APPLYING NOISE COMPENSATION METHODS TO ROBUSTLY PREDICT ACOUSTIC SPEECH FEATURES FROM MFCC VECTORS IN NOISE

Ben Milner[1], Jonathan Darch[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
{b.milner, jonathan.darch}@uea.ac.uk, saeed.vaseghi@brunel.ac.uk

ABSTRACT

This paper examines the effect of applying noise compensation to improve acoustic speech feature prediction from noise contaminated MFCC vectors, as may be encountered in distributed speech recognition (DSR). A brief review of maximum a posteriori prediction of acoustic speech features (voicing, fundamental and formant frequencies) from MFCC vectors is made. Two noise compensation methods are then applied; spectral subtraction and model adaptation. Spectral subtraction is used to filter noise from the received MFCC vectors, while model adaptation is applied to adapt the joint models of acoustic features and MFCCs to account for noise contamination. Experiments examine acoustic feature prediction accuracy in noise and results show that the two noise compensation methods significantly improve prediction accuracy in noise. The technique of model adaptation was found to be better than spectral subtraction and could restore performance close to that achieved in matched training and testing.

Index Terms—robustness, distributed speech recognition, acoustic feature prediction, noise adaptation, spectral subtraction

1. INTRODUCTION

In recent years there has been considerable interest in distributed speech recognition (DSR) for applications operating over mobile networks. In the first version of the ETSI Aurora DSR standard, MFCC feature extraction is performed on the terminal device and a stream of feature vectors is transmitted to the remote back-end for decoding at a bit rate of 4800bps [1]. A later version of the ETSI standard also transmitted voicing and fundamental frequency which increased the bit rate to 5600bps [2]. One motivation for transmitting this extra information is to enable audio speech reconstruction at the back-end. This is achieved using a sinusoidal model that uses an MFCC-derived spectral envelope together with fundamental frequency to reconstruct the audio.

In recent work we have shown that the voicing and fundamental frequency of a frame of speech can be predicted solely from the MFCC representation of that frame [3]. This is achieved by modelling the joint density of fundamental frequency and MFCC vectors which allows a maximum a posteriori (MAP) prediction of fundamental frequency to be made from an MFCC vector. This removes the need to transmit voicing and fundamental frequency and allows speech to be reconstructed solely from the MFCC vectors. We have also extended this work to predict other acoustic features such as formant frequencies and speech class from MFCCs vectors [4].

The aim of this work is to extend previous work by improving acoustic feature prediction from MFCC vectors contaminated by acoustic noise. Without noise compensation, prediction accuracy reduces as the signal-to-noise ratio (SNR) decreases. Two methods of noise compensation are investigated. The first is spectral subtraction which removes noise from the MFCC vectors. The second method adapts the models of the joint density of acoustic features and MFCC vectors to model noisy speech.

Section 2 gives a brief review of MAP prediction of acoustic speech features from MFCCs. The application of spectral subtraction and model adaptation to acoustic feature prediction from noisy MFCCs is presented in section 3. Section 4 shows experimental results on the effectiveness of noise compensation for speaker-dependent and speaker-independent databases.

2. ACOUSTIC FEATURE PREDICTION

This section describes how acoustic speech features are predicted from MFCC vectors. Further details of the procedure are given in [3][4]. Prediction of acoustic speech features (voicing, fundamental frequency and formant frequencies) from MFCC vectors is achieved by modeling the joint density of acoustic features and MFCC vectors. Then, given an MFCC vector, the joint density enables a MAP prediction of the associated acoustic features.

The procedure begins by defining a joint feature vector, $\mathbf{z}_t$, comprising an MFCC vector, $\mathbf{x}_t$ and an acoustic speech vector, $\mathbf{f}_t$,

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

where t indicates vector number. In this work the acoustic vector, $\mathbf{f}_t$, comprises the fundamental frequency, F0, and the first four formant frequencies, F1 to F4, i.e. $\mathbf{f} = \left[F0, F1, F2, F3, F4\right]$. For unvoiced speech and non-speech, F0 is set to zero and for non-speech, F1 to F4 are set to zero. The MFCC vector conforms to the ETSI Aurora standard and comprises static MFCCs 0 to 12 [1].

