16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

COMPARING NOISE COMPENSATION METHODS FOR ROBUST PREDICTION OF ACOUSTIC SPEECH FEATURES FROM MFCC VECTORS IN NOISE

Ben Milner[1], *Jonathan Darch*[1], *Ibrahim Almajai*[1] *and Saeed Vaseghi*[2]

[1]School of Computing Sciences, University of East Anglia, UK
[2]Dept. of Electronic and Computer Engineering, Brunel University, UK
email: {b.milner, jonathan.darch, i.almajai}@uea.ac.uk, saeed.vaseghi@brunel.ac.uk

ABSTRACT

The aim of this paper is to investigate the effect of applying noise compensation methods to acoustic speech feature prediction from MFCC vectors, as may be required in a distributed speech recognition (DSR) architecture. A brief review is made of maximum a posteriori (MAP) prediction of acoustic features from MFCC vectors using both global and phoneme-specific modeling of speech. The application of spectral subtraction and model adaptation to MAP acoustic feature prediction is then introduced. Experimental results are presented to compare the effect of noise compensation on acoustic feature prediction accuracy using both the global and phoneme-specific systems. Results across a range of signal-to-noise ratios show model adaptation to be better than spectral subtraction and able to restore performance close to that achieved in matched training and testing.

1. INTRODUCTION

There has been considerable interest in using distributed speech recognition (DSR) for applications operating over mobile and IP networks. The first version of the ETSI Aurora DSR standard specified MFCC feature extraction on the terminal device to provide a stream of feature vectors to the remote back-end at a bit rate of 4800bps [1]. The standard was later updated to include transmission of voicing and fundamental frequency which increased the bit rate to 5600bps [2]. The primary motivation for transmitting this extra information was to allow audio speech to be reconstructed at the remote back-end. Within the Aurora standard, speech is reconstructed using a sinusoidal model that uses an MFCC-derived spectral envelope and harmonic information derived from the voicing and fundamental frequency. A further application of the fundamental frequency is to enhance speech recognition in tonal languages such as Mandarin and Cantonese.

In our recent work we have examined the correlation between MFCC vectors and acoustic speech features (voicing, speech/non-speech, fundamental frequency and formant frequencies) [3]. This led to a maximum a posteriori (MAP) method of predicting the acoustic features of a frame of speech from its MFCC vector representation. Such a scheme removed the need to transmit additional voicing and fundamental frequency and, by using the predicted voicing and fundamental frequency, allows speech to be reconstructed solely from the MFCC vectors [4].

The aim of this work is to extend previous work by improving acoustic speech feature prediction accuracy from MFCC vectors that are extracted from noisy speech. Without noise compensation, prediction accuracy of the acoustic features deteriorates as the signal-to-noise ratios (SNR) decreases. This deterioration is attributed to the noise distorting the MFCC vectors, which moves their statistics away from the distributions in the statistical model. To improve prediction accuracy in noise it is necessary to remove the mismatch between the training data derived distributions in the model and the input noisy MFCCs.

Many methods have been proposed in the area of robust speech recognition to reduce the mismatch between clean trained speech models and noisy input speech features [5,6]. In this work, two such methods are examined. The first removes (or filters) noise from the input MFCCs to match them to the clean-trained distributions in the models. Examples of this include spectral subtraction, Wiener filtering, etc. The second method involves modifying the clean-trained distributions in the models to model noise contaminated speech. This has the advantage of not only compensating for mean shifts (as filtering does) but also compensating for changes in variance. Examples of this include model adaptation and matched condition training.

The remainder of this paper is arranged as follows. Section 2 gives a brief review of MAP prediction of acoustic speech features from MFCC vectors using both a global and a phoneme-specific method of speech modeling. The application of the two noise compensation methods to MAP prediction of acoustic features from noisy MFCC vectors is presented in section 3. Section 4 presents experimental results which examine the effectiveness of the noise compensation methods for both the global and phoneme-specific methods of acoustic feature prediction.

2. ACOUSTIC FEATURE PREDICTION

This section briefly describes the procedure for predicting acoustic speech features (voicing, fundamental frequency, formant frequencies, speech/non-speech) from MFCC vectors [3][4]. This begins by first modeling the joint density of MFCC vectors and acoustic speech features. Secondly, using the joint density, a maximum a posteriori (MAP) prediction of acoustic features can be made from an input MFCC vector. Two methods of modelling the joint density are considered. The first utilises phoneme-specific models whereby joint densities are created for each phoneme. A second, more simple alternative, uses a single joint density for all speech sounds.

