+1)]^{\mathrm{T}}$. On this scale the value of the normalized cross correlation of two identical signals would be one, while different signals would yield a value less than one. A full derivation may be found in [18].

III. Statistical Framework

The crosstalk classifier consists of a four state eHMM in which each state corresponds to one of the four categories given in Table I.

The probability density function $p(x)$ of each state is modeled by a Gaussian mixture model (GMM)

$$p(x) = \sum_{i=1}^{G} p_i \Phi_i(X, \mu_i, \Sigma_i) \tag{2}$$

where X is the multidimensional feature vector and G is the number of Gaussian densities Φ_i, each of which has a mean vector μ_i, covariance matrix Σ_i and mixing coefficient p_i. For simplicity, we assume a diagonal covariance matrix. For each state labeled S, C, SC and SIL the value of G was 20, 5, 20, and 4, respectively, determined by tests on a development set. Each GMM was trained using the expectation-maximization (EM) algorithm (e.g., [19]).

The likelihood of each state k having generated the data X_t at time frame t is combined with transition probabilities to determine the mostly likely state S_t

$$S_t = \underset{k}{\operatorname{argmax}} P(X_t \mid S_k) P(S_k \mid S_{t-1}). \tag{3}$$

The transition probabilities were computed directly from labels in the training set.

During classification of multichannel meeting data, each channel is classified by a different eHMM in parallel (see Fig. 2). This allows a set of transition constraints to be dynamically applied, such that only legal combinations of channel classifications are possible. For example, it is illegal for more than one channel to be classified as S (speaker alone): if more than one speaker is active the correct classification would be SC (speaker plus crosstalk) for channels containing active speakers. Such constraints are applied in two stages. The first stage determines the likelihood of each cross-channel classification combination from the legal combinations. In other words, we define an eHMM state space in which the observations correspond to the per-channel eHMM states. When considering m observations (audio channels), the state space contains all permutations of:

- $\mathrm{S}(m-1)\mathrm{C}$;
- $q\mathrm{SC}n\mathrm{C}$;
- $m\mathrm{SIL}$.

where $2 \leq q \leq m$ and $n = m - q$. For example, a legal combination for a four channel meeting could be "S C C C."

The second stage reduces the size of the state space. If at least one channel is classified as nonsilence by the initial GMM classifier, it is assumed that none of the other channels can be silent because crosstalk will occur. Furthermore, it was observed empirically on a validation set that speaker-alone GMM classification had a significantly higher accuracy than the other three categories. Hence, if any GMM-based frame classification was speaker-alone, the eHMM state space was limited to those states including a speaker-alone label.

When considering the discrimination results of a classifier over two classes, it is unlikely that a perfect separation between the two groups will occur, hence a decision boundary is necessary. A receiver operating characteristic (ROC) curve shows the discriminatory power of a classifier for a range of decision boundary values. Each point on the ROC represents a different decision boundary value. We base our feature selection approach on the area under the ROC curve (AUROC) for a particular classifier. Rather than consider all possible feature subsets, we use the sequential forward selection (SFS) algorithm (e.g., [20]). This approach computes the AUROC for GMM classifiers trained on each individual feature. The feature with the highest AUROC is retained and GMMs are retrained using all two-feature sets which include the winning feature. Again, the feature set resulting in the highest AUROC is selected. This process continues until the gain in the AUROC is less than a threshold value (1% in our case) at which point the algorithm terminates and the current feature set is selected. In our experiments, the SFS algorithm always terminated with fewer than six features for all crosstalk categories.

IV. Corpus

Experiments were conducted using data from the ICSI meeting corpus [2]. The training data consisted of one million frames per crosstalk category of conversational speech extracted at random from four ICSI meetings (*bro012, bmr006,*

TABLE II
INDIVIDUAL FEATURE PERFORMANCE FOR EACH CLASSIFICATION CATEGORY. VALUES INDICATE THE PERCENTAGE OF TRUE POSITIVES AT EQUAL ERROR RATES, WITH THE BEST PERFORMING FEATURE FOR EACH CLASSIFICATION CATEGORY HIGHLIGHTED. GPn DENOTES THE THREE GENETIC PROGRAMMING FEATURES DESCRIBED IN SECTION II.F. XC DENOTES CROSS-CORRELATION AND S-NORM REFERS TO SPHERICAL NORMALISATION AS DESCRIBED IN SECTION II-G

