ROBUST CONTINUOUS SPEECH RECOGNITION USING PARALLEL MODEL COMBINATION

M. J. F. Gales & S. J. Young

CUED/F-INFENG/TR 172

March 1994

Cambridge University Engineering Department
Trumpington Street
Cambridge CB2 1PZ
England

Email: mjfg@eng.cam.ac.uk / sjy@eng.cam.ac.uk

Abstract

This paper addresses the problem of automatic speech recognition in the presence of interfering noise. It focuses on the Parallel Model Combination (PMC) scheme, which has been shown to be a powerful technique for achieving noise robustness. However, most experiments reported on PMC to date have been on small, 10-50 word vocabulary systems. In this paper, PMC is applied to the Resource Management (RM) 1000 word continuous speech recognition task. This reveals compensation requirements not highlighted by the smaller vocabulary tasks, in particular, it is necessary to compensate the differential as well as the static parameters to achieve good recognition performance.

The database used for these experiments was the RM speaker independent task with Lynx helicopter noise from the NOISEX-92 database added. The experiments reported here used the HTK RM recogniser developed at CUED modified to include PMC based compensation for the static, delta and delta-delta parameters. After training on clean speech data, adding noise at 18-20dB signal to noise ratio was found to seriously degrade the performance of the recogniser. However, using PMC the performance was restored to a level comparable with that obtained when training directly in the noise corrupted environment. Additionally, PMC is shown to be robust to convolutional noise for this task.

Keywords: speech recognition, noise compensation, HMM, PMC, Resource Management.

Contents

1 Introduction

In recent years the size and complexity of speech recognition tasks has greatly increased. However, the vast majority of work has involved 'clean' speech collected in quiet environments. For practical systems it is necessary to make large vocabulary systems robust to interfering noise. Many different approaches to achieving noise robustness have been studied [8]. These approaches may be split into two groups.

Firstly, the corrupted waveform may be preprocessed in such a way that the resulting parameters are closely related to those of clean speech. Techniques in this category include spectral subtraction [5, 15], spectral mapping [6], and inherently robust parameterisations [16]. These methods only use statistical information about the interfering noise in the compensation process, no account is taken of what was said. Some schemes have also attempted to estimate the clean speech signal using information about the speech. These include inhomogeneous estimators using HMMs [9] and minimum mean square error estimators [7]. Additionally, techniques have attempted to estimate the clean speech under additive and convolutional noise conditions [1].

The second class of methods attempt to modify the pattern matching stage in order to account for the interfering noise. Methods using this approach include noise masking [14, 17], state based filtering [2], cepstral mean compensation [4, 3], HMM decomposition [19], and Parallel Model Combination (PMC) [10, 11, 12, 13].

This paper is concerned with the latter approach to noise robustness, in particular, the scheme based on PMC. The basic concept behind PMC is that the performance of speech recognition systems is optimal when there is no mis-match between training and test conditions. Invariably in real applications, there is some mis-match, either in the form of additive noise, or variations in the channel conditions. Some method for compensating the parameters of the models, or for re-training of the models is required. If the effect of the mis-match is known, for example in interfering additive noise, it should be possible to modify the training data to match this new test condition and then re-train the models. This would require that the whole database be stored and modified whenever the conditions change, a highly computationally expensive task. It is therefore preferable to compress the training data into a more manageable form. One method is to store statistics derived from the training data. The task is then to train a new set of models using these training statistics. PMC adopts this approach and uses HMMs to model both the clean speech database and the additive interfering noise. 'Mis-match' functions for static, delta and delta-delta parameters have previously been derived and implemented [12, 13]. The technique has been shown to work well on small vocabulary tasks, however, to date little work has been done on medium to large vocabulary systems. This paper describes experiments involving a medium sized vocabulary task, the Resource Management (RM) Task with noise from the NOISEX-92 database artificially added.

2 Parallel Model Combination

The objective of any HMM based noise compensation scheme is to estimate, according to some objective function, the corrupted speech model given information about the clean speech and interfering noise. If the corrupted speech is to be modelled by a standard HMM, then to obtain the ML estimate of the noise compensated speech model it is necessary to estimate the mean and covariance of the corrupted speech. The mis-match is assumed to be caused by interfering additive and convolutional noise. It is therefore necessary to define the 'mis-match' function for this case. To obtain this function, a series of assumptions are made.

1. The speech and noise are independent.
2. The speech and noise are additive in the time domain. In addition, it is assumed that there is sufficient smoothing on the spectral estimate so that the speech and noise may be assumed to be additive at the power spectrum level.
3. A single Gaussian or a mixture Gaussian per model state contains sufficient information to represent the distribution of the observation vectors in the log domain.

4. The frame/state allocation is not altered by the addition of noise.

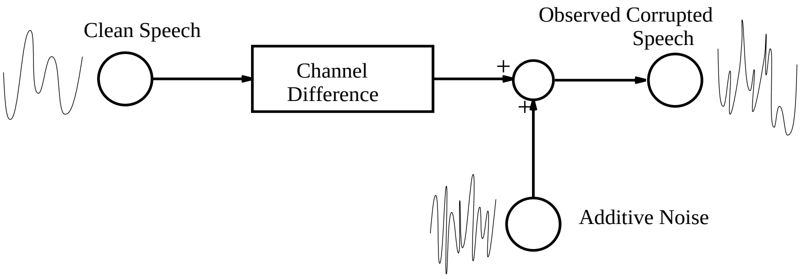

Figure 1: Overall Structure of System

Figure 1 shows how the convolutional noise and the additive noise are assumed to affect the clean speech. The additive noise is shown being combined after the channel difference. This is because the only noise estimate available is the noise observed in the test channel condition, so no spectral tilt will be present on the additive noise.

