SPEECH RECONSTRUCTION FROM MEL-FREQUENCY CEPSTRAL COEFFICIENTS USING A SOURCE-FILTER MODEL

Ben Milner and Xu Shao

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

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

ABSTRACT

This work presents a method of reconstructing a speech signal from a stream of MFCC vectors using a source-filter model of speech production. The MFCC vectors are used to provide an estimate of the vocal tract filter. This is achieved by inverting the MFCC vector back to a smoothed estimate of the magnitude spectrum. The Wiener-Khintchine theorem and linear predictive analysis transform this into an estimate of the vocal tract filter coefficients. The excitation signal is produced from a series of pitch pulses or white noise, depending on whether the speech is voiced or unvoiced. This pitch estimate forms an extra element of the feature vector.

Listening tests reveal that the reconstructed speech is intelligible and of similar quality to a system based on LPC analysis of the original speech. Spectrograms of the MFCC-derived speech and the real speech are included which confirm the similarity.

1. INTRODUCTION

Speech communication from mobile devices has traditionally been made through the use of low bit-rate speech codecs. The low bit-rates at which these codecs operate causes a slight distortion to be introduced onto the speech signal. When input into a speech recogniser this distortion causes a noticeable reduction in accuracy. To avoid this problem the technique of distributed speech recognition (DSR) [1] has been introduced. This involves replacing the codec on the terminal device with the front-end processing component of the speech recogniser.

The limited bandwidth of channels over which mobile devices operate prevents simultaneous transmission of both codec vectors and speech recognition feature vectors. Therefore for voice-to-voice communication codec vectors are transmitted and for speech recognition applications the feature vectors are transmitted – as controlled by a higher level protocol. Traditionally codec vectors have been designed to maximize speech quality while feature vectors are designed to maximize discrimination. It would be more useful to transmit a single vector which can be used for both speech coding and speech recognition. There are essentially two schemes which have been proposed to achieve this – those based upon low bit-rate speech codecs and those based on speech recognition feature vectors.

The codec-based scheme is illustrated in figure 1 and has principally been designed for voice-to-voice communication purposes. Many low bit-rate speech codecs have been developed [2] and vary in terms of their bit-rate, quality and underlying algorithm. A simple method of performing speech recognition from a speech codec is to take the decoded speech and pass it through feature extraction into the speech recogniser. This method is straightforward, but generally leads to reduced recognition performance as a result of distortions made by the codec [3].

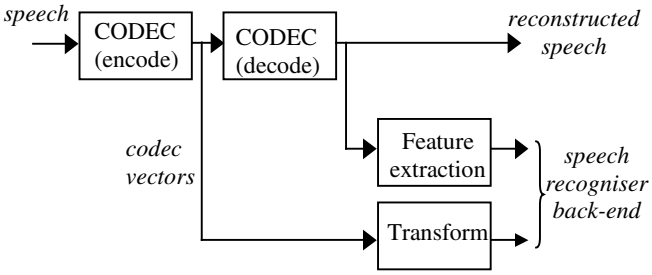

Figure 1: Codec-based speech reconstruction and recognition

An improvement is to transform the codec vectors themselves into features for speech recognition without using the time-domain signal. A number of techniques [3,4] have been suggested for this. For example, in [3] codec vectors based on line spectral pairs (LSPs) are converted into linear predictive coding (LPC) coefficients. A Fourier transform then converts these into the magnitude spectrum, from which MFCCs can be estimated via the standard mel-filterbank analysis, logarithm and discrete cosine transform.

The alternative scheme to provide both speech recognition and speech reconstruction from a common feature vector is based on the feature extraction component a speech recogniser. This is shown in figure 2. On the terminal device the feature extraction component of the speech recogniser produces a stream of static feature vectors which are transmitted to the remote recogniser.

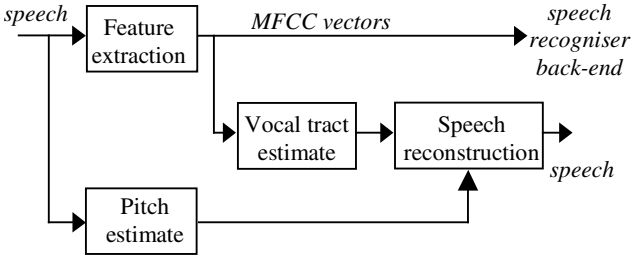

Figure 2: FE-based speech reconstruction and recognition.

