A PRACTICAL PERCEPTUAL FREQUENCY AUTOREGRESSIVE HMM ENHANCEMENT SYSTEM

Beth Logan　　*Tony Robinson*

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.

ABSTRACT

We have previously developed a speech enhancement scheme which can adapt to unknown additive noise. We model speech and noise using perceptual frequency or 'warped' autoregressive HMMs (AR-HMMs) and estimate the clean speech and noise parameters within this framework. In this current work, we investigate the use of our system as a front end to a MFCC recognition system trained on clean speech. To use our system as a front end, we make two modifications. First, we use minimum mean squared error (MMSE) spectral rather than time domain estimators for enhancement. Second, for computational reasons, we form estimators from non-warped AR-HMMs. To avoid mismatch introduced when converting between warped and non-warped models, we use parallel sets of models.

Results are presented for small and medium vocabulary tasks. On the simple task, we are able to approach the performance of a matched system when language model information is included. On the second task, we are not able to incorporate a language model due to modelling deficiencies in AR-HMMs. However, we still demonstrate substantial improvements over baseline results.

1. INTRODUCTION

Speech processing systems can suffer unacceptable performance degradation in the presence of background noise. In this paper, we consider the effect of additive noise on clean speech recognition systems and use speech enhancement as a front end to improve their performance. In particular, we are interested in the case where the noise statistics are unknown. Recent approaches to adaptive speech enhancement are summarised in [2]. Many of the techniques use Kalman filters.

We base our system on work by Ephraim [1]. This technique models speech and noise using autoregressive HMMs (AR-HMMs) and uses these to form a compensated model. This compensated model is used to determine the probability of each compensated state given the noisy observation. These probabilities are then used to weight estimators of the clean speech given that state. The extension to multiple mixture systems is straightforward.

The basic technique uses Wiener filter estimators which have been shown to be inferior to Kalman filters [5]. However, it is possible to use other estimators within the same framework. These may be more applicable when the enhancement system is used as a front end to a recognition system. The incorporation of prior information in the form of trained clean speech models is another advantage of this technique.

We have previously presented two extensions to the work in [1]. In the original system, the noise models were trained on given examples. In [3], we show that maximum likelihood (ML) estimates can instead be made of the noise statistics. It is possible to use ML parameter estimation in the AR-HMM domain to adapt to additive noise because AR-HMMs model features which are additive. This is more difficult if for example cepstral-based HMMs are used.

In later work [4], we show that the modelling power of AR-HMMs is improved by the incorporation of perceptual frequency. Here, the bilinear transform is used to produce autocorrelation coefficients on a warped frequency scale which is a good approximation to the perceptually meaningful Bark scale. These autocorrelation coefficients are then used to construct warped AR-HMMs which are shown to give superior recognition performance to non-warped models.

To date we have only presented limited evaluations of our system. In this current work we present substantial qualitative support for our technique by investigating its potential as a front end to recognition systems trained on clean speech. Some modifications to our previous algorithm are needed and are described in the following section.

2. THE ENHANCEMENT SYSTEM

Our enhancement system models speech and noise using warped AR-HMMs. Within this framework, the noise statistics are determined and estimators formed for enhancement.

Previously, we studied Wiener filter estimators. These give the MMSE time domain estimate of the clean speech. Several other estimators are proposed in [1] however.

In this work, we derive enhanced MFCC parameters directly from the enhanced spectra rather than from an enhanced wave resynthesised in the time domain. We thus found improved performance when we used the MMSE power spectral density (PSD) estimator described in [1] since errors in the spectral domain more strongly influence the enhanced cepstral features. This estimator was therefore used for all the experiments described here.

A further modification to our previous system is to decouple the processes of calculating the posterior probability from the process of applying the estimators. As described, we use warped AR-HMMs for probability calculations since these are better at modelling speech [4]. However, for computational reasons, it is desirable to use non-warped AR-HMMs to form the estimators for enhancement.

We must therefore convert between estimators in the warped and non-warped domains. It is possible to unwarp the estimators for enhancement directly. However, there is mismatch introduced by the unwarping process [8].

Since in this work we use our system as a front end to a standard recognition system, we wish to minimise this mismatch. We achieve this by training warped and non-warped models in parallel and using the warped models to calculate probabilities and the non-warped models to form estimators. The system is shown in Figure 1.

