Noise Robust Speech Recognition using Higher-lag Autocorrelation Coefficients

Benjamin J. Shannon and Kuldip K. Paliwal

***Abstract*— In this paper, we introduce a noise robust spectral estimation technique for speech signals, which we refer to as Higher-lag Autocorrelation Spectral Estimation (HASE). By utilising only the higher-lag portion of the autocorrelation sequence to compute a spectral estimate, the HASE method reduces the contribution of noise components. We also introduce a high dynamic range window design method called DDR, and utilise both the HASE and DDR techniques in a modified Mel Frequency Cepstral Coefficient (MFCC) algorithm to produce noise robust speech recognition features. We call the new features Autocorrelation Mel Frequency Cepstral Coefficients (AMFCCs). We compare the recognition performance of AMFCCs to MFCCs for a range of stationary and non-stationary noises on the Aurora II database. We show that the AMFCC features perform as well as MFCCs in clean conditions and have higher noise robustness.**

I. Introduction

When designing Automatic Speech Recognition (ASR) front-ends, we model the human speech production mechanism using a two-part source-system model. The typical model contains two sub-processes. Only one sub-process is active at a time based on whether the speech sound is voiced or unvoiced. For the voiced speech case, the production model is composed of a variable response filter excited by a train of periodic impulses, and for the unvoiced case, the variable response filter is excited by white noise.

In an ASR system front-end, we are interested in extracting the response of the variable filter. This response is typically estimated over short-time frames (32 ms) using a spectral estimation method, and then transformed into low dimensional, uncorrelated features for processing in the pattern recognition stage. Features of this kind capture sufficient linguistic detail to give high recognition rates. The state-of-the-art Mel Frequency Cepstral Coefficient (MFCC) front-end is one such example.

The period of the periodic pulse train in the voiced speech model determines the pitch of the speech. The human pitch period is typically confined to a range between 2 ms and 12 ms due to bio-mechanics. If we assume that a voiced phoneme is stationary over a 32 ms analysis frame, we can expect to capture between 2 and 16 highly correlated observations of the impulse response of the model filter. We propose to use this predictable correlation pattern for noise reduction.

Noise signals are often the combination of a number of complex sources, thus they tend to lack the predictable periodicity of speech signals. Noise signals are more random, thus tend to have high magnitude lower-lag autocorrelation coefficient and low magnitude higher-lag coefficients. Alternatively, due to their periodicity, voiced speech signals have high magnitude periodic higher-lag coefficients. Since uncorrelated signals are additive in the autocorrelation domain, we can reduce the contribution from noise signals in a spectral estimate by utilising only the coefficients from the higher-lag region.

The potential for computing noise robust speech recognition features from the autocorrelation domain has attracted a lot of attention. A number of speech recognition feature extraction 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 [1], 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 [1] [2] [3], or reduce the contribution from the first few coefficients [4] [5]. All of these methods are based on linear prediction (LP) processing and provide some robustness to noise, but their recognition performance for clean speech is much worse than the unmodified or conventional LP approach [5]. Due to inherent problems associated with LP based methods in noisy conditions [6], we use an alternative approach to process the autocorrelation sequence. We compute the magnitude spectrum of the one-sided higher-lag autocorrelation sequence using the Fourier transform and refer to the technique as Higher-lag Autocorrelation Spectral Estimation (HASE). We further propose to process the HASE 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).

Speech recognition feature extraction algorithms are typically designed assuming stationary broadband (usually white) noise. For the evaluation of the AMFCC technique, we consider stationary noise signals as well as non-stationary noises, such as emergency vehicle sirens and chirp signals. We show that higher-lag autocorrelation processing is particularly robust against these types of noise disturbances. In section II we discuss some properties of the autocorrelation sequence in relation to speech and noise signals. We then describe, in section III, the newly proposed higher-lag autocorrelation spectral estimation technique and test its effectiveness for noise robust speech feature extraction using the Aurora II database in section IV.

II. Properties of Autocorrelation Sequences

In this section, we demonstrate briefly how the short-time autocorrelation sequence captures the smooth spectral

The authors are with the School of Microelectronic Engineering, Griffith University, Brisbane, QLD 4111, Australia.

