EFFECT OF SPEECH AND NOISE CROSS CORRELATION ON AMFCC SPEECH RECOGNITION FEATURES

Benjamin J. Shannon and Kuldip K. Paliwal

Signal Processing Laboratory
Griffith University
Brisbane, QLD 4111, Australia
Ben.Shannon@student.griffith.edu.au, K.Paliwal@griffith.edu.au

ABSTRACT

When designing noise robust speech recognition feature extraction algorithms, it is common to assume that the noise and speech signal are uncorrelated. This assumption allows the cross correlation terms to be ignored in the equations that describe the operation of these algorithms, thus making the mathematics more tractable. In this paper, we investigate the validity of this assumption in the context of the Autocorrelation Mel Frequency Cepstral Coefficient (AMFCC) feature extraction algorithm. To carry out the investigation, we designed a modified AMFCC algorithm that forces the cross terms in the noisy signal autocorrelation equation to be zero. We then compared the performance of the modified algorithm to the un-modified algorithm in recognition experiments performed using the AURORA II database. From these evaluations, we show that the assumption is fair in 5 out of six tested noise cases. The difference in recognition accuracy between the AMFCC and modified AMFCC for these five noises was less than 5%.

Index Terms— Robust speech recognition, feature extraction, autocorrelation function

1. INTRODUCTION

For Automatic Speech Recognition (ASR) systems to be practical they need, among other factors, a level of robustness to changes in the speaking environment. One of the environmental changes that has a large impact on the performance of current ASR systems is background noise. There are several approaches that one can take to improve an ASR systems robustness to changes in background noise. One of these approaches is to address the problem at the feature extraction stage of the system. That is, to use a speech feature extraction algorithm that produces features that are as invariant as possible to background noise changes, while simultaneously capturing the salient speech information.

Many feature extraction algorithms have been proposed that are designed specifically to have a low sensitivity to background noise [1, 2, 3, 4, 5, 6, 7]. When designing such algorithms, assumptions are typically made about the noise signal. This is generally done to make the mathematics more tractable. One such assumption that is commonly assumed is that the speech signal and disturbing noise signal are uncorrelated. The impact of this assumption on algorithms that perform processing in the autocorrelation domain can be described as follows. Assume we have a short-time segment of uncorrupted speech $s(n)$ and a noise signal $d(n)$ that are both wide-sense stationary. A corrupt speech signal $x(n)$ can then be formed if we assume $d(n)$ is an additive noise, therefore $x(n) = s(n) + d(n)$. If we assume nothing further, the autocorrelation of the noisy signal $x(n)$ is,

$$R_{xx} = R_{ss} + R_{dd} + \underbrace{R_{sd} + R_{ds}}_{\text{cross terms}} \tag{1}$$

From this equation, we can see that the autocorrelation is composed of R_{ss}, which is the autocorrelation of the clean speech signal alone, R_{dd}, which is the autocorrelation of the noise signal alone and two cross correlation terms. Typically we further assume that the speech signal and noise signal are uncorrelated. This has the effect of removing the cross correlation terms from the autocorrelation of the noisy speech signal. Therefore, the autocorrelation of the noisy speech signal simplifies to,

$$R_{xx} = R_{ss} + R_{dd} \tag{2}$$

If we use the uncorrelated assumption (2) we can proceed to design feature extraction algorithms by considering the properties of the speech signal and noise signal in isolation. For example, if the speech signal gives non-zero coefficients at all lags in the autocorrelation domain and the noise signal gives coefficients concentrated in the lower-lags, we can design a noise robust feature extraction algorithm by using only the higher-lags of the autocorrelation sequence to compute the speech features. This view of the signal and noise relationship in the autocorrelation domain is what motivated us to design the Autocorrelation Mel Frequency Cepstral Coefficient (AMFCC) [1] method for feature extraction.

