ROBUST PROSODIC FEATURES FOR SPEAKER IDENTIFICATION

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

Ensigma Ltd, Turing House, Station Road, Chepstow, Gwent, NP6 5PB, U.K.
michael,eluned,harvey,stephenb@ensigma.com.

ABSTRACT

This paper describes the use of prosodic features for speaker identification. Features based on the pitch and energy contours of speech are described and the relative importance of each feature for speaker identification is investigated. The mean and variance of the pitch period in voiced sections of speech are shown to be particularly useful at discriminating between speakers. Fusing these features with a Hidden Markov Model speaker identification system gave a marked improvement in figure of merit, over 30% gain was achieved on the six NIST 1995 Evaluation tests presented. Handset variability is known to have an adverse effect on performance when traditional spectral features are used e.g. cepstra. Results are presented showing that the prosodic features are more robust to handset variability.

1. INTRODUCTION

Prosodic features, that is features based on the pitch and energy contours of speech, are known to give information about the identity of a speaker. Several authors, for example [1,2] reported on the use of pitch parameters in speaker recognition in the 1970's and early 1980's. However interest in research in the use of prosodic features appears to have diminished in recent years because these features alone could not give the level of performance required for speaker identification and verification in text dependent systems and it was difficult to see how they could be incorporated in a text independent system. Pitch extraction was also error prone and computationally expensive.

Also successful results have been achieved with systems using Hidden Markov Models of the spectral envelope, for example Gaussian Mixture HMMs [3] or Ergodic HMMs [4]. Consequently little work has been carried out recently on this approach. However advances in speech coding have resulted in more reliable pitch extraction algorithms for example[5] and the computational requirements of these algorithms are easily met by presently available digital signal processors and workstations.

The effect of channel distortions and noise on the performance of speaker identification system is a serious concern. Prosodic features are known to be less effected by these impairments than spectral envelope features such as the low order cepstral coefficients. Prosodic features are therefore worth re-examining for speaker identification particularly when used to improve the performance of algorithms using Hidden Markov Model techniques.

In this paper we first show in section 2 that simple parameters such as the mean and variance of the pitch period in voiced sections of an utterance contain useful speaker discriminative information. In section 3 we describe the results achieved using a wider range of features selected from the statistics of the pitch and energy of the speech and their first and second derivatives. Section 4 briefly describes our Hidden Markov Model system which uses ergodic HMMs. In section 5 the output data of the Hidden Markov Model system are fused with the results of the prosodic model to give improved overall performance. This is demonstrated on the NIST June 1995 and February 1996 speaker evaluation test data. Section 6 discusses the robustness of the features.

2. PITCH

Recent work we have carried out on gender identification[6] indicates that a speaker's gender can be identified with 98% accuracy using the mean pitch parameter alone. This led us to believe that useful information about a speaker's identity may be contained in the speaker's mean pitch. A pitch estimation algorithm based on the IMBE speech coder [5] was therefore incorporated into our system. This was used to extract values of pitch period for the segments of speech marked as vowels by the pattern matching process in the classifier stage.

The mean pitch period was then estimated as follows. An initial estimate of the mean pitch period was made by averaging all pitch values extracted over each 10 ms frame. It was found that this estimate was biased by outliers in the pitch distribution caused by pitch halving and/or pitch doubling. The estimate was therefore refined by recalculating the mean pitch using only the samples of the pitch period found to be within plus or minus 35% of the initial estimate of the mean pitch value. This process was repeated until successive iterations gave no change in the mean pitch value. The standard deviation of the pitch period was also estimated.

