Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition Letters 24 (2003) 2167–2173

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Improving speaker identification in noise by subband processing and decision fusion

R.I. Damper *, J.E. Higgins

Image, Speech and Intelligent Systems (ISIS) Research Group, Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK

Abstract

We investigate speaker identification in narrowband noise using subband processing. The output of each subband is used to train and test individual hidden Markov models (HMMs), each making a preliminary decision on speaker identity. Subsequently, these are combined to produce a final decision. For sufficient numbers of filters, subband processing outperforms traditional wideband techniques by an enormous margin.

Keywords: Speaker recognition; Hidden markov model; Subband processing; Decision fusion

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 N bandpass filters to give a set of N 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 et al., 2001). We believe that this is likely to be a major advantage of subband processing: we can expect to produce much better and more robust models for each of the N subband signals from the (always limited) example data than the single model produced from the wideband signal.

The subband approach has also become popular in recent years in *speech* recognition (Bourlard and Dupont, 1996; Tibrewala and Hermansky, 1997; Morris et al., 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 test utterances decoded using hidden Markov models (HMMs) trained on clean speech.

* Corresponding author. Fax: +44-023-80-594-577.
E-mail addresses: rid@ecs.soton.ac.uk (R.I. Damper), jeh97r@ecs.soton.ac.uk (J.E. Higgins).

0167-8655/03/$ - see front matter
doi:10.1016/S0167-8655(03)00082-5

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 fusion rule used for the final identification. Section 6 gives results and Section 7 concludes.

2. Subband processing

Fig. 1 shows a schematic of the subband system used here. The bandpass filters are sixth-order Butterworth with infinite impulse response, implemented using the MATLAB function `butter`. As well as the filter order, this function takes the −3 dB points as arguments. The filterbank was arranged to have the −3 dB crossover points between adjacent filters equally spaced on the mel scale. This is a psychophysically motivated frequency scale intended to reflect the frequency selectivity of human hearing (Stevens and Volkmann, 1940; Warren, 1999).

Fig. 2 shows the filter profiles for two representative cases: $N = 4$ and $N = 16$. Note that the first filter (i.e., that with lowest upper −3 dB frequency) is designed to have a low-pass rather than a bandpass characteristic. Also, as a consequence of the equal spacing in mel frequency, the last filter (i.e., that with highest upper −3 dB frequency) has

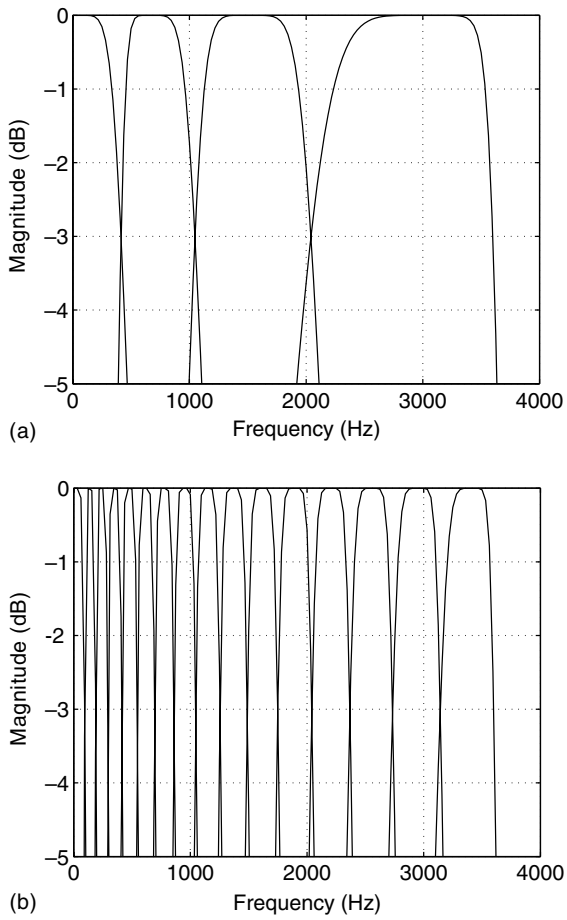

Fig. 2. Filter profiles for the two representative cases of: (a) $N = 4$ and (b) $N = 16$. Filters are sixth-order Butterworth with −3 dB crossover frequencies equally spaced in mel frequency over the range 0–4 kHz.

