INTERSPEECH 2009 BRIGHTON

Reconstructing Clean Speech from Noisy MFCC Vectors

Ben Milner, Jonathan Darch and Ibrahim Almajai

School of Computing Sciences, University of East Anglia, Norwich, UK

b.milner@uea.ac.uk, jonathan_darch@hotmail.com, i.almajai@uea.ac.uk

Abstract

The aim of this work is to reconstruct clean speech solely from a stream of noise-contaminated MFCC vectors, as may be encountered in distributed speech recognition systems. Speech reconstruction is performed using the ETSI Aurora back-end speech reconstruction standard which requires MFCC vectors, fundamental frequency and voicing information. In this work, fundamental frequency and voicing are obtained using maximum a posteriori prediction from input MFCC vectors, thereby allowing speech reconstruction solely from a stream of MFCC vectors. Two different methods to improve prediction accuracy in noisy conditions are then developed. Experimental results first establish that improved fundamental frequency and voicing prediction is obtained when noise compensation is applied. A series of human listening tests are then used to analyse the reconstructed speech quality, which determine the effectiveness of noise compensation in terms of mean opinion scores.

Index Terms: DSR, fundamental frequency, MAP, noise compensation

1. Introduction

Distributed speech recognition (DSR) operates by transmitting MFCC vectors from a terminal device to a remote back-end for decoding. However, within this architecture no facility is available for obtaining a time-domain speech waveform at the back-end. In several applications this may be useful, such as manual checking of errors during speech recognition, or to avoid deploying both a speech codec and a DSR front-end on a terminal device. A later version of the ETSI Aurora standard facilitated back-end speech reconstruction by transmitting the fundamental frequency and voicing of each frame at the expense of an additional 800bps [1]. The standard employs a sinusoidal model for reconstruction, which requires spectral envelope and harmonic information. The spectral envelope is obtained by inverting the MFCC vector into a coarse power spectrum [1], while harmonics are provided by the fundamental frequency.

An alternative method for speech reconstruction has been proposed whereby fundamental frequency and voicing are predicted from the received MFCC vectors [2]. This removes the need to transmit voicing and fundamental frequency and enables speech to be reconstructed solely from MFCC vectors. Motivation for this arose from the high level of correlation that exists between fundamental frequency and MFCC vectors [3]. For prediction, this correlation is exploited by creating a model of the joint density of fundamental frequency and MFCC vectors. This allows a maximum a posteriori (MAP) prediction of the fundamental frequency of a frame of speech to be made from its MFCC vector representation.

This work extends previous work by first developing methods to compensate fundamental frequency and voicing prediction, within a phoneme-specific framework rather than a globally framework, when the input MFCC vectors are corrupted by noise. Secondly, speech is then reconstructed from the compensated fundamental frequency and voicing and an analysis made into the effectiveness of noise compensation at improving speech quality. Earlier studies showed that without noise compensation, prediction accuracy using a global model of speech reduces as signal-to-noise ratios (SNRs) fall [4]. This is attributed to the contaminating noise distorting the MFCC vectors, causing their statistics to shift away from the clean-trained models. This is similar to the effect of noise in speech recognition, where noisy MFCC vectors become mismatched to clean-trained hidden Markov models (HMMs) leading to poor accuracy. Many compensation methods have been proposed to improve speech recognition accuracy in such noisy conditions [5] and many of these can be applied to the problem of fundamental frequency prediction from MFCC vectors.

The remainder of this paper is organised as follows. An overview of fundamental frequency and voicing prediction from MFCC vectors is given in section 2. Section 3 proposes two methods of noise compensation for integration with prediction. Experimental results are presented in section 4 that first examine the effectiveness of the noise compensation methods at improving prediction accuracy. Secondly, mean opinion score (MOS) results from human listening tests are presented which examine the quality of speech reconstructed from noisy MFCC vectors.