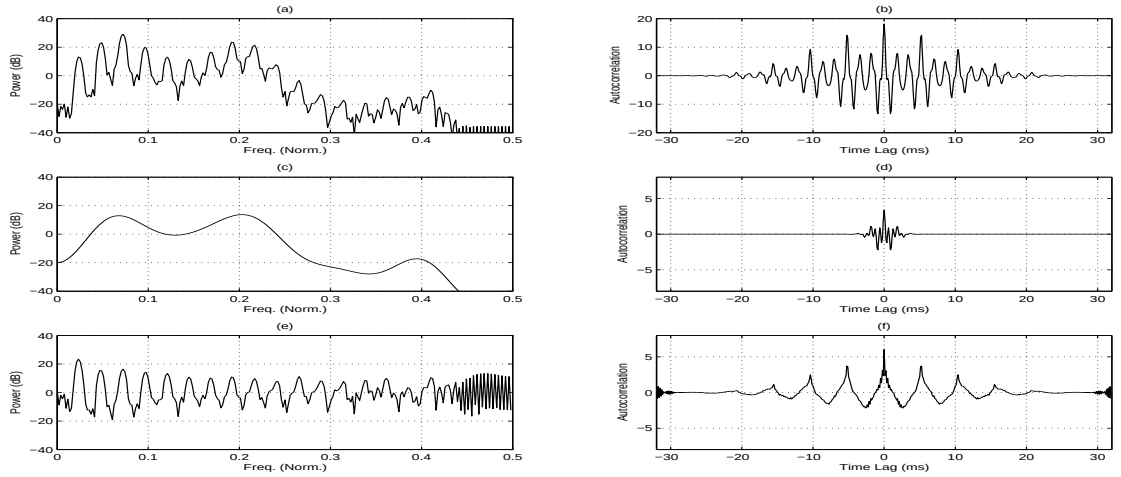

Fig. 1. Decomposition of a 32 ms voiced speech frame, containing an /r/ sound. (a) The original logarithmic power spectrum. (b) Autocorrelation sequence associated with the spectrum in (a). (c) The smooth logarithmic spectral envelope computed by retaining the first 12-cepstral coefficients. (d) The autocorrelation sequence associated with the spectrum shown in (c). (e) The logarithmic excitation spectrum. (f) Autocorrelation sequence associated with the logarithmic spectrum shown in (e).

envelope information of a voiced speech signal. We then discuss the autocorrelation distribution for noise signals.

A. Speech Signals

A depiction of how the smooth spectral envelope information is distributed in the autocorrelation sequence is shown in Fig.1. The logarithmic power spectrum of an /r/ sound is shown in plot (a). This plot shows the harmonic structure typical of voiced speech, along with the information-bearing envelope. Plot (b) shows the autocorrelation sequence associated with the spectrum in (a). By using cepstral processing, we decomposed the spectrum in (a) into the smooth spectrum in (c) and the excitation spectrum shown in (e). (d) and (f) show the corresponding autocorrelation sequences of these two spectra respectively.

Figure 1(d) shows that the smooth power spectrum information is contained in a small number of autocorrelation coefficients. The full autocorrelation sequence shown in (b) is equal to the convolution of the autocorrelation sequences in (d) and (f). This demonstrates that the smooth power spectrum envelope information is spread throughout the whole autocorrelation sequence of the original speech signal frame. Therefore, we should be free to estimate the smooth spectral envelope using any region of the autocorrelation sequence.

B. Noise Signals

The autocorrelation sequences of noise signals vary much more than the autocorrelation sequences of speech signals. This variation can be attributed to the larger range of production mechanisms for noise signals compared to the simple production model applicable to speech signals. All autocorrelation sequences have the largest absolute value at the zero lag location. This coefficient represents the energy of the signal. The shape of the autocorrelation envelope moving away from the zero lag location is directly related to the noise source. The biased autocorrelation estimation algorithm causes some of the decay, but generally, the decay is faster than the algorithm imposed rate. As an example of non-stationary noise, an emergency vehicle siren and its analysis is shown in Fig.2. In this figure, plot (a) shows the spectrogram for a two second segment of the noise. Plots (b), (c) and (d) show the logarithmic power spectrum at times 0.5, 1.0 and 1.5 seconds respectively. Plots (e), (f) and (g) show the autocorrelation sequence associated with the spectra in plots (b), (c) and (d) respectively.

When uncorrelated noise is added to a speech signal, the combination in the autocorrelation domain can be described as follows.

- The zero-lag coefficient is corrupted.
- The lower-lag coefficients are generally more corrupted than the higher-lag coefficients.

If the spectral envelope information is sufficiently contained in the higher-lag autocorrelation coefficients, a more noise robust spectral estimate should result if the more corrupt lower-lag coefficients are de-emphasised during spectral estimation.

