INFLUENCE OF AUTOCORRELATION LAG RANGES ON 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

It is generally believed that the lower-lag autocorrelation coefficients carry information about the spectral envelop and the higher-lag autocorrelation coefficients are more related to pitch information. In this paper, we use lower-lag and higher-lag ranges of the autocorrelation function separately for deriving speech recognition features, and investigate their role in terms of speech recognition performance. The state-of-the-art MFCC features use the whole autocorrelation function in their computation and are used here as a benchmark in our experiments. Our recognition results from the Aurora II corpus show that the higher-lag autocorrelation coefficients perform as well as the whole autocorrelation function for clean speech, and provide better performance for noisy speech, while lower-lag autocorrelation coefficients are not as effective in this aspect.

1. INTRODUCTION

The Mel Frequency Cepstral Coefficient (MFCC) features have become a de facto standard feature in current speech recognition technology. These features are derived from the speech signal in terms of the following steps: 1) compute the short-time power spectrum (through FFT algorithm), 2) apply a Mel filter bank to get energies in the individual filter channels and , 3) take DCT of the logarithm of the resulting filter bank energies. The power spectrum used in this procedure can be interpreted as a Fourier transform of the whole autocorrelation sequence. Linear Prediction Cepstral Coefficient (LPCC) features are another feature set that has also been widely used for speech recognition in the past. This feature set uses the first few autocorrelation coefficients in its computation. These MFCC and LPCC features have been developed on the basis of our understanding that the lower-lag range of the autocorrelation function is mainly useful for speech recognition, while the higher-lag range is more relevant for pitch information.

In this paper, we investigate the relative contribution of the lower-lag and higher-lag ranges of the autocorrelation function for robust speech recognition. The Aurora II corpus is used to carry out speech recognition experiments. Our results show that higher-lag autocorrelation coefficients describe smooth spectral envelop as well as the whole autocorrelation function, and also provide increased robustness to noise; while lower-lag autocorrelation coefficients are not as effective in this aspect.

This paper is organised as follows. In section 2, the specific algorithms for the four different speech feature sets used in the study are introduced. Following this in section 3, the experimental framework is described, along with the results and discussion. Finally conclusions are given in section 4.

2. AUTOCORRELATION DERIVED FEATURES

To evaluate the effect of different ranges of the autocorrelation function on a speech feature set's robustness to noise, four different feature sets are investigated. These include Linear Prediction Cepstral Coefficients (LPCCs), Mel Frequency Cepstral Coefficients (MFCCs), and our newly proposed Higher and Lower lag Autocorrelation Mel Frequency Cepstral Coefficients (HL-AMFCCs, LL-AMFCCs) [4].

LPCCs, in comparison, use the fewest coefficients of the features in the study, with only 13 lower-lag autocorrelations (order 12 model). MFCCs use 256 unique autocorrelations for a 32 ms frame sampled at 8 kHz, and Lower and Higher lag AMFCCs use 24 and 232 coefficients respectively for the same 32 ms / 8 kHz system. A diagram comparing the regions of autocorrelation used in each of the four different features is shown in Fig. 1. Each of the features used in the study are introduced next, along with the proposed AMFCC features.

2.1. Linear Prediction Cepstral Coefficients (LPCC)

Beginning with the speech signal, 32 ms speech frames are formed that overlap by 22 ms. A Hamming window is applied to each of these frames, before a biased autocorrelation estimate is made. Using these autocorrelation co-
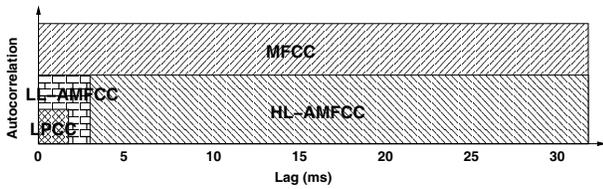
Fig. 1. Comparison of Autocorrelation Ranges in Speech Features.

efficients, the Yule-Walker equations are solved using the Levinson-Durbin algorithm, then converted to cepstral coefficients using a recursion relation [1][2]. A block diagram of the LPCC algorithm is shown in Fig. 2.
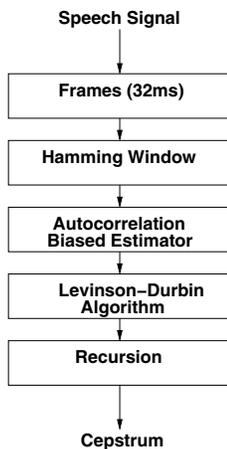
Fig. 2. LPCC block diagram.

