Reconstruction-based speech enhancement from robust acoustic features

Philip Harding and Ben Milner

School of Computing Sciences, University of East Anglia, UK

Abstract

This paper proposes a method of speech enhancement where a clean speech signal is reconstructed from a sinusoidal model of speech production and a set of acoustic speech features. The acoustic features are estimated from noisy speech and comprise, for each frame, a voicing classification (voiced, unvoiced or non-speech), fundamental frequency (for voiced frames) and spectral envelope. Rather than using different algorithms to estimate each parameter, a single statistical model is developed. This comprises a set of acoustic models and has similarity to the acoustic modelling used in speech recognition. This allows noise and speaker adaptation to be applied to acoustic feature estimation to improve robustness. Objective and subjective tests compare reconstruction-based enhancement with other methods of enhancement and show the proposed method to be highly effective at removing noise.

Keywords: speech enhancement, noise reduction, adaptation, sinusoidal model

1. Introduction

This work proposes a reconstruction-based approach to speech enhancement that aims to produce a noise-free signal. This moves away from conventional enhancement methods that use filtering to remove noise. Instead, the enhanced signal is reconstructed from a model of speech production and a set of acoustic features that are estimated from the noisy speech.

Historically, most approaches to speech enhancement are filtering methods implemented as analysis-modification-synthesis systems. Many filtering approaches have been proposed and have been categorised into spectral subtraction, Wiener filtering, statistical and subspace methods [1]. Spectral subtraction is the most simple and requires just a noise spectral estimate but is prone to musical noise and speech distortion and leaving residual noise [2, 3]. Wiener filtering produces higher quality speech,

[1, Ch.11.2], although its implementation is more complex and requires an estimate of the SNR in each frequency bin. Iterative approaches, decision directed methods, nonnegative matrix factorisation and Gaussian mixture models (GMMs) have all been used to estimate the required *a priori* SNR [4, 5, 6]. The Wiener filter is the optimal, in the mean square error sense, linear estimator of the complex speech spectrum. Statistical methods extend this by identifying optimal non-linear minimum mean square error (MMSE) estimators of spectral magnitudes or log spectral magnitudes which give further improvement in speech quality [7, 8]. Extensions use probability density functions that better model the distribution of the speech spectra [9], and by taking into account speech presence uncertainty in estimating spectral amplitudes [10]. Subspace methods transform the noisy speech into speech and noise subspaces [11]. Truncation of the vector space aims to retain elements containing speech and remove noise components, although retaining too few speech components oversmooths the speech, while retaining too many components leaves residual noise. Evaluation across a range of speech enhancement methods shows these filtering methods to be effective in improving speech quality but susceptible to the accuracy of noise and SNR estimates which can introduce unwanted artefacts into the enhanced speech such as musical noise, residual noise and distortion [1, Ch.11.2]. Most filtering methods enhance the magnitude spectrum and combine this with the noisy phase, although some recent work has considered the importance of phase. Phase spectrum compensation is one example and adjusts the phase spectrum according to the noise magnitude spectrum and has been combined with MMSE spectral magnitude estimation [12, 13].

An alternative to filtering the noisy speech is to reconstruct or synthesise a clean speech signal. This is motivated by a desire to reduce artefacts introduced by the filtering process. Reconstruction approaches can be loosely divided into those that reconstruct the speech using a model of speech production and those that use a corpus or inventory of clean speech segments to synthesise an enhanced speech signal.

The main challenge for model-based approaches is to obtain a set of noise-free speech parameters that can be applied to a model of speech production. The sinusoidal model has been used for speech enhancement where an initial set of model parameters is extracted from the noisy speech and then refined iteratively by Wiener filtering and smoothing [14]. In [15], noisy speech is first pre-cleaned and then decomposed into

excitation and vocal tract components. A harmonic plus noise model (HNM) models clean excitation while formants are tracked using a combined Viterbi/Kalman filter, which are then combined using a linear predictive model. Chen et al [16] adopted a similar approach although uses a HNM to model the speech signal rather than the excitation. The noisy speech is again pre-cleaned and HNM analysis applied to extract fundamental frequency, spectral envelope and spectral gain with Kalman filtering tracking the parameters over time. The recent work in hidden Markov model (HMM)-based speech synthesis, [17], has also been applied to speech enhancement [18, 19]. These methods use a network of HMMs to decode noisy speech into a model and state sequence which is then input into an HMM-based synthesiser to output a clean speech signal.

Corpus-based approaches use a large database of clean speech and assume that noise varies more slowly than speech [20]. Given noisy speech, the speech corpus is searched for segments that when added to a stationary noise segment best resembles the noisy speech. Selected segments are then concatenated to form the enhanced signal. A related technique is inventory-style enhancement which utilises clean and noisy codebooks [21]. The two codebooks are created from speech data taken from a single speaker and are matched as the noisy codebook is formed from the speech data but with noise added. During enhancement a hidden Markov model (HMM) finds the optimal sequence of codebook entries from the noisy codebook which in turn identifies waveform units that are concatenated to form the enhanced speech signal. A similar method uses two GMMs, one trained on noisy MFCCs and the other on clean MFCCs [22]. During enhancement the noisy input speech is matched to mixture components from the noisy GMM and then mapped to the clean GMM which outputs a stream of MFCC vectors that are inverted to form the enhanced signal.

The work presented here uses the reconstruction approach to speech enhancement which, given a sufficiently good speech model and a set of noise-free acoustic features, should produce speech that is free from residual noise and artefacts. This gives rise to two main challenges: i) to find a sufficiently good model for speech reconstruction and ii) to develop robust methods to estimate accurately noise-free acoustic speech features. Our approach is based upon a variant of the sinusoidal model and differs from previous approaches as no pre-filtering of the speech is needed before parameter

estimation. Instead we propose an integrated statistical method that estimates the set of acoustic features needed for reconstruction within a single statistical framework. This relates to earlier work that reconstructed speech solely from a sequence of MFCC vectors within a distributed speech recognition (DSR) architecture using a sinusoidal model [23]. Spectral envelope parameters were obtained by inverting the MFCC vectors while a maximum a posteriori (MAP) estimate of fundamental frequency was made from the MFCC vector. This generated good quality speech although was designed for clean speech input, was speaker-dependent and constrained to operate on 23-D MFCC vectors. Later work [24], also within a DSR architecture, considered statistical methods for estimating each acoustic speech feature separately from noisy MFCC vectors by including some noise compensation. The proposed work now uses a single statistical model to estimate the set of acoustic features and improves noise robustness by considering the effect of phase within a mismatch function which is applied using an unscented transform to adapt the clean model statistics to noisy speech. Furthermore, speaker adaptation is also applied to adapt model parameters to the speaker under test which removes the speaker dependence constraint of earlier systems. This set of robust acoustic speech features is then input into a variant of the sinusoidal model to reconstruct a speech signal and forms the proposed method of enhancement.

Section 2 examines speech production models for their suitability in reconstruction-based speech enhancement. These are driven by excitation and vocal tract features that include voicing, fundamental frequency, spectral envelope and phase. Estimating these accurately from noisy speech is problematic and Section 3 develops methods that are robust to both noise and speaker variability. Experiments are presented in Section 4 that compare the effectiveness of reconstruction-based enhancement to other methods of speech enhancement using PESQ and three-way mean opinion score (MOS) listening tests that evaluate signal quality, background noise intrusiveness and overall quality [25].

2. Speech enhancement framework

The proposed speech enhancement method aims to reconstruct a noise-free speech signal from a model of speech production using a set of acoustic features extracted

from noisy speech. This section first examines candidate models of speech production and then considers how acoustic features can be estimated robustly from noisy speech.

2.1. Speech reconstruction model

Many engineering models of speech production have been proposed and form candidates for the model in this work. Source-filter models have found numerous applications in speech coding and typically use linear predictive coding (LPC) to model the vocal tract [26]. A variety of methods has been used to represent the source, which as a minimum require fundamental frequency and voicing, or for more detailed representations use techniques such as vector quantisation (VQ) codebooks as employed in algebraic code excited linear prediction (ACELP) [27]. The relatively small set of parameters lends itself well to low bit-rate speech coding. A variant of the source-filter model is STRAIGHT which was proposed originally for speech modification and now has widespread use in HMM-based speech synthesis [28, 29]. Alternative models of speech are the sinusoidal model and HNM [30, 31]. These model speech as a summation of sinusoids with frequencies equal to harmonic frequencies while a noise term is introduced by the HNM. These usually require more parameters than source-filter models and used in text-to-speech synthesis [31].

