Information Fusion for Subband-HMM Speaker Recognition

J. E. Higgins R. I. Damper T. J. Dodd

Image, Speech and Intelligent Systems Research Group,
Department of Electronics and Computer Science,
University of Southampton,
Southampton SO17 1BJ, UK.
email:{jeh97r,rid,tjd}@ecs.soton.ac.uk

Abstract

Previous work has demonstrated the performance gains that can be obtained in speaker recognition by applying subband processing, together with hidden Markov modelling and multiple classifier recombination. Two recombination rules have been investigated: the sum of log likelihoods, which corresponds to the optimal Bayes' rule under certain constraints, and multilayer perceptrons (MLP), which are not subject to these constraints. It was found that for two spoken digits in the presence of a single case of narrowband noise the sum of log likelihoods and MLP achieved comparable performance. In this paper, the previous work is extended in the direction of investigating the robustness of the recognition system to different narrowband noise. Two approaches are taken towards this aim. Firstly, narrowband noise is added at different centre frequencies. Secondly, a Bayesian MLP approach is investigated using automatic relevance determination (ARD) on the subband inputs to the MLP. From this it is possible to assess the relative importance of the subbands to recognition performance. Results for the new noise conditions show that the sum of log likelihoods generally does better than the (average) MLP fusion.

1 Introduction

Automatic speaker recognition is an important, emerging technology with many potential applications in commerce and business, security, surveillance etc. [6]. Recent attention in speaker recognition has focussed on the use of subband processing, whereby the wideband signal is fed to a bank of bandpass filters to give a set of time-varying outputs, which are individually processed before using multiple classifier techniques to produce a combined, overall decision [2, 3, 22, 13], see Figure 1. Because the subband signals vary slowly relative to the wideband signal, the problem of representing them by some data model should be simplified [9]. Previous work has demonstrated the performance gains that can be obtained in speaker recognition by applying subband processing, together with hidden Markov modelling and multiple classifier recombination. Two recombination rules have been investigated: the sum of log likelihoods, which corresponds to the optimal Bayes' rule under certain constraints, and multilayer perceptrons (MLP), which are not subject to these constraints. It was found that for two spoken digits in the presence of a single case of narrowband noise the sum of log likelihoods and MLP achieved comparable performance. In this paper the previous work is extended in the direction of investigating the robustness of the recognition system to different narrowband noise. Two approaches are taken towards this aim. Firstly, narrowband noise is added at different centre frequencies. Secondly, a Bayesian MLP approach is investigated using automatic relevance determination (ARD) on the subband inputs to the MLP. From this it is possible to assess the relative importance of the subbands to recognition performance.

The subband, or multiple classifier, approach has also become popular in recent years in *speech* recognition [5, 24, 16]. 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, an important aim of the paper is to investigate the robustness of subband speaker recognition with different recombination rules to instances of narrowband noise with various centre frequencies.

0-7803-7044-9/01/$10.00 ©2001 IEEE

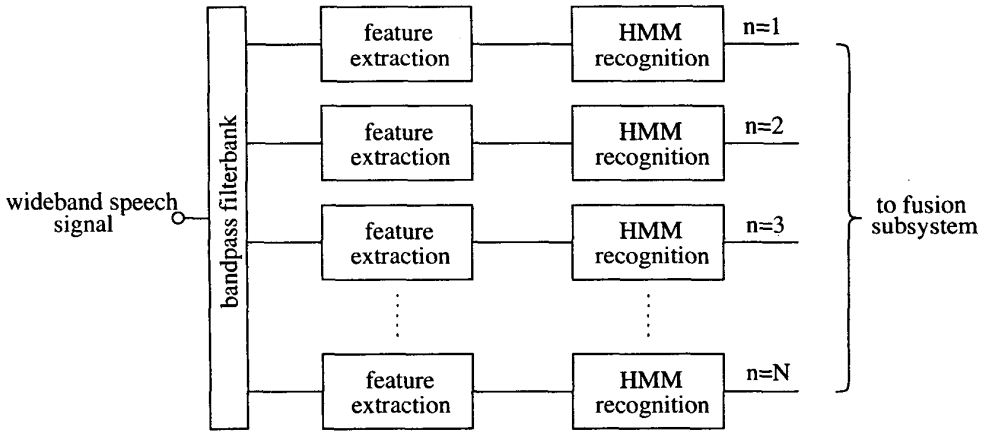

Figure 1: Schematic diagram of the subband processing subsystem used to model each speaker-word combination. Each subband (filter) has its own HMM recogniser. In this work, we use either 2, 4, 6, 8 or 10 6th-order Butterworth filters with centre frequencies equally spaced on the mel scale. There is one such subsystem for each speaker.

