PMC FOR SPEECH RECOGNITION IN ADDITIVE AND CONVOLUTIONAL NOISE

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

CUED/F-INFENG/TR154

December 1993

Cambridge University Engineering Department
Trumpington Street
Cambridge CB2 1PZ
England

Email: mjfg@eng.cam.ac.uk

Abstract

This paper addresses the problem of speech recognition in the presence of both additive and convolutional noise. A new scheme is described, which is a simple extension to the standard Parallel Model Combination (PMC) technique. A modified 'mismatch' function is introduced which accounts for the effects of convolutional noise. This 'mismatch' function is then used to estimate the difference in channel conditions between training and test environments. Having estimated the tilt parameters, Maximum Likelihood (ML) estimates of the corrupted speech model may be obtained. The scheme is evaluated using the NOISEX-92 database. The performance in the presence of both interfering additive noise and convolutional noise shows only slight degradation compared with that obtained when no convolutional noise is present.

Contents

1 Introduction

As speech recognition technology moves from the laboratory to real applications, there is a need to make systems which are robust to a wide variety of background noises. Many different approaches to achieving additive 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 [15, 13] spectral mapping [8] and inherently robust parameterisations [6]. These methods only use statistical information about the interfering noise in the compensation process, no account is made of what was said.

The second class of methods attempt to modify the pattern matching stage in order to account for the interfering noise. Methods in this approach include noise masking [7, 5], state based filtering [17], cepstral mean compensation [2, 18] and HMM decomposition [4].

In general, both sets of techniques have concentrated on solely additive noise. In many situations there may also be a mismatch in the channel conditions between training and test conditions, for example the microphone may change. This mismatch is referred to as convolutional noise or channel distortion. To make a system truly robust to environmental conditions, it is necessary to compensate for this channel distortion. Work in this area has used both inherently convolutional noise robust frontends [12] modified for the effects of additive noise and explicit estimation of the spectral tilt [1]. In previous work [9, 10, 11], Parallel Model Combination (PMC) has been shown to be effective against additive noise. However, to date all the work on PMC has assumed that there is no difference in the channel conditions between training and testing. This report introduces a new method of explicitly estimating the spectral tilt and incorporating this estimate into the PMC framework. The performance of the complete system is evaluated on the NOISEX-92 database.

The layout of this report is as follows. The next section describes the basic theory behind PMC when there is additive noise present. The theory behind compensating models in the presence of both additive and convolutional noise is then described. Section 5 details the experiments performed on the NOISEX-92 database. This section is split into two subsections. The first describes a set of experiments carried out in the presence of just additive noise[1]. The second subsection describes a series of experiments performed in the presence of both additive and convolutional noise. Finally conclusions are drawn from the results.

2 Additive Noise Compensation

Standard PMC attempts to estimate the parameters of a corrupted speech model in the presence of additive interfering noise. It has previously been shown to yield a Maximum Likelihood (ML) estimate of the corrupted speech model [11] given the statistics about the interfering additive noise, the clean speech and a 'mismatch' function describing the effects of the interfering noise on the model parameters. In order to obtain this mismatch function various assumptions are required.

1. The speech and noise are independent.

2. The speech and noise are additive in the linear 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 set of Gaussian mixtures contain 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.

The speech and the noise are most naturally combined in the Log Spectral or Linear Spectral domains. If speech is parameterised using Cepstral coefficients it is necessary to convert the

[1] These results supersede previously published results

parameters to the Log Spectral domain. This is achieved using an inverse Cosine transform, as described in previous work [9]. For static parameters the 'mismatch' function is

$$O_i^l(t) = \mathcal{F}(S_i^l(t), N_i^l(t)) = \log(g \exp(S_i^l(t)) + \exp(N_i^l(t))) \tag{1}$$

where g is a gain matching term introduced to account for level differences between the clean speech and the noisy speech, $O_i^l(t)$, $S_i^l(t)$ and $N_i^l(t)$ are the i^{th} component of the Log Spectral domain feature vector of the corrupted speech, clean speech and noise at time t. The notation adopted throughout this report is to use the superscript to denote the domain of the variable. Thus $O_i^c(t)$, $O_i^l(t)$ and $O_i(t)$ are in the Cepstral , Log Spectral and Linear Spectral domains respectively. Where probability distribution parameters are given those related to the noise have a '$\tilde{}$', those of the corrupted model a '$\hat{}$' and those of the clean speech are unmarked. Thus, $\tilde{\mu}_i^l$ is the mean vector associated with the noise in the Log Spectral domain. The actual estimation of the parameters may be performed in two ways.

1. An approximation to the ML estimate of the corrupted speech model may be obtained by assuming that the sum of two Log-normally distributed variables is itself approximately Log-normally distributed [9]. This will be referred to as the Log-Normal approximation.
2. An alternative estimate may be obtained using numerical integration to obtain $\mathcal{E}\{O_i^l(t)\}$ and $\mathcal{E}\{O_i^l(t)O_j^l(t)\}$, thus giving the mean and variance of the corrupted model [11].