2.1 Modeling phoneme-specific joint densities

Three stages are involved in the phoneme-specific modelling of the joint density of acoustic features and MFCC vectors. First, a set of phoneme-based hidden Markov models (HMMs) are trained. Second, joint feature vectors, specific to each state of each HMM are pooled into voiced, unvoiced and non-speech vector pools. Finally, voiced, unvoiced and non-speech Gaussian mixture models (GMMs) are trained which model the state and model specific joint densities of acoustic features and MFCC vectors.

2.1.1 HMM training

To identify each phoneme in the phoneme-specific prediction, a set of 44 phoneme HMMs and a silence HMM are created and incorporated into an unconstrained phoneme grammar. Each HMM has 3 states with 8 modes per state and diagonal covariance matrices. The static feature vector used comprises MFCCs 0 to 12 and this augmented with velocity and acceleration derivatives [1].

2.1.2 Phoneme-specific vector pools

For each state, *s*, of each HMM, *w*, vector pools, $\Psi^{v}_{s,w}$, $\Psi^{u}_{s,w}$ and $\Psi^{ns}_{s,w}$, corresponding to voiced speech, unvoiced speech and non-speech are created. These are created by force aligning the training data utterances to the correct sequence of phoneme HMMs using Viterbi decoding and reference annotations. For each training data utterance, $\mathbf{X}$=[$\mathbf{x}_1$, $\mathbf{x}_2$, .., $\mathbf{x}_N$], comprising *N* MFCC vectors, this provides a model allocation, $\mathbf{m}$=[m_1, m_2, …, m_N], and state allocation, $\mathbf{q}$=[q_1, q_2, …, q_N]. Therefore, for the t^{th} static MFCC vector, $\mathbf{x}_t$ the model allocation, m_t, and state allocation, q_t, together with the reference voicing, indicate to which vector pool the vector should be allocated. Reference voicing classifications are made by the voicing classifier in the ETSI front-end [1].

Vector pools are created using all vectors in the training set. Each static MFCC vector is also augmented by its corresponding acoustic feature vector, $\mathbf{f}_t$, to create vector pools of joint feature vectors, $\mathbf{z}_t$, where,

$$\mathbf{z}_t = \left[\mathbf{x}_t, \mathbf{f}_t\right] \qquad (1)$$

The acoustic vector, $\mathbf{f}_t$, comprises the fundamental frequency, F0, and first four formant frequencies, F1 to F4, i.e. $\mathbf{f} = \left[F0, F1, F2, F3, F4\right]$. In non-speech, both fundamental frequency and formant frequencies are zero. During unvoiced speech the fundamental frequency is zero while formant frequencies take non-zero values. For voiced speech, both fundamental frequency and formant frequencies are non-zero.

2.1.3 Phoneme-specific GMMs

The state and model specific vectors pools can now be used to create state and model specific GMMs that model the joint density of acoustic features and MFCC vectors. This is achieved by applying expectation-maximisation (EM) clustering to each vector pool to create voiced, $\Phi^{v,\mathbf{z}}_{s,w}$, unvoiced, $\Phi^{u,\mathbf{z}}_{s,w}$, and non-speech, $\Phi^{ns,\mathbf{z}}_{s,w}$, GMMs for each state *s* of each phoneme model *w*. For example, the voiced GMM is represented as,

$$p(\mathbf{z}_t) = \Phi^{v,\mathbf{z}}_{s,w}(\mathbf{z}_t) = \sum_{k=1}^{K^v} \alpha^{v}_{k,s,w}\, N\left(\mathbf{z}_t; \mu^{v,\mathbf{z}}_{k,s,w}, \Sigma^{v,\mathbf{zz}}_{k,s,w}\right) \qquad (2)$$

where K^v is the number of clusters in the voiced GMM. $\mu^{v,\mathbf{z}}_{k,s,w}$, $\Sigma^{v,\mathbf{zz}}_{k,s,w}$ and $\alpha^{v}_{k,s,w}$ are the mean vector, covariance matrix and prior probability in the k^{th} cluster of the voiced GMM for state *s* of model *w*. The means and covariances can be decomposed into their MFCC vector and acoustic vector components as,