The remainder of this paper is organised as follows. In Section 2 we describe essential background on the problem of speaker recognition. The speech database used in obtaining the results is described in Section 3. The subband processing system, feature extraction and data modelling are introduced in Section 4. In Section 5 we detail the various recombination techniques studied, with results presented in Section 6. Finally, various conclusions are drawn in Section 7.

2 The Speaker Recognition Problem

Speaker recognition is concerned with the problems of *verification* and *identification*, each of which may in turn be *text-dependent* or *text-independent* [6, 12]. In verification, the aim is to determine if a given utterance was produced by a claimed speaker. This is most directly done by testing the utterance against a model of the claimed speaker, comparing the score to some threshold. On the basis of this comparison a decision is made whether or not to accept the claimant. In identification however, the aim is to determine which speaker, from amongst a known group, produced the utterance. The test utterance is scored against all possible speaker models, with the best score determining the speaker identity. Of the two tasks, identification is generally accepted to be the harder, especially for large speaker populations [8, p. 1660].

In text-independent recognition, there are no limits on the vocabulary employed by speakers. This is in contrast to text-dependent recognition, where the tested utterance comes from a set of predetermined words or phrases. As text-dependent recognition only models the speaker for a limited set of speech sounds ('phones') in a fixed context, it generally achieves higher performance than text-independent recognition, which must model a speaker for a variety of phones and contexts.

Since identification is simply a matter of selecting among speakers, typically using a minimum distance decision rule, performance is easily quantified by a single measure. There are only two possible outcomes—correct or incorrect—so that the identification error fully specifies the situation. Things are a little more complicated with verification where the system has to accept or reject a claimed speaker identity in the face of potential impersonation. Hence, there are four possible outcomes, of which two—false acceptance and false rejection—are errors. Thus, some decision threshold must be set which effects a balance between the two types of error. Because of the slightly increased difficulty of quantifying error in verification, we focus exclusively on identification in this paper. Also, we restrict attention to text-dependent recognition because of its more obvious applicability [8, p. 1660].

3 Speech Database

We use the text-dependent British Telecom Millar database, specifically designed and recorded for text-dependent speaker recognition research. It consists of 60 (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 an important aspect of speaker recognition [10].

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 as this reduces simulation times and is more typical of the data which might be encountered in a real application. For the work reported here we are only considering test-dependent identification and therefore investigate utterances *seven* and *nine* only. Data from the first two sessions (i.e., 10 repetitions of each word) were used for training and data from the remaining three sessions (15 repetitions) were used for testing.

In order to achieve good performance, manual editing of the start and end points of each utterance was necessary. This was done by author JEH. This was a time-consuming task: For a fully automatic system, we would obviously need to implement a high performance automatic endpointing algorithm.

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. Previously a single case of narrowband noise has been considered for which subband processing was demonstrated to show a significant improvement in recognition over the wideband system. In order to further investigate the effects of noise we consider here identification in the presence of two instances of narrowband noise. Following [3], Gaussian noise, filtered using a 6th-order Butterworth filter, with centre frequencies 987 Hz and 2500 Hz and bandwidth 365 Hz were added. In each case the noise was added to the test tokens at a signal-to-noise ratio of 10 dB.

4 Subband Processing

The subband system used in the paper to model each different word for each individual speaker is shown in Figure 1. For each speaker-word combination we use 2, 4, 6, 8 or 10 bandpass filters (6th-order Butterworth) with the filter centre frequencies equally spaced on the psychophysically-motivated mel scale [23]. Feature extraction is performed on each subband with the resulting sequences of feature vectors passed to each subband's HMM recognition algorithm.

