ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 4, No. 2, 2009, *pp*. 131-141

Novel Feature Vector Set Extraction using Spectral Peaks in Autocorrelation Domain

Poonam Bansal [1, +], Amita Dev [2], Shail Bala Jain [3]

[1]Department of Computer Science and Engineering, Amity School of Engineering and Technology, Guru Gobind Singh Indraprastha University, New Delhi, India

[2]Department of Computer Science and Engineering, Ambedkar Institute of Technology, New Delhi, India

[3]Department of Electronics and Communication Engineering, IGIT, Guru Gobind Singh Indraprastha University, New Delhi, India

(*Received December 13, 2008, accepted January 8, 2009*)

Abstract. This paper presents a new feature vector set for noisy speech recognition in autocorrelation domain. The autocorrelation domain is well known for its pole preserving and noise separation properties. In this paper we will use the autocorrelation domain as an appropriate candidate for robust feature extraction. In our approach, extraction of mel frequency cepstral coefficients (MFCC) of the speech signals are proposed based on novel Differentiated Relative Higher Order Autocorrelation Coefficient Sequence Spectrum (DRHOASS). In this approach, initially the lower lags of the noisy speech autocorrelation sequence are discarded and then, the effect of noise is further suppressed using a high pass filter in autocorrelation domain. Finally, the feature vector set of the speech signal is found using the spectral peaks of the filtered autocorrelation sequence. We tested our features on the Hindi isolated-word task and found that it led to noticeable improvements over other autocorrelation-based and differential spetral-based methods.

Keywords: Autocorrelation domain, Feature vector set, spectral peaks

1. Introduction

Today, a major concern in the design of speech recognition systems is their performance in noisy conditions. Therefore, a substantial amount of research is devoted to the development of noise-robust speech features. The main approaches taken to improve the performance of automatic speech recognition (ASR) systems could be roughly divided into three main categories, namely, robust speech feature extraction, speech enhancement and model-based compensation for noise. In the case of speech enhancement, some initial information about speech and noise is needed to allow the estimation of noise and clean up of the noisy speech. Widely used methods in this category include spectral subtraction (SS) [1] and Wiener filtering [2]. In the framework of model-based compensation, statistical models such as Hidden Markov Models (HMMs) are usually considered. The compensation techniques try to remove the mismatch between the trained models and the noisy speech to improve the performance of ASR systems. Methods such as parallel model combination (PMC) [3][4], vector Taylor series (VTS) [5][6][7] [8][9] and weighted projection measure (WPM) [10] can be classified into this category.

Use of the autocorrelation domain in speech feature extraction has recently proved to be successful for robust speech recognition. Among the techniques introduced that exploit the autocorrelation properties are Short-time Modified Coherence (SMC) [11] and One-Sided Autocorrelation LPC (OSALPC) [12]. Pole preserving is an important property of the autocorrelation domain, i.e. if the original signal can be modelled by an all-pole sequence which has been excited by an impulse train and a white noise, the poles of the autocorrelation sequence would be the same as the poles of the original signal [13]. This means that the features extracted from the autocorrelation sequence could replace the features extracted from the original speech signal. Extracting appropriate speech features is crucial in obtaining good performance in ASR systems since all of the succeeding processes in such systems are highly dependent on the quality of the extracted features. Therefore, robust feature extraction has attracted much attention in the field.

[+] Corresponding author. *E-mail address: pbansal89@yahoo.co.in*

Another property of autocorrelation sequence is that for many typical noise types, noise autocorrelation sequence is more significant in lower lags. Therefore, noise-robust spectral estimation is possible with algorithms that focus on the higher lag autocorrelation coefficients such as autocorrelation mel-frequency cepstral coefficient (AMFCC) method [14]. Moreover, as the autocorrelation of noise could in many cases be considered relatively constant over time, a high pass filtering of the autocorrelation sequence, as done in relative autocorrelation sequence (RAS)[15], 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 [16][17],[18]. Methods such as peak-to-valley ratio locking [19] and peak isolation (PKISO) [20] have been found very useful in speech recognition error rate reduction. In differential power spectrum (DPS)[21], as an example, differentiation in the spectral domain is used to preserve the spectral peaks while the flat parts of the spectrum, that are believed to be more vulnerable to noise, are almost removed.

