ROBUST CONTINUOUS SPEECH RECOGNITION USING PARALLEL MODEL COMBINATION

M. J. F. Gales & S. J. Young
EDICS Number: SA 1.6.8
Cambridge University Engineering Department, Trumpington Street
Cambridge CB2 1PZ, England
Telephone (+44) 223 332754, Fax (+44) 223 332662
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. Most experiments reported on PMC to date have been on small, 10-50 word vocabulary systems. Experiments on the Resource Management (RM) database, a 1000 word continuous speech recognition task, reveal compensation requirements not highlighted by the smaller vocabulary tasks. In particular, that it is necessary to compensate the dynamic parameters as well as the static parameters to achieve good recognition performance.

The database used for these experiments was the RM speaker independent task with either Lynx Helicopter noise or Operation Room 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,the performance of the recogniser was found to be severely degraded when noise was added to the speech signal at between 10dB and 18dB. However, using PMC the performance was restored to a level comparable with that obtained when training directly in the noise corrupted environment.

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 [14]. 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, 16], spectral mapping [6], and inherently robust parameterisations [17]. These methods only use statistical information about the interfering noise in the compensation process. No account is taken of what was said. Other schemes have also attempted to estimate the clean speech signal using information about the speech. These include inhomogeneous estimators using HMMs [8] 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 [15, 18], state based filtering [2], cepstral mean compensation [3, 4], HMM decomposition [21], and Parallel Model Combination (PMC) [9, 10].

This paper is concerned with the latter approach to noise robustness, in particular 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 mismatch, either in the form of additive noise, or variations in the channel conditions. Some method for compensating the parameters of the models is therefore required. PMC is a method for compensating the parameters in a computationally efficient manner. 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 first decision to be made in a model-based scheme is the form of the corrupted-speech model. In PMC, the model used is a standard HMM with Gaussian output probability distributions. This model therefore requires no modification of the recognition software and allows the standard HMM estimation formulae to be applied. The disadvantage of using Gaussian distributions is that the corrupted speech distributions are known to be non-Gaussian [19]. This problem may be overcome by using multiple component Gaussian distributions to model the corrupted speech distribution [12]. However, for this work the simpler assumption is made that using a single Gaussian to model the corrupted speech distribution will yield sufficiently good results.

Having decided on the form of the models, it is necessary to choose an approach to estimate the new model parameters. The best technique for additive noise, in terms of building a matched system, would be to add samples of the background noise to the clean training data at the waveform level and retrain the models. This will normally be impractical for a large database. However, if the clean speech models are assumed to contain sufficient information about the statistics of the training data, they may be used in the compensation scheme in place of the data itself. Moreover, a model of the background noise can be generated using whatever noise samples are available to represent the background noise conditions. The problem then is to find a method of combining the two models to accurately estimate the corrupted-speech models.