2.1 Static Parameters

For the static parameters in the presence of interfering additive noise, but no convolutional noise, the 'observations' are given by the 'mis-match' function

$$O^l_i(t) = \log(g\exp(S^l_i(t)) + \exp(N^l_i(t))) \tag{1}$$

where g is a gain matching term introduced to account for level differences between the clean speech and the noisy speech, $\mathbf{S}^l(t)$ is the clean speech and $\mathbf{N}^l(t)$ is the interfering noise. Throughout the rest of this paper the superscript will be used to indicate the domain of the variable. Thus $\mathbf{O}^c(t)$ is the corrupted speech observation in the cepstral domain, $\mathbf{O}^l(t)$ is in the log spectrum domain and $\mathbf{O}(t)$ is in the linear spectrum domain. Furthermore, $\mathbf{O}^c(t)$ will represent the observation at time t and the associated random variable will be $\mathbf{O}^c$. For simplicity of notation, only one state of each model will be considered. The way in which multi-state models are combined is detailed in previous work [11].

If only a HMM with single Gaussian output distributions is to be estimated, or in the mixture Gaussian case the frame/component allocation is assumed to be unaltered by the noise, then the mean, $\hat{\mu}^l$, is given by

$$\hat{\mu}^l_i = \int_{\mathcal{R}^n} d\mathbf{S}^l \int_{\mathcal{R}^n} d\mathbf{N}^l \;\log(g\exp(S^l_i) + \exp(N^l_i))p(\mathbf{S}^l)p(\mathbf{N}^l) \tag{2}$$

The covariance, $\hat{\Sigma}^l$, estimation can be written in the same form.

$$\hat{\Sigma}^l_{ij} = \int_{\mathcal{R}^n} d\mathbf{S}^l \int_{\mathcal{R}^n} d\mathbf{N}^l \log(g\exp(S^l_i) + \exp(N^l_i))\log(g\exp(S^l_j) + \exp(N^l_j))p(\mathbf{S}^l)p(\mathbf{N}^l) - \hat{\mu}^l_i\hat{\mu}^l_j \tag{3}$$

It is not necessary to re-estimate the transition probabilities or mixture component weights, as these are assumed to be unaltered by the addition of the noise.

If the speech and noise are modelled by separate HMMs trained on cepstral feature vectors having Gaussian distributions with parameters $\{\mu^c, \mathbf{\Sigma}^c\}$ and $\{\tilde{\mu}^c, \tilde{\mathbf{\Sigma}}^c\}$ respectively, it is necessary to map these parameters to the log spectrum domain. For the speech

$$\begin{aligned} \mu^l &= \mathbf{C}^{-1}\mu^c & (4)\\ \mathbf{\Sigma}^l &= \mathbf{C}^{-1}\mathbf{\Sigma}^c(\mathbf{C}^{-1})^T & (5) \end{aligned}$$

Similarly the noise parameters $\{\tilde{\mu}^c, \tilde{\mathbf{\Sigma}}^c\}$ may be mapped to $\{\tilde{\mu}^l, \tilde{\mathbf{\Sigma}}^l\}$ where $\mathbf{C}$ is the matrix representing the discrete cosine transform. No additional assumptions are required at this stage, as the linear combination of Gaussian distributed random variables is itself Gaussian.

There is no closed form expression for either the compensated mean, $\hat{\mu}^l$, or covariance, $\hat{\mathbf{\Sigma}}^l$, as described by equations 2 and 3. To obtain exact forms for the above expressions, it is necessary to perform multi-dimensional numerical integration. Alternatively, an approximate estimate of the mean and variance may be obtained by assuming that the sum of two Log-normally distributed variables is itself Log-normally distributed [11].

2.2 Delta and Delta-Delta Parameters

For large vocabulary speech recognition it is necessary to incorporate dynamic coefficients in the speech parameterisation to achieve good performance. The mis-match function for the static parameters relies on the fact that the speech and noise are additive in the linear spectral domain. When dynamic coefficients are used this simple combination is not possible. Hence to implement delta coefficient compensation within the PMC framework, it is necessary to obtain a new 'mis-match' function [12]. If the speech is now parameterised using

$$\mathbf{O}^{\Delta c}(t)^T = \left[\mathbf{O}^c(t)^T, \mathbf{\Delta O}^c(t)^T\right] \tag{6}$$

where $\mathbf{\Delta O}^c(t)$ are the simple delta coefficients, then

$$\begin{aligned} \mathbf{\Delta O}^c(t) &= (\mathbf{O}^c(t+w) - \mathbf{O}^c(t-w)) \\ &= \mathbf{C}(\mathbf{O}^l(t+w) - \mathbf{O}^l(t-w)) \\ &= \mathbf{C}\log\left[(\mathbf{O}(t+w))./(\mathbf{O}(t-w))\right] \end{aligned} \tag{7}$$

where ./ is element-wise division and w is the difference offset. Using the assumption that the speech and noise are additive, $\mathbf{O}(t) = g\mathbf{S}(t) + \mathbf{N}(t)$, and substituting this in the above equation yields

$$\mathbf{\Delta O}^c(t) = \mathbf{C}\log\left[(g\mathbf{S}(t+w) + \mathbf{N}(t+w))./(g\mathbf{S}(t-w) + \mathbf{N}(t-w))\right] \tag{8}$$

This may be expressed in terms of the delta coefficients of the speech, $\mathbf{\Delta S}(t)$, and noise, $\mathbf{\Delta N}(t)$, in the linear domain

$$\begin{aligned} \Delta O_i(t) &= \left(\frac{gS_i(t+w)}{gS_i(t-w) + N_i(t-w)}\right) + \left(\frac{N_i(t+w)}{gS_i(t-w) + N_i(t-w)}\right) \\ &= \Delta S_i(t)\left(\frac{\frac{gS_i(t-w)}{N_i(t-w)}}{\frac{gS_i(t-w)}{N_i(t-w)} + 1}\right) + \Delta N_i(t)\left(\frac{1}{\frac{gS_i(t-w)}{N_i(t-w)} + 1}\right) \end{aligned} \tag{9}$$