2.1 Training

Three voicing classes exist for the speech; voiced, unvoiced and non-speech. As such, training data vectors are pooled into three sets, Ψ^{v}, Ψ^{u} and Ψ^{ns}, according to their reference voicing. From each of these three sets of joint feature vectors, Gaussian mixture models (GMMs), Φ^{v}, Φ^{u} and Φ^{ns}, are trained. Considering the voiced GMM (unvoiced and non-speech GMMs follow similar training procedures – see [4]), expectation-maximisation (EM) clustering is used to create a GMM comprising K^{v} clusters,

$$p(\mathbf{z}_t) = \Phi^{v}(\mathbf{z}_t) = \sum_{k=1}^{K^{v}} \alpha_k^{v}\, \phi_k^{v}(\mathbf{z}_t) = \sum_{k=1}^{K^{v}} \alpha_k^{v}\, N\left(\mathbf{z}_t; \mu_k^{v}, \Sigma_k^{v}\right) \tag{2}$$

Each cluster comprises a prior probability, α_k^v, and a Gaussian probability density function, N, with mean vector, μ_k^v, and covariance matrix, Σ_k^v. The mean vector comprises two components, the mean of the voiced MFCC vectors, $\mu_k^{v,\mathbf{x}}$, and the mean of the acoustic feature vector, $\mu_k^{v,\mathbf{f}}$. Similarly, the covariance matrix comprises four components; the covariance matrix of the MFCC vectors, $\Sigma_k^{v,\mathbf{xx}}$, the covariance matrix of the acoustic features, $\Sigma_k^{v,\mathbf{ff}}$, and the covariances of the MFCCs and acoustic features, $\Sigma_k^{v,\mathbf{xf}}$ and $\Sigma_k^{v,\mathbf{fx}}$. The components of μ_k^v and Σ_k^v can be represented as,

$$\mu_k^v = \begin{bmatrix} \mu_k^{v,\mathbf{x}} \\ \mu_k^{v,\mathbf{f}} \end{bmatrix} \text{ and } \Sigma_k^v = \begin{bmatrix} \Sigma_k^{v,\mathbf{xx}} & \Sigma_k^{v,\mathbf{xf}} \\ \Sigma_k^{v,\mathbf{fx}} & \Sigma_k^{v,\mathbf{ff}} \end{bmatrix} \tag{3}$$

2.2 Prediction

The voiced, unvoiced and non-speech GMMs can now be used to predict the voicing, fundamental frequency and formant frequencies associated with an input MFCC vector. First a voicing decision is made by computing the probability of the MFCC vector from each of the three marginalised GMMs, $\Phi^{v,\mathbf{x}}$, $\Phi^{u,\mathbf{x}}$ and $\Phi^{ns,\mathbf{x}}$,

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

For MFCC vectors classified as voiced, fundamental and formant frequencies are computed using MAP prediction, while for unvoiced speech only formant frequencies are predicted. The predicted acoustic feature vector, $\hat{\mathbf{f}}_t(k)$, from cluster k of the voiced GMM, ϕ_k^v, is computed as,

$$\hat{\mathbf{f}}_t(k) = \arg\max_{\mathbf{f}_t} \left(p\left(\mathbf{f}_t \middle| \mathbf{x}_t, \phi_k^v\right) \right) \tag{5}$$

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

$$\hat{\mathbf{f}}_t = \sum_{k=1}^{K^v} h_k(\mathbf{x}_t) \left(\mu_k^{v,\mathbf{f}} + \Sigma_k^{v,\mathbf{fx}} \left(\Sigma_k^{v,\mathbf{xx}} \right)^{-1} \left(\mathbf{x}_t - \mu_k^{v,\mathbf{x}} \right) \right) \tag{6}$$