$$\boldsymbol{\mu}^{v,\mathbf{z}}_{k,s,w} = \begin{bmatrix} \boldsymbol{\mu}^{v,\mathbf{x}}_{k,s,w} \\ \boldsymbol{\mu}^{v,\mathbf{f}}_{k,s,w} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}^{v,\mathbf{z}}_{k,s,w} = \begin{bmatrix} \boldsymbol{\Sigma}^{v,\mathbf{xx}}_{k,s,w} & \boldsymbol{\Sigma}^{v,\mathbf{xf}}_{k,s,w} \\ \boldsymbol{\Sigma}^{v,\mathbf{fx}}_{k,s,w} & \boldsymbol{\Sigma}^{v,\mathbf{ff}}_{k,s,w} \end{bmatrix} \qquad (3)$$

Phoneme-specific GMMs are also created for unvoiced speech and non-speech, $\Phi^{u,\mathbf{z}}_{s,w}$ and $\Phi^{ns,\mathbf{z}}_{s,w}$.

2.2 Prediction of acoustic features

The first stage in predicting acoustic features from a stream of MFCC vectors is to determine their state and phoneme model sequence using Viterbi decoding. Secondly, for each MFCC vector, a state and model specific voicing classification is made. Finally, for MFCC vectors classified as voiced, formant and fundamental frequencies are predicted, while for unvoiced MFCC vectors only formant frequencies are predicted.

2.2.1 Voicing prediction

From an input stream of MFCC vectors, $\mathbf{X}$=[$\mathbf{x}_1$, $\mathbf{x}_2$, .., $\mathbf{x}_N$], Viterbi decoding is used to determine their state and model sequence, $\mathbf{q}$=[q_1, q_2, …, q_N] and $\mathbf{m}$=[m_1, m_2, …, m_N]. The probability of each MFCC vector from the voiced, unvoiced and non-speech GMMs is then computed and used to make a voicing prediction,

$$voicing_t = \begin{cases} voiced & \Phi^{v,\mathbf{x}}_{q_t,m_t}(\mathbf{x}_t) \geq \Phi^{u,\mathbf{x}}_{q_t,m_t}(\mathbf{x}_t) \quad and \quad \Phi^{v,\mathbf{x}}_{q_t,m_t}(\mathbf{x}_t) \geq \Phi^{ns,\mathbf{x}}_{q_t,m_t}(\mathbf{x}_t) \\ unvoiced & \Phi^{u,\mathbf{x}}_{q_t,m_t}(\mathbf{x}_t) \geq \Phi^{v,\mathbf{x}}_{q_t,m_t}(\mathbf{x}_t) \quad and \quad \Phi^{u,\mathbf{x}}_{q_t,m_t}(\mathbf{x}_t) \geq \Phi^{ns,\mathbf{x}}_{q_t,m_t}(\mathbf{x}_t) \\ non-speech & otherwise \end{cases} \qquad (4)$$

$\Phi^{v,\mathbf{x}}_{q_t,m_t}$, $\Phi^{u,\mathbf{x}}_{q_t,m_t}$ and $\Phi^{ns,\mathbf{x}}_{q_t,m_t}$ represent the voiced, unvoiced and non-speech GMMs associated with state q_t and model m_t which have been marginalised to the MFCC vector component, $\mathbf{x}$.

2.2.2 Acoustic feature prediction

For MFCC vectors classified as voiced, fundamental and formant frequencies are predicted. Using the state and model allocation for MFCC vector $\mathbf{x}_t$, a MAP prediction of the acoustic feature vector, $\hat{\mathbf{f}}_t(k)$, can be made from cluster *k* of the voiced GMM to which MFCC vector $\mathbf{x}_t$ is allocated, $\phi^{v,\mathbf{z}}_{k,q_t,m_t}$, as,

$$\hat{\mathbf{f}}_t(k) = \underset{\mathbf{f}_t}{\arg\max}\left(p\left(\mathbf{f}_t \middle| \mathbf{x}_t, \phi^{v,\mathbf{z}}_{k,q_t,m_t}\right)\right) \qquad (5)$$