Rewriting the above equation in the same form as the mis-match function for the static parameters gives

$$\Delta O_i^l(t) = \log\left(\frac{\exp(\Delta S_i^l(t) + S_i^l(t-w) + g^l - N_i^l(t-w)) + \exp(\Delta N_i^l(t))}{\exp(S_i^l(t-w) + g^l - N_i^l(t-w)) + 1}\right) \tag{10}$$

$$\begin{aligned} &= \log\left(\exp(\Delta S_i^l(t) + S_i^l(t-w) + g^l) + \exp(\Delta N_i^l(t) + N_i^l(t-w))\right) \\ &\quad - \log\left(\exp(S_i^l(t-w) + g^l) + \exp(N_i^l(t-w))\right) \end{aligned} \tag{11}$$

where $g^l = \log(g)$. The corrupted speech cepstral delta coefficients have been rewritten in terms of the static and delta coefficients of the clean speech and interfering noise. The expression for the delta coefficient at time t is dependent on the static coefficients at time $t - w$. This is contrary to one of the assumptions behind the use of HMMs for speech recognition, that the speech waveform may be split into stationary segments with instantaneous transitions between them. However, if the segments are assumed to be long enough then the statistics of $\mathbf{S}(t - w)$ will be approximately the same as those of $\mathbf{S}(t)$, and those of $\mathbf{N}(t-w)$ will be approximately the same as $\mathbf{N}(t)$. With this assumption, statistics exist for all the variables of equation 9. Alternatively, it is possible to generate an additional set of models built on statistics at time $t-w$.

An ML estimate of the delta parameters of the HMM is needed. Again this requires the mean and the covariance of the signal in the log spectrum domain to be calculated. If the speech is parameterised in the cepstral domain, it must be mapped to the log spectrum domain. For the speech

$$(\mu^{\Delta l})^T = \left[(\mathbf{C}^{-1}\mu^c)^T, (\mathbf{C}^{-1}\boldsymbol{\Delta}\mu^c)^T\right] \tag{12}$$

and

$$\boldsymbol{\Sigma}^{\Delta l} = \left[\begin{array}{cc} \mathbf{C}^{-1}\boldsymbol{\Sigma}^c(\mathbf{C}^{-1})^T & \mathbf{C}^{-1}\delta\boldsymbol{\Sigma}^c(\mathbf{C}^{-1})^T \\ \mathbf{C}^{-1}(\delta\boldsymbol{\Sigma}^c)^T(\mathbf{C}^{-1})^T & \mathbf{C}^{-1}\boldsymbol{\Delta}\boldsymbol{\Sigma}^c(\mathbf{C}^{-1})^T \end{array}\right] \tag{13}$$

where $\delta\boldsymbol{\Sigma}^c$ is the covariance matrix representing the correlation between the static and delta coefficients. A similar mapping converts the noise parameters $\{\tilde{\mu}^{\Delta c}, \tilde{\boldsymbol{\Sigma}}^{\Delta c}\}$ to $\{\tilde{\mu}^{\Delta l}, \tilde{\boldsymbol{\Sigma}}^{\Delta l}\}$. An ML estimate of the delta parameters may then be obtained using numerical integration [12]. The form of the delta parameters given by equation 11 is similar to that of the static parameters. Hence, it is possible to use the same approximation to calculate values of the delta parameters. For the delta parameters two Gaussian distributions are estimated and then combined. However, the estimation of the covariance matrix will be poor as the two Gaussians are correlated, $S_i^l(t-w)$ and $N_i^l(t-w)$ appear in both expressions.

The 'mis-match' function given in equation 11 has been derived for difference coefficients. It is simple to extend this to acceleration or delta-delta parameters, $\boldsymbol{\Delta}^2\mathbf{O}^c(t)$, where simple differences of delta parameters are calculated, so

$$\boldsymbol{\Delta}^2\mathbf{O}^c(t) = \boldsymbol{\Delta}\mathbf{O}^c(t+w_a) - \boldsymbol{\Delta}\mathbf{O}^c(t-w_a) \tag{14}$$

where w_a is the acceleration difference offset. Thus the same form of the 'mis-match' function as with the delta parameters may be used. If linear regression coefficients are used to calculate the delta and delta-delta parameters no simple 'mis-match' function is possible and the technique described here cannot be used to compensate the delta and higher order parameters.

2.3 Spectral Tilt Estimation

In situations where there is channel distortion as well as additive noise, the corrupted observations are described by [13]

$$O_i(t) = H_i S_i(t) + N_i(t) \tag{15}$$

where $\mathbf{H}$ represents the channel difference between training and testing. Note that $\mathbf{H}$ now subsumes the gain term g used in equation 1. Expressing this equation in the same form as equation 1 yields

$$O_i^l(t) = \mathcal{F}(S_i^l(t), N_i^l(t), H_i^l) = \log(\exp(S_i^l(t) + H_i^l) + \exp(N_i^l(t))) \tag{16}$$

This expression assumes that the noise model statistics are obtained in the test channel conditions.

Similar to the standard PMC function, equation 16 has been obtained in terms of the clean speech and noise, but it also includes the channel difference $\mathbf{H}^l$. There is now therefore the additional problem of obtaining the statistics of $\mathbf{H}^l$. At run time, the transcription of each utterance is unknown. Hence, it is preferable to estimate $\mathbf{H}^l$ assuming no knowledge of the transcription. To achieve this, a single state, single component model of the speech corrupted by additive and convolutional noise is computed. As all frames of the corrupted speech are assumed to belong to this distribution, by altering $\mathbf{H}^l$ and minimising an appropriate distance measure between the corrupted speech and the corrupted speech model, it is possible to estimate the spectral tilt without a transcription of the corrupted speech. An inherent assumption in this estimation process is that the set of acoustic vectors on which the clean model is trained is representative of the acoustic vectors observed in the test channel conditions, after being corrupted by convolutional and additive noise. As only a single state corrupted speech model is to be estimated, it is not necessary to use models of individual words or phones, provided that a good model of all the acoustic data is available. Hence, a global single-state clean speech and silence model, $\mathcal{M}_G$, is combined with a single state noise model to yield the single state corrupted speech model. The statistics used to estimate the tilt are therefore:

1. the global speech and silence model, $\mathcal{M}_G$;
2. the additive noise model measured down the test channel, $\mathcal{M}_N$;
3. the corrupted speech mean measured down the test channel, $\overline{\mathbf{O}}^c$.