The posterior probability, $h_k(\mathbf{x}_t)$, of the MFCC vector belonging to the k^{th} cluster of the GMM is computed as,

$$h_k(\mathbf{x}_t) = \frac{\alpha_k^v \, p\left(\mathbf{x}_t \middle| \phi_k^{v,\mathbf{x}}\right)}{\sum_{k=1}^{K^v} \alpha_k^v \, p\left(\mathbf{x}_t \middle| \phi_k^{v,\mathbf{x}}\right)} \tag{7}$$

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

3. NOISE COMPENSATION

When noise is present in the speech signal the accuracy of acoustic feature prediction reduces. The effect of additive noise in the time-domain will alter the resulting MFCC vector which leads to a mismatch with the clean speech GMMs and inaccurate prediction of the acoustic features. To improve prediction accuracy in the presence of noise, two methods are considered to reduce this mismatch between the clean trained GMMs and noise contaminated input MFCC vectors. The first method removes noise from the input MFCC vectors. This is achieved by spectral subtraction where a noise estimate is subtracted from the noisy speech [5]. The second method adjusts the statistics of the GMMs to model noisy speech. Such adaptation methods have been successfully applied to speech recognisers for 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 a linear spectral domain where speech and noise are additive. This is achieved by first zero padding the MFCC vector to the dimensionality of the log filterbank vector and applying an inverse discrete cosine transform (DCT) 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,

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

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

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

It is not necessary 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, leads to more stability and reduces the likelihood of processing distortions 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, which gives superior performance over other implementations of spectral subtraction. 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 input into the acoustic feature prediction system of section 2.

3.2 Model adaptation

The second noise compensation method adapts the statistics of the 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. 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, $\mu_k^{\mathbf{x},lfb}$ and $\Sigma_k^{\mathbf{x},lfb}$,

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

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 exhibit a log normal distribution. The log filterbank means and covariances can be transformed into the linear filerbank domain, $\mu_k^{\mathbf{x},fb}$ and $\Sigma_k^{\mathbf{x},fb}$, [6],

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

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

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

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

The noisy filterbank means and covariances can be transformed back 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, $\mu_k^{\mathbf{y}}$ and $\Sigma_k^{\mathbf{y}}$,

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

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.

4. EXPERIMENTAL RESULTS

The aim of these experiments is to examine the effectiveness of the noise compensation methods when applied to acoustic feature prediction from MFCC vectors in noise. Two speech databases are used for evaluation to allow comparisons between speaker-dependent and speaker-independent prediction. The speaker-dependent database is taken from a single female US English speaker and comprises 589 sentences for training. A further 246 sentences, containing approximately 130,000 vectors, are used for testing. Reference fundamental frequency and voicing is obtained from a laryngograph while formant frequencies were obtained using LPC and Kalman filtering [7]. The speaker-independent database is the VTR (vocal tract resonances) database which is a subset of the TIMIT database [8]. Training uses 324 sentences taken from 173 male and female speakers. For testing a set of 192 sentences are used, comprising 57,823 vectors, which are recorded from a different set of 24 male and female speakers. Hand corrected formant frequencies are provided while the YIN algorithm was used to obtain reference fundamental frequency [9]. Both databases were downsampled to 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.

Before presenting experimental results the error measures used to measure prediction accuracy of the acoustic features must be defined. The accuracy of identifying voiced frames is measured using the percentage voicing classification error, E_{vc}, defined as,

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

$N_{v|nv}$ 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,

$$E_p = \frac{1}{N_V}\sum_{t=1}^{N_V}\frac{\left|\hat{F}0_t - F0_t\right|}{F0_t} \times 100\% \tag{18}$$

$\hat{F}0_t$ and $F0_t$ are the predicted and reference fundamental frequency of the t^{th} frame. E_p is measured for all frames labeled as voiced, N_v, according to the reference voicing. This ensures voicing classification errors do not influence E_p which is likely in noisy speech. Classification of frames as speech or non-speech is measured by the percentage speech activity classification error, E_{sc},

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

$N_{s|ns}$ 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. Finally, formant frequency prediction errors are averaged across all four formants to give the percentage formant frequency error, E_f,

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

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 reference frames labeled as speech, N_s, to ensure classification errors do not influence E_f.