For speech enhancement there is no requirement to operate at low bit-rates. Instead the priority is to reconstruct high quality speech from a set of parameters that can be estimated robustly from noisy speech. In preliminary tests, a comparison of synthesised speech from a range of speech models found harmonic-type models to give highest quality speech [32]. Using the HNM a time-domain speech signal, $x(n)$, is synthesised as

$$x(n) = \sum_{l=1}^{L} a_l \, cos(2\pi f_l n + \vartheta_l) + d(n) \tag{1}$$

where L is the number of sinusoids and a_l, f_l and ϑ_l are the amplitude, frequency and phase of each sinusoid, while $d(n)$ is the noise term.

2.2. Application to speech enhancement

For speech enhancement, the parameters of the model must be estimated from noisy speech. This can be simplified by making various assumptions of the model to reduce the number of parameters to be estimated. First, for voiced speech, a

harmonic approximation determines the frequencies of the sinusoids based upon the fundamental frequency, f_0, i.e. $f_l = lf_0$. Furthermore, to simplify estimation of sinusoid amplitudes, a_l, these are obtained by sampling a spectral envelope function at sinusoid frequencies, i.e. $a_l = |\hat{X}(lf_0)|$ where $|\hat{X}(f)|$ is an estimate of the clean speech spectral envelope. The sinusoid phases are obtained by sampling the phase spectrum of the noisy speech, $\angle Y(f)$, i.e. $\vartheta_l = \angle Y(lf_0)$. To synthesise unvoiced speech, best quality was produced by inputting white noise, $w(n)$, into a filter, $\eta_X(n)$, with frequency response determined by the spectral envelope. With these assumptions the enhanced speech, $\hat{x}(n)$, is reconstructed using a variant of the sinusoidal model that operates in two states – harmonic for voiced speech or noise for unvoiced speech

$$\hat{x}(n) = \begin{cases} \sum_{l=1}^{L} |X(lf_0)| \, cos(2\pi lf_0 n + \vartheta_l) & \text{voiced} \\ w(n) * \eta_X(n) & \text{unvoiced} \end{cases} \tag{2}$$

To allow sinusoid amplitudes to be estimated at specific (integer) harmonic frequencies the spectral envelope is sampled at 1Hz intervals – this is discussed in Section 3.3. Further improvements in quality were obtained using formant enhancement to compensate for over-smoothing of formants and sub-frame reconstruction to avoid discontinuities in fundamental frequency between frames [33, 34].

2.3. Parameter estimation

Given the speech model, the challenge for speech enhancement is to estimate robustly the voicing, fundamental frequency and spectral envelope. Simple methods of voicing classification use parameters such as signal energy, spectral slope and harmonicity, while more advanced methods have used machine learning [35]. Many methods for fundamental frequency estimation have been proposed and operate in the time, frequency or cepstral domains [36, 37, 38]. Spectral envelope estimation seeks a smooth contour that joins spectral peaks and is estimated typically using LPC analysis, cepstrum processing or filterbank analysis [26, 39, 40]. These methods are generally accurate in clean conditions but deteriorate as SNRs fall.

To provide robust estimates of the acoustic features, and to have an integrated method that provides all the acoustic features needed for reconstruction, this work builds on previous work [23] and uses a statistical framework to estimate voicing,

fundamental frequency and spectral envelope. The next section presents the integrated statistical framework and further extends previous work by adapting the models to the current noise and speaker which is necessary for robust operation across different noise environments and speakers.

3. Acoustic feature estimation

To estimate voicing, fundamental frequency and spectral envelope a statistical approach is proposed. Several studies have shown correlation to exist between feature vectors (MFCCs, filterbank, LPC coefficients) and acoustic features (fundamental frequency, voicing, spectral envelope and formants) [32, 41, 42]. This correlation can be exploited and a model of the joint density of feature vectors and acoustic features created, from which an acoustic feature can be estimated from an input feature vector. Separate models could be produced to estimate each acoustic feature individually or a single model produced to estimate all acoustic features. Both methods have been investigated and slightly higher accuracy was observed when using a single model, which forms the basis of the acoustic feature estimation used in this work.

Modelling the joint density begins by defining a joint vector, $\boldsymbol{z}_i$, that contains a feature vector component, $\boldsymbol{x}_i$, and acoustic feature vector, $\boldsymbol{\theta}_i$, for frame i. The acoustic feature vector comprises a fundamental frequency value, f_{0i}, and spectral envelope vector, $\boldsymbol{\chi}_i$, as needed for speech reconstruction

$$\boldsymbol{z}_i = [\boldsymbol{x}_i, \boldsymbol{\theta}_i]^T = [\boldsymbol{x}_i, f_{0i}, \boldsymbol{\chi}_i]^T \tag{3}$$

The spectral envelope vector, $\boldsymbol{\chi}$, contains elements $\chi(m)$ which are the log amplitudes of an M-channel filterbank – note that for simplification of notation the frame subscript i is now omitted. Several choices exist for the feature vector although tests found MFCC vectors to give highest estimation accuracy. These were computed from 20ms frames of Hamming windowed speech, extracted every 10ms, which were transformed to a power spectrum upon which a mel-scale filterbank, log and DCT were applied. Previous work on speech reconstruction from MFCC vectors within a DSR architecture was constrained to the Aurora standard which specifies a 23 channel filterbank and truncation to a 13-D MFCC vector [23, 43]. With application to speech enhancement

this restriction is removed. Tests were performed that considered both linear-spaced and mel-spaced filterbanks with from 16 to 128 channels and found that highest speech quality was obtained using a 32 channel mel-filterbank and no subsequent truncation [32, Ch. 3.5.2]. Other variants of MFCC vectors have also been proposed and aim to provide a more robust representation. These include better modelling of the human auditory system by including masking effects, and more noise robust representations [44, 45]. These may give better performance, although at present noise and speaker robustness is achieved through explicit compensation, introduced in Sections 3.5 and 3.6.

3.1. Voicing classification

Voicing classification begins by creating three training vector pools, Υ^{v}, Υ^{uv} and Υ^{ns}, that contain joint vectors, $\boldsymbol{z}$, from voiced speech, unvoiced speech and non-speech respectively. Applying expectation-maximisation (EM) training to each vector pool creates Gaussian mixture models (GMMs), Φ_v, Φ_{uv} and Φ_{ns}, that model voiced, unvoiced and non-speech. The joint vectors used to train the unvoiced and non-speech GMMs do not have meaningful fundamental frequency components and these can be set to zero in those vector pools (Υ^{uv} and Υ^{ns}), although a modification to the EM training is necessary to avoid overflow with the zero-valued variance of the fundamental frequency. Alternatively, fundamental frequency can be removed from the joint feature vector with Φ_{uv} and Φ_{ns} trained on the remaining components. Both approaches were tested with no significant difference in accuracy.

Considering, for example, the voiced GMM, this models the joint density of the MFCC vector and acoustic feature vector for voiced speech

$$p(\boldsymbol{z}|\Phi_v^z) = \sum_{k=1}^{K} \alpha_{k,v}\, \phi_{k,v}^z(\boldsymbol{z}) = \sum_{k=1}^{K} \alpha_{k,v}\, \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{k,v}^z, \boldsymbol{\Sigma}_{k,v}^{zz}) \tag{4}$$

The kth mixture component, $\phi_{k,v}^z$, within the voiced GMM has mean vector $\boldsymbol{\mu}_{k,v}^z$ and covariance matrix $\boldsymbol{\Sigma}_{k,v}^{zz}$

$$\boldsymbol{\mu}_{k,v}^z = \begin{bmatrix} \boldsymbol{\mu}_{k,v}^x \\ \boldsymbol{\mu}_{k,v}^\theta \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_{k,v}^{zz} = \begin{bmatrix} \boldsymbol{\Sigma}_{k,v}^{xx} & \boldsymbol{\Sigma}_{k,v}^{x\theta} \\ \boldsymbol{\Sigma}_{k,v}^{\theta x} & \boldsymbol{\Sigma}_{k,v}^{\theta\theta} \end{bmatrix} \tag{5}$$

where the mean vector comprises means of the MFCC vector, $\boldsymbol{\mu}^{x}_{k,v}$, and acoustic feature vector, $\boldsymbol{\mu}^{\theta}_{k,v}$. The covariance matrix contains covariances of the MFCC vector, $\boldsymbol{\Sigma}^{xx}_{k,v}$, and acoustic feature vector, $\boldsymbol{\Sigma}^{\theta\theta}_{k,v}$, and their cross-covariances, $\boldsymbol{\Sigma}^{x\theta}_{k,v}$ and $\boldsymbol{\Sigma}^{\theta x}_{k,v}$. Prior probabilities, $\alpha_{k,v}$, reflect the proportion of training data in each cluster.

An estimate of the voicing, $\hat{v}$, of an input feature vector, $\boldsymbol{x}$, is made from the probability of the vector being voiced, unvoiced and non-speech using each GMM, marginalised to the MFCC vector component, and selecting the highest probability

$$\hat{v} = \underset{vc\in\{v,uv,ns\}}{\arg\max}\ p(\boldsymbol{x}|\Phi^{x}_{vc}) \tag{6}$$

where Φ^{x}_{vc} is the GMM marginalised to the MFCC vector.