The LPCC feature set is the only one out of the four feature sets being compared that does not employ a perceptually motivated warped frequency axis (eg. Mel scale). Some performance degradation may be attributed to this fact.

2.2. Mel Frequency Cepstral Coefficients (MFCC)

The MFCC feature extraction algorithm starts in the same way as the LPCC analysis. The speech signal is broken into 32 ms Hamming windowed time frames, which overlap by 22 ms. The power spectrum of the windowed time frames (computed through FFT algorithm) is then found before a filter bank is applied. In this analysis, a 23 channel Mel warped filter bank is applied to the estimated power spectrum as done in [3]. The resulting filter bank energies are converted to cepstral coefficients by taking the discrete cosine transform (DCT) of their logarithm values, then retaining 12 cepstral coefficients after discarding C0. Figure 3 shows the MFCC feature block diagram.
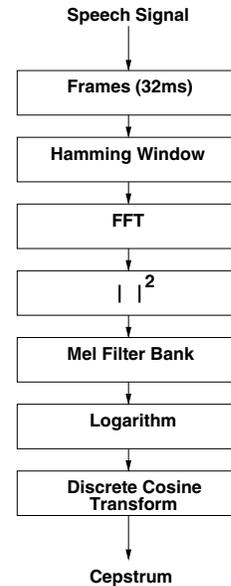
Fig. 3. MFCC block diagram.

2.3. Autocorrelation Mel Frequency Cepstral Coefficients (AMFCC)

The Autocorrelation Mel Frequency Cepstral Coefficients (AMFCCs) are proposed here as features for speech recognition, motivated by the assumption that higher-lag autocorrelation coefficients are less effected by noise than the original signal [4][5][6]. The algorithm for computing the feature proceeds as follows.

The speech signal is broken into 32 ms Hamming windowed overlapping time frames as with the previous two methods. An unbiased autocorrelation sequence is then computed for each frame. Of the autocorrelation sequence from each frame, only a desired region is retained for further processing. In these experiments, two regions are used. 1) The 0 ms to 3 ms lag coefficients are retained for computing the Lower-Lag AMFCC (LL-AMFCC) features, and 2) The 3 to 32 ms region is retained for computing the Higher-Lag AMFCC (HL-AMFCC) features. For LL-AMFCC computation, the symmetry property of the autocorrelation function is used to extend the range of autocorrelation lags from -3 to 3 ms.

A Kaiser window with high side-lobe attenuation is next applied to the extracted coefficient lags. This is necessary since the dynamic range of the autocorrelation sequence is twice the dynamic range of the original time sequence, as discussed in [7][8]. In AMFCCs, the Kaiser window function in Eq.(1) is used, where α is set to 10. Also due to the dynamic range of the autocorrelation sequence, the magnitude spectrum of the Kaiser windowed autocorrelation coefficients is found. From this step onwards, the algorithm is

the same as MFCCs as shown in Fig. 4.

