LANGUAGE INDEPENDENT GENDER IDENTIFICATION

Eluned S Parris and Michael J Carey.

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

ABSTRACT

This paper describes a novel technique specifically developed for gender identification which combines acoustic analysis and pitch. Two sets of Hidden Markov models, male and female, are matched to the speech using the Viterbi algorithm and the most likely sequence of models with corresponding likelihood scores are produced. Linear discriminant analysis is used to normalise the models and reduce bias towards a particular gender. An enhanced version of the pitch estimation algorithm used for IMBE speech coding is used to give an average pitch estimate for the speaker. The information provided by the acoustic analysis and pitch estimation are combined using a linear classifier to identify the gender of the speech. The system was tested on three British English databases giving less than 1% identification error rate with two seconds of speech. Further tests without optimisation on eleven languages of the OGI database gave error rates less than 5.2% and an average of 2.0%

1. INTRODUCTION

There are relatively few papers published on gender identification to date and it is often treated as a by-product of research in other areas e.g. speech recognition [1]. This paper describes a novel technique specifically developed for gender identification which combines acoustic analysis and pitch. Male and female speaker independent Hidden Markov Models (HMMs) are used to model the spectral envelope. These are matched to the speech to produce acoustic likelihood scores. Broad class models and linear discriminant analysis are used to achieve accurate segmentation. The pitch of speech is a key indicator to the gender of the speaker. Pitch estimation techniques based on speech coding ideas [2] have been enhanced to produce likelihood scores for each gender. Finally, the scores produced by the two techniques are fused using a linear classifier similar to that developed on our work on language identification [3]. This combines the acoustic and pitch information to produce better results than either technique in isolation.

An accurate gender identification system is useful for many speech identification and recognition problems. In our previous work on speaker verification [4] we found that nearly half of the false acceptances were caused by impostors of the opposite sex. Using the extra gender information would reduce the error rates considerably for verification and speaker recognition. Gender identification can also be used in speech recognition where better performance is generally achieved by using separate acoustic models for men and women.

This paper describes experiments carried out on gender identification using small amounts of speech, typically one to five seconds. Section 2 describes the acoustic analysis and the use of linear discriminant analysis for normalising the models. Section 3 describes the enhancements made to the IMBE speech coding pitch estimation algorithm. Section 4 describes how the information provided by acoustic analysis and pitch estimation were combined to give improved results. Section 5 describes initial experiments on British English producing an identification rate of over 99%. Further experiments are also presented on eleven other languages in the Oregon Graduate Institute (OGI) Multi Lingual Corpus and show that the technique is language independent.

2. ACOUSTIC ANALYSIS

This section describes the acoustic analysis used for gender identification. The speech data was sampled at 8 kHz and filtered using a filterbank analyser with nineteen mel spaced filters and a frame rate of 10 ms. Twenty six features were calculated each frame, twelve cepstra, twelve transitional cepstra, energy and transitional energy. The transitional cepstra were calculated over a 50 ms window by taking a weighted sum of cepstra over a series of frames.

0-7803-3192-3/96 $5.00©1996 IEEE

In training two sets of HMMs, male and female, were built from the Subscriber database [5]. This is a large British English database collected over the telephone network and includes over one thousand talkers from throughout the British Isles. Initially a set of forty four subword models were built representing all of the British English phonemes. The HMMs were three state, left to right models with no skipping of states allowed. Two confusion matrices were then produced for the forty four male and female models and used to give a reduced set of classes. Eight broad classes were chosen where each class contained subwords which were regularly confused. Most of the confusions occurred as expected between phonemes belonging to the same broad class e.g. nasals, stops. Between nine and fourteen mixture densities were used in each state to produce the best segmentation of the data.

Previous work has shown that linear discriminant analysis (LDA) [6,7] can be used to normalise models and also reduce the number of parameters required. In gender identification we transformed the male and female models using LDA thus reducing bias towards any particular models. Each mixture density in each male and female model was treated as a separate class. The speech data in Subscriber was pooled by aligning frames to the mixture densities in the models. An LDA transform was produced from the pooled data and new male and female models constructed directly from the pools.