Feature	S	C	SC	SIL
MFCC	58.98	56.98	54.58	59.86
Energy	72.16	64.56	70.17	71.25
ZCR	55.53	52.18	50.97	54.61
Kurtosis	68.05	66.59	67.50	71.14
Freq. Kurtosis	53.59	53.56	58.46	53.45
Fundamentalness	63.71	63.31	60.43	58.21
SAPVR	52.21	52.12	46.02	51.28
PPF	63.55	60.63	58.39	57.81
GP1	59.58	58.60	62.49	62.19
GP2	72.80	64.94	64.42	56.19
GP3	53.47	50.67	41.43	51.68
Max XC	52.62	55.32	70.84	75.25
Min XC	62.19	57.22	71.05	68.05
Mean XC	55.84	57.33	**72.19**	**75.54**
Max Norm XC	**78.11**	75.07	50.83	56.07
Min Norm XC	60.14	54.08	53.30	51.12
Mean Norm XC	77.61	**75.19**	50.49	55.21
Max S-Norm XC	55.06	49.95	64.41	67.19
Min S-Norm XC	48.15	51.27	59.35	71.46
Mean S-Norm XC	56.30	53.25	66.28	70.27

bed008, bed010). For each channel, a label file specifying the four different crosstalk categories (see Table I) was automatically created from the existing ASR word-level transcriptions. For the feature selection experiments, the test data consisted of 15 000 frames per crosstalk category extracted at random from one ICSI meeting (*bmr001*).

Note that frames were labeled as crosstalk (C) or speaker plus crosstalk (SC) on the basis of comparisons between word-level alignments generated by ASR for each channel. In practice, the audibility of the crosstalk was sometimes so low that, upon listening, the frames appeared to be silent.

V. FEATURE SELECTION EXPERIMENTS

Before describing the selected feature sets, it is insightful to examine the performance of the individual features on each crosstalk classification category. Table II shows the true positive rate for each GMM feature-category classifier. These performance values are taken from the ROC operating point at which the false negative rate (1-true positive rate) and false positive rate are equal.

It is interesting to note that although some features have a high accuracy for one or more classification categories (e.g., maximum normalized cross-correlation), some features perform relatively poorly on all categories (e.g., SAPVR). The poor performance of the zero crossing rate (ZCR) feature is most likely explained by the varying degree of background noise in each channel. Also, a number of meeting participants were inexperienced in the use of head-mounted microphones and frequently generated breath noise. Such breath noise causes high ZCR values irrespective of whether the microphone wearer is speaking or not.

Surprisingly, two features previously described in the literature and expected to perform well—PPF and SAPVR—both gave mediocre results. However, we note that Lewis and Ramachandran [16] only evaluated the PPF on synthetic mixtures of utterances drawn from the TIMIT database. Our results suggest that the PPF is not robust for real acoustic mixtures which contain a substantial noise floor, such as the recordings used here. Similarly, Krishnamachari *et al.* report good performance for the SAPVR when evaluated on mixtures of TIMIT sentences in which no background noise was present [11] but its performance on our noisy data is poor. Note, though, that the SAPVR was developed as a measure for determining which portions of a target utterance, when mixed with corrupting speech, remain usable for tasks such as speaker identification. In other words, this measure determines when a target speaker is dominating a segment of speaker plus crosstalk (SC). This is a different task to the one presented here, in which the goal is to distinguish between single speaker speech and multiple speaker speech.

Note that the equal error rates presented in Table II cannot be used to estimate the performance of various feature combinations directly due to the nature of the selection process. As described in Section III, a feature is added to the currently selected feature set only if it increases the AUROC by more than 1%. Therefore, this measure relies on the shape of the ROC curve (i.e., the performance at all operating points) rather than the performance at the equal error rate point.

A. Feature Selection Using Full Feature Set

The feature sets derived by the SFS algorithm were as follows:

- local channel speaker alone (S): kurtosis, and maximum normalized cross-channel correlation;
- local channel speaker concurrent with one or more speakers (SC): energy, kurtosis, maximum normalized cross-channel correlation, and mean spherically normalized cross-channel correlation;
- one or more nonlocal speakers (C): energy, kurtosis, mean cross-channel correlation, mean normalized cross-channel correlation, maximum spherically normalized cross-channel correlation;
- silence (SIL): energy and mean cross-channel correlation.

It is interesting to note that some features used in previous studies (such as MFCCs, SAPVR, and PPF) did not perform well enough to be included in any of the optimal feature sets.