4.1 Speaker-dependent acoustic feature prediction

This section examines noise compensation for speaker-dependent acoustic feature prediction where GMM training and testing uses the speaker-dependent database. Table 1 shows voicing classification error, E_{vc}, fundamental frequency error, E_p, speech classification error, E_{sc}, and formant frequency error, E_f, for clean speech and speech contaminated with white noise at SNRs of 20dB, 10dB and 0dB. Results for no noise compensation (NNC), spectral subtraction (SS) and model adaptation are shown and these all use 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 noise conditions. In practice matched condition training and testing is not feasible but it does provide an upper bound on performance.

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
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 1. Speaker-dependent acoustic feature prediction errors on clean and noisy speech for no noise compensation (NNC), spectral subtraction (SS), model adaptation and matched training/testing.

The results show that prediction of all acoustic features deteriorates as SNR reduces. With no noise compensation, voicing and fundamental frequency errors increase, although at a lower rate than speech classification and formant frequency errors. In particular, noise causes speech/non-speech classification to return nearly all frames as being speech, hence the convergence of E_{sc} to around 18% as SNR falls. Both noise compensation methods improve prediction accuracy with model adaptation generally outperforming spectral subtraction. Matched conditions shows the upper bound on performance and in many cases adaptation performs close to this level. Compensation against speech classification errors is most difficult to achieve, with even matched training/testing performing poorly.

4.2 Speaker-independent acoustic feature prediction

This section examines noise compensation on acoustic feature prediction for speaker-independent speech. Table 2 shows acoustic feature prediction errors in the same format as table 1. In clean speech, errors are higher for the speaker-independent system than for the speaker-dependent system. This is explained by the larger variances of the speaker-independent models arising from the larger variation in speech sounds. In noise, the speaker-independent system generally performs worse than the speaker-dependent system, with the exception of formant frequency prediction. Applying noise compensation increases acoustic feature prediction accuracy in noise at all SNRs, although the amount of improvement is less than with speaker-dependent prediction. This is attributed to the wider variances of the speaker-independent models already providing some noise robustness, making the application of explicit noise compensation less marked. This is confirmed by the smaller differences between NNC and matched training/testing on the speaker-independent system in comparison to the speaker-dependent system.

Error	Noise	NNC	SS	Adapt	Match
E_{vc}	Clean	11.72	11.72	11.72	11.72
	20dB	12.15	11.72	11.66	12.22
	10dB	13.77	12.76	11.81	12.71
	0dB	24.23	25.15	21.54	13.89
E_p	Clean	10.37	10.37	10.37	10.37
	20dB	15.35	14.09	13.82	12.96
	10dB	26.39	23.64	21.98	19.60
	0dB	33.72	33.41	29.16	25.52
E_{sc}	Clean	8.14	8.14	8.14	8.14
	20dB	13.37	13.31	13.04	12.10
	10dB	13.38	13.36	13.36	16.84
	0dB	13.38	13.38	13.42	20.01
E_f	Clean	11.71	11.71	11.71	11.71
	20dB	13.07	12.94	12.90	12.58
	10dB	14.82	14.27	14.09	13.53
	0dB	16.59	16.42	16.24	14.78

Table 2. Speaker-independent acoustic feature prediction errors on clean and noisy speech for no noise compensation (NNC), spectral subtraction (SS), model adaptation and matched training/testing.

5. CONCLUSION

This work has shown that noise compensation can be successfully applied to MAP prediction of acoustic features from MFCC vectors. Filtering the noise using spectral subtraction generally performs less effectively than adapting the speech models to model noisy speech. It is interesting to note that a model adaptation type of compensation cannot be implemented in most traditional methods of fundamental frequency and formant frequency estimation [7][8]. However, the statistical modeling approach used here can benefit from adaptation. Similar results in robust speech recognition have also been observed where adaptation techniques generally outperform filtering methods for noise robustness.

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] 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

[4] 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

[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] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing", Proc. ICASSP, 2006

[9] 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