III. Spectral Estimation from Higher-Lag Autocorrelation

To compute the HASE from the one-sided autocorrelation sequence, we first designed a suitable high dynamic range window function. This was necessary since the dynamic range of the magnitude spectrum of the autocorrelation sequence is equal to the dynamic range of the power spectrum of the time domain signal. To maintain the dynamic range after windowing, we need to use a window function on the autocorrelation sequence that has twice the dynamic range of the window function used on the time domain signal. To address this problem, we devised a novel window function design method,

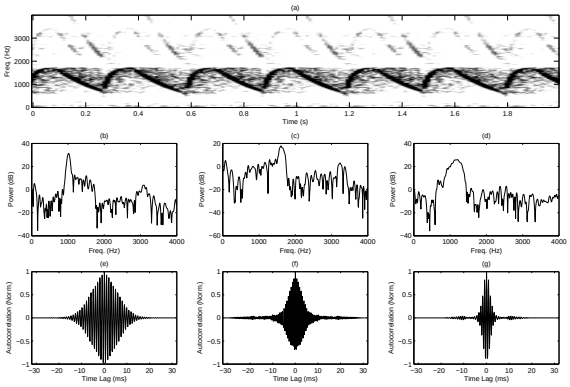

Fig. 2. Analysis of siren noise signal using 32 ms frames. (a) Spectrogram of a 2 second sample of siren noise. (b)(c)(d) The logarithmic power spectrum of frames taken at 0.5, 1.0 and 1.5 seconds respectively. (e)(f)(g) The autocorrelation sequences corresponding to the spectra in (b)(c)(d) respectively.

called Double Dynamic Range (DDR). DDR is an application specific alternative to the more complex and general window design methods such as Kaiser or Dolph-Chebyshev [7].

The proposed DDR window design method computes a window that has twice the dynamic range of a seed window, by computing the seed windows autocorrelation sequence. This technique also results in a side-lobe profile of the new window that matches the side-lobe profile of the seed window function. For example, we computed the DDR-Hamming window function used on the autocorrelation sequence in the following experiments, by finding the autocorrelation of a Hamming window.

A comparison of the cepstral smoothed spectral estimates computed using both the Periodogram and the HASE algorithm are shown in Fig.3. Here we have used a speech frame, an emergency vehicle siren frame and an artificial chirp noise frame to highlight the noise reduction properties of the HASE algorithm. First, we created plot (a) by computing the spectral estimate using both methods, then adjusted the gain of each so that the smooth spectral envelops magnitudes matched well. Keeping these gain settings, we then processed the two noise frames. For the siren noise case (plot (b)), the peak of the siren has been attenuated by 10 dB compared to the Periodogram method. In the artificial chirp noise case (plot (c)), approximately 40 dB attenuation is measured. If we assume that these noises are uncorrelated and additive with the speech signal, the HASE spectrum of a corrupt frame will have a higher spectral SNR than the corresponding Periodogram spectrum.

IV. Recognition experiments

In these experiments, we compared the noise robustness of the new speech recognition feature with MFCCs. For the evaluation, we used the Aurora II database, recognition scripts and the HTK software. We also used a range of stationary and non-stationary noise samples, which included Gaussian white noise, car noise, siren noise (Fig.2 and Fig.3(b)), and an

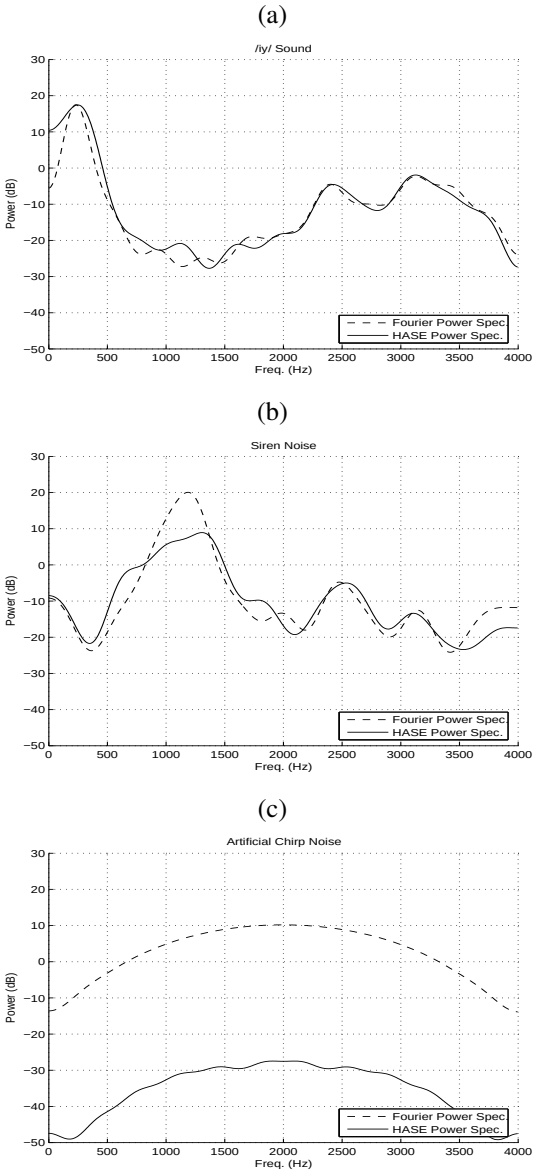

Fig. 3. Comparison of spectra computed using both the Fourier method and the HASE method. (a) Voiced speech frame. (b) Emergency vehicle siren noise. (c) Artificial chirp noise.

artificial chirp noise (Fig.3(c)), which repeatedly sweeps from 0 to 4 kHz in every 32 ms.