If a multiple state noise model is required to deal with non-stationary noise, it must be mapped to a single state model to estimate the channel difference. For an ergodic noise model, this involves simply using the normalised expected durations for the weights of the mixture components associated with the state.

There are several possible optimisation criteria which may be used to estimate the difference in channel conditions. Here a Euclidean distance is used and the spectral tilt $\mathbf{H}^l$ is found from the following optimisation

$$\mathbf{H}^l \quad = \quad \arg\min\left[\left(\hat{\mu}^l(\mathbf{H}^l) - \overline{\mathbf{O}}^l\right)^T \left(\hat{\mu}^l(\mathbf{H}^l) - \overline{\mathbf{O}}^l\right)\right] \tag{17}$$

where $\overline{\mathbf{O}}^l$ is the noise global mean and

$$\hat{\mu}^l(\mathbf{H}^l) = \int\limits_{\mathcal{R}^n}\int\limits_{\mathcal{R}^n} \log\left(\exp(S_i^l + H_i^l) + \exp(N_i^l)\right) p(\mathbf{S}^l)p(\mathbf{N}^l)\, d\mathbf{S}^l\, d\mathbf{N}^l \tag{18}$$

where $\hat{\mu}^l(\mathbf{H}^l)$ is the estimated mean of the corrupted speech. The statistics for the noise and speech, $p(\mathbf{N}^l)$ and $p(\mathbf{S}^l)$, are obtained from the models, $\mathcal{M}_N$ and $\mathcal{M}_G$ respectively. As with standard PMC, it is assumed that the global model $\mathcal{M}_G$ accurately models the output probability distribution of the training data. In order to achieve this modelling a mixture Gaussian distribution may be used. If no additive noise is present, then this extended PMC compensation scheme becomes equivalent to the standard cepstral mean subtraction scheme used by many current systems.

2.4 Summary of the Noise Robust System

The complete procedure used to achieve noise robustness in additive and convolutional noise may be summarised as follows:

1. estimate a set of 'clean' speech models, thus obtaining $\{\mu^c, \mathbf{\Sigma}^c\}$;
2. estimate the noise in the test channel condition, $\{\tilde{\mu}^c, \tilde{\mathbf{\Sigma}}^c\}$;
3. estimate the spectral tilt difference, $\mathbf{H}^l$;
4. calculate the new set of models for the noise corrupted speech, $\{\hat{\mu}^c, \hat{\mathbf{\Sigma}}^c\}$.

If desired, a-priori information such as knowledge that there is no spectral tilt, or that the spectral tilt is smooth may be incorporated.

3 Database

The database used for the clean speech models was the ARPA Resource Management database (RM) [18]. This is a 1000 word task with a vocabulary based on a naval resource management domain. There are 3990 training sentences and a set of four 300 sentence test sets, of which three

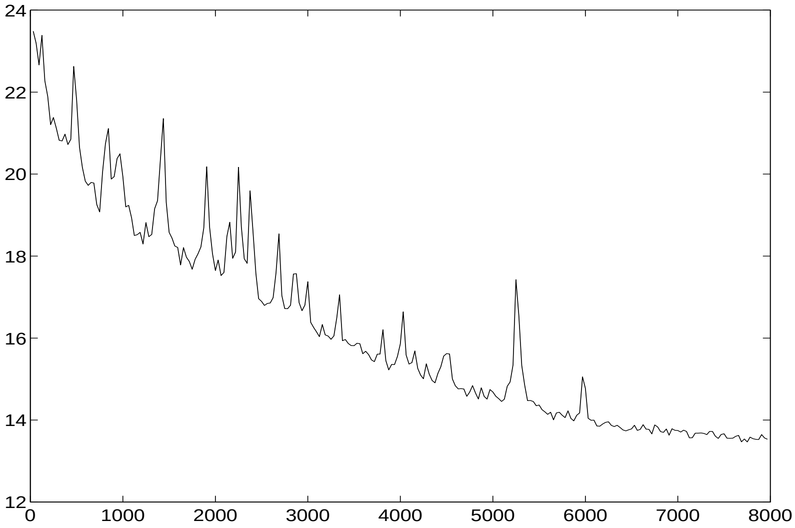

Figure 2: Log Power Spectral Density against Frequency of Lynx Helicopter Noise taken from the NOISEX-92 database

are considered in this work, the February 1989, October 1989 and February 1991 ARPA evaluation sets. The recordings were made in a sound isolated recording booth, yielding a Signal to Noise Ratio, SNR, of > 40dB.

The second database used was the NOISEX-92 database [20]. This database contains various noise sources of which the Lynx Helicopter noise was used here. The power spectral density of the noise source is shown in figure 2.

Test Set	Clean SNR (dB)	Noisy SNR (dB)
Feb'89	50.4	19.6
Oct'89	48.5	19.0
Feb'91	48.6	20.3

Table 1: Average SNR for the RM Test Sets adding Lynx Helicopter Data scaled by 0.1

The Lynx helicopter noise was added to the clean RM data to give a SNR of approximately 20dB. The exact SNRs are given in table 1. The mis-match between the clean and the noisy conditions can be seen to be about 30dB. To obtain alternative SNR values the noise was scaled accordingly and added.

