ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

FEATURE FUSION FOR MUSIC DETECTION

Eluned S. Parris, Michael J. Carey and Harvey Lloyd-Thomas

Ensigma Ltd., Turing House, Station Road, Chepstow, Monmouthshire, U.K.
{michael harvey}@ensigma.com

10.21437/Eurospeech.1999-485

ABSTRACT

Automatic discrimination between music, speech and noise has grown in importance as a research topic over recent years. The need to classify audio into categories such as music or speech is an important part of the multimedia document retrieval problem. This paper extends work previously carried out by the authors which compared performance of static and transitional features based on cepstra, amplitude, zero-crossings and pitch for music and speech discrimination. Two approaches are described to combine the features to improve overall performance. The first approach uses separate GMM classifiers for each feature type and fuses the outputs of the classifiers. The second approach combines different features into a single vector prior to modelling the data with a GMM. Significant improvements in performance have been observed using both approaches over the results achieved by a single type of feature. An equal error rate of 0.3% is achieved for the best system on ten second tests using seventeen hours of test material. The performance is maintained as the length of test file is reduced with an equal error rate of less than 1% being achieved with only two seconds of data.

1. INTRODUCTION

Automatic discrimination between music, speech and noise and combinations of these three events has grown in importance as a research topic over recent years. The demand to carry out multimedia document retrieval is increasing, particularly with the growth of the world wide web. The need to classify audio into categories such as music or speech is an important part of this problem. Several approaches to audio classification have been described in recent literature [1,2,3]. Each of these uses different features and pattern classification techniques and describes results on different material. In a previous paper [4] the authors examined the discrimination achieved by several different features using common training and test sets and the same classifier. The database assembled for these tests included speech from thirteen languages and music from all over the world. In each case the distributions in the feature space were modelled by a Gaussian mixture model (GMM).

Experiments were carried out on four types of feature, amplitude, cepstra, pitch and zero-crossings. The pitch and cepstral coefficients encompass the fine and broad spectral features respectively. The zero-crossing parameters and the amplitude were believed worthy of investigation as a computationally inexpensive alternative to the other features. In each case the derivative of the feature was also used and found to improve performance. The best performance resulted from using the cepstra and delta cepstra which gave an equal error rate (EER) of 1.2%. This was closely followed by normalised amplitude and delta amplitude. This however used a much less complex model. The pitch and delta pitch gave an EER of 4% which was better than the zero-crossing which produced an EER of 6%.

In the previous paper [4] we reported that the zero-crossing rates of the signals were not a good discriminator between speech and music. Therefore, the zero-crossing feature was omitted from further experiments. This paper describes experiments that have been carried out to exploit the independence between the cepstra, amplitude and pitch features to improve overall performance. The first approach uses separate GMM classifiers for each feature type and fuses the outputs of the classifiers. The second approach combines different features into a single vector prior to modelling the data with a GMM. Results for both of these approaches are given in this paper along with plots comparing performance for different lengths of test data.

2. FEATURE ESTIMATION

2.1 Cepstral Coefficients

The cepstral analysis used in the experiments was as follows. The data was sampled at 8kHz and was then filtered using a filterbank containing nineteen filters. The filterbank had a mel scale characteristic. The log power outputs of the filterbank were transformed into twelve cepstral coefficients and twelve delta cepstral coefficients at a frame rate of 10ms. The delta cepstra were calculated by estimating the trend of the cepstra over five successive frames. Cepstral mean subtraction was applied to each of the test files to ensure that the classifier did not use channel information to distinguish between the three types of signal.

2.2 Amplitude Features

These coefficients were the filterbank energy and delta energy. The features were normalised over a test file so that the absolute amplitude of the material did not effect the results by allowing the classifier to use level information to distinguish between the three types of signal. The delta amplitude was calculated by estimating the trend of the amplitude over five successive frames.

2.3 Pitch Features

The pitch estimation algorithm was similar to that used for IMBE speech coding [5]. This has been found to be an effective technique for pitch estimation in our previous work on gender and speaker identification [6,7].