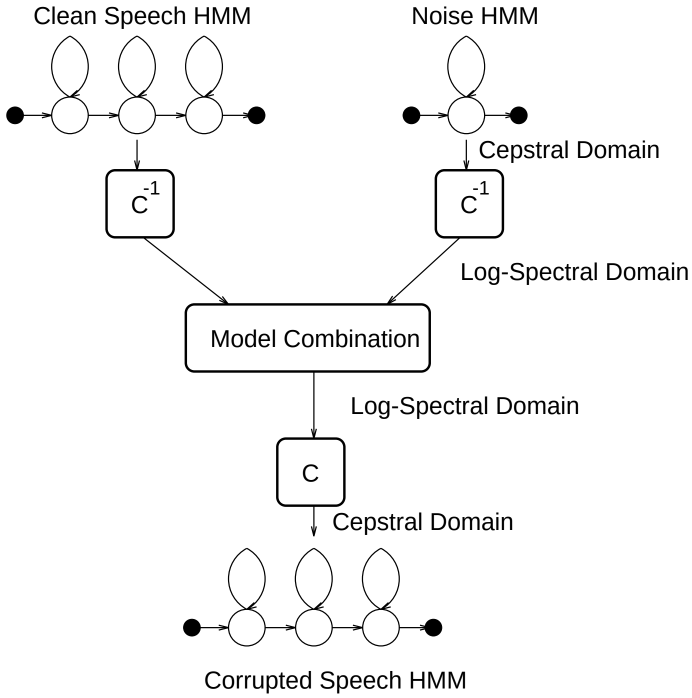


Figure 1: The basic parallel model combination process

The basic PMC process is illustrated in figure 1. The inputs to the scheme are clean speech models and a noise model. Since the combination of the speech and noise is most naturally expressed in the Linear-Spectral or Log-Spectral domains, it is simplest to model the effects of the additive noise on the speech parameters in one of these domains. The function that approximates this will be referred to as the *mismatch function*. If Cepstral parameters are used, the model parameters must be transformed into the appropriate combination domain (Log-Spectral in figure 1). The clean speech models and the noise model are then combined according to this mismatch function. After the models have been combined, the estimate of the corrupted-speech model is transformed back into the Cepstral domain, if required.

3 The Mismatch Function

The first stage in a model combination process is to produce some function describing how noise affects each of the speech parameters. In order to describe the effects of the noise on the clean speech parameters, a series of assumptions are required:

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 multiple Gaussian mixture component model contains sufficient information to represent the distribution of the observation vectors in the Cepstral or Log-Spectral domain.

4. The frame/state alignment used to generate the speech models from the clean speech data is not altered by the addition of noise.

Having made these assumptions it is possible to write expressions for the effects of the additive noise on each of the parameters of the feature vector. In order to use these expressions in a compensation process, it is necessary to have statistics for all the variables and the various correlations between the statistics must be adequately described.

3.1 Mismatch Function for the Static Parameters in Additive Noise

The static noise-corrupted speech "observations" in the Log-Spectral domain are given by the mismatch function

$$\begin{aligned} O_i^l(\tau) &= \mathcal{F}(S_i^l(\tau), N_i^l(\tau)) \\ &= \log(g\exp(S_i^l(\tau)) + \exp(N_i^l(\tau))) \end{aligned} \quad (1)$$

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 [10].

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

$$\hat{\mu}_i^l = \mathcal{E}\left\{\mathcal{F}(S_i^l, N_i^l)\right\} \quad (2)$$

and

$$\hat{\Sigma}_{ij}^l = \mathcal{E}\left\{\mathcal{F}(S_i^l, N_i^l)\mathcal{F}(S_j^l, N_j^l)\right\} - \hat{\mu}_i^l\hat{\mu}_j^l \quad (3)$$

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.

3.2 Mismatch Function for Dynamic Parameters in Additive Noise

For medium to large vocabulary speech recognition it is necessary to incorporate dynamic coefficients in the parameter set to achieve good performance. These parameters are added in an attempt to model the correlation between successive frames. If the dynamic coefficients are calculated using simple differences over a given window width, w for the deltas and w_a for the delta-delta parameters, then the observation feature vector will be

$$\mathbf{O}^{\Delta^2 c}(\tau)^T = \left[\begin{array}{ccc} \mathbf{O}^c(\tau)^T & \mathbf{\Delta O}^c(\tau)^T & \mathbf{\Delta}^2\mathbf{O}^c(\tau)^T \end{array} \right]^T \tag{4}$$

where

$$\mathbf{\Delta O}^c(\tau) = (\mathbf{O}^c(\tau + w) - \mathbf{O}^c(\tau - w)) \tag{5}$$

and

$$\mathbf{\Delta}^2\mathbf{O}^c(\tau) = \mathbf{\Delta O}^c(\tau + w_a) - \mathbf{\Delta O}^c(\tau - w_a) \tag{6}$$

are the delta and delta-delta parameters respectively. If statistics about the correlation between successive frames are known, for example $\mathcal{E}\{\mathbf{S}^c(\tau+w)\mathbf{S}^c(\tau-w)^T\}$, then statistics, including correlations, exist for the above expressions and the corrupted delta and delta-delta parameters may be estimated using the standard static parameter mismatch function. However, in order to determine the correlation between successive frames it is necessary to use full, or partial full, covariance matrices for the clean speech, with an associated increase in model storage requirements.