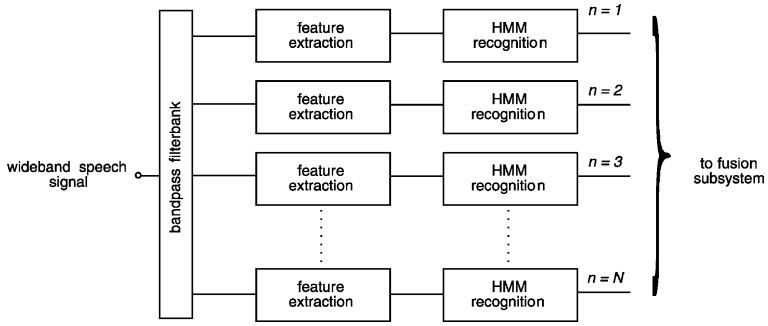

Fig. 1. Schematic diagram of the subband processing system. Each subband (filter) has its own recognition subsystem, whose output is fed to a fusion algorithm which makes the final, overall decision about speaker identity.

its upper −3 dB frequency some way below 4 kHz. The filter bandwidths increase with frequency in a way which reflects the frequency resolution of human hearing (Moore and Glasberg, 1986).

Filtering was performed in the time domain by direct calculation from the difference (recurrence) equation. Feature extraction was performed on each subband, and the resulting sequences of feature vectors passed on to each subband's recognition algorithm for a preliminary decision on speaker identity. Note that this preliminary decision is 'soft' in the sense that information about the match between the input speech and the models for *all* speakers is retained and fed forward. Thereafter, the outputs from each separate recogniser were fused to produce an overall decision as to the identity of the speaker—a form of decision fusion (Dasarathy, 1993).

Successful speaker identification is critically dependent on obtaining good speaker models from the training data. 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 et al., 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 et al., 2001). This in our view is another potentially very strong advantage of subband processing.

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 60 (46 male and 14 female) native English speakers saying the digits *one* to *nine*, *zero*, *nought* and *oh* 25 times each. Here, we present results for words *seven* and *nine* only both to limit simulation times and because we believe this should still yield representative and meaningful results. Recordings were made in five sessions spaced over three months, to capture the variation in speaker's voices over time.

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) 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. It was found to be relatively easy to achieve 100% accuracy on 'clean' or noise-free speech from the Millar database, for all systems tested. Accordingly, we consider only the noise-added situation here.

4. Data modelling

Initially, pseudo-cepstral features are extracted on a frame-by-frame basis. Cepstral analysis is motivated by, and designed for, problems centred on voiced speech (Deller et al., 1993), but 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 words *seven* and *nine* spoken by the

different speakers. For this, we use the popular HMMs. HMMs are powerful statistical models of sequential data that have been used extensively for many speech applications (Rabiner, 1989; Knill and Young, 1997). They assume an underlying (hidden) stochastic process that can only be observed through another 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, trained and tested using the HTK software of Young et al. (2000). (Training used the forward-backward algorithm and decoding used the token-passing algorithm.)

5. Decision fusion rule

In this work, we have used a simple fusion rule which avoids any need to estimate weighting parameters from training data. (See Higgins et al., 2001a,b, 2002 for related work using trainable fusion.) Kittler et al. (1998) developed a common theoretical framework for such simple rules of combination which use distinct pattern representations (as here). 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 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:

$$i = \arg\max_{s} \sum_{n=1}^{N} \log p(\boldsymbol{x} \,|\, \omega(n,s)) \quad 1 \leqslant s \leqslant S(=60)$$

where $p(\boldsymbol{x} \,|\, \omega(n,s))$ is the probability that model $\omega(n,s)$ for subband n and model speaker s produced the observed data sequence $\boldsymbol{x}$. Because of the use of logarithms, this is effectively the product rule but other rules tried (e.g., sum, max) worked no better.

6. Results

To evaluate the effectiveness of the subband system, all 60 speakers in the database were tested speaking words *seven* and *nine*.

6.1. Fixed noise condition

Following Besacier and Bonastre (2000), we use Gaussian noise 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.

For the results in this subsection, the number of HMM states (including start and end states) was varied from 5 to 10 (although a wider range was used to optimise the wideband results). Apart from self-loops (staying in the same state), only left-to-right transitions were allowed. The frames of speech data were vector quantised and each HMM had its own binary tree codebook of size 32. The number of subbands was varied from 2 to 22.

Results are depicted in Fig. 3 which shows a general increase in speaker identification performance with the number of subbands. For comparison, the best results for a traditional, wideband system (for which a wider range of states was studied in an attempt to optimise performance) were 54.7% for *seven* with 9 HMM states and 19.8% for *nine* with 3 states. Note that the chance level is 1/60, or 1.67%.