2. Fundamental frequency prediction

Prediction of fundamental frequency and voicing from MFCC vectors is achieved by first creating a joint feature vector, $\boldsymbol{y}_i$,

$$\boldsymbol{y}_i = [\boldsymbol{x}_i, f_i]^T \tag{1}$$

$\boldsymbol{x}_i$ is the ith MFCC vector and f_i the corresponding fundamental frequency. For unvoiced speech and nonspeech (i.e. non-voiced), $f_i = 0$. Using a set of training data, a model of the joint density, $\Phi^{\boldsymbol{y}}$, is created. Given an input MFCC vector, $\boldsymbol{x}_i$, a MAP prediction of the fundamental frequency can be made, $\hat{f}_i$:

$$\hat{f}_i = \arg\max_{f_i} \{p\left(f_i|\boldsymbol{x}_i, \Phi^{\boldsymbol{y}}\right)\} \tag{2}$$

In practice, localising prediction to individual phonemes gives higher accuracy than a single model of all speech [2].

2.1. Phoneme-specific modelling

Training phoneme-specific models involves three stages. First, a set of phoneme HMMs is created. Second, the HMMs are used to create state and phoneme-specific pools of MFCC vectors. Finally, Gaussian mixture models (GMMs) are trained from each vector pool to create state and phoneme-specific models of the joint density of fundamental frequency and MFCC vectors.

Phoneme localisation in training and testing is achieved using a set of $W+1$ left-right HMMs, which model the W phonemes in the database and nonspeech. These comprise $S=3$ states with $H=8$ modes per state and are arranged in an unconstrained grammar to place no constraints on the speech.

The HMMs are used to force align the training data to reference phoneme sequences which provides, for each training data utterance $\boldsymbol{X}=[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N]$, comprising N MFCC vectors, state and phoneme alignments where MFCC vector $\boldsymbol{x}_i$ is allocated to state q_i of phoneme m_i. Each MFCC vector, $\boldsymbol{x}_i$, is then joined by its fundamental frequency value, f_i, to form a joint feature vector, $\boldsymbol{y}_i$, as in equation 1.

The state and phoneme allocation for each MFCC vector, and the voicing class, enable voiced and nonvoiced vector pools, $\Omega^{\mathrm{v}}_{s,w}$ and $\Omega^{\mathrm{nv}}_{s,w}$, to be created in each state s in each phoneme w

$$\begin{aligned}\Omega^{\mathrm{v}}_{s,w} &= \{\boldsymbol{y}_i \in Z \;:\; f_i \neq 0,\; q_i = s, m_i = w\}\\ \Omega^{\mathrm{nv}}_{s,w} &= \{\boldsymbol{y}_i \in Z \;:\; f_i = 0,\; q_i = s, m_i = w\}\\ &\qquad 1 \leq s \leq S, \quad 0 \leq w \leq W\end{aligned} \tag{3}$$

Z represents the complete set of training data vectors.

The joint density of fundamental frequency and MFCC vectors can be modelled by applying expectation-maximisation clustering to each of the vector pools. This results in voiced and nonvoiced GMMs, $\Phi^{\mathrm{v}}_{s,w}$ and $\Phi^{\mathrm{nv}}_{s,w}$, associated with each state s of each phoneme model w. The voiced GMMs take the form

$$\Phi^{\mathrm{v}}_{s,w}(\boldsymbol{y}) = \sum_{k=1}^{K} \alpha^{\mathrm{v}}_{k,s,w}\, \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}^{\mathrm{v},\boldsymbol{y}}_{k,s,w}, \boldsymbol{\Sigma}^{\mathrm{v},\boldsymbol{yy}}_{k,s,w}) \tag{4}$$

where the kth Gaussian distribution in the voiced GMM for state s and model w has prior probability $\alpha^{\mathrm{v}}_{k,s,w}$, and mean $\boldsymbol{\mu}^{v,\boldsymbol{y}}_{k,s,w}$, and covariance $\boldsymbol{\Sigma}^{v,\boldsymbol{yy}}_{k,s,w}$, where:

$$\boldsymbol{\mu}^{\mathrm{v},\boldsymbol{y}}_{k,s,w} = \begin{bmatrix} \boldsymbol{\mu}^{\mathrm{v},\boldsymbol{x}}_{k,s,w} \\ \boldsymbol{\mu}^{\mathrm{v},f}_{k,s,w} \end{bmatrix}, \quad \boldsymbol{\Sigma}^{\mathrm{v},\boldsymbol{yy}}_{k,s,w} = \begin{bmatrix} \boldsymbol{\Sigma}^{\mathrm{v},\boldsymbol{xx}}_{k,s,w} & \boldsymbol{\Sigma}^{\mathrm{v},\boldsymbol{x}f}_{k,s,w} \\ \boldsymbol{\Sigma}^{\mathrm{v},f\boldsymbol{x}}_{k,s,w} & \boldsymbol{\Sigma}^{\mathrm{v},ff}_{k,s,w} \end{bmatrix} \tag{5}$$

Similar to equation 4, nonvoiced GMMs, $\Phi^{\mathrm{nv}}_{s,w}$, are created.

2.2. Prediction of fundamental frequency and voicing

For prediction, state and phoneme allocations, q_i and m_i, for each each input MFCC vector, $\boldsymbol{x}_i$, are determined by the network of HMMs. The voicing of the ith frame is computed as

$$\widehat{voicing}(i) = \begin{cases} \text{voiced} & \Phi^{\mathrm{v},\boldsymbol{x}}_{q_i,m_i}(\boldsymbol{x}_i) \geq \Phi^{\mathrm{nv},\boldsymbol{x}}_{q_i,m_i}(\boldsymbol{x}_i) \\ \text{nonvoiced} & \Phi^{\mathrm{nv},\boldsymbol{x}}_{q_i,m_i}(\boldsymbol{x}_i) > \Phi^{\mathrm{v},\boldsymbol{x}}_{q_i,m_i}(\boldsymbol{x}_i) \end{cases} \tag{6}$$

$\Phi^{\mathrm{v},\boldsymbol{x}}_{q_i,m_i}$ and $\Phi^{\mathrm{nv},\boldsymbol{x}}_{q_i,m_i}$ represent voiced and nonvoiced GMMs that have been marginalised to the MFCC vector component, $\boldsymbol{x}$.

For MFCC vectors classified as voiced, a weighted MAP prediction of fundamental frequency, $\hat{f}_i$, is made using each of the K clusters, $\phi^{\mathrm{v}}_{k,q_i,m_i}$, in the voiced GMM, $\Phi^{\mathrm{v}}_{q_i,m_i}$:

$$\hat{f}_i = \sum_{k=1}^{K} h_{k,q_i,m_i}(\boldsymbol{x}_i) \arg\max_{f_i} \left\{ p\left(f_i | \boldsymbol{x}_i, \phi^{v}_{k,q_i,m_i}\right) \right\} \tag{7}$$

$h_{k,q_i,m_i}(\boldsymbol{x}_i)$ is the posterior probability of $\boldsymbol{x}_i$ belonging to the kth cluster of the GMM in state q_i and phoneme m_i:

$$h_{k,q_i,m_i}(\boldsymbol{x}_i) = \frac{\alpha_{k,q_i,m_i}\; p\left(\boldsymbol{x}_i | \phi^{\boldsymbol{x}}_{k,q_i,m_i}\right)}{\sum_{k=1}^{K} \alpha_{k,q_i,m_i}\; p\left(\boldsymbol{x}_i | \phi^{\boldsymbol{x}}_{k,q_i,m_i}\right)} \tag{8}$$

$p\left(\boldsymbol{x}_i | \phi^{\boldsymbol{x}}_{k,q_i,m_i}\right)$ is the marginal distribution of the MFCC vector for the kth cluster of the GMM in state q_i and model m_i.