4 Recognition System

The baseline speech recognition system was built using the Resource Management Toolkit distributed with HTK Version 1.5 [22] and used for the November 1992 ARPA Evaluation [21]. A state clustered triphone system was built consisting of 1805 tied states forming 1959 distinct triphones. Two modifications to the parameterisation used in the November 1992 ARPA evaluation were made. In order to use PMC to compensate a set of models, it is necessary to have the zeroth cepstra in the parametrisation. The standard HTK system uses a normalised log energy and is thus unsuitable for use with PMC. In addition to modifying the energy term the method for calculating the delta and delta-delta parameters was altered. As previously mentioned it is not possible to 'correctly' compensate delta parameters when they are calculated using Linear Regression. HTK was therefore extended to allow for the use of simple difference expressions to calculate the deltas and delta-deltas.

The data was preprocessed using a 25 msec Hamming window and a 10 msec frame period. Additionally the data was pre-emphasised with a factor of 0.97 and liftered with a factor of 22. 24 Log-energy filter banks were used to generate the basic acoustic analysis using an observation vector consisting of 13 Mel frequency cepstral coefficients, including the zeroth cepstra. In addition to the static parameters delta and delta-delta parameters were added, the deltas and delta-deltas being calculated using

$$\begin{aligned} \mathbf{\Delta O}^c(t) &= \frac{1}{2w}\left(\mathbf{O}^c(t+w) - \mathbf{O}^c(t-w)\right) &(19) \\ \mathbf{\Delta}^2\mathbf{O}^c(t) &= \frac{1}{2w_a}\left(\mathbf{\Delta O}^c(t+w_a) - \mathbf{\Delta O}^c(t-w_a)\right) &(20) \end{aligned}$$

where in these experiments, $w = 2$ and $w_a = 2$. As all the cepstral parameters are liftered and the delta and delta-delta parameters are scaled, it is necessary to inverse lifter and rescale to achieve the distributions required for compensation. In addition the mapping of 24 Log-energy values down to 13 cepstral parameters means that when performing the inverse cosine transform it is necessary to zero pad the feature vector and similarly when mapping back it is necessary to truncate the cepstral vector to 13 again. In order to add spectral tilt to the observations for the experiments with convolutional noise, the normal pre-emphasis was suppressed when parameterising the test data. This causes the low frequency components of the corrupted speech to be amplified and the high frequency components attenuated compared to the clean speech used for training

5 Results

Energy Term	Delta Terms	Num. Mix.	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
Log	Regr.	1	93.7	4.3	2.1	0.9	7.2	35.3
Cepstra	Regr.		93.9	4.1	2.0	0.7	6.8	37.7
Cepstra	Diff.		93.0	4.6	2.3	1.0	8.0	40.3
Log	Regr.	2	94.6	3.5	1.6	0.5	5.9	32.0
Cepstra	Regr.		94.8	3.2	2.0	0.5	5.7	32.3
Cepstra	Diff.		94.6	3.4	2.0	0.7	6.1	32.7

Table 2: Performance on Clean RM Feb'89 Test Set for Various Parametrisations

As noted above, in order to use PMC to compensate the models it is necessary to alter the parameterisation of the speech compared with that used in the standard HTK RM recogniser. To test the effect of this, a series of systems were built using various parametrisations and tested on the February 1989 test set. The results are shown in table 2. A slight difference in performance may be seen, particularly for the single mixture component model case, however these differences are greatly reduced in the two mixture component case. From these results, it appears that there is no significant difference in performance between using Log-energy with Linear regression as used in the standard system and the Zeroth Cepstra with simple differences as used here.

5.1 Additive Noise Only

5.1.1 Single Component System

A single-component state-clustered recognition system was built and tested on the three clean test sets. The results are shown in table 3. These results represent the best possible performance achievable by an enhancement scheme, though this could only be achievable by a perfect scheme with no additional estimation variance.

Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
Feb'89	93.0	4.6	2.3	1.0	8.0	40.3
Oct'89	92.3	5.5	2.3	1.3	9.0	41.3
Feb'91	93.8	4.6	1.7	1.3	7.5	33.7

Table 3: Performance of a Single Component System on Clean RM

Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
Feb'89	88.6	8.9	2.5	2.3	13.7	53.0
Oct'89	90.5	7.0	2.5	2.1	11.6	44.3
Feb'91	90.8	7.9	1.3	1.9	11.2	42.7

Table 4: Single Component System Trained on Additive Noise Corrupted RM at 18-20dB

A more realistic upper bound on performance for a model compensation scheme such as PMC is obtained by training and testing in the same environment. A set of models were therefore trained on the noisy data and the results are shown in table 4. As can be seen, the noise degrades the performance, even at relatively high SNR.

Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
Feb'89	88.9	8.5	2.6	1.6	12.6	53.0
Oct'89	90.7	6.8	2.5	1.4	10.7	43.3
Feb'91	90.3	7.9	1.9	1.9	11.6	45.3

Table 5: Performance of a Single Component System Trained on Additive Noise at 18-20dB Corrupted RM using Complete Dataset of the Clean System

One of the assumptions behind PMC is the that state frame allocation (ie the complete dataset) does not alter when noise is added. To test this assumption, a set of HMMs were trained on the noise corrupted data using the complete dataset of the clean system. The performance on the various test sets is shown in table 5. Comparing these results with those in table 4, shows there is no significant difference in performance. These figures represent the upper bound on the performance of model based schemes, which do not alter the complete dataset, such as the implementation of PMC used here.

Table 6 shows the performance of various PMC compensation schemes. The first set of results is for the uncompensated system where the performance is very poor. Compensating just the static parameters improves the performance, for example reducing the word error rate on the Feb'89 task from 55.8% to 23.1%, a reduction of 59%. Compensating all the parameters, static, deltas and delta-deltas, further reduces the error rate to 15.2%, indicating that the delta parameters are not inherently noise robust as is often assumed.