The posterior probability, $h_{k,q_t,m_t}(\mathbf{x}_t)$, of the MFCC vector belonging to the k^{th} cluster of the GMM can be used to make a weighted MAP prediction from all K^v clusters,

$$\hat{\mathbf{f}}_t = \sum_{k=1}^{K^v} h_{k,q_t,m_t}(\mathbf{x}_t)\left(\mu^{v,\mathbf{f}}_{k,q_t,m_t} + \Sigma^{v,\mathbf{fx}}_{k,q_t,m_t}\left(\Sigma^{v,\mathbf{xx}}_{k,q_t,m_t}\right)^{-1}\left(\mathbf{x}_t - \mu^{v,\mathbf{x}}_{k,q_t,m_t}\right)\right) \qquad (6)$$

where the posterior probability is computed as,

$$h_{k,q_t,m_t}(\mathbf{x}_t) = \frac{\alpha^{v}_{k,q_t,m_t}\, p\left(\mathbf{x}_t \middle| \phi^{v,\mathbf{x}}_{k,q_t,m_t}\right)}{\sum_{k=1}^{K^v} \alpha^{v}_{k,q_t,m_t}\, p\left(\mathbf{x}_t \middle| \phi^{v,\mathbf{x}}_{k,q_t,m_t}\right)} \qquad (7)$$

$p\left(\mathbf{x}_t \middle| \phi^{v,\mathbf{x}}_{k,q_t,m_t}\right)$ is the marginal distribution of the MFCC vector for the k^{th} cluster in the voiced GMM.

For MFCC vectors classified as unvoiced, a similar procedure is followed using the unvoiced GMMs to predict only formants.

2.3 Global prediction of acoustic features

A more simple approach to predicting acoustic speech features from MFCC vectors is to use just one voiced, unvoiced and non-speech GMM which together model the joint density of MFCC vectors and acoustic features across all speech sounds. In effect this is a specific implementation of the phoneme-specific prediction, where just a single state HMM is used. This avoids having to decode the MFCC vectors into a state and model sequence as all acoustic features will be predicted from the same global state comprising voiced, unvoiced and non-speech GMMs. This simple alternative provides a useful comparison to the more sophisticated phoneme-specific methods and is evaluated in the experimental results.

3. NOISE COMPENSATION

Previous work has shown that acoustic feature prediction errors increase when noise contaminates the speech [3]. This is attributed to the noise distorting the MFCC vectors which leads to a mismatch with the joint densities trained on clean speech. In this work two noise compensation methods are applied to acoustic speech feature prediction to reduce this mismatch between clean trained models and noisy input MFCC vectors. The first method examined is spectral subtraction [5]. The second method adapts the statistics of the joint densities to model noisy speech. Such adaptation methods have been successfully applied to speech recognition systems to improve noise robustness [6]. The remainder of this section describes the application of spectral subtraction and model adaptation to acoustic feature prediction.

3.1 Spectral subtraction

To apply spectral subtraction, the MFCC vectors received at the DSR back-end must be returned to the linear spectral domain where speech and noise are additive. The MFCC vectors are first zero padded to the dimensionality, J, of the log filterbank and an inverse discrete cosine transform (DCT) applied to obtain a log filterbank vector, $\mathbf{x}_t^{lfb}$,

$$\mathbf{x}_t^{lfb} = \mathbf{C}^{-1}\mathbf{x}_t \tag{8}$$

Matrix $\mathbf{C}$ contain the basis vectors of the DCT, where each element c_{ij} is given as,

$$c_{ij} = \cos\left[\frac{i\pi(j+0.5)}{J}\right] \quad 0 \le i, j \le J-1 \tag{9}$$

Applying an exponential gives linear filterbank vectors, $\mathbf{x}_t^{fb}$,

$$\mathbf{x}_t^{fb} = \exp\left(\mathbf{x}_t^{lfb}\right) \tag{10}$$

In this application it is unnecessary to return the filterbank vector to a magnitude or power spectrum for subtraction. In fact the wider bandwidths of filterbank channels, over those of the spectral bins, gives more stability and reduces the chance of processing distortion as a result of over subtraction. Of the many variants of spectral subtraction, this work uses linear subtraction with an over-subtraction factor, α. Spectral distortion is reduced by a maximum attenuation threshold, β, rather than a noise floor. The clean speech filterbank estimate, $\hat{s}_t^{fb}(i)$, for the i^{th} channel of the t^{th} frame is given as,