The GMM classification performance for each feature set is shown in Fig. 3. For equal false positive and false negative rates, the performance of each classifier is approximately 80%.

B. Feature Selection Excluding Energy Feature

In the second set of experiments, we assumed that the channel energy is unreliable (as it may be for corpora using lapel microphones) and removed it from the set of potential features available to the feature selection process. Using this reduced set, the features derived by the SFS algorithm were the following:

- local channel speaker alone (S): kurtosis and maximum normalized cross-channel correlation;

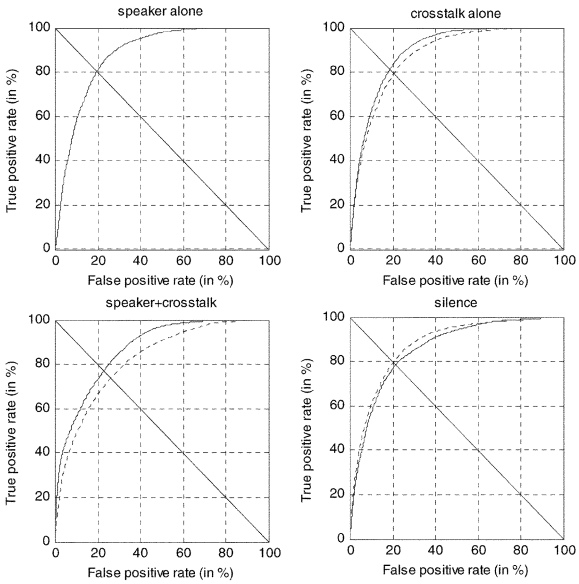


Fig. 3. ROC performance curves for each crosstalk category's optimum feature set. Diagonal lines indicate equal error rates. Dashed curves indicate performance when log energy is excluded from the set of potential features.

- local channel speaker concurrent with one or more speakers (SC): kurtosis, fundamentalness, maximum normalized cross-channel correlation and mean spherically normalized cross-channel correlation;
- one or more nonlocal speakers (C): mean cross-channel correlation and mean spherically normalized cross-channel correlation;
- silence (SIL): kurtosis, mean cross-channel correlation and mean spherically normalized cross-channel correlation.

The GMM classification performance for each feature set is shown in Fig. 3. The removal of log energy has little effect on the ROC curves, and overall classification performance of the system remains at approximately 80%. This is most likely due to the high performance of the cross-correlation features which dominate the ROC curves. It is also interesting to note that the fundamentalness feature, which was developed for a different task but was expected to discriminate well between single speaker and multiple speaker speech, also contributes to the feature set for speaker plus crosstalk.

VI. Multistream eHMM Classification Experiments

The previous section identified the subset of features which were best suited to classifying isolated frames of audio data. Here, we investigate whether the eHMM framework shown in Fig. 2 can improve performance by exploiting contextual constraints. Each channel classification is represented by a state within the eHMM which, in turn, is modeled by a GMM of the form used in the feature selection experiments. Contextual constraints are embodied in the transition probabilities between states, which were estimated from the training data.

To ensure that likelihoods generated by each state of the eHMM were in the same range, each state employed the union of the four winning feature sets described in Section V-A. The test data consisted of all the transcribed channels from 27 ICSI meetings.[1] The eHMM classification performances are shown in Fig. 4.

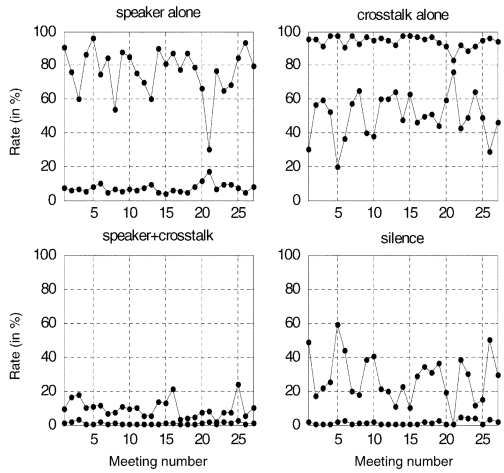


Fig. 4. True positive rate (upper-line) and false positive rate (lower-line) per meeting for each channel classification.