The speech signal is modelled using cepstral coefficients, obtained on a frame-by-frame basis, as features which are obtained from linear prediction [15]. Cepstral analysis is motivated by, and designed for, problems centred on voiced speech [7]. In practice, it also works well for unvoiced sounds and should therefore be applicable to our database. Cepstral coefficients have been used extensively in speaker recognition [11, 21], partly because a simple recursive relation exists that approximately transforms the easily-obtained linear prediction coefficients into 'pseudo' cepstral ones [1]. The analysis frame used was 16 ms long, Hamming windowed and overlapping by 50%. The first 12 cepstral coefficients were used, excluding the zeroth coefficient (as is usual).

Subsequently we apply hidden Markov models (HMMs) as recognition models for the utterances of the different speakers. HMMs are powerful statistical models of sequential data that have been used extensively for many speech applications [20]. They embody an underlying (hidden) stochastic process that can only be observed through a set of stochastic processes that produces an observation sequence. In speech processing applications, this observation sequence is the series of feature vectors that have been extracted from an utterance.

Discrete HMMs were used with 4 states for word *seven* and 3 states for word *nine*, plus a start and end state in each case. This structure was found to give the best results in preliminary tests. Apart from self-loops (staying in the same state), only left-to-right transitions are allowed. Speech frames were vector quantised, and each HMM has its own linear codebook of size 32. Therefore, in the wideband case there are 60 codebooks (equal to the number of speakers) and in the subband system there are $60 \times N$ codebooks (where N is the number of subbands), which were constructed using a Euclidean distance metric. HMMs were trained and tested using the HTK software of [25].

5 Subband Recombination

Our earlier work has demonstrated the performance gains that can be obtained in speaker recognition by applying subband processing, together with hidden Markov models and multiple classifier recombination. The HMMs deliver log likelihood values, so that sum-rule fusion corresponds to taking products of likelihoods. Under assumptions of conditional independence and equal priors, this strategy is optimal (e.g. [18]). Using this rule, the identified speaker, i, is that for whom:

$$i = \arg\max_s [y^s] = \arg\max_s \sum_{n=1}^{N} \log L_n^s \qquad (1)$$

where N is the number of classifiers (subbands), $1 \leq s \leq S = 60$, L_n^s is the likelihood that classifier n and model speaker s produced the observed data sequence, and y^s is the recombined (final) score for speaker s from the set of S speakers.

The formulation in equation (1) is linear with constant (unity) weights. However, according to [5] in their work on subband speech recognition: "... it is often argued that the recombination mechanism should be nonlinear" (p. 427). Also, the assumption of conditional independence is unsatisfactory. Accordingly, [5] used a multilayer perceptron (MLP) trained to estimate posterior probabilities of speech units (HMM states, phones, syllables or words) given the log-likelihoods of all subbands and all speech units. It is also intuitively-attractive for a recombination scheme to have variable weights [17] and the MLP offers this.

Hence, MLPs have been used for the ANN recombination in this work (see [4] for relevant background). We apply a single 'local' MLP for which the structure has N inputs and only a single output. It is trained on outputs from all 60 speaker subsystems (as in Fig. 1). During test, output from each speaker subsystem is passed in turn to the MLP, and the identified speaker is that producing the largest output activation.

Each MLP was trained 10 times from different initial points in the search space, with the initial weights drawn from a zero-mean, unit-variance isotropic Gaussian distribution. The single output had a logistic activation function. For the results reported here, all MLPs had a single hidden layer of 5 tanh nodes. (It was found that using either 10 or 15 hidden nodes did not significantly affect the results.) Training minimised the cross-entropy error function using a conjugate-gradient algorithm. Outputs were trained to 0 or 1, with the latter indicating that the MLP classified the utterance as belonging to the speaker model. A weight decay scheme (with $\alpha = 0.2$) was used to prevent over-training. The order of the training data was randomised to avoid bias in the learning (in terms of all the positive examples being presented in a single block). Training used noise-free speech data only.

To assess the relative importance of the different subbands we also applied a Bayesian MLP with automatic relevance determination (ARD). ARD was proposed by [14] as a method for feature selection in neural networks. Weights connected to irrelevant inputs are automatically set to small values by assigning large ARD parameters.

Input data were scaled to be in the unit-interval in each input axis. This was to make the weight initialisation easier (as above) and also to avoid slow convergence of the weights in the presence of highly imbalanced data (less than 2% of the examples were positive). Without this scaling it was found that the weights could not converge in the number of iterations allowed.