3. Noise compensation

Fundamental frequency and voicing prediction work well in clean conditions but become less accurate in noise, which introduces a mismatch between the noisy MFCCs and clean models. This is similar to the problem of speech recognition in noise, where many solutions have been proposed to reduce the mismatch and can be categorised as feature-domain or model-domain methods [5]. Many of these methods can be applied to prediction. Due to their success in robust speech recognition, the feature-domain method of spectral subtraction and model-domain method of model adaptation have been selected as two alternative compensation methods to integrate with prediction.

3.1. Spectral subtraction

Spectral subtraction is applied to the MFCC vectors by inverting them to the filterbank domain where speech and noise are additive. MFCC vectors are zero-padded to the dimensionality, J, of the filterbank and an inverse discrete cosine transform (DCT) applied, followed by an exponential operation to obtain linear filterbank features, $\boldsymbol{x}^{\mathrm{fb}}_i$:

$$\boldsymbol{x}^{\mathrm{fb}}_i = \exp\left(\boldsymbol{C}^{-1}\boldsymbol{x}_i\right) \tag{9}$$

$\boldsymbol{C}$ is the DCT matrix, where each element c_{ij} is given as:

$$c_{ij} = \cos\left\{\frac{i\pi(j+0.5)}{J}\right\} \quad 0 \leq i,j \leq J-1 \tag{10}$$

Of the many variants of spectral subtraction, this work uses linear spectral subtraction with an over-subtraction factor, α, and a maximum attenuation threshold, β. The clean speech filterbank estimate, $\hat{s}^{\mathrm{fb}}_i(j)$, for the jth channel of the ith frame is given as

$$\hat{s}^{\mathrm{fb}}_i(j) = \begin{cases} x^{\mathrm{fb}}_i(j) - \alpha\hat{d}^{\mathrm{fb}}_i(j) & x^{\mathrm{fb}}_i(j) - \alpha\hat{d}^{\mathrm{fb}}_i(j) > \beta\, x^{\mathrm{fb}}_i(j) \\ \beta\, x^{\mathrm{fb}}_i(j) & \text{otherwise} \end{cases} \tag{11}$$

$\hat{d}^{\mathrm{fb}}(j)$ is the noise estimate and is computed from the first few nonspeech MFCC vectors in the utterance. Following spectral subtraction, the clean speech filterbank estimate, $\hat{\boldsymbol{s}}^{\mathrm{fb}}_i$, is transformed to the MFCC domain using log and DCT operations.

3.2. Model adaptation

Model adaptation transforms the statistics of the clean-trained GMMs to model noise-contaminated MFCC vectors. From equation 5, the GMMs comprise a joint mean vector and joint covariance matrix, which must be adapted to model noisy speech. For clarity, state and phoneme indices, s and w, and the voicing superscript are dropped from the notation in this section. In the mean vector, the MFCC component, $\boldsymbol{\mu}^{\boldsymbol{x}}_k$, is affected by noise and must be adapted. The fundamental frequency component, $\boldsymbol{\mu}^{f}_k$, is independent of noise and can be left unadapted. In the joint covariance matrix, the variance of fundamental frequency, $\boldsymbol{\Sigma}^{ff}_k$, is independent of noise and requires no adaptation. The covariance of the MFCCs, $\boldsymbol{\Sigma}^{\boldsymbol{xx}}_k$, is affected by noise and must be adapted. The cross-covariances of MFCCs and fundamental frequency, $\boldsymbol{\Sigma}^{\boldsymbol{x}f}_k$ and $\boldsymbol{\Sigma}^{f\boldsymbol{x}}_k$, are equal due to the symmetry property of the covariance matrix and are computed:

$$\boldsymbol{\Sigma}^{\boldsymbol{x}f}_k = E[(\boldsymbol{x} - \boldsymbol{\mu}^{\boldsymbol{x}}_k)(f - \mu^{f}_k)] \tag{12}$$