In this paper, we test the assumption that the cross correlation between a speech signal and noise signal is negligible enough that its effect can be ignored when designing noise robust speech recognition features. To test this assumption we use two versions of the AMFCC algorithm. One version is the un-modified algorithm that has all the cross correlation terms included and the other version is a modified algorithm that has all of the cross correlation terms removed. We also designed an artificial noise that can be described as a repeating chirp noise. If the cross terms in the un-modified AMFCC algorithm are zero, then the AMFCC algorithm should be immune to this noise signal. Five other more typical noises are also included in the evaluation. These include four natural noises from the AURORA II database (subway, babble, car and exhibition) and artificial white Gaussian noise.

An outline of the paper is as follows. In Section 2, we briefly describe both the un-modified AMFCC algorithm and the modified AMFCC algorithm that we use in the experiments. Section 3 gives an analysis of the artificial chirp noise, and finally in Sections 4 and 5 we finish with the recognition experiment results and conclusions.

2. AUTOCORRELATION MEL FREQUENCY CEPSTRAL COEFFICIENTS

The Autocorrelation Mel Frequency Cepstral Coefficients (AMFCC) were proposed as noise robust features for speech recognition. In proposing this algorithm, we were motivated by the assumption that higher-lag autocorrelation coefficients are less effected by noise than the original signal [1, 2, 3, 8]. In this investigation, we use two versions of the AMFCC algorithm. Both of these versions are depicted in the block diagram shown in Fig. 1. The top path shown in Fig. 1 we refer to as the "all cross terms" path and the path shown below that we refer to as the "zero cross terms" path.

The "all cross term" path is the algorithm that has been proposed in [1]. This algorithm uses the full autocorrelation expression shown in (1) and assumes the cross terms are zero. This algorithm has a lot of steps in common with the MFCC algorithm [9]. The main difference can be found in the method of estimating the speech spectrum. The MFCC algorithm typically uses the squared Fourier magnitude spectrum as the estimate of the signals power spectrum. In the case of AMFCC, we first compute the autocorrelation coefficients. We then apply a high dynamic range (86 dB) window function to the higher-lag coefficients (2 to 32 ms lag). And finally, we compute the magnitude spectrum of the resulting sequence as an estimate of the power spectrum of the clean speech signal. For further details of the algorithm, please refer to [1].

It is apparent from the block diagram that the "zero cross terms" path is not realisable in a practical situation. This path requires the speech and noise signals to be separate up until they are combined after the autocorrelation step. Since we are combining the speech and noise artificially in this evaluation, we are free to realise this path. This method results in the cross correlation components being forced to zero, as was previously discussed.

3. ARTIFICIAL CHIRP NOISE

For this evaluation, we needed a noise signal that produced high magnitude lower-lag short-time autocorrelation coefficients and very low magnitude higher-lag coefficients. If we use a noise signal with these characteristics and the zero cross correlation assumption was good, then the AMFCC algorithm should be immune to this noise.

We identified three basic signal types that give a delta function like short-time autocorrelation sequence. These were 1) ideal white noise, 2) a pulse train where the separation between the pulses is greater than the analysis window width and 3) a repeating chirp noise signal. For this evaluation we chose to use the repeating chirp signal to design an ideal noise.

The artificial chirp noise designed for testing the AMFCC algorithm has a period equal to the window size used on the speech signal (32 ms). One period of this noise is generated as a sinusoidal signal whose frequency changes linearly from zero to half of the sampling frequency over the period. An analysis of this noise is presented in Fig. 2. These plots show that this type of noise gives the desired high magnitude lower-lag coefficients and very low magnitude higher-lag coefficients.

4. RECOGNITION EXPERIMENTS

4.1. Speech database

In these experiments we measured the noise robustness of the un-modified AMFCC algorithm and the modified AMFCC algorithm in the same conditions. To conduct this evaluation, we used speech from the Aurora II database, the Aurora II experiment scripts, and HTK software [1]. We also tested the baseline MFCC features as a reference.

All the experiments conducted used clean training speech. The noisy test speech was generated by corrupting the clean test speech from the subway noise case. To set the signal-to-noise ratio (SNR) of the noisy test speech, we adjusted the level of the noise sample so the global SNR after degradation was the desired value. Each recognition experiment was repeated using six different noises at seven different SNRs. The six noises were artificial repeating chirp noise, Gaussian white noise and the four AURORA noises, subway, babble, car and exhibition. The seven SNRs were -5 dB to 20 dB in 5 dB steps and clean.