The number of HMM states does not appear to have a major impact on the results. Accordingly, to make the variation of performance with the number of subbands clearer, we present results in Fig. 4 for a fixed number of HMM states. For word *seven*, the number of HMM states is fixed at 6. For word *nine*, the number of HMM states is fixed at 5. These are effectively 2D slices through the 3D plots of Fig. 3(a) and (b).

It is abundantly clear that the subband/fusion systems outperform the wideband systems by an

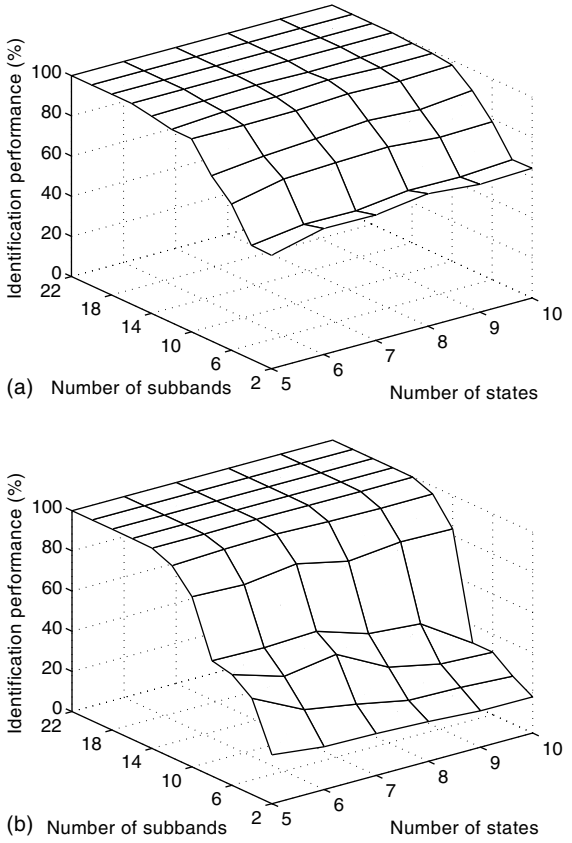

Fig. 3. Percentage correct identification for 60 speakers saying the words: (a) *seven* and (b) *nine* as the number of subbands is varied between 2 and 22 and the number of HMM states for each subband recogniser is varied from 5 to 10. Noise condition: 987 Hz centre frequency, 365 Hz bandwidth.

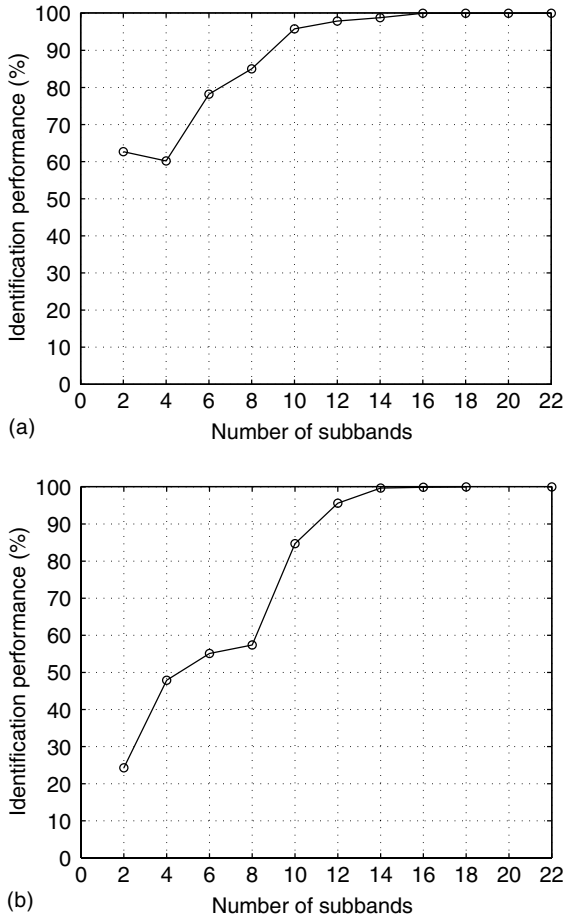

Fig. 4. Percentage correct identification for 60 speakers saying the words: (a) *seven* and (b) *nine* as the number of subbands is varied between 2 and 22. For word *seven*, the number of HMM states is fixed at 6. For word *nine*, the number of HMM states is fixed at 5. Noise condition: 987 Hz centre frequency, 365 Hz bandwidth.

enormous degree, easily achieving 100% correct only provided a sufficient number of subbands ($\geqslant \sim 14$) is used. In particular, the increase from 19.8% speakers correct to 100% for the word *nine* dramatically illustrates the improvements which can be gained.