Each of the above-mentioned autocorrelation-based methods has its own disadvantages. RAS, while working well in low SNRs, does not perform as well in higher SNRs and clean condition. The main reason is that while filtering the lower frequency parts of noisy autocorrelation sequence can lead to the suppression of noise in low SNR conditions (large noise energies), it in fact filters out parts of the signal autocorrelation sequence in high SNRs. However, the removal of the lower lags of the sequence leads to a good performance in high SNRs in comparison to high-pass filtering. According to [14], AMFCC works well for car and subway noises, but in babble and exhibition noises does not work as well. This is attributed to high similarities between the properties of the latter two and speech. The reason could be that for the two latter noise types, the noise properties are very similar to speech properties, i.e. the autocorrelation sequence is not mostly concentrated in lower lags. Therefore, the removal of the lower lags might not be that much helpful.

In this paper we propose a novel Differentiated Relative Higher Order Autocorrelation sequence spectrum (DRHOASS) method for computing MFCC feature vector set. In this method, removal of lower lags of the autocorrelation sequence has been proposed with the additional high-pass filtering of the autocorrelation sequence to provide doublefold noise suppression. In addition to this, the resultant spectral peaks are extracted as a third noise suppression step. The remainder of this paper is organized as follows. Properties of the short-time autocorrelation function are described in section 2. Calculation of coefficients in autocorrelaation domain has been explained in section 3. Extraction of feature vector set is explained in section 4. In section 5. experiments conducted using different front-ends are discussed and compared with the proposed feature vector set. Finally a conclusion is drawn in section 6.

2. Properties of Short-time Autocorrelation Function

The autocorrelation function of a signal contains the same information about the signal as its power spectrum [22]. In the power spectrum domain, the information is presented as a function of frequency and in the autocorrelation domain it is represented as a function of time. In this section, we demonstrate some of the properties of the short-time autocorrelation function relevant in the context of DRHOASS method of feature extraction. Since the DRHOASS method discards the zeroth and lower-lag autocorrelation coefficients and uses only the higher-lag autocorrelation coefficients for spectral estimation, it is necessary to know whether and how these coefficients contain the spectral information necessary for speech recognition. We are proposing the DRHOASS method as a robust feature extraction procedure on the basis that the additive noise distortion has most of its autocorrelation coefficients concentrated near the lower time- lags and their higher-lag autocorrelation coefficients are zero (or, very small). Theoretically (and asymptotically), the autocorrelation function should be zero for all the lags except for the zeroth lag [23]. We want to know whether this is true for short-time analysis. We take 3 s long computer-generated (artificial) white Gaussian noise and perform a short-time analysis (with Hamming window) using a frame length of 16 ms. For illustration, we take three different frames starting from 0.5, 1 and 1.5 s. In Fig.1 (a)(b) and (c), we show the waveform of the frame at 0.5 s, its power spectra and its autocorrelation spectra. Similarly Fig.1 (d)(e) and (f) and Fig.1(g),(h) and (i) show waveform of the frame, their respective power spectrum and their respective autocorrelation spectrum at 1 and 1.5 s. As expected, the higher-lag autocorrelation coefficients are smaller in magnitude than the zeroth autocorrelation coefficient, but they have non-zero values due to short-time analysis.

Additional filtering which is used in the proposed method is high-pass filtering of Higher order Autocorrelation sequence and extraction of spectral peaks. This filtering combines the advantages of RAS and DPS. Fig.2 depicts a sample speech signal, its short-time autocorrelation spectrum and the differentiated short-time autocorrelation spectrum. This sample signal is one frame of sample speech. In order to simplify the representation, only the significant lower-frequency parts of the spectrum have been shown and the non-significant parts omitted. As shown in Fig.2 and mentioned above, the flat parts of the filtered autocorrelation spectrum have been transformed to zero by differentiation and each peak of it split into two positive and negative parts. Since the spectral peaks convey the most important information in speech signal, this fact that the differential power spectrum retains spectral peaks means that we will not lose the important information of the speech signal.

Furthermore, since noise spectrum is often flat and the differentiation either reduces or omits the relatively flat parts of the spectrum, it will lead to omission of the effect of noise on the signal leading to more robust features.

