Robust Features for Noisy Speech Recognition using MFCC Computation from Magnitude Spectrum of Higher Order Autocorrelation Coefficients

Dr. Amita Dev
Bhai Parmanad Institute of Business Studies,
New Delhi, India.

Poonam Bansal
Department of Computer Science and Engineering
Amity School of Engineering and Technology
New Delhi, India.

ABSTRACT
Noise robustness is one of the most challenging problem in automatic speech recognition. The goal of robust feature extraction is to improve the performance of speech recognition in adverse conditions. The mel-scaled frequency cepstral coefficients (MFCCs) derived from Fourier transform and filter bank analysis are perhaps the most widely used front-ends in state-of-the-art speech recognition systems. One of the major issues with the MFCCs is that they are very sensitive to additive noise. To improve the robustness of speech front-ends we introduce, in this paper, a new set of MFCC vector which is estimated through three steps. First, the relative higher order autocorrelation coefficients are extracted. Then magnitude spectrum of the resultant speech signal is estimated through the fast Fourier transform (FFT) and it is differentiated with respect to frequency. Finally, the differentiated magnitude spectrum is transformed into MFCC-like coefficients. These are called MFCCs extracted from Differentiated Relative Higher Order Autocorrelation Sequence Specrum (DRHOASS). Speech recognition experiments for various tasks indicate that the new feature vector is more robust than traditional mel-scaled frequency cepstral coefficients (MFCCs) in additive noise conditions.

KEYWORDS
MFCC, Autocorrelation domain, magnitude spectrum

1. INTRODUCTION
Acoustic features may greatly affect the performance of a speech recognizer. A great deal of work has been done for feature extraction [1]. In the literature, various approaches have been proposed to improve the tolerance of an ASR system with respect to noise, such as Wiener filtering [2], spectral subtraction [3], RASTA [4], lin-log RASTA [5], parallel model compensation (PMC) [6], vector Taylor series approximation based model compensation [7] etc. Noise robust spectral estimation using MFCC has been discussed in with the name of autocorrelation mel-frequency cepstral coefficients (AMFCC). As the autocorrelation of noise could in many cases be considered relatively constant over time, a high pass filtering of the autocorrelation sequence (RAS) , could lead to substantial reduction of the noise effect. Furthermore, it has been shown that preserving spectral peaks is very important in obtaining a robust set of features for ASR [8]. Methods such as peak-to-valley ratio locking [9] and peak isolation (PKISO) [10] have been found very useful in speech recognition error rate reduction. In the present paper, we present a novel technique of computing speech coefficients by using the magnitude spectrum of the relative one-sided higher-order autocorrelation sequence, differentiating it and then processing it through a Mel filter bank and finally parameterized it in terms of MFCCs. The paper organization is as follows. Section 2 gives a description of the newly proposed technique of feature extraction. Section 3 explains in detail the proposed method along with the block diagram. Finally, an experimental comparison of the proposed feature set with MFCCs is presented in section 4, followed by conclusion in section 5.

2. EXTRACTION IN AUTOCORRELATION DOMAIN
If u (m, n) is the additive noise, x (m, n) noise-free speech signal and h (n) impulse response of the channel, then the noisy speech signal y (m, n) can be written as:

$$y(m,n) = [x(m,n) + u(m,n)] \otimes h(n), \quad 0 \le m \le M-1 , \; 0 \le n \le N-1 \qquad (1)$$

Where M denotes the number of frames in an utterance and N denotes the number of samples in a frame and ⊗ denotes the convolution operation. As we intend to use our method to remove or reduce additive noise from noisy speech signal, therefore the channel effect will not be considered here. We will then have

$$y(m,n) = [x(m,n) + u(m,n)], \quad 0 \le m \le M-1 , \; 0 \le n \le N-1 \qquad (2)$$

If the noise is uncorrelated with the speech, it follows that the autocorrelation of the noisy speech y(m,n) is the sum of autocorrelation of the clean speech x(m,n) and autocorrelation of the noise u(m,n),i.e.