6.2. Varying noise bandwidth

Subsequently, to verify the generality of our results for other noise conditions, the bandwidth of the added noise was varied, keeping the centre frequency fixed at 987 Hz. The number of HMM states was fixed at 6 for word *seven* and 5 for word *nine*.

Fig. 5 shows the results. As before, the general trend is for performance to increase dramatically with the number of subbands, showing that this result is not an artifact of a particular choice of noise bandwidth. Looking at the results in more detail, those for word *seven* (Fig. 5(a)) are entirely as expected. That is, as the noise bandwidth is increased, the performance deteriorates. However, the pattern of results for word *nine* (Fig. 5(b)) is unexpected, showing the opposite trend. The reason for this is unknown. Yet still, for both words,

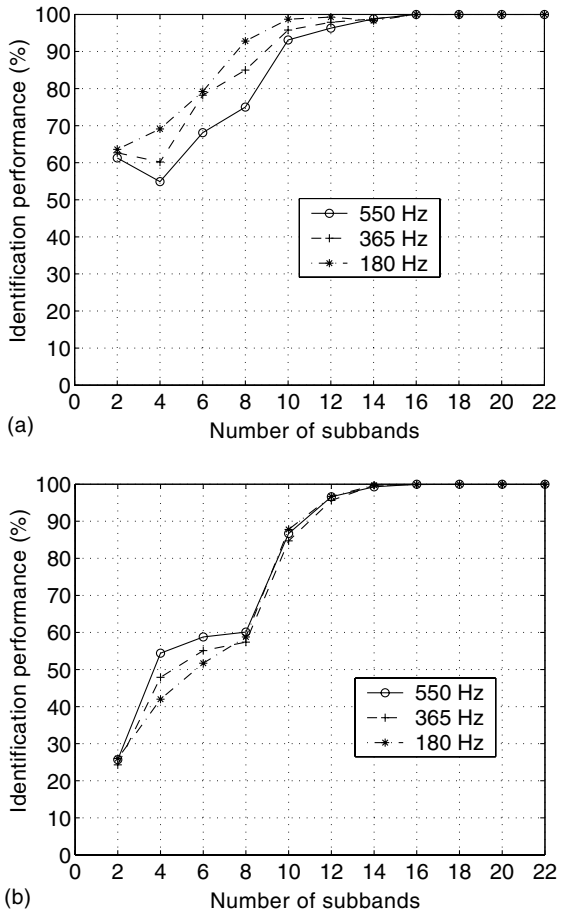

Fig. 5. Effect of varying the noise bandwidth on performance for words: (a) *seven* and (b) *nine*. The noise centre frequency is fixed at 987 Hz. For word *seven*, the number of HMM states is fixed at 6. For word *nine*, the number of HMM states is fixed at 5.

100% correct performance is achieved for a sufficient number of subbands.

6.3. Varying noise centre frequency

We have also done some simulations in which the centre frequency of the added noise was varied from 500 to 3000 Hz in steps of 500 Hz, keeping the bandwidth fixed at 365 Hz. The number of HMM states was fixed at 6 for word *seven* and 5 for word *nine*. As we do not have a full set of results (each simulation takes some considerable time), in a form comparable to those above, we do not present them in detail here. Generally, the lowest identification rate was seen for noise centre frequencies of 1000–1500 Hz for low numbers of subbands. However, the dramatic performance increase with the number of subbands seen in earlier results was confirmed. In particular, 100% performance was robustly achieved for adequate numbers of subbands, showing that this result is not an artifact of a particular choice of noise centre frequency.

7. Conclusions and discussion

Results presented in this paper demonstrate that subband processing used with decision fusion offers enormously improved speaker identification performance, compared to a wideband system, in the face of narrowband noise. For the subband system and two spoken digits tested here, *seven* and *nine*, 100% correct speaker identification was easily and robustly achieved for sufficient numbers of subbands, unlike the traditional, wideband system which produced a best score as low as 19.8% for *nine*, in spite of our efforts to optimise the HMM model structure for the wideband case. Performance improvements were relatively insensitive to other experimental variables, such as noise centre frequency, noise bandwidth and number of HMM states. Hence, we are confident that the improvements seen are real, and not a consequence of a fortuitous choice of conditions.

