Speech enhancement by reconstruction from cleaned acoustic features

Philip Harding, Ben Milner

University of East Anglia, Norwich, UK
p.harding@uea.ac.uk, b.milner@uea.ac.uk

Abstract

This paper proposes a novel method of speech enhancement that moves away from conventional filtering-based methods and instead aims to reconstruct clean speech from a set of speech features. Underlying the enhancement system is a speech model which at present is based on a sinusoidal model. This is driven by a set of speech features, comprising voicing, fundamental frequency and spectral envelope, that are extracted from the noisy speech. A maximum *a posteriori* approach is proposed for estimating clean spectral envelope features from the noisy spectral envelope. A set of subjective tests, measuring speech quality, noise intrusiveness and overall quality, found the proposed method to be highly effective at removing noise. Comparison against conventional speech enhancement methods found performance to be equivalent to Wiener filtering.

Index Terms: speech enhancement, MAP, sinusoidal model

1. Introduction

Speech enhancement is typically implemented as a two stage filtering process of first noise estimation and then noise removal. Numerous methods have been proposed for speech enhancement and can be broadly categorised into spectral subtraction, Wiener filtering, statistical methods and subspace methods [1]. These can provide good improvements in speech quality but suffer from effects such as residual noise, musical noise and speech distortion. The aim of this work is to move away from these traditional filtering-based enhancement methods and instead reconstruct clean speech from a set of clean speech features that have been estimated from noisy speech. The motivation behind this approach is that given a sufficiently good speech model and a noise-free set of features, then the reconstructed speech should itself be free from noise and distortion.

This leads on to two main challenges: i) finding a sufficiently good speech model from which to reconstruct speech and ii) establishing a set of speech features that can be estimated robustly from noisy speech. Clearly the two challenges are linked as the speech feature set must contain sufficient information for the model to synthesise speech.

Many different speech models have been proposed and these typically have application in speech coding and synthesis. The more successful and flexible models are based on the source-filter model and include the vocoder, sinusoidal model and STRAIGHT [2, 3, 4]. An evaluation of these is made in Section 2 to find the best synthesis method for enhancement. These models are typically driven by a combination of excitation and vocal tract features such as fundamental frequency, voicing and spectral envelope. Robust estimation of fundamental frequency and voicing has been the subject of much research with a comparison of methods given in [5], and as such this work will use the method proposed by the ETSI Aurora standard [2]. To obtain an estimate of the clean spectral envelope from noisy speech a maximum *a posteriori* (MAP) approach is proposed and this is discussed in Section 3.2.

The optimal configuration of the proposed reconstruction-based speech enhancement system is determined through a series of objective tests that are presented in Section 5. These are followed by further analysis using subjective listening tests that also compare the proposed enhancement method with three conventional filtering-based methods, namely spectral subtraction, Wiener filtering and log MMSE.

2. Speech reconstruction model

The method of speech enhancement proposed in this work first estimates clean speech features from noisy speech and then applies them to a speech model for reconstruction. The model must therefore transform the speech features into a time-domain signal with as little distortion as possible. Models based on the source-filter model have been effective at this and are found in many different speech processing applications such as speech coding, voice conversion and HMM synthesis [3, 4].

The most simple source-filter model is the vocoder which is typically used for low bandwidth applications and as such is limited in terms of quality. For the proposed speech enhancement application compression is not an issue and although improved representation of the excitation component is possible, the overall speech quality is not sufficient. The sinusoidal model is an improvement over the vocoder and represents speech as a combination of sinusoids which model the speech harmonics. This offers good flexibility in terms of adjustment of the parameters driving the model and has found widespread use in a range of speech processing applications. A further speech model is STRAIGHT which was designed for speech manipulation and has also found application in HMM-based synthesis [3]. A useful comparison of these speech models was made in [4], which established through a set of listening tests that the sinusoidal model attained best quality. Based on this finding, the sinusoidal model has been chosen for this work although other speech models could equally well be applied.

