A COMPARISON OF FEATURES FOR SPEECH, MUSIC DISCRIMINATION.

Michael J. Carey, Eluned S. Parris and Harvey Lloyd-Thomas
Ensigma Ltd., Turing House, Station Road, Chepstow, Monmouthshire, U.K.
{michael eluned harvey}@ensigma.com

ABSTRACT

Several approaches have previously been taken to the problem of discriminating between speech and music signals. These have used different features as the input to the classifier and have tested and trained on different material. In this paper we examine the discrimination achieved by several different features using common training and test sets and the same classifier. The database assembled for these tests includes 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. Experiments were carried out on four types of feature, amplitude, cepstra, pitch and zero-crossings. 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%.

1. INTRODUCTION

Automatic discrimination between speech and music has become a research topic of interest in the last few years. Several approaches have been described in the recent literature [1,3,4]. Each of these uses different features and pattern classification techniques and describes results on different material. In this paper we make a comparison of several of the different features previously suggested in addition to some we believe have useful properties. We carry out these tests on the same data and use the same type of classifier. In this way we attempt to provide a comparative view of the value of the different types of features in speech music discrimination. We start by reviewing the published material and then justify the inclusion of prosodic features which we believe to be important.

Saunders [1] has described a speech music discriminator based on zero-crossings. Its suggested application is for discrimination between advertisements and programmes in radio broadcasts. Since it is intended to be incorporated in consumer radios it is intended to be low cost and simple. It is mainly designed to detect the characteristics of speech which are described as,

1. Limited Bandwidth
2. Alternate Voiced And Unvoiced Sections
3. Limited Range Of Pitch
4. Syllabic Duration Of Vowels
5. Energy Variations Between High And Low Levels

It is indirectly using the amplitude, pitch and periodicity estimate of the waveform to carry out the detection process since zero-crossings give an estimate of the dominant frequency in the waveform [2].

In reference [3] Zue and Spina use an average of the cepstral coefficients over a series of frames. This is shown to work well in distinguishing between speech and music when the speech is band-limited to 4kHz and the music to 16kHz but less well when both signals occupied a 16kHz bandwidth.

Scheier and Slaney [4] use a variety of features. These are

1. Four Hertz Modulation Energy
2. Low Energy
3. Roll Off Of The Spectrum
4. The Variance Of The Roll Off Of The Spectrum
5. Spectral Centroid
6. Variance Of The Spectral Centroid
7. Spectral Flux
8. Variance Of The Spectral Flux
9. Zero-Crossing Rate
10. The Variance Of The Zero-Crossing Rate
11. The Cepstral Residual
12. The Variance Of The Cepstral Residual
13. The Pulse Metric

The first two features are amplitude related. The next six features are derived from the fine spectrum of the input signal and therefore are related to the techniques described in the reference [3]. Features 9 and 10 use the zero-crossing rate in common with reference [1].

Considering that the chief difference between speech and music, at least in the form of singing, is the difference in the prosody of the signal it is surprising that none of the work so far has used pitch and amplitude features explicitly. A preliminary investigation of a selection of typical speech and music files showed that the distribution of the first differential of the pitch is different for speech and music. The music distribution has a strong concentration about zero delta pitch corresponding to steady notes and a significant occurrence of large pitch changes corresponding to shifts between notes. The pitch changes in speech were more evenly distributed. A similar but less pronounced difference is also observable for the delta amplitude. Hence a comparison is made between the use of the following features in a music detector,

1. Cepstral Coefficients
2. Delta Cepstral Coefficients
3. Amplitude
4. Delta Amplitude
5. Pitch
6. Delta Pitch
7. Zero-Crossing Rate
8. Delta Zero-Crossing Rate

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.

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 two 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 two 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.

The estimation of pitch produced by the algorithm is most reliable in voiced regions of speech and musical notes of reasonable duration.

2.4 Zero-Crossings

The zero-crossing feature was calculated by summing the zero-crossings over a 10ms frame. The delta zero-crossing was calculated by estimating the trend of the zero-crossing over five successive frames.

3. EXPERIMENTAL CONFIGURATION

3.1 Database

The experiments described in the following section were carried out using a database of music and speech. 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.

