TOWARDS A BETTER AUDITORY REPRESENTATION FOR SPEECH RECOGNITION

S W Beet and L Baghai-Ravary

s.beet@shef.ac.uk, l.baghai-ravary@shef.ac.uk

Department of Electronic & Electrical Engineering
The University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK

ABSTRACT

This paper compares a number of different auditory power spectral density representations of speech signals in a phoneme recognition task. The numerical properties of the various representations are quite different even though they are calculated from the same intermediate representation. The results presented here clearly indicate that the degree of variability in results is large, even when it is ostensibly the same parameter which is being estimated. Thus, it is not merely 'what' is calculated, but 'how' its value is estimated, which ultimately may determine recognition performance.

Two similar but different sets of comparisons have been made to confirm that a significant difference does indeed exist. In both cases, the maximum entropy method of power spectrum estimation is significantly better than the others, even though both this and the maximum likelihood method are based on the same initial linear prediction analysis of the signal. The maximum likelihood method's performance is very nearly the same as that of the Blackman-Tukey method.

1. INTRODUCTION

Mainstream speech recognition systems, such as those based on hidden Markov models (HMMs) [1] and artificial neural networks (ANNs) [2], are unable to use full-scale auditory models due to a combination of inappropriate data characteristics, excessive volumes of output data, and high computational cost. Consequently, the closest they get is to make use of a pseudo-auditory frequency scale in mel-frequency cepstral coefficient (MFCC) or perceptual linear prediction (PLP) analysis methods [3]. However, many other computationally-efficient and HMM/ANN-compatible representations are made possible by modifying standard signal analysis methods.

The most computationally-efficient models of the peripheral auditory system are based on power-spectrum models of perceptual phenomena ranging from basilar membrane tuning curve variation to masking and dynamic range compression. While imperfect in many respects, such models are simple to implement and largely predictable in their behaviour.

The most powerful methods for analysing temporal sequences are based on estimates of the signal's autocorrelation function (ACF). This can be thought of as the inverse Fourier transform of the power spectral density (PSD). Thus any of these methods could be implemented in an auditory form by applying an appropriate transformation in the frequency (Fourier) domain before estimating the ACF from this modified PSD estimate. Nonetheless, this is not always the best approach. For example, speech is known to be well-suited to analysis in terms of autoregressive (AR) model parameters, and if the ACF is distorted prior to such analysis, the validity of the AR model becomes suspect. It is therefore often more appropriate to apply the auditory transformation at a later stage in the processing.

2. PSD ESTIMATES

There are many well-established methods for estimating the PSD of a signal [4]. Some of these are particularly well-suited to the analysis of speech signals. Others are not, and will not be discussed any further here. The three approaches investigated for this paper are briefly described below:

2.1. The Blackman-Tukey Method

One method for controlling the resolution of a periodogram, and hence suppressing pitch in both time and frequency domains, is to window the autocorrelation function before taking its inverse Fourier transform. This allows the frequency resolution to be reduced without reducing the frame length, and it also produces an unbiased PSD estimate. However, the window must have a non-negative Fourier transform for negative PSD estimates to be avoided. In the work reported here, the window was made equal to the autocorrelation function of a suitable prototype window, simultaneously ensuring that the created window has finite duration and non-negative Fourier transform at all frequencies.

2.2. The Maximum Likelihood Method

This is variously referred to as the minimum variance PSD estimate, the maximum likelihood method (MLM) or Capon's method. It is equivalent to the design of an FIR filter for each frequency where an estimate of the PSD is required, although it can be implemented more simply and efficiently using the method in [5]. This uses a standard linear prediction analysis, and after some simple modification of the coefficients, yields a complete PSD estimate via the discrete Fourier transform.

Each MLM filter has unity gain at the design frequency, but with minimal output power. In effect, the technique attempts to attenuate all but the frequency component of interest, and can be considered as a data-adaptive periodogram.