The performance of the two methods has been found to yield comparable results [11]. Unless otherwise stated numerical integration is used throughout this report.

3 Convolutional Noise Compensation

In the previous section the 'mismatch' function for static parameters has been described assuming that there is no mismatch in the channel conditions. For situations where there is channel distortion as well as additive noise

$$O_i(t) = H_i S_i(t) + N_i(t) \tag{2}$$

where $\mathbf{H}$ represents the channel difference between training and testing, noting that convolutional noise in the time domain is multiplicative in the frequency domain. $\mathbf{H}$ now subsumes the gain term g used in equation 1 since that latter can be regarded as just a very simple form of channel distortion, that of a uniform gain difference. Expressing this equation in the same form as equation 1 yields

$$O_i^l(t) = \mathcal{F}(S_i^l(t), N_i^l(t), H_i^l) = \log(\exp(S_i^l(t) + H_i^l) + \exp(N_i^l(t))) \tag{3}$$

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

Having obtained the 'mismatch' function, it is possible to examine the effects of additive noise on a standard tilt compensation scheme. The scheme considered is cepstral mean subtraction. Here the mean of each cepstral parameter is subtracted from the feature vector. As the Discrete Cosine Transform is a linear operation, subtracting means in the Log Spectral domain is identical to subtracting means in the cepstral domain. Hence

$$\hat{O}_i^l(t) = O_i^l(t) - \overline{O}_i^l \tag{4}$$

where

$$\overline{O}_i^l = \sum_{t=1}^{T} O_i^l(t) \tag{5}$$

For clean speech with no additive no noise

$$\hat{O}_i^l(t) = S_i^l(t) + H_i^l - (\overline{S}_i^l + H_i^l) = S_i^l(t) - \overline{S}_i^l \tag{6}$$

which is independent of the spectral tilt. Thus, provided that mean subtraction is applied to the clean training data and that T is large enough to give a sufficiently accurate estimate of the mean, this scheme will be effective for convolutional noise alone. However when additive noise is present

$$\hat{O}_i^l(t) = S_i^l(t) - \overline{S}_i^l + \log\left(\frac{\exp(S_i^l(t) + H_i^l) + \exp(N_i^l(t))}{\exp(S_i^l(t) + H_i^l)}\right) - \overline{\log\left(\frac{\exp(S_i^l + H_i^l) + \exp(N_i^l)}{\exp(S_i^l + H_i^l)}\right)} \quad (7)$$

which is dependent on both the spectral tilt and the additive noise. The use of such a scheme in the presence of additive noise will seriously degrade the performance. This scheme will be referred to as mean compensation.

A scheme that is robust to both additive and convolutional noise is required. Equation 3 has been obtained in terms of the clean speech and noise, similar to the standard PMC function, but additionally includes the channel difference $\mathbf{H}^l$. The problem is how to obtain the statistics of $\mathbf{H}^l$? Firstly, it is necessary to define what variables are available to make these estimates. At run time the transcription of what is said in the channel conditions is, naturally, not known. Hence, it is preferable to estimate the spectral tilt assuming no knowledge of the transcription. To achieve this end, the model of the speech corrupted by additive and convolutional noise is required to have one state and a single Gaussian distribution associated with that state. If a mixture Gaussian is used for the corrupted speech model, then there are no sufficient statistics available, unless the frame to state/mixture allocation (the transcription) is known. This problem is discussed later. It is therefore not necessary to use models of individual words or phones, provided that a good model of all the acoustic data is available. Instead, a global single-state clean speech and silence model, $\mathcal{M}_G$ is used. By mapping this global model of the speech and silence to a single Gaussian corrupted speech model and assuming that the set of acoustic vectors on which the model is trained is representative of the acoustic vectors observed in the test channel conditions, after being offset, it is not necessary to perform any recognition before estimating the tilt. 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, $\overline{\mathbf{O}}^c$, and higher order statistics if required.

A single state noise model will be assumed in this work. If a multiple state noise model is required for good recognition, 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. Though the models are generated in the Cepstral domain the data is most naturally combined in the Log or Linear Spectral domains. The models are converted between the two domains as described in previous work [9]. Now, it is necessary to chose an appropriate optimisation criterium to estimate the difference in channel conditions. One such criterion is the channel difference that minimises the Euclidean distance between the measured speech and noise global mean, $\overline{\mathbf{O}}^l$, and the mean of the corrupted model obtained after shifting the means of the clean model and compensating for the noise. Hence

$$\mathbf{H}^l = \arg\min\left[\left(\hat{\mu}^l(\mathbf{H}^l) - \overline{\mathbf{O}}^l\right)^T \left(\hat{\mu}^l(\mathbf{H}^l) - \overline{\mathbf{O}}^l\right)\right] \quad (8)$$

where

$$\hat{\mu}^l(\mathbf{H}^l) = \int_{\mathcal{R}^n}\int_{\mathcal{R}^n} \log\left(\exp(S_i^l + H_i^l) + \exp(N_i^l)\right) p(\mathbf{S}^l) p(\mathbf{N}^l)\, d\mathbf{S}^l\, d\mathbf{N}^l \quad (9)$$

and the statistics for the noise and speech, $p(\mathbf{N}^l)$ and $p(\mathbf{S}^l)$, are obtained from the models, $\mathcal{M}_G$ and $\mathcal{M}_N$ respectively. As with the 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 the PMC tilt compensation becomes equivalent to the standard mean compensation described in equation 6.