These results show that the speaker alone and crosstalk alone channel classifications exhibit a high true positive rate across all meetings. A number of meetings exceed the 90% true positive rate, with mean true positive rate for speaker alone being 76.5%, and 94.1% for crosstalk alone. Furthermore, the average false positive rate for the speaker alone class is only 7% but 50.2% for crosstalk alone. The true positive rate for the remaining classes (speaker plus crosstalk and silence) are significantly lower but do exhibit a consistently small false positive rate. Upon examining the confusion matrix (a grid showing which and how many classes have been misclassified), it was discovered that many of the silence frames were misclassified as crosstalk alone, thus explaining the low true positive rate for silence and relatively high false positive rate for crosstalk alone.

Fluctuations in classifier performance can also be seen in Fig. 4. Transcription notes from the ICSI corpus indicate that some channels of the meetings on which we achieved lower performance suffer from poor recording. For example, test meeting *bmr014* (meeting number 13 in Fig. 4) suffered from "spikes" and low gains on some channels which we believe caused single speaker true positives to fall to 60% (significantly lower than the average of 76.5%). Meeting 21 (*bro008*) exhibits poor classification performance for all four channel classifications due to unusually low channel gains during recording. Other recording issues ranged from fluctuating channel gains to corrupted audio buffers which also affected subsequent channel synchronization.

As stated in the introduction, two applications for such a classification system are speech recognition preprocessing and speaker turn analysis. Both of these rely on accurate detection of local speaker activity, which is largely equivalent to the speaker alone (S) channel classification since class SC occurs relatively infrequently (accounting for 2.4% of the ICSI data). As described above, speaker alone classification at the frame level can

[1]The 27 test meetings were 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*.

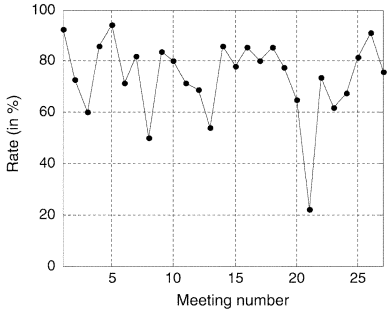


Fig. 5. Speaker alone segment recognition rate per meeting.

be as high as 96%. However, these applications require the accurate classification of contiguous segments of audio, rather than individual frames. To this end, we have also assessed the ability of the classifier to detect segments of speaker-alone activity.

We define a segment to be a contiguous region in which all frames have the same channel classification. A segment is deemed to have been correctly identified if a certain proportion of classified frames within the segment boundaries agree with the manual transcription. Segment-based classification for the speaker alone class is shown in Fig. 5.

Here it was assumed that a segment was correctly classified if more than 50% of the constituent frames were correctly classified. The segment-level performance is similar to that of the frame-level approach, with a mean recognition rate of 74% and recognition rates approaching 94% for some meetings.

VII. Evaluation Using ASR

An evaluation of ASR performance using the segments described above was conducted on a number of ICSI meetings (*bmr001*, *bro018*, and *bmr018*) totalling 2.5 hours of multichannel speech. On these meetings, the eHMM classifier has a segment recognition accuracy of between 83% and 92% for single speaker detection. The ASR system is the publicly available version of HTK [21] trained on 40 hours of the ICSI meetings data. On unseen test data, this recognizer has a word accuracy of approximately 50% without speaker adaptation. To evaluate the eHMM classifier we compare results of ASR on the ground truth segments versus ASR on the eHMM segments.

It is also interesting to compare ASR performance using the eHMM segments against those produced by a classical voice activity detector (VAD) such as [6]. For the purposes of this evaluation, we do not wish to make the distinction between voiced and unvoiced speech so only the first stage of the VAD algorithm is used, which distinguishes between silence and nonsilence. The average energy is measured (from the training set described above) for each of the two voice activity classes and is used to determine the appropriate classification for each test frame based on a normalized Euclidean distance.

Table III shows the ASR results on the various segment types. The word accuracy achieved using eHMM segments is close to that obtained using the ground truth segments. In *bmr001* and *bro018* there is only a small drop in eHMM ASR word accuracies compared to the ground truth word accuracies (relative factors of 98.50% and 99.29%, respectively) despite lower segment accuracies of 92% and 89%. *Bmr018* has a relative factor that is close to the segment accuracy. The results indicate that

TABLE III
Segment and ASR Accuracies (%) on Whole Meetings. Results in Brackets Are as a Percentage of the Baseline Ground Truth Segments