The information needed for speech reconstruction must be extracted from this feature vector stream. Some recent work [5,

6] proposed a method of achieving reconstruction using the sinusoidal model of speech coding [7]. They converted the MFCCs to a smoothed spectral estimate using an inverse DCT and exponential operation. A estimate of pitch is also computed at the terminal device from the speech and included as an extra component of the feature vector. This enables the location of the spectral peaks in the sinusoidal model to be determined. The amplitude of the peaks were estimated from the smoothed spectral estimate and the phases from a phase model.

In this work a different approach for speech reconstruction is developed, based upon the source-filter model of speech production. Section 2 briefly reviews this model and in particular identifies parameters of the model which need to be extracted from the feature vectors. Section 3 considers how the vocal tract filter coefficients can be estimated from an MFCC vector, while section 4 considers the problem of creating an appropriate excitation signal. Results of the speech reconstruction are presented in section 5 and a conclusion made in section 6.

2. MODEL OF SPEECH PRODUCTION

A common model of speech production, used in many low bit-rate speech codecs, is the source-filter model [2]. The filter models the spectral envelope of the vocal tract and is typically implemented as an all-pole filter. The input, or source, is the excitation signal, *e(n)*, and this models the signal generated by the vocal chords. The resulting speech signal, *s(n)*, is produced as

$$s(n) = \sum_{i=1}^{p} a_i\, s(n-i) + Ge(n) \qquad (1)$$

where a_i is the i^{th} coefficient of the p^{th} order vocal tract filter and G is a gain term. The coefficients of the vocal tract filter are calculated using linear predictive analysis which minimises the mean square prediction error. Using the autocorrelation method to determine the coefficients leads to the following set of equations

$$\sum_{i=1}^{p} a_i\; r_{|j-i|} = r_j \qquad 1 \le j \le p \qquad (2)$$

where r_j is the j^{th} autocorrelation coefficient of the windowed speech signal. This is often written in matrix notation as

$$\mathbf{aR} = \mathbf{p} \qquad (3)$$

The matrix of autocorrelation coefficients, **R**, has a Toeplitz structure which allows the filter coefficients to be extracted using the Levinson-Durbin recursive procedure.

Many techniques have been proposed for encoding a low bit-rate representation of the excitation signal. These range from simple pitch pulse/white noise excitation used in vocoders to the more sophisticated codebook approaches found in analysis-by-synthesis schemes [2].

The next section describes how the parameters of the vocal tract filter can be estimated from an MFCC feature vector, while section 4 shows how an extra element of the feature vector can provide sufficient information to generate an excitation signal.

3. ESTIMATION OF VOCAL TRACT

The time-varying coefficients of the vocal tract filter must be estimated from the MFCC feature vectors. Figure 3 illustrates the stages through which a speech signal passes to be transformed into an MFCC vector. This conforms to the Aurora standard proposed by ETSI [1] and is used throughout this work.

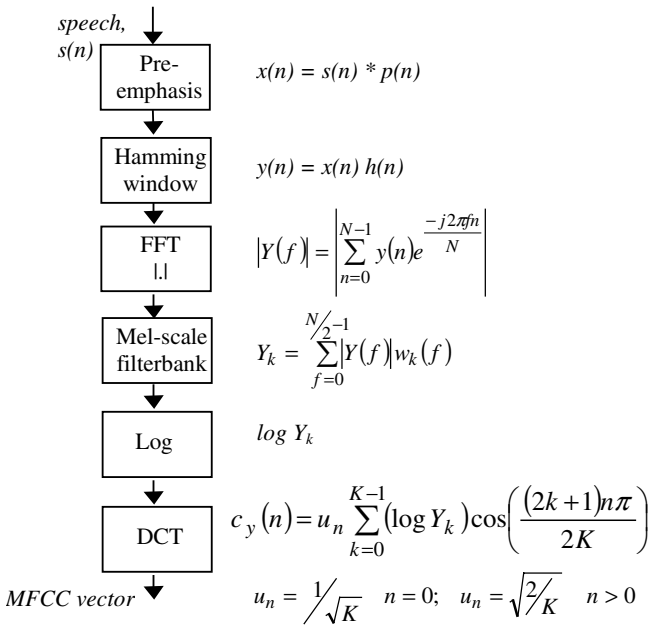

Figure 3: Procedure for calculating MFCC vectors.