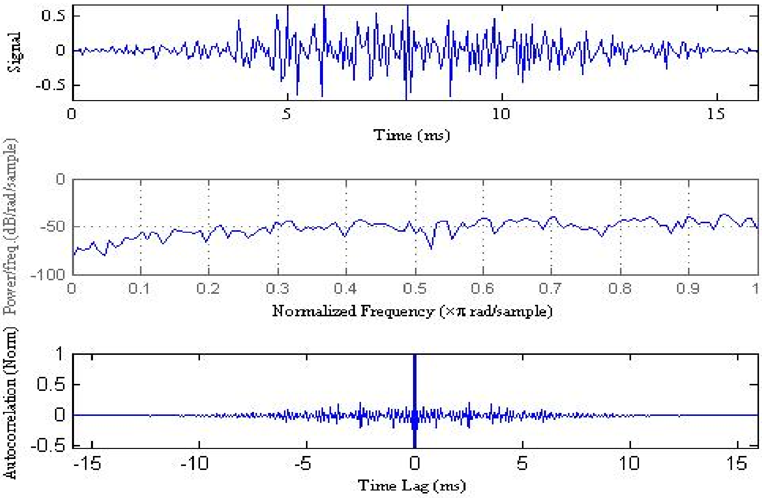

Fig. 1. Short-time analysis of the artificial white noise signal using 16ms frame. (a) Waveform of noise frame taken at 0.5 sec.(b) Power spectrum estimate of given frame (c) Autocorrelation spectrum corresponding to power

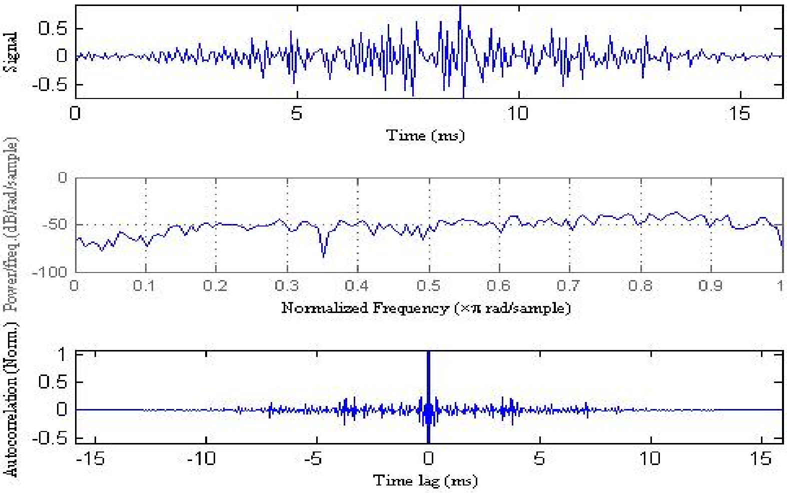

Fig. 1. Short-time analysis of the artificial white noise signal using 16ms frame. (d) Waveform of noise frame taken at 1 sec.(e) Power spectrum estimate of given frame (f) Autocorrelation spectrum corresponding to power

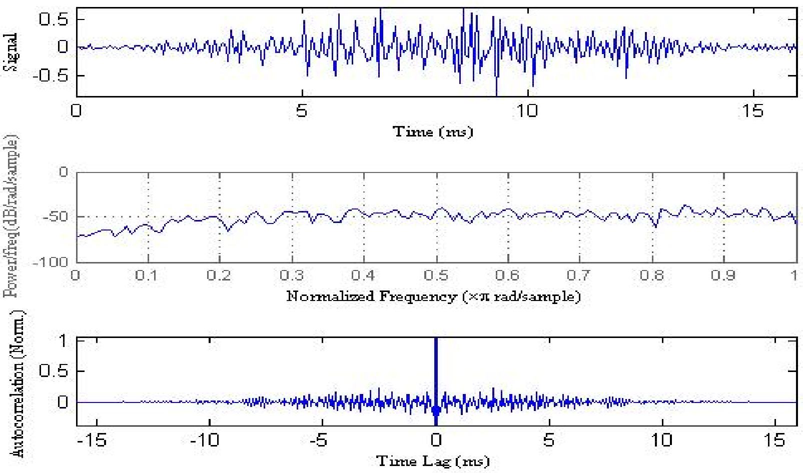

Fig. 1. Short-time analysis of the artificial white noise signal using 16ms frame. (g) Waveform of noise frame taken at 1.5 sec.(h) Power spectrum estimate of given frame (i) Autocorrelation spectrum corresponding to power