For verification, an unknown speaker's pitch score was computed as the square of the distance between the mean for the unknown speaker's pitch and the average of the known speaker's speech, weighted by the inverse of the target speaker's variance.

$$S_p = \frac{(p_o - p_t)^2}{\sigma_t^2}$$

Where S_p is the pitch score p_o, p_t are the means of the pitch in voiced sections of speech in testing and training and σ_t^2 is the variance of the pitch observed in training. Tests were carried out on the male speakers in the NIST95 Speaker Identification Evaluation. Thirty seconds of training from a single file and ten

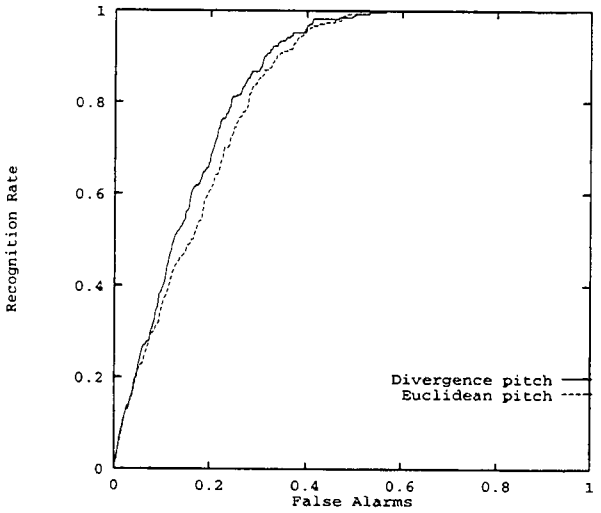

Figure 1: ROC Curve Produced by Pitch Score Using Euclidean and Divergence Measures.

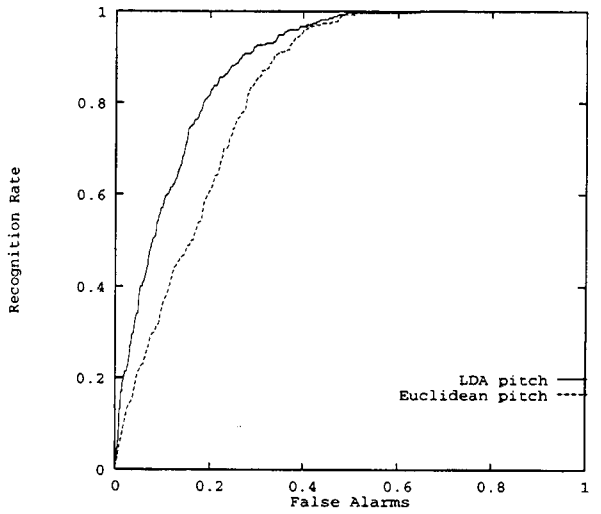

Figure 2: ROC Curve for LDA Transformed Prosodic Parameter Set.

second tests were used. The Receiver Operating Characteristic (ROC) plot is shown in Figure 1.

While this shows an impressive level of discrimination for a single parameter estimated over a whole speech file the performance can be improved by also using the variance of the test data. To do this we again assume that the distributions are gaussian and the Divergence or Kullback Leibler distance is used as a measure of the difference between the training and test distributions. The score is then computed as:

$$S_p = \frac{(p_o - p_t)^2 + \sigma_t^2}{\sigma_o^2} + \frac{(p_o - p_t)^2 + \sigma_o^2}{\sigma_t^2} - 1$$

where σ_o^2 is the variance of the test observations. This gives the improvement shown in Figure1.

3. PROSODIC FEATURES

Having established that the mean pitch alone was useful in discriminating between speakers we decided to experiment with a more extensive set of features. We therefore chose the first four statistics, mean, variance, skew and kurtosis of the pitch and energy and their first two derivatives. The derivatives were estimated using:

$$\delta p_i = \sum_{k=-2}^{k=2} k p_{i+k}$$

for the first derivative and:

$$\delta^2 p_i = \delta p_{i+1} - \delta p_{i-1}$$

for the second derivative where p_i is the estimated pitch period at frame *i*.