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. For example, taking the magnitude of the complex spectrum loses phase information. Similarly the mel-filterbank quantises the frequency spectrum into a reduced resolution and the truncation of the DCT smoothes the log filterbank.

Too much information is lost to allow the MFCC vectors to be inverted back into a time-domain signal through a reversal of the procedures illustrated in figure 3. However it is possible to recover a reasonable, although smoothed, estimate of the magnitude spectrum. From this magnitude spectrum an estimate of the vocal tract filter coefficients can be made. Thus, computation of vocal tract filter coefficients from MFCCs can be considered as a two-stage process; i) inversion of the MFCC vector to its magnitude spectrum, and ii) calculation of vocal tract filter coefficients from the MFCC-derived magnitude spectrum.

3.1. Inversion of MFCCs to Magnitude Spectrum

Figure 3 has shown that the process by which an MFCC vector is computed from a speech signal contains a number of stages which are not invertible. It is, however, possible to make certain approximations to the information which has been discarded to allow an inverse to be estimated.

The first stage of inverting an MFCC vector, $\boldsymbol{c}_y$, into a magnitude spectral representation requires an estimate of the log filterbank vector. At this stage higher-order MFCCs have been truncated

(in this work from 23 dimensions down to 13) and these represent the finer detail of the log filterbank. Zero-padding the truncated MFCC vector to the dimensionality of the filterbank ($K=23$) allows an inverse DCT to be taken. This results in a smoothed estimate of the log filterbank vector, $log\,\hat{Y}_k$,

$$\log \hat{Y}_k = \sum_{n=0}^{K-1} u_n \; c_y(n) \cos\left(\frac{(2k+1)n\pi}{2K}\right) \tag{4}$$

The log operation is straightforward to invert, through the use of the exponential operator, and provides an estimate of the (linear) mel-filterbank vector.

The next stage is to estimate the M-dimensional magnitude spectrum – in this work $M=128$ spectral bins are computed. This requires M linearly spaced magnitude estimates to be computed from the K mel-spaced filterbank channels. As $M>K$ some form of interpolation is necessary. Work at Motorola [6] utilised the inverse DCT to perform the interpolation. This was achieved by taking a high resolution (3933 dimensional) IDCT to transform the truncated MFCC vector into the log filterbank domain. This gave a very fine mel-scale resolution from which the linearly spaced spectral bins could be estimated to recreate the M-dimensional magnitude spectrum.

At this point a high-frequency tilt of the magnitude spectrum is present. This comes from both the pre-emphasis filter and as a result of the mel-spacing of the filterbank analysis.

The area under the triangular filters used in the mel-filterbank analysis increases at higher frequencies. The effect of this is to impose a high frequency tilt on the resulting mel-filterbank channels which distorts the estimated magnitude spectrum. This tilt can be equalized in the frequency domain by scaling the mel-filterbank outputs, $\hat{Y}_k$, by the area of the corresponding triangular mel-filter, w_k.

An alternative implementation for equalization is to transform the vector, $\boldsymbol{w}$, (comprising the areas, w_k, of the mel-spaced triangular windows) into the cepstral domain through a log and DCT

$$c_w(n) = u_n \sum_{k=0}^{K-1} (\log w_k) \cos\left(\frac{(2k+1)n\pi}{2K}\right) \tag{5}$$

This cepstral representation of the distortion, $\boldsymbol{c}_w$, can be subtracted from original speech MFCC vector, $\boldsymbol{c}_y$.

The effect of the pre-emphasis filter is also additive in the cepstral domain. Computing an MFCC vector, $\boldsymbol{c}_p$, from the pre-emphasis filter enables equalization to also be made as a subtraction operation in the cepstral domain.

Therefore the equalized MFCC vector, $\boldsymbol{c}_x$, is computed as

$$\boldsymbol{c}_x = \boldsymbol{c}_y - \boldsymbol{c}_w - \boldsymbol{c}_p \tag{6}$$

This now forms that starting point from which the inversion to the magnitude spectrum begins using equation (1).