2.1. Sinusoidal model implementation

A speech signal, $s(n)$, is synthesised from the sinusoidal model as a summation of L sinusoids with frequencies, f_l, amplitudes, a_l and phases, θ_l,

$$s(n) = \sum_{l=1}^{L} a_l \cos(2\pi f_l n + \theta_l) \tag{1}$$

Each synthesised frame is assumed to be either voiced, unvoiced or of mixed voicing. For voiced frames the sinusoid frequencies are assumed to have a harmonic relationship to the fundamental frequency, f_0, i.e. $f_l = l f_0$. Sinusoid amplitudes are computed by sampling the speech spectral envelope estimate, $A(f)$, at harmonic frequencies, i.e. $a_l = A(f_l)$. Unvoiced frames are synthesised using a dense sampling of the spectral envelope, while mixed voicing frames are synthesised as voiced up to 1.2kHz and unvoiced at higher frequencies [2].

The quality of the sinusoidal model is improved by harmonic tracking whereby each frame is divided into four subframes and the fundamental frequency interpolated across each subframe. This improved harmonic tracking reduces the inter-frame frequency difference between harmonics and removed a slight buzzyness in the synthesised speech.

Equation (1) shows that the sinusoidal model requires robust estimates of the spectral envelope, $A(f)$, the fundamental frequency, f_0, and voicing, and the phase, θ_l. The next two sections discuss estimation of these features from noisy speech.

3. Spectral envelope estimation

This section describes the proposed method of estimating a clean speech spectral envelope from noisy speech. This operates as a three stage process of first extracting features from the speech signal, secondly removing the effect of noise and finally transforming them into a spectral envelope representation.

3.1. Speech features

A wide range of both parametric and nonparametric speech features have been proposed that retain spectral envelope information and have found application across a wide range of speech processing applications. This work considers three of the more common features, namely linear predictive coding (LPC) coefficients, line spectral frequencies (LSF) and mel-frequency cepstral coefficients (MFCC) features [1].

In LPC the spectral envelope is represented as an all-pole filter. Extraction of the filter coefficients is straightforward but accuracy degrades substantially in noisy conditions, leading to poor representation of formant positions. LSFs are derived from LPC coefficients and generally provide a more stable representation of formant positions and bandwidths. MFCCs provide a compact nonlinear frequency representation of the spectral envelope with more detail given to low frequency regions. In deciding which speech feature to use in this work a series of preliminary tests were carried out to measure the accuracy of spectral envelope estimation from these different features. MFCCs were found to offer the most accurate estimates and have been selected as the speech feature in this work, although the proposed method of speech enhancement could be applied to other features.

The MFCC vectors were extracted according to the ETSI Aurora standard [2] but with the number of filterbank channels left as a parameter to optimise. For all configurations the number of MFCCs was made equal to the number of filterbank channels, thereby not introducing any smoothing. Spectral envelope estimates, $\hat{A}(f)$, were obtained from the MFCC vectors by applying an inverse discrete cosine transform, exponential operation and then cubic spline interpolation.

3.2. MAP estimation of clean spectral envelope

Following MFCC extraction it is likely that the resulting feature is contaminated by noise. To remove the effects of noise a MAP estimate of the clean MFCC vector is made from the noisy MFCC vector. This approach is similar to SPLICE, [6], which makes clean feature estimates from noisy features for a robust speech recognition system. A model is first made of the joint density of the clean and noisy MFCC vectors using a Gaussian mixture model (GMM). Training begins by first creating a joint feature vector, $\mathbf{z}_i$

$$\mathbf{z}_i = [\mathbf{c}_i, \mathbf{n}_i] \tag{2}$$

where $\mathbf{c}_i$ and $\mathbf{n}_i$ are clean and noisy MFCC vectors representing frame i. From a set of training vectors expectation-maximisation (EM) clustering is applied to create a GMM, Φ, to model the joint density

$$\Phi(\mathbf{z}_i) = \sum_{k=1}^{K} \alpha_k \phi_k(\mathbf{z}_i) = \sum_{k=1}^{K} \alpha_k N(\mathbf{z}_i, \mu_k, \mathbf{\Sigma}_k) \tag{3}$$

