CLEAN SPEECH RECONSTRUCTION FROM NOISY MEL-FREQUENCY CEPSTRAL COEFFICIENTS USING A SINUSOIDAL MODEL

Xu Shao and Ben Milner

School of Information Systems, University of East Anglia, Norwich, UK

x.shao@uea.ac.uk , bpm@sys.uea.ac.uk

ABSTRACT

This paper extends the technique of speech reconstruction from MFCCs by considering the effect of noisy speech. To reconstruct a clean speech signal from noise contaminated MFCCs an estimate of the clean mel-filterbank vector is required together with a robust estimate of the pitch. This work applies spectral subtraction to the mel-filterbank vector (derived from noisy MFCCs) to provide a clean speech spectral estimate. To obtain a reliable estimate of pitch a robust extraction technique is used.

Spectrograms and informal listening tests reveal that a clean speech signal can be successfully reconstructed from the noisy MFCCs. Pitch errors are shown to manifest themselves as artificial sounding bursts in the reconstructed speech signal. Incorrect estimates of the spectral envelope introduce periods of noise into the reconstructed speech.

1. INTRODUCTION

The increasing deployment of mobile devices in combination with advances in speech recognition technology has resulted in a substantial increase in the number of automated speech-based services being made available. To a large extent the success of these automated services relies on their ability to perform robust speech recognition from mobile devices in a range of environmental conditions.

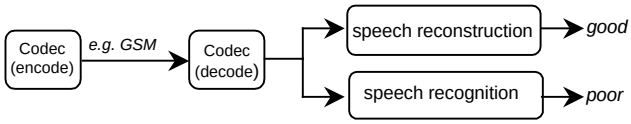

Figure 1-a: Codec-based speech communication.

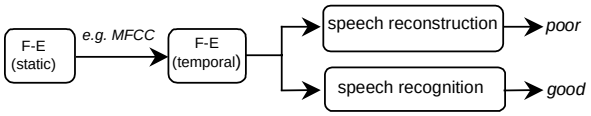

Figure 1-b: Distributed speech recognition communication

Speech communication from mobile devices has traditionally been made using low bit-rate speech codecs as illustrated in figure 1-a. The low bit-rates at which these codecs operate introduces a slight distortion onto the speech signal which becomes more severe in noisy conditions. When input into a speech recogniser this distortion causes a noticeable reduction in accuracy. The technique of distributed speech recognition (DSR) [1] has been proposed by the ETSI Aurora group to overcome this problem - as illustrated in figure 1-b. DSR replaces the codec on the terminal device with the front-end processing component of the speech recogniser which thereby removes codec-based distortion from the speech recogniser input. This results in a significant improvement in speech recognition accuracy. However, because speech feature vectors are designed to be a compact representation, for discriminating between speech sounds, they do not contain sufficient information to enable reconstruction of the original speech signal. In particular valuable speaker information is lost, such as pitch. However, several schemes have recently been proposed for reconstructing speech from MFCC vectors and pitch [2,3].

This work builds on techniques for speech reconstruction, but now considers reconstructing a clean speech signal from noise contaminated MFCCs. Speech reconstruction can be considered as requiring both speech excitation (source) and vocal tract (filter) information. Therefore to achieve clean speech reconstruction the noise contaminated MFCC-derived spectral envelope must be enhanced and a reliable pitch estimate made from the noisy speech signal. Section 2 of this paper reviews the sinusoidal model of speech and introduces a pitch smoothing which improves speech quality. Section 3 describes the use of spectral subtraction for obtaining a clean spectral envelope from noise contaminated MFCCs. A method for reliably extracting pitch from noisy speech is described in section 4. Results of the clean speech reconstruction are presented in section 5 and a conclusion is made in section 6.

2. SPEECH RECONSTRUCTION

This section reviews the MFCC extraction process and shows how it can be integrated into a sinusoidal model of speech.

2.1. MFCC-based Feature Extraction

The stages involved in creating a stream of MFCC vectors are illustrated in figure 2 - based on the ETSI Aurora standard [1].