The gender identification was performed by matching the twenty six features produced each frame to the male and female models. The Viterbi algorithm was used to produce the most likely sequence of models and a corresponding set of likelihood scores. It was hypothesised that the female models would be matched in preference to the male models for a female speaker's speech and vice versa. The number of matches to female models and number of matches to male models were accumulated over the segment of speech to be identified. This technique can be used in isolation for gender identification by choosing the gender of the set of models which produced most matches.

3. PITCH ESTIMATION

A major difference between male and female speech is the pitch. In general, female speech has higher pitch (120 - 200 Hz) than male speech (60 - 120 Hz) and could therefore be used to discriminate between men and women if an accurate pitch estimate could be calculated. Pitch has been widely used in speech coding research but rarely with any success in speech recognition or identification. Therefore the pitch estimation algorithm used for IMBE speech coding [2] was used for gender identification and a number of enhancements were made to the algorithm.

The IMBE technique calculates an initial pitch estimate by correlating the 1 kHz low pass filtered speech with delayed versions of the same signal. The range of delays considered spans the pitch period applicable to speech (20 to 120 samples). 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 interpulse interval. The high resolution results from the spectral match at the high frequency harmonics.

The estimation of pitch produced by the IMBE algorithm is most reliable in voiced regions of speech. In particular, it is reliable in steady state regions e.g. vowels. The pattern matcher described in the previous section identifies the vowel and diphthong classes with over 90% accuracy. Therefore, it was decided for gender identification to use only the pitch estimates in regions identified by the pattern matcher as being in these classes.

Gender identification experiments were performed using pitch by estimating the pitch for each frame of the speech. These estimates were stored until the pattern matcher identified regions of interest. An initial estimate of the average pitch was calculated across these regions and then refined by calculating a new average from pitch estimates within a percentage of the original average. This removed outliers produced by pitch doubling, tripling etc. and errors in region classification. These refinements improved the gender identification performance considerably. This technique using pitch can be used in isolation for gender identification by comparing the average pitch estimate with a preset threshold. Estimates below the threshold are identified as male and those above as female.

4. COMBINING KNOWLEDGE SOURCES

The techniques described in the previous sections provide different information about the gender of a speaker. Previous work in data fusion [8] and our quadrant classifier used for language identification [3] have shown how independent techniques can be combined to improve overall identification performance. The knowledge sources used must produce differences in classifier output. The errors produced by the acoustic analysis and pitch estimation were found to differ in the majority of cases. It was also found that files assigned a low confidence measure by one technique were generally assigned a high confidence measure by the other technique.The techniques

Technique	Error Rate %
Acoustic Analysis without LDA	12.5
Acoustic Analysis with LDA	4.2
Pitch Period Estimation	2.8
Combined Knowledge Sources	0.7

Table 1. Gender Identification Results for British English

were combined for gender identification as follows. The outputs produced by the acoustic analysis and pitch estimation were mapped onto two new parameters. When the two parameters agreed about the gender then the speaker was identified to be of that gender. Otherwise a weighted summation of the parameters was made. The probability that the speech is uttered by a man $p(m)$ is given by

$$p(m) = \alpha\, p(m \mid AC) + \beta\, p(m \mid PE)$$

where

$p(m \mid AC)$ is the probability of the speech being a man given the acoustic data,

$p(m \mid PE)$ is the probability of the speech being a man given the pitch data,

α and β are weightings optimised for the training set.

Similarly for $p(w)$, the probability that the speech is uttered by a woman.

The value for the slope of the mapping function was chosen to optimise the performance on the training data where the two parameters disagreed. This focused the algorithm on discriminating between parameters where errors occurred eliminating from the process files with no errors.

5. EXPERIMENTS

The Subscriber database was used to develop the techniques and to set the thresholds for gender identification. Three other British English databases were used to evaluate the performance of each technique, SCRIBE, Bramshill and a telephone database collected in-house. Table 1 shows the identification error rates achieved on the test databases using five seconds of speech. The use of LDA to remove bias towards the models improved the gender identification performance