The GMM comprises a set of K Gaussian probability density functions (PDFs), ϕ_k, that localise the joint density of the clean and noisy MFCC vectors. α_k represents the prior probability of the kth cluster and μ_k and $\mathbf{\Sigma}_k$ represent the mean and covariance of the joint vector within the kth Gaussian distribution

$$\mu_k^z = \begin{bmatrix} \mu_k^c \\ \mu_k^n \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma}_k^z = \begin{bmatrix} \mathbf{\Sigma}_k^{cc} & \mathbf{\Sigma}_k^{cn} \\ \mathbf{\Sigma}_k^{nc} & \mathbf{\Sigma}_k^{nn} \end{bmatrix} \tag{4}$$

The mean vector comprises clean and noisy MFCC mean vectors, μ_k^c and μ_k^n. The covariance matrix consists of clean and noisy covariance matrices, $\mathbf{\Sigma}_k^{cc}$ and $\mathbf{\Sigma}_k^{nn}$, and cross-covariances of the clean and noisy MFCCs, $\mathbf{\Sigma}_k^{cn}$ and $\mathbf{\Sigma}_k^{nc}$.

A MAP estimate of the clean MFCC vector can be made from the noisy MFCC vector and their joint density. For the kth cluster in the GMM, ϕ_k, the MAP estimate of the clean MFCC vector, $\hat{\mathbf{c}}_i^k$, from the noisy MFCC vector, $\mathbf{n}_i$ is given

$$\hat{\mathbf{c}}_i^k = \arg\max_{\mathbf{c}_i}(\Pr(\mathbf{c}_i \mid \mathbf{n}_i, \phi_k)) \tag{5}$$

The estimates from each cluster in joint density can be combined by weighting by the posterior probability of the noisy MFCC vector belonging to the kth cluster, to give a weighted estimate of the clean MFCC vector

$$\hat{\mathbf{c}}_i = \sum_{k=1}^{K} h_k(\mathbf{n}_i) \arg\max_{\mathbf{c}_i}(\Pr(\mathbf{c}_i \mid \mathbf{n}_i, \phi_k)) \tag{6}$$

$h_k(\mathbf{n}_i)$ represents the posterior probability of the noisy vector and is defined

$$h_k(\mathbf{n}_i) = \frac{\alpha_k \Pr(\mathbf{n}_i \mid \Phi_k)}{\sum_{k=1}^{K} \alpha_k \Pr(\mathbf{n}_i \mid \Phi_k)} \tag{7}$$

$\Pr(\mathbf{n}_i \mid \Phi_k)$ is the marginalised distribution of the noisy MFCC vector. Finally, the estimate of the clean speech MFCC vector can be calculated

$$\hat{\mathbf{c}}_i = \sum_{k=1}^{K} h_k(\mathbf{n_i}) \Big(\mu_k^c + \mathbf{\Sigma}_k^{cn} (\mathbf{\Sigma}_k^{nn})^{-1} (\mathbf{n}_i - \mu_k^n) \Big) \tag{8}$$

Experiments in Section 5 determine the optimal parameters for both the MFCC feature extraction and MAP estimation.

4. Fundamental frequency, voicing and phase estimation

Of the speech features needed for reconstruction, this work has focussed on clean speech spectral envelope estimation. As such, for fundamental frequency and voicing estimation the method specified in the ETSI Aurora standard is used as this has been shown to offer good robustness to noise [2]. The phase estimates required for the sinusoidal model use the phase of the noisy speech. This has been shown to be the optimal estimate of phase down to SNRs of about 8dB [1]. As we are not presently considering application of our method in very noisy conditions we have not attempted to further process the phase.