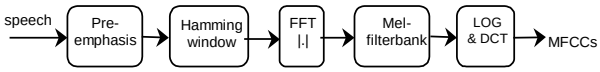

Figure 2: MFCC feature extraction procedure

A number of the transformation stages are invertible - such as pre-emphasis, the Hamming window and logarithm operations. Other stages discard information which makes them non-invertible. The overlapping triangular filters of the mel-filterbank essentially extract a frequency warped spectral envelope from the magnitude spectrum. This loses finer detail of the magnitude spectrum which includes speech excitation information in the form of the pitch harmonics. Similarly taking the magnitude of the complex spectrum loses phase information and the truncation of the DCT smoothes the log filterbank. Too much information

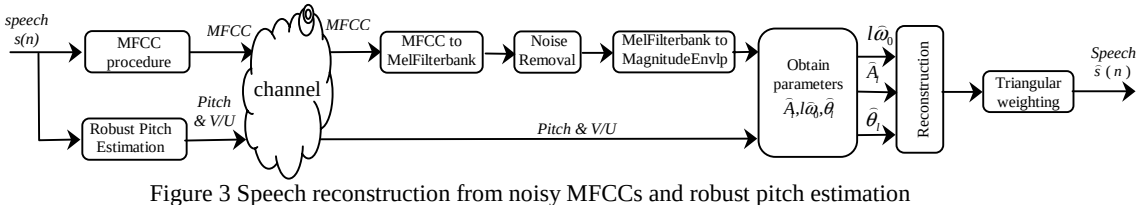

Figure 3 Speech reconstruction from noisy MFCCs and robust pitch estimation

is lost to enable the MFCC vectors to be inverted into a time-domain signal through a reversal of the procedures in figure 2.

However, it is possible to recover a smoothed estimate of the magnitude spectrum from the MFCCs. Combining this with an estimate of the pitch enables either a sinusoidal model or a source-filter model to synthesis the original speech signal [2,3]. The proposed scheme for reconstruction clean speech from noisy speech is illustrated in figure 3.

2.2. Sinusoidal Model of Speech for Reconstruction

The sinusoidal model [4] synthesises a speech signal, *x(n)*, from a summation of a number of sinusoids of varying amplitude, A_l, frequency, ω_l, and phase, θ_l,

$$x(n) = \sum_{l=1}^{L(t)} A_l \cos(\omega_l n + \theta_l) \quad (1)$$

If all the parameters of the model are accurately identified then a good reproduction of the original signal can be synthesised. Reconstructing speech from MFCCs and a pitch estimate requires several simplifications [9] of the model, but a good reproduction of the original speech can still be recovered.

As described in [2] an estimate of the spectral envelope can be obtained from MFCC vectors through zero padding and an inverse DCT operation to give a log mel-filterbank estimate. Applying an exponential operation to this and then interpolation results in a smoothed magnitude spectral estimate, $\left|\hat{X}(\omega)\right|$.

Normalisation is needed to remove the effect of pre-emphasis and the non-linear filterbank channel bandwidths. This can be implemented in either the cepstral domain (as subtraction) or in the frequency domain (as filtering) [3].

A simplification of the sinusoidal model assumes that the frequencies of the sinusoid components, ω_l, are harmonics of the pitch frequency, ω_0, i.e.

$$\omega_l = l\,\omega_0 \quad (2)$$

Therefore from an estimate of the pitch, ω_0, the frequencies of all the sinusoids can be determined. Their amplitudes can be computed from the value of the smoothed magnitude spectrum at the particular harmonic frequency,

$$A_l = \left|\hat{X}(l\,\omega_0)\right| \quad (3)$$

The phase offset, θ_l, is calculated from two components [9]. One component comes from the speech excitation (source) and can be estimated using a linear phase model. The second component comes from the vocal tract and can be estimated using a Hilbert transform based on a minimum phase assumption [9].

Therefore, for each MFCC vector and pitch estimate, a frame of reconstructed speech can be generated. Frames are merged together through the use of overlapping triangular windows.

2.3. Pitch Smoothing

Listening tests and examination of spectrograms identified a buzzing-type noise to be present in the reconstructed speech. It was found that this results from the small pitch changes which occur between successive frames. At mid and higher frequency harmonics these pitch changes cause larger frequency differences and result in confusion of harmonic tracks as shown in figure 4a.

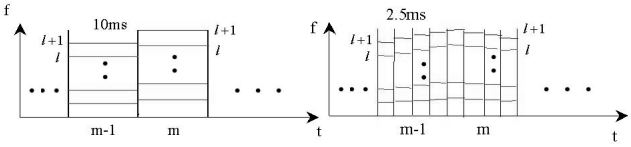

Figure 4: Harmonic confusion a) original model; b) Sub-frame

To reduce the severity of inter-frame pitch changes each frame is divided into a number of sub-frames as shown as figure 4-b. Pitch in each sub-frame is obtained from linear interpolation between adjacent frames. The effect of this is to smooth the pitch and hence reduce the frequency shift that occurs between frames.