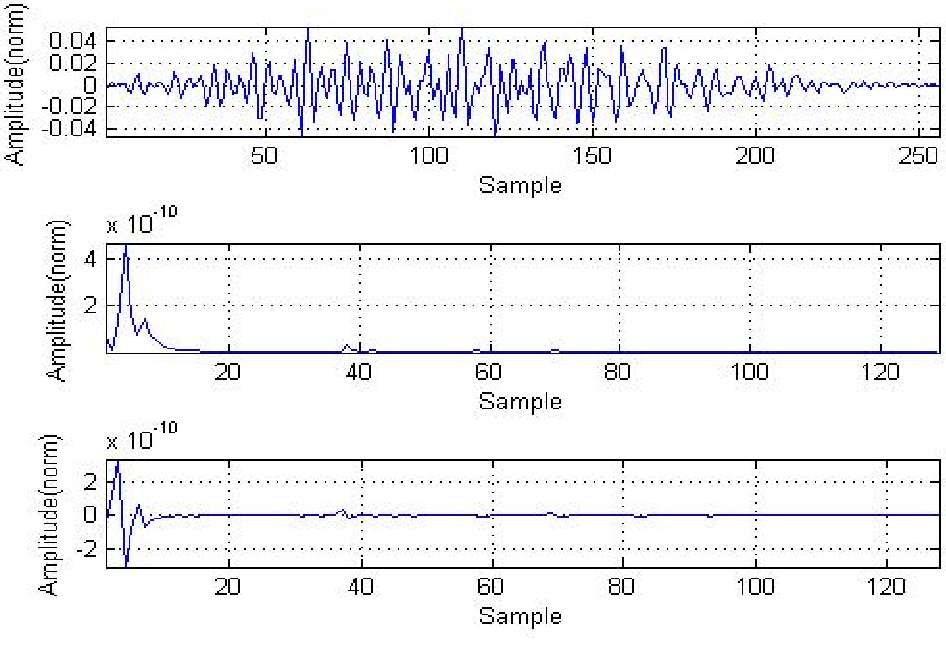

Fig. 2. (a) A sample speech signal, and (b) the autocorrelation spectrum magnitude and (c) differentiated autocorrelation spectrum magnitude of the same signal with a 512-point FFT. Only 128 points of the spectrum are shown for clarity.

3. Calculation of Coefficients 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

$$y(m,n) = [x(m,n) + u(m,n)],\ 0\le m\le M-1\ ,\ 0\le n\le N-1 \tag{2}$$

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.

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

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 (3) becomes

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

The N-point ryy (m,k) is computed from N-point y(m,n) using the following equation,

$$r_{yy}(m,k) = \sum_{i=0}^{N-1-k} y(m,i)\, y(m,i+k) \tag{5}$$

Eliminating the lower lags of the noisy speech signal autocorrelation should lead to removal of the main noise components. The maximum autocorrelation index to be removed is usually found experimentally. The resulting sequence after the removal of lower lags would be

$$\begin{aligned} r_{yy}(m,k) &= r\ _{yy}(m,k), && D \le m \le M-1 \\ r_{yy}(m,k) &= 0, && 0 \le m \le D \end{aligned} \tag{6}$$

Where D is the Elimination threshold (found experimentally).

Differentiating the resultant autocorrelation sequence with respect to m, will remove the noise autocorrelation and gives:

$$\frac{\partial r_{yy}(m,k)}{\partial m} = \frac{\partial r_{xx}(m,k)}{\partial m} + \frac{\partial r_{uu}(k)}{\partial m} \cong \frac{\partial r_{xx}(m,k)}{\partial m} = \frac{\sum_{t=-L}^{L} t.\ r_{yy}(m+t,k)}{\sum_{t=-L}^{L} t^2},\ 0\le m\le M-1\ ,\ 0\le k\le N-1 \tag{7}$$

The sequence, $\left\{\partial ryy(m,k)_{k=0}^{N-1}\right\}$ is named the Relative Autocorrelation Sequence (RAS) of noisy speech at the mth frame. In order to get DRHOASS, we take differentiation of the spectrum of the filtered signal (which we get from previous step i.e. RAS). 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. The differential power spectrum of the filtered signal in discrete domain, can be defined as

$$\text{Diff}\ _{Y}(k) \approx \sum_{l=-Q}^{P} a_l\, Y(k+l),\ 0 \le k \le K-1 \tag{8}$$

where P and Q are the orders of the differential equation, al are some real-valued weighing coefficients and K is the length of FFT.

4. Extraction of Feature vector Set by DRHOASS

In this section we will describe our proposed method to obtain new robust features for speech recognition. First, we pre-emphasis the input speech signal using a pre-emphasis filter H(z) = 1- 0.97 z-1 . In order to carry out short time analysis of the pre-emphasised speech signal, we perform frame blocking with a frame size of 16ms and a frame shift of 8 ms and the signal is then analysed sequentially in a frame-wise manner. The Hamming window is applied to the pre-emphasised signal and then, the autocorrelation sequence 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, a Hamming window is applied to the filtered signal and 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 may in many occasions 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 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 feature vector set of dimension 39 is formed by concatenating energy feature, Delta MFCC and Delta-Delta MFCC. Front-end for extraction of MFCC feature vector set by DRHOASS has been shown in Fig.3.