Since it is apparent that the results vary somewhat for the two different words, future work will need to study all of the spoken digits in our database. However, this variation is much greater for the best wideband system than for the subband/fusion systems, offering the distinct promise that high performance in the latter case will depend largely if not solely on having appropriate numbers of subbands, irrespective of the vocabulary used.

Future work will also study the effect of non-stationary noise (e.g., where the noise is switched on and off alternately, and/or its centre frequency either drifts or is switched suddenly). Here, there is reason to believe that the subband system should display good immunity to non-stationary noise.

Since the speaker models (based on linear prediction and HMMs) are obtained from clean speech, and since the decision fusion rule is not trained from data, the subband system apparently achieves its excellent performance in the present study without exploiting any special characteristics of the added noise, such as the fact that it is stationary. If this is correct, performance in non-stationary noise should also be good but this prediction remains to be confirmed.

References

Atal, B.S., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Amer. 55 (6), 1304–1312.

Besacier, L., Bonastre, J.-F., 1997. Subband approach for automatic speaker recognition: Optimal division of the frequency domain. In: J. Bigün, G. Chollet, G. Borgefors (Eds.), Proc. 1st Internat. Conf. on Audio- and Visual-Based Biometric Person Authentication (AVBPA), Crans-Montana, Switzerland, pp. 195–202.

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

Bourlard, H., Dupont, S., 1996. A new ASR approach based on independent processing and recombination of partial frequency bands. In: Proc. Fourth Internat. Conf. on Spoken Language Process., ICSLP'96, Philadelphia, PA, Vol. 1, pp. 426–429.

Dasarathy, B.V., 1993. Decision Fusion. IEEE Computer Society Press, Silver Spring MD.

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

Finan, R.A., Damper, R.I., Sapeluk, A.T., 2001. Improved data modelling for text-dependent speaker recognition using sub-band processing. Internat. J. Speech Technol. 4 (1), 45–62.

Furui, S., 1981. Cepstral analysis techniques for automatic speaker verification. IEEE Trans. on Acoust., Speech Signal Process. ASSP 29 (2), 254–272.

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

Higgins, J.E., Damper, R.I., Dodd, T.J., 2001a. Information fusion for subband-HMM speaker recognition. In: Proc. INNS-IEEE Internat. Jnt. Conf. on Neural Networks, IJCNN'01, Washington, DC, Vol. 2, pp. 1504–1509.

Higgins, J.E., Dodd, T.J., Damper, R.I., 2001b. Application of multiple classifier techniques to subband speaker identification with an HMM/ANN system. In: Kittler, J., Roli, F. (Eds.), Multiple Classifier Systems, Second International Workshop, MCS 2001. Springer, Berlin, Germany, pp. 369–377.

Higgins, J.E., Damper, R.I., Dodd, T.J., 2002. Improving speaker identification by trainable data fusion and subband processing techniques. In: Proc. Third IEEE Workshop on Automat. Identification Advanced Technologies, AutoID-02, Tarrytown, NY, pp. 109–114.

Kittler, J., Hatef, M., Duin, R.P.W., Matas, J., 1998. On combining classifiers. IEEE Trans. on Pattern Anal. Machine Intell. 20 (3), 226–239.

Knill, K., Young, S., 1997. Hidden Markov models in speech and language processing. In: Young, S., Bloothooft, G. (Eds.), Corpus-Based Methods in Language and Speech Processing. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 27–68.

Moore, B.C.J., Glasberg, B.R., 1986. The role of frequency selectivity in the perception of loudness, pitch and time. In: Moore, B.C.J. (Ed.), Frequency Selectivity in Hearing. Academic Press, London, UK, pp. 251–308.

Morris, A., Hagen, A., Bourlard, H., 1999. The full-combination subbands approach to noise robust HMM/ANN-based ASR. In: Proc. 6th Europ. Conf. on Speech Comm. Technol., Eurospeech'99, Budapest, Hungary, Vol. 2, pp. 599–602.

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

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

Stevens, S.S., Volkmann, J., 1940. The relation of pitch to frequency: A revised scale. Amer. J. Psychol. 55 (3), 329–353.

Tibrewala, S., Hermansky, H., 1997. Sub-band based recognition of noisy speech. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Process., ICASSP'97, Munich, Germany, Vol. 2, pp. 1255–1258.

Warren, R.M., 1999. Auditory Perception: A New Analysis and Synthesis. Cambridge University Press, Cambridge, UK.

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