The previous set of results used the assumption that the statistics for the data were approximately the same for time t and time $t-2$. To test this, a set of models were built using the same complete dataset, but accumulating the statistics for time $t-2$. The models were then compensated using these $t-2$ models to estimate the delta and delta-delta parameters. The results are shown in table 7. Using the $t-2$ models further improves the performance to 14.0% on the Feb'89 task, giving an overall 75% reduction in word error rate compared to the uncompensated system. Comparing these results with those obtained training the models on noise corrupted training data,

Comp Param	Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
—	Feb'89	57.9	37.6	4.5	13.7	55.8	92.6
	Oct'89	67.0	28.9	4.1	13.6	46.5	87.3
	Feb'91	63.3	33.1	3.7	15.5	52.3	91.9
O	Feb'89	80.8	14.8	4.3	3.9	23.1	67.3
	Oct'89	85.0	11.3	3.7	2.5	17.4	55.0
	Feb'91	83.5	13.6	2.8	3.7	20.2	55.3
O	Feb'89	86.5	10.1	3.4	1.7	15.2	57.0
ΔO	Oct'89	89.3	7.8	2.9	1.3	12.0	47.0
$\Delta^2 O$	Feb'91	87.8	9.7	2.5	2.2	14.3	51.3

Table 6: Comparison of No Compensation, Compensating only Static Parameters and Compensating all the Parameters of a Single Component System on Additive Noise Corrupted RM at 18-20dB

Comp Param	Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
O	Feb'89	87.6	9.3	3.2	1.6	14.0	56.0
ΔO	Oct'89	89.7	7.5	2.8	1.3	11.5	45.7
$\Delta^2 O$	Feb'91	88.7	9.2	2.1	1.9	13.2	48.3

Table 7: Performance of a Single Component System on Additive Noise Corrupted RM at 18-20dB using Models Generated using $t-2$ Statistics to Compensate the Delta and Delta-Delta Parameters

table 4 and table 5, shows little degradation in performance. Hence it is possible to obtain a set of models for noise corrupted speech without having to re-parameterise the whole training database and re-estimate the models, a computationally expensive task.

Comp Param	Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
O	Feb'89	87.5	9.3	3.2	1.6	14.0	56.0
ΔO	Oct'89	89.6	7.6	2.8	1.3	11.7	46.0
$\Delta^2 O$	Feb'91	88.7	9.1	2.1	1.9	13.2	48.3

Table 8: Performance of a Single Component System on Additive Noise Corrupted RM at 18-20dB using Models Generated using $t-2$ Statistics to Compensate the Delta and Delta-Delta Parameters using a Reduced Number of Points for the Numerical Integration

All the results so far described used a 20 point numerical integration to obtain the new means and variances. In order to reduce the computational overhead this was reduced to 10 points for the one dimensional integrations required for the means and 6 points for the two dimensional numerical integrations required for the variances. Only six points were used in the latter cases, since it was felt that the performance relies more on an accurate estimate of the means rather than an accurate estimate of the variances. The results of this system using the $t-2$ statistic models for the delta compensation are shown in table 8. As shown, reducing the number of points in the numerical integration does not significantly affect the performance. All remaining results are therefore calculated using 10 points for the one dimensional integration cases and 6 points for the two dimensional numerical integration cases.

An alternative to performing numerical integration is to use a log-normal approximation, as previously used on the small vocabulary systems. Just compensating the static parameters using the log-normal approximation, gives a word error rate of 26.3% on the Feb'89 test set compared to a word error rate of 23.1% for numerical integration. This agrees with the previous small vocabulary results [13], that the use of the Log-normal approximation slightly degrades the performance. The means of the delta and delta-delta parameters were then also compensated using the fast approximation, achieving 23.7% word error rate. Thus, compensating the means and variances of the delta and delta-delta parameters using the Log-normal approximation degrades the performance. This is due to the correlation of the two Gaussians as described in section 2.2. The best performance achieved using the fast compensation was thus only comparable with that achieved using numerical integration to compensate just the static parameters.

5.1.2 Multiple Component Systems

Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
Feb'89	94.6	3.4	2.0	0.7	6.1	32.7
Oct'89	93.6	4.3	2.1	0.8	7.2	34.0
Feb'91	95.2	3.5	1.4	0.7	5.6	27.7

Table 9: Two Component Performance on Clean RM

The single component state clustered triphone system was then 'mixed up' to a two component system using the standard HTK toolkit HMM editor [22]. The performance on the clean RM data is shown in table 9.

Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
Feb'89	90.8	6.8	2.4	1.2	10.5	45.3
Oct'89	91.5	6.1	2.3	1.3	9.8	40.3
Feb'91	92.5	6.3	1.2	1.4	8.9	37.7

Table 10: Performance of a Two Component System Trained on Noise Corrupted RM at 18-20dB

The models trained on the noise corrupted data were also 'mixed up' and evaluated on the various test sets. The results are shown in table 10.

The performance for various two component systems is shown in table 11. For the uncompensated models trained on the clean data, a pruning error occurred for one sentence, so the result quoted for the Feb'91 task is for 299 sentences. The *Noisy* model set was generated using the complete dataset of the clean data and noise corrupted training data. The *PMC* model set used the $t-2$ model set to compensate the delta and delta-delta parameters. Again PMC yields good performance. Comparing the performance of the *Noisy* models and the *PMC* models with the model trained on the noise corrupted data, table 10, shows that there was no degradation for the *Noisy* models and only slight degradation for the *PMC* models. The assumption that the frame state/component allocation is not altered by the addition of noise appears to be justifiable.

The effect of not compensating the variances of a set of models was then examined. If the variances were not compensated on the two component models the word error rate on the Feb'89 task was 20.1% compared with 11.2% when all parameters were compensated. Therefore it is clearly important to compensate the variances of the models to achieve good performance.

