MFCC Computation from Magnitude Spectrum of Higher Lag Autocorrelation Coefficients for Robust Speech Recognition

Benjamin J. Shannon and Kuldip K. Paliwal

School of Microelectronic Engineering
Griffith University, Brisbane, QLD 4111, Australia
Ben.Shannon@student.griffith.edu.au, K.Paliwal@griffith.edu.au

Abstract

Processing of the speech signal in the autocorrelation domain in the context of robust feature extraction is based on the following two properties: 1) pole preserving property (the poles of a given (original) signal are preserved in its autocorrelation function), and 2) noise separation property (the autocorrelation function of a noise signal is confined to lower lags, while the speech signal contribution is spread over all the lags in the autocorrelation function, thus providing a way to eliminate noise by discarding lower-lag autocorrelation coefficients). In this paper, we use these properties to derive robust features for automatic speech recognition. We compute the magnitude spectrum of the one-sided higher-lag autocorrelation sequence, process it through a Mel filter bank and parameterise it in terms of Mel Frequency Cepstral Coefficients (MFCCs). Since the proposed method combines autocorrelation domain processing with Mel filter bank analysis, we call the resulting MFCCs, Autocorrelation Mel Frequency Cepstral Coefficients (AMFCCs). Recognition experiments are conducted on the Aurora II database and it is found that the AMFCC representation performs as well as the MFCC representation in clean conditions and provides more robust performance in the presence of background noise.

1. Introduction

There is a long history of feature extraction algorithms for Automatic Speech Recognition (ASR) that use techniques based on processing in the autocorrelation domain [1][2]. The attraction of autocorrelation domain processing can be illustrated easily by taking a simple example, where a speech signal for a vowel frame (/iy/) is corrupted by additive white noise. Figures 1, 2 and 3 show the autocorrelation sequences of clean speech, white noise, and corrupted speech, respectively, for this example. It is well known that if the speech signal and white noise sequence are uncorrelated, the autocorrelation of their sum is equal to the sum of their autocorrelations. Furthermore, as seen from Fig. 2, the autocorrelation sequence of the white noise sequence is an impulse-like signal. These properties combine to show that the contribution from white noise in the autocorrelation sequence is neatly contained in the zero-lag coefficient, while the contribution from the information carrying speech signal is spread over a broad range of lag indexes (Fig. 1). When we consider more realistic coloured noises occurring in real life (such as car noise, babble noise, etc.), their contribution to the autocorrelation sequence may spread away to lags greater than zero (as shown in Fig. 4), but it is still confined to relatively lower lags. Therefore, noise-robust spectral estimates should be possible through algorithms that focus on higher lag autocorrelation coefficients.

The autocorrelation sequence has another important property, which states that it preserves in it the poles of the original signal sequence, as illustrated by McGinn and Johnson [3]. Assuming the original signal to be an all-pole sequence generated by an all-pole model that has been excited by a single impulse, they showed that the poles of the autocorrelation sequence are the same as the poles of the original sequence. The concept of the pole preserving property was extended by Mansour and Juang [1] to include an impulse train excitation and a white Gaussian noise excitation. These are better approximations of the excitation source for voiced and unvoiced speech signals, respectively. The pole preserving property is important when processing in the autocorrelation domain. It means spectral estimates made with the autocorrelation sequence will show poles in the same place as estimates made with the original time domain signal, thus the autocorrelation domain processing will provide information about the signal similar to that obtained from the original signal directly.

A number of techniques have been proposed in the literature based on autocorrelation domain processing. The first technique proposed in this area was based on the use of High-Order Yule-Walker Equations (HOYWE) [4], where the autocorrelation coefficients that are involved in the equation set exclude the zero-lag coefficient. Other similar methods have been used that either avoid the zero-lag coefficient [4][5][6], or reduce the contribution from the first few coefficients [1][2]. All of these methods are based on linear prediction (LP) approach and provide some robustness to noise, but their recognition performance for clean speech is much worse than the unmodified or conventional LP approach [2].

A potential source of error in using LP methods to estimate the power spectrum of a varying SNR signal is highlighted by Kay [7]. He showed that the model order is not only dependent on the AR process, but also on the prevailing SNR condition. Therefore, in the present paper, we do not use an LP based method. Instead, we compute the magnitude spectrum of the one-sided higher-lag autocorrelation sequence, process it through a Mel filter bank and parameterise it in terms of MFCCs. Since the proposed method combines autocorrelation domain processing with Mel filter bank analysis, we call the resulting MFCCs, Autocorrelation Mel Frequency Cepstral Coefficients (AMFCCs).