The order of the filters determines the maximum number of discrete frequency regions which can be attenuated, and is chosen according to the application. To resolve formant structure while suppressing pitch information, the filter order should be chosen to be slightly more than twice the maximum number of formants, as in linear prediction analysis. An order of 16 was used for the experiments described here.

Since the number of frequencies which can be completely attenuated by the FIR filters is limited, this method actually provides an upper bound on the true PSD. This property may be useful in some applications, but is not exploited here.

2.3. The Maximum Entropy Method

The power spectrum of an autoregressive (AR) process can be obtained by calculating the parameters of the AR model from the autocorrelation function of the signal. The maximum entropy method (MEM) PSD estimate is then obtained by multiplying the innovation power by the power transfer function of the implied recursive filter.

The same linear prediction analysis is used in finding the AR model's transfer function as was used for the MLM. In both cases, the auditory transformation was actually applied to the final PSD estimate, rather than being used to modify the ACF, as mentioned previously.

2.4. Auditory Transformation

The auditory transformation applied here consisted of a form of dynamic range compression and a non-linear scaling and smoothing of the power spectrum, in accordance with observed psychophysical phenomena.

2.4.1. Amplitude Scale

Power spectrum estimates are normally encoded on a logarithmic power scale. Auditory models, on the other hand, generally use an N^{th} root operator. The correct value of N is still a matter for debate, and is known to be frequency-dependent [6].

While 'physiological' or 'psychophysical correctness' would suggest that that frequency-dependence should be replicated here, there are sound mathematical reasons for quantising N to the nearest odd integer. In practice, this means that N will be independent of frequency.

The use of odd integer values for N is convenient because it is then possible to take the required root of any value, even if rounding errors in the PSD estimation algorithm have generated a negative result. In the data presented here, a value of $N = 5$ has been used, which is within 3dB of a normalised logarithmic scale over a dynamic range of 200:1, so any linear scaling of the input signal can be approximated as the addition of an offset to the values in the transformed power spectrum. This is an important (and desirable) feature in automatic speech recognition (ASR) systems since it gives a degree of immunity to changes in signal level.

2.4.2. Frequency Scale

The warping function used here was chosen to be the equivalent rectangular bandwidth (ERB) scale of Moore and Glasberg [7]. This was developed specifically to characterise frequency resolution in a power spectrum model of human perception, and as in the case of amplitude scaling, the use of this near-logarithmic scale provides a useful immunity to moderate scaling of formant frequencies due to variations in effective vocal tract length.

The auditory representations considered here have therefore been calculated on a warped frequency scale, and have been smoothed to exhibit an appropriate frequency resolution. This was done by forming a weighted sum of the PSD estimate for each of a set of equally spaced points on the constant ERB-rate scale. The weighting is triangular, with unit height and width of one ERB.

This form of weighting is like that traditionally used for mel-scale processing, and was used in preference to the one suggested by Moore and Glasberg purely for the sake of simplicity: the exact form of the weighting they suggest changes depending on the signal level and involves more complicated (and thus, time-consuming) calculations.

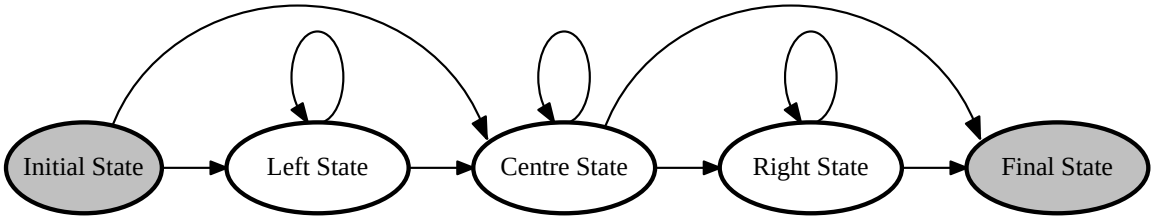

Figure 1: HMM structure used in phoneme recognition experiments.