3.2. Fundamental frequency estimation

Fundamental frequency is estimated from the voiced GMM, Φ_v, marginalised to MFCC and fundamental frequency components to give $\Phi^{xf_0}_{v}$. An estimate of the fundamental frequency, $\hat{f}_{0k}$, from the kth mixture component in the GMM is made from MFCC vector, $\boldsymbol{x}$, using the conditional mean which is calculated

$$\hat{f}_{0k} = \mu^{f_0}_{k,v} + \Sigma^{f_0,x}_{k,v}\left(\Sigma^{xx}_{k,v}\right)^{-1}\left(\boldsymbol{x} - \mu^{x}_{k,v}\right)^{T} \tag{7}$$

Fundamental frequency estimates from each mixture component of the GMM are combined by weighting according to the posterior probability, $h^{x}_{k,v}(\boldsymbol{x})$, of the MFCC vector from the kth mixture component of the voiced GMM

$$\hat{f}_0 = \sum_{k=1}^{K} h^{x}_{k,v}(\boldsymbol{x})\left[\mu^{f_0}_{k,v} + \Sigma^{f_0,x}_{k,v}\left(\Sigma^{xx}_{k,v}\right)^{-1}\left(\boldsymbol{x} - \mu^{x}_{k,v}\right)^{T}\right] \tag{8}$$

where the posterior probability is computed from marginalised distributions $\phi^{x}_{k,v}$

$$h^{x}_{k,v}(\boldsymbol{x}) = \frac{\alpha_{k,v}\ p\left(\boldsymbol{x}|\phi^{x}_{k,v}\right)}{\sum_{j=1}^{J}\alpha_{j,v}\ p\left(\boldsymbol{x}|\phi^{x}_{j,v}\right)} \tag{9}$$

3.3. Spectral envelope estimation

The estimate of the spectral envelope vector, $\boldsymbol{\chi}$, is made from the voiced or unvoiced GMM, depending on the voicing classification. The GMM is marginalised to the MFCC vector and spectral envelope components, $\Phi_v^{x\chi}$ or $\Phi_{uv}^{x\chi}$, depending on the voicing. Estimation from the voiced GMM, for example, is given as

$$\hat{\boldsymbol{\chi}} = \sum_{k=1}^{K} h_{k,v}^{x}(\boldsymbol{x}) \left[\mu_{k,v}^{\boldsymbol{\chi}} + \Sigma_{k,v}^{\boldsymbol{\chi},x} \left(\Sigma_{k,v}^{xx} \right)^{-1} (\boldsymbol{x} - \mu_{k,v}^{x})^{T} \right] \tag{10}$$

where $\phi_{k,v}^{x\chi}$ is the kth mixture component of the voiced GMM marginalised to the MFCC vector and spectral envelope components. The procedure for estimation using the unvoiced GMM is identical. Cubic spline interpolation is applied to the M-channel spectral envelope vector, $\hat{\boldsymbol{\chi}}$, to give a 1Hz frequency resolution estimate of the spectral envelope function, $|\hat{X}(f)|$, used in Eq.(2) to provide the sinusoid amplitudes

$$|\hat{X}(f)| = interp(\hat{\boldsymbol{\chi}}) \tag{11}$$

3.4. Selection of optimal number of mixture components

To determine the optimal number of mixture components, K, GMMs were trained with K from 1 to 512 and the accuracy of the resulting voicing classification, fundamental frequency and spectral envelope estimates measured. Voicing classification reached maximum accuracy with K=16, while for the more complex tasks of fundamental frequency and spectral envelope estimation peak performance was with K=256. Voicing classification did not deteriorate with K=256 mixture components and so K is set to 256.

3.5. Adaptation to noise

To reconstruct a clean speech signal the estimates of acoustic features must be robust to noise. One method would be to pre-filter the noisy speech before estimation, however this returns to problems faced by conventional speech enhancement. Instead, it is proposed to adapt the GMMs modelling the joint densities of MFCC vector and acoustic features to the noise conditions. Noise will affect the MFCC vector component but not the voicing, fundamental frequency or spectral envelope of the talker, so it is only the MFCC component of the joint density that must be adapted.

Several methods for adapting HMM-based speech recognisers to noisy conditions have been used successfully. These adapt clean speech statistics within the states of the HMMs to model noisy speech using methods such as parallel model combination, vector Taylor series (VTS) and the unscented transform [46, 47, 48]. The adaptation required for acoustic feature estimation differs somewhat, as the joint vector, $\boldsymbol{z}$, comprises an MFCC vector component, $\boldsymbol{x}$, that needs to be adapted to model noisy MFCC vector, $\boldsymbol{y}$, and an acoustic feature component, $\boldsymbol{\theta}$, that needs no adaptation – i.e. transforming the GMM $\Phi^{x\theta}$ into $\Phi^{\hat{y}\theta}$. All three adaptation methods could be applied to acoustic feature estimation, although initial tests found the unscented transform to work best. Adaptation is considered as a two stage process that requires first the derivation of a mismatch function to model the effect of noise on the MFCC vector and secondly application of the unscented transform to adapt the GMMs to noise.

3.5.1. Derivation of mismatch function

In the power spectral domain the noisy speech, $|Y(f)|^2$, is formed from the addition of clean speech, $|X(f)|^2$, noise, $|D(f)|^2$, and a phase component

$$|Y(f)|^2 = |X(f)|^2 + |D(f)|^2 + \quad 2|X(f)||D(f)|cos(\varphi(f)) \qquad 0 \leq f \leq F-1 \tag{12}$$

where $\varphi(f)$ is the phase difference between the noise and clean speech in the fth spectral bin. In many noise adaptation methods this phase-related mismatch is ignored. Recent studies have, however, shown that retaining the phase component improves modelling of the noisy speech [49].

Considering the stages of MFCC extraction, M-channel mel-filterbank features, $y^{fb}(m)$, are obtained by multiplying the power spectrum by an $M \times F$ matrix, $\boldsymbol{W}$, with each row corresponding to a mel-filterbank basis function to give

$$y^{fb}(m) = x^{fb}(m) + d^{fb}(m) + \quad 2\beta(m)\sqrt{x^{fb}(m)d^{fb}(m)} \qquad 0 \leq m \leq M-1 \tag{13}$$

where $x^{fb}(m)$ and $d^{fb}(m)$ are the clean and noise mel-filterbank features and $\beta(m)$ is related to the phase difference between the clean speech and noise in the mth

filterbank channel and is defined

$$\beta(m) = \frac{\sum_{f=0}^{F-1} W(m,f) cos\left(\varphi(f)\right) |X(f)||D(f)|}{\sqrt{x^{fb}(m) d^{fb}(m)}} \tag{14}$$

Taking the log of Eq.(13) and multiplying by an $M \times M$ discrete cosine transform (DCT) matrix, $\boldsymbol{C}$, where each row corresponds to a cosine basis function, a mismatch function, $g(.)$, gives noisy MFCC vector, $\boldsymbol{y}$, from clean speech and noise MFCC vectors, $\boldsymbol{x}$ and $\boldsymbol{d}$, and phase term, $\boldsymbol{\beta}$,

$$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{y}^l = \boldsymbol{C}g(\boldsymbol{C}^{-1}\boldsymbol{x}, \boldsymbol{C}^{-1}\boldsymbol{d}, \boldsymbol{\beta}) \tag{15}$$

where the superscript l denotes a log filterbank vector with the mismatch function, $g(.)$, defined

$$g(\boldsymbol{x}^l, \boldsymbol{d}^l, \boldsymbol{\beta}) = \boldsymbol{x}^l + \log\left(1 + \exp^{\boldsymbol{d}^l - \boldsymbol{x}^l} + 2\boldsymbol{\beta}\sqrt{\exp^{\boldsymbol{d}^l - \boldsymbol{x}^l}}\right) \tag{16}$$

An explicit value of $\boldsymbol{\beta}$ is not known, however following [49], an estimate is made using a lookup table that is computed offline during a training stage. For a given $\boldsymbol{x}$ and $\boldsymbol{d}$, the lookup table outputs a phase averaged estimate of $\boldsymbol{\beta}$ that is used in Eq. 16. If the phase component is ignored, i.e. $\boldsymbol{\beta} = [\boldsymbol{0}]$, the mismatch function becomes the conventional phase-independent mismatch function.

3.5.2. Adapting model parameters

The unscented transform is applied to the voiced, unvoiced and non-speech GMMs to adapt them to the joint density of noisy MFCC vectors and acoustic speech features. The unscented transform is an effective method for estimating the statistics of a distribution that has undergone a non-linear transformation as is the case of speech and noise addition in the MFCC domain [48]. As noise affects only the MFCC vector and not the acoustic feature, the unscented transform adapts only the statistics of the MFCC component of the joint vector.

