An HMM-Based Subband Processing Approach to Speaker Identification

J. E. Higgins and R. I. Damper

Image, Speech and Intelligent Systems Research Group
Department of Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ, UK.
[jeh97r|rid]@ecs.soton.ac.uk

Abstract. This paper contributes to the growing literature confirming the effectiveness of subband processing for speaker recognition. Specifically, we investigate speaker identification from noisy test speech modelled using linear prediction and hidden Markov models (HMMs). After filtering the wideband signal into subbands, the output time trajectory of each is represented by 12 pseudo-cepstral coefficients which are used to train and test individual HMMs. During recognition, the HMM outputs are combined to produce an overall score for each test utterance. We find that, for particular numbers of filters, subband processing outperforms traditional wideband techniques.

1 Introduction

Automatic speaker recognition is an important, emerging technology with many potential applications in commerce and business, security, surveillance etc. Recent attention in speaker recognition has focussed on the use of subband processing, whereby the wideband signal is preprocessed by a bank of bandpass filters to give a set of time-varying outputs, which are individually processed (Besacier and Bonastre 1997, 2000). Because these subband signals vary slowly relative to the wideband signal, the problem of representing them by some data model should be simplified (Finan, Damper, and Sapeluk 2001).

The subband approach has also become popular in recent years in *speech* recognition (Bourlard and Dupont 1996; Tibrewala and Hermansky 1997; Morris, Hagen, and Bourlard 1999). In this related area, the main motivation has been to achieve robust recognition in the face of noise. The key idea is that the recombination process allows the overall decision to be made taking into account any noise contaminating one or more of the partial bands. Hence, we investigate subband speaker identification in which narrowband noise is added to test utterances. The speech is modelled using linear prediction and hidden Markov models (HMMs).

The remainder of this paper is organised as follows. Section 2 describes subband processing and its possible benefits to an identification system. Section 3

briefly describes the speech database used and Section 4 details the feature extraction and data modelling processes. In Section 5, we describe the recombination of subband information and the decision rule used for the final identification. Section 6 gives results and Section 7 concludes.

2 Subband Processing

Figure 1 shows a schematic of the subband system used here. The bandpass filters are sixth-order Butterworth with infinite impulse response, designed using the bilinear transform. They are equally spaced on the mel-scale (Stevens and Volkmann 1940). Filtering was performed in the time domain by direct calculation from the difference (recurrence) equation. Feature extraction is performed on each subband, and the resulting sequences of feature vectors are passed on to each subband's recognition algorithm. Thereafter, the outputs from each separate recogniser are fused, using multiple classifier techniques, to produce an overall decision as to the identity of the speaker.

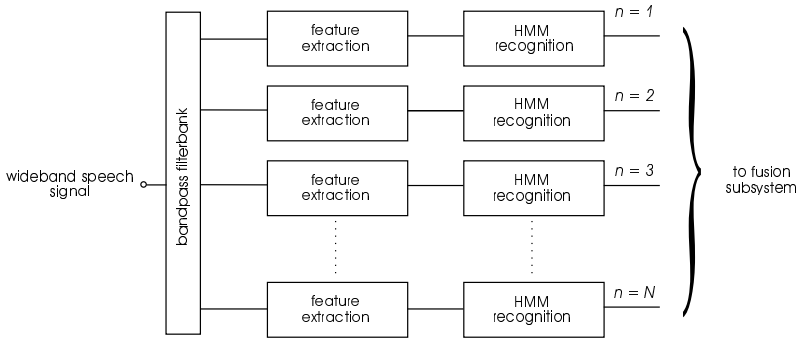

Fig. 1. Schematic diagram of the subband processing system. Each subband (filter) has its own recognition subsystem, whose output is fed to a fusion algorithm which makes the final, overall decision about speaker identity.

A successful recognition system is critically dependent on building good speaker models from the training data. In the system of Fig. 1, the problem arises at two points: extraction of features to represent the signal and building the recognition model. Data modelling, however, is subject to the well-known bias/variance dilemma (Geman, Bienenstock, and Doursat 1992). According to this, models with too many adjustable parameters (relative to the amount of training data) will tend to overfit the data, exhibiting high variance, and so will generalise poorly. On the other hand, models with too few parameters will be over regularised, or biased, and will be incapable of fitting the inherent variability of the data. Subband processing offers a practical solution by replacing a large unconstrained data modelling problem by several smaller (and hence more constrained) problems (Finan, Damper, and Sapeluk 2001).

3 Speech Database

In this work, we use the text-dependent British Telecom Millar database, specifically designed and recorded for text-dependent speaker recognition research. It consists of 46 male and 14 female native English speakers saying the digits *one* to *nine*, *zero*, *nought* and *oh* 25 times each. Recordings were made in 5 sessions spaced over 3 months, to capture the variation in speaker's voices over time which is one of the most important aspects of speaker recognition (Furui 1974).