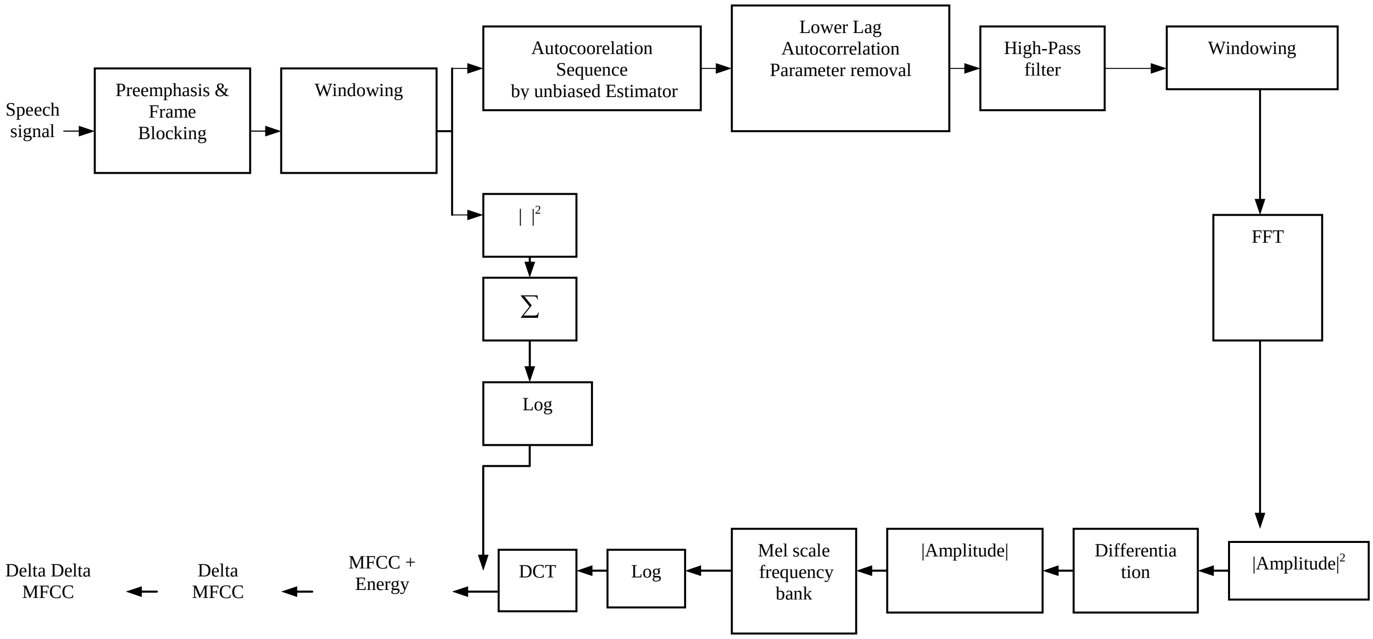

Fig. 3. Block diagram of proposed front end

5. Experiments and Results

The proposed approach was implemented on TIFR Hindi speech database. Database of 200 Hindi words (Table 1) spoken by 30 speakers was used. The spoken samples were recorded by 15 male, 10 female and 5 child speakers in a studio environment condition using Sennheiser microphone model MD421 and a tape recorder model Philips AF6121. Each speaker uttered 5 repetitions of words. Database was divided into training set and testing set.

We evaluate the recognition performance of the proposed feature vector set in the presence of white and colored noises and compare it with other front ends. We compare it with the MFCC, AMFCC, and RAS methods. Features vector sets of size 39 are extracted using different front-ends (MFCC (for comparison purposes), RAS-MFCC, AMFCC and our method DRHOASS-MFCC). With these features vector sets, word models of training database for different front-ends are created by seven state left-right Hidden Markov model. Afterwards word recognition rates for testing database are computed with all the above front-ends and compared with the traditional MFCC.

(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 Table2. 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 methods.

Table1. Hindi speech Database for a vocabulary of 200 words used in the experiment

1.	Language	:	Standard Hindi (Khari Boli)
2.	Vocabulary Size	:	A set of 200 most frequently occurring Hindi words
3.	Speakers	:	30 speakers
4.	Utterances	:	(15 male, 15 female and 5 children) 5 repetitions each
4.	Audio Recording	:	Recording on a cassette tape in studio S/N > 40dB
5.	Digitization	:	16KHz. Sampling, 16 bit quantization.