With the addition of noise, the cross-covariance can be written:

$$\begin{aligned}\boldsymbol{\Sigma}^{(\boldsymbol{x}+\boldsymbol{d})f}_k &= E[(\boldsymbol{x} + \boldsymbol{d} - \boldsymbol{\mu}^{\boldsymbol{x}}_k - \boldsymbol{\mu}^{\boldsymbol{d}}_k)(f - \mu^{f}_k)]\\ &= E[(\boldsymbol{x} - \boldsymbol{\mu}^{\boldsymbol{x}}_k)(f - \mu^{f}_k)] + E[(\boldsymbol{d} - \boldsymbol{\mu}^{\boldsymbol{d}}_k)(f - \mu^{f}_k)]\\ &= \boldsymbol{\Sigma}^{\boldsymbol{x}f}_k + \boldsymbol{\Sigma}^{\boldsymbol{d}f}_k\end{aligned} \tag{13}$$

The term $\boldsymbol{\Sigma}_k^{\boldsymbol{df}}$ is the cross-covariance of noise and fundamental frequency which is zero, as noise and fundamental frequency are independent. Therefore both $\boldsymbol{\Sigma}_k^{\boldsymbol{xf}}$ and $\boldsymbol{\Sigma}_k^{\boldsymbol{fx}}$ are independent of noise and require no adaptation. So the components in equation 5 that require adaptation to noise are the MFCC means and covariances, $\boldsymbol{\mu}_k^{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_k^{\boldsymbol{xx}}$.

Adaptation takes place in the filterbank domain. The MFCC means and covariances are zero-padded to the dimensionality of the filterbank and converted to log filterbank domain means and covariances, $\boldsymbol{\mu}_k^{\boldsymbol{x},\mathrm{lfb}}$ and $\boldsymbol{\Sigma}_k^{\boldsymbol{xx},\mathrm{lfb}}$, by inverse DCTs:

$$\boldsymbol{\mu}_k^{\boldsymbol{x},\mathrm{lfb}} = \boldsymbol{C}^{-1}\boldsymbol{\mu}_k^{\boldsymbol{x}}, \qquad \boldsymbol{\Sigma}_k^{\boldsymbol{xx},\mathrm{lfb}} = \boldsymbol{C}^{-1}\boldsymbol{\Sigma}_k^{\boldsymbol{xx}}\left(\boldsymbol{C}^{-1}\right)^T \quad (14)$$

The MFCCs are assumed to be Gaussian which also holds in the log filterbank domain. In the linear filterbank domain the vectors will be log normally distributed. The log filterbank means and covariances can be transformed into the linear filerbank domain, $\boldsymbol{\mu}_k^{\boldsymbol{x},\mathrm{fb}}$ and $\boldsymbol{\Sigma}_k^{\boldsymbol{xx},\mathrm{fb}}$, as [6]:

$$\boldsymbol{\mu}_k^{\boldsymbol{x},\mathrm{fb}}(i) = \exp\left\{\boldsymbol{\mu}_k^{\boldsymbol{x},\mathrm{lfb}}(i) + \frac{\mathrm{diag}(\boldsymbol{\Sigma}_k^{\boldsymbol{xx},\mathrm{lfb}}(i,i))}{2}\right\} \quad (15)$$

$$\boldsymbol{\Sigma}_k^{\boldsymbol{xx},\mathrm{fb}}(i,j) = \boldsymbol{\mu}_k^{\boldsymbol{x},\mathrm{fb}}(i)\boldsymbol{\mu}_k^{\boldsymbol{x},\mathrm{fb}}(j)\exp\left\{\boldsymbol{\Sigma}_k^{\boldsymbol{xx},\mathrm{lfb}}(i,j) - 1\right\} \quad (16)$$