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{uu}(m,k), \quad 0 \le m \le M-1 , \; 0 \le k \le N-1 \qquad (3)$$

where ryy(m, k), rxx(m, k) and ruu(m, k) are the one-sided autocorrelation sequences of noisy speech, clean speech and noise respectively, and k is the autocorrelation sequence index within each frame. If the additive noise is assumed to be stationary, the autocorrelation sequence of noise can be considered to be identical for all frames. Hence, the frame index m can be dropped out, and equation (3) becomes

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{uu}(k), \quad 0 \le m \le M-1 , \; 0 \le k \le N-1 \qquad (4)$$

Eliminating the lower order of the noisy speech signal autocorrelation coefficients should lead to removal of the main noise components[11]. The maximum autocorrelation index to be removed is usually found experimentally, where D represents elimination threshold.

$$r_{yy}(m,k) = r_{yy}(m,k) \text{ if } D \le k \le N-1 \text{ and } r_{yy}(m,k) = 0 \text{ if } 0 \le k < D \qquad (5)$$

Differentiating the resultant autocorrelation sequence with respect to m, will give Relative Autocorrelation Sequence (RAS) of noisy speech at the mth frame[12]. In order to get DRHOASS, we take differentiation of the spectrum of the filtered signal .This further contributes to immunization against noise. By this approach the flat parts of the spectrum are almost removed while each spectral peak is split into two, one positive and one negative.

3. PROPOSED METHOD

This section describes our novel method to obtain new set of MFCC feature vectors. First, we pre-emphasis the input speech signal using a pre-emphasis filter $H(z) = 1- 0.97\ z^{-1}$. Then we perform frame blocking with a frame size of 16ms and a frame shift of 8 ms so that signal can be analyzed sequentially in a frame-wise manner. The Hamming window is applied to the pre-emphasized signal and then, the autocorrelation sequences of the framed signal are obtained. The lower lags of the autocorrelation sequence less than 1.375 ms (experimentally derived) are removed. A FIR high-pass filter is then applied to the signal autocorrelation sequence to further suppress the effect of additive noise. Then, the short-time Fourier transform of this filtered signal is calculated. In the next step, differential power spectrum of the filtered signal is found. Since the noise spectrum, in many occasions may be considered flat, in comparison to the speech spectrum, the differentiation either reduces or omits these relatively flat parts of the spectrum, leading to even further suppression of the effect of noise. A set of cepstral coefficients (DRHOASS-MFCC) are finally derived from the magnitude of the differentiated high order relative autocorrelation power spectrum by applying it to a conventional mel- frequency filter-bank and passing the logarithm of the output to a DCT block. MFCC vector set of dimension 39 is formed by concatenating energy, Delta MFCC and Delta- Delta MFCC. Front-end for extraction of MFCC vector set by DRHOASS has been shown in Figure. 1.

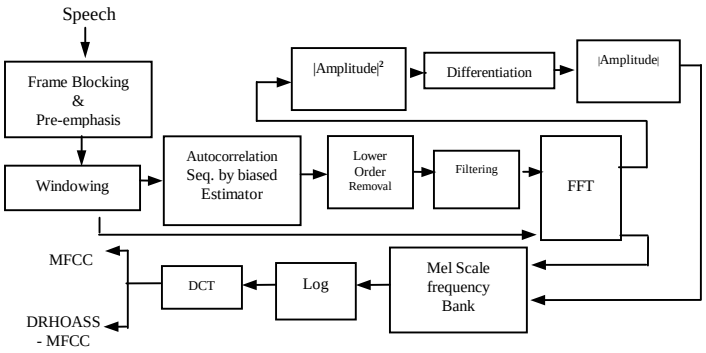

Figure 1. Block diagram for extracting MFCC by DRHOASS

4. RECOGNITION EXPERIMENT

The proposed approach was implemented on TIFR Hindi speech database of 200 Hindi words, spoken by 30 speakers. The spoken samples were recorded by 15 male, 10 female and 5 child speakers (5 repetitions) in a studio environment condition using Sennheiser microphone model MD421 and a tape recorder model Philips AF6121.Database was divided into training set and testing set. Features vector sets of size 39 are extracted using different front-ends (MFCC (for comparison purposes), RAS-MFCC, AMFCC and our method DRHOASS-MFCC) and their performances are compared.

