ISCA Archive
http://www.isca-speech.org/archive

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

HMM RECOGNITION IN NOISE USING PARALLEL MODEL COMBINATION

M.J.F. Gales *S.J. Young*
Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, England.

ABSTRACT

This paper addresses the problem of automatic speech recognition in the presence of interfering noise. The approach adopted is to compensate the parameters of a clean speech model given the statistics of the interfering noise. In this work these statistics are assumed to be modelled with a Hidden Markov Model. The basic theory of static coefficient Parallel Model Combination (PMC) is reviewed and placed within the framework of approximating the Maximum Likelihood (ML) estimate of the corrupted speech model, given the clean speech and interfering noise models. A new form of PMC is described which improves the performance of fixed grand variance based recognition schemes, by compensating the variance to be more representative of the corrupted speech fixed grand variance. In addition, the paper examines the problem of compensating delta coefficients in a PMC framework. Expressions for ML estimates of delta coefficients are derived and computationally efficient approximations of these estimates are given. The effectiveness of compensating delta parameters is discussed.

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

10.21437/Eurospeech.1993-192

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 noise robustness have been studied [11]. 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 [12, 10] spectral mapping [7] and inherently robust parameterisations [5]. 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 [6, 4], state based filtering [14], cepstral mean compensation [1, 15] and HMM decomposition [3].

This paper is concerned with the latter approach to noise robustness. In particular the scheme based on Parallel Model Combination (PMC) [8, 9]. In previous studies, PMC has been shown to provide robust recognition in a variety of noise conditions, both under simulated, as in the NOISEX-92 database, and under real noise conditions, as in the ARS ENST Car Noise database. Two forms of PMC have previously been described, a fixed grand variance scheme and a state based variance scheme. In this paper, a new form is investigated which uses a compensated fixed grand variance. Normally the fixed grand variance is estimated on the clean speech. However using PMC, it is possible to compensate the variance to be representative of the grand variance of the noisy data.

In addition, the paper details work into the compensation of delta coefficients within the framework of PMC. Presently, the theory behind PMC is only applicable to static coefficients. To achieve good recognition performance it is necessary to use delta coefficients in the speech parameterisation. At high SNR, it may be assumed that the delta coefficients are not significantly altered by the noise. However, as the SNR decreases this assumption becomes increasingly invalid and the use of uncompensated delta coefficient distributions may be detrimental to the recognition performance. A new technique is described which allows these parameters to be compensated in the PMC framework.

2. BASIC THEORY

The objective of any HMM based noise compensation scheme is to estimate, according to some objective function, the corrupted speech model given information about the clean speech and interfering noise. The objective function chosen here is a Maximum Likelihood (ML) criterion. The task is then to estimate the probability distribution of the corrupted speech, $\mathbf{O}^c(t)$, given the probability distributions for the clean speech, $\mathbf{S}^c(t)$, and interfering noise, $\mathbf{N}^c(t)$. Throughout 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-energy domain and $\mathbf{O}(t)$ is in the linear energy domain. Furthermore $\mathbf{O}^c(t)$ will represent the observation at time t, the associated random variable will be $\mathbf{O}^c$. All variables in bold are vectors or matrices, subscripts indicating elements of the vector or matrix. 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 [9].

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

$$\mu^l = \mathbf{C}^{-1}\mu^c \tag{1}$$

$$\mathbf{\Sigma}^l = \mathbf{C}^{-1}\mathbf{\Sigma}^c(\mathbf{C}^{-1})^T \tag{2}$$

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

If the corrupted speech is to be modelled by a standard HMM, then to obtain the ML estimate of the noise compensated speech model it is necessary to estimate the mean, $\hat{\mu}^l$, and covariance, $\hat{\mathbf{\Sigma}}^l$ of the corrupted speech. To calculate these values a series of assumptions are made.

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.

Using these assumptions the ML estimate of the mean and covariance are

$$\hat{\mu}_i^l = \int_{\mathcal{R}^n} d\mathbf{S}^l \int_{\mathcal{R}^n} d\mathbf{N}^l \log(g\exp(\mathbf{S}_i^l) + \exp(\mathbf{N}_i^l))p(\mathbf{S}^l)p(\mathbf{N}^l) \qquad (3)$$