$$\hat{s}_t^{fb}(i) = \begin{cases} x_t^{fb}(i) - \alpha\hat{d}^{fb}(i) & x_t^{fb}(i) - \alpha\hat{d}^{fb}(i) > \beta x_t^{fb}(i) \\ \beta x_t^{fb}(i) & otherwise \end{cases} \tag{11}$$

where $\hat{d}^{fb}(i)$ is the noise estimate in the i^{th} filterbank channel. This is estimated in speech inactive periods and computed from received MFCC vectors using an inverse DCT and exponential operation. The clean speech filterbank estimate, $\hat{\mathbf{s}}^{fb}$, is transformed back to the MFCC domain using log, DCT and truncation operations. The resulting noise-reduced MFCC vector is then input into the acoustic feature prediction system.

3.2 Model adaptation

The second noise compensation method adapts the statistics of each of the phoneme-specific voiced, unvoiced and non-speech GMMs to model noise contaminated MFCC vectors. Considering equation 3, the MFCC mean vectors and covariances, $\mu_k^{v,\mathbf{x}}$ and $\Sigma_k^{v,\mathbf{xx}}$, need to be adapted to the noise. Note, for clarity, the state and model indices, q_t and m_t, have been dropped from the notation. The acoustic feature means and covariances, $\mu_k^{v,\mathbf{f}}$ and $\Sigma_k^{v,\mathbf{ff}}$, are independent of the noise and left unchanged. Similarly, the covariances of MFCCs and acoustic features, $\Sigma_k^{v,\mathbf{xf}}$ and $\Sigma_k^{v,\mathbf{fx}}$, can be left unchanged as the noise is uncorrelated with the acoustic features.

The MFCC means and covariances must be adapted so that instead of modeling clean speech they model noisy speech. To allow adaptation, the MFCC-domain means and covariances must be inverted to the linear filterbank domain where speech and noise are additive. First, the MFCC-domain means and covariances are zero padded and inverse DCTs applied to obtain log filterbank domain means and covariances, $\mu_k^{v,\mathbf{x},lfb}$ and $\Sigma_k^{v,\mathbf{xx},lfb}$,

$$\mu_k^{v,\mathbf{x},lfb} = \mathbf{C}^{-1}\mu_k^{v,\mathbf{x}} \qquad \Sigma_k^{v,\mathbf{xx},lfb} = \mathbf{C}^{-1}\Sigma_k^{v,\mathbf{xx}}\left(\mathbf{C}^{-1}\right)^T \tag{12}$$

It is assumed that MFCC vectors exhibit a Gaussian distribution which is also true in the log filterbank domain. However, in the linear filterbank domain the vectors are log normal. The log filterbank means and covariances can be transformed into the linear filerbank domain, $\mu_k^{v,\mathbf{x},fb}$ and $\Sigma_k^{v,\mathbf{xx},fb}$, [6], as,

$$\mu_k^{v,\mathbf{xx},fb}(i) = \exp\left\{\mu_k^{v,\mathbf{xx},lfb}(i) + \frac{diag\left(\Sigma_k^{v,\mathbf{xx},lfb}(i,i)\right)}{2}\right\} \tag{13}$$

$$\Sigma_k^{v,\mathbf{xx},fb}(i,j) = \mu_k^{v,\mathbf{x},fb}(i)\,\mu_k^{v,\mathbf{x},fb}(j)\exp\left\{\Sigma_k^{v,\mathbf{xx},lfb}(i,j) - 1\right\} \tag{14}$$

The linear filterbank means and covariances of noisy speech, $\mu_k^{v,\mathbf{y},fb}$ and $\Sigma_k^{v,\mathbf{yy},fb}$, are computed by adding the clean speech means and covariances to the noise mean and covariance, $\mu^{\mathbf{d},fb}$ and $\Sigma^{\mathbf{dd},fb}$,

$$\mu_k^{v,\mathbf{y},fb} = \mu_k^{v,\mathbf{x},fb} + \mu^{\mathbf{d},fb} \qquad \Sigma_k^{v,\mathbf{yy},fb} = \Sigma_k^{v,\mathbf{xx},fb} + \Sigma^{\mathbf{dd},fb} \tag{15}$$

The noise mean and covariance are provided by a single cluster GMM that has been trained from non-speech periods.

