IMPROVING AUTOREGRESSIVE HIDDEN MARKOV MODEL RECOGNITION ACCURACY USING A NON-LINEAR FREQUENCY SCALE WITH APPLICATION TO SPEECH ENHANCEMENT

B. T. Logan and A. J. Robinson

Cambridge University Engineering Department, Cambridge, United Kingdom

ABSTRACT

A new method to improve the accuracy of Autoregressive Hidden Markov Model (AR-HMM) based recognition systems is proposed. The technique uses the bilinear transform to warp the frequency scale of the observation vectors, hence it uses a better perceptual measure to compare the observation vectors to the trained models. Results presented for the E-set letters from the ISOLET database and the first speaker dependent task of the Resource Management (RM) database show that this technique improves recognition accuracy considerably. However, in the case of the RM system, the recognition results still fall short of those obtained from a similar mel-frequency cepstral (MFCC) based system without delta parameters. Reasons for the inferior performance of the AR-HMM system are proposed and future research directions are suggested. The models built for the RM task are incorporated into an existing enhancement algorithm to form a large vocabulary speaker dependent enhancement system. Preliminary results are presented for this system.

1. INTRODUCTION AND MOTIVATION

The motivation for this work comes from the desire to extend previous enhancement systems [1], [2] to a vocabulary independent system. These techniques estimate the clean speech and noise models within an autoregressive HMM framework [3]. AR-HMMs are used to model the speech and noise and a combined model is built and used to recognise the noisy speech. A new noise model is estimated according to the alignment and the process is repeated until the total likelihood converges to a maximum. Thus enhancement and recognition in unknown noise conditions is possible.

Autoregressive HMMs are used because they segment the speech into clusters of signals with similar autocorrelation parameters. These are used to form Wiener filters to enhance the speech. The task is made easier because for additive noise, the noisy speech observation vector is a linear combination of the speech and noise vectors [4].

However, recognition systems based on AR-HMMs have been neglected in recent years due to their inferior performance compared to MFCC and perceptual linear predictive (PLP) based systems. Yet it is less preferable to use the MFCC and PLP parameterisations for enhancement due to their inherent non-linearities. It would be desirable however to incorporate the advantages of these parameterisations into the AR framework.

AR-HMMs as described in [3] effectively use a linear frequency scale to compare the spectrum of an observation to that of a trained model. Yet is is well known [5] that it is more appropriate to use a warped frequency scale such as the Mel or Bark scale since this corresponds to the frequency resolution of humans. Non-linear frequency scales are used by both MFCC and PLP systems. Therefore it seems reasonable to investigate an AR-HMM system using a non-linear frequency scale.

The bilinear transform [6] has been used in the past to improve the performance of linear prediction coding systems [7]. The procedure transforms a time sequence to a new sequence with a warped spectrum. By adjusting the so-called warping factor, the degree of warping can be made to be a very good approximation to the Bark scale. The work in this paper applies the bilinear transform to an AR-HMM system and shows that it can benefit from the use of a warped frequency scale.

A very nice feature of the bilinear transform is that it can be reversed by applying the same transform with the negative of the warping factor. Thus the use of this technique to improve AR-HMM recognition performance will not interfere with the enhancement part of the system since the feature vector can be "unwarped" after recognition in preparation for enhancement.

2. EXPERIMENTAL SETUP

Experiments were conducted using the ISOLET database (collected for use in [8]) and the Resource Management (RM) database [9].

The ISOLET database contains two isolated tokens of each letter of the alphabet for 150 American English speakers, 75 male and 75 female. 120 speakers were used for training and 30 for testing. The speech is sampled at 16kHz. Experiments were performed using the English E-set letters ({"B", "C", "D", "E", "G", "P", "T" and "V"}) only.

The implementation of the AR-HMM system was as follows. One 13 state HMM was trained for each letter. The order of the AR models was 20. Experiments were conducted using various numbers of mixture components and with and without the frequency warping. The frequency warping was implemented in the frequency domain according to [7].

A baseline MFCC-based system was trained to allow a comparison. This also had one 13 state HMM for each letter with 12 MFCC coefficients and one energy coefficient per state. Recognition was performed with and without the 13 delta coefficients and using various numbers of mixture components.

The RM database is suitable for large vocabulary continuous speech recognition experiments. The speaker dependent part of this database was used for the experiments described in this paper. This consists of 600 training and 100 test sentences. The results presented here are for speaker "bef0_3". Again the speech is sampled at 16kHz.