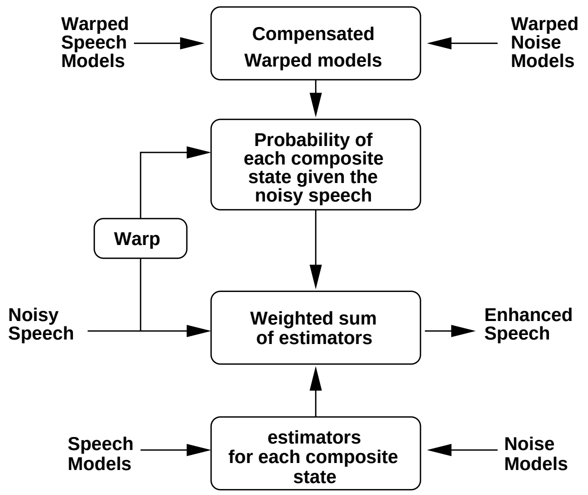

Figure 1: A perceptual frequency enhancement system. The weights for each estimator are determined using warped AR-HMMs. The estimators are formed using non-warped AR-HMMs.

The clean speech non-warped models are trained using single pass retraining. This technique generates a parallel set of models by computing the state probabilities using one set of models and training data, and then switching to different training data to compute parameter estimates for a second model. The noise statistics are estimated in both the warped and non-warped domains in order to implement the adaptive enhancement system.

3. SMALL VOCABULARY SPEAKER DEPENDENT EXPERIMENTS

The first set of experiments use the NOISEX-92 database [9]. We study the male isolated digits task corrupted by the following four stationary noise sources: Lynx helicopter noise, speech noise, car noise and F16 aircraft noise.

We study enhancement systems based on two types of clean speech models: word-based models and general models. We model the noise using a single state AR-HMM with autoregressive order 20. This model is initialised by assuming the whole utterance is noise.

3.1. Clean Speech Models

The clean speech MFCC HMM recognition system contains an 8-emitting state left-to-right HMM model for each digit and a 1-emitting state model for silence. The MFCC feature vectors contain 15 cepstral coefficients including the zeroth coefficient. Diagonal covariance matrices are used.

The standard Baum-Welsh algorithm is used for training. Connected word Viterbi decoding is used for recognition (i.e. not isolated word recognition). The syntax for the recognition network is constrained to be a string of digits each followed by silence.

We also construct warped and non-warped AR-HMM clean speech recognition systems. These are needed for the word-based enhancement system and for baseline experiments. These have the same topology as the MFCC models except there are 2 mixture components per state. The order of the autoregressive models is 20.

All the recognition and enhancement systems use frames of 32ms with overlap of 16ms. These parameters are chosen to be convenient for construction of enhanced time domain waveforms (used for perceptual evaluations not discussed here).

3.2. Baseline Performance

The results in this section are the best results achievable for clean and matched systems. They are obtained by optimising the insertion penalty and another parameter, the silence probability increment, for each test condition. This latter parameter, described in [7], weights the log observation probability of the silence model by a fixed value to improve the chance of low energy frames at word boundaries being recognised correctly as silence.

Table 1 shows the summary recognition error rates for speech corrupted by each of the four noise sources and tested with clean and matched MFCC models. The matched systems are obtained using single pass retraining. In this and subsequent tables, we show the average word error rates for the four noise sources at each SNR. D, S and I are the total number of deletion, insertion and substitution errors respectively. We see that the performance of the clean models degrades rapidly with decreasing SNR.

Table 2 summarises the word error rates for the compensated non-warped and warped AR systems. The compensated models are formed using trained noise models. They are thus the best compensated models available to calculate the probabilities required for enhancement. We see that the warped AR-HMM system has superior performance. Thus the remainder of this paper will focus on enhancement systems based on warped AR-HMMs.

SNR (dB)	% Error (D,S,I)			
	Clean		Matched	
∞	0.00	(0,0,0)	0.00	(0,0,0)
18	54.50	(57,97,64)	0.00	(0,0,0)
12	77.00	(88,160,60)	0.00	(0,0,0)
6	92.00	(300,0,0)	0.25	(0,1,0)
0	95.00	(380,0,0)	2.50	(0,10,0)
-6	95.00	(380,0,0)	32.50	(37,81,12)

Table 1: Word error rates for speech corrupted by the four noises and recognised using clean and matched MFCC models.