An alternative to the very direct modelling given above is to re-express the delta parameters in terms of parameters whose statistics are known and, if necessary, may be assumed to be independent [11]. It may be shown that

$$\begin{aligned} \Delta O_i^l(\tau) &= \mathcal{F}^{\Delta}(S_i^l(\tau - w), N_i^l(\tau - w), \Delta S_i^l(\tau), \Delta N_i^l(\tau)) \\ &= \log\left(\exp(\Delta S_i^l(\tau) + S_i^l(\tau - w) + g^l) + \exp(\Delta N_i^l(\tau) + N_i^l(\tau - w))\right) \\ &\quad - \log\left(\exp(S_i^l(\tau - w) + g^l) + \exp(N_i^l(\tau - w))\right) \end{aligned} \tag{7}$$

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. Note that the expression for the delta coefficient at time τ is dependent on the static coefficients at time $\tau - w$.

The mismatch function given in equation 7 has been derived for delta coefficients. The same form of analysis may be applied to delta-delta parameters. Here, simple differences of delta parameters are calculated. An added constraint is also introduced that $w = w_a$. The mismatch function is of the form

$$\Delta^2 O_i^l(\tau) = \mathcal{F}^{\Delta^2}(S_i^l(\tau), N_i^l(\tau), \Delta S_i^l(\tau - w), \Delta N_i^l(\tau - w), \Delta^2 S_i^l(\tau), \Delta^2 N_i^l(\tau)) \tag{8}$$

this may be shown to be[1]

$$\begin{aligned} \Delta^2 O_i^l(\tau) = \log(&\exp(\Delta^2 S_i^l(\tau) + 2(S_i^l(\tau) - O_i^l(\tau))) + \exp(\Delta^2 N_i^l(\tau) + 2(N_i^l(\tau) - O_i^l(\tau))) + \\ &\exp(\Delta^2 N_i^l(\tau) + \Delta N_i^l(\tau - w) - \Delta S_i^l(\tau - w) + S_i^l(\tau) + N_i^l(\tau) - 2O_i^l(\tau)) + \\ &\exp(\Delta^2 S_i^l(\tau) + \Delta S_i^l(\tau - w) - \Delta N_i^l(\tau - w) + S_i^l(\tau) + N_i^l(\tau) - 2O_i^l(\tau))) \end{aligned} \tag{9}$$

where $O_i^l(\tau)$ is described by the static parameter mismatch function described in equation 1.

Again, by assuming that the frame/component alignment within a state is not altered by the noise, the ML estimates of the corrupted speech dynamic parameters are given by the expected values of the mismatch functions, in the same form as equations 2 and 3. For delta parameters this has the added complication of being a function of both the static and delta parameters. No closed-form solutions are available for these equations.

[1] The use of g has been dropped in this expression to simplify it.

3.3 Mapping from the Cepstral to the Log-Spectral Domain

The mismatch functions described above have assumed that the speech and noise are modelled in the Log-Spectral domain. As the speech is normally parameterised in the cepstral domain, it must be mapped to and from the log spectrum domain. This is simply achieved using the cosine transform, $\mathbf{C}$, and its inverse. For the case where static and delta parameters are used

$$\mu^{\Delta l} = \left[\begin{array}{cc} (\mathbf{C}^{-1}\mu^c)^T & (\mathbf{C}^{-1}\boldsymbol{\Delta}\mu^c)^T \end{array} \right]^T \tag{10}$$

where $\boldsymbol{\Delta}\mu^c$ is the delta parameter mean in the Cepstral domain. The mapping of the covariance matrix is slightly more complex as there are covariance terms relating the static and delta coefficients. If a full Cepstral domain covariance matrix is used then

$$\boldsymbol{\Sigma}^{\Delta l} = \left[\begin{array}{cc} \mathbf{C}^{-1}\boldsymbol{\Sigma}^c(\mathbf{C}^{-1})^T & \mathbf{C}^{-1}\delta\boldsymbol{\Sigma}^c(\mathbf{C}^{-1})^T \\ \mathbf{C}^{-1}(\delta\boldsymbol{\Sigma}^c)^T(\mathbf{C}^{-1})^T & \mathbf{C}^{-1}\boldsymbol{\Delta\Sigma}^c(\mathbf{C}^{-1})^T \end{array} \right] \tag{11}$$