Multiple mixture 3 state triphone-clustered HMMs were trained for this task. Apart from the number of states, the parameterisation for the AR-HMMs and baseline MFCC-HMMs was identical to that used in the ISOLET experiments.

The optimal method of clustering the triphones in the AR-HMM system is still an area of investigation. It was found that improved results could be obtained by the use of a MFCC-based system to dictate the clusters.

3. RECOGNITION EXPERIMENTS

3.1. Determination of the Warping Factor

In order to determine the warping factor for the given sampling rate, the recognition accuracy of the ISOLET system was investigated for various warping factors. These results are plotted in Figure 1 (with a 4th-order polynomial fitted to the points). They indicate that the exact choice of warping factor is not critical and that a factor in the range 0.5 - 0.6 will produce good results.

Smith [10] has presented a formula to determine the optimal warping factor for a given sampling frequency. This formula is reproduced here as Equation 1.

$$\alpha \approx 1.0211 \left[\frac{2}{\pi}\tan^{-1}(0.076 f_s)\right]^{\frac{1}{2}} - 0.19877 \qquad (1)$$

Here α is the warping factor and f_s is the sampling frequency measured in kHz. For the given sampling frequency of 16kHz, Equation 1 gives a warping factor of 0.57. This corresponds well with the minimum of the graph in Figure 1 and was the chosen warping factor for the experiments. Figure 2 shows the approximation of the bilinear transform to the Bark scale for this warping factor.

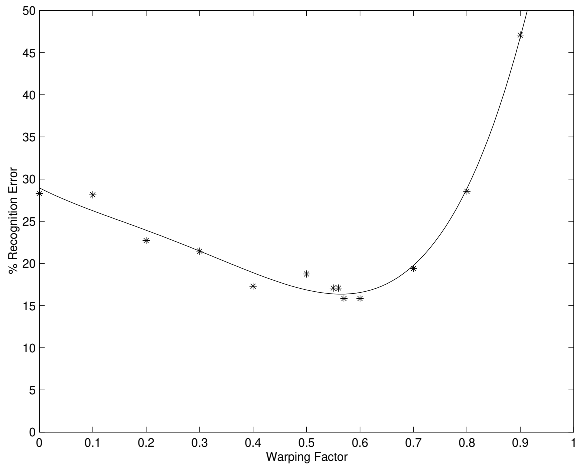

Figure 1: Recognition Error Rates for Various Warping Factors (ISOLET Data)

3.2. ISOLET Database

Recognition results for both the AR-based and MFCC-based systems on the ISOLET database are shown in Table 1. The '% Error' figure in this table was calculated using the following formula.

$$\text{\% Error} = \frac{D + S + I}{N} \cdot 100\% \qquad (2)$$

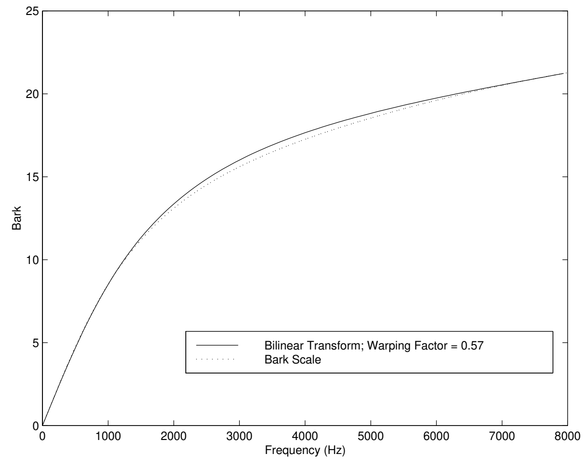

Figure 2: Bilinear Transform Approximation to the Bark Scale

Model	Number Mixture Components	% Error (D,S,I)
AR no warping	1	32.3 (0,154,1)
	2	27.1 (0,130,0)
	3	28.3 (0,136,0)
AR with warping	1	24.0 (0,114,1)
	2	20.6 (0,99,0)
	3	15.8 (0,76,0)
MFCC no deltas	1	21.6 (0,67,0)
	2	21.0 (0,50,1)
	3	17.1 (0,82,0)
MFCC with deltas	1	13.9 (0,67,0)
	2	10.6 (0,50,1)
	3	7.5 (0,36,0)

Table 1: ISOLET E-Set Recognition Results

Here D, I and S represent the number of deletions, insertions and substitutions respectively and N is the number of letters in the test set.

It can be seen that warping the frequency scale decreases the recognition error of the AR-HMM system quite considerably. In fact for this task the error rate is comparable to the MFCC-based system without delta parameters. It would appear then that that in this domain the information in the delta parameters is the now main information missing from this modified AR-HMM system.