Noisy filterbank means and covariances, $\boldsymbol{\mu}_k^{\boldsymbol{z},\mathrm{fb}}$ and $\boldsymbol{\Sigma}_k^{\boldsymbol{zz},\mathrm{fb}}$, are computed by adding noise means and covariances, $\boldsymbol{\mu}^{\boldsymbol{d},\mathrm{fb}}$ and $\boldsymbol{\Sigma}^{\boldsymbol{dd},\mathrm{fb}}$, to the clean speech statistics:

$$\boldsymbol{\mu}_k^{\boldsymbol{z},\mathrm{fb}} = \boldsymbol{\mu}_k^{\boldsymbol{x},\mathrm{fb}} + \boldsymbol{\mu}^{\boldsymbol{d},\mathrm{fb}}, \quad \boldsymbol{\Sigma}_k^{\boldsymbol{zz},\mathrm{fb}} = \boldsymbol{\Sigma}_k^{\boldsymbol{xx},\mathrm{fb}} + \boldsymbol{\Sigma}^{\boldsymbol{dd},\mathrm{fb}} \quad (17)$$

The noisy filterbank means and covariances are transformed to the log filterbank domain using the inverse of equations 15 and 16 and then to noisy MFCC domain means and covariances, $\boldsymbol{\mu}_k^{\boldsymbol{z}}$ and $\boldsymbol{\Sigma}_k^{\boldsymbol{zz}}$, using 1-D and 2-D DCTs. These replace the clean speech means and covariances in equation 5 to give noise-adapted GMMs. The voiced and nonvoiced GMMs in each state of each phoneme model are adapted to noise with the input MFCC vectors left uncompensated. The noise statistics were obtained offline from a 10 second duration noise segment.

4. Experimental results

The experiments first examine the effectiveness of noise compensation on fundamental frequency and voicing prediction from MFCC vectors. Secondly, the quality of speech reconstructed from noisy MFCC vectors is examined. The experiments are conducted on a speaker-dependent database recorded from a female US English speaker. This has 579 sentences for training and 246 sentences for testing (approximately 130,000 test vectors). Reference fundamental frequency and voicing were obtained from laryngograph recordings and MFCC vectors extracted in accordance with the ETSI Aurora standard [1].

4.1. Prediction of voicing and fundamental frequency

The accuracy of voicing prediction is measured using the percentage voicing classification error, E^{vc}, defined as

$$E^{\mathrm{vc}} = \frac{N_{\mathrm{v|nv}} + N_{\mathrm{nv|v}}}{N_T} \times 100\% \quad (18)$$

$N_{\mathrm{v|nv}}$ is the number of unvoiced or nonspeech (nonvoiced) vectors that are incorrectly classified as voiced, $N_{\mathrm{nv|v}}$ is the number of voiced vectors that are incorrectly classified as nonvoiced, and N_T is the total number of vectors in the test set.

Table 1: Voicing and fundamental frequency prediction errors on noisy speech with noise compensation.

Error	SNR	Clean	20dB	10dB	0dB
E^{vc}	NNC	6.37	8.05	31.57	33.42
	SS	6.37	6.08	8.17	29.29
	Adapt	6.37	7.38	6.06	11.89
	Match	6.37	5.76	6.14	9.80
	ETSI	10.34	6.93	7.84	18.41
E^{f0}	NNC	5.39	7.85	10.06	13.37
	SS	5.39	7.52	10.47	13.06
	Adapt	5.39	6.31	8.01	13.38
	Match	5.39	5.92	7.09	11.24
	ETSI	2.50	2.37	2.63	9.62

Fundamental frequency prediction is measured using the percentage fundamental frequency error, E^{f0}, defined as

$$E^{f0} = \frac{1}{N_{\mathrm{v}}}\sum_{i=1}^{N_{\mathrm{v}}}\frac{|\hat{f}_i - f_i|}{f_i} \times 100\% \quad (19)$$

$\hat{f}_i$ and f_i are the predicted and reference fundamental frequencies of the ith frame. E^{f0} is measured for only those N_{v} frames labelled as voiced according to the reference voicing. This ensures voicing classification errors do not influence the measurement of E^{f0} which is likely in noisy speech.