This technique calculates an initial pitch estimate by correlating the 1kHz low pass filtered signal with delayed versions of the same signal. The correlation peaks occur at multiples of the pitch period. This initial estimate is smoothed using backward and forward pitch tracking to restrict inter-frame variations. The algorithm was modified to provide an estimate every 10 ms, the frame rate of the acoustic analysis. The smoothed pitch estimate is refined to produce a final pitch estimate accurate to 0.25 of a sample period. The pitch refinement algorithm uses a frequency domain matching technique to optimise a windowed periodic pulse train to the input speech, the pitch period corresponding to the inter-pulse interval. The high resolution results from the spectral match at the high frequency harmonics.

Four pitch parameters were calculated for each 10ms frame. These parameters were the pitch and delta-pitch estimate, the mean and variance of the pitch estimate over a 1s sliding window.

3. EXPERIMENTAL CONFIGURATION

3.1 Database

The experiments described in the following section were carried out using a database of music, speech and noise. All of the speech material was conversational and included examples from both genders. The following languages were represented, American English, Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. There were 2370 10s training files for speech and 2107 10s test files giving about six hours of speech in each case.

The music was predominately a diverse selection of Western music including classical, popular and jazz. However examples of music from Eastern Asia, the Arab world, Africa, South America and the Indian sub-continent were also included. There were 1529 10s training files and 1388 10s test files for music giving about four hours of material in each case. Both the speech and music signals were band-limited to 4kHz and sampled at an 8kHz rate.

The noise data was taken from the same material as the speech data. The regions of data between speech utterances were concatenated to produce 10s files. This resulted in 2650 10s training files and 2367 10s test files giving about seven hours of material in each case.

3.2 Experimental System

Pattern classification was carried out using Gaussian Mixture Models (GMM) [8]. Each of the possible classes of signal, music, speech or noise, were represented by a GMM trained on the training set using the expectation maximisation algorithm. The variances of the distributions were modelled by a diagonal covariance matrix. The score for a test file S_D for a given feature set was computed as the difference in log likelihood ratio between the music score and the best of speech and noise scores,

$$S_D = L_m - \max\left(L_s, L_n\right)$$

where L_m, L_s and L_n are the likelihood scores for music, speech and noise. This was found to be the most successful way of combining the individual music, speech and noise scores into a single score.

Linear weighting was used to combine the scores from the detectors using different front end features to try and improve system performance. The scores from two detectors were combined into a single score S as follows:

$$S = \alpha S_{D1} + \left(1-\alpha\right) S_{D2} \qquad (1)$$

where S_{D1} and S_{D2} are the scores from the two detectors and α is a weighting factor.

The combined scores were then used to generate Receiver Operating Characteristics and plotted as a Detection Error Trade-off curve[9]. The test files for music were used as the target data and the test files for speech and noise were used as the impostor data.

4. EXPERIMENTS

4.1 Linear Combination of Cepstral and Amplitude Features

Initially, the test files were processed using separate detectors for the cepstral and amplitude features. The distributions of the cepstra and delta-cepstra were modelled using sixty-four mixture GMMs and the amplitude and delta-amplitude features were modelled using four mixture GMMs. The number of distributions used in the GMMs were the optimal found in the experiments reported in the previous paper [4]. The scores from the two detectors on each test file were combined using (1). The best results were achieved using equal weighting, i.e. α set to 0.5, and are shown in Figure 1. The combined results are better than either of the individual detectors and achieve an equal error rate of 0.7%.

4.2 Linear Combination of Cepstral and Pitch Features

The test files were processed using separate detectors for the cepstral and pitch features. The distributions of the cepstra and delta-cepstra were modelled using sixty-four mixture GMMs and the four pitch features were modelled using sixteen mixture GMMs. The scores from the two detectors on each test file were combined using optimal weighting The best results were achieved using a weighting of α set to 0.23 for the pitch features. The results are shown in Figure 2. The combined results are again better than either of the individual detectors and achieve an equal error rate of 0.5%.