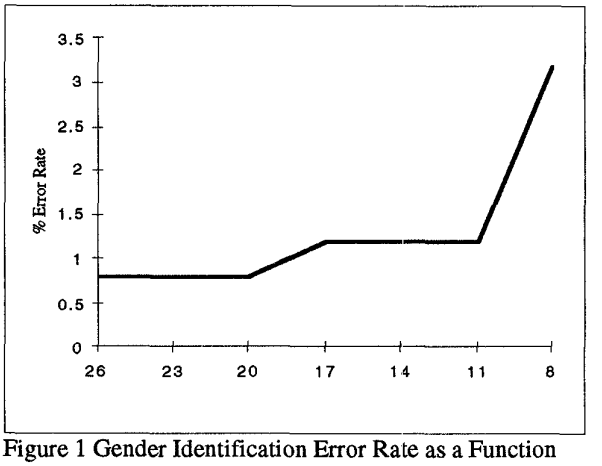

Figure 1 Gender Identification Error Rate as a Function of LDA Dimension for British English, 5s Speech.

by reducing the error rate by two thirds. The pitch estimation technique alone produced results of 2.8% error. Combining the two techniques gave 0.7% error, a statistically significant improvement over any single technique. The performance of the combined knowledge sources identifier was maintained down to two seconds of speech with 2.0% identification error achieved with just one second of speech. The system also maintained its performance when the dimensionality of the LDA feature vector was reduced as Figure 1 illustrates.

A series of experiments were carried out using the British English gender identification system with foreign speech files. No changes were made to the system to optimise thresholds for any language. The story files from eleven languages in the OGI database were used to test the system. Up to five seconds of speech were used in each test. Table 2 shows the identification rates achieved on each language. The identification error rates vary across languages from 5.2% for Tamil down to no errors observed for Mandarin and Vietnamese. It was found that for the languages with the poorest results the errors exhibited bias towards one gender, e.g. in Tamil and Hindi all of the errors were due to misclassification of male speakers. This implied that higher pitch and higher formant frequencies were present than those seen in British English speakers. Similarly the results for American English were all misclassification of women as men. A simple change in threshold for the pitch estimation improves the performance.

The gender identification system can be optimised further for a given language by using Hidden Markov models built from speech taken from the language. This improves the acoustic likelihood score input to the classifier. A second test of American English on the Switchboard database using British English models and pitch

Language	Error Rate %
American English	2.7
Farsi	3.1
French	1.0
German	1.9
Japanese	1.2
Korean	2.3
Hindi	3.6
Mandarin	0.0
Spanish	0.9
Tamil	5.2
Vietnamese	0.0
Average	2.0

Table 2 Gender Identification Results for Other Languages

thresholds gave similar results to OGI. However, changing the thresholds in the system and using American English models reduced the error rate to 0.7%, close to that achieved with British English.

6. CONCLUSIONS

This paper has described a novel technique for performing highly accurate gender identification. Two separate knowledge sources, acoustic analysis matches to broad class HMMs for each gender and the speaker's mean pitch, were combined to give less than 1% identification error rate on three British English databases with two seconds of speech. The technique has also been shown to work well on eleven other languages without modification giving a worst case identification error rate of 5.2% for Tamil and an average error rate of 2%. It has also been shown that optimisation for a specific language, American English, improves these results which are then close to those achieved with British English.

7. ACKNOWLEDGEMENT

The authors would like to thank BT Laboratories, Martlesham, U.K. for the Subscriber database.

8. REFERENCES

[1] L. Lamel and J. Gauvain. 'Cross-Lingual Experiments with Phone Recognition', Proc ICASSP 1993, Minneapolis.

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

[3] E. S. Parris and M. J. Carey. 'Language Identification Using Multiple Knowledge Sources', Proc ICASSP 1995, Detriot.

[4] M. J. Carey and E. S. Parris. 'A Speaker Verification System Using Alphanets', Proc ICASSP 1991, Toronto.

[5] A. D. Simons and K. Edwards. 'Subscriber - A Phonetically Annotated Telephony Database', Proc Institute of Acoustics 1992, Vol. 14, Part 6.

[6] M. Hunt et al. 'An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination', Proc ICASSP 1991, Toronto.

[7] E. S. Parris and M. J. Carey. 'Estimating Linear Discriminant Parameters for Continuous Density Hidden Markov Models', Proc ICSLP 1994, Yokohama.

[8] R. Ricart et al. 'Speaker Recognition in Tactical Communications', Proc ICASSP 1994, Adelaide.