Enhancing Speech by Reconstruction from Robust Acoustic Features

Philip Harding and Ben Milner

University of East Anglia, Norwich, UK

Abstract

A method of speech enhancement is developed that reconstructs clean speech from a set of acoustic features using a sinusoidal model of speech. This is a significant departure from traditional filtering-based methods of speech enhancement. A major challenge with this approach is to estimate accurately the acoustic features (voicing, fundamental frequency, spectral envelope) from noisy speech. This is achieved using maximum a-posteriori estimation methods that operate on the noisy speech. Objective results are presented to optimise the proposed system and a set of subjective tests compare the approach with traditional enhancement methods.

Index Terms: speech enhancement, MAP, sinusoidal model

1. Introduction

Most speech enhancement methods operate by first making an estimate of the contaminating noise and then filtering this from the noisy signal to leave an estimate of the clean speech. Many methods have been proposed for speech enhancement and can be broadly grouped as spectral subtraction, Wiener filtering, statistical methods and sub-space methods [1]. Evaluations of these methods show them capable of improving speech quality but often at the expense of introducing unwanted artifacts such as musical noise, residual noise and distortion.

This work addresses these issues by proposing an alternative method of speech enhancement whereby a speech model is used to reconstruct clean speech from a set of acoustic speech features that have been estimated from the noisy speech. Given a suitably accurate speech model and noise-free set of features the reconstructed speech should be free from both noise and distortion. Two challenges arise from this approach – i) to find a sufficiently good speech model and ii) to develop methods to robustly estimate noise-free speech features.

The idea of using a speech model for speech enhancement is not a new one, with previous approaches using traditional methods such as Wiener filtering to clean the spectral envelope before a speech model, typically the sinusoidal model, is used to reconstruct the signal [2]. Our method improves on these methods by using a method similar to the SPLICE feature enhancement process [3] to robustly estimate noise-free speech features. Of the various speech models proposed, our initial work used a sinusoidal model [4]. This is now improved by including harmonic tracking and formant enhancement to the model, and is discussed in Section 2. To reconstruct speech the sinusoidal model requires a set of acoustic features that include voicing, fundamental frequency and spectral envelope. For speech enhancement these acoustic features need to be estimated from noisy speech which in itself is problematic. We now extend our previous work [4] to develop better methods of estimating these acoustic features from noisy speech, with the result that the quality of the enhanced speech is improved. Class-based methods of estimating spectral envelope features are developed whilst a maximum a-posteriori (MAP) estimation of fundamental frequency and voicing is applied. These are discussed in Sections 3, 4 and 5, before experimental results are presented in Section 6.

2. Speech reconstruction model

The estimate of clean speech, \( \hat{s}(n) \), is reconstructed from a set of acoustic speech features within the framework of a sinusoidal model. Speech is synthesised from a summation of \( L \) sinusoids with frequencies, \( f_l \), amplitudes, \( a_l \), and phases, \( \theta_l \), on a frame-by-frame basis, with overlap-and-add used to join frames,

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

Frames of speech are labelled as voiced or unvoiced. Voiced frames assume the sinusoids have a harmonic relationship to fundamental frequency \( f_0 \), i.e. \( f_l = lf_0 \). Peak-picking to determine amplitudes can be problematic in noisy speech [2] and so sinusoid amplitudes are given by sampling the spectral envelope, \( A(f_l) \), at harmonic frequencies, i.e. \( a_l = A(lf_0) \). Unvoiced frames are synthesised by exciting an all-pole filter, constructed using the spectral envelope, with white noise.

Speech quality is improved using harmonic tracking whereby each frame is divided into four subframes and fundamental frequency interpolated across each subframe. This reduces inter-frame frequency differences between harmonics and removed a slight buzzyness in the synthesised speech. A further improvement in speech quality was achieved using a postfilter to emphasise speech formants and attenuate spectral valleys [5]. This removed a slight muffling sound that was caused by an averaging of formant structures resulting from spectral envelope estimation.

Thus, to reconstruct clean speech, robust estimates of the spectral envelope, voicing, fundamental frequency and phase are required. The following sections discuss robust estimation of these features from noisy speech.

3. Spectral envelope estimation

Estimating the clean spectral envelope from noisy speech is a three stage process. First, features are extracted from the noisy speech. Second, MAP estimation is applied to obtain clean features from the noisy features. Finally, the clean features are transformed into a spectral envelope.

3.1. Spectral envelope feature selection

Two features from which the spectral envelope can be estimated are considered, namely MFCCs and linear-frequency cepstral coefficients (LFCCs). Both features utilise an \( H \)-channel filterbank – mel spaced for MFCCs and linearly spaced for LFCCs. The mel-spacing of filterbank channels is effective in speech recognition although this is questionable for accurate spectral envelope
estimation, hence the consideration of LFCCs. Preliminary tests were carried out using the two features for reconstructing speech across a range of signal-to-noise ratios (SNRs) [4]. Applying a log-likelihood ratio (LLR) test to the enhanced speech showed LFCCs to outperform MFCCs by on average 11% across all SNRs. As a result the LFCC feature is used in this work.