a) Testing on Clean Speech

This experiment is to evaluate the performance of MFCC, RAS-MFCC, AMFCC and DRHOASS-MFCC, when training data & the testing data are in clean (40 dB) environment. The results are shown in Table 1. These are the baseline results for comparison purposes. Performance on the basis of recognition rates is observed to be more or less same if we use either MFCC, RAS-MFCC, AMFCC or DRHOASS-MFCC. This shows that the spectral information derived by DRHOASS method captures the speech information to the same extent as that by other method.

b) Testing on Noisy Speech

The polluted testing utterances are generated by adding the artificial noises at five SNR levels. The white noise is generated by using a random number generation program, and other colored noises, i.e., factory noise, F16 noise, and babble noise, are extracted from the NATO RSG-10 corpus [13]. The noises are added to the clean speech signal at 20, 15, 10 5 and 0 dB SNRs. Figure 2(a)-(d) shows the results obtained using MFCC, RAS-MFCC, AMFCC and DRHOASS front-ends. For the case of speech sounds corrupted by white noise as shown in Fig. 2(a), the performance of MFCC degrades most significantly among all features, and found to be worst among RAS-MFCC, AMFCC and DRHOASS-MFCC. As evident from Figure 2(a) DRHOASS-MFCC are quite robust to the additive noise.

Figures 2(b), (c) and (d) compares the performance of different front ends when the testing speech is corrupted by factory, babble, and f16 noise respectively. The figures depict that the performance of MFCC degrades significantly as compared to other feature vectors. The best performance comes from DRHOASS-MFCC. Improvement in recognition score of 5.62% at 20dB, 10.14% at 15 dB, 12.55% at 10 dB, 15.87% at 5dB and 13.3% at 0dB was seen in comparison to RAS-MFCC. It was primarily due to the removal of lower order autocorrelation coefficients in DRHOASS-MFCC which was not in the case of RAS-MFCC. Similarly improvement in recognition scores of the tune of 6.92% at 20dB, 14.9% at 15 dB, 17.8% at 10dB, 23.2% at 5dB and 15.8% at 0dB was analyzed in comparison to AMFCC feature vector set. Although in both the feature vector sets (DRHOASS-MFCC and AMFCC) lower autocorrelation coefficients are discarded but due to the high pass filtering as the additional step in DRHOASS-MFCC, it showed a remarkable improvement.

Table 1. Comparison of clean-train and clean test recognition rates for various features

Feature Type	MFCC	AMFCC	RAS-MFCC	DRHOASS-MFCC
Recognition rate % at 40 dB	98.241	98.246	98.30	99.64

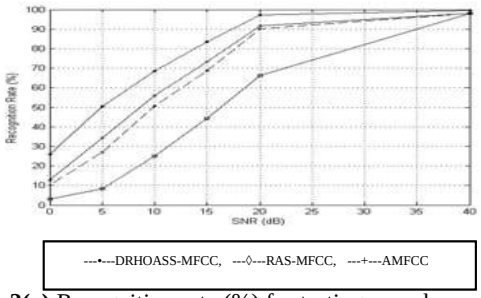

Figure 2(a) Recognition rate (%) for testing speech corrupted by white noise.