Assuming the various statistics are known, it is necessary to perform the optimisation. Using the optimisation criterion of equation 8 all the channel differences may be estimated independently. For this task a simple line search may be used. It may be shown that the mismatch function increases monotonically as $\mathbf{H}^l$ increases. Thus, there is only, at most, one minimum of equation 8. As $\mu^l(\mathbf{H}^l)$ spans all possible values for $\overline{\mathbf{O}}^l$, this minimum will occur when $\hat{\mu}^l(\mathbf{H}^l) = \overline{\mathbf{O}}^l$. Given this minimum there is no difference between using the simple Euclidean distance measure and a Mahanolobis distance measure. There are no real minima to equation 8 when the noise mean is greater than the speech and noise mean. This situation is highly unlikely, unless there are long term variations in the noise signal and the estimates for the speech and noise and the noise are made at different times.

An alternative distance measure to that given in equation 8 is to maximise the probability of the tilt estimation data given the speech and silence model and the noise model. Thus

$$\mathbf{H}^l = \arg\max\left[p(\mathbf{O}_T^l|\mathbf{H}^l, \mathcal{M}_G, \mathcal{M}_N)\right] \tag{10}$$

where $\mathbf{O}_T^l$ is the data available to estimate the spectral tilt and assuming that there is an even prior on all the possible spectral tilts. Again the compensated model, $\mathcal{M}_C$, may only have a single Gaussian distribution associated with it, thus avoiding any problems with sufficient statistics. This expression may be related to equation 8. As the log() function increases monotonically it is possible to maximise

$$\begin{aligned}\log\left(p(\mathbf{O}_T^l|\mathbf{H}^l, \mathcal{M}_G, \mathcal{M}_N)\right) &= \\ K - \frac{T}{2}\log(|\hat{\mathbf{\Sigma}}^l(\mathbf{H}^l)|) &- \frac{1}{2}\sum_{t=1}^{T}\left(\mathbf{O}^l(t) - \hat{\mu}^l(\mathbf{H}^l)\right)^T \hat{\mathbf{\Sigma}}^l(\mathbf{H}^l)^{-1}\left(\mathbf{O}^l(t) - \hat{\mu}^l(\mathbf{H}^l)\right)\end{aligned} \tag{11}$$

where the output probability distribution for the corrupted model has mean and variance $\hat{\mu}(\mathbf{H}^l)$ and $\hat{\mathbf{\Sigma}}^l(\mathbf{H}^l)$ respectively. Ignoring the variation of the variance with the spectral tilt, this expression is maximised when

$$\hat{\mu}(\mathbf{H}^l) = \sum_{t=1}^{T}\mathbf{O}^l(t) = \overline{\mathbf{O}}^l \tag{12}$$

This is identical to minimising equation 8. However, the fact that the variance is also a function of the spectral tilt has been ignored. If this effect is included the two schemes will yield different results. It would be preferable to maximise the probability, equation 10, but by using the Euclidean distance measure all the dimensions may be optimised independently, a far easier optimisation task. Additionally, it is simple to prove that there is only one global maximum.

In the previous section two methods of estimating the parameters of the corrupted model were described. Both techniques may be used to estimate the tilt. However, unlike the numerical integration described above, the parameters estimated using the Log-normal approximation are not guaranteed to be monotonically increasing as H_i^l increases. Using the approximation is computationally cheaper than using the numerical integration, as described in section 4

Finally it is necessary to examine the effects of parameter variation on the estimates of the tilt. The clean model, $\mathcal{M}_G$ is trained on all the clean data. Also, the noise model may be trained on periods of noise alone in the test conditions. These models should have small variances on their parameter estimates. The greatest errors in the tilt estimation will result from variations in the estimate of $\overline{\mathbf{O}}^c$. $\overline{\mathbf{O}}^c$ is estimated at run time using all the observed test data. As speech has temporal structure, it is incorrect to assume that the statistics of a short period of speech are the same as those estimated for a long period of speech. To make a good estimate of the tilt it is preferable to use as much data as possible. However, this delays the recognition process, so in a practical system it may be necessary to initially estimate the tilt on limited data. If the estimate of the mean is incorrect then how does this effect the estimate of the tilt? Assuming that the optimisation technique yields the optimal value, so $\hat{\mu}^l(\mathbf{H}^l) = \overline{\mathbf{O}}^l$,

$$\frac{\partial \overline{\mathbf{O}}^l}{\partial H_i^l} = \frac{\partial}{\partial H_i^l}\left(\hat{\mu}^l(\mathbf{H}^l)\right) = \int_{\mathcal{R}^n}\int_{\mathcal{R}^n} \frac{\exp(S_i^l + H_i^l)}{(\exp(S_i^l + H_i^l) + \exp(N_i^l))} p(\mathbf{S}^l) p(\mathbf{N}^l)\, d\mathbf{S}^l\, d\mathbf{N}^l \tag{13}$$