3.2. MAP estimation of clean spectral envelope features

For clean spectral envelope estimation, features extracted from noisy speech must have the noise removed. This is achieved by making a MAP estimate of the clean feature from the noisy feature. Previous work used a single model of the joint density of clean and noisy features to achieve this [4]. This work extends the estimation to use a localised joint density. Many different partitions of the feature space exist, but for this work two partitions are considered – by phoneme class or by articulation class. The articulation classes used in this work are: affricate, approximant, fricative, nasal, plosive, vowel and non-speech.

To construct class-based models of the joint density of clean and noisy features a joint feature vector, \( z_i \), is first defined,

\[
z_i = [c_i, n_i]
\]

where \( c_i \) and \( n_i \) are clean and noisy feature vectors representing the \( i \)th frame of speech. Using class label data a set of training data vectors, \( Z \), is distributed into \( M \) vector pools, \( \Omega^m \),

\[
\Omega^m = \{ z_i \in Z : \text{class}(z_i) = m \}
\]

where \( \text{class}(z_i) \) is the class allocated to the \( i \)th feature vector. The localised joint density of clean and noisy features is modelled by applying expectation-maximisation clustering to each class pool. This gives a set of \( M \) GMMs, \( \Phi^m \), that model the joint density of clean and noisy features for each class, \( m \),

\[
P(z_i|m) = \Phi^m(z_i) = \sum_{k=1}^{K} \alpha^m_k N(z_i, \mu^m_k, \Sigma^m_k)
\]

Each GMM comprises \( K \) Gaussian probability density functions with means, \( \mu^m_k \), variances, \( \Sigma^m_k \), and prior probabilities, \( \alpha^m_k \).

A MAP estimate of the clean speech feature, \( e_i \), can now be made from a noisy speech feature and the class-specific GMM

\[
e_i = \arg \max_{e} \{ Pr(e | n_i, \Phi^m) \}
\]

The class label, \( m \), that specifies which GMM to use in the estimation is obtained from an HMM speech recogniser that classifies each noisy feature vector into one of \( M \) classes. Section 6 compares the performance of using phoneme classes (\( M = 41 \)), articulation classes (\( M = 7 \)) and a global model (\( M = 1 \)).

3.3. Inverting spectral envelope features

Estimation of the sinusoidal amplitudes in the sinusoidal model requires the speech features (LFCCs or MFCCs) to be transformed into a spectral envelope. This is achieved through an inverse discrete cosine transform and exponential operation before applying cubic spline interpolation to give a spectral envelope, \( A(f) \). For MFCC features linear filters the non-linear spacing and areas of the filterbank channels must be equalised, and is implemented as a subtraction in the MFCC-domain [6]. Recursive first-order averaging is applied to reduce inter-frame discontinuities:

\[
A(f)_i = \beta \cdot A(f)_{i-1} + (1-\beta) \cdot A(f)_{i-1}
\]

A value \( \beta = 0.85 \) was determined in preliminary testing.

4. Fundamental frequency and voicing

The sinusoidal model is reliant on accurate voicing and fundamental frequency and many methods are effective at giving accurate estimates of these parameters in clean conditions. However, as SNRs fall these methods become erroneous leading to voicing errors and inaccurate fundamental frequency estimates. These both degrade significantly the quality of reconstructed speech.

Our previous work used autocorrelation within the ETSI Aurora XFE tool [6] to estimate voicing and fundamental frequency but this was found to be inaccurate at low SNRs. In this work a MAP method of estimation is used which is significantly more robust at low SNRs than autocorrelation methods [7]. This operates in a similar manner to the spectral envelope estimation of Section 3.2, but instead models the joint density, \( \Psi \), of fundamental frequency and spectral feature. For accurate estimation the number of channels in the filterbank was set to \( H=64 \). Using this joint density, a MAP estimate of fundamental frequency, \( f_0 \), is made from a speech feature,

\[
f_0_i = \arg \max_{f_0} \{ Pr(f_0 | n_i, \Psi) \}
\]

5. Phase estimation

The final parameter required by the sinusoidal model is the phase of the sinusoids, which is set equal to the phase of the noisy speech. Noise distorts the phase spectrum of speech, although phase distortions up to \( \pi/4 \) have been shown to be perceptually inaudible whilst distortion beyond these limits is perceived as a “roughness” in quality [8]. Given an assumption that the audibility threshold of phase distortion is \( \pi/4 \), measurements of the phase error between the reconstructed speech and original speech found 94% of harmonics to be below the audible level at an SNR of 15dB. This decreases to 91% at an SNR of 5dB and 90% at 0dB. In more detail, the first harmonic remains robust to phase distortions whilst higher frequency harmonics are more affected by noise. For example, harmonics between 1500-2500Hz are distorted by a perceptually audible amount 55% of the time at 0dB SNR whilst the first harmonic is affected in only 4% of frames. This is attributed to the higher SNR at lower harmonics. This distortion causes some loss of detail in the reconstructed signal with noticeable spectral smearing present around affected harmonics which can cause the roughness in quality of voiced frames. Unvoiced frames are unaffected due to the random nature of the original phase. Given these investigations, the assumption of using the noisy phase to give sinusoid phases, whilst not ideal, gives an acceptable level of perceptual quality.