where $\boldsymbol{\Delta\Sigma}^c$ is the covariance matrix of the delta parameters and $\delta\boldsymbol{\Sigma}^c$ is the covariance matrix representing the correlation between the static and delta coefficients. A similar mapping converts the noise parameters in the Cepstral domain, $\{\tilde{\mu}^{\Delta c}, \tilde{\boldsymbol{\Sigma}}^{\Delta c}\}$, to the Log-Spectral domain, $\{\tilde{\mu}^{\Delta l}, \tilde{\boldsymbol{\Sigma}}^{\Delta l}\}$. If diagonal covariance matrices are used for the speech and noise models, the cross correlation terms between the static and delta parameters may be ignored. The extension of this mapping to incorporate delta-delta parameters is straightforward.

When the feature vector in the Cepstral domain is truncated, ie the higher order Cepstral parameters removed, then an exact mapping from the Cepstral domain to the Log-Spectral domain is not possible. For the experiments described here, where such truncation is used, both the mean and variance vectors are zero padded. This has been found to produce little degradation in performance compared to modelling all the Cepstral parameters and reduces both the memory and computational requirements.

4 Estimating the Model Parameters

Even when the component alignment within a state is assumed unaltered by the noise, the expressions for estimating the corrupted speech model parameters do not have closed-form solutions[2]. Various approximations have previously been used. The corrupted-speech parameters may be estimated using *numerical integration* techniques [13]. Though this will yield "good" estimates, it is computationally expensive for dynamic coefficients. Alternatively, the *log-normal* approximation [9] may be used. Unfortunately this technique is only applicable to static parameters compensation. The simplest approach is to assume that the variances have little affect on the estimation of the means and use the *log-add* approximation [12]. This approach is limited as it may not be used to compensate the variance parameters. Finally, the approach used in this work, *Data-driven PMC* (DPMC) [12] may be used[3]. Here "observations" are generated from the speech and noise distributions and are then combined, using the appropriate mis-match function to form a corrupted-speech

[2] This is not a necessary assumption in PMC as described in [12].

[3] Throughout the rest of this paper the term PMC will be used to refer to DPMC.

"observation". Generating a set of these observations and calculating the mean and variance will yield estimates of the corrupted-speech distribution[4]. This technique gives the same results as numerical integration at a far lower computational cost.

So far, nothing has been discussed about the actual statistics of the speech and noise that are stored. All the mismatch functions described may be expressed in terms of five variables for each of the noise and speech sources. For the speech models these are

$$\mathbf{V}^c(\tau) = \left[\begin{array}{ccccc} \mathbf{S}^c(\tau)^T & \mathbf{\Delta S}^c(\tau)^T & \mathbf{\Delta}^2\mathbf{S}^c(\tau)^T & \mathbf{S}^c(\tau-w)^T & \mathbf{\Delta S}^c(\tau-w)^T \end{array} \right]^T \tag{12}$$

The first three are the standard elements stored in a HMM system, referred to as the τ model set. The remaining two are additional elements required for the dynamic coefficient compensation schemes, the $\tau - w$ model set. An important question is what correlations and statistics are required to achieve good compensation performance. Strong correlations are known to exist between the static and delta-delta parameters. Additionally strong correlations exist between the τ and $\tau - w$ model sets. In this paper various correlation approximations are examined.

The best solution would be to model all the correlations. This results in a full covariance matrix based on $\mathbf{V}^c(\tau)$. For many applications this is not feasible. A simplifying assumption is therefore made, that the cosine transform decorrelates the elements of the static parameters. It is now only necessary to model correlations between the static and dynamic parameters for each element of $\mathbf{V}^c(\tau)$, not between elements. This is referred to as the *Extended* covariance approximation. Alternatively, a diagonal covariance approximation may be used, where all elements of $\mathbf{V}^c(\tau)$ may be assumed to be independent. This is referred to as the *Diagonal* covariance approximation. The simplest approximation possible is that the statistics at time τ are the same as at time $\tau - w$. It is therefore only necessary to store the standard HMM parameters, the τ model set. Furthermore the standard assumption of independence between elements of the feature vector is made. This covariance approximation will be referred to as *Standard Diagonal*.

