SPECTRAL ESTIMATION USING HIGHER-LAG AUTOCORRELATION COEFFICIENTS WITH APPLICATIONS TO 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

In this paper, we introduce a noise robust spectral estimation technique for speech signals that is derived from a windowed one-sided higher-lag autocorrelation sequence. We also introduce a new high dynamic range window design method, and utilise both 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 in noisy conditions.

1. INTRODUCTION

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].

A potential source of error in using LP methods to estimate the power spectrum of a varying SNR signal is highlighted by Kay [6]. Kay showed that the model order is not only dependent on the AR process, but also on the prevailing SNR condition. Therefore, in this paper, we do not use an LP based method to process the autocorrelation sequence. Instead, we compute the magnitude spectrum of the one-sided higher-lag autocorrelation sequence using the Fourier transform, 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).

Speech recognition feature extraction algorithms are typically designed assuming stationary broadband (usually white) noise. In this work, we consider stationary noise signal as well as non-stationary noises, such as emergency vehicle sirens and chirp signals. We show that higher-lag autocorrelation processing is robust against these types of noise disturbances.

The paper organisation is as follows. In section 2 we discuss some properties of the autocorrelation sequence in relation to speech and noise signals showing examples. We then describe, in section 3, 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 4. This is then followed by conclusions in section 5.

2. PROPERTIES OF AUTOCORRELATION SEQUENCES

In this section, we demonstrate briefly how the smooth spectral envelope information of a voiced speech signal is distributed within its short-time autocorrelation sequence. We then discuss the autocorrelation distribution for noise signals giving an example of a non-stationary noise.

2.1. Speech Signals

In automatic speech recognition, we model the human speech production system using a simple source-system model. The model consists of a variable response filter, excited by either a white noise source or a periodic pulse train source. We model unvoiced speech as the output of the variable response filter excited by the white noise source and voiced speech as the output of the variable response filter excited by the periodic pulse train. For speech recognition, we are typically interested in extracting the magnitude response of the variable response filter over time. We assume that this carries the speech information sufficiently for accurate recognition.

Most of the popular speech recognition features, such as LPCCs and MFCCs, are derived from an estimate of the smooth power spectrum of the speech signal. We can consider the smooth power spectrum in both of these cases

0-7803-9243-4/05/$20.00 ©2005 IEEE

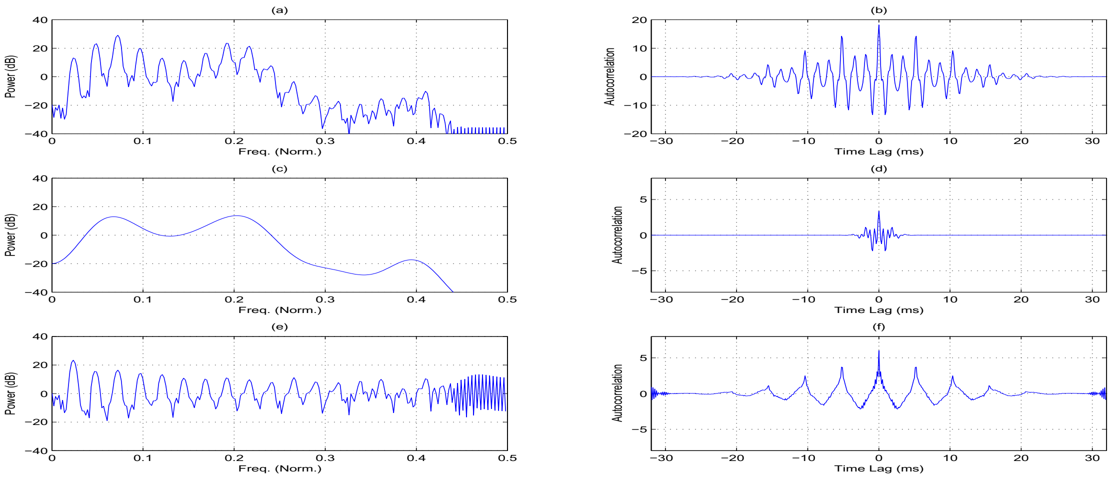

Fig. 1. Decomposition of a 32 ms voiced speech frame, containing an /r/ sound. (a) The original logarithmic power spectrum. (b) Autocorrelation sequence associated with the spectrum in (a). (c) The smooth logarithmic spectral envelope computed by retaining the first 12 cepstral coefficients. (d) The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

as being computed from the autocorrelation sequence. In the LPCC algorithm, the smooth spectral estimate is computed from the first few autocorrelation coefficients, and in the MFCC algorithm, the smooth spectral estimate is computed using the whole autocorrelation sequence. 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 Fig.1(a). This 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). The corresponding autocorrelation sequences of these two spectrums are shown in (d) and (f), 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) can be considered as the convolution of the autocorrelation sequences in (d) and (f). This process 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.

2.2. 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. Some general comments about autocorrelation sequences are made below.