Table 1 shows voicing and fundamental frequency errors for clean speech and speech contaminated with car noise at SNRs of 20dB, 10dB and 0dB. For each noise condition, no noise compensation (NNC), spectral subtraction (SS), model adaptation (Adapt) and matched training/testing conditions (Match) performance is shown. With no noise compensation, voicing errors increase substantially below 20dB, due to the mismatch between the noisy MFCC vectors and clean models. Spectral subtraction retains good performance down to 10dB before breaking down. Model adaptation is significantly more effective at lower SNRs, achieving performance close to matched training/testing which indicates the effectiveness of adaptation. For fundamental frequency prediction, spectral subtraction is less effective, but model adaptation is able to improve performance at 20dB and 10dB. At 0dB, only matched condition testing reduces errors significantly.

As a comparison, the rows labelled ETSI show voicing and fundamental frequency estimation using the ETSI front-end [1]. In an analysis involving a range of fundamental frequency and voicing estimation methods that have access to the time-domain waveform this was found to be the most robust [7]. Comparing E^{f0} shows the ETSI method to perform significantly better than prediction and this is attributed to the ETSI method having access to the time-domain waveform, rather than just the MFCC vectors as with prediction. For E^{vc}, prediction with model adaptation is more accurate than the ETSI method, which is not unexpected as several low-order MFCCs are well suited to voicing classification, particularly after noise compensation.

Noise contributes to prediction errors in two ways: directly – by distorting the MFCC vectors which affects MAP prediction in equation 7, and indirectly – by reducing Viterbi decoding accuracy, leading to prediction from incorrect state and phoneme-specific GMMs. These effects can be examined by comparing prediction errors when the MFCC vectors are decoded using either the unconstrained phoneme grammar or by being forced to the correct phoneme sequence. Table 2 shows prediction errors

Table 2: Voicing and fundamental frequency prediction errors using unconstrained and forced phoneme grammars.

Test	Grammar	Clean	20dB	10dB	0dB
Acc	Unconst.	73.73	46.36	18.98	6.46
	Forced	100.00	100.00	100.00	100.00
E^{vc}	Unconst.	6.37	8.05	31.57	33.42
	Forced	7.04	8.83	17.64	23.09
E^{f0}	Unconst.	5.39	7.85	10.06	13.37
	Forced	5.39	7.94	10.92	16.68

with no noise compensation on clean and noisy speech using unconstrained decoding and forced alignment. Phoneme decoding accuracies are shown in the top two rows.

In clean speech, fundamental frequency and voicing prediction is similar whether using unconstrained or forced decoding. As SNRs reduce, the less accurate alignment provided by unconstrained decoding leads to significantly lower voicing classification accuracy in comparison to forced alignment. This is caused by many voiced vectors being incorrectly aligned to phonemes that are primarily unvoiced or nonspeech. These have a high prior probability of being nonvoiced (equation 6), which leads to incorrect voicing classification. For fundamental frequency, incorrect phoneme alignment has less effect where from clean speech down to 10dB, prediction accuracy is almost identical between the unconstrained and forced decoding.

4.2. Reconstructed speech quality

This section examines the quality of reconstructed speech using a set of human listening tests to compute mean opinion scores (MOS). To reconstruct speech, the ETSI Aurora backend speech reconstruction standard is used [1], which is based on a sinusoidal model of speech. Sinusoid frequencies are obtained as harmonics of the fundamental frequency and sinusoid amplitudes from a spectral envelope obtained by inverting the MFCC vectors to a magnitude spectrum.