and

$$\hat{\boldsymbol{\Sigma}}_{ij}^l = \int_{\mathcal{R}^n} d\mathbf{S}^l \int_{\mathcal{R}^n} d\mathbf{N}^l \Big\{ \log(g\exp(\mathbf{S}_i^l) + \exp(\mathbf{N}_i^l)) \log(g\exp(\mathbf{S}_j^l) + \exp(\mathbf{N}_j^l))p(\mathbf{S}^l)p(\mathbf{N}^l) \Big\} - \hat{\mu}_i^l\hat{\mu}_j^l \qquad (4)$$

where g is a gain matching term introduced to account for level differences between the clean speech and the noisy speech. In the above expressions a log() compression function has been used, as is normal with the use of cepstral coefficients. Appropriately modified expressions may be used with any compression function, however, the following approximations are specific to the log() compression. There is no closed form for either the compensated mean, $\hat{\mu}^l$ or covariance, $\hat{\boldsymbol{\Sigma}}^l$. So to obtain exact forms for the above expressions would require multidimensional numerical integration. However, if it is assumed that the sum of two lognormally distributed variables is itself approximately lognormally distributed then

$$\hat{\mu} = g\mu + \tilde{\mu} \qquad (5)$$

$$\hat{\boldsymbol{\Sigma}} = g^2\boldsymbol{\Sigma} + \tilde{\boldsymbol{\Sigma}} \qquad (6)$$

where the parameter set $\{\mu, \boldsymbol{\Sigma}\}$ are the mean and covariance respectively of the lognormal distribution associated with the Gaussian distribution $\{\mu^l, \boldsymbol{\Sigma}^l\}$ and similarly $\{\tilde{\mu}, \tilde{\boldsymbol{\Sigma}}\}$ and $\{\tilde{\mu}^l, \tilde{\boldsymbol{\Sigma}}^l\}$. The parameters of the clean speech are related by

$$\mu_i = \exp(\mu_i^l + \boldsymbol{\Sigma}_{ii}^l/2) \qquad (7)$$

$$\boldsymbol{\Sigma}_{ij} = \mu_i\mu_j\left[\exp(\boldsymbol{\Sigma}_{ij}^l) - 1\right] \qquad (8)$$

and similarly for the noise. Using the inverse of the above expressions $\{\hat{\mu}, \hat{\boldsymbol{\Sigma}}\}$ may be mapped back into $\{\hat{\mu}^l, \hat{\boldsymbol{\Sigma}}^l\}$. If cepstral parameters are to be used in the recognition stage, then we use the final mapping

$$\mu^c = \mathbf{C}\mu^l \qquad (9)$$

$$\boldsymbol{\Sigma}^c = \mathbf{C}\boldsymbol{\Sigma}^l\mathbf{C}^T \qquad (10)$$

This process can be viewed as an approximation to the ML estimate of a Gaussian distribution of the observed corrupted speech signal, $\mathbf{O}^c$ given the Gaussian distributions of the clean speech and interfering noise, with minimal computational overhead.

3. COMPENSATED FIXED GRAND VARIANCE

As previously described PMC has been used in two forms. A fixed grand variance scheme [8] and a state based compensated variance scheme [9] have been implemented. Although the performance of the state based compensated variance scheme is better than that of the fixed variance scheme it is computationally more expensive. In order to improve the performance of the fixed variance scheme, whilst introducing no computational overhead, a new version of PMC is proposed, the compensated fixed variance scheme. The fixed grand variance scheme uses the means calculated by PMC and the fixed variance computed from the clean speech. However using PMC it is possible to estimate a fixed grand variance based on the corrupted speech. It is then possible to use this new estimate of the fixed grand variance in the compensated models.

A single Gaussian mixture, single state HMM is generated using all the clean speech training data. The variance of this model is the clean speech grand variance and the mean is the global mean of the clean speech. In addition a single Gaussian mixture, single state noise model is generated. Using the standard assumptions behind PMC, as described in the previous section, it is possible to combine the two models to form a composite global corrupted speech model. The variance of this new model should be more representative of the grand variance of the corrupted speech signal. Hence the use of this new compensated grand variance should yield improved results over the use of an uncompensated grand variance.