For each GMM and for each mixture component, k, a set of $2 \times (M + M_\theta)$ sigma points, $\boldsymbol{s}_{i,k}^z$, are chosen

$$\boldsymbol{s}_{i,k}^z = \left[\boldsymbol{s}_{i,k}^x, \boldsymbol{s}_{i,k}^\theta\right]^T \tag{17}$$

where M_θ is the dimensionality of the acoustic feature vector. The sigma points are chosen so that their mean and covariance equal the mean and covariance of the kth mixture component in the GMM, i.e. $\boldsymbol{\mu}_k^z$ and $\boldsymbol{\Sigma}_k^{zz}$.

A single Gaussian is also trained on MFCC vectors extracted from noise which has mean and covariance $\boldsymbol{\mu}^d$ and $\boldsymbol{\Sigma}^{dd}$. A new mean and covariance, $\tilde{\boldsymbol{\mu}}^d$ and $\tilde{\boldsymbol{\Sigma}}^{dd}$, are then created which have the same dimensionality as the joint vector, i.e. $M + M_\theta$

$$\tilde{\boldsymbol{\mu}}^d = \begin{bmatrix} \boldsymbol{\mu}^d_{(M)} \\ \mathbf{0}_{(M_\theta)} \end{bmatrix} \quad \text{and} \quad \tilde{\boldsymbol{\Sigma}}^{dd} = \begin{bmatrix} \boldsymbol{\Sigma}^{dd}_{(M\times M)} & \mathbf{0}_{(M\times M_\theta)} \\ \mathbf{0}_{(M_\theta\times M)} & \mathbf{0}_{(M_\theta\times M_\theta)} \end{bmatrix} \tag{18}$$

A set of $2 \times (M + M_\theta)$ sigma points, $\boldsymbol{s}_i^{\tilde{d}}$, representing noise are now generated from the noise distribution and chosen to have the same mean and covariance as Eq.(18)

$$\boldsymbol{s}_i^{\tilde{d}} = \left[\boldsymbol{s}_i^d, \mathbf{0}\right]^T \tag{19}$$

Using the mismatch function, $g(.)$, in Eq.(16) the clean MFCC sigma points, $\boldsymbol{s}_{i,k}^x$, and noise MFCC sigma points, $\boldsymbol{s}_i^d$, are combined to give a set of noisy MFCC sigma points, $\boldsymbol{s}_{i,k}^{\hat{y}}$. These then replace the clean MFCC points in the joint vector

$$\boldsymbol{s}_{i,k}^{\hat{z}} = \left[\boldsymbol{s}_{i,k}^{\hat{y}}, \boldsymbol{s}_{i,k}^{\theta}\right]^T = \left[g(\boldsymbol{s}_{i,k}^x, \boldsymbol{s}_i^d,\ \boldsymbol{\beta}), \boldsymbol{s}_{i,k}^{\theta}\right]^T \tag{20}$$

For each mixture component the mean and covariance of the new set of sigma points, $\boldsymbol{s}_{i,k}^{\hat{z}}$, are computed and these provide the estimate of the noise adapted statistics for the GMM that now models the joint density of noisy MFCC vector and acoustic feature, i.e. $\phi_k^{\hat{y}\theta}$.

Adaptation requires statistics of the noise and many methods have been proposed to provide these which include voice activity detection, minimum statistics and speech presence probability methods [50, 51, 52, 53]. In this implementation 50 frames of noise (0.5 seconds) are taken from the start of the utterance, with the assumption that there is a period of non-speech prior to the speech, and these provide estimates of the noise mean and covariance, $\boldsymbol{\mu}^d$ and $\boldsymbol{\Sigma}^{dd}$. Tests found that the noise estimate from using only 50 frames was able to give a substantial reduction in acoustic feature estimation error as the mismatch between the clean acoustic models and noisy MFCC

vectors was largely removed. In a practical implementation, particularly for longer duration continuous speech, other noise estimation methods could be used to provide updated noise estimates such as those discussed in [54, Ch. 6].

3.6. Adaptation to speaker

When acoustic models are trained on the speaker under test, feature estimation accuracy is high, but to achieve accurate estimation across all potential speakers it is necessary adapt the acoustic models. Speaker adaptation has been successful in HMM-based speech recognition where speaker-independent acoustic models are adjusted to match the speaker under test. The close similarity between the statistical modelling used in acoustic feature estimation and that in the states of HMMs for speech recognition allows speaker adaptation methods to be applied. In speech recognition, adaptation is needed to adapt the statistics of the many hundreds or thousands of GMMs within the states of the HMMs to a new speaker. In comparison, for acoustic feature estimation it is only the voiced and unvoiced GMMs that need to be adapted.

Several approaches to speaker adaptation have been developed and applied successfully to HMM-based speech recognition and include MAP and maximum likelihood linear regression (MLLR) methods [55, 56, 57]. Of these methods, early investigations found MAP adaptation to be more effective than MLLR even with small amounts of adaptation data as there was sufficient coverage of the voiced and unvoiced classes [32]. This approach is similar to speaker adaptation used in GMM-based speaker identification based on a universal background model (UBM) that is trained on a large number of speakers. This models all speakers while a MAP adapted version of the UBM models the speaker under test [58]. The remainder of this section explains the application of MAP speaker adaptation to acoustic feature estimation.

3.6.1. Application of MAP speaker adaptation

MAP adaptation is applied to the means, covariances and prior probabilities of the GMMs to adapt them to model the speaker under test. From this speaker a set of adaptation vectors is extracted, $\boldsymbol{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N\}$. These take the same form as the joint feature vector defined in Eq.(3). For each adaptation vector, $\boldsymbol{a}_i$, the probability of each mixture component in the GMM, $\gamma_k(i)$, is computed (the

procedure is explained for the voiced GMM but is identical for the unvoiced GMM)

$$\gamma_k(i) = \frac{\alpha_k \, p(\boldsymbol{a}_i|\phi^z_{k,v})}{\sum_{j=1}^{K} \alpha_j p(\boldsymbol{a}_i|\phi^z_{j,v})} \tag{21}$$

The MAP adapted mean, $\boldsymbol{\mu}^{\hat{z}}_{k,v}$, for the kth mixture component in the GMM is calculated as a weighted combination of the prior mean from the original speaker-independent model, $\boldsymbol{\mu}^{z}_{k,v}$, and the estimated mean of the adaptation data in the kth mixture component [56]

$$\boldsymbol{\mu}^{\hat{z}}_{k} = \frac{\tau \boldsymbol{\mu}^{z}_{k,v} + \sum_{i=1}^{N} \gamma_k(i) \boldsymbol{a}_i}{\tau + \sum_{i=1}^{N} \gamma_k(i)} \tag{22}$$

τ is a tuneable parameter that determines the weighting between the mean of the adaptation data and the prior mean. With larger amounts of adaptation data τ can be increased.

Similarly the covariances, $\boldsymbol{\Sigma}^{\widehat{zz}}_{k,v}$, and mixture weights, $\hat{\alpha}_{k,v}$, can be updated using the adaptation vectors as

$$\boldsymbol{\Sigma}^{\widehat{zz}}_{k,v} = \frac{\boldsymbol{\Sigma}^{zz}_{k,v} + \sum_{i=1}^{N} \gamma_k(i)(\boldsymbol{a}_i - \boldsymbol{\mu}^{\hat{z}}_{k})(\boldsymbol{a}_i - \boldsymbol{\mu}^{\hat{z}}_{k})^T + \boldsymbol{\Psi}}{\omega_k - (M + M_\theta) + \sum_{i=1}^{N} \gamma_k(i)} \tag{23}$$

$$\text{where} \;\; \boldsymbol{\Psi} = \tau(\boldsymbol{\mu}^{z}_{k} - \boldsymbol{\mu}^{\hat{z}}_{k})(\boldsymbol{\mu}^{z}_{k} - \boldsymbol{\mu}^{\hat{z}}_{k})^T$$

$$\hat{\alpha}_{k,v} = \frac{\alpha_{k,v} - 1 + \sum_{i=1}^{N} \gamma_k(i)}{\sum_{j=1}^{K} \left(\alpha_{k,v} - 1 + \sum_{i=1}^{N} \gamma_j(i) \right)} \tag{24}$$

ω_k relates to the summed probability of adaptation vectors for the kth mixture component and is defined in [56]. The means and covariances can be calculated using the same or independent values of τ, with values typically in the range two to twenty [55]. For the experiments in Section 4, a value of τ=12 was used throughout as this was found to give best performance.