4.3 Linear Combination of Cepstral, Amplitude and Pitch Features

The best results achieved by combining non-cepstral features into a single feature vector prior to GMM modelling were obtained using six features: amplitude, delta-amplitude and the four pitch features. The scores from this detector were combined with the scores from the twenty-four cepstral feature detector for each test file. The best results were achieved using α set to 0.67 for the cepstral features and are shown in

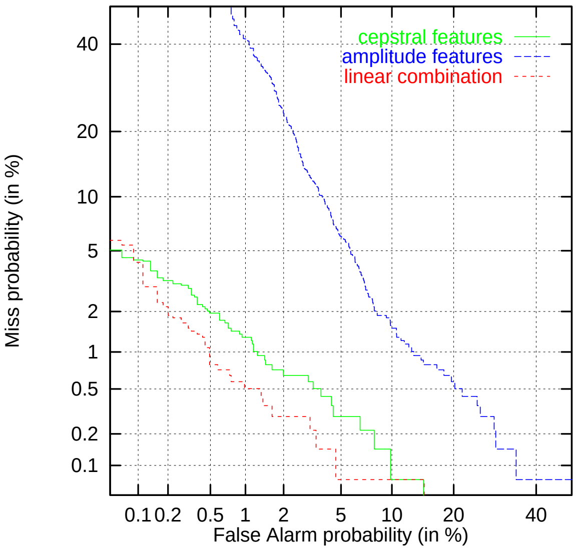

Figure 1: DET Plot for Linear Combination of Cepstral and Amplitude Features

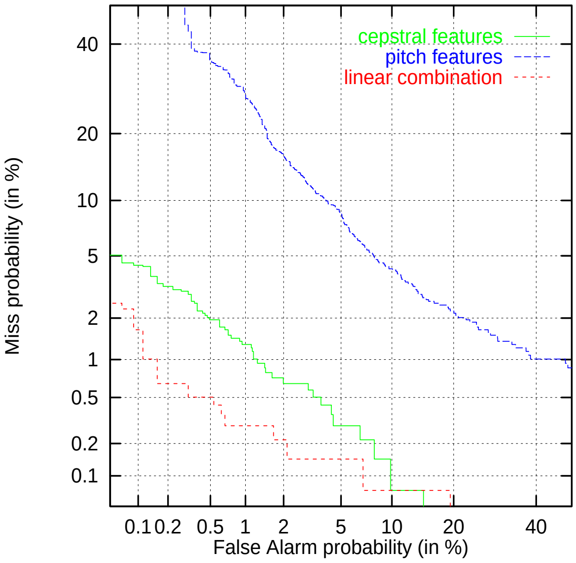

Figure 2: DET Plot for Linear Combination of Cepstral and Pitch Features

Figure 3. An equal error rate of 0.3% was achieved, representing errors for sixteen out of the 5862 test files in the database.

4.4 Combined Feature GMMs

The aim of the work described in this paper was to improve the performance of a music detector by combining different features. The previous three sections describe the results achieved by taking linear combinations of detectors using separate feature sets. An alternative approach is to combine different features into a single vector prior to GMM modelling. This results in one GMM for each of music, speech and noise and no need for fusion on the output of the detector. This is a simpler approach and also requires less computational resource during testing.

The cepstral and amplitude features are an obvious choice for combining into a single feature vector since the vector of both

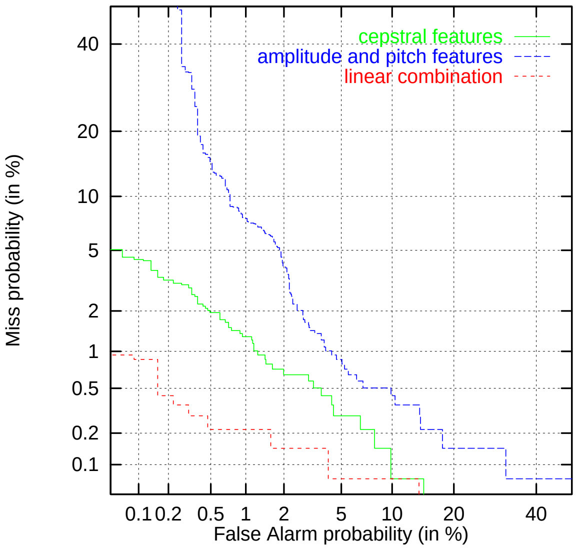

Figure 3: DET Plot for Linear Combination of Cepstral, Amplitude and Pitch Features