Table 2.Comparison of clean-train and clean test recognition rates for various features

Feature type	Recognition rate (%) at 40 dB
MFCC	98.241
AMFCC	98.246
RAS-MFCC	98.30
DRHOASS-MFCC	99.642

(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 [24]. The noises are added to the clean speech signal at 20, 15, 10 5 and 0 dB SNRs. RAS-MFCC, AMFCC and DRHOASS-MFCC are evaluated and word recognition rates are compared with the traditional MFCC front end. Fig. 4(a)-(d) shows the results obtained using MFCC, RAS-MFCC, AMFCC and DRHOASS front-ends. For the case of white noise corruption, i.e., in Fig. 4(a), the performance of MFCC degrades most significantly among all features, its performance is worse than RAS-MFCC, AMFCC and DRHOASS-MFCC. It is obvious that DRHOASS-MFCC are quite robust to the additive noises.

Fig. 4(b), (c) and (d), show the performance when the testing speech is corrupted by factory, babble, and f16 noises, respectively. The figures depict that the performance of MFCC degrades significantly. The best performance comes from DRHOASS-MFCC. This is due to peak preserving property of power spectrum domain, which helps in better recognition in noisy environment. The experiments show the better performance of the new feature vector set in comparison to the other autocorrelation based robust speech recognition parameters.

In order to find the most suitable autocorrelation lag for discarding, we have tested different lag values. The results are shown in Fig. 5. As seen in figure, best results are obtained when lags of lower than 22 samples (1.375ms) were discarded. Hence this value is taken as threshold in our experiments.