3.2 Experimental System

Pattern classification was carried out using Gaussian Mixture Models (GMM) [8]. Each of the possible classes of signal, speech or music, 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. Tests were carried out to establish the optimum number of mixtures in the models for each of the features. The score for each test file was computed as the difference in log likelihood ratio,

$$S_m = L_m - L_s$$

where L_m, *and* L_s are the likelihood scores for music and

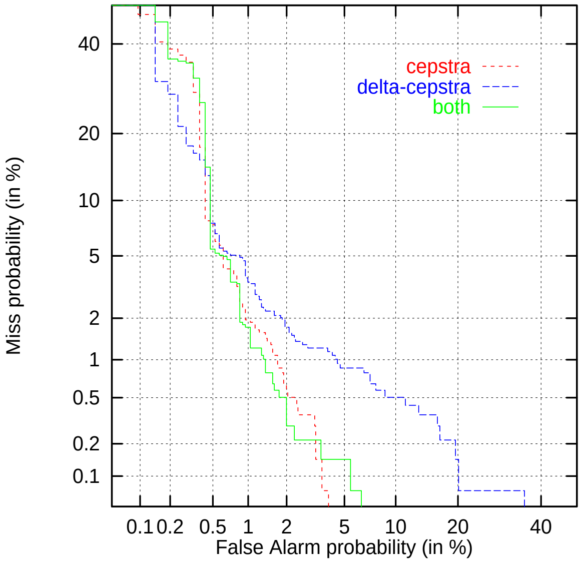

Figure 1 DET Plot for Cepstral Features

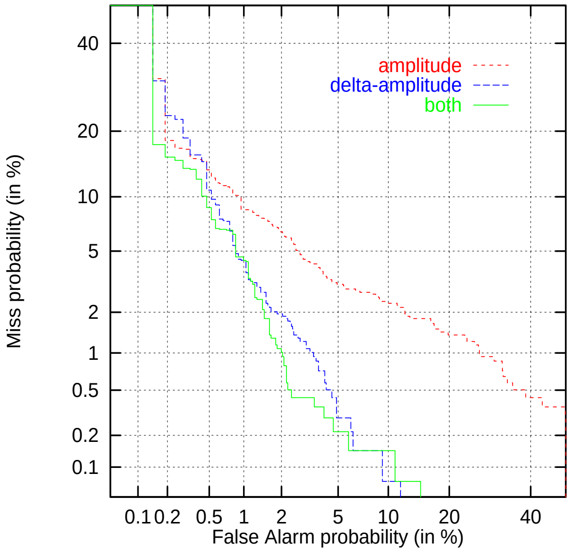

Figure 2 DET Plot for Amplitude Features

speech. The scores were then used to generate Receiver Operating Characteristics and plotted as a Detection Error Trade-off curve[9].

4. EXPERIMENTS

4.1 Cepstral Coefficients

The number of mixtures in the GMMs used to model the speech and music signals was increased from one to sixty for at which point little performance improvement was seen. The distributions of the cepstra and delta cepstra were hence modelled using a 64 mixture GMM. As can be seen from the DET plots of Figure 1, the spectral derivative represented by the delta cepstra outperformed the static feature while using both features gave a further improvement. However this was small indicating a strong correlation between the information represented by these feature sets.

Feature	Statistic	Music	Speech	EER %
Amp.	μ	0.00	0.00	3.5
	σ	1.14	2.50	
Delta Amp.	μ	0.00	0.00	2
	σ	0.02	0.53	
Zero-Crossing	μ	0.18	0.17	20
	σ	0.10	0.13	
Delta Zero-Crossing	μ	0.00	0.00	13
	σ	0.19	0.33	
Pitch	μ	75	52	9
	σ	27.3	21.7	
Delta Pitch	μ	0.28	0.05	7.5
	σ	61.5	53.5	

Table 1 Statistics of the Gaussian Mixtures for Three Types of Features

4.2 Amplitude Features

Figure 2 shows the DET plots for the amplitude features. The amplitude feature alone gave surprisingly good performance particularly since the models had single Gaussian mixtures. The delta-amplitude alone achieved an equal error rate below 2% and equal to the delta cepstra with a fraction of the computational complexity. Combining both parameters into a two dimensional feature vector and modelling the feature space with a single Gaussian resulted in an EER of 1.7%.

4.3 Pitch Features

Figure 3 shows the DET plots for the pitch features modelled by a sixteen mixture GMM. Above sixteen mixtures the performance of the GMMs used to model the pitch features deteriorated. Both the pitch and the delta pitch give similar performance. However, the combination of the features results in a significant improvement in performance with an EER of about 4%.