6 Results

The results for the sum of log likelihood and MLP are shown in Figure 2 for data sets incorporating narrowband noise at the two different centre frequencies. The performance for the wideband system for comparison. The sum of log likelihood and MLP are plotted as functions of the number of subbands. Error bars are shown for the MLP, as a measure of the variability of the results starting from different random initial weight settings. Although not shown, the results for the Bayesian MLP with ARD were almost identical to those for the standard MLP.

In each case, we see the general trend for increased performance with increasing number of subbands. There are, however, some substantial dips in performance for particular combinations of spoken digit, number of subbands, noise centre frequency and fusion technique. The most obvious case is for 4 subbands, word *seven* and 987 Hz noise. This is most likely a result of the specific way that the filter profiles overlap in frequency, relative to the noise frequency and the important spectral components which differentiate speakers. In particular, however, there is a very significant improvement in performance over the wideband system. In each case performance close, or equal, to 100% is achieved for certain combinations of subbands.

In general, the sum of log likelihood fusion generally does about as well as the average MLP although, of course, the best MLP is always better than the average. Sometimes, the sum of log likelihoods does significantly better, for instance, achieving 100% on *seven* with 2500 Hz noise whereas the average MLP only achieves just over 90%. This suggests that the conditional independence assumption in equation (1) is reasonable. We can therefore assume that the subbands are approximately conditionally independent and therefore provide different information in making the overall identification decision. This is further demonstrated by the ARD parameters shown in Figure 3. These are for utterance *seven*. These do not show significant variation

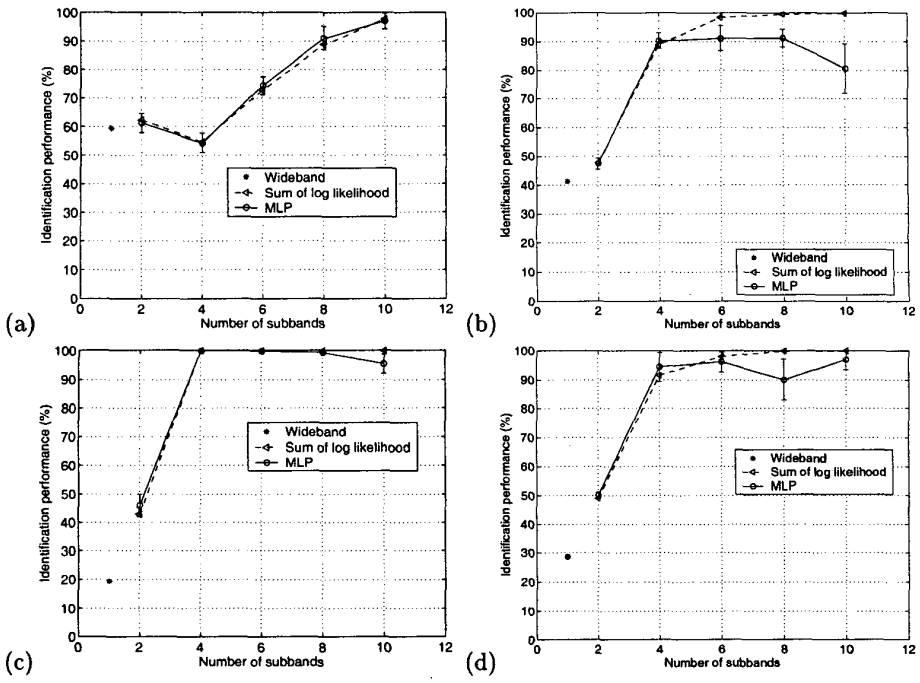

Figure 2: Results as a function of the number of subbands for noisy test utterances *seven* and *nine*. The plots show (a) utterance *seven* with narrowband noise centred at 987 Hz, (b) utterance *seven*, 2500 Hz, (c) utterance *nine*, 987 Hz, and (d) utterance *nine*, 2500 Hz.