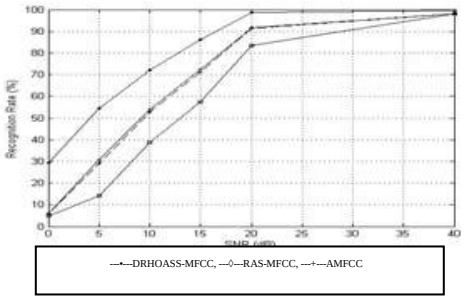

Figure 2(b) Recognition rate (%) for testing speech corrupted by factory noise.

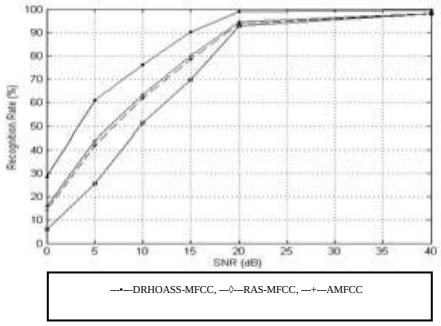

Figure 2(c) Recognition rate (%) for testing speech corrupted by babble noise.

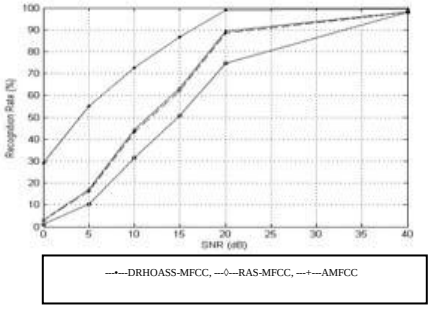

Figure 2(d) Recognition rate (%) for testing speech corrupted by F16 noise.

5. CONCLUSION

In this paper, we proposed a novel feature extraction technique using Differentiated Relative Higher Order Autocorrelation sequence spectrum (DRHOASS) for computing MFCC feature vector set. The DRHOASS- MFCCs showed remarkable increase in word recognition rate as compared to other traditional methods utilizing MFCC. It was found that higher order autocorrelation coefficients along with additional filtering improved the robustness of the speech recognition system under different background noises.

REFERENCES

1. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoustics, Speech Signal Process. 28, 357–366 (1980).
2. Vaseghi, S.V., Milner, and B.P.: Noise compensation methods for hidden Markov model speech recognition in adverse environments. IEEE Trans. Speech Audio Process. 5 (1), 11–21 (1997).
3. Boll, S.F.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoustics, Speech Signal Process. 27 (2), 113–120 (1979).
4. Hermansky, H., Morgan, N.: RASTA of processing of speech. IEEE Trans. Speech Audio Process. 2 (4), 578–589 (1994).
5. Hermansky, H., Morgan, N., Bayya, A., Kohn, P.: Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In: EUROSPEECH, Genova, p.p. 1367–1370 (1991).
6. Gales, M.J.F., Young, S.J.: Robust speech recognition using parallel model combination. IEEE Trans. Speech Audio Process. 4 (5), 352–359 (1996).
7. Moreno, P.J., Raj, B., Stern, R.M.: A vector Taylor series approach for environment independent speech recognition. In: ICSLP, Philadelphia, PA, pp. 733–736 (1996).
8. Padmanabhan, M. : Spectral peak tracking and its use in speech recognition. In: ICSLP (2000).
9. Sujatha, J., Prasanna K.R., Ramakrishnan, K. R., Balakrishnan, N.: Spectral maxima representation for robust automatic speech recognition. In: Eurospeech, pp. 3077-3080 (2003).
10. You, K.H., Wang, H.C.: Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences. Speech Communication, (28),13-24 (1999).
11. Strope, B., Alwan, A.: A model of dynamic auditory perception and its application to robust word recognitionIEEE Trans. on Speech and Audio Processing. 5 (5), 451-464 (1997).
12. Shannon B.J., Paliwal, K. K.: Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition. Speech Communication, 48(11), 1458-1485 (2006).
13. Varga, A., Steeneken, H.J.M.: Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun., 12, 247–251 (1993).