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. Generally, the envelope decays when moving away from the zero lag coefficient. Some of the decay can be attributed to the biased autocorrelation estimation algorithm, 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 spectrums 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. The lower-lag coefficients can be significantly attenuated by using a tapered window

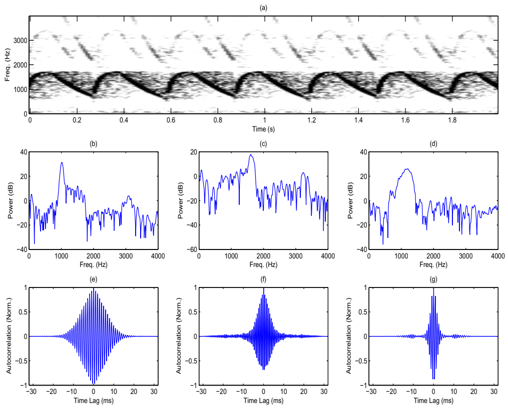

Fig. 2. Analysis of siren noise signal using 32 ms frames. (a) Spectrogram of a 2 second sample of siren noise. (b)(c)(d) The logarithmic power spectrum of frames taken at 0.5, 1.0 and 1.5 seconds respectively. (e)(f)(g) The autocorrelation sequences corresponding to the spectrums in (b)(c)(d) respectively.

function. This also has the added effect of attenuating the very high-lag coefficients, which have high estimation variance.

3. SPECTRAL ESTIMATION FROM HIGHER-LAG AUTOCORRELATION

Based on the previously discussed motivation, we compute a spectral estimate as the magnitude spectrum of the windowed one-sided autocorrelation sequence. A new speech recognition feature is then computed by substituting the new spectral estimate for the power spectrum in the MFCC algorithm.

To compute the new spectral estimate from the one-sided autocorrelation sequence, we first designed a suitable high dynamic range window function. Since the dynamic range of the magnitude spectrum of the autocorrelation sequence is the same as the dynamic range of the power spectrum of the time domain signal, we need to use a window function on the autocorrelation sequence that has twice the dynamic range of the window function that is normally used on the time domain signal. We devised a novel window function design method for this application as an alternative to more complex general design methods such as Kaiser or Dolph-Chebyshev.

A window function that has twice the dynamic range of a seed window function can be computed as the autocorrelation of the seed window. This technique also results in a side-lobe profile of the new window that matches the side-lobe profile of the seed window function. In the following experiments, the window function used on the autocorrelation sequence was computed as the autocorrelation of a Hamming window.

4. 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 used a range of stationary and non-stationary noise samples, which included Gaussian white noise, car noise, siren noise (as featured in Fig.2), and an artificial chirp noise, which repeatedly swept from 0 to 4 kHz in 32 ms.

Recognition accuracy curves for the four noise cases are shown in Fig.3. 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 least improvement was displayed in the car noise case, and the most improvement was displayed in the artificial chirp noise case.

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 improvement for AMFCCs for these types of noise.

5. CONCLUSIONS

In this paper, we have 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.

We also introduced a new high dynamic range window function design approach. This technique is specifically suited to designing windows for the autocorrelation domain. This method involved computing the high dynamic range window as the autocorrelation of a seed window function used in the time domain.

The new spectral estimate was used in 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.

6. 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.

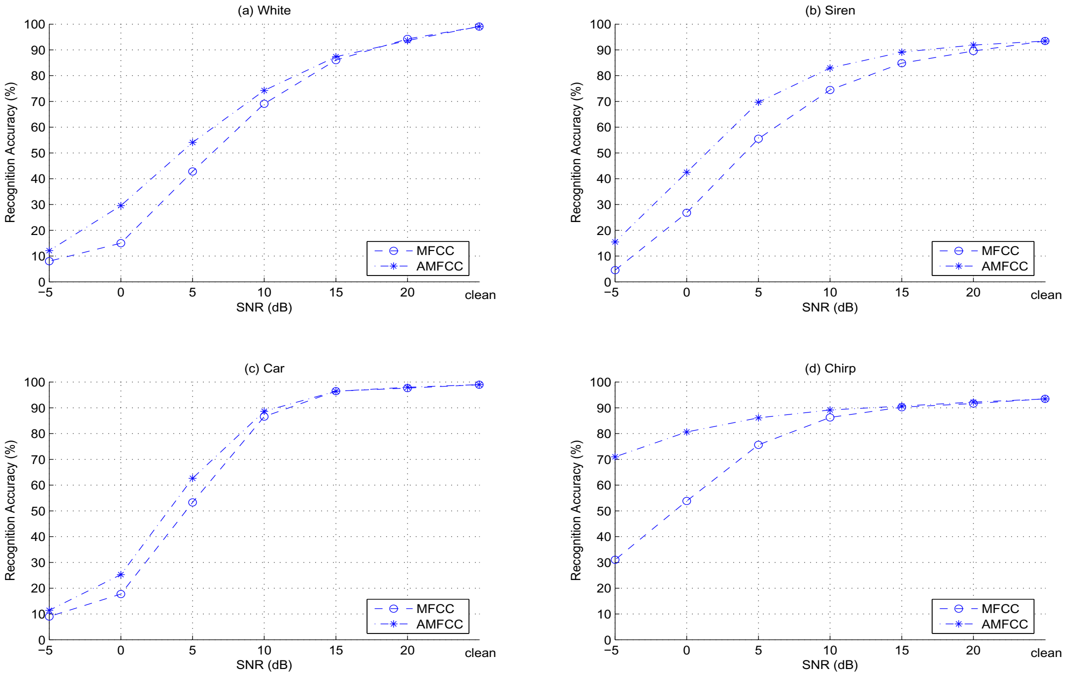

Fig. 3. Recognition accuracy results from the Aurora II database for MFCC and AMFCC features. (a) White Gaussian noise. (b) Emergency vehicle siren noise. (c) Car noise. (d) Artificially generated chirp noise.

[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.