Proceedings of the Institute of Acoustics

NOISE MASKING IN THE MFCC DOMAIN FOR THE RECOGNITION OF SPEECH IN BACKGROUND NOISE

B. A. Mellor and A. P. Varga

Speech Research Unit, Defence Research Agency, St Andrews Road, Malvern.

1. INTRODUCTION

A previous paper [Varga and Ponting, 1989] presented results for various noise compensation algorithms, concluding that the technique of noise masking [Klatt, 1976] could give a particularly robust speech recognition performance down to signal–to–noise ratios as poor as 3dB (for the case of stationary pink noise). Varga and Ponting formulated the noise masking algorithm for a hidden Markov model (HMM) based recogniser with a filter bank front end [Holmes, 1980]. Their filter bank generated observation vectors consisting of the log energy in critical band spaced band–pass filters. It has been widely noted however, that under various circumstances improvements in recognition performance can be achieved by transforming such filter bank data; for example, with a cosine transform (leading to mel scaled frequency cepstral coefficients) [Russell, 1992], or with a transform based on a linear discriminate analysis (e.g IMELDA [Hunt, 1989]). It is therefore of interest to examine the use of the noise masking algorithm in such a transformed domain. This paper describes an approach to the use of noise masking in transformed domains and it reports experimental results comparing the performance of a recogniser working on filter bank observation data and transformed observation data, both with and without noise masking.

2. NOISE COMPENSATION AND NOISE MASKING

Noise compensation techniques work by modifying the recognition process to take account of background noise which is inextricably embedded in an input speech signal; this contrasts with pre–processing approaches in which attempts are made to "clean up" the signal before recognition. Noise masking is one such compensation technique. The masking algorithm was developed in detail in [Varga and Ponting, 1989]. In summary, a tracking estimate of the background noise is maintained; each band mean of the active speech model is examined in turn; if the value of the noise estimate for that channel is greater than the model mean then that mean is replaced (masked) by the noise estimate. The input speech frame is similarly masked, i.e. if the noise estimate for a band is greater than the observation then the observation is replaced (masked) by the noise estimate. The masking process applied in the filter bank domain (modifying observation energies

and model means on the basis of knowledge about the noise and the speech) acts on the observation probability evaluation process in a way that improves recognition robustness to background noise.

3. NOISE MASKING IN TRANSFORM DOMAINS

The masking algorithm works by modifying model and observation spectra using knowledge of the noise. However, in general, in a transform domain any component within the observation vector is a weighted combination of the components of the observation vector in the original domain (in the case of the filter bank front end used here, mel scaled frequency cepstral coefficients (MFCCs) are obtained by applying a cosine transformation to the log energies output by the filter bank). Therefore the non–linear masking operation can not be applied in such transform domains. So the approach developed here is to carry out the masking in the spectral domain and subsequently to carry out the transformation; observation probabilities may then be evaluated in the transform domain on data which has been masked in the spectral domain. The initial experiments reported below examine whether this technique offers the same degree of noise robustness in the transformed domain as in the spectral (or filter bank domain).

In detail the masking operation is carried out as follows. Masking the observation vectors is straight forward, the spectral domain data is generated before any transformation can be carried out; it is therefore simple to both maintain a tracking noise estimate and to apply the masking before the transformation. Masking the models is a little more complex. The approach is to maintain a spectral domain version of the models on which to carry out the masking, the masked model means can then be transformed for use in evaluation of the observation probabilities. This process can be carried out with reasonable efficiency by only re–masking and transforming the model means on a demand basis and then only when the noise estimate has been updated. Pre–calculation can also be used, for instance in the case of known noise, or mutli–state non–tracking noise masks (c.f. those used with decomposition [Varga and Moore, 1990]).

4. EXPERIMENTAL SETUP

4.1 Experimental data

The speech data used were isolated digits extracted from the NATO RSG–10 isolated digit database [Vonusa et al., 1982]. It consists of five continuous tables each of 100 digits spoken in isolation. One table was used to train the models, one table was used for parameter optimisation and the remaining three tables for tests. Pink noise data was taken from the NATO RSG–10 noise database [Steeneken and Geurtsen, 1988]. The signals were sampled at 20Khz.

