# A FAST AND FLEXIBLE IMPLEMENTATION OF PARALLEL MODEL COMBINATION

*M.J.F. Gales* *S.J. Young*

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

## ABSTRACT

In previous papers the use of Parallel Model Combination (PMC) for noise robustness has been described. Various fast implementations have been proposed, though to date in order to compensate all the parameters of a system it has been necessary to perform Gaussian integration. This paper introduces an alternative method that can compensate all the parameters of the recognition system, whilst reducing the computational load of this task. Furthermore, the technique offers an additional degree of flexibility, as it allows the number of components to be chosen and optimised using standard iterative techniques. The new technique is referred to as Data-driven PMC (DPMC). It is evaluated on the Resource Management database, with noise artificially added from the NOISEX-92 database. The performance of DPMC is found to be comparable to PMC, at a far lower computational cost. In complex noise environments, by more accurately modelling the noise source, using multiple components, and then reducing the number of components to the original number a slight improvement in performance is obtained.

## 1. INTRODUCTION

Practical speech recognition systems must be able to achieve good performance in a wide variety of noise environments. These environments may vary in both the additive interfering noise, such as fans, engine noise, and the channel conditions over which the speech is being recorded, such as microphone variation. These problems effect all speech recognition systems. However, the task of achieving noise robustness for medium to large vocabulary systems is far harder, as in order to achieve good recognition performance it is necessary to incorporate dynamic coefficients into the feature vector. The effect of interfering noise on these dynamic coefficients is complicated and makes many techniques used for small vocabulary systems ineffective for larger vocabulary systems.

Most existing techniques for making medium to large vocabulary systems robust to noise, rely on making collections of 'stereo' data where one track is noise-free and the other is recorded in a typical noisy environment [4, 5]. By learning sets of mappings between the two, noisy test speech from an unknown environment can be recognised by using the most appropriate of the pre-trained mappings. However, these techniques require new noise data corrupted data in order to handle new noise sources. It would be preferable to have a scheme that adapted to the current environment. This is the approach adopted by Parallel Model Combination (PMC).

In PMC the speech models are modified to be representative of the speech in the new acoustic environment given an estimate of the additive noise. However, a major problem with the technique is the time taken to adapt the parameters of the models. For small vocabulary systems, where using only static parameters achieves acceptable recognition performance, very fast approximate techniques can be used [1]. However, such approximations are not applicable when dynamic parameters are required to be compensated. This need to increase the speed of the compensation process has lead to the development of the Data-driven Parallel Model Combination (DPMC) technique described here. In this implementation of PMC the speech models are used to generate separate samples of 'speech' and 'noise'. These are then combined according to the appropriate 'mis-match' function to obtain the noise corrupted speech samples which are then used to estimate the corrupted models. Using this technique, the compensation time may be dramatically reduced. In addition to increasing the speed of compensation the technique also allows the generation of an arbitrary number of components per state, thus removing the problem of the combinatorial expansion of states or mixture components which occurs in standard PMC when the noise source is non-stationary.

This paper describes the DPMC technique and presents results for a noise corrupted Resource Management recognition task.

## 2. PARALLEL MODEL COMBINATION

### 2.1. Static Parameters

PMC attempts to estimate the parameters of a corrupted speech model. It assumes that a speech model trained on clean speech log spectra $\mathbf{S}^l(t)$ is available and that a model of the noise log spectra $\mathbf{N}^l(t)$ can be estimated *on-line*. Note that in this and subsequent equations, the super-script $l$ is used to indicate that the variable is in the log spectral domain, variables without super-scripts are in the linear spectral domain. Under the assumption that the speech and noise are additive in the linear spectral domain

$$O_i^l(t) = \log(g \exp(S_i^l(t)) + \exp(N_i^l(t))). \quad (1)$$

where the subscript $i$ indexes the $i^{th}$ component of the spectral feature vector and $g$ is a gain matching function introduced to account for the difference between the training and testing speech levels. PMC then estimates the statistics of $\mathbf{O}^l(t)$ using the statistics of $\mathbf{S}^l(t)$ and $\mathbf{N}^l(t)$. For recognition, the speech is typically modelled in the log domain using Hidden Markov Models in which each state is a multivariate Gaussian or a mixture of Gaussians. For



0-7803-2431-5/95 $4.00 

this case, therefore, the PMC method involves applying the above mis-match function to each pair of speech and noise states to yield a new compensated *noisy speech* state whose means and variances are computed from the expected values of $O_i^l(t)$. If the noise HMM has $N$ states then the number of states in the compensated HMM is increased by a factor of $N$ compared to the original [2]. However, in many cases, including previously reported experiments, it is sufficient to model the noise with a single state HMM and in this case the PMC-based compensation does not increase the computational complexity of the recogniser. Notice that the implicit assumption is being made that the addition of the noise does not change the state/frame-component alignment between the speech and the HMM.