The mean and variance of the length of the voiced speech segments were added to these. Each of these features was then evaluated by generating a figure of merit on the NIST95 test set. The best seven parameters when tested individually were pitch mean, variance and skew, delta pitch mean and variance, energy variance and the delta energy kurtosis. Linear Discriminant Analysis was carried out on these parameters to reduce the dimensionality of the data and to weight the features optimally. The first three LDA dimensions were then retained. The effectiveness of the feature set is illustrated in Figure 3 which shows a scatterplot of the first two LDA dimensions including ellipses of points two standard deviations from the training mean. While some data from other speakers falls within the ellipses much does not showing that most impostor speakers are rejected by this measure. Figure 2 shows the ROC curve given by this measure again for the male speakers. Performance on women is worse since, as Figure 3 indicates, the points representing women are more tightly clustered.

4. SPECTRAL ENVELOPE PARAMETERS

An important objective of this work was to combine the prosodic score with that produced by a system using spectral envelope parameters, cepstra, to improve the overall performance. The system used in this experiment is now briefly described.

The acoustic 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 log power outputs of the filterbank were transformed into twelve cepstral coefficients and twelve delta cepstral coefficients at a frame rate of 10ms. These coefficients were augmented by energy and delta energy parameters to give a twenty six element feature vector. The mean of each of the cepstral parameters was estimated for each segment of speech and subtracted from each of the feature vectors. The subword models used were three state Hidden Markov Models with continuous mixture distributions and a left to right topology and no skipping of states allowed. The Expectation Maximisation