3.6.2. Practical considerations

An important difference between adaptation for speech recognition and its application to acoustic feature estimation is the joint feature vector. In speech recognition the feature vector is typically an MFCC vector that needs to be adapted

to the new speaker. For feature estimation the joint feature vector comprises both an MFCC component and an acoustic feature component. This leads to two approaches that have been considered for adaptation: to adapt only the MFCC component or to adapt both the MFCC and acoustic feature components.

Adaptation of only the MFCC component of the GMMs is achieved by marginalising Eq.(21) so only the MFCC vector determines the contribution of each mixture component in the GMM. Similarly Eqs.(22), (23) and (24) are marginalised so only the MFCC vector components of the GMM are adapted. Adapting the acoustic feature component of the GMM as well can be more difficult as voicing and fundamental frequency errors can occur in the adaptation data which may affect adaptation accuracy. Tests comparing both approaches found that full adaptation gave lower estimation errors than marginalised adaptation. Consequently, for the experimental evaluation presented in Section 4 the full adaptation method is used.

In this work adaptation is supervised and applied statically using 20 seconds of data for each new speaker under test. Adaptation was found to be faster than reported for application to speech recognition, primarily as fewer Gaussians require adaptation. After 20 seconds of adaptation data, 75% of the performance gain that is attained with fully adapted models was reached, and after even a few seconds of adaptation data, improvement was considerable. Adaptation could also be unsupervised and applied dynamically, and the models updated continuously as increasing amounts of speaker data are processed during the utterance [59].

4. EXPERIMENTAL RESULTS

Experiments first examine the effect that voicing, fundamental frequency and spectral envelope parameters have on speech quality and establish the configuration of the speech reconstruction model. A second set of experiments then compares reconstruction-based enhancement with a range of conventional methods of enhancement using PESQ, subjective listening tests and spectrogram analysis. Speech data is taken from the WSJCAM0 database and is downsampled to 8kHz [60]. Forty-eight female speakers and 63 male speakers provide training data and a further 5 female and 5 male speakers are used for testing with utterances ranging from 5 to 10 seconds in duration. Each training speaker provides approximately 100 sentences and each

Table 1: Spectral envelope, voicing and f_0 configurations.

Method	Spectral envelope	Voicing and f_0
REC(noisy, noisy)	Noisy	Noisy
REC(adapt, adapt)	Adaptation	Adaptation
REC(adapt, noisy)	Adaptation	Noisy
REC(noisy, adapt)	Noisy	Adaptation

testing speaker provides 45 sentences.

4.1. Voicing, fundamental frequency and spectral envelope

Using the speech reconstruction model, four configurations of voicing, fundamental frequency and spectral envelope estimation are examined as shown in Table 1. The first configuration, REC(noisy,noisy), extracts spectral envelope and fundamental frequency from noisy speech with no adaptation of the acoustic models to speaker and noise and represents a basic system. Conversely, REC(adapt, adapt) extracts parameters from noisy speech but adapts the acoustic models to the speaker and noise and represents a fully adapted system. REC(adapt, noisy) and REC(noisy, adapt) examine the effect when estimation of one of the acoustic features is degraded, by applying adaptation to only the spectral envelope or the fundamental frequency/voicing components and leaving the other unadapted. Specifically REC(adapt, noisy) adapts the spectral envelope component of the acoustic models but leaves voicing and fundamental frequency unadapted, while REC(noisy, adapt) adapts the voicing and fundamental frequency components, but leaves the spectral envelope component unadapted. The four configurations were analysed using three-way mean opinion score (MOS) tests. Twenty listeners took part in the tests which were conducted in a sound-proof room with utterances played through headphones in accordances with the ITU-T recommendations [25]. Each utterance was played three times, and after each instance listeners rated the sample in terms of first signal quality, then noise intrusiveness and finally overall quality on a scale of 1 to 5 with five indicating good performance. Table 4 shows results in car noise at SNRs of 20dB, 10dB and 5dB.

Extracting spectral envelope and voicing/fundamental frequency from noisy speech, REC(noisy,noisy), gave lowest quality which was expected. Applying adaptation to both spectral envelope and voicing/fundamental frequency estimation, REC(adapt, adapt), improved overall quality and in particular reduced background

Table 2: Three-way MOS scores for different speech enhancement configurations in car noise at SNRs of 20dB, 10dB and 5dB showing signal quality, background noise intrusiveness and overall quality. .

	SNR	REC(noisy,noisy)	REC(adapt, adapt)	REC(adapt, noisy)	REC(noisy, adapt)
Signal	20dB	3.20	3.40	3.45	3.95
	10dB	2.95	2.90	2.80	2.80
	5dB	2.30	2.15	1.95	2.15
Noise	20dB	2.60	4.20	4.10	2.50
	10dB	1.80	3.95	3.50	1.80
	5dB	1.30	3.50	3.40	1.20
Overall	20dB	2.90	3.70	3.65	3.05
	10dB	2.30	3.10	2.90	2.20
	5dB	1.50	2.25	2.00	1.50

noise. Changing to using the noisy fundamental frequency, REC(adapt, noisy), introduced a small deterioration in performance. However, using the noisy spectral envelope, REC(noisy, adapt), caused a larger reduction in overall quality and background noise. These tests confirm that adaptation improves acoustic features and consequently the reconstructed speech. The speech is more sensitive to spectral envelope errors than fundamental frequency errors. These often lead to artifacts or distortion, while errors in fundamental frequency tend to be small variations in f_0 that do not introduce such perceptible changes to the speech.

Investigating adaptation times, by measuring acoustic feature estimation errors, found that adapting to noise was fast, requiring between 1 and 3 seconds of noise data to be fully adapted. Speaker adaptation required more data and converged after 120 seconds. However, after only 5-10 seconds of speaker adaptation data, performance reached around 50% of that of the fully adapted model [32, Ch. 5.4.1].

4.2. Comparative objective evaluation

Experiments now compare reconstruction-based enhancement with other methods of enhancement using PESQ analysis. Tests were carried out in car noise, babble noise and destroyer operations room noise at SNRs from 0dB to +15dB [61] and shown in Table 3. Three variants of reconstruction-based enhancements are considered with the aim of further investigating the impact that noise and speaker adaptation has on speech quality. The first variant (REC0) has no noise or speaker adaptation, the second variant (REC1) includes noise adaptation while the third variant (REC2) uses both noise and speaker adaptation. Baseline performance is given with no noise compensation (NNC) and seven other enhancement methods are also included that cover a range of different techniques:

1. LOG: MMSE log spectral magnitude estimation with speech probability uncertainty [10]
2. MPC: MMSE spectral magnitude estimation with phase spectrum compensation [13]
3. SUB: Generalised subspace method [11]
4. CSM: Enhancement using a constrained iterative sinusoidal model [14]
5. WNR: Wiener filter using GMM gain estimation [6]
6. INV: GMM mapping of noisy MFCCs to clean MFCCs followed by inversion to time-domain [22]
7. NMF: Non-negative matrix factorisation [5].

These cover broadly the various methods discussed in the Introduction although the HMM-based enhancement has not been included. Tests found this method to be sensitive to the decoding accuracy of the HMMs which was rather high for the noisy WSJCAM0 task. In many cases at lower SNRs, the model sequence had too many errors to synthesise sufficiently intelligible speech. Methods LOG, MPC, SUB and CSM are considered unsupervised techniques that have no prior model of the speech or noise, and in these tests obtain noise estimates using unbiased MMSE-based noise power estimation [53]. Conversely the WNR, INV, NMF and REC methods can be considered as being supervised as they require some kind of prior model of the speech that is obtained during a training stage.

The results show unadapted reconstruction-based enhancement (REC0) to perform relatively poorly and in some instances below that of NNC. The acoustic models are trained on clean speech but, with noisy speech input, are poorly matched which results in acoustic feature estimates that are highly erroneous leading to low speech quality. For example, in babble noise at an SNR of 5dB, voicing classification error is 44%. Applying noise adaptation (REC1) gives a substantial increase in speech quality which is attributed to more accurate estimates of acoustic features as the models are now more closely matched to the input speech – voicing classification error in babble noise at 5dB is reduced to 13%. Finally, applying noise and speaker adaptation (REC2) sees more improvements in speech quality, although the gains are smaller, and are again attributed to the acoustic models better matching the input speech – voicing classification error in babble noise at 0dB is now reduced to 11%.

Table 3: PESQ results in car noise, babble noise and destroyer noise at SNRs of 15dB, 5dB and 0dB.