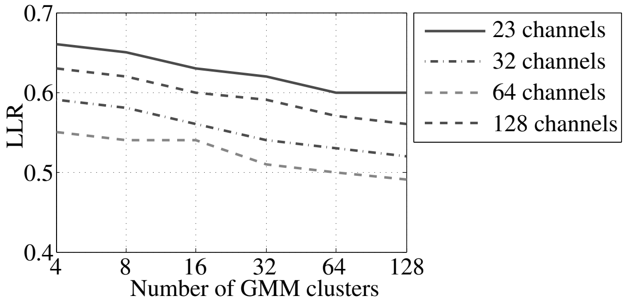

Figure 1: *Comparison of LLR scores for differing filterbank sizes and numbers of GMM clusters at an SNR of 10dB.*

5. Experimental results

This section first describes a set of objective tests that determine the final configuration of the proposed speech enhancement system. Secondly, subjective listening test results are presented that investigate the effectiveness of the reconstruction-based speech enhancement and compare this to conventional filtering-based methods. The speech for the experiments was taken from a single female US talker at a sampling rate of 8kHz, with 586 utterances used for training and a separate set of 246 for testing. Car noise from the AURORA database was used to create the noisy speech.

5.1. Optimisation of speech enhancement configuration

The number of clusters in the GMM modelling the joint density used in MAP estimation, and the number of channels in the filterbank, are now determined.

5.1.1. GMM parameter selection

MAP estimation of clean MFCC vectors is reliant on the accuracy of the joint density of clean and noisy MFCC vectors which is modelled by a GMM. To establish the optimal number of clusters in the GMM, speech has been reconstructed using the proposed enhancement system and the log likelihood ratio (LLR) of the enhanced speech computed to give a measure of speech distortion. This has been performed at SNRs of 20dB, 10dB, 5dB and 0dB and also for filterbank sizes of 23, 32, 64 and 128 channels. For illustration, Figure 1 shows the LLR at an SNR of 10dB with the number of GMM clusters varied from 4 to 128. In general, for all filterbank sizes, the LLR reduces as the number of clusters is increased. Beyond 64 clusters the level of improvement reduces and as such K=64 clusters are used in the GMMs. This result was also observed at the other SNRs.

5.1.2. Filterbank size

The second parameter to be optimised is the number of channels in the filterbank used in MFCC feature extraction. This is analysed by calculating the RMS error between clean filterbank vectors and MAP estimated filterbank vectors across a range of different filterbank sizes. Table 1 shows the RMS filterbank estimation error for filterbank sizes of 23, 32, 64 and 128 and at SNRs of 20dB, 10dB, 5dB and 0dB, averaged across all test utterances. Estimation error is seen to reduce as filterbank size increases from 23 to 64 channels and then increases with 128 channels, making 64 the optimal number of channels.

A second objective test was also carried out to investigate the effect of filterbank size. In this experiment the effect of filterbank size on the enhanced speech was examined using the LLR distortion measure. Table 2 shows LLR scores using the same filterbank sizes and SNR configurations as previously. An additional column also shows LLR scores for speech reconstructed from features estimated from clean speech. For comparison purposes, the final row in the table shows LLR scores for the original unprocessed noisy speech. The results follow closely the RMS filterbank error results with minimum distortion again attained using 64 channels in the filterbank. As such, a filterbank size of 64 channels is chosen for the remaining work. Comparing the enhanced speech to the unprocessed noisy speech shows the reconstruction to give a significant reduction in LLR disortion.

Table 1: *RMS filterbank estimation error for filterbank sizes from 23 to 128 channels in SNRs from 20dB to 0dB.*

	20dB	10dB	5dB	0dB
23	7.69	11.18	13.17	15.24
32	6.48	9.03	10.97	13.27
64	6.87	8.98	10.63	12.61
128	8.85	11.47	13.38	15.57

Table 2: *LLR scores of enhanced speech for filterbank sizes from 23 to 128 channels in SNRs from 20dB to 0dB and clean speech.*