	Seg	Seg	ASR	ASR	ASR
Meeting	eHMM	VAD	Ground truth	eHMM	VAD
bmr001	~92%	~30%	44.44 (100)	43.77 (98.50)	27.91 (62.80)
bro018	~89%	~77%	61.75 (100)	61.31 (99.29)	51.84 (83.95)
bmr018	~83%	~61%	63.97 (100)	53.90 (84.20)	49.73 (77.73)

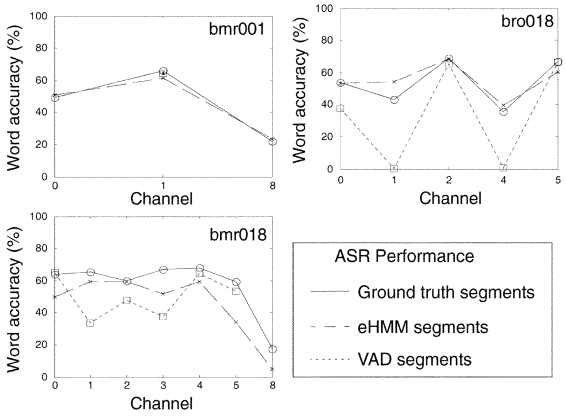


Fig. 6. ASR performance for meetings *bmr001, bro018*, and *bmr018*. Note that the VAD classifier failed on a number of channels and hence, some data points (channels 0 and 8 from *bmr001* and channel 8 from *bmr018*) are missing.

the eHMM classifier is capable of detecting most of the frames required for optimal ASR. In comparison, the word accuracy on VAD segments is much lower due to significantly lower segment accuracy.

Fig. 6 shows the ASR results obtained using the three different types of segment by channel. The inconsistent VAD ASR results emphasise that an energy based measure for speaker detection is highly unreliable: some channels can be so noisy that the VAD classifier labels all frames as speech activity. For example *bmr001* was particularly problematic, since the whole of channels 0 and 8 were labeled as speech activity.

VIII. General Discussion

Two experiments are described in this paper, both relating to the broad classification of audio data from meeting recordings. Our goals were to produce accurate labels corresponding to the number of speakers active at a particular time and to indicate if the local speaker is active. The first experiment identified the optimal feature set for each channel classification, each of which achieved approximately 80% frame accuracy when considering equal true-positive and false-positive error rates. Several cross-channel correlation measures were selected in addition to conventional features such as short-term energy and kurtosis. Additionally, we found that features which were originally designed for a different purpose can also play a role in crosstalk analysis (e.g., "fundamentalness" [14]). Furthermore, features which have previously been used to identify overlapping speakers such as MFCCs, PPF [16], and SAPVR (e.g., [8]) were rejected.

In the second set of experiments, the optimal feature set was used to train a number of eHMM classifiers (one per meeting channel) which operated in parallel. This allowed transition constraints to be dynamically applied depending on the previous state and the unconstrained GMM classifications. This approach improved performance for some classes, notably speaker alone (S). Indeed, automatic speech recognition results using the automatically generated speaker alone segments indicate performance equal to that obtained using the ground truth segments.

To conclude, a multichannel activity classification system has been described which can distinguish between the four activity categories shown in Table I. Furthermore, the segmentation of speaker alone activity has been shown to be particularly reliable for speech recognition applications: ASR performances using the eHMM segments and the transcribed ground truth segments are extremely similar.

REFERENCES

[1] N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "Meetings about meetings: Research at ICSI on speech in multiparty conversations," in *Proc. ICASSP*, 2003, pp. 740–743.

[2] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. ICASSP*, 2003, pp. 364–367.

[3] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech style," in *Proc. ICSLP*, 2002, pp. 301–304.

[4] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proc. EUROSPEECH*, vol. 2, 2001, pp. 1359–1362.

[5] S. Renals and D. Ellis, "Audio information access from meeting rooms," in *Proc. ICASSP*, 2003, pp. 744–747.

[6] L. R. Rabiner and M. R. Sambur, "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem," *IEEE Trans. Acoust. Speech.*, vol. 25, no. 4, pp. 338–343, Aug. 1977.

[7] J. LeBlanc and P. de Leon, "Speech separation by kurtosis maximization," in *Proc. IEEE ICASSP*, 1998, pp. 1029–1032.

[8] K. R. Krishnamachari, R. E. Yantorno, J. M. Lovekin, D. S. Benincasa, and S. J. Wenndt, "Use of local kurtosis measure for spotting usable speech segments in co-channel speech," in *Proc. IEEE ICASSP*, 2001, pp. 649–652.