3. RECOGNITION EXPERIMENTS

The task selected for the recognition experiments was that of speaker-independent phoneme recognition, using the TIMIT database [8]. There were 60 phonetic categories in the recogniser's vocabulary, and a simple 3-state hidden Markov model was used for each phoneme, as shown in figure 1. Two sets of experiments were conducted, one with the raw data augmented with delta coefficients, the other including delta-delta coefficients as well. A baseline was set by performing the same experiments with a standard (DFT-based) MFCC analysis. The reduction in error rate (expressed as a percentage of the MFCC error rate) is shown in figures 2 and 3 for the experiments without and with delta-delta coefficients, respectively.

In both cases, the auditory power spectrum estimates out-performed the MFCC analysis. In the case of the MEM PSD, this improvement was quite marked. The variation between the results with and without delta-delta coefficients was (on average) 5%, and this gives some indication of the likely variability in the results due to factors unrelated to the issues of interest here.

Following on from these reults, it was decided to assess the change in recognition performance of the various representations due to the use of the ERB frequency scale and smoothing. These results are shown in figure 4.

It is clear, given the likely variability in the results noted previously, that the advantage offered by auditory frequency scaling, and imposing the associated auditory frequency resolution, is nearly independent of the method used to calculate the PSD estimate (around 10% on average).

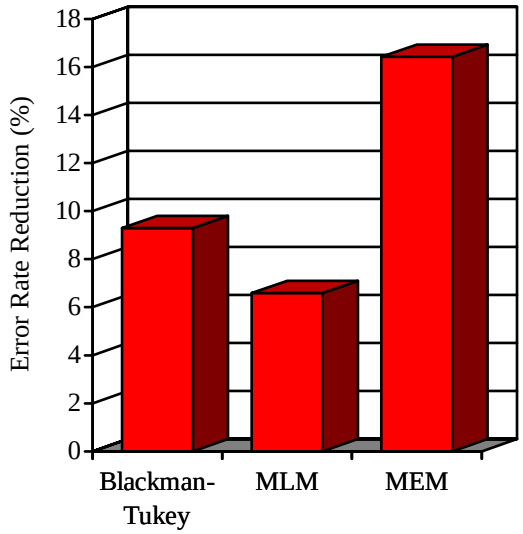

Figure 2: Relative reduction in error rate when replacing MFCCs by 'auditory' PSD estimates.

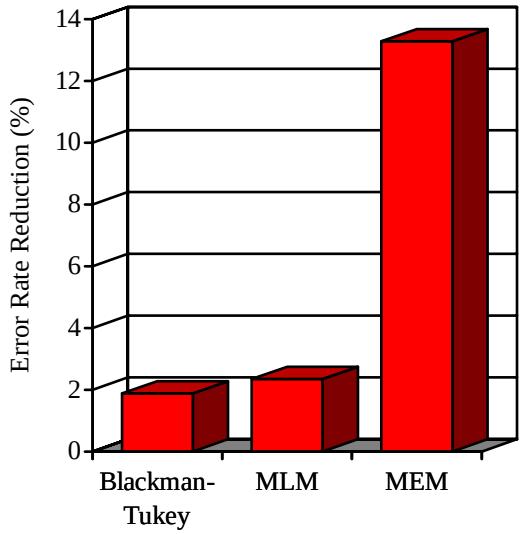

Figure 3: Relative reduction in error rate when replacing MFCCs by 'auditory' PSD estimates, using delta-delta coefficients in both cases.