SNR (dB)	% Error (D,S,I)			
	AR Models		Warped AR Models	
18	1.00	(0,4,0)	0.00	(0,0,0)
12	3.25	(0,13,0)	0.00	(0,0,0)
6	8.00	(3,29,0)	0.75	(0,3,0)
0	22.5	(13,67,10)	4.75	(2,17,0)
-6	38.75	(30,100,25)	20.75	(14,57,12)

Table 2: Word error rates for corrupted speech recognised using compensated AR models.

3.3. Word-Based Models

The first set of enhancement experiments investigates systems using word-based HMMs. For these experiments, Viterbi alignment is used to obtain the most likely speech and noise state for each frame given the noisy observation. The most likely mixture component given this state is then determined. The speech and noise statistics for this mixture component are then used to to reestimate the noise parameters and to construct estimators for enhancement.

We found that the optimal insertion penalty used during Viterbi alignment varied according to the SNR. In order to automatically choose this parameter, the NIST tool *wavmd* was used to approximate the SNR for each test file. This was then mapped to an insertion penalty.

The first column of Table 3 summarises the enhancement results obtained using this scheme. We see that substantial improvements have been made over baseline results and that the error rates are comparable to the matched model results in Table 1.

SNR (dB)	% Error (D,S,I)			
	Word Models		General Models	
18	0.00	(0,0,0)	2.50	(3,6,1)
12	0.00	(0,0,0)	15.75	(16,44,3)
6	0.50	(0,2,0)	51.50	(93,103,10)
0	6.25	(1,14,10)	78.75	(274,41,0)
-6	26.50	(30,53,23)	85.75	(332,11,0)

Table 3: Word error rates for corrupted speech enhanced adaptively. The general models contain 128 mixture components in the speech state.

3.4. General Models

We now investigate the effect of removing the language model information from the enhancement system. We model the clean speech and silence using a two-state ergodic HMM. The first state models speech using 128 mixture components and the second state models silence using a single mixture component. For this system, the forward-backward equations are used instead of Viterbi alignment to calculate the likelihood of each mixture component of the compensated model.

The second column of Table 3 summarises the results for this system. These results are inferior to the word-based system despite having a comparable number of mixture components. Thus we conclude that some performance is sacrificed by the use of simpler models.

4. MEDIUM VOCABULARY SPEAKER INDEPENDENT EXPERIMENTS

In this section, we investigate the performance of our algorithm on the more challenging Resource Management (RM) task [6]. Since the RM database contains clean speech only, Lynx noise from the NOISEX-92 database was added to the test sets at 18dB and 12dB.

In [4], we found that on the speaker dependent RM task, the performance of a clean speech AR-HMM system was worse than a standard MFCC system. This was because currently AR-HMM systems cannot incorporate delta features and because the MFCC system uses a superior distortion measure.

The speaker independent task is significantly harder. Typically, it is necessary to incorporate acceleration features in addition to delta features. We were therefore unable to construct a clean AR-HMM medium vocabulary speaker independent recognition system with acceptable performance. This implied that we could not incorporate language model information into our enhancement system.

Therefore a scheme based on general speech models is studied. A similar non-adaptive enhancement system based on compensated MFCC models without delta parameters has been shown to perform well on this task [7].

4.1. Clean Recognition Systems

The enhanced MFCC parameters are evaluated using a clean speech recognition system as before. This is trained using the RM Toolkit [10] as a template. The system models 3-state left-to-right clustered triphones with 5 mixture components per state. The feature vectors contain 13 cepstral coefficients including the zeroth coefficient augmented with delta and acceleration coefficients. These are modelled using diagonal covariance matrices. The data is pre-emphasised by the filter $H(z) = 1 - 0.97z^{-1}$.

The frame rate and frame size are 16ms and 32ms respectively as used as in the previous experiments. These differ from the standard parameters used in the RM Toolkit. The non-standard frame rate affects the modelling of short phones by increasing the minimum duration. This problem is alleviated by the introduction of

Model	SNR (dB)	% Error				
		Feb89	Oct89	Feb91	Sep92	Avg.
Clean	∞	6.3	7.3	5.9	11.0	7.6
	18	38.9	30.4	35.8	43.1	37.0
	12	80.4	81.0	77.7	85.2	81.1
Matched	18	16.7	14.8	14.1	21.0	16.7
	12	40.8	31.3	34.0	40.4	36.6

Table 4: Baseline results for the RM database speaker independent test sets for clean speech and speech corrupted using Lynx noise. Performance using clean and matched models is shown.