The speech was recorded in a quiet environment using a high-quality microphone, and a sampling rate of 20 kHz with 16-bit resolution. The speech data used here were downsampled to 8 kHz sampling rate, both to reduce the computation time necessary for our simulations and because this bandwidth is more typical of a real application. Data from the first two sessions (i.e., 10 repetitions of *seven*) were used for training and data from the remaining three sessions (15 repetitions) were used for testing.

As so far described, the speech data are essentially noise-free. However a major motivation behind subband processing has been the prospect of achieving good recognition performance in the presence of narrowband noise. Such noise affects the entire wideband model but only a small number of subbands. Hence, we have conducted identification tests with added noise. Following Besacier and Bonastre 2000, Gaussian noise was filtered using a sixth-order Butterworth filter with centre frequency 987 Hz and bandwidth 365 Hz. It was added to the test tokens at a signal-to-noise ratio of 10 dB.

4 Data Modelling

In this work, we first have to model the speech signal. This is done by extracting features on a frame-by-frame basis. Many possible features could be extracted from the speech but here the feature set is based on cepstral coefficients. Cepstral analysis is motivated by, and designed for, problems centred on voiced speech (Deller, Proakis, and Hansen 1993). It also works well for unvoiced sounds. Cepstral coefficients have been used extensively in speaker recognition (Furui 1981; Reynolds and Rose 1995), mainly because a simple recursive relation exists that approximately transforms easily-obtained linear prediction coefficients into 'pseudo' cepstral ones (Atal 1974). The analysis frame was 20 ms long, Hamming windowed and overlapping by 50%. The first 12 coefficients were used (ignoring the zeroth cepstral coefficient, as usual).

Subsequently, we have to derive recognition models for the word *seven* spoken by the different speakers. For this, we use the popular hidden Markov models (HMMs). HMMs are powerful statistical models of sequential data that have been used extensively for many speech applications (Rabiner 1989). They consist of an underlying (hidden) stochastic process that can only be observed through a set of stochastic processes that produces an observation sequence. In the case of speech, this observation sequence is the series of feature vectors that have been extracted from an utterance (Section 4). Discrete HMMs were used with

four states, plus a start and end state. Apart from self-loops (staying in the same state), only left-to-right transitions are allowed. The frames of speech data were vector quantised and each HMM has its own linear codebook of size 32. Codebooks were constructed using a Euclidean distance metric. HMMs were trained and tested using the HTK software of Young, Kershaw, Odell, Ollason, Valtchev, and Woodland (2000).

5 Score Combination and Decision Rule

Kittler, Hatef, Duin, and Matas (1998) developed a common theoretical framework for combining classifiers which use distinct pattern representations. They outlined a number of possible combination schemes such as product, sum, min, max, and majority vote rules, and compared their performance empirically using two different pattern recognition problems. They found that the sum rule outperformed the other classifier combination schemes, in spite of theoretical assumptions apparently stronger than for the product rule. Further investigation indicated that the sum rule was the most resilient to estimation errors, which almost certainly explains its superior performance.

In this work, the HMM recognisers produce log probabilities as outputs. The use of logarithms is conventional, to avoid arithmetic underflow during computation. The fusion rule used here is that the identified speaker, i, is that for whom:

$$i = \arg\max_s [y^s] = \arg\max_s \sum_{n=1}^{N} \log p(\mathbf{x}|\omega(n,s)) \qquad 1 \leq s \leq S$$

where N is the number of classifiers (subbands), and $p(\mathbf{x}|\omega(n,s))$ is the probability that model $\omega(n,s)$ for classifier n and model speaker s produced the observed data sequence $\mathbf{x}$, and y^s is the recombined (final) score for speaker s from the set of S speakers. Because of the use of logarithms, this is effectively the product rule but other rules tried worked no better.

6 Results

To test the subband system all 60 speakers in the database were used and the number of subbands was varied from 2 to 10. We compare the performance with the wideband (unfiltered) speaker identification system. The results are depicted in Figure 2.