The speech and the noise signals were recorded separately and added together digitally at seven different signal–to–noise ratios: 21, 15, 9, 3, –3, –9 and –15 dB; clean speech with no added noise was used for training and the base line test. The signal–to–noise ratio was calculated on the basis of signal level measurements made using the British Telecom SV6 speech voltmeter. The SV6 conforms to the CCITT standard [CCITT, 1984] for speech level measurement.

4.2 The recogniser

The recogniser is a one pass fully continuous system with a 27 band filter bank front end. Ten state left–right whole word speaker dependent hidden Markov models were used, the output distribution for each state was multi–variate single mode Gaussian with diagonal covariance matrix. The speech models were trained under the noise free condition on ten repetition of each digit. A simple and computationally cheap noise tracking algorithm was used to generate an automatic estimate of the noise for use in the recogniser. The estimate was calculated from the non–speech periods in the data. A single–state background noise model was used to "recognise" the non–speech periods, the means of this model were simply set to be the current noise estimate and the model used had no duration penalty

5. THE EFFECT OF VARYING THE NUMBER COEFFICIENTS

Recognition performance has been found to be sensitive to the number of MFCC coefficients used [Russell, 1992]; so a small study was carried out examining the performance for various numbers of coefficients; the results are summarised in figure 1. This experiment was carried out on the optimization data set. The noise masking algorithm was used together with a noise tracking background noise model. It can be seen that overall eight MFCC coefficient gave the best performance; therefore eight MFCCs were used in the following tests.

NOISE MASKING IN THE MFCC DOMAIN

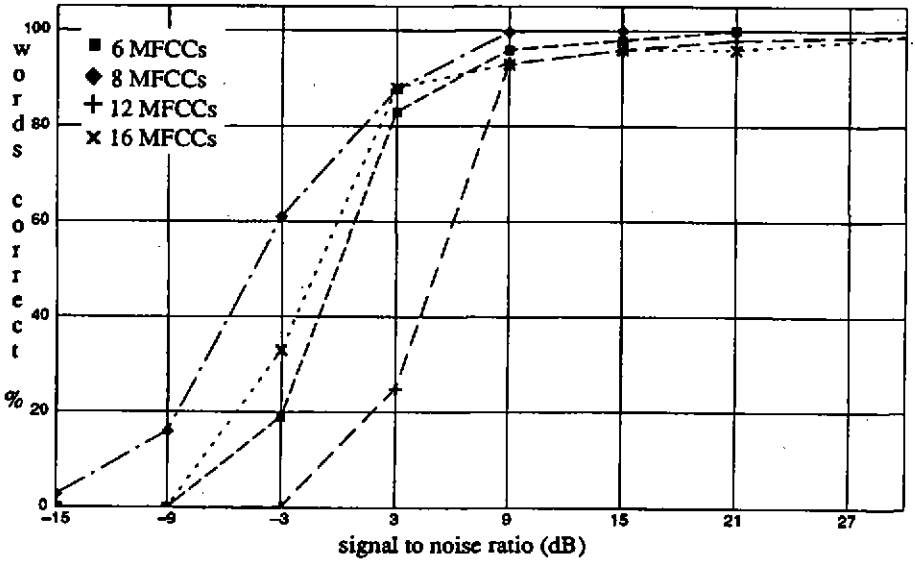

Figure 1. Variation of percentage words correct with number of cepstral parameters for the noise masking algorithm operating in the MFCC domain.

6. EXAMINATION OF VARIOUS ASPECTS OF THE RECOGNISER IMPLEMENTATION

The use of a background noise model is essential to for "good" recognition performance. This can be seen from the results comparison shown in figure 2 for the MFCC based recogniser. The worst performance was obtained for the case where a the background noise model was based on the low level background for the clean speech (i.e. a very poor model of the background noise for the case of added pink noise). The simple addition of an HMM that modelled closely the background noise provided a significant improvement in performance. However, the addition of noise masking gave a further improvement in performance, equivalent to a 12dB improvement in SNR.