Hence

$$\delta H_i^l \approx \frac{\delta \overline{O}_i^l}{\int_{\mathcal{R}^n}\int_{\mathcal{R}^n} \frac{\exp(S_i^l + H_i^l)}{(\exp(S_i^l + H_i^l) + \exp(N_i^l))} p(\mathbf{S}^l) p(\mathbf{N}^l)\, d\mathbf{S}^l\, d\mathbf{N}^l} \tag{14}$$

If the variances on the noise and speech are small then

$$\delta H_i^l \approx \delta \overline{O}_i^l \left(1 + \frac{1}{SNR}\right) \tag{15}$$

where SNR is the Signal to Noise Ratio (SNR) in the test channel. From this expression it can be seen that as the SNR increases, so $\delta H_i^l \approx \delta \overline{O}_i^l$, and as the SNR decreases variations in the estimate of the speech and noise mean have a greater effect on the spectral tilt estimate.

Errors in the estimation of $\overline{\mathbf{O}}^c$ and higher order statistics, if required to be estimated, may result from two sources. Firstly, errors may occur due to using insufficient data points to correctly estimate the channel mean. Increasing the number of samples used to estimate the mean will reduce this error. Additionally, the errors may occur from a different set of acoustic vectors being uttered in the test conditions to those said in training. For example all the digits may be spoken for training, whereas in the test conditions a telephone number such as 'NINE NINE NINE' is uttered. Simply increasing the number of samples used in the estimate will not reduce this error. In order to solve this problem an iterative EM scheme must be used.

In order to accurately estimate the spectral tilt when the acoustic vectors differ between training and testing an estimate of the transcription is required. This transcription is found given the present estimate of $\mathbf{H}^l$. With this new transcription, $\mathbf{H}^l$ is re-estimated. By this it is meant that instead of mapping the possibly multiple mixture clean speech model to a single mixture corrupted speech model, the corrupted model has the same number of mixtures. In order to have sufficient statistics about the estimation data it is necessary to have the frame/mixture allocation, the transcription. For simplicity it is assumed that the addition of convolutional noise and additive noise does not alter the frame/mixture allocation significantly. The scheme is thus

1. Make an initial estimate of $\mathbf{H}^{l(0)}$, with no knowledge of the transcription.
2. **Estimate** the complete data set, the observed data and the transcription. This gives the frame/state,mixture alignment, $p(\mathbf{O}_T^l, s_T^{(i)} | \mathbf{H}^{l(i)}, \mathcal{M}_G, \mathcal{M}_N)$.
3. **Maximise** the log likelihood of the data.

$$\mathbf{H}^{l(i+1)} = \arg\max \left[\sum_s p(\mathbf{O}_T^l, s_T^{(i)} | \mathbf{H}^{l(i)}, \mathcal{M}_G, \mathcal{M}_N) \log(p(\mathbf{O}_T^l, s_T^{(i)} | \mathbf{H}^{l(i+1)}, \mathcal{M}_G, \mathcal{M}_N))\right] \tag{16}$$

4. Stop if convergence has occurred, otherwise goto step 2.

By allowing the complete data set to vary it allows the mismatch between training data and tilt estimation data to be modelled. This scheme is guaranteed to always increases the probability of the data with each iteration, thus $p(\mathbf{O}_T^l | \mathbf{H}^{l(i+1)}, \mathcal{M}_G, \mathcal{M}_N) \geq p(\mathbf{O}_T^l | \mathbf{H}^{l(i)}, \mathcal{M}_G, \mathcal{M}_N)$. However, it is not guaranteed to obtain a global maximum for the data, just a local maximum. The exact local maximum position will depend on the initial complete data set. The scheme described above uses $\mathcal{M}_G$ to represent all the clean data. This is computationally efficient, however it removes all the transition information and any language modelling. If these are considered important it is necessary to use all the models to estimate the complete data set and the new tilt. These schemes will not be further considered in this report.