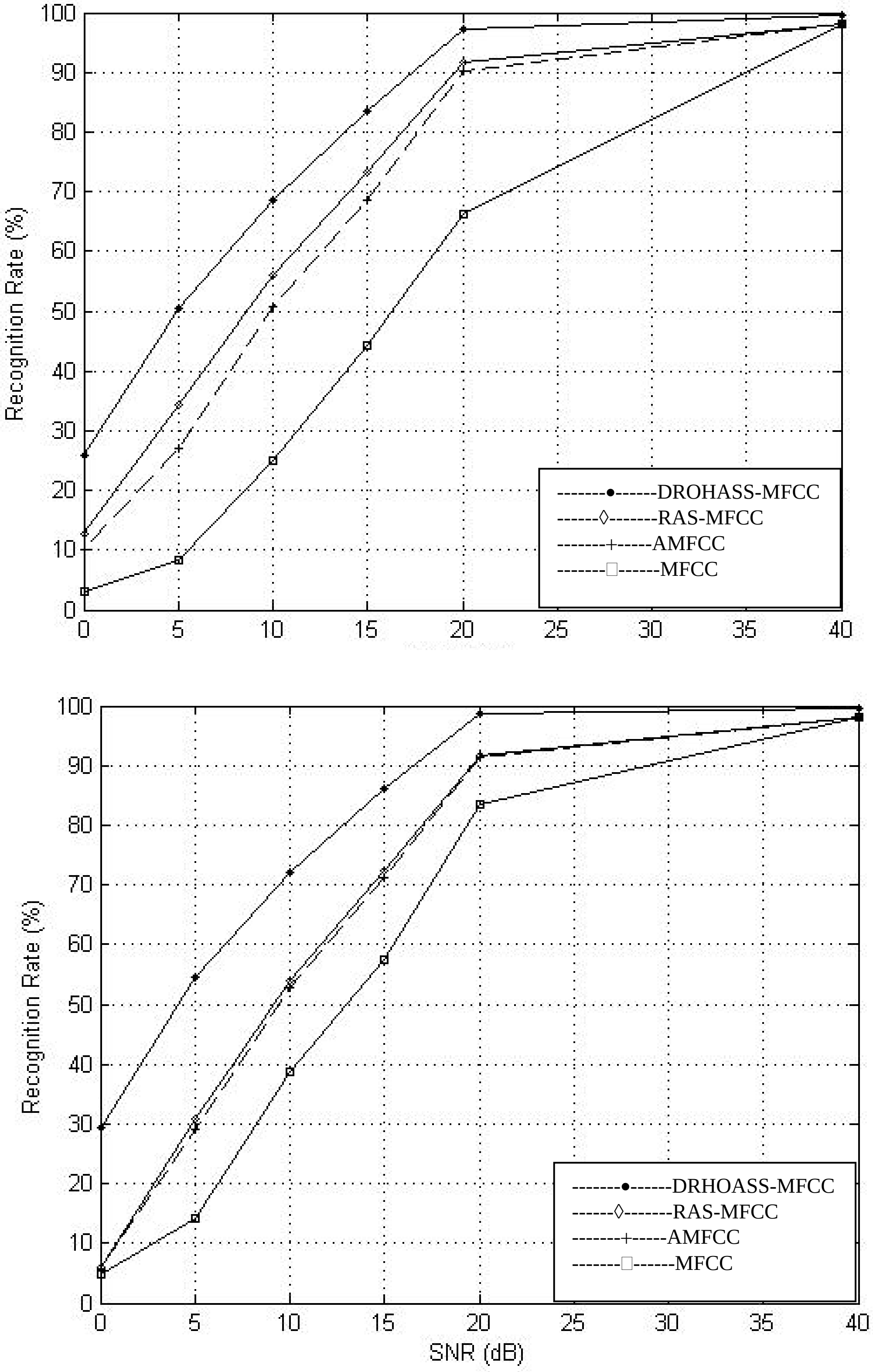

Fig. 4(b). Recognition rate (%) for testing speech corrupted by factory noise

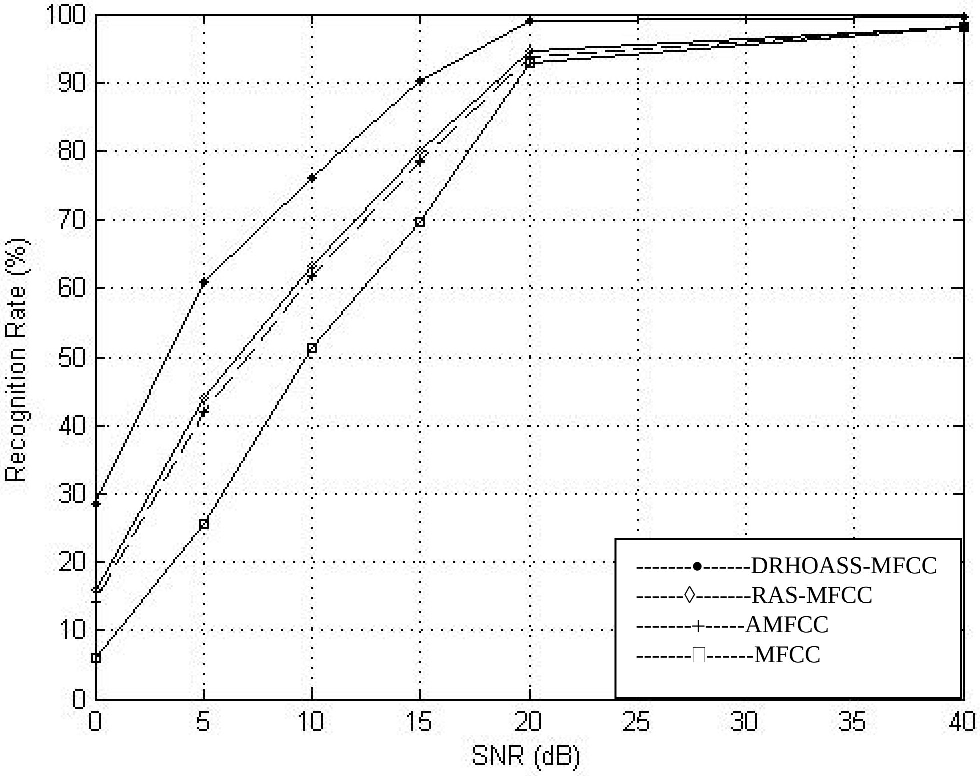

Fig. 4(c). Recognition rate (%) for testing speech corrupted by babble noise

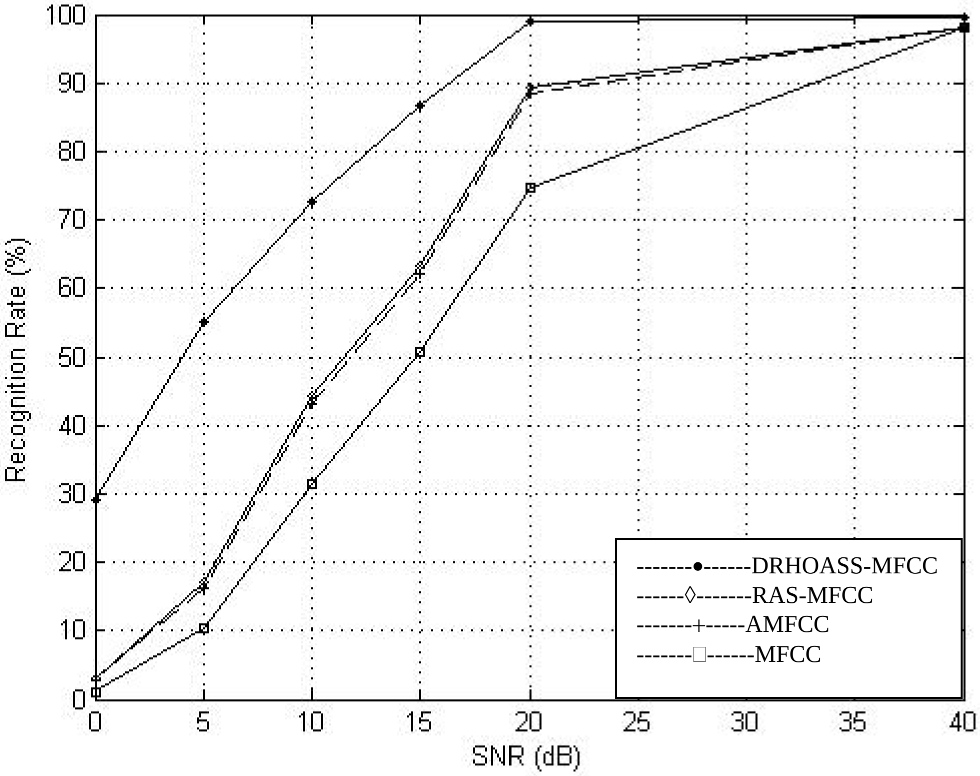

Fig. 4(c). Recognition rate (%) for testing speech corrupted by F16 noise

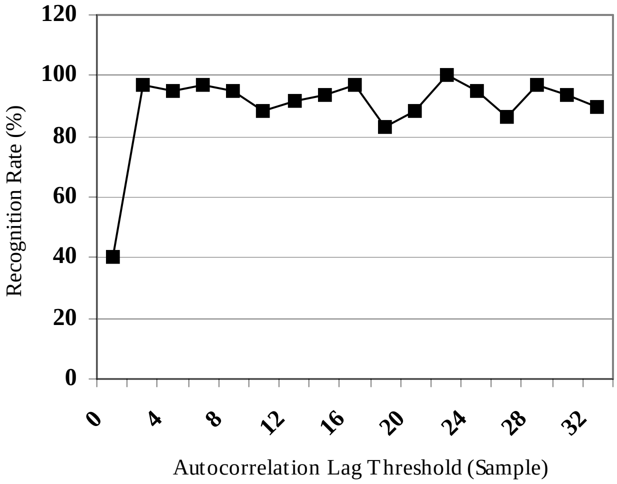

Fig. 5. Average recognition rate (%) vs autocorrelation lag threshold

6. Conclusion

In this paper a new feature vector set is proposed based on DRHOASS to improve the performance of ASR systems. We have improved the performance of existing ASRs based on higher lag autocorrelation coefficients by including additional filtering and picking the peaks in the spectral domain. Filtering stage has helped to reduce the effects of additive noises. The concept of spectral peaks has introduced a new set of cepstral features for improving the robustness of speech recognition. We note that just like the power spectrum, picking the spectral peaks can also preserve spectral information to discriminate among words.

7. References