3.3. RM Database

Recognition results for both the AR-HMM and MFCC-based systems on the RM Database are shown in Table 2. It can been seen from these results that again recognition has been improved considerably using the bilinear transform. However, for this task, the performance of the AR-HMM systems still fall far short of the performance achievable using a MFCC-based system. It appears then that for this large vocabulary system, while the use of a perceptual frequency scale does improve recognition, it

Model	Number Mixture Components	% Error (D,S,I)
AR no warping	4	37.8 (48,204,57)
	5	32.6 (45,168,54)
AR with warping	4	24.0 (39,129,28)
	5	20.4 (34,112,21)
MFCC no deltas	4	11.1 (12,44,35)
	5	9.8 (12,40,28)
MFCC with deltas	4	6.9 (12,23,22)
	5	5.1 (3,20,19)

Table 2: RM Recognition Results - speaker BEF0_3

does not completely explain the difference between the information provided by the non-warped AR-HMM system and the MFCC-no-delta system.

3.4. Discussion

One variation between the MFCC-no-delta and AR-HMM system is the lack of energy information in the AR-HMM system. To investigate the effect of this, energy information was incorporated into the AR-HMM framework along the lines of [11]. The likelihood expression for an observation given the current state was modified by the addition of a term describing the probability of the state having a given energy. This extra term is scaled by an empirically chosen factor. It was found that some slight gains in recognition accuracy of the order of 1-2% absolute could be made by the use of this technique which tended to mostly correct substitution errors. Hence there is still an unaccounted for difference between the two systems.

The major source of deviation between the two systems is caused by the differing distortion measures used. Both systems essentially use spectral difference measures to compare an utterance to trained templates [12]. However the MFCC-based system uses a more flexible distortion measure. While the AR-HMM system treats errors in any part of the spectrum equally, the MFCC-based system effectively weights the error in each part of the spectrum by the (trained) variance of each estimate.

It should be noted that this is done at the expense of twice the number of system parameters. (i.e. the MFCC-based system has both a mean and a variance for each state whereas the AR-HMM system has a mean and an implicit variance). However as shown in Table 3 increasing the number of AR-HMM system parameters by increasing the number of mixture components does not solve this problem. This latter result was also observed in [11] on a simpler task and essentially the same conclusion was reached.

Therefore it seems reasonable that future directions of research should address improving the distortion measure of the AR-HMM system and somehow incorporating confidence into the estimate. Other areas of investigation include improving clustering techniques and energy incorporation. Clearly the superior performance of MFCC-based systems cannot be ignored and it is hoped that a careful study of their exact workings will show further ways of improving AR-HMM systems, as it already has done in the case of introducing the frequency warping.

Model	Number Mixture Components	% Error (D,S,I)
AR with warping	5	20.4 (34,112,21)
	6	20.1 (39,101,24)
	7	20.3 (39,102,25)

Table 3: RM Recognition Results - speaker BEF0_3 - Increasing Mixture Components

4. ENHANCEMENT EXPERIMENTS

The ultimate aim of this work was the construction of a vocabulary independent enhancement system along the lines of [2]. This system is able to perform enhancement in unknown noise conditions. Previously it had only been tested on a small vocabulary system.

Gaussian noise at approximately 6dB signal-to-noise ratio was added to test utterances from the RM database. These were then enhanced using the method of [2] and the models developed in Section 3.3. Some initial observations are described below.

The algorithm appears to be reasonably sensitive to initialisation of the noise statistics. In [2], the noise statistics were initialised using all the frames of the utterance to be enhanced. The nature of the data in [2] (isolated digits) meant that proportionally more frames were noise than in the experiments in this paper and thus the initial noise estimate was superior. Since the enhancement algorithm only converges to a likelihood local maximum, a good initial estimate is required. Thus it was found that enhancement could be improved by initialising the noise statistics from either the first frame (this makes the assumption though that the utterance is preceded by speech-free section) or the frame with the least power.

A further observation was that the system could perform quite effective enhancement. Figures 3, 4 and 5 show the clean, noisy and enhanced spectrums for a typical utterance.

5. CONCLUSIONS

It has been shown that the accuracy of an AR-HMM based recognition system can be improved considerably by the use of the bilinear transform to warp the frequency scale. Results are presented for both a small vocabulary system speaker-independent system and a large vocabulary speaker-dependent system. For the small vocabulary system, the recognition results for the warped AR-HMM system were comparable with those of a MFCC-based system without delta parameters. This was not the case for the large vocabulary system. It was reasoned that this was due to the difference in distortion measures between the two systems. Thus the improvement of the AR-HMM distortion measure would seem to be the most obvious path for future recognition experiments although this is by no means the only path.