4 Computational Overhead

As noted two methods of compensating the model parameters are available. The first assumes that the sum of two Log-normally distributed variables is itself Log-normally distributed. Alternatively, numerical integration may be used. The two schemes have different computational load. The implementation of the approximate scheme is straight forward and has been described in previous work [10]. For the numerical integration it is necessary to map the equations to a standard integration form. The equations were mapped to a Gaussian integration using the abscissas and weights obtained from the Gauss-Hermite polynomial. If performed directly the estimation of the mean is a two dimensional integration and that of the variance a four dimensional integration. By simple manipulation, these values may be estimated using one and two dimensional integrations respectively, as shown in appendix A. The computational cost of this scheme will then depend on the number of points used. For an N point numerical integration in each dimension the computa-

Parameter	Scheme	log()	exp()	Mult/Div	Add/Sub
μ_i	Approx	1	1	2	2
Σ_{ii}	Approx	1	1	4	2
Σ_{ij}	Approx	1	1	4	2
μ_i	Num.	N	N	$3N/2$	$3N$
Σ_{ii}	Num.	N^2+N	N^2+N	$7N^2/2$+$5N/2$	$5N^2+4N$
Σ_{ij}	Num.	$4N^2$	$4N^2$	$33N^2/2$	$19N^2$

Table 1: Computational Load

tional load is shown in table 1. The figures quoted do not include the mapping from the Cepstral domain to the Log-Spectral domain. This cost is the same for both methods. For the numerical integration, terms of of order less than N are ignored. For a P dimensional vector the complete cost, $\mathcal{C}_T$, to estimate the parameters of a single state are

$$\mathcal{C}_T = P\mathcal{C}_{\mu_i} + P\mathcal{C}_{\Sigma_{ii}} + \frac{1}{2}P(P-1)\mathcal{C}_{\Sigma_{ij}} \tag{17}$$

where $\mathcal{C}_{\mu_i}$, $\mathcal{C}_{\Sigma_{ii}}$ and $\mathcal{C}_{\Sigma_{ij}}$ are the cost of estimating μ_i, Σ_{ii} and Σ_{ij}, where $i \neq j$, respectively.

From the table there is a significant computational overhead associated with the use of the numerical integration. The exact cost being dependent on the number of points used, the cost rising as $\mathcal{O}(N^2)$. Various values of N were tried, with no significant variation in results down to $N = 6$. For the results quoted in section 5, N was set to 20.

5 Evaluation on NOISEX-92

In this section a number of experiments using the NOISEX-92 database [3] are reported. The data was preprocessed using a 25 msec Hamming window and a 10 msec frame period. For each frame a set of 15 MFCC were computed. The zeroth cepstral coefficients is computed and stored since it is needed in the PMC mapping process.

For each digit, a single mixture continuous density HMM with 8 emitting states was trained using the clean data only. Where spectral tilt experiments were performed the clean models were tested on the tilted data. The tilt applied was a flat frequency response up to a break point frequency of 250Hz followed by a +3dB/oct tilt above 250Hz. The topology for all models was left-right with no skips and diagonal covariance matrices were assumed throughout. For each test condition, a single state diagonal covariance noise HMM was trained using the silence intervals of the test files. Recognition used a standard connected word Viterbi Decoder constrained by a syntax consisting of silence followed by a digit in a loop. Thus no explicit end-point detector was used and insertion/deletion errors occurred as well as classification errors. The results are in terms of % accuracy where for N tokens, S substitution errors, D deletion errors and I insertion errors, accuracy is calculated as $[(N-S-D-I)/N] \times 100\%$. The error counts themselves were calculated by using a DP string matching algorithm between the recognised digit sequence and the reference transcription. All training and testing used version 1.4 of the portable HTK HMM toolkit [16] with suitable extensions to perform PMC. All results quoted are for the male speaker on the isolated digit or triplet tasks.

5.1 Baseline Results

SNR	*Lynx*		*F16*		*Car*	
dB	MFCC	MFCC_E	MFCC	MFCC_E	MFCC	MFCC_E
-06	12	10	12	5	23	15
+00	25	17	17	5	32	30
+06	59	49	50	27	75	59
+12	97	72	82	62	96	73
+18	100	98	95	96	99	82

Table 2: Baseline Recognition Rates

Table 2 shows the baseline performance for models trained and tested under the same convolutional noise conditions. The effects of additive noise can be seen to be detrimental to the recogniser's performance, even at high signal to noise ratios. The performance of systems where the zeroth cepstra is incorporated, those labelled MFCC_E, are seen to be more sensitive to additive noise than those without energy.

SNR	*Lynx*		*F16*		*Car*	
dB	MFCC	MFCC_E	MFCC	MFCC_E	MFCC	MFCC_E
-06	68	67	63	59	81	76
+00	98	98	96	94	97	97
+06	100	99	99	100	97	97
+12	100	100	100	100	100	100
+18	100	100	100	100	100	100

Table 3: Static Parameter Standard Performance

Table 3 shows the performance of the recogniser when PMC is used to estimate the corrupted speech model parameters. The performance with and without the energy channel, MFCC and MFCC_E respectively are seen to be approximately the same and significantly better than the uncompensated performance. For the rest of this report decoding is performed using MFCC parameters only, the energy, zeroth cepstra, channel is not used unless otherwise stated.