NOISE MASKING IN THE MFCC DOMAIN

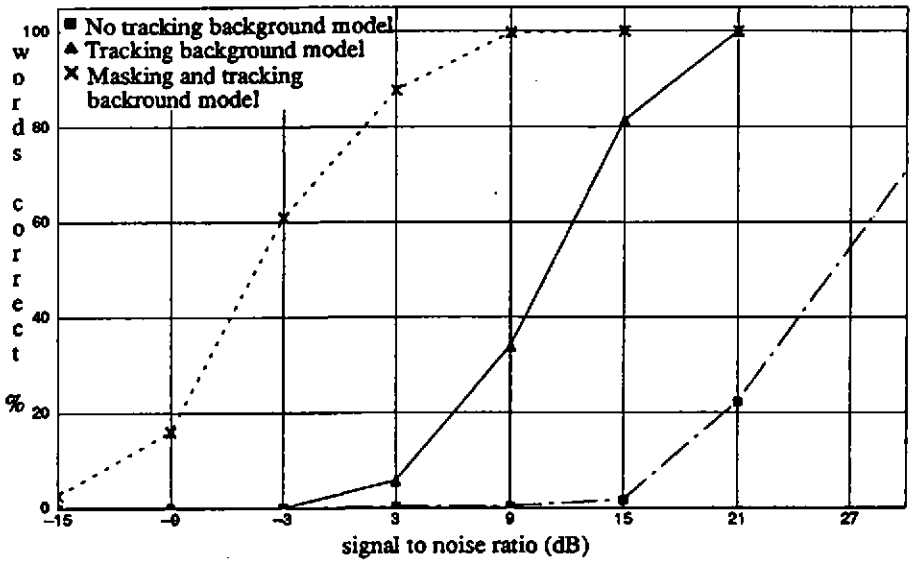

Figure 2. Comparative effect on performance of the use of a tracking background noise model and noise masking.

6. COMPARISON OF FILTER BANK AND MFCC FRONT ENDS

The final experiment was to compare the performance of noise masking in the MFCC domain with the results reported in [Varga and Ponting, 1989] for noise masking in the filter bank log energy domain. For the purposes of verification the experiment carried out in [Varga and Ponting, 1989] was repeated with the newer version of the recognition software used for the main experimental work here. New model sets were re–estimated on the original clean speech data and the original test set was used with the noise mixed as before, the recognition was carried out using noise masking in the filter bank log energy space with a noise tracking background model. The results from the new version of the recogniser matched those reported in the earlier experiment. The comparison between the log energy domain representation and the eight MFCC representation is shown in figure 3. Also shown are the results for the recognition experiments without noise masking in the MFCC and filter bank domains, both with noise tracking background models. It can be seen that noise masking in the filter bank log energy domain provides robust speech recognition in noise, giving good performance down to 3dB signal–to–noise ratio. The performance of the MFCC domain noise masking tracks that of the filter bank down to 9dB SNR, however below this the MFCCs front end gives a words correct performance approximately 10% down

on the filter bank front end. It is interesting to note that when noise masking is not used MFCC domain observation vectors give a performance enhancement over the filter bank log energy domain observations.