Recognition accuracy curves for the four noise cases are shown in Fig.4. These results show that the AMFCC features performed as well as the MFCC features in clean conditions. Secondly, these results show that the AMFCC features are more noise robust than the MFCC features in all the tested cases. The extent of the robustness improvement shown by the AMFCCs appears to be dependent on the type of noise. The car noise case displayed the least improvement, and the artificial chirp noise case showed the most improvement.

The artificial chirp noise case shows a dramatic improvement in noise robustness for AMFCCs over MFCCs. This type of signal produces large magnitude lower-lag autocorrelation coefficients and very low magnitude higher-lag coefficients over a short analysis window. This explains the large improve-

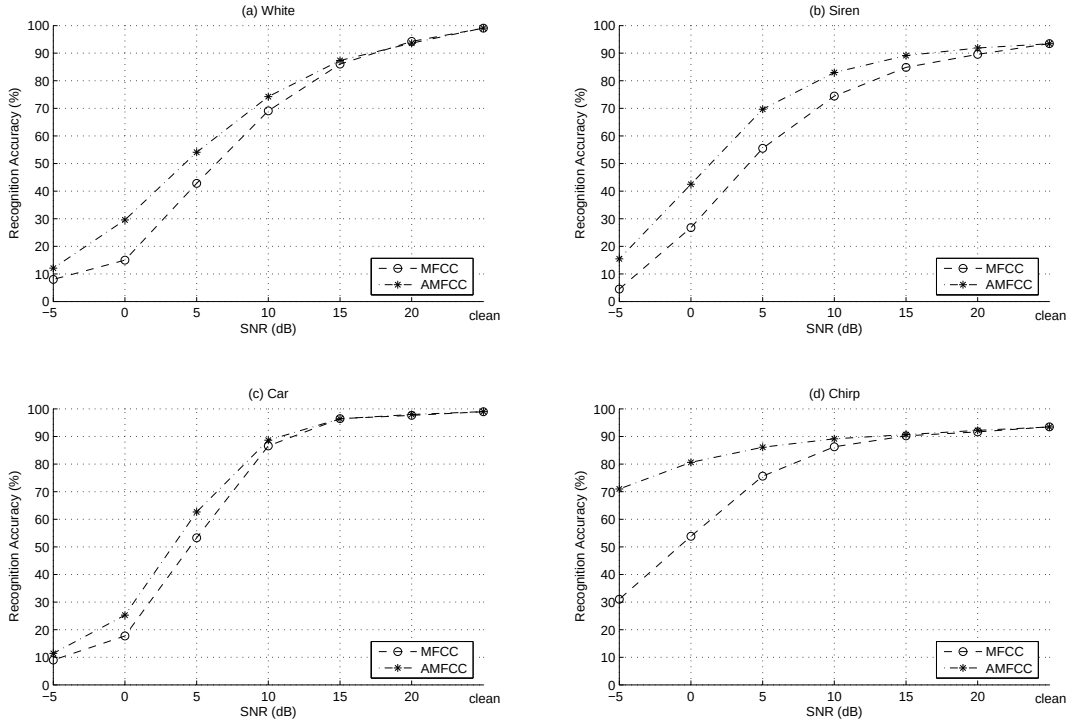

Fig. 4. Recognition accuracy results from the Aurora II database for MFCC and AMFCC features. (a) White Gaussian noise. (b) Emergency vehicle siren noise. (c) Car noise. (d) Artificially generated chirp noise.

ment for AMFCCs for these types of noise.

V. Conclusions

In this paper, we introduced a new noise robust spectral estimation technique for speech signals. This method was computed as the magnitude spectrum of the windowed one-sided higher-lag autocorrelation sequence, which we referred to as HASE.

During the development of the HASE method, we introduced a new high dynamic range window function design approach, which we called DDR. This technique was proposed specifically to design windows to be used in the autocorrelation domain. The DDR method involved computing the high dynamic range window as the autocorrelation sequence of a seed window function that had half the desired dynamic range.

The HASE method was combined with the MFCC algorithm to produce speech recognition features called AMFCCs. On the Aurora II database, the AMFCC features gave higher recognition accuracy scores than MFCCs over a range of SNRs using both stationary and non-stationary noises. The AMFCC features also matched the clean condition performance of MFCC features.

References

[1] 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.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," in *Proc. of the IEEE*, Jan. 1978, vol. 66, pp. 51–83.