In practical systems, cepstral coefficients are often used in preference to log spectra. This does not affect the PMC method since the discrete cosine transform is linear and hence the cepstral parameters can be easily mapped to and from the log spectral domain [1].

### 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 [3] for the delta parameters, $\Delta O_i^l(t)$. If

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

where $w$ is the difference offset, then

$$\begin{aligned} \Delta O_i^l(t) \;=\; & \log\left(\exp(\Delta S_i^l(t) + S_i^l(t-w) + g^l)\right. \\ & \left. + \exp(\Delta N_i^l(t) + N_i^l(t-w))\right) \\ & - \log\left(\exp(S_i^l(t-w) + g^l) + \exp(N_i^l(t-w))\right) \end{aligned} \tag{3}$$

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 3. Alternatively, it is possible to generate an additional set of models built on statistics at time $t-w$.

## 3. DATA DRIVEN PMC

DPMC is a new technique to estimate the parameters of the models. It addresses two problems associated with PMC. The first is the computational overhead associated with compensating large vocabulary systems, especially where delta and delta-delta parameters are required to be compensated. When performing numerical integration of the mismatch functions, it is necessary to integrate in the Log-Spectral domain, hence all elements of a full covariance matrix must be calculated. For a 24 dimensional Log-normal model this requires 300 separate numerical integrations. This problem is overcome in DPMC by generating a set of noise corrupted speech vectors for each speech and noise component pairing. These vectors may be generated in either the Log-Spectral or Cepstral domains. Once the data has been generated it is only necessary to calculate the means and variances to obtain the ML estimate of the corrupted speech model; a simple and fast task. The computational overhead is now in synthesising the data and is dependent on the number of points generated.

The second problem associated with PMC is in situations where a complex noise model is required. In standard PMC every speech-noise component pair must be separately modelled. Thus a two component per state noise model and a six component speech model results in a twelve components per state corrupted speech model with an associated run-time computational overhead. Since a set of corrupted speech vectors has been generated for each speech noise state pairing, it is now a standard HMM training problem to obtain the ML estimate of the data, with as many or few components as desired.

The compensation process thus uses equations 1 and 3 to generate the samples of the noise corrupted speech. This yields a set of $T$ data points in the cepstral domain. The weights, means and variances are then estimated using [3]

$$\hat{c}_m^{(n+1)} = \frac{1}{T}\sum_{\tau=1}^{T} \mathcal{K}_m(\tau) \tag{4}$$

$$\hat{\mu}_m^{(n+1)} = \frac{\sum_{\tau=1}^{T} \mathcal{K}_m(\tau)\mathbf{O}^c(\tau)}{\sum_{\tau=1}^{T} \mathcal{K}_m(\tau)} \tag{5}$$

and

$$\hat{\mathbf{\Sigma}}_m^{(n+1)} = \left( \frac{\sum_{\tau=1}^{T} \mathcal{K}_m(\tau)\mathbf{O}^c(\tau)\mathbf{O}^c(\tau)^T}{\sum_{\tau=1}^{T} \mathcal{K}_m(\tau)} \right) - \hat{\mu}_m^{(n+1)}(\hat{\mu}_m^{(n+1)})^T \tag{6}$$

where

$$\mathcal{K}_m(\tau) = \frac{c_m \mathcal{N}(\mathbf{O}^c(\tau); \hat{\mu}_m^{(n)}, \hat{\mathbf{\Sigma}}_m^{(n)})}{\sum_{i=1}^{M} c_i \mathcal{N}(\mathbf{O}^c(\tau); \hat{\mu}_i^{(n)}, \hat{\mathbf{\Sigma}}_i^{(n)})} \tag{7}$$

In order to initialise the estimation process it is necessary to give an initial estimate of the state components. If the number of components is to be the number of noise components times the number of speech components, the initial estimates are generated by knowledge of which component pair generated the corrupted observation. This is the 0 iteration estimation and is equivalent to assuming that the frame/component allocation does not change. If the number of components is to be reduced the initial estimate may be made by merging components, or taking the heaviest set of components.

It is only possible to combine components within a state, as the states contain the only temporal information at the model level. Thus, it is not possible to take a three state left to right model and estimate a two state left to right model in any sensible fashion.

## 4. RESULTS

### 4.1. Lynx Helicopter Noise

A set of six component models, similar to those used for the ARPA RM system developed at CUED [6], were generated. Additionally, a set of *Noisy* models were generated by corrupting the RM training database with Lynx helicopter noise. These models were generated in a single pass using the complete dataset from the clean speech. By generating models in this way they should be the true model set that PMC attempts to estimate.