Finally, it is also necessary to decide on the form of covariance matrix for the corrupted speech distribution. Using PMC will result in a full corrupted-speech covariance matrix, even when diagonal covariance speech and noise models are used. If used, this would result in the standard memory and computation problems associated with full covariance matrices. To overcome this problem the corrupted speech output probabilities are diagonalised by setting all the off-diagonal terms to zero.

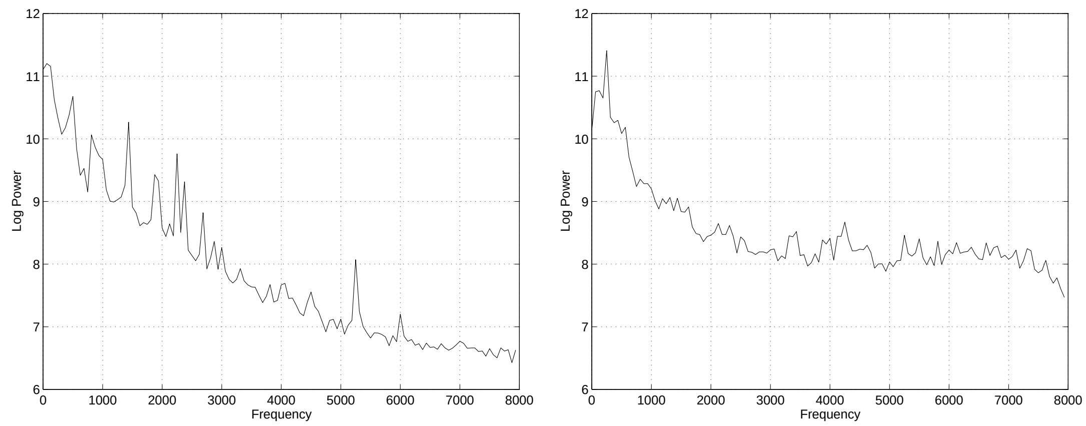
There are two questions that need to be answered in assessing the performance of the compensation system described here. Firstly, how "close" is the estimated model set to the "matched" system trained in the new noise environment? This measures the accuracy of the compensation process. Secondly, given the crude nature of the modelling of the corrupted speech distributions with single Gaussian distributions, is this approximation good enough to achieve acceptable noise robustness? The performance of the system will be compared to that of "matched" systems built on the noise corrupted speech data. Two methods for building matched systems may be considered. The first is to train from "scratch" on the noise corrupted data, this is referred to as the *Multi-Pass* system. The second system is trained assuming that the addition of noise does not change the frame/state component alignment, the *Single-Pass* system. As this single-pass system is

[4] Of course, the accuracy of the estimations will depend on the number of "observations". Though not optimal computationally the sample set generated was sufficiently large to give a small variance on the parameters.

built using the same assumption as the model compensation process, it is the target set of models for the compensation schemes considered. The same performance criteria as in [13] will be used, the average KL number between the single-pass system and the PMC system, and the more standard (%) word error rate. The use of the average KL number allows the performance of the compensation scheme on a per component level to be examined, which is particularly useful when examining dynamic coefficient compensation.

5 Database

The database used for the clean speech models was the RM database [20]. 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 only three were considered in this work, the February 1989, October 1989 and February 1991 DARPA evaluation sets[5]. The recordings were made in a sound isolated recording booth, yielding a Signal-to-Noise Ratio (SNR) of > 40dB. For all tests performed in this chapter a word pair grammar, perplexity 60, was used.



(a) Lynx Helicopter noise (b) Operations Room noise

Figure 2: Log power spectral density against frequency for Lynx Helicopter noise and Operations Room noise taken from the NOISEX-92 database

The second database used was the NOISEX-92 database [22]. This database has a variety of noise sources, two of which were chosen for experiments, the Lynx Helicopter noise, the power spectral density of which is shown in figure 2(a), and the Operations Room noise, figure 2(b).