3 SPECTRAL ESTIMATION FROM NOISY MFCCs

To achieve clean speech reconstruction it is necessary to obtain an estimate of the clean speech magnitude spectrum from the noisy magnitude spectrum, extracted from the noisy MFCC vectors. Many techniques have been proposed for achieving this with the most successful being spectral subtraction and Wiener filtering [5]. Currently, spectral subtraction is used to obtain a clean speech magnitude spectral estimate, i.e.

$$\left|\hat{X}_t(f)\right| = \begin{cases} \left|Y_t(f)\right| - \alpha\left|\hat{N}_t(f)\right| & \left|\hat{X}_t(f)\right| > \beta\left|Y_t(f)\right| \\ \beta\left|Y_t(f)\right| & \textit{otherwise} \end{cases} \quad (4)$$

where $\left|Y_t(f)\right|$, $\left|\hat{N}_t(f)\right|$ and $\left|\hat{X}_t(f)\right|$ represent the magnitude spectra of the noisy speech, the noise and the clean speech estimate. The variables α and β are the over-subtraction factor and maximum attenuation of the filter, respectively.

Spectral subtraction is known to suffer from processing distortions which occur when spectral magnitudes reach a spectral floor. This results in certain frequencies being turned on and off and causes the so called "musical noise". In this work such processing distortions are minimised by implementing the subtraction in the mel-filterbank domain, rather than the magnitude spectral domain. The averaging of the magnitude spectrum made by the triangular windows of the mel-filterbank means that channel estimates are less likely to reach flooring values which would introduce distortion. For example, figure 5-a shows the mel-filterbank and magnitude spectrum of a clean speech signal. Figure 5-c shows the same signal but contaminated by noise (shown in figure 5-b). Figure 5-d shows

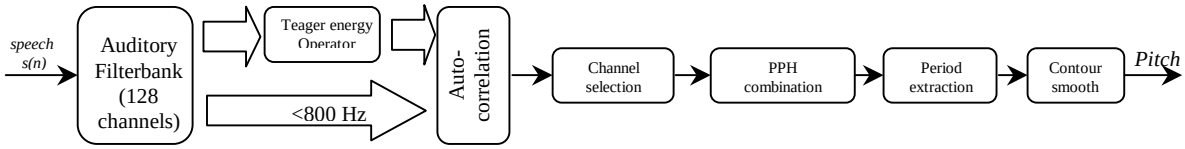

Figure 6: Robust pitch estimation from noisy speech

the result of subtracting the noise from the noisy signal to give a clean speech estimate.

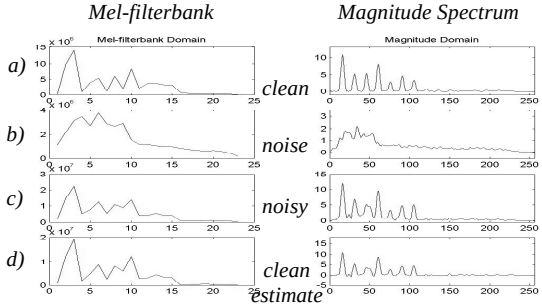

Figure5: Spectral subtraction – filterbank & magnitude domains

Comparing the mel-filterbank domain clean speech estimate with the original speech signal reveals a closer structure than is obtained in the magnitude spectrum.

4. ROBUST PITCH ESTIMATION

To accurately reconstruct the speech signal it is vital to have a reliable estimate of the pitch (for voiced sounds). Previous work [6] used a comb function to determine the pitch from the magnitude spectrum of the speech signal. This delivers good pitch estimates for clean speech but is less accurate when estimating pitch from noise contaminated speech. The pitch detection algorithm in this work is based on robust pitch estimation techniques [7,8].

Figure 6 illustrates the robust pitch estimation scheme. The noisy speech signal is split into 128 frequency channels by the auditory filterbank. A Teager energy operator extracts an energy envelope for mid and high frequency regions of the signal. A set of normalized auto-correlation functions, $R_i(\tau)$, are then obtained for each of the channels, i, at varying time lags, τ. Auto-correlation values from channels identified as being noisy are discarded, while channels corresponding to clean speech are summed together at the pseudo-periodic histogram (PPH) stage [7]. This produces a waveform which varies at the pitch period.