4. DELTA COEFFICIENT COMPENSATION THEORY

For large vocabulary speech recognition it is necessary to incorporate dynamic coefficients in the speech parameterisation to achieve good recognition performance. The basic theory for PMC has relied on the fact that the speech and noise are additive. Hence the corrupted speech signal is a simple combination of the speech and noise signal. When dynamic coefficients are used this simple combination is not possible. If the speech is now parameterised using

$$\mathbf{O}^{\Delta c}(t)^T = \left[\mathbf{O}^c(t)^T, \Delta\mathbf{O}^c(t)^T\right] \qquad (11)$$

where $\Delta\mathbf{O}^c(t)$ are the simplest form of dynamic coefficients, delta coefficients, then

$$\begin{aligned} \Delta\mathbf{O}^c(t) &= (\mathbf{O}^c(t+1) - \mathbf{O}^c(t-1)) \\ &= \mathbf{C}(\mathbf{O}^l(t+1) - \mathbf{O}^l(t-1)) \\ &= \mathbf{C}\log\left[(\mathbf{O}(t+1)) ./ (\mathbf{O}(t-1))\right] \qquad (12) \end{aligned}$$

where $./$ is elementwise division. Using the assumption that the speech and noise are additive, $\mathbf{O}(t) = \mathbf{S}(t) + \mathbf{N}(t)$, and substituting this in the above equation

$$\Delta\mathbf{O}^c(t) = \mathbf{C}\log\left[(\mathbf{S}(t+1) + \mathbf{N}(t+1)) ./ (\mathbf{S}(t-1) + \mathbf{N}(t-1))\right] \qquad (13)$$

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

$$\begin{aligned} \Delta\mathbf{O}_i(t) & \qquad (14) \\ &= \left(\frac{\mathbf{S}_i(t+1)}{\mathbf{S}_i(t-1) + \mathbf{N}_i(t-1)}\right) + \left(\frac{\mathbf{N}_i(t+1)}{\mathbf{S}_i(t-1) + \mathbf{N}_i(t-1)}\right) \\ &= \Delta\mathbf{S}_i(t)\left(\frac{\frac{\mathbf{S}_i(t-1)}{\mathbf{N}_i(t-1)}}{\frac{\mathbf{S}_i(t-1)}{\mathbf{N}_i(t-1)} + 1}\right) + \Delta\mathbf{N}_i(t)\left(\frac{1}{\frac{\mathbf{S}_i(t-1)}{\mathbf{N}_i(t-1)} + 1}\right) \end{aligned}$$

The corrupted speech cepstral delta coefficients have been rewritten in terms of the static and delta coefficients of the clean speech and interfering noise. Examining the observation time the delta coefficient at time t is dependent on the static coefficients at time $t-1$. 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 $\mathbf{S}(t-1)$ will be approximately the same as those of $\mathbf{S}(t)$ and $\mathbf{N}(t-1)$ the same as $\mathbf{N}(t)$. With this assumption statistics exist for all the variables of equation 14.

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

$$(\mu^{\Delta l})^T = \left[(\mathbf{C}^{-1}\mu^c)^T, (\mathbf{C}^{-1}\Delta\mu^c)^T\right] \tag{15}$$

and

$$\Sigma^{\Delta l} = \begin{bmatrix} \mathbf{C}^{-1}\Sigma^c(\mathbf{C}^{-1})^T & \mathbf{C}^{-1}\delta\Sigma^c(\mathbf{C}^{-1})^T \\ \mathbf{C}^{-1}(\delta\Sigma^c)^T(\mathbf{C}^{-1})^T & \mathbf{C}^{-1}\Delta\Sigma^c(\mathbf{C}^{-1})^T \end{bmatrix} \tag{16}$$

where $\delta\Sigma^c$ is the covariance matrix representing the correlation between the static and delta coefficients. A similar mapping converts the noise parameters $\{\tilde{\mu}^{\Delta c}, \tilde{\Sigma}^{\Delta c}\}$ to $\{\tilde{\mu}^{\Delta l}, \tilde{\Sigma}^{\Delta l}\}$. The ML estimates of the static coefficients are unaltered. Those for the delta coefficients are given by

$$\Delta\hat{\mu}_i^l = \int_{\mathcal{R}^n} d\mathbf{S}^l \int_{\mathcal{R}^n} d\mathbf{N}^l \int_{\mathcal{R}^n} d\Delta\mathbf{S}^l \int_{\mathcal{R}^n} d\Delta\mathbf{N}^l \tag{17}$$

$$p(\mathbf{S}^l, \Delta\mathbf{S}^l)p(\mathbf{N}^l, \Delta\mathbf{N}^l)\log\left(\gamma_i \exp(\Delta\mathbf{S}_i^l) + \eta_i \exp(\Delta\mathbf{N}_i^l)\right)$$

and

$$\Delta\hat{\Sigma}_{ij}^l = \int_{\mathcal{R}^n} d\mathbf{S}^l \int_{\mathcal{R}^n} d\mathbf{N}^l \int_{\mathcal{R}^n} d\Delta\mathbf{S}^l \int_{\mathcal{R}^n} d\Delta\mathbf{N}^l \tag{18}$$