In these experiments, the speaker-independent word models had 16 emitting states. The modelled acoustic feature vec-

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

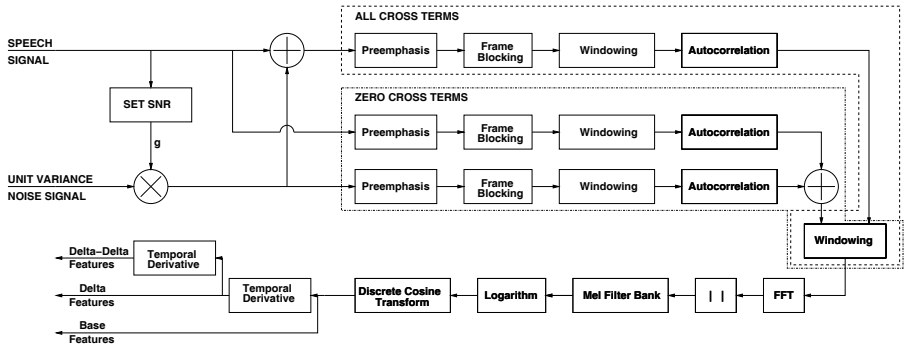

Fig. 1. Block diagram showing how the noisy speech signal is generated and processed using the two AMFCC algorithms. The box labelled "ALL CROSS TERMS" shows the typical implementation of the AMFCC algorithm. The box labelled "ZERO CROSS TERMS" shows the modified algorithm that eliminates the cross terms from the autocorrelation sequence. The "ZERO CROSS TERMS" path is not realisable in a practical system since it requires exact knowledge of the clean speech signal and the noise signal separately. We can implement it in an artificial framework though, since we do have exact knowledge of both of these signals.

tor was composed of a 12 dimensional base feature. These did not include a logarithmic energy coefficient or zero-th cepstral coefficient. These base features were then concatenated with delta and acceleration coefficients to produce a 36-dimensional feature vector.

4.2. Results

The results from the experiments are shown in Fig. 3. These plots compare the three tested features. The AMFCC-X curves are the artificial AMFCC features where the cross correlation has been forced to zero. The AMFCC and MFCC features are the un-modified algorithms.

From the chirp noise results, there is a clear difference between the AMFCC and AMFCC-X features. In this case, the AMFCC-X features show near ideal behaviour. The AMFCC features still show a very significant improvement over MFCC features.

The other five noise cases show the AMFCC and AMFCC-X performance to be very similar. In these cases, the AMFCC-X features performed better than the AMFCC features, but the improvement was generally less than 5%.

5. CONCLUSIONS

In this paper, we investigated the validity of the assumption that there is negligible cross correlation over a short-time between speech and noise signals. This assumption, which is commonly used during the design of noise robust speech recognition feature extraction algorithms, was shown to be fair for the AMFCC algorithm for five of the six tested noises. For the sixth noise, which was a periodic chirp signal, it was shown that the assumption was poor. We showed by using the Aurora II database and a modified AMFCC algorithm that the cross correlation between the speech and the chirp noise signal was strong enough to produce up to 30% difference in recognition accuracy over the normal AMFCC algorithm. This difference was contrasted with other natural noises where the recognition accuracy differences were less than 5%.

6. REFERENCES

[1] B. J. Shannon and K. K. Paliwal, "MFCC Computation from Magnitude Spectrum of Higher Lag Autocorrelation Coefficients for Robust Speech Recognition," in *Proc. ICSLP*, 2004.

[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, 1997.

[3] 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, 1989.

[4] Y. T. Chan and R. P. Langford, "Spectral Estimation via the High-Order Yule-Walker Equations," *IEEE Trans. on ASSP*, , no. 5, pp. 689–698, 1982.

[5] K. K. Paliwal and M. M. Sondhi, "Recognition of noisy speech using cumulant-based linear prediction analysis," in *Proc. ICASSP*, 1991, pp. 429–432.

[6] K. K. Paliwal and Sagisaka, "Cyclic autocorrelation-based linear prediction analysis of speech," in *Proc. Eurospeech*, 1997, pp. 279–282.