The noisy filterbank means and covariances can be transformed into the MFCC domain using the inverse of equations 13 and 14. Finally, the noisy log filterbank means and covariances are transformed to the MFCC domain, $\mu_k^{v,\mathbf{y}}$ and $\Sigma_k^{v,\mathbf{yy}}$,

$$\mu_k^{v,\mathbf{y}} = \mathbf{C}\mu_k^{v,\mathbf{y},lfb} \qquad \Sigma_k^{v,\mathbf{yy}} = \mathbf{C}\Sigma_k^{v,\mathbf{yy},lfb}\mathbf{C}^T \tag{16}$$

These noisy MFCC means and covariances replace the clean speech means and covariances, $\mu_k^{v,\mathbf{x}}$ and $\Sigma_k^{v,\mathbf{xx}}$, in equation 3. Similar adaptations are made for the means and covariances in the unvoiced and non-speech GMMs.

4. EXPERIMENTAL RESULTS

These experiments investigate the effectiveness of the noise compensation methods when applied to acoustic feature prediction from MFCC vectors in noise. Their effectiveness is examined for both the global (GMM) and phoneme-specific (HMM-GMM) based methods of acoustic feature prediction.

The experiments use a speaker-dependent speech database recorded from a single female US English speaker. This comprises 589 sentences for training and further 246 sentences for testing. This provides a test set of approximately 130,000 vectors. Reference fundamental frequency and voicing is obtained from a laryngograph. Speech/non-speech classification is derived from hand corrected phoneme annotations. Formant frequencies were obtained from a combined linear predictive-Kalman filtering approach [7]. The speech was sampled at 8kHz and 13-D MFCC vectors extracted from 25ms frames at a rate of 100 vectors per second in accordance with the ETSI Aurora standard [1].

The first set of experiments examine the effectiveness of noise compensation on voicing classification and fundamental frequency prediction. This is followed by a second set of experiments which apply noise compensation to speech/non-speech classification and formant frequency prediction.

4.1 Voicing and fundamental frequency prediction

This section examines the effectiveness of the noise compensation methods on voicing and fundamental frequency prediction using both global and phoneme-specific prediction. Before presenting experimental results, the measures used to evaluate voicing and fundamental frequency prediction errors must be defined. The accuracy of identifying voiced frames is measured using the percentage voicing classification error, E_{vc}, defined as,

$$E_{vc} = \frac{N_{v|nv} + N_{nv|v}}{N_T} \times 100\% \qquad (17)$$

$N_{v|nv}$ is the number of unvoiced or non-speech vectors that are incorrectly classified as voiced, $N_{nv|v}$ is the number of voiced vectors that are incorrectly classified and N_T is the total number of vectors in the test set. Fundamental frequency prediction is measured using the percentage fundamental frequency error, E_p,

$$E_p = \frac{1}{N_V} \sum_{t=1}^{N_V} \frac{\left|\hat{F0}_t - F0_t\right|}{F0_t} \times 100\% \qquad (18)$$

$\hat{F0}_t$ and $F0_t$ are the predicted and reference fundamental frequency of the t^{th} frame. E_p is measured for all N_v frames labelled as voiced according to the reference voicing. This ensures voicing classification errors do not influence E_p .

Table 1 shows voicing classification error, E_{vc}, and fundamental frequency error, E_p, obtained using the global (GMM) system. The table shows prediction accuracies in clean speech and at SNRs of 20dB, 10dB and 0dB in white noise. The columns of the table show results for no noise compensation (NNC), spectral subtraction (SS) and model adaptation. These are all based on clean speech trained GMMs. To indicate likely best performance in noise, the final column (Match) shows performance when the GMMs are trained and tested in the same matched noise conditions. In practice matched condition training and testing is not feasible but it does provide a guide to best performance. Table 2 presents a similar set of voicing and fundamental frequency prediction errors but these are produced using the phoneme-specific (HMM-GMM) system.