The paper organisation is as follows. Section 2 gives a description of the newly proposed algorithm, followed by a discussion on the discarded autocorrelation coefficients and the choice of window function in Sections 3 and 4. Finally, an experimental comparison of the proposed feature set with MFCCs is presented in Section 5.

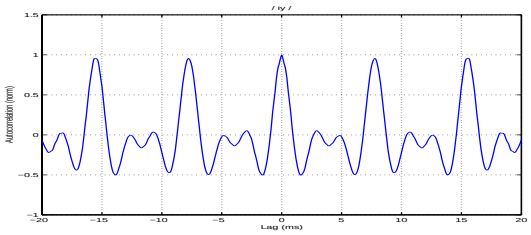

Figure 1: Autocorrelation sequence of vowel /iy/.

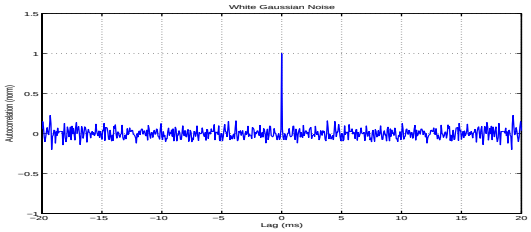

Figure 2: Autocorrelation sequence of white Gaussian noise.

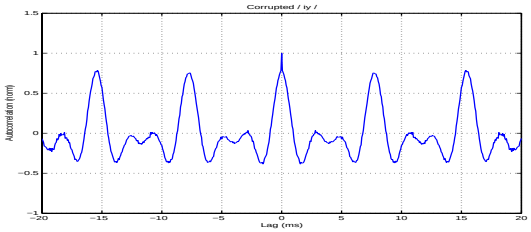

Figure 3: Autocorrelation sequence of corrupted /iy/ with white Gaussian noise.

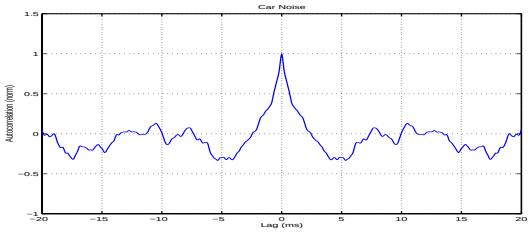

Figure 4: Autocorrelation sequence of car noise.

2. AMFCC procedure

Figure 5 shows the block diagram of the proposed AMFCC feature compared to the standard MFCC feature. The following subsections describe each step of the proposed AMFCC algorithm in more detail.

2.1. Framing and windowing

The speech signal is pre-emphasised, then segmented into 32 ms frames every 10 ms. A Hamming window is then applied to each frame before finding the autocorrelation sequence.

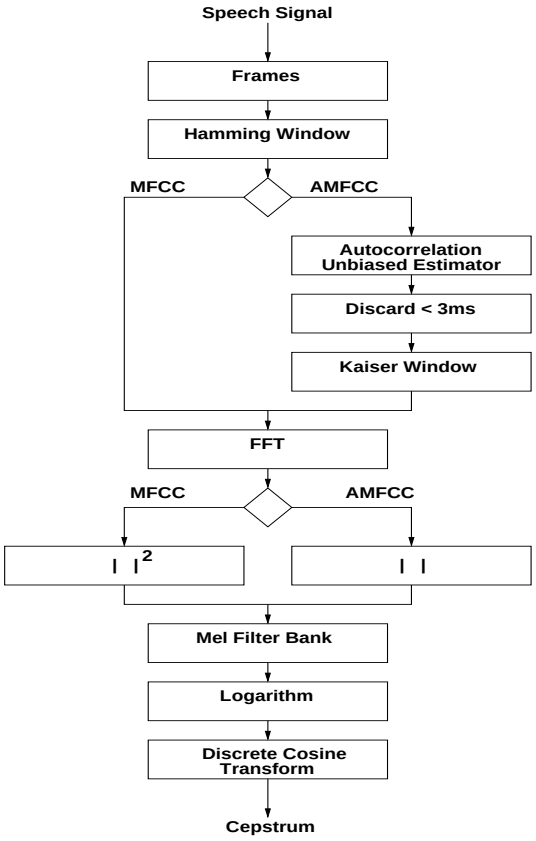

Figure 5: MFCC and AMFCC block diagram.