[7] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environments," *Computer Language and Speech*, vol. 1, pp. 109–130, 1986.

[8] N. A. Anstey, "Correlation Techniques," *Canadian Journal of Exploration Geophysics*, vol. 2, no. 1, pp. 55–82, 1966.

[9] 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, 1980.

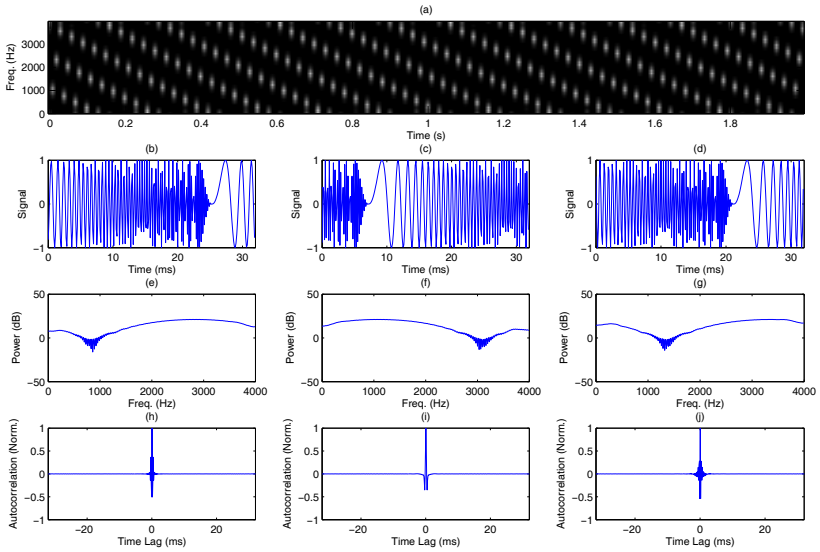

Fig. 2. Short-time analysis of the artificial periodic chirp noise signal using 32 ms frames. (a) Spectrogram of a 2 s sample of long noise signal, (b)(c)(d) Waveform of noise frames taken at 0.5, 1.0 and 1.5 s, respectively, (e)(f)(g) Power spectra (periodogram estimate with a Hamming window) of the frames shown in (b)(c)(d), respectively. (h)(i)(j) Autocorrelation sequences corresponding to the power spectra shown in (e)(f)(g), respectively.

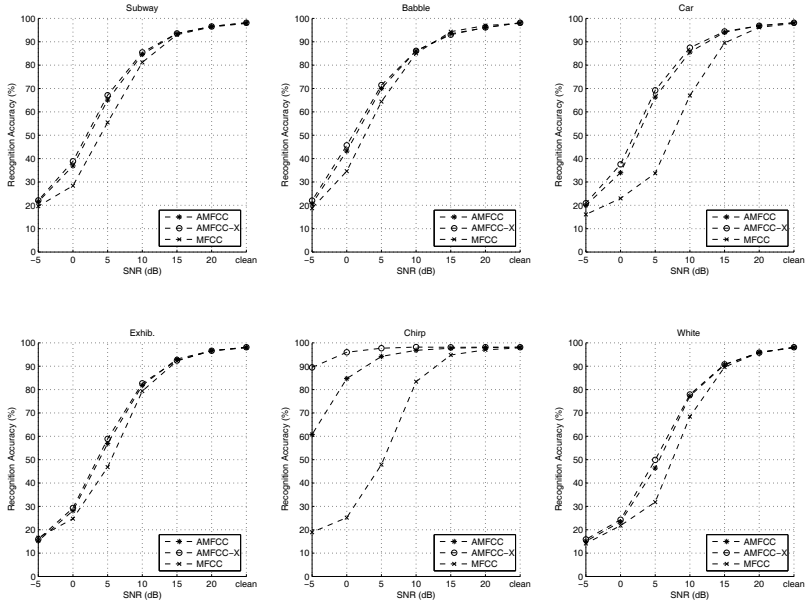

Fig. 3. Recognition accuracy results for six noises comparing the performance of AMFCC features (AMFCC), modified AMFCC features that have no cross correlation terms in the autocorrelation domain (AMFCC-X) and MFCC features.