Several improvements have been made to the system for reliably extracting this pitch period. First, a low pass filter is used to remove fluctuations at the output of the PPH stage. A comb function is then applied to this signal to identify the pitch period. Depending on the fit of the comb function to the waveform, decisions can be made as to the level of voicing present.

A five-point median filter and a series of post-processing rules are used to smooth the resulting pitch contour. These rules are derived from observations of measured pitch contours. These rules include that voiced frames last for at least three frames and that voiced frames are unlikely to be interrupted by short bursts of unvoiced frames.

To illustrate the effectiveness of the robust pitch extraction algorithm, figure 7-a shows a signal contaminated by tonal noise.

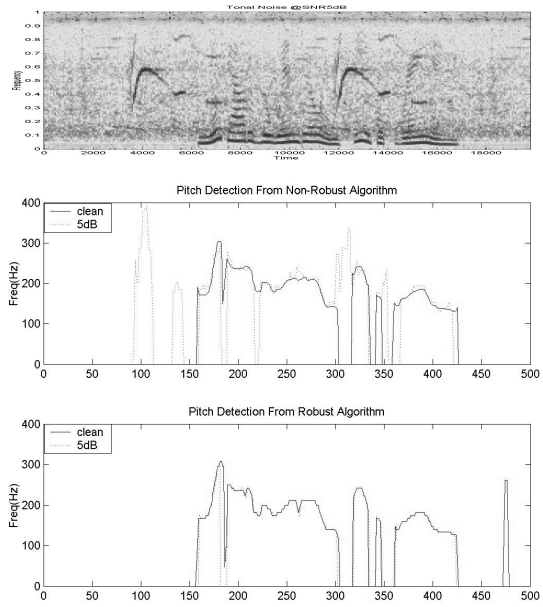

Figure 7: a) Noise contaminated speech signal; b) Pitch contour from comb function; c) Pitch contour from robust pitch estimator

Figure 7-b shows the pitch contour extracted using the comb function for both clean speech and speech contaminated at an SNR of 5dB. For clean speech the contour is accurate but for noisy speech the estimator introduces many pitch errors. Figure 7-c shows the pitch contours estimated by the robust pitch estimator. These are clearly less affected by noise and provide an essential robust estimate of the pitch.

5. EXPERIMENTAL RESULTS

To analyse the effectiveness of the clean speech reconstruction scheme a set of speech utterances based on Messiah sentences has been used. These have been sampled at 8kHz and also contain accurate pitch measurements taken from a laryngograph at the recording sessions. To simulate noisy speech, wideband noise from the ETSI Aurora database has been added artificially at varying signal to noise ratios (SNRs).

Figure 8-a shows the spectrogram of the sentence “Look out of the window and see if it’s raining” spoken by a female speaker. Figure 8-b shows the same sentence contaminated by wideband noise at an SNR of 10dB.

From this noisy speech a pitch contour is extracted using the robust pitch estimation technique described in section 4 and also a set of MFCC vectors as described in section 2.1. It is from these two sets of parameters that clean speech reconstruction

takes place. The overall quality of the reconstructed speech is dependent on the accuracy of pitch estimation and the effectiveness of spectral subtraction to remove the noise.

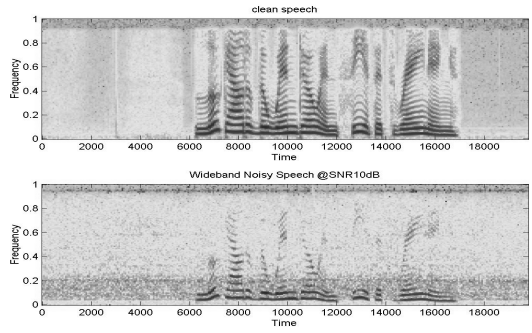

Figure 8: a) Original clean speech; b) Noisy speech - 10dB

Figure 9-a shows the spectrogram of the reconstructed speech signal using the robust pitch estimate (section 4) and MFCC vectors. No spectral subtraction has been employed at this stage. Figure 9-b shows the same signal but reconstructed from mel-filterbank vectors which have had an estimate of the noise subtracted from them.