SNR	*Lynx*		*F16*		*Car*	
dB	MFCC	MFCC_E	MFCC	MFCC_E	MFCC	MFCC_E
-06	53	58	57	54	75	74
+00	95	94	95	93	95	96
+06	100	99	100	100	97	97
+12	100	100	100	100	99	100
+18	100	100	100	100	100	100

Table 4: Baseline Recognition Rates using Log-Normal Approximation

As a confirmation that the use of the Log-Normal approximation does not seriously effect the performance, models using this approximation were generated. The results are shown in table 4. These results agree with the observation that the performance is not seriously degraded [11]. However the results may be seen to be slightly poorer. To investigate this further the experiments were repeated using the triplet test set.

SNR	*Lynx*			*F16*			*Car*		
dB	Base	PMC	Approx	Base	PMC	Approx	Base	PMC	Approx
-06	10	59	52	9	55	57	21	85	69
+00	19	87	83	15	88	78	34	97	95
+06	53	97	96	38	97	93	57	98	97
+12	68	99	98	77	98	98	77	97	95
+18	89	99	98	96	98	99	83	96	93
∞	99	99	99	99	99	99	99	99	99

Table 5: Triplet Results

The results for the triplet data are shown in table 5. Again the use of PMC greatly enhances the robustness of the system, achieving good performance down to 0dB SNR for both the numerical integration, *PMC*, and the Log-normal approximation, *Approx*. However, similarly to the digit test, the performance of the Log-normal approximation is generally slightly worse than that of the numerical integration.

5.2 Spectral Tilt Results

SNR	*Lynx*		*F16*		*Car*	
dB	Base	PMC	Base	PMC	Base	PMC
-06	18	33	5	42	20	44
+00	20	51	18	75	27	74
+06	47	75	62	86	45	74
+12	85	82	86	82	73	80
+18	90	85	95	97	76	93

Table 6: Baseline Static Parameter Performance with Spectral Tilt

Table 6 shows the baseline performance figures. The *Base* column shows the performance of uncompensated models in the presence of additive and convolutional noise. Column *PMC* shows the performance with no tilt compensation, just the standard PMC additive noise compensation. The performance can be seen to be badly degraded by the spectral tilt.

Table 7 shows the performance of a standard tilt compensation scheme, that of removing the mean from the data. The *Mean* column indicates that the standard mean compensation scheme was used. The *Mean_c* column indicates that a clean speech mean, with no additive noise, was used in the compensation. The *Mean_c* performance is better than using the mean of the additive noise corrupted speech and similar to that of baseline models where there was no spectral tilt.

Table 8 shows the performance of the tilt compensation scheme with various numbers of mixtures in the speech model, $\mathcal{M}_G$. The corrupted mean estimates were obtained using all the test data. From the table it can be seen that as the number of mixtures increases, so the performance increases. Comparing the performance of the ten mixture clean speech model to the no spectral tilt conditions, shows only a very slight degradation at SNRs of 0dB and higher. This is expected, as from the theory the tilt estimates are effected by errors in the parameters more at lower signal to noise ratios. This indicates that for the database chosen, a ten mixture model encapsulates sufficient information about the speech to allow a good tilt estimate to be made.

SNR	*Lynx*		*F16*		*Car*	
dB	Mean	Mean_c	Mean	Mean_c	Mean	Mean_c
-06	15	17	8	5	16	21
+00	16	20	26	14	27	33
+06	35	53	46	44	30	82
+12	81	97	66	86	42	97
+18	98	100	85	99	53	94
∞	99	99	99	99	99	99

Table 7: Mean Compensation Performance with Spectral Tilt

SNR	*Lynx*				*F16*				*Car*			
dB	1	2	5	10	1	2	5	10	1	2	5	10
-06	52	47	58	49	52	57	58	65	61	73	78	83
+00	80	88	96	96	86	85	90	95	78	89	96	95
+06	94	99	100	100	96	98	99	99	80	95	97	97
+12	98	100	100	100	100	100	100	100	81	96	99	97
+18	99	100	100	100	100	100	100	100	94	100	100	100

Table 8: Static Parameter Performance against Clean Speech Mixtures

SNR	*Lynx*				*F16*				*Car*			
dB	2	5	20	100	2	5	20	100	2	5	20	100
-06	40	38	35	49	51	29	57	65	72	79	80	83
+00	75	92	91	96	82	88	88	95	85	94	95	95
+06	88	100	100	100	93	100	100	99	95	96	95	97
+12	100	100	100	100	99	100	100	100	100	97	96	97
+18	100	100	100	100	100	100	100	100	99	100	100	100

Table 9: Static Parameter Performance against Training Digits

Table 9 shows the performance of the tilt compensation scheme against the number of digits used to estimate the cepstral means, $\overline{\mathbf{O}}^c$. Thus the tilt compensation data in the case of the *20* column was the first 20 digit labels and twenty silence labels. The noise model used for decoding is trained on all the noise data to make results directly comparable with those in other tests. For both the *20* and *100* digit columns all the digits occur in both the training and the test conditions. The performance of the schemes in both cases is good down to 0dB. For the *2* and *5* digit columns not all the digits were spoken in the test conditions. As the iterative scheme, where an estimate of the transcription is made, is not used, this table illustrates the effect of the mismatch between words uttered for training and those uttered for the tilt adaptation data. The performance for the *5* digit case is comparable to those using larger number of digits. However, as expected using only *2* digits to estimate the tilt degrades the performance, particularly at low SNRs.