The Lynx helicopter noise was added to the clean RM data to give a SNR of approximately 18dB[6]. To achieve this SNR, the noise was attenuated by 20dB. The exact SNRs for the three test sets, and the "clean" SNRs, are shown in table 1. The mismatch between the clean and the noisy conditions was about 30dB. To obtain alternative SNRs the noise was scaled accordingly and added.

[5] Only three of the four possible RM test sets were considered. The Sep'92 test set was not used. A variety of "spot checks" were performed with the Sep'92 and no difference in the trends described were observed.

[6] NIST "wavmd" code was used to determine the SNR.

Test Set	Clean SNR (dB)	Noisy SNR (dB)
Feb'89	48.9	18.2
Oct'89	48.7	18.4
Feb'91	48.7	18.6

Table 1: Average SNR for the RM test sets adding Lynx Helicopter noise attenuated by 20dB

6 Recognition System

The baseline speech recognition system was built using the Resource Management Toolkit distributed with HTK Version 1.5 [24] and used for the November 1992 ARPA Evaluation [23]. Two modifications to the parameterisation used for the evaluation were made. In order to use PMC to compensate a set of models, it is necessary to have the zeroth Cepstra to represent the energy whereas the standard HTK system uses a normalised log-energy. 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 calculate deltas and delta-deltas using simple difference expressions.

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-Spectral parameters were used to generate the basic acoustic analysis. From this vector, an observation vector consisting of 12 MFCCs, appended by the zeroth Cepstra, was generated. In addition to the static parameters, delta and delta-delta parameters were added with $w = 2$ and $w_a = 2$. Since all the Cepstral parameters were liftered, it was necessary to inverse lifter to achieve the distributions required for compensation. In addition, Cepstral smoothing was applied. The 24 Log-Spectral values were mapped to 13 Cepstral parameters. This meant that when performing the inverse DCT it was necessary to zero pad the feature vector. Similarly when mapping back it was necessary to truncate the Cepstral vector to 13 again.

7 Results

7.1 Clean System Performance

Energy Term	Dynamic Terms	Test Set			Average
		Feb'89	Oct'89	Feb'91	
Log [23]	Regr.	4.5	5.1	4.0	4.5
Cepstra	Diff.	4.3	5.4	3.9	4.6

Table 2: Word error rates (%) of a six component model set on clean RM

As noted above, in order to use PMC to compensate the models it was necessary to alter the speech parameter set from that used in the standard HTK RM recogniser. A set of six component models, similar to those used for the ARPA RM system developed at CUED [23], was therefore generated. Table 2 shows the performance of this six component system on the clean RM test sets compared with those published using the standard HTK parameterisation [23]. There was little difference in performance, 4.6% compared to 4.5% for the standard HTK RM feature set. This six component model set with zeroth Cepstra energy and simple difference dynamic parameters was the standard model set used throughout this paper.

7.2 Matched System Performance

Comp Mean Param	Comp Var Param	Test Set			Average
		Feb'89	Oct'89	Feb'91	
—	—	38.7	32.0	33.4	34.7
O	—	19.9	17.7	16.9	18.2
O,ΔO	—	15.3	13.5	12.0	13.6
O,ΔO,$\Delta^2 O$	—	10.9	11.1	9.0	10.4
O	O	16.1	13.7	13.5	14.4
O,ΔO	O,ΔO	10.2	10.1	8.4	9.6
O,ΔO,$\Delta^2 O$	O,ΔO,$\Delta^2 O$	7.3	8.6	6.9	7.6

Table 3: Word error rates (%) compensating various parameter groups using single-pass training on additive Lynx noise-corrupted RM at 18dB

A six component system was trained in a single-pass on 18dB Lynx Helicopter additive noise-corrupted data. Various parameters of this system were merged with the clean system, to examine the importance of compensating each "group" of parameters in a model-based compensation scheme, i.e. static, delta and delta-delta parameters. The results on the various test sets are shown in table 3. The first column shows which of the means are "compensated", and the second column which variances. In both cases O, ΔO and $\Delta^2 O$ represent static, delta and delta-delta parameters respectively.