Figure 4 compares the recovered magnitude spectrum with that computed from the original speech samples. The magnitude spectrum of a frame of 200 speech samples is shown as the dotted line. The magnitude spectrum estimated from an non-truncated (23-D) MFCC vector is shown as the solid line. The dashed line shows the magnitude estimate from a truncated 13-D MFCC vector.

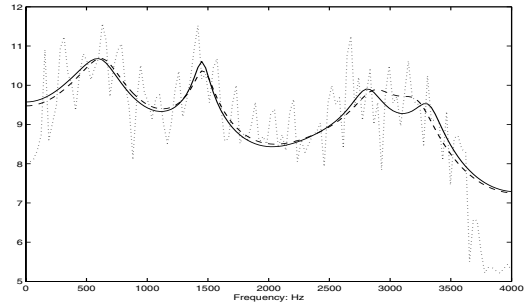

Figure 4: Magnitude spectrum reconstruction.

The figure shows that the envelope of the magnitude spectrum has been reasonably well preserved. As expected the 23-D MFCC vector provides a more detailed reconstruction than the truncated 13-D MFCC.

3.2. Magnitude Spectrum to Vocal Tract Filter

In conventional LPC analysis the coefficients of the vocal tract filter, $\boldsymbol{a} = \{a_1, a_2, \ldots, a_p\}$, are calculated by minimizing the mean square error of a p^{th} order linear predictor [2]. The autocorrelation solution for finding the filter coefficients requires the first $p+1$ autocorrelation coefficients, $\{r_0, r_1, \ldots, r_p\}$, which are normally computed from a frame of windowed speech.

The Wiener-Khintchine theorem relates the autocorrelation coefficients to the power spectrum through an inverse Fourier transform. Squaring and reflecting the magnitude spectrum estimated from the MFCC vector and then computing its inverse Fourier transform provides the set of autocorrelation coefficients

$$\hat{r}_j = \frac{1}{N} \sum_{f=0}^{N-1} \left|\hat{X}(f)\right|^2 e^{\frac{j2\pi fj}{N}} \tag{7}$$

An estimate of the vocal tract filter coefficients can now be made by inserting the estimated autocorrelation coefficients into equation (3).

Figure 5 shows the frequency response of the vocal tract filter estimated from both a 23-D non-truncated MFCC vector (solid line) and a 13-D truncated MFCC vector (dashed line). For comparison, the dotted line shows the frequency response of the vocal tract filter estimated directly from the speech samples using LPC analysis.

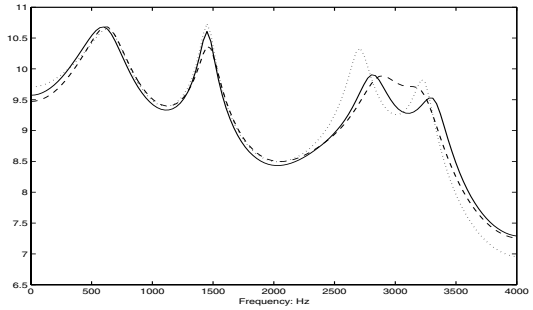

Figure 5: MFCC-derived vocal tract frequency response.

The figure shows that the MFCC-derived vocal tract filter successfully models the first two formants – for both the truncated and non-truncated MFCC vectors. The third and fourth formants are merged into a single spectral peak by the vocal tract filter estimated from the truncated MFCC. However, the non-truncated MFCC vector retains sufficient detail to differentiate the peak into two separate formants. This matches the original vocal tract frequency response. Analysis made on other frames of speech gave similar comparisons.

4. SPEECH EXCITATION

The excitation signal for the source-filter model of speech production can be found by inverse filtering the original speech signal by the inverse vocal tract filter. Many suggestions have been proposed for encoding this excitation signal [2]. At present in this work the excitation signal is reconstructed simply using a series of pitch pulses or white noise, depending on whether the speech is voiced or unvoiced.

Pitch is estimated on the terminal device using the comb filter method described in [8]. To signify unvoiced speech the pitch is set to zero. This pitch feature is transmitted as an extra element of the feature vector. A suitable value for the gain, *G*, can be estimated from the log energy element of the feature vector.