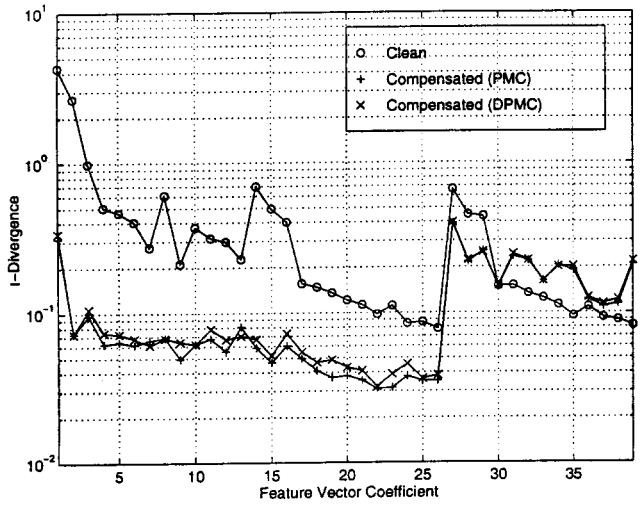


Figure 1. I-Divergence between the Model Set trained on Noise Corrupted data and each of the Clean Model Set, the PMC Compensated Model Set and the DPMC Compensated Model Set on Lynx Additive Noise Corrupted RM at 18-20dB

To investigate how closely models obtained using PMC and DPMC are to this *Noisy* model set, the average I-Divergence over all the Gaussian components was calculated. The results for each feature vector component are shown in figure 1. Feature vector coefficients 1-13 are the static parameters $C_0$ to $C_{12}$, 14-26 are the deltas and 27-39 are the delta-deltas. For the uncompensated models the static parameters are the most effected by additive noise, with the low order cepstra being distorted to a greater extent than the higher order Cepstra. The use of both PMC and DPMC show comparable performance and for both the static and the delta parameters have a significantly lower average I-Divergence than the uncompensated models. However for the delta-delta parameters the reduction in the I-Divergence is not so dramatic, indeed for the higher Cepstra the I-Divergence is higher. PMC does however reduce the maximum and the variance of the I-Divergence of the model set.

Table 1 shows the performance of these models on a noise corrupted version of the RM database, Lynx helicopter noise was added at 18-20dB. As described in previous reports, the use of PMC achieves good performance in the noise corrupted environment, comparable with that of training the model in the noise environment. The performance of the clean models *Clean* is poor. From figure 1 the advantages in compensating the delta-delta parameters appear to be small. On the Feb'89 test set not compensating the delta-delta parameters increased the word error rate from 8.3% to 10.4%. Thus, for this task there is an advantage in compensating the delta-delta parameters. This may be explained by examining the distance between the means of the PMC compensated models and the *Noisy* models. PMC improves the accuracy of means of the delta-deltas, however the variance compensation appears to be poor. As the recognition performance is effected by the means to a greater extent than the variances, the performance may improve, despite the increase in the I-Divergence.

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

Table 1. Comparison of the Performance of Clean models, Models trained on noise corrupted date and PMC compensated models for a Six Component System on Lynx Additive Noise Corrupted RM at 18-20dB

| Test Set | No. Pts. | Mean | Std.Dev. |
|---|---|---|---|
| Feb'89 | 1200 | 8.34 | 0.22 |
| Oct'89 | 1200 | 7.96 | 0.18 |
| Feb'91 | 1200 | 8.14 | 0.32 |

Table 2. Word Error Rate Statistics for Data Driven PMC at 18-20dB Lynx Additive Noise Corrupted RM

The results for DPMC using the same set of clean speech models and noise model is shown in table 2. Comparing these results with standard PMC shows a slight degradation in performance. However, the time taken to compensate the model set is an order of magnitude faster than the numerical integration. In addition to the average performance there is a standard deviation quoted, as the corrupted speech models are estimated using randomly generated points and are thus dependent on the random seed value.

### 4.2. Machine Gun Noise

| Model Set | Test Set | Words Corr | Subs Err | Del Err | Ins Err | Word Err |
|---|---|---|---|---|---|---|
| Clean | Feb'89 | 65.3 | 31.3 | 3.4 | 17.0 | 51.6 |
| | Oct'89 | 65.7 | 31.0 | 3.3 | 16.8 | 51.1 |
| | Feb'91 | 67.2 | 29.7 | 3.1 | 19.5 | 52.3 |
| 1 comp. | Feb'89 | 86.9 | 9.6 | 3.5 | 1.7 | 14.8 |
| | Oct'89 | 87.1 | 7.5 | 4.2 | 2.1 | 14.9 |
| | Feb'91 | 89.7 | 8.0 | 2.3 | 1.7 | 12.0 |
| 2 comp. | Feb'89 | 89.5 | 7.1 | 3.4 | 0.8 | 11.3 |
| | Oct'89 | 88.9 | 7.3 | 3.8 | 0.9 | 12.0 |
| | Feb'91 | 91.2 | 6.7 | 2.1 | 0.8 | 9.6 |