SNR (dB)	Nr. Mixes	% Error				
		Feb89	Oct89	Feb91	Sep92	Avg.
18	128	23.6	20.1	21.7	27.6	23.2
	256	18.7	16.5	18.9	23.9	19.5
	512	18.1	15.5	18.1	24.2	19.0
12	512	42.8	35.7	37.9	46.7	40.8

Table 5: Enhancement results for the RM speaker independent test sets for Lynx noise. The speech is enhanced using general speech models with varying numbers of mixture components.

a skip state into each triphone model. The frame rate also affects the period of time used to calculate the delta and acceleration coefficients. This effect was not considered.

4.2. Baseline Performance

Table 4 shows the word error rates for the clean and noisy speech on the four test sets. The clean baseline is worse than the published performance on this database because of the decreased frame rate as discussed. We see that the addition of noise has a substantial effect on the error rate. Also in this table are the word error rates for a matched system.

4.3. Enhancement Performance

We first investigate the 18dB noise condition. Table 5 shows the word error rates for various numbers of mixture components in the models. We see that a substantial improvement has been made on the baseline performance. The performance improves as the number of mixture components increases although the difference between the 512-mixture and 256-mixture systems is not significant. The last row of this table shows the error rate for the 512-mixture system at 12dB. Again substantial improvements have been made over the baseline performance.

From these two test conditions, it appears that the improvement gained by the enhancement technique halves the error rate. However this performance is significantly worse than the matched model results suggesting that there is a modelling deficiency.

5. CONCLUSIONS

We have shown that the enhancement system developed in in [3] and [4] can be used as a front end to a recogniser trained on clean speech to improve recognition performance in the presence of unknown noise.

We modified our existing system in two ways to use it as a front end. First, we form MMSE spectral estimates rather than MMSE waveform estimates of the enhanced wave. Second, although we use perceptual frequency or warped AR-HMMs to model the speech, we use a parallel set of non-warped models to form estimators. This minimises the mismatch between the enhanced cepstral parameters and the clean speech model.

Our results are encouraging. On the small vocabulary, speaker dependent task, we were able to approach the performance of a matched model when a language model was used. We were however unable to use a language model on the medium vocabulary, speaker independent task because currently AR-HMMs do not incorporate delta parameters. Despite this, we were still able to substantially reduce the error rate compared to unprocessed speech.

6. REFERENCES

1. Y. Ephraim. A Bayesian estimation approach for speech enhancement using hidden Markov models. *IEEE Trans. on Signal Processing*, 40(4):725–735, April 1992.
2. S. Gannot, D. Burshtein, and E. Weinstein. Iterative and sequential Kalman filter-based speech enhancement algorithms. *IEEE Trans. on Speech and Audio Processing*, 6(4):373–385, July 1998.
3. 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 *Proc. ICASSP*, pages 843–846, 1997.
4. B. T. Logan and A. J. Robinson. Improving autoregressive hidden Markov model recognition accuracy using a non-linear frequency scale with application to speech enhancement. In *Proc. EUROSPEECH*, pages 2103–2106, September 1997.
5. K. K. Paliwal and A. Basu. A speech enhancement method based on Kalman filtering. In *Proc. ICASSP*, pages 6.3.1–6.3.4, 1987.
6. P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett. The DARPA 1000-word Resource Management database for continuous speech recognition. In *Proc. ICASSP*, 1988.
7. C. W. Seymour. *Model-Based Speech Enhancement*. PhD thesis, Univeristy of Cambridge, 1996.
8. H. W. Strube. Linear prediction on a warped frequency scale. *J. Acoust. Soc. Am.*, 68(4):1071–1076, October 1980.
9. A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones. The noisex-92 study on the effect of additive noise on automatic speech recognition. Technical report, DRA Speech Research Unit, 1992.
10. S. J. Young, P. C. Woodland, and W. J. Byrne. *HTK: Hidden Markov Model Toolkit V1.5*. Cambridge University Engineering Department Speech Group and Entropic Research Laboratories Inc., September 1993.