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

Speech Signal

Frames (32ms)

Hamming Window

Autocorrelation Unbiased Estimator

Lower Lag   **Higher Lag**

Kaiser Window −3ms < n < 3ms   **Kaiser Window 3ms < n < 32ms**

FFT

| |

Mel Filter Bank

Logarithm

Discrete Cosine Transform

Cepstrum
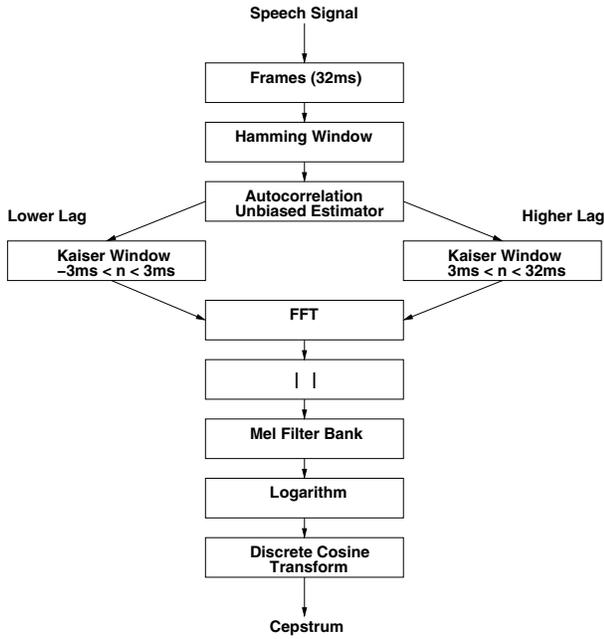
Fig. 4. AMFCC block diagram.

3. RECOGNITION EXPERIMENT

3.1. Speech database

To evaluate the importance of different regions of the autocorrelation sequence in regards to noise robustness, we used the Aurora II database, Aurora II experiment scripts, and HTK software [1]. 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.

In 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.

---

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

3.2. Results

Recognition accuracy curves for subway, babble, car and exhibition noise can be seen in Fig. 5, 6, 7 and 8, respectively. The first thing to note from these results is that all the features performed well in the uncorrupted case, regardless of the range of autocorrelation used in their computation. This result is significant since it demonstrates that all regions of the autocorrelation sequence convey information about the power spectral envelop of the speech signal.
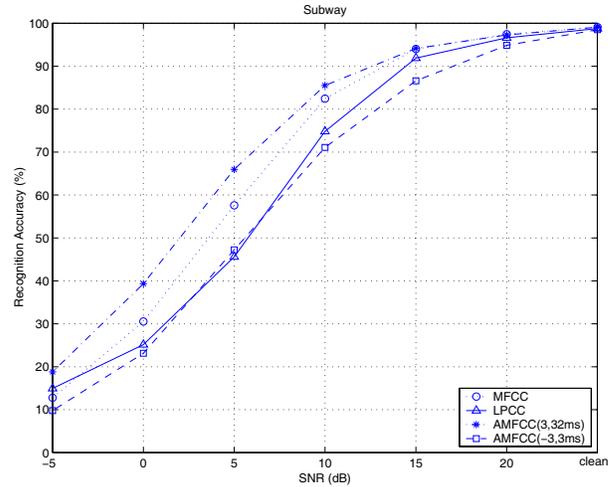
Fig. 5. Subway noise.
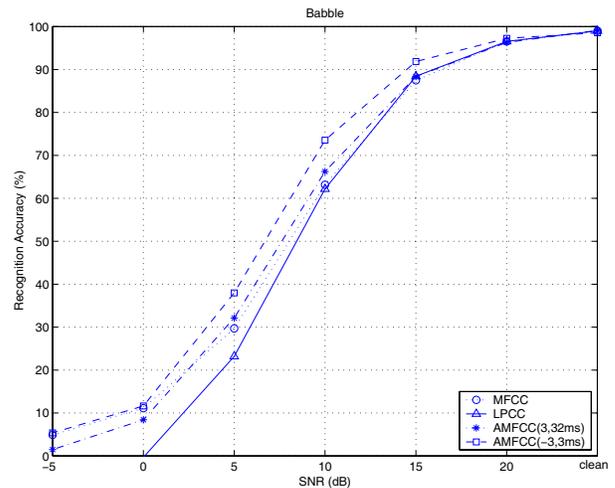
Fig. 6. Babble noise.