	SNR	NNC	LOG	MPC	SUB	CSM	WNR	INV	NMF	REC0	REC1	REC2
Car	15dB	2.47	2.72	2.82	2.85	2.76	2.70	2.74	2.85	2.40	2.87	**2.94**
	5dB	1.86	1.84	2.14	2.06	2.25	2.13	2.24	2.24	1.90	2.35	**2.39**
	0dB	1.60	1.43	1.71	1.59	1.85	1.75	1.88	1.83	1.65	1.83	**1.89**
Babble	15dB	2.57	2.85	2.89	2.87	2.75	2.72	2.72	2.82	2.57	2.86	**2.91**
	5dB	1.91	1.98	2.18	2.13	2.28	2.16	2.20	2.20	1.96	2.19	**2.37**
	0dB	1.56	1.59	1.72	1.68	1.81	1.78	1.74	1.76	1.65	1.78	**1.87**
Dest.	15dB	2.64	2.97	**3.09**	3.08	2.85	2.81	2.76	2.97	2.64	2.90	2.99
	5dB	1.91	2.04	2.39	2.30	2.38	2.25	2.23	2.34	2.03	2.28	**2.41**
	0dB	1.75	1.76	2.01	1.87	2.00	1.91	1.87	1.97	1.64	1.85	**2.10**

Comparing reconstruction-based enhancement with the other methods shows the proposed method (REC2) to give the largest increase in PESQ score over no noise compensation in eight out of the nine noise and SNR combinations, with an average increase of 0.40 over NNC. With only noise adaptation (REC1) the increase over NNC is 0.29 while with no adaptation (REC0) the increase is 0.02. The enhancement method giving next best performance varies depending on noise type and SNR, although averaged across all conditions the next best methods were MPC and NMF which both attained an average increase of 0.30 over NNC. Interestingly, these represent one unsupervised method (MPC) and one supervised method (NMF), and have performance approximately equivalent to the REC1 method.

4.3. Comparative subjective evaluation

Subjective tests are now used to compare the reconstruction-based enhancement with competing methods of enhancement under the same noise and SNR conditions as in Section 4.2. Three-way MOS listening tests are used, but with a different set of 20 listeners from those in Section 4.1. Tables 4, 5 and 6 show MOS results for car noise, babble noise and destroyer noise, with each table showing scores for signal quality, background noise and overall.

Considering first noise removal, all methods reduce noise intrusiveness in comparison to no compensation. Reconstruction-based enhancement (REC2) achieves highest levels of noise removal in six out of the nine noise and SNR combinations. In four of these six configurations (car at 15dB, babble at 15dB and 5dB, and destroyer at 5dB), REC2 is statistically better than all other methods at the 95% confidence level, while in the other two configurations one other method is within the 95% confidence interval for REC2 – LOG in car noise at 10dB and INV in babble noise at 0dB. In the

Table 4: Three-way MOS results in car noise showing signal quality, background noise intrusiveness and overall quality.

	SNR	NNC	WNR	LOG	MPC	SUB	INV	CSM	REC2
Signal	15dB	**4.73**	4.19	4.38	4.23	4.15	3.65	3.92	3.88
	5dB	**4.50**	3.04	2.54	2.65	3.04	2.85	2.92	3.00
	0dB	**4.08**	1.88	1.15	1.73	1.88	1.92	1.98	1.88
Noise	15dB	2.88	3.38	4.09	3.65	3.58	3.81	3.35	**4.59**
	5dB	1.54	2.46	3.27	2.58	2.42	2.92	2.58	**3.53**
	0dB	1.15	1.65	**2.46**	2.19	1.88	2.23	1.89	2.42
Overall	15dB	3.35	3.54	**4.23**	3.88	3.85	3.46	3.58	3.96
	5dB	2.65	2.73	2.31	2.58	2.69	2.62	2.79	**3.04**
	0dB	1.69	1.54	1.27	1.65	**1.77**	1.58	**1.77**	1.73

Table 5: Three-way MOS results in babble noise showing signal quality, background noise intrusiveness and overall quality.

	SNR	NNC	WNR	LOG	MPC	SUB	INV	CSM	REC2
Signal	15dB	**4.81**	4.04	4.46	4.23	4.50	3.58	3.63	3.92
	5dB	**4.35**	2.77	2.85	2.81	3.42	2.92	3.15	2.85
	0dB	**4.12**	1.81	1.81	1.88	2.08	1.77	2.38	2.04
Noise	15dB	2.65	3.23	3.31	3.15	3.42	3.71	3.62	**4.35**
	5dB	1.88	2.19	2.23	2.23	2.31	2.67	2.31	**3.39**
	0dB	1.23	1.73	1.88	1.69	1.42	2.46	1.54	**2.54**
Overall	15dB	3.50	3.69	3.77	3.77	3.81	3.65	3.88	**4.00**
	5dB	2.77	2.38	2.46	2.58	2.58	2.46	**2.96**	**2.96**
	0dB	**2.19**	1.54	1.58	1.58	1.77	1.69	1.62	1.77

Table 6: Three-way MOS results in destroyer operations room noise showing signal quality, background noise intrusiveness and overall quality.

	SNR	NNC	WNR	LOG	MPC	SUB	INV	CSM	REC2
Signal	15dB	**4.58**	4.04	4.54	4.31	4.50	3.62	4.33	4.15
	5dB	**4.65**	2.88	2.69	3.00	3.12	2.69	3.18	3.08
	0dB	**4.27**	2.23	1.69	2.23	2.35	1.88	2.31	2.31
Noise	15dB	3.00	3.31	**4.31**	4.00	4.08	4.04	3.92	4.19
	5dB	1.88	2.23	3.15	2.81	2.69	2.92	2.76	**3.73**
	0dB	1.38	1.96	2.54	**2.69**	2.23	2.31	2.17	2.62
Overall	15dB	3.77	3.58	**4.46**	4.27	4.23	3.73	4.23	4.19
	5dB	2.69	2.50	2.65	2.77	2.71	2.62	2.69	**3.38**
	0dB	**2.38**	2.00	1.73	2.04	**2.38**	1.69	2.18	2.31

three remaining conditions REC2 is ranked second of the enhancement methods and is significantly better than many of the other methods tested. This is attributed to the underlying speech production model that is constrained to output a speech-like signal, which, when given robust parameters should be largely free from the original noise. An interesting observation with reconstruction-based enhancement is that the MOS for noise removal at a given SNR is almost constant across the three different types of noise. This suggests that the reconstruction removes successfully the original noise and then introduces a fairly consistent, SNR-dependent noise which can be likened to an artefact of the reconstruction process. Listening to reconstructed speech across all SNRs reveals no evidence of the original noise and confirms that a wideband-like noise is introduced, which becomes more noticeable at lower SNRs. This arises mainly from errors in spectral envelope estimation and also from voicing errors, typically occurring at word boundaries, which can introduce a short, but audible, burst of noise.

The filtering methods (LOG, MPC, SUB and WNR) all leave increasing amounts of residual noise from the original noise as SNRs reduce. These methods also introduce artefacts, which for log MMSE (LOG) and MMSE with phase spectrum compensation (MPC) is musical noise, for Wiener filtering (WNR) is a wideband-like noise and for the subspace method (SUB) is a swirling-like noise. The MFCC inversion method (INV) is often better at noise removal than filtering methods and leaves little residual noise but instead introduces a wideband-like noise which has similarities to that introduced by reconstruction-based enhancement. In fact, both methods reconstruct a new speech signal, for INV by inverting clean MFCC vectors estimated using a GMM, which accounts for the similar characteristics of the enhanced speech they produce. The constrained iterative sinusoidal model (CSM) method is also based upon a model of speech production although has less constraints in its parameter estimation than the proposed (REC2) method. This was found to leave more residual noise in the enhanced speech which led to lower scores for noise intrusiveness compared to REC2.

In terms of signal quality, reconstruction-based enhancement (REC2) performs slightly worse than the best performing filtering methods at 15dB but at lower SNRs has similar performance to the better performing filtering methods. The comparatively lower signal quality at higher SNRs is again attributed to the underlying speech production model, where in cleaner conditions the imperfect reconstruction from the

model is more evident. Conversely, at lower SNRs the distortion introduced by the model is masked by the increased level of noise and is no more apparent than with the other methods of enhancement. When applied to clean speech, reconstruction introduces a distortion above that of filtering methods which leave the signal unchanged given a sufficiently low amplitude noise estimate. There is no method that gives consistently highest signal quality, with SNR and noise type causing the best performing method to change.

For overall quality, the enhancement methods perform similarly, again with variations across noise type and SNR. At 5dB, reconstruction-based enhancement performs slightly better across all noise types, and is statistically best in destroyer noise at 95% confidence. This is attributed to its superior noise removal over other methods at that SNR and that at 5dB the signal distortion introduced is beginning to be masked.