SNR	*Lynx*		*F16*		*Car*	
dB	MFCC	MFCC_E	MFCC	MFCC_E	MFCC	MFCC_E
-06	48	69	58	58	83	76
+00	96	98	95	87	95	96
+06	100	100	99	100	97	97
+12	100	100	100	100	97	100
+18	100	100	100	100	100	100

Table 10: Recognition Rates using Log-Normal Approximation

As mentioned in the previous section a Log-Normal approximation may be used in combining the two models. Table 10 shows the performance of a scheme using such an approximation to estimate the spectral tilt. All the test data, 100 digits, were used to estimate the spectral tilt. Having estimated the spectral tilt the models were compensated using the numerical integration scheme. The use of the approximation does not appear to have significantly effected the results.

SNR	*Lynx*			*F16*			*Car*		
dB	Base	Mean	PMC_t	Base	Mean	PMC_t	Base	Mean	PMC_t
-06	14	16	48	9	11	67	15	18	76
+00	19	18	78	22	26	89	23	25	95
+06	44	47	94	41	50	95	39	45	97
+12	60	81	98	65	86	97	59	68	98
+18	78	94	99	79	95	96	73	88	99
∞	95	97	97	95	97	97	95	97	97

Table 11: Triplet Results with Various Tilt Compensation Techniques

In addition the tilt compensation was evaluated on the triplet test set. The results comparing mean compensation, labelled *Mean*, with PMC, *PMC_t*, and no compensation, *Base*, are shown in table 11. The mean compensation can be seen to achieve improved performance over the baseline performance for low SNRs. However as the SNR decreases the performance degrades rapidly. The performance of the PMC compensation is slightly worse than that achieved when no spectral tilt is present and PMC is used.

6 Conclusion

This paper has introduced a new approach to achieving robust speech recognition in the presence of both convolutional and additive noise. Two basic schemes have been proposed. The first assumes that the acoustic information is approximately the same in both the training and test channels.

An alternative scheme, where this assumption is not required, using an EM style algorithm is also proposed. The first scheme has been shown to perform well on the NOISEX-92 database, where the training and test acoustic data is matched. The second scheme introduced, though not implemented, allows for differing acoustic data in the training and test conditions. The actual variation of true speech, for example on a sentence level, has not been investigated in this report, but will be examined in future work. This will give a clear indication whether such an iterative scheme is required in a real implementation.

Acknowledgement

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

References

[1] Acero A and Stern RM. Environmental robustness in automatic speech recognition. In *IEEE Proc. ICASSP*, pages 849–852, 1990.

[2] Berstein AD and Shallom ID. An hypothesized Wiener filtering approach to noisy speech recognition. In *Proceedings ICASSP*, pages 913–916, 1991.

[3] Varga AP, Steeneken HJM, Tomlinson M, and Jones D. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. In *Technical Report, DRA Speech Research Unit*, 1992.

[4] Varga AP and Moore RK. Hidden Markov model decomposition of speech and noise. In *Proceedings ICASSP*, pages 845–848, 1990.

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

[6] Mansour D and Juang BH. The short-time modified coherence representation and noisy speech recognition. In *IEEE Trans. Acoust., Speech Signal Processing*, volume 37, pages 795–804, 1989.

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

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

[9] Gales MJF and Young SJ. An improved approach to the hidden Markov model decomposition of speech and noise. In *Proceedings ICASSP*, 1992.

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

[11] Gales MJF and Young SJ. Parallel model combination for speech recognition in noise. *Technical Report CUED/F-INFENG/TR135*, 1993.

[12] Morgan N and Hermansky H. RASTA extensions: Robustness to additive and convolutional noise. In *Proceedings ESCA Workshop on Speech Processing in Adverse Conditions*, volume 67, pages 115–118, 1992.

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

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

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

[16] Young SJ. *HTK Version 1.4: Reference Manual and User Manual.* Cambridge University Engineering Department Speech Group, 1992.

[17] Beattie VL and Young SJ. Noisy speech recognition using hidden Markov model state based filtering. In *Proceedings ICASSP*, pages 917–920, 1991.

[18] Beattie VL and Young SJ. Hidden Markov model state-based cepstral noise compensation. In *Proceedings ICSLP*, pages 519–522, 1992.

A Numerical Integration

As mentioned in the text the parameter estimation may be performed using numerical integration. If implemented directly this would involve a two dimensional integration to estimate the mean and a four dimensional integration to estimate the covariance matrix. This would be computationally very expensive, so a more efficient form for the integration is required. Firstly it is necessary to select an appropriate numerical integration approximation. Given the form of the integration over Gaussian probability distributions, the numerical integration approximation used is

$$\int_{-\infty}^{\infty} f(x)\exp(-x^2)\,dx = \sum_{i=1}^{I} w_i f(x_i) \tag{18}$$

where x_i and w_i are given by the Gauss-Hermite abscissa and weights respectively. For simplicity of notation single Gaussian mixture distributions will be considered. The extension to multiple mixture Gaussian distributions is trivial.