2.2. Autocorrelation sequence

Unbiased autocorrelation is found for each frame using equation (1). This form of unbiased estimate can be efficiently calculated by utilising Wiener-Khinchin theorem. If a frame of length N is zero padded to length $2N$, a biased autocorrelation can be found by taking the inverse Fourier transform of the power spectrum of the $2N$ length sequence. The resulting autocorrelation estimate can be made unbiased by dividing each term by $N - i$, where i is the lag index in the autocorrelation sequence. This process yields a sequence that is numerically identical to that given by the following equation:

$$R_{xx}(i) = \frac{1}{N-i} \sum_{n=0}^{N-i-1} x(n)x(n+i), \quad i = 0, 1, ..., N-1 \tag{1}$$

2.3. Discard lower lag autocorrelation coefficients

As with previous methods, the zero-lag autocorrelation coefficient is discarded from the analysed sequence. As an extension of this, all lower lag coefficients up to 3 ms lag are also discarded, which is further discussed in Section 3.

2.4. Window function

A Kaiser window function that has a side lobe attenuation of approximately 80 dB is applied to the one-sided higher-lag autocorrelation sequence. Due to the high dynamic range of the autocorrelation sequences, a Hamming window is not used, which is discussed further in Section 4.

2.5. Magnitude spectrum of the higher-lag autocorrelation sequence

The magnitude spectrum of the windowed autocorrelation fragment is found as opposed to the more typical power spectrum. By finding the magnitude spectrum, the dynamic range of the resulting spectrum estimate is of the same order as the power spectrum of the original speech signal. This is explained by the relationships shown in Eq. (2). As shown, the time autocorrelation sequence and the power spectrum are related through a Fourier transform. Likewise, the power spectrum and the time signal are related through the absolute square of the Fourier transform.

$$P(\omega) = |\mathcal{F}[x(n)]|^2 = \mathcal{F}[R_{xx(n)}] \tag{2}$$

2.6. Filter bank and cepstrum

Just as in the MFCC algorithm, a Mel filter bank analysis is performed to obtain a perceptually meaningful spectral estimate. From this step onwards, the AMFCC procedure and the MFCC procedure are identical.

3. Choice of discarded autocorrelation coefficients

Spectral leakage can be reduced in the resulting Fourier spectrum by moving the analysis window away from the large value coefficients around the zero-lag index. The autocorrelation sequence of a voiced speech frame (Fig. 1) has a maximum value at zero-lag, then oscillates outwards with a cosine like nature, typically peaking again at the index corresponding to the pitch period. Since the autocorrelation sequence is a zero phase sequence [8], the opportunity exists to position the analysis window in a way, which on average, gives the autocorrelation coefficients at the two ends of the analysis window that have absolute values nearest to zero. By doing this, the autocorrelation sequence implicitly tapers towards zero (i.e., discontinuities at the boundaries are minimised), thus reducing the spectral leakage [9].

Figure 6 shows a plot of average absolute autocorrelation function for speech. To produce this plot, the autocorrelation coefficients of each of the speech frames from the TIMIT database were calculated, and their absolute values were averaged over all frames. This plot suggests that the autocorrelation window to be used for spectral analysis should begin at 3 ms lag in order to minimise spectral leakage.

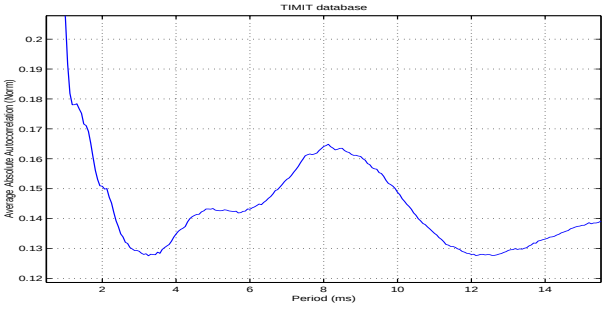

Figure 6: Measured average absolute autocorrelation.