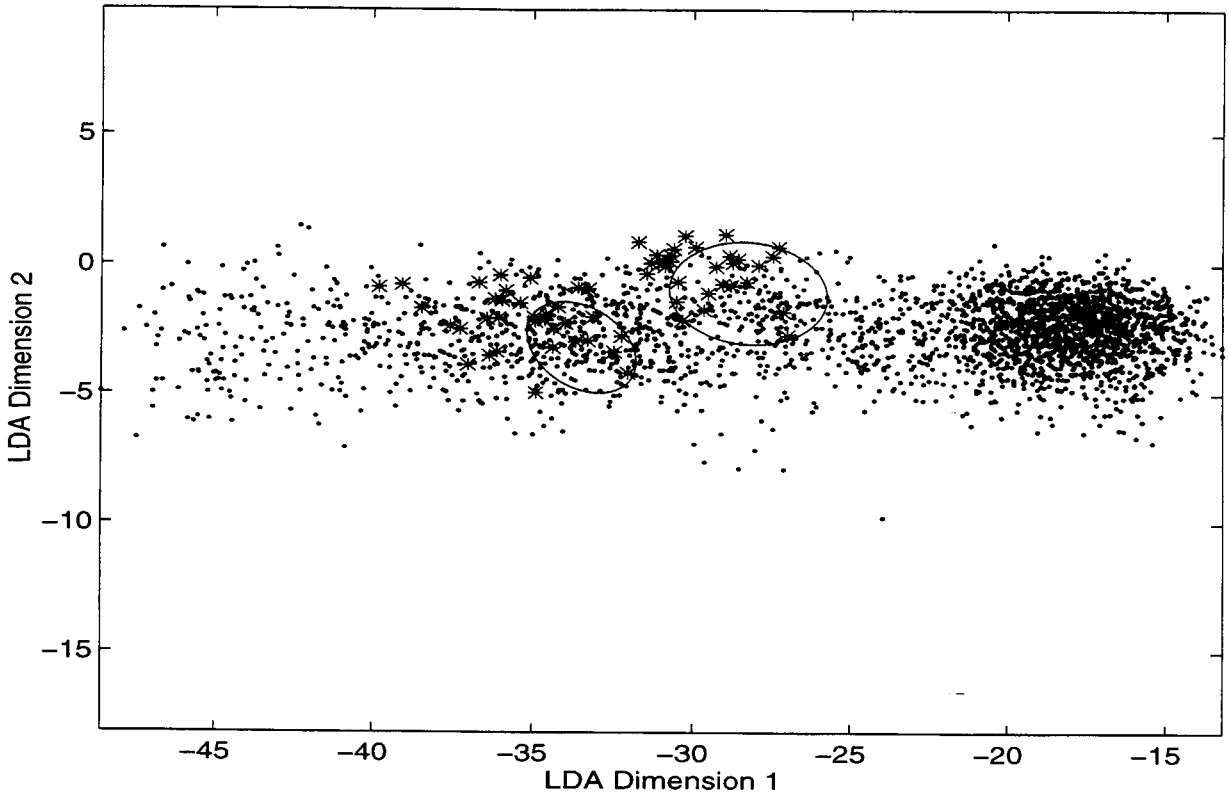

Figure 3: Training Ellipses and Scatterplot for Two Speakers and Background Speakers. The stars are true speaker test points, dots are background speakers. The cluster of background dots to the right represent the female speakers, the more scattered dots are the men.

algorithm was used for parameter estimation and the Viterbi algorithm performed pattern matching.

Speaker independent models were built as follows. A set of subword models corresponding to the forty one phonemes of American-English was built using the TIMIT database and the American-English part of the OGI Multilingual Corpus. Recognition was then performed on the training material and the results of the recognition were compared with the annotation files to give a confusion matrix between the subword models.

The number of subword classes was then reduced by combining subword units likely to be confused reducing the number of classes to twenty eight. The trained classes so combined were then used to build a new set of speaker independent models. Each model state had seven guassian mixture modes. At training time these speaker independent models were used to segment the training speech for each of the test speakers and speaker dependent models were then built from this speech. Each of the speaker dependent models had a single mode per state.

During recognition an unknown speaker's speech was matched to a set of models comprising each of the hypothesised target speaker's dependent models and a set of speaker independent models. A score was generated for each of the target speakers which was the percentage of the total matches achieved by that speaker's models. Figure 4 shows the ROC curve for the spectral envelope parameters using the NIST February 1996 data.

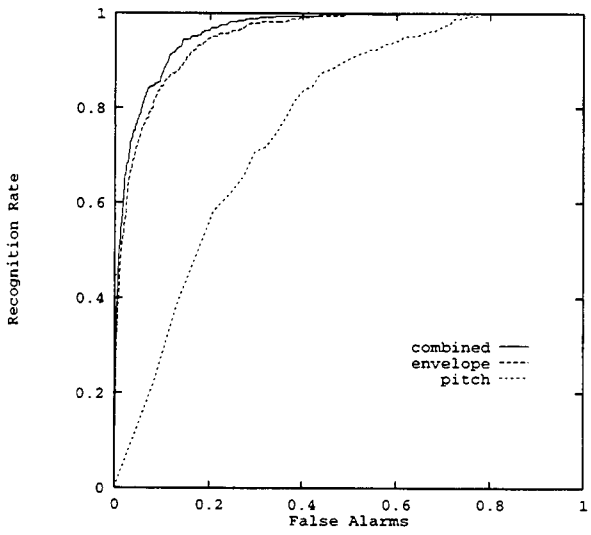

Figure 4: Improvement in ROC Curve Produced by Pitch.

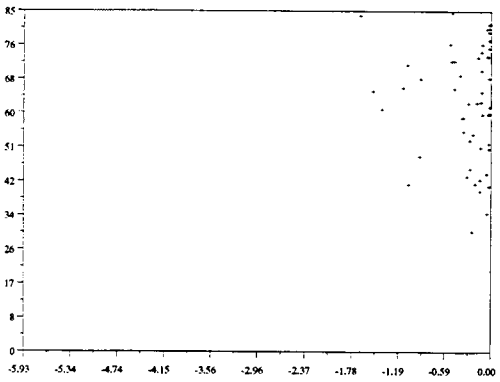

Figure 5a: Scatter Plot for True Speakers.

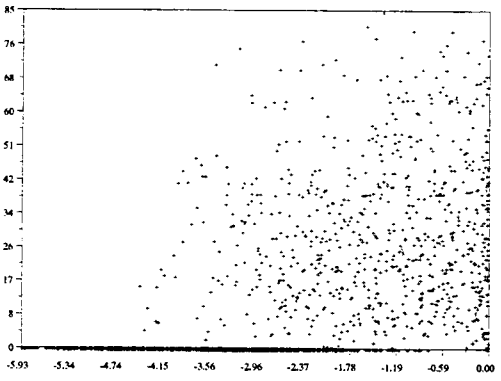

Figure 5b: Scatter Plot for Impostors.
Vertical axis is Hidden Markov Model score horizontal axis is log(1-pitch score).