5. SPEECH RECONSTRUCTION

The MFCC-derived vocal tract filter and the pitch-based excitation signal can now be combined to reconstruct the speech signal using equation (1). The MFCC vectors in this work conformed to the ETSI standard [1]. These were generated at a rate of 100 frames per second which gave an update period of 10ms to the vocal tract filter and pitch estimate. The bandwidth of the speech was 4kHz and the mel-filterbank had 23 channels. To conform with the ETSI Aurora standard the feature vectors were compressed to a bit rate of 4.8kbps.

Informal listening tests revealed that an intelligible speech signal was produced. A slight deterioration in quality was heard when moving from 23-D MFCCs to 13-D MFCCs. As a comparison a speech signal was produced by generating the vocal tract filter from the original speech samples using LPC analysis. The resultant quality was comparable to the MFCC-derived speech which indicates that the MFCCs are able to produce a good estimate of the vocal tract filter.

Figure 6a illustrates the spectrogram of a sequence of digits (9-8-1-8-8-3-0) taken from the original speech samples. Figure 6b shows the spectrogram of the same sequence but reconstructed from a 13-D MFCC vector stream and pitch estimate.

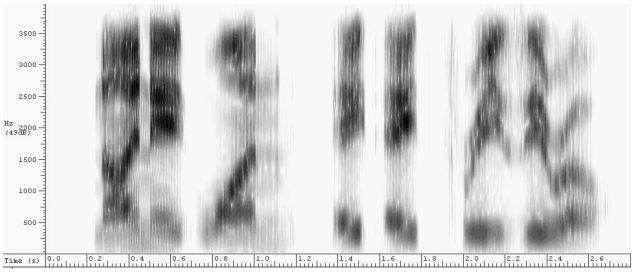

Figure 6a: Spectrogram of original speech signal

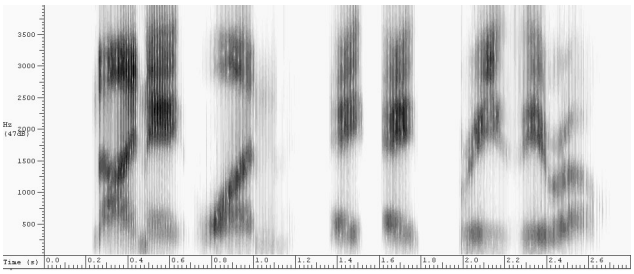

Figure 6b: Spectrogram of MFCC reconstructed speech.

The spectrograms show that a good representation of the formats of the original speech has been reconstructed. Problems do occur in places where the smoothed resolution of the magnitude spectrum results in formants merging into a single peak.

6. CONCLUSION

This work has shown that speech reconstruction is possible from a stream of MFCC vectors using a source-filter model of speech production. Additional information regarding the excitation signal has needed to be included in the feature vector, but can be as simple as a single element indicating the pitch period.

Inversion of the MFCC vector to a smoothed version of the magnitude spectrum has provided sufficient information to estimate the vocal tract filter coefficients. Analysis has shown that a reasonable estimate of the spectral envelope of the vocal tract frequency response can be attained, although in some cases higher frequency formants can be lost or merged. Listening tests reveal that the resulting speech quality is comparable to that produced from LPC analysis of the original speech signal.

7. REFERENCES

1. ESTI document - ES 201 108 – STQ: DSR – Front-end feature extraction algorithm; compression algorithm, 2000.
2. W.B. Kleijn and K.K. Paliwal, “Speech coding and synthesis”, Elsevier, 1995.
3. R. Tucker et al, “Compression of acoustic features – are perceptual quality and recognition performance incompatible goals?”, Proc. Eurospeech, 1999.
4. H.K. Kim and R.V. Cox, “A bitstream-based front-end for wireless speech recognition on IS-136 communication systems”, IEEE Trans. Speech and Audio Processing, volume 9, number 5, pp. 558-568, 2001.
5. D. Chasan et al, “Speech reconstruction from mel frequency cepstral coefficients and pitch”, Proc. ICASSP, 2000.
6. T. Ramabadran et al, “Enhancing distributed speech recognition with back-end speech reconstruction”, Proc. Eurospeech, 2001.
7. R.J. McAuley and T.F. Quatiery, “Speech analysis/synthesis based on a sinusoidal representation”, IEEE Tran. ASSP, vol. 34, pp. 744-754, 1986.
8. D. Chazan et al, "Efficient Periodicity Extraction Based on Sine-wave Representation and its Application to Pitch Determination of Speech Signals", Proc Eurospeech, 2001.