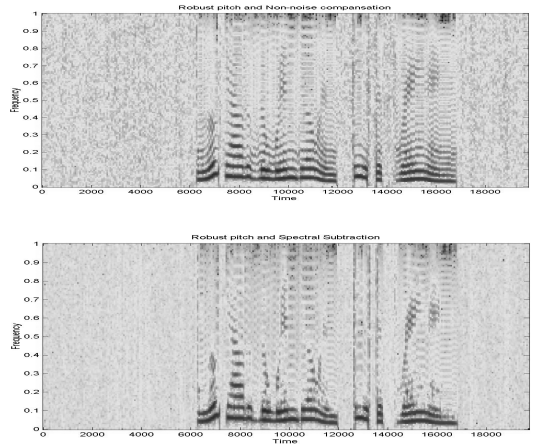

Figure 9: a) Reconstructed speech using robust pitch estimate; b) reconstruction using robust pitch and spectral subtraction.

Figure 9-a shows that robust pitch estimates have enabled the resulting pitch harmonics to be correctly positioned in comparison with the original signal in figure 8. The inversion of the MFCC vectors to a spectral envelope has produce a good reproduction of the original spectral envelope of the speech. The spectrogram in figure 9-b clearly shows that spectral subtraction has removed the wideband noise present in the speech signal shown in figure 8-b. Informal listening tests on a number of utterances confirm the effective removal of the noise.

As a comparison, figure 10 shows the spectrogram of the reconstructed speech signal using the measured pitch (taken from laryngograph signal) and spectrally subtracted estimates of the clean speech mel-filterbank vectors.

Again the spectrogram and listening tests show a relatively clean speech signal. It is interesting to observe how similar the pitch harmonics are between those derived from the robust pitch estimate and the measured pitch from the laryngograph.

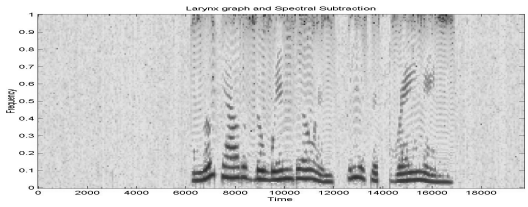

Figure 10: Reconstruction using laryngograph and subtraction.

These spectrograms, together with listening tests, demonstrate that both robust pitch estimation and noise removal from the spectral envelope are necessary conditions for clean speech reconstruction. Errors in pitch estimation manifest themselves as artificial sounding bursts in the reconstructed speech signal. Incorrect estimates of the spectral envelope are perceived as part of the contaminating noise remaining in the reconstructed speech. A downloadable result is available at http://www.uea.ac.uk/~a169838/

6. CONCLUSION AND DISCUSSION

This work has demonstrated that it is possible to reconstruct a clean speech signal from a series of noisy MFCC vectors. To achieve this both a robust pitch estimate and an estimate of the clean magnitude spectrum is necessary. Results have shown that performing spectral subtraction in the mel-filterbank domain enables a sufficiently good estimate of the clean speech magnitude spectrum to be derived for clean speech reconstruction. Using a robust pitch estimation method, together with some post-processing, gives a very close value to that obtained from a laryngograph across a range of SNRs.

7. REFERENCES

1. ESTI document - ES 201 108 – STQ: DSR – Front-end feature extraction algorithm; compression algorithm, 2000.
2. D. Chasan et al, “Speech reconstruction from mel frequency cepstral coefficients and pitch”, Proc. ICASSP, 2000.
3. B. P Milner and X. Shao, “Speech Reconstruction from MFCCs using a source-filter model”, Proc. ICSLP, 2002
4. R.J. McAulay and T.F. Quatiery, “Speech analysis/synthesis based on a sinusoidal representation”, IEEE Tran. ASSP, vol. 34, pp. 744-754, 1986.
5. S. V. Vaseghi and B. P. Milner, "Noise compensation methods for HMM speech recognition in adverse environment," IEEE Trans. SAP, Vol. 5, pp.11-21, 1997.
6. D. Chazan et al, "Efficient Periodicity Extraction Based on Sine-wave Representation and its Application to Pitch Determination of Speech Signals", Proc Eurospeech, 2001.
7. J. Rouat, Y. C. Liu and D. Morissette, "Pitch determination and voiced/unvoiced decision algorithm for noisy speech", Speech Communication Journal, pp. 191-207., 1997
8. M. Wu, D. L. Wang and G. J. Brown, "A multi-pitch tracking algorithm for noisy speech", Proc. ICASSP, 2002.
9. R. McAulay and T. Quatieri, “Sinsoidal Coding,” Ch. 4, Speech Coding and Synthesis, Elsevier, 1995.