The models built for the large vocabulary recognition system were incorporated into an existing enhancement algorithm to form a large vocabulary speaker- dependent enhancement system.

6. ACKNOWLEDGEMENTS

B. T. Logan gratefully acknowledges funding from the Cambridge Commonwealth Trust and an ORS Award.

7. REFERENCES

[1] B. T. Logan and A. J. Robinson, "Noise estimation for enhancement and recognition within an autoregressive hidden-Markov-model framework," in *Proceedings Sixth Australian International Conference on Speech Science and Technology*, pp. 85–90, 1996.

[2] B. T. Logan and A. J. Robinson, "Enhancement and recognition of noisy speech within an autoregressive hidden Markov model framework using noise estimates from the noisy signal," in *Proceedings ICASSP*, pp. 843–846, 1997.

[3] B. Juang, "On the hidden Markov model and dynamic time warping for speech recognition - a unified view," *AT&T Bell Laboratories Technical Journal*, vol. 63, pp. 1213–1243, Sept. 1984.

[4] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 40, pp. 725–735, Apr. 1992.

[5] J. Deller, J. Proakis, and J. Hansen, *Discrete–Time Processing of Speech Signals*. Macmillan, 1993.

[6] A. V. Oppenheim and D. H. Johnson, "Discrete representation of signals," *Proceedings of the IEEE*, vol. 60, pp. 681–691, June 1972.

[7] H. W. Strube, "Linear prediction on a warped frequency scale," *J. Acoust. Soc. Am.*, vol. 68, pp. 1071–1076, Oct. 1980.

[8] M. Fanty and R. Cole, "Spoken letter recognition," in *Proceedings Neural Information Processing System Conference*, 1990.

[9] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The DARPA 1000-word Resource Management database for continuous speech recognition," in *Proceedings ICASSP*, pp. 651–654, 1988.

[10] J. O. Smith and J. S. Abel, "The bark bilinear transform," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 1995.

[11] B. Juang and L. R. Rabiner, "Mixture autoregresive hidden markov models for speech signals," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-33, pp. 1404–1413, Dec. 1985.

[12] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Prentice–Hall, 1993.

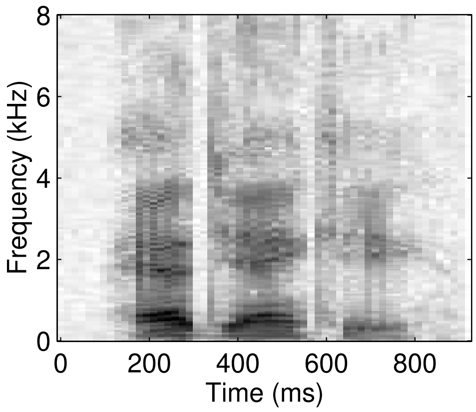

Figure 3: Clean Speech ("Add Yankee") [sound A0775S01.WAV]

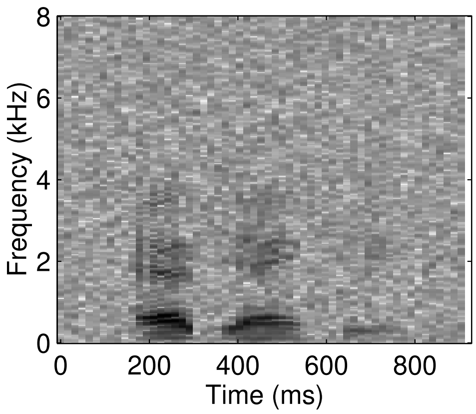

Figure 4: Noisy Speech ("Add Yankee") with 6dB Gaussian Noise [sound A0775S02.WAV]

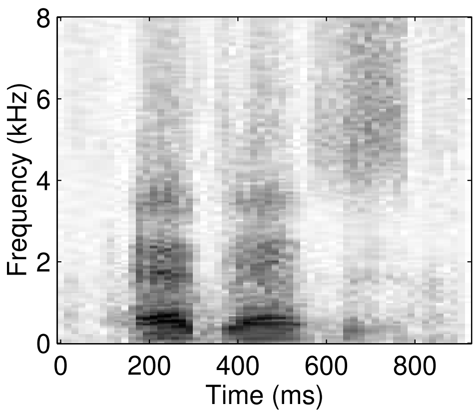

Figure 5: Enhanced Speech ("Add Yankee") from 6dB Gaussian Noisy Speech [sound A0775S03.WAV]