Initially examining the estimation of the corrupted mean, rewriting

$$\mathcal{E}\{\log(\exp(S_i^l(t)) + \exp(N_i^l(t)))\} = \mathcal{E}\{N_i^l(t) + \log(\exp(S_i^l(t) - N_i^l(t)) + 1)\} \tag{19}$$

and noting that the sum or difference of two independent Gaussian distributions is itself Gaussian distributed then

$$\hat{\mu}_i^l = \tilde{\mu}_i^l + \int_{-\infty}^{\infty} \mathcal{N}(x; \mu_i^l - \tilde{\mu}_i^l, \Sigma_{ii}^l + \tilde{\Sigma}_{ii}^l)\log(\exp(x)+1)\,dx \tag{20}$$

where x denotes $S_i^l - N_i^l$. It is a simple transformation to convert the above equation to the correct form.

Having dealt with the mean, it is necessary to obtain the covariance estimate.

$$\begin{aligned}
&\mathcal{E}\{\log(\exp(S_i^l(t)) + \exp(N_i^l(t)))\log(\exp(S_j^l(t)) + \exp(N_j^l(t)))\} = \\
&\quad \mathcal{E}\{N_i^l(t)N_j^l(t) + N_i^l(t)\log(\exp(S_j^l(t) - N_j^l(t)) + 1) + N_j^l(t)\log(\exp(S_i^l(t) - N_i^l(t)) + 1) \\
&\quad + \log(\exp(S_i^l(t) - N_i^l(t)) + 1)\log(\exp(S_j^l(t) - N_j^l(t)) + 1)\}
\end{aligned} \tag{21}$$

The expected value of the above expression is required. When dealing with the leading diagonal terms the above expression may be simplified to

$$\mathcal{E}\{(O_i^l)^2\} = \mathcal{E}\{(N_i^l)^2 + 2N_i^l\log(\exp(S_i^l - N_i^l) + 1) + \left(\log(\exp(S_i^l - N_i^l) + 1)\right)^2\} \tag{22}$$

The first term is known. The second term is a directly implementable two dimensional integration with variables N_i^l and S_i^l. The final term is a single dimension integration having variable $S_i^l - N_i^l$. For the off diagonal terms examining the terms individually.

$$\mathcal{E}\{N_i^l(t)N_j^l(t)\} = \tilde{\Sigma}_{ij}^l + \tilde{\mu}_i^l\tilde{\mu}_j^l \tag{23}$$

by definition. Examining the second term

$$\mathcal{E}\{N_i^l(t)\log(\exp(S_j^l(t) - N_j^l(t)) + 1)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} N_i^l\log(\exp(x)+1)p(x, N_i^l)\,dx\,dN_i^l \tag{24}$$

where x denotes $S_j^l - N_j^l$. The speech and the noise are known to be uncorrelated. However, the various noise channels may be correlated depending on the form of the noise modelling. Thus $p(x, N_i^l)$ may have a full covariance matrix associated with it. This covariance matrix will be

$$\mathbf{\Sigma} = \begin{bmatrix} \Sigma_{ii}^l + \tilde{\Sigma}_{ii}^l & -\tilde{\Sigma}_{ij}^l \\ -\tilde{\Sigma}_{ij}^l & \tilde{\Sigma}_{jj}^l \end{bmatrix} \tag{25}$$

and the mean

$$\mu = \left[\begin{array}{cc} \mu_i^l - \tilde{\mu}_i^l & \tilde{\mu}_j^l \end{array} \right]^T \tag{26}$$

This new feature vector space may then be rotated to diagonalise $\mathbf{\Sigma}$ and then the standard numerical integration performed. The third term is of the same form as the second term. Examining the final term will yield an expression to be integrated over the variables $S_i^l - N_i^l$ and $S_j^l - N_j^l$. The covariance matrix for such an integration is given by

$$\mathbf{\Sigma} = \left[\begin{array}{cc} \Sigma_{ii}^l + \tilde{\Sigma}_{ii}^l & \Sigma_{ij}^l + \tilde{\Sigma}_{ij}^l \\ \Sigma_{ij}^l + \tilde{\Sigma}_{ij}^l & \Sigma_{jj}^l + \tilde{\Sigma}_{jj}^l \end{array} \right] \tag{27}$$

and the mean by

$$\mu = \left[\begin{array}{cc} \mu_i^l - \tilde{\mu}_i^l & \mu_j^l - \tilde{\mu}_j^l \end{array} \right]^T \tag{28}$$

Again the feature vector space may be rotated to allow integration over two independent variables. By breaking the integration up into four separate components it is possible to perform a series of two dimensional integrations, as opposed to the performing a direct four dimensional integration.