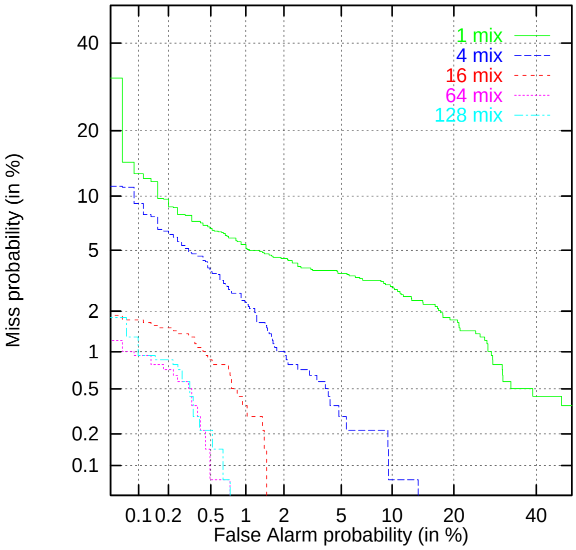

Figure 4: DET Plot for Combined Cepstral and Amplitude Feature GMMs

features describes the spectral envelope of the signal. The cepstral front end processor also automatically generates both sets of features. GMMs were trained for music, speech and noise for the combined cepstral and amplitude features. One, four, sixteen, sixty-four and one hundred and twenty-eight mixture models were built and the results are shown in Figure 4. An equal error rate of about 0.4% has been achieved using the combined feature vector for both the sixty-four and one hundred and twenty-eight mixture GMMs. This is in comparison to an equal error rate of 0.7% for the best linear combination presented in Figure 1.

The test files for which detection errors occurred were examined. Most of the music files rejected by the system corresponded to fast machine generated music. The speech files labelled as music either contained interference or music playing in the background. Similarly, most of the noise files labelled as music contained a television or radio in the

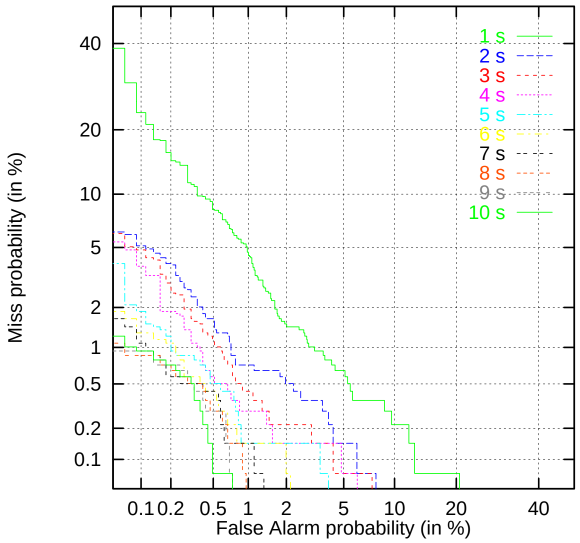

Figure 5: DET Plot Using Combined Cepstral and Amplitude Feature GMMs for Different Quantities of Test Data

background. The scores for these files were near to zero indicating a marginal decision between music, speech and noise.

4.5 Comparison of Music Detector Performance for Different Quantities of Test Data

The performance of the best combined GMM and linearly combined detectors was compared for different quantities of test data. The best combined GMM was achieved using the twenty-four cepstral features, amplitude and delta-amplitude in a single feature vector. Figure 5 shows the performance of this music detector as the quantity of test data is reduced from ten to one second. An equal error rate of less than 1% is achieved with at least two seconds of test data and an equal error rate of less than 2% with one second of data.