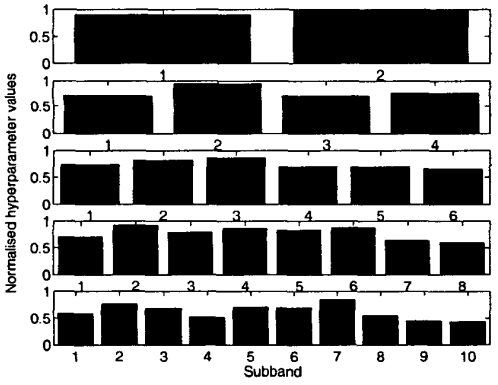

Figure 3: Mean normalised automatic relevance determination parameters for utterance *seven*.

over the subbands indicating that all the subbands are approximately equally relevant. This supports the notion that they are conditionally independent.

However, we must be careful in interpreting the ARD results. In [19] an extensive empirical investigation was made of ARD for feature selection. It was found that ARD is only effective in networks having many hidden units and where there are many irrelevant inputs. The latter point is likely given that the MLPs were trained on clean speech and therefore no subbands were corrupted by the narrowband noise (this only occurred in the testing phase). However, we also used MLPs with only 5 hidden units. Therefore there is a possibility that the lack of any variation in the ARD parameters may simply be a result of not enough hidden units to model the redundancies in the subbands. However, given that all the other evidence points towards the subbands all being relevant we believe this to be unlikely.

7 Conclusions and Future Work

In this paper, we have extended our earlier work on subband speaker recognition using multiple classifier techniques. We have studied speaker identification in two narrowband noise conditions for spoken digits *seven* and *nine* for 60 speakers. Results show that the sum of log likelihoods generally does better than the (average) MLP fusion, which we interpret to indicate that the assumption of conditional independence between subbands is reasonable. This interpretation is supported by our automatic relevance determination (ARD) anal-

ysis. Our priorities for future work are to include other fusion techniques in our performance comparisons, to explore other kinds of noise contamination, to attempt to understand the difference in the pattern of results for the two different spoken digits studied here, and beyond that to study all ten spoken digits in the database.

References

[1] B. S. Atal. 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, 1974.

[2] L. Besacier and J.-F. Bonastre. 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)*, pages 195–202, Crans-Montana, Switzerland, 1997.

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

[4] C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, UK, 1995.

[5] H. Bourlard and S. Dupont. 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, pages 426–429, Philadelphia, PA, 1996.

[6] J. P. Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997.

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

[8] G. Doddington. Speaker recognition – identifying people by their voices. *Proceedings of the IEEE*, 73(11):1651–1664, 1985.

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

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

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

[12] S. Furui. Recent advances in speaker recognition. *Pattern Recognition Letters*, 18:859–872, 1997.

[13] J. E. Higgins, R. I. Damper, and C. J. Harris. A multi-spectral data-fusion approach to speaker recognition. In *Proceedings of 2nd International Conference on Information Fusion, Fusion 99*, volume II, pages 1136–1143, Sunnyvale, CA, 1999.

[14] D. J. C. MacKay. The evidence framework applied to classification problems. *Neural Computation*, 4(3):415–447, 1992.

[15] J. D. Markel and A. H. Gray. *Linear Prediction of Speech.* Springer-Verlag, Berlin, Germany, 1976.

[16] A. Morris, A. Hagen, and H. Bourlard. 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, pages 599–602, Budapest, Hungary, 1999.

[17] S. Okawa, T. Nakajima, and K. Shirai. A recombination strategy for multi-band speech recognition based on mutual information criterion. In *Proceedings of 6th European Conference on Speech Communication and Technology, Eurospeech'99*, volume 2, pages 603–606, Budapest, Hungary, 1999.

[18] M. Pavel and H. Hermansky. Information fusion by humans and machines. In *Proceedings of First European Conference on Signal Analysis and Prediction*, pages 350–353, Prague, Czech republic, 1997.

[19] W. D. Penny and S. J. Roberts. Bayesian neural networks for classification: How useful is the evidence framework? *Neural Networks*, 12:877–892, 1999.

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

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

[22] P. Sivakumaran, A. M. Ariyaeeinia, and J. A. Hewitt. Sub-band speaker verification using dynamic recombination weights. In *Proceedings of Fifth International Conference on Spoken Language Processing, ICSLP'98*, Sydney, Australia, 1998. Paper 1055 on CD-ROM.

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

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

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