$$\Big\{p(\mathbf{S}^l, \Delta\mathbf{S}^l)p(\mathbf{N}^l, \Delta\mathbf{N}^l)\log\left(\gamma_i \exp(\Delta\mathbf{S}_i^l) + \eta_i \exp(\Delta\mathbf{N}_i^l)\right)$$

$$\log\left(\gamma_j \exp(\Delta\mathbf{S}_j^l) + \eta_j \exp(\Delta\mathbf{N}_j^l)\right)\Big\} - \Delta\hat{\mu}_i^l\Delta\hat{\mu}_j^l$$

and

$$\delta\hat{\Sigma}_{ij}^l = \int_{\mathcal{R}^n} d\mathbf{S}^l \int_{\mathcal{R}^n} d\mathbf{N}^l \int_{\mathcal{R}^n} d\Delta\mathbf{S}^l \int_{\mathcal{R}^n} d\Delta\mathbf{N}^l \tag{19}$$

$$\Big\{p(\mathbf{S}^l, \Delta\mathbf{S}^l)p(\mathbf{N}^l, \Delta\mathbf{N}^l)\log\left(g\exp(\mathbf{S}_i^l) + \exp(\mathbf{N}_i^l)\right)$$

$$\log\left(\gamma_j \exp(\Delta\mathbf{S}_j^l) + \eta_j \exp(\Delta\mathbf{N}_j^l)\right)\Big\} - \hat{\mu}_i^l\Delta\hat{\mu}_j^l$$

where

$$\gamma_i = \left(\frac{\exp(\mathbf{S}_i^l - \mathbf{N}_i^l)}{\exp(\mathbf{S}_i^l - \mathbf{N}_i^l) + 1}\right) \tag{20}$$

$$\eta_i = \left(\frac{1}{\exp(\mathbf{S}_i^l - \mathbf{N}_i^l) + 1}\right) \tag{21}$$

If diagonal covariance matrices for the corrupted speech are to be estimated it is not necessary to estimate $\delta\hat{\Sigma}^l$. To calculate the full forms of equation 17 and equation 18 again requires multi-dimensional numerical integration, which is computationally expensive. However by making an additional assumption that the variances on γ and η are negligible then the form of the ML estimates of the delta parameters are the same as those of the static coefficients. Hence

$$\Delta\hat{\mu}_i = \bar{\gamma}_i\Delta\mu_i + \bar{\eta}_i\Delta\tilde{\mu}_i \tag{22}$$

$$\Delta\hat{\Sigma}_{ij} = \bar{\gamma}_i\bar{\gamma}_j\Delta\Sigma_{ij} + \bar{\eta}_i\bar{\eta}_j\Delta\tilde{\Sigma}_{ij} \tag{23}$$

where

$$\mathcal{E}[\gamma_i] = \mathcal{E}\left[\frac{\frac{\mathbf{S}_i}{\mathbf{N}_i}}{\frac{\mathbf{S}_i}{\mathbf{N}_i} + 1}\right] \approx \left(\frac{\frac{\mu_i}{\tilde{\mu}_i}}{\frac{\mu_i}{\tilde{\mu}_i} + 1}\right) = \bar{\gamma}_i \tag{24}$$

and

$$\mathcal{E}[\eta_i] = \mathcal{E}\left[\frac{1}{\frac{\mathbf{S}_i}{\mathbf{N}_i} + 1}\right] \approx \left(\frac{1}{\frac{\mu_i}{\tilde{\mu}_i} + 1}\right) = \bar{\eta}_i \tag{25}$$

The mean and covariance can now be mapped back into the cepstral domain in a similar way to the static coefficients

$$\Delta\mu^c = \mathbf{C}\Delta\mu^l \tag{26}$$

$$\Delta\Sigma^c = \mathbf{C}\Delta\Sigma^l\mathbf{C}^T \tag{27}$$

5. EVALUATION ON NOISEX-92

In this section a number of experiments using the NOISEX-92 database [2] 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. Where delta coefficients are used all 15 delta MFCC are calculated.

For each digit, a single mixture continuous density HMM with 8 emitting states was trained using the clean data only. 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 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 [13] with suitable extensions to perform PMC.

SNR	*Lynx*	*F16*	*Car*
dB	BFGV	BFGV	BFGV
-06	20	11	15
+00	25	28	36
+06	59	47	80
+12	91	78	100
+18	100	88	100

Table 1. Baseline Fixed Grand Variance Performance

Table 1 shows the baseline performance with no noise compensation and a fixed grand variance.