	Clean	20dB	10dB	5dB	0dB
23 channels	0.47	0.43	0.60	0.73	0.88
32 channels	0.44	0.37	0.53	0.65	0.80
64 channels	0.40	0.34	0.50	0.62	0.77
128 channels	0.33	0.38	0.57	0.72	0.87
Unprocessed	0.00	0.71	1.12	1.34	1.52

5.2. Subjective quality measurement

This section presents the results of subjective tests that compare the speech quality of the proposed speech enhancement method to that of a range of conventional methods. A three-point mean opinion score (MOS) was used for the tests with listeners asked to rate utterances on signal quality, background noise intrusiveness and overall quality, each on a five point scale. The tests were performed in a sound-proof room with headphones, in accordance to the ITU-T P.835 recommendations.

Seven different speech enhancement configurations were investigated together with the original, unprocessed audio (NNC). Four systems were based on the sinusoidal model generated speech. Methods SIN($f0$, A^N) and SIN($\hat{f}0$, A^N) examine the effect of using the either the reference, $f0$, or estimated, $\hat{f}0$, fundamental frequency and voicing. Both methods use noisy, unprocessed filterbank amplitudes, A^N. Methods SIN($f0$, $\hat{A}$) and SIN($\hat{f}0$, $\hat{A}$) now use MAP estimates of clean filterbank amplitudes, $\hat{A}$. Method SIN($\hat{f}0$, $\hat{A}$) represents the final enhancement system, with the other SIN(.) methods included for analysis purposes. For comparison, the conventional speech enhancement methods of spectral subtraction (SS), Wiener filtering (WF) and log MMSE are also included in the tests. Listening tests were carried out on clean speech and at SNRs of 20dB, 10dB and 5dB. Twenty listeners took part in the tests with a random selection of 32 speech utterances taken from the testing set, played in a random order. Results from these tests are displayed in Figure 2.

Considering first the overall quality of the clean speech, a

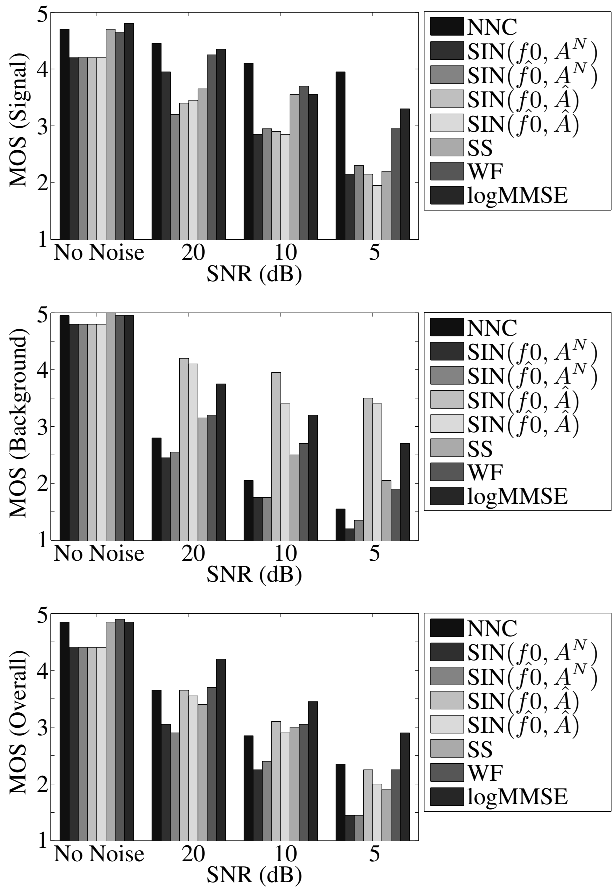

Figure 2: *MOS results showing a) signal quality, b) background noise intrusiveness and c) overall quality for different speech enhancement methods at varying SNRs.*