4.4. Spectrogram analysis

Reconstruction-based enhancement is demonstrated in figure 1 which shows four spectrograms of the utterance '*Look out of the window and see if it's raining*'. The spectrograms show a) clean speech, b) speech contaminated by street noise at an SNR of 5dB, c) reconstruction-based enhancement, d) log MMSE enhancement with speech presence uncertainty (LOG). Figure 1b, shows street noise to be complex, containing several stationary and non-stationary noises coming from sources such as passing cars, people talking and a warning siren from a pedestrian crossing. The warning siren is visible on the spectrogram as tones in the first two-thirds of the utterance. Figure 1d shows LOG to be effective in removing stationary components of the noise although not as effective at removing the non-stationary noises. In contrast, reconstruction-based enhancement shown in figure 1c reconstructs an almost noise-free signal with none of the stationary or non-stationary noises which correlates closely to the listening test results. Some differences between the reconstructed speech and clean speech can be seen in the spectrograms and these are perceived as speech distortion.

4.5. Discussion

The PESQ analysis showed reconstruction-based enhancement to perform best in eight out of the nine configurations. Adaptation of the acoustic models to noise

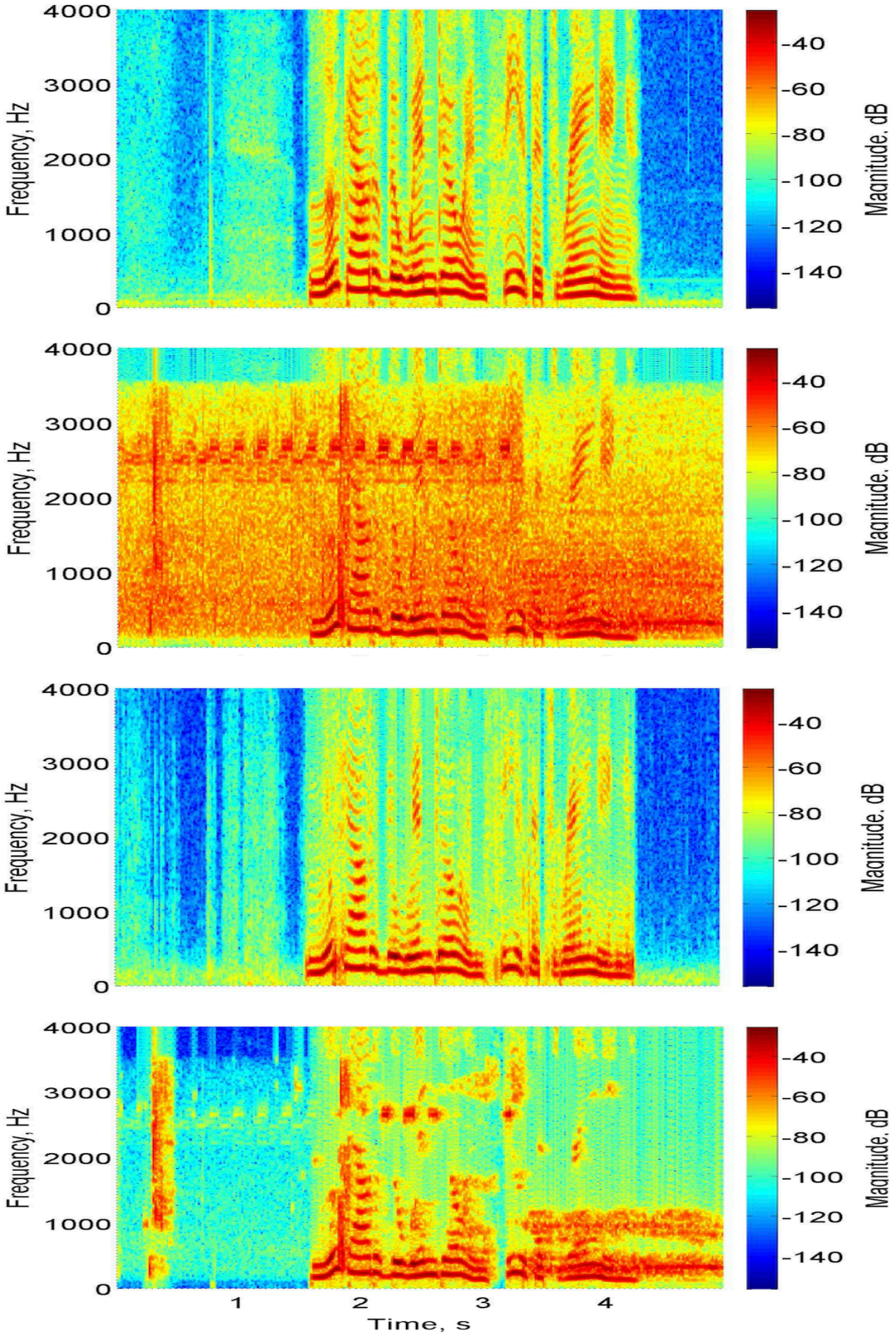

Figure 1: Spectrograms of a) clean, b) noisy, c) REC2 enhancement and d) LOG enhancement of the utterance '*Look out of the window and see if it's raining*'.

improved their match to the input noisy speech and gave substantial gains in speech quality due to the improved estimates of the acoustic features. Further adapting the models to the speaker gave more improvement in quality, although the gain was less than from noise adaptation. This suggests that noise adaptation is more important than speaker adaptation, and needed to provide at least some match between the the original noise-free models and the noise conditions. Matching to the specific speaker under test is less critical in terms of speech quality as the acoustic models are matched to some extent to the speaker through the speaker-independent training. The more detailed breakdown of performance made in the listening tests has given some insight into the attributes of reconstruction-based enhancement and found its strongest aspect to be its ability to reduce noise. This is attributed to the constraints imposed in reconstruction in terms of parameter estimation and by the sinusoidal model. The strength in noise reduction is also clearly evident in the spectrogram analysis which showed REC2 to leave very little residual noise. Some differences between the reconstructed speech and clean speech were observed in the spectrograms and are perceived as speech distortion which is also revealed in the listening tests where signal quality can be reduced at higher SNRs.

The pattern of these results is generally in line with a detailed investigation made into the effectiveness of speech enhancement algorithms [1, Ch. 11.2]. Although somewhat different enhancement methods were used in that work, and we consider lower SNRs, both studies found that relatively few methods made significant increases in overall quality. For noise reduction, both studies showed that significant reductions in background noise can be achieved by enhancement. In terms of signal quality, the study in [1, Ch. 11.2] suggested that a positive result is when the enhancement method introduces no reduction compared to the noisy signal quality and reported several enhancement methods able to achieve this. Our studies have shown comparable signal quality at high SNRs but a reduction at lower SNRs in comparison to the noisy signal. This we attribute to the scores for the noisy signal in our tests being significantly higher than reported in [1] making them more difficult to equal after enhancement.

5. Conclusion

This work has proposed a method of speech enhancement that reconstructs a clean speech signal from a sinusoidal model and a set of acoustic speech features estimated from noisy speech. The motivation is that by constraining the enhanced signal to be produced by a model of speech production, the resulting signal should be free from the original noise. Listening tests have confirmed this and comparing the proposed method to six other methods of speech enhancement has found it to be the most effective at removing noise from the noisy speech signal with only small amounts of residual noise apparent. Analysis has revealed two weaknesses of the proposed method. First, although residual noise is largely eliminated, errors in estimating the spectral envelope and voicing classification can lead to artefacts being introduced in the reconstructed speech. Secondly, the reconstructed signal quality can be slightly lower than filtering methods at higher SNRs. This is attributed to the speech model being unable to reconstruct the speech signal without introducing some distortion over the original clean speech. The difference is however slight and the relative difference becomes unnoticeable as SNRs reduce. Audio examples are available at *www.uea.ac.uk/computing/speech-language-and-audio-processing/se-results*.

References

[1] P. Loizou, Speech Enhancement: Theory and Practice, CRC Press, Inc., 2007.

[2] S. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoustics, Speech and Signal Processing 27 (2) (1979) 113–120.

[3] H.-T. Hu, F.-J. Kuo, H.-J. Wang, Supplementary schemes to spectral subtraction for speech enhancement, Speech Communication 36 (3) (2002) 205–218.

[4] P. Scalart, J. Vieira-Filho, Speech enhancement based on a priori signal to noise estimation, in: ICASSP, 1996, pp. 629–632.

[5] M. Mohammadiha, P. Smaragdis, A. Leijon, Supervised and unsupervised speech enhancement using nonnegative matrix factorisation, IEEE Trans. Audio, Speech and Language Processing 21 (10) (2013) 2140–2151.

[6] N. Hadir, F. Faubel, D. Klakow, A model-based spectral envelope Wiener filter for perceptually motivated speech enhancement, in: Interspeech, 2011, pp. 213–216.

[7] Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, IEEE Trans. Acoustics, Speech and Signal Processing 32 (6) (1984) 1109–1121.

[8] Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, IEEE Trans. Acoustics, Speech and Signal Processing 33 (2) (1985) 443–445.

[9] R. Martin, Speech enhancement based on minimum mean square error estimation and supergaussian priors, IEEE Trans. Speech Audio Process. 13 (5) (2005) 845–856.