[9] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Am.*, vol. 60, pp. 911–918, 1976.

[10] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 407–424, Sep. 1997.

[11] K. Krishnamachari, R. Yantorno, D. Benincasa, and S. Wenndt, "Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions," in *Proc. Int. Symp. Intelligent Signal Process Communication Systems*, 2000, pp. 710–713.

[12] M. A. Zissman, C. J. Weinstein, and L. D. Braida, "Automatic talker activity labeling for co-channel talker interference suppression," in *Proc. IEEE ICASSP*, 1990, pp. 813–816.

[13] T. Pfau, D. P. W. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the ICSI meeting recorder," in *Proc. IEEE ASRU Workshop*, 2001, pp. 107–110.

[14] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "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 Commun.*, vol. 27, pp. 187–207, 1999.

[15] R. E. Yantorno, "A Study of the Spectral Autocorrelation Peak Valley Ratio (SAPVR) as a Method for Identification of Usable Speech and Detection of Co-Channel Speech," Temple Univ., Philadelphia, PA, Speech Processing Lab. Tech. Report, 2000.

[16] M. A. Lewis and R. P. Ramachandran, "Cochannel speaker count labeling based on the use of cepstral and pitch prediction derived features," *Pattern Recognit.*, vol. 34, pp. 499–507, 2001.

[17] W. Banzhaf, P. Nordin, R. Keller, and F. Francone, *Genetic Programming: An Introduction*. San Mateo, CA: Morgan Kaufmann, 1998.

[18] V. Wan, "Speaker verification using support vector machines," Doctoral thesis, Dept. Comput. Sci., Univ. Sheffield, Sheffield, UK, 2003.

[19] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford, UK: Clarendon Press, 1995.

[20] F. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection," in *Pattern Recognition in Practice IV, Multiple Paradigms, Comparative Studies and Hybrid Systems*: Elsevier, 1994.

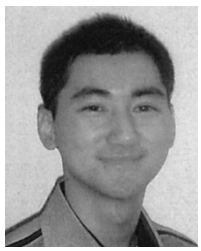
[21] S. Young *et al.*, The HTK Book (for HTK Version 3.2.1), [Online]. Available: http://htk.eng.cam.ac.uk/, Cambridge, U.K., 2002.

Stuart N. Wrigley (M'03) received the B.Sc. and Ph.D. degrees in computer science from the University of Sheffield, Sheffield, U.K., in 1998 and 2002, respectively.

Since 2002, he has worked in the Department of Computer Science, University of Sheffield, as a Research Associate on the EU Multimodal Meeting Manager (M4) Project. He has research interests in auditory selective attention, short term memory, feature binding, binaural hearing, and multisource signal analysis.

Guy J. Brown received the B.Sc. degree in applied science from Sheffield Hallam University, Sheffield, U.K., in 1988, the Ph.D. degree in computer science from the University of Sheffield in 1992, and the M.Ed. degree from the University of Sheffield, in 1997.

He is currently a Senior Lecturer in Computer Science at the University of Sheffield. He has a long-standing interest in computational models of auditory perception, and also has research interests in robust automatic speech recognition, spatial hearing, and music technology.



Vincent Wan received the B.A. degree in physics from the University of Oxford, Oxford, U.K., in 1997, and the Ph.D. degree in speaker verification and support vector machines from the University of Sheffield, U.K., in 2003.

In 1998 and 1999, he worked on hybrid speech recognition at the University of Sheffield, and then spent the year 2000 working at the Motorola Human Interface Labs at Palo Alto, CA, on speech and handwriting recognition. He presently holds a postdoctoral position at the Department of Computer Science, University of Sheffield. His interests include machine learning, biometrics, and speech processing.

Steve Renals (M'91) received the B.Sc. degree in chemistry from the University of Sheffield, Sheffield, U.K., the M.Sc. degree in artificial intelligence, and the Ph.D. degree in speech recognition and neural networks from the University of Edinburgh, Edinburgh, U.K.

He held postdoctoral fellowships at the International Computer Science Institute, Berkeley and the University of Cambridge. For nine years, he was a Lecturer, then Reader, in computer science at the University of Sheffield. In 2003, he was appointed Professor of speech technology in the School of Informatics at the University of Edinburgh, where he is also director of the Center for Speech Technology Research. He has published about 100 papers in the area of spoken language processing.