4. Choice of window function

As highlighted previously by Mansour and Juang [1], if the time domain signal has a power spectrum dynamic range of 40 dB, then the corresponding autocorrelation sequence will have a dynamic range of 80 dB. For this reason, a window function with high side lobe attenuation is required when processing the autocorrelation sequence, in order to maintain the larger dynamic range. It is noted that the Kaiser window was tested in the Short-time Modified Coherence (SMC) method development without success [1]. Our experience has not been the same during the development of the proposed method. Both the Kaiser window and Dolph-Chebyshev window displayed a performance advantage over the Hamming window in our experiments.

The Kaiser window expression is given in Eq. (3). In this expression N is the window length of the autocorrelation sequence (after 3 ms lag), n is the lag index, I_0 is a modified Bessel function of the first kind, and α is the design parameter that sets the side lobe attenuation level. For the AMFCC, α was chosen to be 10.

$$w(n) = \begin{cases} \frac{I_0\left(2\alpha\sqrt{\frac{n}{N-1}-(\frac{n}{N-1})^2}\right)}{I_0(\alpha)} & ,0 \leq n < N \\ 0 & ,\text{otherwise} \end{cases} \tag{3}$$

By choosing α to be 10, the resulting window function has a dynamic range of approximately 80 dB. When this window function is applied to the autocorrelation fragment, the resulting power spectrum like estimate has a dynamic range equivalent to that of conventional MFCCs.

5. Recognition experiment

Using the Aurora II database, Aurora II experiment scripts, and HTK software [1], the noise robustness of AMFCC features was compared with MFCC features. The experiments conducted used the clean training, test set A scenario. With this scenario, noise robustness is evaluated using four different noise types; subway, babble, car and exhibition, at seven different SNRs, ranging from clean, then 20dB to -5dB in 5dB steps.

For these experiments, the speaker independent word models had 16 emitting states. The modelled acoustic feature vector was composed of a 12 dimension base feature concatenated with a logarithmic energy coefficient. This was then concatenated with delta and acceleration coefficients to produce a final 39-dimensional feature vector.

Figures 7, 8, 9 and 10 show word recognition accuracy curves for MFCC and AMFCC features. In the uncorrupted or clean case, MFCC and AMFCC performance is equal. As noise is introduced, MFCC's word accuracy degrades faster than AMFCC for all types of noise. For noise types that have autocorrelation sequences similar to speech (babble, exhibition), AMFCCs word accuracy improvement over MFCCs was lower than types with dissimilar autocorrelation sequences (car, subway).

6. Conclusions

In this paper, a new noise robust speech recognition feature extraction algorithm was proposed. The new algorithm used the magnitude spectrum of higher-lag autocorrelation coefficients, as opposed to similar algorithms that operate on the autocorrelation sequence, which all employ linear prediction. Experiments so far have shown the new representation to be more robust to

[1] Hidden Markov Tool Kit (HTK), http://htk.eng.cam.ac.uk

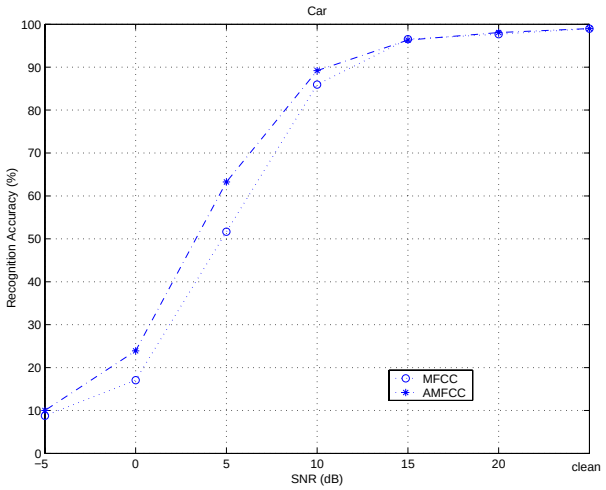

Figure 7: AMFCC - MFCC (Car noise).