Error	Noise	NNC	SS	Adapt	Match
E_{vc}	Clean	5.50	5.50	5.50	5.50
	20dB	6.06	6.83	5.28	5.33
	10dB	10.88	8.02	7.14	6.43
	0dB	41.45	14.92	16.17	11.10
E_p	Clean	5.26	5.26	5.26	5.26
	20dB	9.49	9.01	6.91	5.80
	10dB	13.95	12.93	9.17	7.71
	0dB	22.13	19.04	14.46	11.34

Table 1 - Voicing and fundamental frequency prediction errors on clean and noisy speech for no noise compensation (NNC), spectral subtraction (SS), model adaptation and matched training/testing using global speech modeling.

Error	Noise	NNC	SS	Adapt	Match
E_{vc}	Clean	5.95	5.95	5.95	5.95
	20dB	6.28	5.55	5.58	5.43
	10dB	12.39	6.92	7.25	5.87
	0dB	34.41	31.23	8.35	7.93
E_p	Clean	5.58	5.58	5.58	5.58
	20dB	9.78	8.36	6.68	6.23
	10dB	13.57	13.18	9.99	8.06
	0dB	16.32	16.90	14.23	11.81

Table 2 - Voicing and fundamental frequency prediction errors on clean and noisy speech for no noise compensation (NNC), spectral subtraction (SS), model adaptation and matched training/testing using phoneme-specific speech modeling.

The results show that as SNR reduces, the accuracy of voicing and fundamental frequency prediction deteriorates for both the global and phoneme-specific systems as indicated by the NNC columns in tables 1 and 2. Spectral subtraction gives significant reductions in prediction errors, in particular for voicing classification. In general, model adaptation further reduces prediction errors over those achieved by spectral subtraction. In fact, at higher SNRs the errors rates of model adaptation approach those of matched training/testing which can be considered the optimal adaptation.

For the phoneme-specific system, the noisy MFCC vectors contribute to prediction errors in two ways. Directly, by their distorted values affecting the MAP prediction, and indirectly, by reducing Viterbi decoding accuracy which corrupts the model and state sequence used in phoneme-specific prediction. The second of these effects can be investigated by comparing prediction errors when the sequence of phonemes is generated using the unconstrained phoneme decoding to when it is forced to the correct phoneme sequence. Table 3 shows voicing and fundamental frequency prediction errors on clean and noisy speech, using unconstrained decoding and forced alignment. The last rows of the table indicate the phoneme accuracy.

Test	Grammar	Clean	20dB	10dB	0dB
E_{vc}	Unconstrained	5.95	6.28	12.39	34.41
	Forced	6.03	7.59	10.12	13.13
E_p	Unconstrained	5.58	9.78	13.57	16.32
	Forced	5.62	8.98	13.61	19.08
%Acc	Unconstrained	73.7	33.9	15.3	11.3
	Forced	100.0	100.0	100.0	100.0

Table 3 – Voicing and fundamental frequency prediction errors with no noise compensation for clean and noisy speech using unconstrained and forced phoneme grammars.

The results show that even when the correct phoneme sequence is given (by the forced grammar), voicing and fundamental frequency errors increase considerably as SNR reduces. Moving to unconstrained phoneme decoding gives a large increase in voicing classification errors, particularly at 0dB. For fundamental frequency prediction, the effect of unconstrained decoding is much less, and at 0dB gives lower errors than forced decoding.

4.2 Speech classification and formant prediction

This section examines the effectiveness of the noise compensation methods on speech/non-speech prediction and formant frequency prediction using both the global and phoneme-specific systems. Classification of vectors as speech or non-speech is measured by the percentage speech activity classification error, E_{sc},

$$E_{sc} = \frac{N_{s|ns} + N_{ns|s}}{N_T} \times 100\% \qquad (19)$$

$N_{s|ns}$ is the number of non-speech vectors that are incorrectly classified as speech, $N_{ns|s}$ is the number of speech vectors that are incorrectly classified as non-speech. Formant frequency prediction errors are averaged across all four formants to give the percentage formant frequency error, E_f,

$$E_f = \frac{1}{4 \times N_V} \sum_{t=1}^{N_S} \sum_{q=1}^{4} \frac{\left|\hat{F}(q)_t - F(q)_t\right|}{F(q)_t} \times 100\% \qquad (20)$$

where $\hat{F}(q)_t$ and $F(q)_t$ are the predicted and reference frequency of the q^{th} formant for the t^{th} frame. Similar to E_p, formant frequency errors are measured for all N_s reference frames labelled as speech to ensure classification errors do not influence E_f.