SNR	*Lynx*		*F16*		*Car*	
dB	FGV	CFGV	FGV	CFGV	FGV	CFGV
-06	30	35	38	47	49	60
+00	68	75	87	86	85	89
+06	97	97	98	98	96	96
+12	100	100	100	100	100	100
+18	100	100	100	100	100	100

Table 2. PMC Fixed Grand Variance Performance

Two forms of PMC were then compared. The first used PMC to compensate the means of the models, but used the fixed grand variance computed from the clean speech, *FGV*. The second version of PMC again used PMC to compensate the means. However it used the grand variance based on the PMC estimate, *CFGV*. The results are shown in table 2.

For both *FGV* and *CFGV* variance schemes the performance is better than baseline performance. At higher signal to noise ratios, $> 06dB$, the performance of both the schemes is the same. However, at lower signal to noise ratios, where the grand variance will be most distorted by the interfering noise, the compensation shows a marked improvement.

In order to investigate the use of delta coefficient compensation in the PMC framework, it was decided to use only the delta coefficients in the recognition. If static coefficients are incorporated they tend to dominate the recognition, having typically around 90% accuracy at $+00dB$ on this database. The first set of experiments on the delta coefficients were run assuming that the true corrupted speech variance was known. Thus only the delta coefficient means were compensated and the variance was set to the true variance of the corrupted speech. Table 3 shows the baseline performance of the clean delta coefficient

SNR	Lynx		F16		Car	
dB	Clean	Noisy	Clean	Noisy	Clean	Noisy
-06	26	29	13	22	28	33
+00	29	54	28	59	34	43
+06	66	82	77	90	68	82
+12	90	97	97	99	84	87
+18	95	97	100	100	95	98

Table 3. Baseline Delta Mean Performance

means, *Clean*, and the true corrupted speech means, *Noisy*. The performance of the *Clean* parameters drops off rapidly below +12dB, whilst the *Noisy* parameters achieve good performance down to +06dB.

SNR	Lynx		F16		Car	
dB	Fast	Num.	Fast	Num.	Fast	Num.
-06	30	29	27	25	35	33
+00	44	44	62	61	53	43
+06	82	84	88	89	79	81
+12	95	95	99	99	88	88
+18	97	97	100	100	96	97

Table 4. Compensated Delta Mean Performance

Table 4 shows the performance of the compensated means. Two compensation schemes were examined. The first used the approximation given in equation 22, labelled *Fast*. Secondly, the true ML estimate, given in equation 17, was implemented, labelled *Num.*, using Gaussian numerical integration. Again the variances were taken from the true variances of the corrupted speech. The *Fast* and *Num.* performance is approximately the same and comparable to the training of the parameters in noise, *Noisy*, in table 3.

SNR	Lynx		F16		Car	
dB	Base	Comp	Base	Comp	Base	Comp
-06	20	32	8	20	25	29
+00	27	44	28	53	37	37
+06	49	64	34	82	61	50
+12	84	80	88	96	86	71
+18	97	94	96	98	95	88

Table 5. Full Delta Compensation Performance

Table 5 shows a comparison of uncompensated delta coefficients, *Base*, with compensated delta coefficients, *Comp*. For the *Comp* scheme both means and variances were estimated using the fast approximation. The performance of both schemes are worse than if the true variance of the delta coefficients is used.

6. CONCLUSIONS

A new implementation of PMC for static coefficient compensation has been proposed for situations where a fixed grand variance is desired. It has been shown that given the statistics of the interfering noise, a new fixed grand variance can be estimated, which better approximates the fixed grand variance of the corrupted speech. This improves performance at no recognition time overhead compared to a standard fixed grand variance scheme.

Maximum likelihood estimates for both delta and static coefficients have been derived and computationally efficient approximations for these expressions have been implemented. The estimation scheme for the compensated delta means has been found to achieve good performance. However, the performance when the compensated variances are used is not as good, indicating a poor estimation of the corrupted speech delta coefficient variance. Methods of improving the estimates of the delta variances are being investigated.

ACKNOWLEDGEMENT

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

REFERENCES

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

[2] 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.

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

[4] 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.

[5] 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.

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

[7] 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.

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

[9] Gales MJF and Young SJ. Cepstral parameter compensation for hmm recognition in noise. *Speech Communication*, 1993. To be published.

[10] 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.

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

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

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

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

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