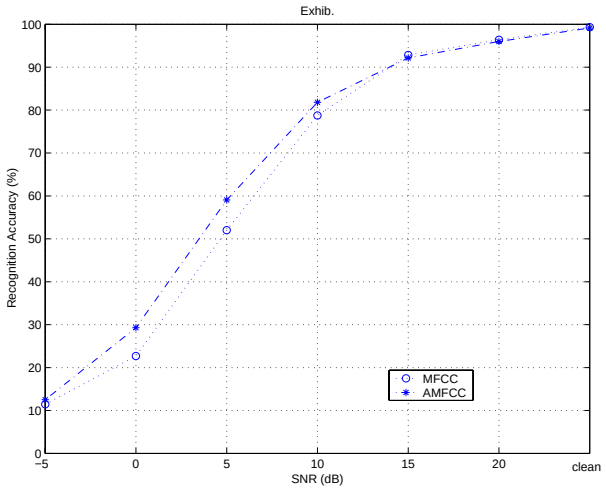

Figure 9: AMFCC - MFCC (Exhibition noise).

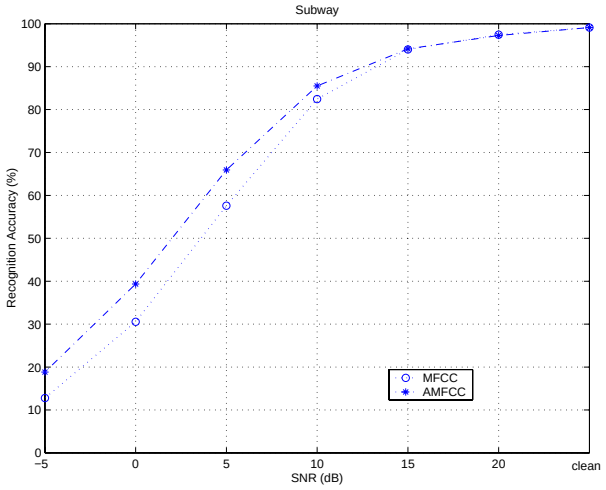

Figure 8: AMFCC - MFCC (Subway noise).

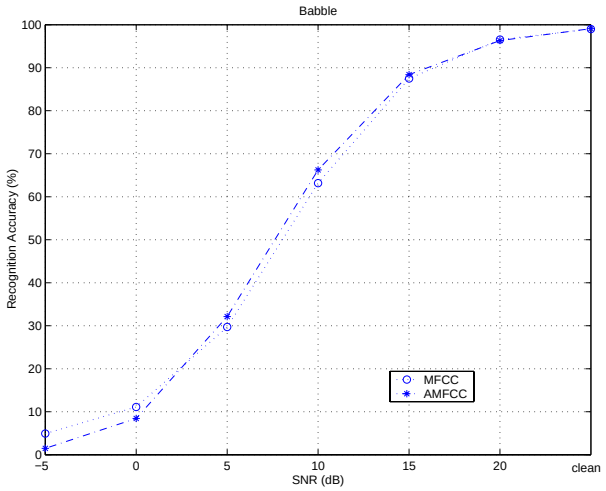

Figure 10: AMFCC - MFCC (Babble noise).

background noise than MFCCs, and more significantly, without compromised performance in clean conditions.

7. References

[1] D. Mansour and B. H. Juang, "The short-time modified coherence representation and noisy speech recognition," *IEEE Transactions on ASSP*, vol. 37, no. 6, pp. 795–804, Jun 1989.

[2] J. Hernando and C. Nadeu, "Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 80–84, Jan. 1997.

[3] D. McGinn and D. H. Johnson, "Reduction of all-pole parameter estimator bias by successive autocorrelation," in *Proc. ICASSP*, 1983, pp. 1088–1091.

[4] Y. T. Chan and R. P. Langford, "Spectral estimation via the high-order yule-walker equations," *IEEE Trans. on ASSP*, vol. ASSP-30, no. 5, pp. 689–698, Oct. 1982.

[5] K. K. Paliwal, "A noise-compensated long correlation matching method for ar spectral estimation of noisy signals," in *Proc. ICASSP*, 1986, pp. 1369–1372.

[6] J. A. Cadzow, "Spectral estimation: An overdetermined rational model equation approach," in *Proc. IEEE*, vol. 70, Sep. 1982, pp. 907–939.

[7] S. M. Kay, "The effects of noise on the autoregressive spectral estimator," *IEEE Transactions on ASSP*, vol. ASSP-27, no. 5, pp. 478–485, Oct. 1979.

[8] J. Suzuki, "Speech processing by splicing of autocorrelation function," in *Proc. ICASSP*, Apr. 1976, pp. 713–716.

[9] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," in *Proc. of the IEEE*, vol. 66, no. 1, Jan. 1978, pp. 51–83.