6. Experimental results

This section first shows objective tests that optimise the performance of components in the proposed enhancement method. Subjective listening tests are then presented that compare the proposed method with traditional filtering-based enhancement 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. Street noise from the Aurora database [9] was added to create the noisy speech. This was chosen because of its highly nonstationary character, making evaluations more realistic of real conditions.
6.1. Spectral envelope estimation

This section examines the performance of spectral envelope estimation using phoneme classes, articulation classes and a single global model. First the accuracy of the speech recogniser is measured on the phoneme and articulation classes and the results shown in Table 1. This shows articulation class classification to be more robust to noise, having only seven possible class labels compared to the 41 phonemes. An investigation is now made of the spectral envelope estimation accuracy made by the phoneme class, articulation class and global systems. Table 2 shows mean RMS spectral envelope estimation error when compared to the original clean features. To show the effect of frame classification accuracy in spectral envelope estimation, the RMS error of the phoneme and articulation class systems are shown first using reference labels (no classification errors) and then using the noisy HMM-based classifications (as shown in Table 1). Best performance is given by the phoneme class system using reference labels as this has the most accurate localisation of the feature space. When the HMM speech recogniser provides class information the errors it introduces cause a deterioration of spectral envelope estimation which increases above both the articulation class and global systems at 0dB. Despite the articulation class labels being more accurate than phoneme labels, the less detailed localisation of the feature space yields performance roughly equal to that of the phoneme-based system at 0dB and 5dB.

6.2. Fundamental frequency estimation

This section examines the accuracy of MAP fundamental frequency and voicing estimation across a range of SNRs and makes a comparison to the ETSI XFE method [6]. Three error measures are shown in Figure 1 – the percentage of voicing classification errors, the mean percentage fundamental frequency error and the percentage of voiced frames with an RMS error greater than 20% (gross errors). Reference data was provided by laryngograph (REF) and labels obtained from the HMM on noisy speech (HMM). Three sources of voicing/fundamental frequency were also tested – laryngograph (LAR), XFE and the proposed MAP method. This gave six configurations of this system to be evaluated, with, for example, the source SM, REF, LAR referring to the sinusoidal model-based method using reference class labels and f0 obtained from the laryngograph. For comparison the conventional enhancement methods of spectral subtraction (SS), Wiener filtering (WF) and log MMSE were also evaluated along with no noise compensation (NNC). The tests were carried out at SNRs of 15dB, 5dB and 0dB. Twenty listeners took part in the tests with speech from each method played twice (using different utterances) to give a total of 62 utterances, played in a random order. Results from these tests are displayed in Figure 2. The most obvious result is the effectiveness of the proposed method to remove noise. This is attributed to the sinusoidal mode reconstructing speech only and not introducing any noise. Results for signal quality show the proposed method to have introduced some distortion to the speech. This is attributed to distortions introduced by the spectral envelope estimation, as reported in Table 2. The overall quality of the proposed system is good and higher than the traditional methods, particularly at lower SNRs.

Comparing performance using either reference or HMM-based phoneme labels shows little difference at higher SNRs but a greater difference at low SNRs. This is explained by the deteriorating phoneme accuracy as SNRs fall as shown in Table 1 which increases the spectral envelope estimation error. Examining the effect of voicing/fundamental frequency estimation shows the more accurate MAP method to produce better enhanced speech in comparison to the poorer performing XFE method. This difference increases as SNRs reduce due to the large increase in errors by the XFE method.

In summary, all configurations using reference phoneme labels are found to outperform log MMSE. When using labels generated from noisy speech, performance degrades in high levels of noise, but remains at least comparable with log MMSE.
6.4. Spectrogram analysis

To illustrate the effectiveness of the proposed enhancement method, Figure 3 shows four spectrograms of the utterance ‘Look out of the window and see if it’s raining’ contaminated by street noise at an SNR of 5dB under different processing – (a) clean, (b) noisy, (c) enhanced using the SM_HMM_MAP configuration and (d) log MMSE. As seen in Figure 3(b), street noise is highly complex and contains a number of stationary and non-stationary noises from sources including passing cars, people talking and warning sirens. The warning sirens are visible on the spectrogram as tones in the first two-thirds of the utterance whilst the noise of a car accelerating can be seen as low frequency noise in the latter third. Log MMSE is successful at removing the stationary components of the noise, although it is clear that almost none of the non-stationary noise has been removed. In contrast the proposed method is almost noise-free with none of the stationary or non-stationary noise reconstructed by the model.

7. Conclusion

This work has proposed a speech enhancement method that operates by reconstructing clean speech from a set of robust acoustic features extracted from noisy speech. Listening tests have shown the method to be significantly better than traditional filtering methods at removing noise, comparable in terms of signal quality and better in terms of overall quality. The success of the method relies on obtaining robust estimates of acoustic speech features. Whilst the proposed methods are able to make reasonably good estimates of these features further work must concentrate on providing more reliable estimates.

8. References