Several interesting points can be gleaned from this figure. First, for six or more subbands, the subband system outperforms the wideband system. The best correct identification of 97.7% is obtained using 10 subbands, which was the maximum number used. The wideband system only achieves 65.2%. These results confirm the advantage of using a subband system in the face of narrowband noise. Second, using a small number of filters (< 6), subband performance is worse than the wideband system. The reason for this is currently unknown but

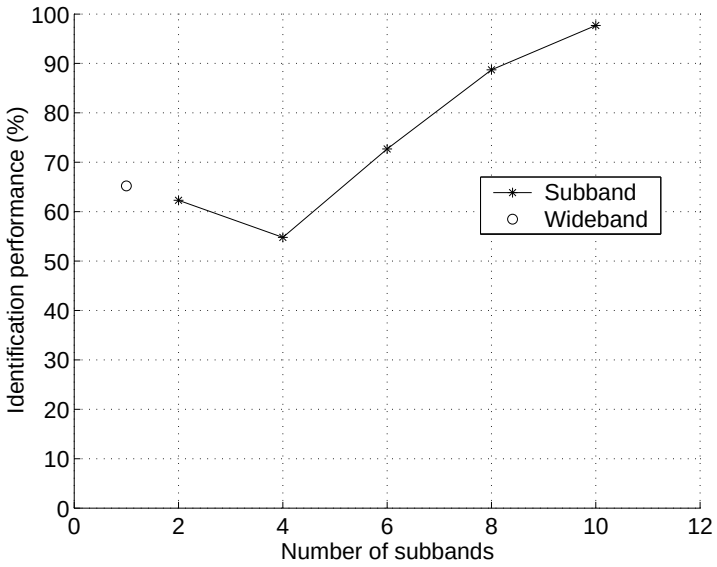

Fig. 2. Results for a 4-state HMM using 60 speakers, product fusion rule and various numbers of subbands. For comparison, the performance of a single, wideband HMM recogniser is also shown.

is possibly because of the shape and location of the filters relative to the centre frequency of the noise and/or those frequency regions which are important for discriminating speakers.

We are currently running further experiments in which larger numbers of subbands are being used as well as testing words other than just *seven*. These more complete results will be reported at the conference.

7 Conclusions

This paper contributes to the growing literature confirming the effectiveness of subband processing for speaker identification. The results confirm that subband processing offers improved performance (compared to the wideband system) in the face of narrowband noise. For the subband system tested here, ten subbands gave the best result. Future work will explore the use of more subbands and different words. We will also attempt to understand why the performance dips with four subbands.

Acknowledgements

The filter design program used here was written by Robert Finan. Author JEH is supported by a research studentship from the UK Engineering and Physical Science Research Council.

References

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America 55*(6), 1304–1312.

Besacier, L. and J.-F. Bonastre (1997). Subband approach for automatic speaker recognition: Optimal division of the frequency domain. In *Proceedings of 1st International Conference on Audio- and Visual-Based Biometric Person Authentication (AVBPA)*, Crans-Montana, Switzerland, pp. 195–202.

Besacier, L. and J.-F. Bonastre (2000). Subband architecture for automatic speaker recognition. *Signal Processing 80*(7), 1245–1259.

Bourlard, H. and S. Dupont (1996). A new ASR approach based on independent processing and recombination of partial frequency bands. In *Proceedings of Fourth International Conference on Spoken Language Processing, ICSLP'96*, Volume 1, Philadelphia, PA, pp. 426–429.

Deller, J. R., J. P. Proakis, and J. H. L. Hansen (1993). *Discrete-Time Processing of Speech Signals*. Englewood Cliffs, NJ: MacMillan.

Finan, R. A., R. I. Damper, and A. T. Sapeluk (2001). Text-dependent speaker recognition using sub-band processing. *International Journal of Speech Technology 4*(1), 45–62.

Furui, S. (1974). An analysis of long-term variation of feature parameters of speech and its application to talker recognition. *Electronic Communications 57-A*, 34–42.

Furui, S. (1981). Cepstral analysis techniques for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-29*(2), 254–272.

Geman, S., E. Bienenstock, and R. Doursat (1992). Neural networks and the bias/variance dilemma. *Neural Computation 4*(1), 1–58.

Kittler, J., M. Hatef, R. P. W. Duin, and J. Matas (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*(3), 226–239.

Morris, A., A. Hagen, and H. Bourlard (1999). The full-combination sub-bands approach to noise robust HMM/ANN-based ASR. In *Proceedings of 6th European Conference on Speech Communication and Technology, Eurospeech'99*, Volume 2, Budapest, Hungary, pp. 599–602.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–285.

Reynolds, D. A. and R. C. Rose (1995). Robust text-independent speaker identification using Gaussian mixture models. *IEEE Transactions on Speech and Audio Processing 3*(1), 72–83.

Stevens, S. S. and J. Volkmann (1940). The relation of pitch to frequency: A revised scale. *American Journal of Psychology 53*(3), 329–353.

Tibrewala, S. and H. Hermansky (1997). Sub-band based recognition of noisy speech. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'97*, Volume II, Munich, Germany, pp. 1255–1258.

Young, S., J. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland (2000). The HTK Book. Available from URL `http://htk.eng.cam.ac.uk/`.