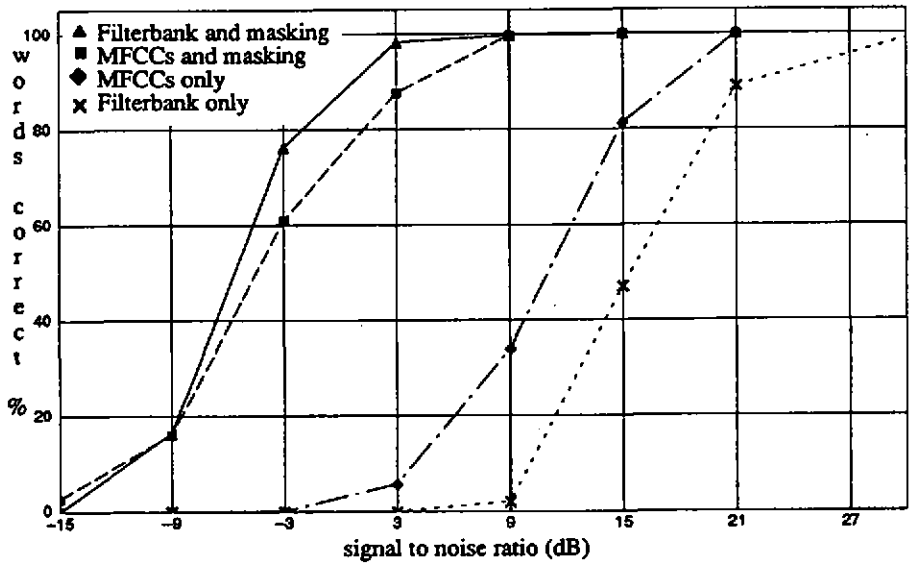

fig 3. Comparison of Noise Masking in MFCC Space With Log Energy Domain

CONCLUSION

It can be seen in this experiment that the use of the noise masking technique can provide a good degree of noise robustness in both the filter bank domain and the mel scaled frequency cepstral (MFCC) domain. The masking technique is computationally cheaper than the more comprehensive decomposition technique [Varga and Moore, 1990], however it provides poorer recognition performance at very low signal–to–noise ratios (i.e. below 3dB). Noise masking may therefore offer a cheaper alternative to decomposition for higher signal–to–noise ratios.

When used in a transformed domain masking offers a theoretically tractable alternative to the full decomposition algorithm which has not been extended for operation in a transform domain (though recently Gales and Young [1992] have suggested an approach similar in philosophy to that developed in this paper). So the technique may be of use where such transformations are required for other performance reasons. The relatively poorer performance of the MFCC front end (c.f. the filter bank) is still the subject of investigation, it is hoped that it will prove possible to obtain the same performance from both front ends.

8. REFERENCES

A. P. Varga and K. M. Ponting, 1989. *Control Experiments on Noise Compensation in Hidden Markov Model Based Continuous Word Recognisers.* ESCA Proc. EUROSPEECH, 1989.

J.N. Holmes, 1980. *The JSRU channel vocoder,* IEE Proc. F, vol 127, no. 1, pp53–60, Feb 1980.

D.H. Klatt, 1976. *A Digital Filterbank For Spectral Matching.* IEEE Proc. ICASSP, 1979.

M.J. Russell, 1992. *The Development Of The Speaker Independent ARM Continuous Speech Recognition System.* RSRE Memo # 4473, January 1992.

M. J. Hunt and C Lefebvre, 1989. *Distance measures for speech recognition.* National Aeronautical Establishement, Canada, NAE–AN–57, NRR no. 30144, March 1989.

The International Telegraph and Telephone Consultative Committee, CCITT, 1984. *Objective Measurement of Active Speech Level.* Suppl #8, Red Book, vol *V, VIII*th Plenary Assembly, pp242–247, Malaga, Oct. 1984.

R. S. Vonusa, J. T. Nelson, S. E. Smith, and J. G. Parker, 1982 . *NATO AC/243 (Panel 111 RSG–10) Language database,* Proc. US National Bureau of Standards workshop on "Standards for Speech I/O Technology" pp.223—228, 1982.

H. J. M. Steeneken and F. W. M. Geurtsen, 1988. *Description of the RSG10 noise database,* TNO Institute for Perception, report no IZF 1988–3, 1988.

A. P. Varga and R. K. Moore, 1990. *Hidden Markov Model Decomposition of Speech and Noise.* Proc. IEEE Int. Conf. Acoust. Speech & Signal Proc. ICASSP'90, pp845–848, 1990.

M. J. F. Gales and S. Young, 1992. *An improved approach to the hidden Markov model decomposition of Speech and Noise.* Proc. IEEE Int. Conf. Acoust. Speech & Signal Proc. ICASSP'92, ppI–233 – I–236, 1992.