A set of six component models, similar to those used for the ARPA RM system developed at CUED [21], were then generated. Table 12 shows the performance of this six component system on the clean RM test sets. Comparing these results with those published using the standard HTK parameterisation [21] shows no significant difference in performance, for example on the Feb'89

Model Set	Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
Clean	Feb'89	66.1	29.4	4.6	9.1	43.1	84.0
	Oct'89	75.0	21.1	4.0	8.5	33.5	78.3
	Feb'91	71.9	24.1	4.0	9.8	37.8	76.6
Noisy	Feb'89	90.7	6.8	2.5	0.9	10.2	47.3
	Oct'89	91.8	6.2	2.0	0.9	9.1	38.3
	Feb'91	92.6	6.0	1.4	1.4	8.8	37.0
PMC	Feb'89	89.9	7.4	2.7	1.2	11.2	48.7
	Oct'89	91.2	6.5	2.3	1.1	9.9	40.7
	Feb'91	91.6	6.9	1.4	1.2	9.5	40.3

Table 11: Comparison for a Two Component System of Uncompensated Models, Models trained on the Noise Corrupted Data using the Complete Data Set of the Clean Models and a Set of Compensated Models, Compensating all the Parameters on Additive Noise Corrupted RM at 18-20dB

Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
Feb'89	96.1	2.3	1.5	0.4	4.3	25.3
Oct'89	95.3	3.1	1.7	0.6	5.4	28.0
Feb'91	96.7	2.4	1.0	0.6	3.9	21.3

Table 12: Performance of a Six Component System on Clean RM

test set using simple differences and the zeroth cepstra the word error rate was 4.3% compared to 4.5% using log-energy and linear regression parameters.

The six component system was then compensated. A comparison of the performance of these compensated models (*PMC*) with uncompensated models (*Clean*) is shown in table 13. The compensated models are more noise robust, for example compensating the models has reduced the word error rate of the Feb'89 test set by 78.3%, from 38.7% to 8.3% word error rate. Comparing these results with training on the noise corrupted data using the clean complete data set (*Noisy*), shows little difference in performance, 7.3% word error rate compared to 8.3% word error rate, though again the performance of PMC is slightly worse.

The SNR was then decreased by 8dB and a new set of models generated. The results are shown in table 14. Even at 10-12dB the performance of the system is still good, 22.8% word error rate, compared to 86.6% word error rate when no compensation is applied, showing that PMC works even at poor SNRs.

5.2 Additive and Convolutional Noise

Convolutional noise was incorporated into the system, by removing the pre-emphasis on the test data. This results in the corrupted speech being attenuated at high frequencies and amplified at low frequencies compared to the clean speech. The six component system was then tested on RM with this additive and convolutional noise present. No *a-priori* knowledge, such as smoothness of the spectral difference was made about the convolutional noise, only that it was constant with time. The noise model used was trained under the new spectral tilt condition, as this model is expected to be generated *on-line*. A ten component global speech and silence model and cepstral mean parameters estimated using all the Feb'89 test set were used to calculate the spectral tilt difference. Table 15 shows the results. If no account is made of the convolutional noise it degrades performance, the word error rate rising from 8.3% to 15.2% on the Feb'89 test set. When the

Model Set	Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
Clean	Feb'89	71.2	25.3	3.6	9.9	38.7	80.0
	Oct'89	77.6	19.1	3.3	9.5	32.0	74.3
	Feb'91	75.5	21.9	2.6	8.9	33.4	72.3
Noisy	Feb'89	93.4	4.6	2.0	0.8	7.3	38.0
	Oct'89	92.3	5.7	2.0	0.9	8.6	36.3
	Feb'91	94.5	4.4	1.0	1.4	6.9	30.0
PMC	Feb'89	92.6	5.2	2.2	0.9	8.3	41.7
	Oct'89	92.8	5.1	2.0	0.9	8.1	36.3
	Feb'91	93.8	5.1	1.1	1.1	7.3	34.0

Table 13: Comparison of the Performance of Clean, Compensating all the parameters and training on Noisy Data given the Clean Complete Data Set for a Six Component System on Additive Noise Corrupted RM at 18-20dB

Approx SNR	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
50	96.1	2.3	1.5	0.4	4.3	25.3
20	92.6	5.2	2.2	0.9	8.3	41.7
12	79.0	15.5	5.5	1.9	22.8	67.0

Table 14: Performance of a Six Component System, Compensating all the Parameters, on RM Feb'89 Test Set at Various SNR

Test Set	Spec Tilt	Comp Tilt	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
Feb'89	No	No	92.6	5.2	2.2	0.9	8.3	41.7
	Yes	No	87.0	10.1	2.9	2.2	15.2	57.7
	Yes	Yes	92.2	5.7	2.1	0.9	8.7	42.3
Oct'89	No	No	92.8	5.1	2.0	0.9	8.1	36.3
	Yes	No	88.6	8.8	2.6	2.2	13.6	50.3
	Yes	Yes	92.5	5.6	1.9	1.2	8.7	36.0
Feb'91	No	No	93.8	5.1	1.1	1.1	7.3	34.0
	Yes	No	90.1	8.1	1.8	2.1	12.0	47.0
	Yes	Yes	91.7	7.1	1.2	1.8	10.1	38.0

Table 15: Comparison of the performance with and without Convolutional Noise on Additive Noise Corrupted RM at 18-20dB

spectral tilt is estimated using the extended version of PMC the error rate decreases back to 8.7%, only a slight degradation compared to the no spectral tilt case. It is worth noting that for simple difference parameters, as described here, the delta-delta parameters are robust to spectral tilt in the presence of additive noise. However delta parameters, which are robust to spectral tilt in quiet environments are not robust when additive noise is present. The degradation in performance of the Feb'91 test set due to the convolutional noise was greater than for the other test sets. Examining the results shows that speaker *ALK0* had a word error rate of 37.2% under additive and convolutional noise conditions, compared to 21.2% on additive noise conditions. Looking at the cepstral means associated with this speaker, shows a significantly different shift compared to the other speakers in the test set. Using the cepstral means from speaker *ALK0* to estimate the spectral tilt differences, reduced the error rate on *ALK0* to 21.9%, though the overall error rate on the Feb'91 test set rises to 11.2%. It would therefore be preferable to be able to adapt the models to variations in the values of the cepstral mean estimates, for example at the sentence level. This problem will be addressed in future work.