slight reduction is observed for speech reconstructed from the sinusoidal model. This is attributed to the reconstruction introducing a slight distortion and is consistent with results in [4]. Very little difference is observed between speech reconstruction using either the reference or estimated fundamental frequency and voicing. This is attributed to the robustness of estimation which has a fundamental frequency error of just 5% at an SNR of 5dB. Examining now the effect of MAP estimation of clean filterbank channels, a substantial improvement is observed in the reduction of background noise. Reconstruction using the sinusoidal model is highly effective at removing inter-harmonic noise. In addition, when using the clean spectral envelope estimate, noise added to the harmonics energies is also attenuated resulting in an almost noise-free signal. This is illustrated in Figure 3 which shows spectrograms of clean, noisy (10dB SNR) and enhanced speech. The speech enhanced by the sinusoidal model retains almost none of the original noise. However, wideband spectral differences in comparison to the clean speech do arise from reconstruction and contribute to the lower speech signal quality. Comparing the sinusoidal model to conventional speech enhancement methods shows it to be better at reducing noise but worse in terms of speech quality. Overall, reconstruction-based synthesis is comparable to Wiener filtering, outperforming spectral subtraction, but not as good as log MMSE.

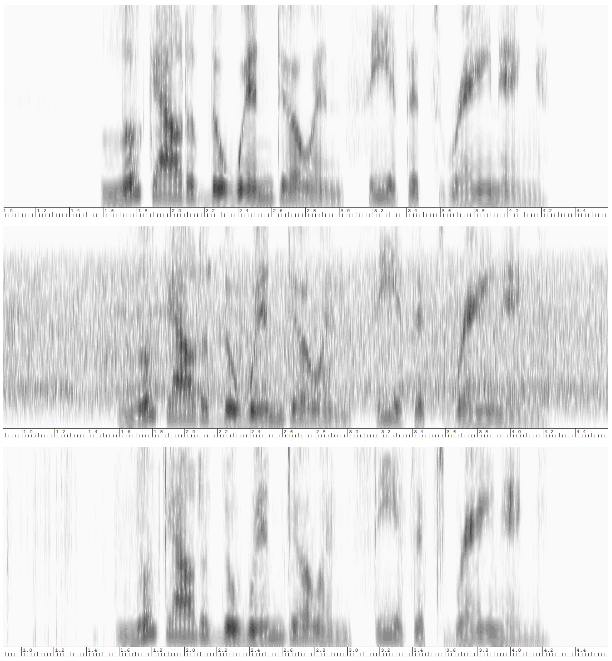

Figure 3: *Wideband spectrograms of utterance "Look out of the window and see if it's raining" for a) clean speech, b) noisy speech at an SNR of 10dB and c) enhanced*

6. Conclusion

This work has shown that speech enhancement can be achieved by reconstructing speech from a set of clean speech features extracted from noisy speech. Listening tests have established that such reconstruction-based synthesis is highly effective at removing noise but does introduce some distortion to the speech. The work at present uses a sinusoidal model as the basis for speech reconstruction but analysis in clean speech conditions shows that this does reduce speech quality. Comparison against three conventional methods showed overall quality to be comparable to Wiener filtering but as yet not as effective as log MMSE enhancement. Further work will examine improved speech models with the aim of reducing distortion.

7. References

[1] P. Loizou, *Speech Enhancement: Theory and Practice (Signal Processing and Communications).* CRC, 2007.

[2] A. Sorin and T. Ramabadran, "Extended advanced front end (XAFE) algorithm description, Version 1.1," ETSI STQ-Aurora DSR Working Group, Tech. Rep., 2003.

[3] H. Kawahara, I. Masuda-Katsuse, and 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*, vol. 27, no. 3, pp. 187–208, 1999.

[4] D. Chazan, R. Hoory, A. Sagi, S. Shechtman, A. Sorin, Z. Shuang, and R. Bakis, "High quality sinusoidal modeling of wideband speech for the purposes of speech synthesis and modification," in *ICASSP*, 2006.

[5] J. Darch and B. Milner, "A comparison of estimated and MAP-predicted formants and fundamental frequencies with a speech reconstruction application," in *ICSLP*, Aug. 2007, pp. 542–545.

[6] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *ICSLP*, vol. 3, 2000, pp. 806–809.