Four signal variants are compared. First is the original, unprocessed speech. Second is speech reconstructed from MFCC vectors with fundamental frequency and voicing estimated from the time-domain signal – as specified in the ETSI Aurora standard. Third is speech reconstructed solely from MFCC vectors, with fundamental frequency and voicing predicted from MFCC vectors as described in section 2. The final method again reconstructs speech solely from MFCC vectors, but now applies the model adaptation noise compensation method described in section 3 to improve fundamental frequency and voicing prediction. Model adaptation was selected due to its superior performance over spectral subtraction. Within a sound-proof room, twenty listeners were each played 40 speech utterances in a random order which covered the four signal variants in both clean speech and at SNRs of 20dB, 10dB and 5dB. MOS results were computed and these are displayed in figure 1.

In clean conditions, ETSI reconstruction of speech has a MOS of 4.1 which is 0.8 points below the original speech. Reconstruction solely from MFCC vectors (using predicted fundamental frequency/voicing) has a MOS of 3.6 which is 0.5 points below ETSI and is attributed to the prediction errors seen in table 1. As SNRs reduce, the MOS results show that reconstructing speech with noise compensation (PREDA) improves speech quality over the uncompensated reconstruction (PRED). Comparing the ETSI reconstructed speech with speech reconstructed solely from MFCC vectors with noise compensation reveals relatively small MOS differences – about 0.5 in clean speech and at 20dB, and about 0.2 at lower SNRs of 10dB and 5dB.

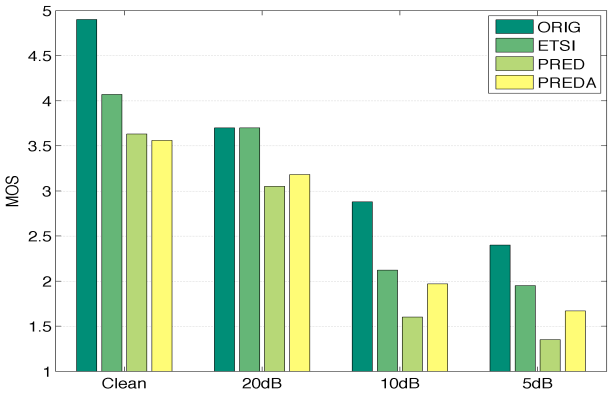

Figure 1: *MOS for: original speech (ORIG), reconstructed from MFCC vectors and fundamental frequency/voicing (ETSI), reconstructed using predicted fundamental frequency/voicing (PRED), and reconstructed with model adaptation (PREDA).*

5. Conclusion

This work has shown that noise compensation methods, developed for speech recognition, can be successfully applied to fundamental frequency and voicing prediction from MFCC vectors. Superior performance came from model adaptation which attained performance approaching that obtained with matched training/testing. Listening tests demonstrated that using these robustly predicted fundamental frequency/voicing values for speech reconstruction produced higher quality speech in noisy conditions than without. In fact the reconstructed speech quality achieved solely from MFCC vectors approached that produced by the ETSI method which has access to the time-domain signal for fundamental frequency/voicing estimation.

6. References

[1] ETSI, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm," ETSI STQ-Aurora DSR Working Group, ES 202 211 version 1.1.1, Nov. 2003.

[2] B. Milner and X. Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 24–33, Jan. 2007.

[3] J. Darch, B. Milner, and S. Vaseghi, "Analysis and prediction of acoustic speech features from mel-frequency cepstral coefficients in distributed speech recognition architectures," *JASA*, vol. 124, no. 6, pp. 3989–4000, Dec. 2008.

[4] B. Milner, J. Darch, and S. Vaseghi, "Applying noise compensation methods to robustly predict acoustic speech features from MFCC vectors in noise," in *ICASSP*, 2008, pp. 3945–3948.

[5] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions," *Computer Speech and Language*, vol. 23, no. 3, pp. 389–405, July 2009.

[6] M. Gales and S. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech and Language*, vol. 9, no. 4, pp. 289–307, Oct. 1995.

[7] J. Darch, "Robust acoustic speech feature prediction from mel frequency cepstral coefficients," Ph.D. dissertation, Norwich, UK, 2008.