Simply compensating the static means, as is done with hypothesised Wiener filtering for example, reduced the error rate by 48%. A total reduction of 60% was achieved by also compensating the delta means and, finally, incorporating the delta-delta compensated means yielded a 70% reduction. For this noise it can clearly be seen that both the delta and the delta-delta means must be compensated to achieve good performance. If in addition to the means, the variances were compensated the following reductions in error rate were obtained: for the statics 58%, the static and deltas 72%, and all the means and variances of the system 78%. Compensating the static means gave the largest reduction in error rate, however it can be seen that to obtain the best performance both means and variances must be compensated.

The performances of a multi-pass trained system and a single-pass trained system with additional iterations of Baum-Welch re-estimation were found to be marginally worse than that of the single-pass trained system. This degradation in performance, despite the fact that the non-Gaussian nature of the corrupted-speech distribution can be modelled, is felt to be due to the poor alignments achieved when corrupted data is used.

Throughout this paper the single-pass trained system will be referred to as the *matched system*.

7.3 PMC-Compensated System Performance

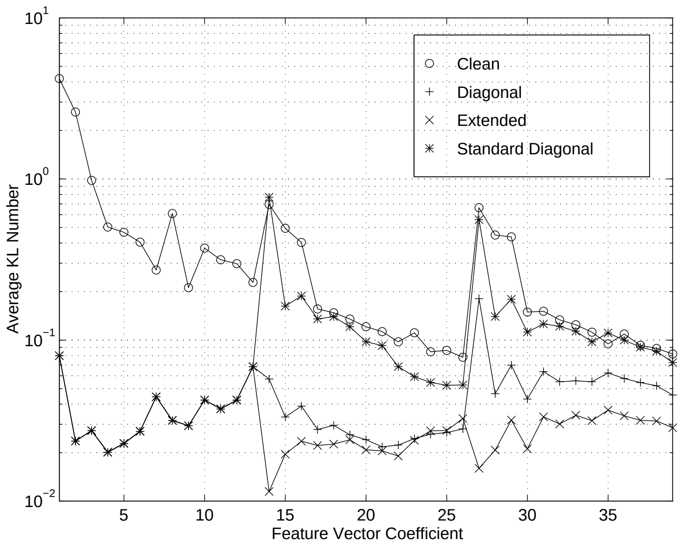


Figure 3: Average KL number for clean model set and PMC-compensated model sets using Diagonal, Extended and Standard Diagonal covariance approximations on Lynx Helicopter additive noise-corrupted RM at 18dB

Figure 3 shows the average KL number for an uncompensated system and various PMC compensated systems. In the figure, feature vector coefficients 1 to 13 are C_0 to C_{12}, 14 to 26 are the respective delta parameters, and 27 to 39 the delta-delta parameters. For the clean model set, *Clean*, the static parameters were the most affected by additive noise, with the low-order static Cepstra being distorted to a greater extent than the higher-order Cepstra. The delta parameters were less affected than the static parameters, though the same general shape was observed. The delta-delta parameters were affected, approximately, to the same extent as the delta parameters. This agrees with the word error rate results in table 3, where best performance was achieved when all the parameters were compensated. The use of PMC with the Diagonal covariance approximation, labelled *Diagonal*, reduced the average KL number, in particular for the static and the delta parameters. However, the delta-delta parameters were not so well compensated, especially the lower-order delta-delta parameters. Improving the correlation modelling using the Extended covariance approximation, *Extended*, decreased the average KL number for the dynamic coefficients. The use of the Standard Diagonal covariance approximation, *Standard Diagonal*, shows poor dynamic parameter compensation.

The recognition performance of the various covariance approximations is shown in table 4. Comparing the Diagonal covariance approximation with the matched system showed approximately the same level of performance. The Extended approximation slightly improved the performance. If the Standard Diagonal covariance approximation was used there was a marked drop in performance. Hence, for good recognition performance on this database, it was necessary to explicitly model the additional parameters required for the dynamic coefficient mismatch functions, agreeing with the observations using the average KL number.