5. DATA FUSION

The fusion of the scores from the pitch and envelope parameters will only be beneficial if the errors are uncorrelated. This can be assessed with reference to the scatter plots of Figures 5a and 5b. In Figure 5a where the true speakers' speech has been input to the system the majority of points are in the upper right-hand part of the plot indicating that the speaker scored well for both pitch and spectral envelope. There are very few points in the upper left-hand corner of the plot indicating that the speaker scored well for the spectral envelope but poorly for pitch. This is not the case in Figure 5b which shows impostors where a large number of points occur in the upper left-hand corner of the envelope. These points are rejected as speech from impostors using the pitch score alone.

The fusion of the two techniques is achieved by a linear classifier in which the divergence is subtracted from the spectral envelope score described in the previous section to give an overall score for that speaker on the unknown speaker's speech. The performance of the system was tested using male speaker data selected for the February 1996 NIST Speaker Verification evaluation. This comprises twenty one target speakers, 653 target trials and 24190 male impostor trials taken from the Switchboard Corpus. Training data was a two minute section of a single file for each speaker.

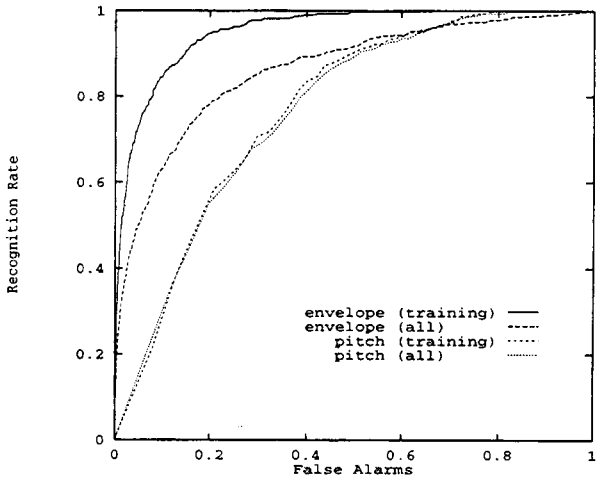

Figure 6: ROC Curves Produced by Spectral Envelope and Pitch Measures for Test Speech from the Training Handset and All Handsets.

Test data comprised, 10s sections in length and similar sections for the impostors.

6. ROBUSTNESS

The effects of the impairments introduced by telephone channels and handsets is of increasing interest as these effects are a major obstacle to higher performance[7]. Prosodic features particularly those based on pitch should be less susceptible to handset and channel effects. To test this assertion we tested the system using target test material from the training handsets alone and from both training and other handsets. The resulting ROC curves are shown in Figure 6. While the performance of the envelope parameters has been degraded the performance achieved by the pitch divergence score is little effected.

7. REFERENCES

[1] B S Atal, 'Automatic Speaker Recognition Based on Pitch Contours', JASA Vol.52 No.6 pp 1687-1697.

[2] J D Markel, B T Oshika and A H Gray, 'Long-Term Feature Averaging for Speaker Recognition', IEEE Trans ASSP Vol. ASSP-25, pp 330-337.

[3] D A Reynolds and R C Rose, 'Robust Text Independent Speaker Identification', IEEE Trans Speech and Audio Processing Vol. 3. Jan 1995 pp 72-83.

[4] E S Parris and M J Carey, 'Discriminative Phonemes for Speaker Identification.', Proc. ICSLP 1994, Yokohama pp. 1843-1846.

[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 Atlanta, pp 685-688.

[7] D. A Reynolds et al. 'The Effects of Telephone Transmission Degradations on Speaker Recognition Performance', Proc ICASSP 1995, Detroit pp 329-332.