In three out of the four noise cases, the features that were either derived exclusively from higher-lag autocorrelation range (HL-AMFCC) or the whole autocorrelation function (MFCC) (which is dominated by higher-lag autocorre-
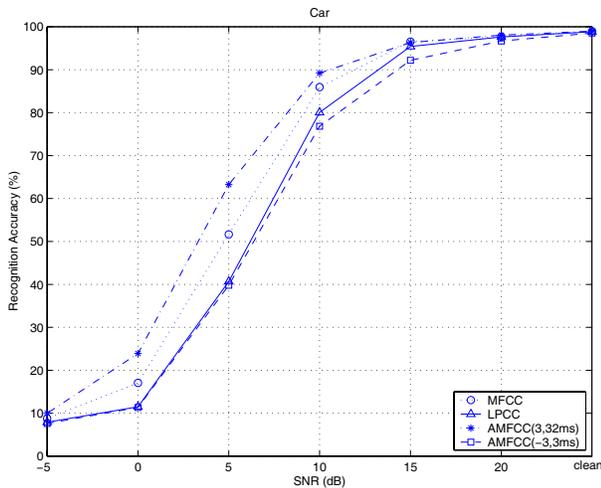
Fig. 7. Car noise.
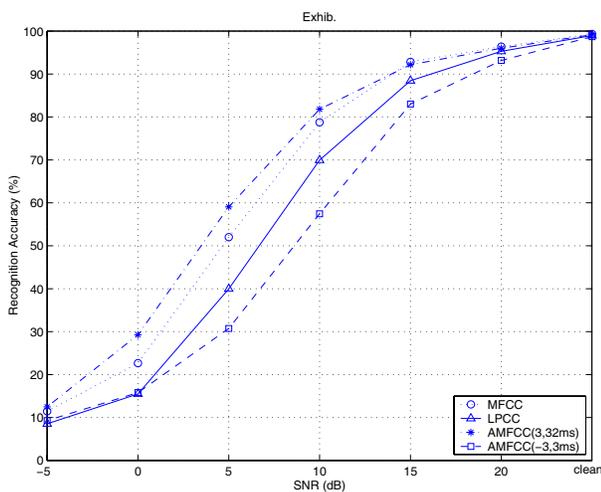
Fig. 8. Exhibition noise.

lation coefficients) displayed a higher noise robustness. In the babble noise case, the advantage was either lost (HL-AMFCC vs. LL-AMFCC) or became negligible (MFCC vs. LPCC). This suggests that the contribution to noise robustness that different autocorrelation regions make is a function of the noise signal's autocorrelation function. These experiments show that speech recognition features that are derived from higher-lag autocorrelation coefficients are more robust than features that use all lags (MFCC) for all tested noise types. They also show that higher-lag derived features are more robust than lower-lag features for most of the tested noise types (3/4).

4. CONCLUSIONS

In this paper, several features that are derived from different ranges of the autocorrelation sequence are evaluated for their robustness to noise for a speech recognition task. It is shown that all regions of the autocorrelation sequence (higher-lag as well as lower-lag) produce features that yield high recognition accuracy in clean conditions. It is also shown that features that are derived from the higher-lag range of the autocorrelation function are always more robust to noise than features that are derived from the whole autocorrelation function. In addition, in three out of four noise cases, these features are more noise-robust than the feature derived from the lower-lag autocorrelation coefficients.

5. REFERENCES

[1] J. Makhoul, "Spectral analysis of speech by linear prediction," *IEEE Transactions on Audio Electroacoust.*, vol. 21, pp. 140–148, June 1973.

[2] J. Makhoul, "Spectral linear prediction, properties and applications," *IEEE Trans. Acoustics, Speech and Signal Processing*, pp. 283–296, June 1975.

[3] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-28, no. 4, pp. 357–365, Aug. 1980.

[4] B. J. Shannon and K. K. Paliwal, "Mfcc computation from magnitude spectrum of higher lag autocorrelation coefficients for robust speech recognition," in *Accepted to Proc. ICSLP*, 2004.

[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] 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.

[7] 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.

[8] 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.