Spec Tilt	Comp Tilt	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
No	No	79.0	15.5	5.5	1.9	22.8	67.0
Yes	Yes	81.1	15.3	3.6	3.6	22.6	67.3

Table 16: Comparison of the performance with and without Convolutional Noise on Additive Noise Corrupted RM at 10-12dB on the Feb'89 Test Set

The SNR was then decreased to 10-12dB and the performance of the six component system with convolutional noise was examined. Table 16 compares the performance with and without convolutional noise for this case. Again little degradation in performance is observed when convolutional noise is present.

6 Conclusions

This paper has described the application of Parallel Model Combination (PMC) to a medium size vocabulary speech recognition task. Using this larger vocabulary system has highlighted the need to compensate the delta and delta-delta parameters to achieve good performance. For example a single component system achieves 23.1% word error rate on the Feb'89 task at 18-20dB when only the static parameters are compensated. Whereas compensating all the parameters, reduces the word error rate to 14.0% on the same task, a reduction of 39% in the word error rate. In order to effectively compensate the delta parameters, it is better to use an additional model set based on the statistics at time $t-2$ to compensate the delta and delta-delta parameters. This was found to reduce the word error rate by about 8% on a single component Feb'89 task at 18-20dB. PMC has also been shown to work at lower SNRs, 10-12dB, showing that it can be useful as a compensation technique at poor SNRs.

An extended form of PMC which allows an explicit estimate of the spectral tilt difference to be made, even in the presence of additive interfering noise, was also investigated. For additive noise at 18-20dB and spectral tilt caused by not pre-emphasising the test data the system achieved a word error rate of 8.7% compared to 8.3% when no spectral tilt was present. This compares with 15.2% word error rate when only the additive noise is compensated for using PMC.

In this paper a series of experiments have shown that PMC may be used to achieve robust speech recognition on a medium vocabulary task under artificially corrupted convolutional and additive noise conditions. The performance under real conditions has not presently been evaluated. However, given that very few assumptions are made and that no tuning of the algorithms is required for specific noise conditions, it is felt that similar improvements in performance will be obtained under real test environments.

Acknowledgement

M. Gales is funded by a SERC studentship and a CASE award with DRA Malvern.

References

[1] A Acero and R M Stern. Robust speech recognition by normalization of the acoustic space. In *Proceedings ICASSP*, pages 893–896, 1991.

[2] V L Beattie and S J Young. Noisy speech recognition using hidden Markov model state based filtering. In *Proceedings ICASSP*, pages 917–920, 1991.

[3] V L Beattie and S J Young. Hidden Markov model state-based cepstral noise compensation. In *Proceedings ICSLP*, pages 519–522, 1992.

[4] A D Berstein and I D Shallom. An hypothesized Wiener filtering approach to noisy speech recognition. In *Proceedings ICASSP*, pages 913–916, 1991.

[5] S F Boll. Suppression of acoustic noise in speech using spectral subtraction. In *IEEE Transactions ASSP*, volume 27, pages 113–120, 1979.

[6] H M Cung and Y Normandin. Noise adaptation algorithms for robust speech recognition. In *Proceedings ESCA Workshop on Speech Processing in Adverse Conditions*, pages 171–174, 1992.

[7] A Erell and M Weintraub. Filterbank-energy estimation using mixture and markov models for recognition of noisy speech. *IEEE Transactions SAP*, pages 68–76, 1993.

[8] S Furui. Toward robust speech recognition under adverse conditions. In *Proceedings ESCA Workshop in Speech Processing in Adverse Conditions*, pages 31–42, nov 1992.

[9] L Gagnon. A noise reduction approach for non-stationary additive interference. *Proceedings ESCA Workshop in Speech Processing in Adverse Conditions*, pages 139–142, 1992.

[10] M J F Gales and S J Young. An improved approach to the hidden Markov model decomposition of speech and noise. In *Proceedings ICASSP*, pages 233–236, 1992.

[11] M J F Gales and S J Young. Cepstral parameter compensation for HMM recognition in noise. *Speech Communication*, 12:231–240, 1993.

[12] M J F Gales and S J Young. Parallel model combination for speech recognition in noise. *Technical Report CUED/F-INFENG/TR135*, 1993.

[13] M J F Gales and S J Young. PMC for speech recognition in additive and convolutional noise. *Technical Report CUED/F-INFENG/TR154*, 1993.

[14] D H Klatt. A digital filterbank for spectral matching. In *Proceedings ICASSP*, pages 573–576, 1979.

[15] P Lockwood and J Boudy. Experiments with a non linear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars. In *Proceedings Eurospeech*, pages 79–82, 1991.

[16] D Mansour and B H Juang. The short-time modified coherence representation and noisy speech recognition. In *IEEE Transactions ASSP*, volume 37, pages 795–804, 1989.

[17] B A Mellor and A P Varga. Noise masking in the MFCC domain for the recognition of speech in background noise. In *Proceedings IOA*, volume 14, pages 503–510, 1992.

[18] 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*, pages 651–654, 1988.

[19] A P Varga and R K Moore. Hidden Markov model decomposition of speech and noise. In *Proceedings ICASSP*, pages 845–848, 1990.

[20] 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. In *Technical Report, DRA Speech Research Unit*, 1992.

[21] P C Woodland and S J Young. The HTK tied-state continuous speech recogniser. In *Proceedings Eurospeech*, pages 2207–2210, 1993.

[22] S J Young, P C Woodland, and W J Byrne. *HTK: Hidden Markov Model Toolkit V1.5.* Entropic Research Laboratories Inc., 1993.