Table 3. Comparison of Uncompensated, a Single Gaussian Noise Model and a Two Gaussian Component Noise Model for Data Driven PMC at 4-6dB Machine Gun Noise using 1200 points per state

To investigate the ability to alter the number of components in complex noise environments, machine gun noise was added to the RM database at 4-6dB. This is a very distinct two state/component noise. The performance of the clean, single component noise model and two component noise model are shown in table 3. The single component noise model performs surprisingly well considering the nature of the noise, achieving 14.8% on the Feb'89 test set. By incorporating an additional component in the noise model this figure may be reduced to 11.3%. However the number of components in the compensated system is 12, resulting in a run-time overhead.

| Test Set | Words Corr | Subs Err | Del Err | Ins Err | Word Err |
|---|---|---|---|---|---|
| Feb'89 | 87.6 | 8.5 | 3.9 | 1.7 | 14.1 |
| Oct'89 | 88.0 | 7.9 | 4.1 | 1.8 | 13.8 |
| Feb'91 | 90.1 | 7.4 | 2.5 | 1.1 | 11.1 |

Table 4. Performance of a DPMC Compensated System using a Two Component Noise Model and Re-combining the Component Associated with each Speech Components

The simplest method to reduce the number of components associated with each state is to recombine all components which were generated from the same speech component. The results of this scheme are shown in table 4. By improving the accuracy of the noise model, then reducing the number of components in this simple fashion, has reduced the word error rate, for example on the Feb'89 test set the word error rate has been reduced from 14.8% to 14.1%.

| Initial Data-set | Test Set | Words Corr | Subs Err | Del Err | Ins Err | Word Err |
|---|---|---|---|---|---|---|
| Heaviest | Feb'89 | 88.8 | 7.8 | 3.3 | 0.8 | 11.9 |
| | Oct'89 | 88.2 | 8.4 | 3.4 | 1.3 | 13.2 |
| | Feb'91 | 90.2 | 7.6 | 2.3 | 1.4 | 11.2 |
| M. Info. | Feb'89 | 87.3 | 9.1 | 3.6 | 1.1 | 13.8 |
| | Oct'89 | 86.7 | 9.5 | 3.8 | 1.0 | 14.3 |
| | Feb'91 | 90.4 | 7.2 | 2.4 | 1.0 | 10.6 |

Table 5. Comparison of Various Initial Data-set Techniques with a Two Gaussian Component Noise Model using Data Driven PMC at 4-6dB Machine Gun Noise with 1200 points per State, Mapping back to 6 Components using 10 iterations of EM

Using DPMC it is possible to choose the number of components in the compensated models and to use standard EM techniques to obtain maximum likelihood estimates. For this work the number is the original number of speech components, hence there will be no additional run-time overhead. It is necessary to decide on the initial set of components to be used in the iterative scheme. Two techniques for estimating the initial set of components were investigated. Firstly the six "heaviest" components, *Heaviest*, may be used as initial estimates. Alternatively, the closest components according to a mutual information measure may be merged to achieve the required number,*M. Info.* 10 iterations of EM were then used to smooth the distributions and obtain the ML estimates. The results are shown in table 5. The best performance was achieved using the *Heaviest* initial components. With this system the performance on Feb'89 test set was reduced from 14.8% for the single component noise model and 14.1% for the speech combined two component system to 11.9%. However, overall the performance is worse than that of the standard two component system.

## 5. CONCLUSIONS

Initial results have been presented on the use of a new technique for implementing PMC, called Data-driven PMC. This technique has been shown to achieve comparable performance with standard PMC at a far lower computational cost. The performance of DPMC in *complex* noise environments, was then investigated. By using multiple component distributions to improve the accuracy of the noise model, recognition performance was found to improve, albeit at the computational cost associated with doubling, in the case of a two component noise model, the number of components in the system. To overcome this problem, various merging and smoothing schemes, were examined. A simple scheme of combining according to the original speech component, was found to improve performance. Further slight improvements were obtained by smoothing the set of the heaviest components with EM. These performance gains were obtained with a very distinct two state/component interfering noise. Whether similar gains may be obtained when the noise source is not so distinctly multi-state needs to be investigated.

## ACKNOWLEDGEMENT

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

## REFERENCES

[1] 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.

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

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

[4] F Liu, P J Moreno, R M Stern, and A Acero. Signal processing for robust speech recognition. In *Proceedings ARPA Workshop on Human Language Technology*, pages 309–314, 1994.

[5] L Neumeyer and M Weintraub. Probabilistic optimum filtering for robust speech recognition. In *Proceedings ICASSP*, pages 417–420, 1994.

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