4.4 Zero-Crossings

Results for the two zero-crossing parameters are shown in Figure 4. The zero-crossing distribution was modelled by a single Gaussian while four Gaussian mixtures were used to model the derivative distribution and the distribution of both parameters. Although computationally inexpensive these parameters performed least well. Even when the zero-crossing features were combined, the resultant EER barely matched the performance of a single parameter set of the other feature types.

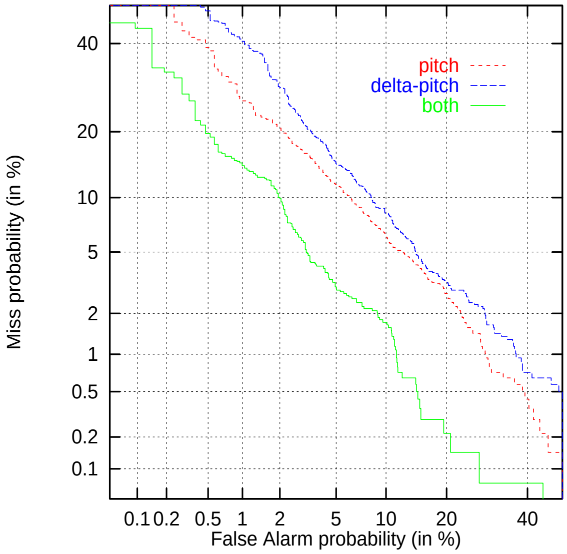

Figure 3 DET Plot for Pitch Features

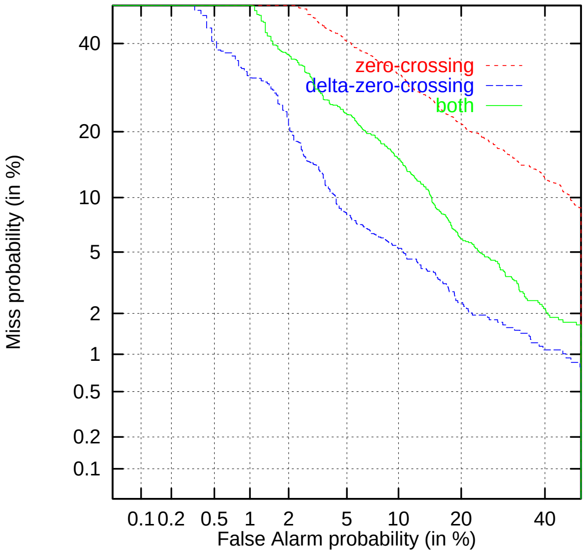

Figure 4 DET Plot for Zero-Crossing Features

5. DISCUSSION

Table 1 shows the means and standard deviations of a single mixture model for the uni-dimensional features. As expected the means of the delta features and the normalised amplitude are zero for both speech and music. Discrimination in these cases is provided by the variance of the distributions alone. The ratio of the variances of the delta amplitude for speech and music is particularly high accounting for the excellent performance of this feature. It appears that music in general has very little amplitude variation between frames when compared with the speech signal. There is less difference in variances for the other parameters resulting in less discriminating ability.

For the pitch, signal discrimination is mainly produced by the difference between the means of the speech and music distribution. The pitch and the delta pitch are extracting two different aspects of the signals, their difference in average value and the expected rate of change. This would explain the larger improvement shown by the combination of these features.

The zero-crossing rates of the signals are not a good discriminator between speech and music. The difference in the statistics of the zero-crossing rate for the two signals is small accounting for its poor performance.

Because of the higher dimensionality of the feature space and the increased complexity of the model it is more difficult to infer which of the cepstral information is providing the discrimination between the signals. However visual examination of the spectrograms for speech and music indicate that the music spectrum changes much more slowly from frame to frame than the speech spectrum. Hence the very good performance from the delta features. The cepstra require more computation than the amplitude features or zero crossings. This disadvantage will often be nullified since the cepstra will be required for other processing of the signal for example speech or speaker recognition.

We are of the opinion that there is a degree of independence between the different features we have reported in the paper. Hence the features could be combined into a larger vector or the scores from the different models can be fused to give improved performance. We intend to report on this in a future paper.

6. REFERENCES

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

[2] B. Kedam, ‘Spectral Analysis and Discrimination by Zero-Crossings’, Proc. IEEE Vol. 74 No. 11 Nov 1986, pp1477-1493.

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

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

[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. Carey, E. 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.