[10] I. Cohen, Optimal speech enhancement under signal presence uncertainty using log-spectra amplitude estimator, IEEE Signal Processing Letters 9 (4) (2002) 113–1116.

[11] Y. Hu, P. Loizou, A generalized subspace approach for enhancing speech corrupted by colored noise, IEEE Trans. Speech Audio Process. 11 (4) (2003) 334–341.

[12] A. Stark, K. Wójcicki, J. Lyons, K. Paliwal, Noise-driven short-time phase spectrum compensation procedure for speech enhancement, in: Interspeech, 2008, pp. 549–552.

[13] K. Paliwal, K. Wójcicki, B. Shannon, The importance of phase in speech enhancement, Speech Communication 53 (2011) 465–494.

[14] J. Jensen, J. Hansen, Speech enhancement using a constrained iterative sinusoidal model, IEEE Trans. Audio, Speech and Language Processing 9 (7) (2001) 731–740.

[15] Q. Yan, S. Vaseghi, E. Zavarehei, B. Milner, J. Darch, P. White, I. Andrianakis, Kalman tracking of linear predictor and harmonic noise models for noisy speech enhancement, Computer Speech and Language 22 (1) (2008) 69–83.

[16] R. Chen, C.-F. Chan, H. So, Model-based speech enhancement with improved

spectral envelope estimation via dynamics tracking, IEEE Trans. Audio, Speech and Language Processing 20 (4) (2012) 1324–1336.

[17] H. Zen, K. Tokuda, A. Black, Statistical parametric speech synthesis, Speech Communication 51 (11) (2009) 1039–1064.

[18] J. Carmona, J. Barker, A. Gomez, N. Ma, Speech spectral envelope enhancement by HMM-based analysis/resynthesis, IEEE Signal Processing Letters 20 (6) (2013) 563–566.

[19] A. Kato, B. Milner, Using hidden Markov models for speech enhancement, in: Interspeech, Singapore, 2014, pp. 2695–2699.

[20] J. Ming, D. Crookes, Speech enhancement from additive noise and channel distortion - a corpus-based approach, in: Interspeech, 2014, pp. 2710–2714.

[21] X. Xiao, R. Nickel, Speech enhancement with inventory style speech resynthesis, IEEE Trans. Audio, Speech and Language Processing 18 (6) (2010) 1243–1257.

[22] L. Boucheron, P. D. Leon, Low SNR, speaker-dependent speech enhancement using GMMs and MFCCs, in: Interspeech, 2012, pp. 575–578.

[23] B. Milner, X. Shao, Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction, IEEE Trans. Audio, Speech and Language Processing 15 (1) (2007) 24–33.

[24] B. Milner, J. Darch, Robust acoustic speech feature prediction from noisy mel-frequency cepstral coefficients, IEEE Trans. Audio, Speech and Language Processing 19 (2) (2011) 338–347.

[25] ITU-T, P.835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms, ITU-T recommendation, 2003.

[26] J. Makhoul, Linear prediction: a tutorial review, Proc. IEEE 63 (4) (1975) 561–580.

[27] 3G TS 26.071, 3rd generation partnership project (3GPP) TSG-SA codec working group mandatory speech codec speech processing functions amr speech codec; general description (1999).

[28] H. Kawahara, I. Masuda-Katsuse, A. de Cheveigné, Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds, Speech Communication 27 (1999) 187–207.

[29] S. Takaki, K. Sawada, K. Hashimoto, K. Oura, K. Tokuda, Overview of NITECH HMM-based speech synthesis system for Blizzard Challenge 2013, in: Proc. Blizzard Challenge, 2013.

[30] R. McAulay, T. Quatieri, Speech analysis/synthesis based on a sinusoidal representation, IEEE Trans. Acoustics, Speech and Signal Processing 34 (4) (1986) 744–754.

[31] Y. Stylianou, Applying the harmonic plus noise model in concatenative speech synthesis, IEEE Trans. Speech Audio Process. 9 (1) (2001) 21–29.

[32] P. Harding, Model based speech enhancement, Ph.D. thesis, University of East Anglia, UK (2013).

[33] A. Chen, S. Vaseghi, P. McCourt, State based sub-band LP Wiener filters for speech enhancement in car environments, in: ICASSP, Vol. 1, Istanbul, Turkey, 2000, pp. 213–216.

[34] X. Shao, B. Milner, Clean speech reconstruction from noisy mel-frequency cepstral coefficients using a sinusoidal model, in: ICASSP, Vol. 1, Hong Kong, 2003, pp. 704–707.

[35] P. Harding, B. Milner, On the use of machine learning methods for voice activity detection, in: Interspeech, Portland, USA, 2012, pp. 709–712.

[36] A. de Cheveigné, H. Kawahara, YIN, a fundamental frequency estimator for speech and music, Journal of the Acoustical Society of America 111 (4) (2002) 1917–1930.

[37] S. Gonzalez, M. Brookes, PEFAC - a pitch estimation algorithm robust to high levels of noise, IEEE/ACM Trans. Audio, Speech and Language Processing 22 (2) (2014) 518–530.

[38] S. Ahmadi, A. Spanias, Cepstrum-based pitch detection using a new statistical V/UV classification algorithm, IEEE Trans. Audio, Speech and Language Processing 7 (3) (1999) 333–338.

[39] A. Oppenheim, R. Schafer, Discrete-Time Signal Processing, Prentice Hall, 1989.

[40] O. Cappe, E. Moulines, Regularization techniques for discrete cepstrum estimation, IEEE Signal Processing Letters 3 (4) (1996) 100–102.

[41] A. Syrdal, S. Steele, Vowel F1 as a function of speaker fundamental frequency, 110th Meeting of ASA, Journal of the Acoustical Society of America 78 (S1) (1985) S56.

[42] J. Darch, B. Milner, S. Vaseghi, Analysis and prediction of acoustic speech features from mel-frequency cepstral coefficients in distributed speech recognition architectures, Journal of the Acoustical Society of America 124 (6) (2008) 3989–4000.

[43] ETSI, Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm, ES 202 212 version 1.1.1, ETSI STQ-Aurora DSR Working Group (Nov. 2003).

[44] X. Luo, I. Y. Soon, C. K. Yeo, An auditory model for robust speech recognition, in: International Conference on Audio, Language and Image Processing (ICALIP), 2008, pp. 1105–1109.

[45] C. Kim, R. Stern, Power-normalized cepstral coefficients (PNCC) for robust speech recognition, in: ICASSP, 2012, pp. 4101–4105.

[46] M. Gales, S. Young, Robust continuous speech recognition using parallel model combination, IEEE Trans. Speech Audio Process. 4 (5) (1996) 352–359.

[47] P. Moreno, B. Raj, R. Stern, A vector Taylor series approach for environment-independent speech recognition, in: ICASSP, Vol. 2, 1996, pp. 733–736.

[48] Y. Hu, Q. Huo, An HMM compensation approach using unscented transformation

for noisy speech recognition, Lecture Notes in Computer Science, Chinese Spoken Language Processing 4274 (2006) 346–357.

[49] F. Faubel, J. McDonough, D. Klakow, A phase-averaged model for the relationship between noisy speech, clean speech and noise in the log-mel domain, in: Interspeech, 2008, pp. 553–556.

[50] R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics, IEEE Trans. Speech Audio Process. 9 (5) (2001) 504–512.

[51] S. Rangachari, P. Loizou, A noise estimation algorithm for highly nonstationary environments, Speech Communication 48 (2) (2006) 22—231.

[52] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, R. Martin, An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments, in: ICASSP, 2011, pp. 4640–4643.

[53] T. Gerkmann, R. Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay, IEEE Trans. Audio, Speech and Language Processing 20 (4) (2012) 1383–1393.

[54] R. Hendriks, T. Gerkmann, J. Jensen, DFT-domain based single-microphone noise reduction for speech enhancement, Morgan-Claypool, 2013.

[55] P. Woodland, Speaker adaptation for continuous density HMMs: A review, in: ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition (Adaptation-01), 2001, pp. 11–19.

[56] J.-L. Gauvain, C.-H. Lee, Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, IEEE Trans. Speech Audio Process. 2 (2) (1994) 291–298.

[57] C. Leggetter, P. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density HMMs, Computer Speech and Language 9 (2) (1995) 171–185.

[58] D. Reynolds, T. Quatieri, R. Dunn, Speaker verification using adapted Gaussian mixture models, Digital Signal Processing 10 (1–3) (2000) 19–41.

[59] F. Geiger, F. Wallhoff, G. Rigoll, GMM-UBM based open-set online speaker diarization, in: Interspeech, 2010.

[60] T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals, WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition, in: ICASSP, Vol. 1, 1995, pp. 81–84.

[61] A. Varga, H. Steeneken, Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems, Speech Communication 12 (3) (1993) 247–251.