The best linear combination of detectors was given by combining the output of the cepstral detector with the pitch and amplitude detector. Figure 6 shows the performance of this combination of detectors as the quantity of test data is reduced from ten to one second. An equal error rate of less than 1% is achieved for this system with at least five seconds of test data. Although the detection rate of the linearly combined system is good, the performance deteriorates more rapidly than the combined GMM system as the amount of test data is reduced.

5. DISCUSSION

This paper has examined the benefits accruing from fusing the outputs of several music detectors. Combining these detectors has resulted in a substantial improvement in performance. However combining the amplitude and cepstral features into a single input vector produced a performance equal to that given by a linear combination of amplitude, pitch and cepstral scores.

The omission of the pitch feature simplifies the system and reduces the computation required for the feature extraction stage by more than a half. Since the amplitude and cepstral features are often also required for later stages of processing of

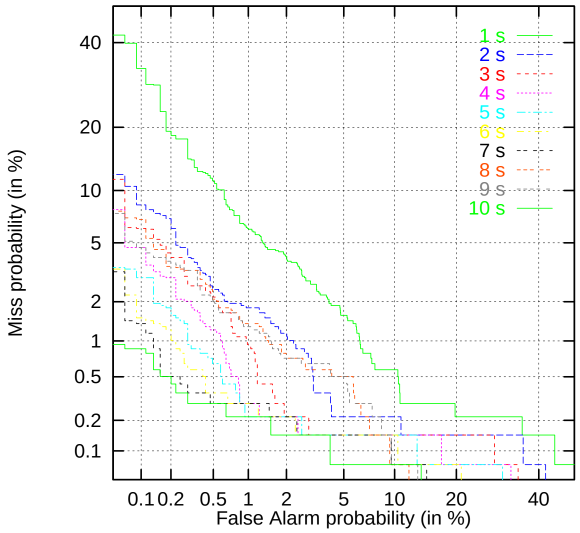

Figure 6: DET Plot Using Linear Combination of Cepstral, Amplitude and Pitch Features for Different Quantities of Test Data

the signal the use of these features for music detection results in little extra overhead. Also the pattern matching stage of the algorithm runs thirty times faster than real-time on a 400MHz Pentium II PC under the Linux operating system.

The system works well with short segments of speech leading to the possibility of adapting the algorithm to work with a finite duration window. Integrating the frame by frame score over the window length enables the identification of the start and the end of music segments within the file.

6. REFERENCES

[1] J. Saunders, 'Real–Time Discrimination of Broadcast Speech/Music', Proc. ICASSP 1996, pp993-996.

[2] M. S. Spina and V.W. Zue, 'Automatic Transcription of General Audio Data: Preliminary Analyses', Proc. ICSLP 1996, pp594-597.

[3] E. Scheier and M Slaney, 'Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator', Proc. ICASSP 1997, pp1331-1334.

[4] M. J. Carey, E. S. Parris and H. Lloyd-Thomas, 'A Comparison of Features for Speech, Music Discrimination', Proc. ICASSP 1999, pp 149-152

[5] Inmarsat - M, Voice Coding System Description. Draft Version 1.3, February 1991, Inmarsat.

[6] E. S. Parris and M. J. Carey, 'Language Independent Gender Identification', Proc. ICASSP 1996, pp685-688.

[7] M. J Carey, E. S Parris, S. Bennett, and H Lloyd-Thomas, 'Robust Prosodic Features For Speaker Identification' Proc. ICSLP 1996, pp1800-1803.

[8] R. Rose and R Reynolds, 'Text Independent Speaker Identification Using Automatic Acoustic Segmentation', Proc. ICASSP 1990, pp293-296.

[9] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, 'The DET Curve in Assessment of Detection Task Performance', Proc. Eurospeech 1997, pp1895-1898.