Error	Noise	NNC	SS	Adapt	Match
E_{sc}	Clean	3.58	3.58	3.58	3.58
	20dB	18.16	18.16	17.90	11.80
	10dB	18.84	18.11	23.21	16.43
	0dB	18.10	18.13	18.31	22.51
E_f	Clean	10.00	10.00	10.00	10.00
	20dB	21.74	20.27	18.41	14.24
	10dB	25.07	24.22	20.68	16.47
	0dB	26.11	31.78	25.52	20.74

Table 4 - Speech/non-speech and formant frequency prediction errors on clean and noisy speech for no noise compensation (NNC), spectral subtraction (SS), model adaptation and matched training/testing using global speech modeling.

Error	Noise	NNC	SS	Adapt	Match
E_{sc}	Clean	1.90	1.90	1.90	1.90
	20dB	10.67	14.49	3.93	3.33
	10dB	17.31	14.48	3.58	3.47
	0dB	17.29	17.30	4.54	4.55
E_f	Clean	10.30	10.30	10.30	10.30
	20dB	23.28	20.28	15.92	13.60
	10dB	30.22	25.95	16.81	15.89
	0dB	42.35	41.91	26.29	20.45

Table 5 - Speech/non-speech and formant frequency prediction errors on clean and noisy speech for no noise compensation (NNC), spectral subtraction (SS), model adaptation and matched training/testing using phoneme-specific speech modeling.

Table 4 shows speech/non-speech classification error, E_{vc}, and formant frequency error, E_p, obtained using the global (GMM) system. As in tables 1 and 2, errors are shown in clean and noisy speech at SNRs from 20dB to 0dB in white noise. Results are again shown for no noise compensation, spectral subtraction, model adaptation and matched training/testing. Table 5 presents similar speech/non-speech and formant frequency prediction errors but uses the phoneme-specific (HMM-GMM) system.

For speech/non-speech classification, phoneme-specific prediction consistently outperforms global prediction. The most significant differences are for model adaptation and matched conditions at low SNRs, where the phoneme-specific system hardly deteriorates from the no noise performance. This is attributed to both model adaptation and matched conditions maintaining higher phoneme accuracies as SNR falls. For example, in clean speech, phoneme accuracy is 74%. With no noise compensation this falls to 14% at 10dB, but with matched conditions is increased to 50%.

For formant frequency prediction, the phoneme-specific system is generally more accurate than global prediction, particularly at lower SNRs. The results also show model adaptation to be more effective at noise compensation than spectral subtraction, and approaching matched condition performance. This is consistent with prediction of the acoustic features in the previous section.

5. CONCLUSION

This work has shown that noise compensation can be successfully applied to both phoneme-specific and global MAP prediction of acoustic features from MFCC vectors. Adapting the clean speech models to model noisy speech performs better than removing the noise using spectral subtraction. It is interesting to observe that the model adaptation method of noise compensation, which has been shown to be more effective than filtering, cannot be implemented in traditional fundamental frequency and formant frequency estimation methods [7,8]. However, the statistical modeling approach used here is able to benefit from adaptation.

6. REFERENCES

[1] European Telecommunications Standards Institute – ES 201 108 STQ, Front-end feature extraction algorithm, 2000

[2] European Telecommunications Standards Institute – ES 202 212 STQ – Extended advanced front-end, back-end reconstruction, 2003

[3] J. Darch and B.P. Milner, "MAP prediction of formant frequencies and voicing from MFCC vectors in noise", Speech Communication, vol. 48, no. 11, pp. 1556-1572, Nov. 2006

[4] B.P. Milner and X. Shao, "Prediction of fundamental frequency and voicing from MFCCs for unconstrained speech reconstruction", IEEE Trans. ASLP, no. 1, pp. 24-33, Jan.2007

[5] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise", Proc. ICASSP, 1979

[6] M.J.F. Gales and S.J. Young, "Cepstral parameter compensation for HMM recognition in noise", Speech Communication, vol. 12, 1993

[7] Q. Yan, S. Vaseghi, E. Zavarehei, and B. Milner, "Formant tracking linear prediction models for speech processing in noisy environments", Proc. Interspeech, 2005

[8] Yin A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music", JASA, vol. 111, no. 4, pp. 1917-1930, April 2002