The SNR was lowered by 8dB to about 10dB SNR, an attenuation of the original noise signal by 12dB. The word error rates for the various PMC covariance approximations are

Model Set	Test Set			Average
	Feb'89	Oct'89	Feb'91	
Clean	38.7	32.0	33.4	34.7
Matched	7.3	8.6	6.9	7.6
Diagonal	7.5	8.1	6.4	7.4
Extended	6.7	7.9	6.6	7.1
Standard Diagonal	9.6	9.0	8.2	8.9

Table 4: Word error rates (%) using various PMC covariance approximations on Lynx Helicopter additive noise-corrupted RM at 18dB

Model Set	Test Set			Average
	Feb'89	Oct'89	Feb'91	
Clean	86.6	84.7	85.1	85.5
Matched	18.3	15.2	15.7	16.4
Diagonal	20.0	15.8	16.3	17.4
Extended	17.3	14.2	15.3	15.6
Standard Diagonal	24.8	19.3	20.4	21.5

Table 5: Word error rates (%) using various PMC covariance approximations on Lynx Helicopter additive noise-corrupted RM at 10dB

shown in table 5. At this lower SNR the need to explicitly model all the parameters of the mismatch functions and some correlations became more obvious. The Extended covariance approximation reduced the word error rate by about 10% over the Diagonal case. It is interesting to note that the Extended covariance approximation performed slightly better than the matched system (this is also true for the results in table 4), though the difference was only slight. A possible explanation is that in the compensation schemes all clean speech models are effectively observed in all noise conditions. For components which are not observed many times in the training data, the matched system may not have observations in all noise conditions.

A second noise source was examined, the Operations Room noise from the NOISEX-92 database. Again, the noise was attenuated by 20dB and added to give about 18dB SNR. The word error rates for uncompensated, matched, and various PMC compensated systems are shown in table 6. The same general trends as for the Lynx Helicopter noise-corrupted data were observed, despite the very different nature of the interfering noise source. However, for this case there was little improvement in terms of word error rate in using the more complex covariance approximations.

8 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

Model Set	Test Set			Average
	Feb'89	Oct'89	Feb'91	
Clean	44.3*	39.6*	37.4*	40.5
Matched	9.2	11.4	8.4	9.7
Diagonal	9.7	11.5	8.3	9.9
Extended	9.5	10.6	8.9	9.7

Table 6: Word error rates (%) using various PMC covariance approximations on Operations Room additive noise-corrupted RM at 18dB, * indicates some sentences ran out of hypotheses

performance. For example a single-pass trained, matched, six component system achieves 18.2% 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 7.6% on the same task, a reduction of 58% in the word error rate. In order to effectively compensate the delta parameters, it is necessary to use an additional model set based on the statistics at time $\tau - 2$ to compensate the delta and delta-delta parameters. Using PMC and the Diagonal covariance approximation a 7.4% word error rate on the same task, comparable with the matched system, compared with 8.9% when no $\tau - w$ models were used, the Standard Diagonal covariance approximation. Improving the correlation modelling using the Extended covariance approximation further reduced the error rate of the system by 4%. Similar trends were observed on the Operations Room noise at the same SNR. PMC also gave performance comparable to the matched system at 10dB SNR.

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 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

During the period of this work M. Gales was 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. *IEEE Transactions ASSP*, 27:113–120, 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*, 1:68–76, 1993.

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

[9] 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.

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

[11] M J F Gales and S J Young. HMM recognition in noise using parallel model combination. In *Proceedings Eurospeech*, pages 837–840, 1993.

[12] M J F Gales and S J Young. A fast and flexible implementation of parallel model combination. In *Proceedings ICASSP*, pages 133–136, 1995.

[13] M J F Gales and S J Young. Robust speech recognition in additive and convolutional noise using parallel model combination. *Computer Speech and Language*, 9:289–307, 1995.

[14] Y Gong. Speech recognition in noisy environments: A survey. *Speech Communication*, 16:261–291, 1995.

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

[16] 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.

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

[18] 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.

[19] J P Openshaw and J S Mason. On the limitations of Cepstral features in noise. In *Proceedings ICASSP*, volume 2, pages 49–52, 1994.

[20] 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.

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

[22] 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.

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

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