4. CONCLUSIONS

In the experiments reported here, it appears that auditory PSD estimates consistently out-perform DFT-derived MFCC analysis as a representation of speech for ASR. It has been shown, however, that the magnitude of the improvement is variable, depending both on the method employed to estimate the PSD and that used to characterise any dynamic aspects of the data. In

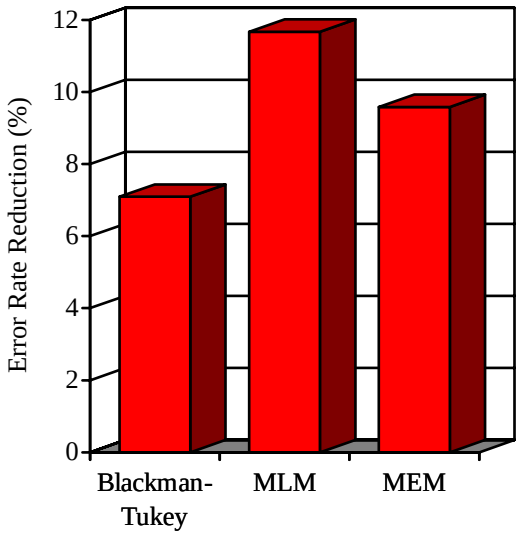

Figure 4: Relative reduction in error rate due to use of auditory frequency-domain transformation.

particular the advantage offered by the auditory PSD estimates was less marked when delta-delta coefficients were used, ***except*** when the PSD was estimated with the maximum entropy method. In this latter case, the advantage was very significant (similar to that reported in [9]) and was almost independent of the use of delta-delta coefficients.

This serves to act as a caveat for those who would interpret published results claiming superiority for any representation, be it auditory or non-auditory. There is a complicated interaction between the representation and the recognition algorithm, and future developments in the latter may invalidate any current results assessing the former.

Having established that, in all the experiments described here, the auditory version of the maximum entropy method gave around 15% reduction in phoneme recognition errors, compared to the MFCC representation. This advantage could not, however, be attributed merely to the fact that the MEM is based on AR modelling of the speech signal: the MEM and MLM methods were both based on the same AR model analysis of the data, but the MLM results were more like those obtained with the Blackman-Tukey method, which is based on an ACF/DFT approach. In the case when delta-delta coefficients were used, there was little difference between MFCC and either Blackman-Tukey or MLM auditory PSD estimates.

For all the PSD estimates used here, there was about 10% improvement to be gained by the use of an auditory frequency scale and an auditory frequency resolution.

REFERENCES

[1] Bahl, L. R., Brown, P. F., De Souza, P. V., and Mercer, R. L., "Speech recognition with continuous-parameter hidden Markov models", in "Readings in Speech Recognition" eds: A. Waibel and K.-F. Lee, Morgan Kaufmann, 1990, pp. 332-339.

[2] Robinson, A. J., "Artificial neural networks: the mole-grips of the speech scientist" in "Visual Representations of Speech Signals", eds: M. P. Cooke, S. W. Beet and M. D. Crawford, John Wiley and Sons, Ltd., 1993, pp. 83-94.

[3] Hermansky, H., Tsuga, K., Makino, S., and Wakita, H., "Perceptually based processing in automatic speech recognition", Proc. ICASSP '86, Tokyo, pp. 1971-1974, 1986.

[4] Musicus, B. R., "Fast MLM power spectrum estimation from uniformly spaced correlations", IEEE Trans. ASSP, vol. 33, pp. 1333-1335, 1985.

[5] Proakis, J. G., Rader, C. M., Ling, F., and Nikias, C. L., "Advanced Digital Signal Processing", Macmillan, 1992.

[6] Takeshima, H., Kumagai, M., Suzuki, *et al.*, "Some experimental results for a full-scale revision of equal-loudness level contours", Proc. 15th International Congress on Acoustics, Trondheim, June 1995, pp. 301-304.

[7] Moore, B. C. J., and Glasberg, B. R., "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns", Hearing Research, vol. 28, pp. 209-225, 1987.

[8] Price, P. J., Fisher, W., Bernstein, J., *et al.*, "A database for continuous speech recognition in a 1000-word domain", Proc. ICASSP '88, New York, vol. 1, pp. 651-654, 1988.

[9] Mashao, D. J., Gotoh, Y., and Silverman, H. F., "Analysis of LPC/DFT features for an HMM-based alphadigit recogniser", IEEE Signal Processing Letters, vol. 3, pp. 103-106, 1996.