[1] J.Beh and H. Ko. A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech. *in Proc. ICASSP'*. 2003, **I** : 648-651.

[2] C.H.Lee, F.K. Soong and K.K. Paliwal. Automatic speech and speaker recognition. Kluwer Academic Publishers, 1996.

[3] M.J.F. Gales and S.J. Young. Robust speech recognition in additive and convolutional noise using parallel model combination. *Computer Speech and Language.*1995, **9**: 289-307.

[4] M.J.F. Gales and S.J. Young. Robust Continuous Speech Recognition Using Parallel Model Combination. *IEEE Trans. Speech Audio Processing*. 1996, 4 (5): 352-359.

[5] A. Acero, L. Deng, T. Kristjansson & J. Zhang. HMM adaptation using vector Taylor series for noisy speech recognition. *in Proc. ICSLP'*. 2000, **3**: 869-872.

[6] D.Y. Kim, C.K. Un and N.S. Kim. Speech recognition in noisy environments using first-order vector Taylor series. *Speech Communication.* 1998, **24** (1): 39-49.

[7] P.J. Moreno. *Speech recognition in noisy environment.* PhD Thesis, Carnegie-Mellon University, 1996.

[8] P.J. Moreno, B. Raj and R.M. Stern. A vector Taylor series approach for environment independent speech recognition. *in Proc. ICASSP'96.* pp.733-736.

[9] J.L. Shen, J.W. Hung and L.S. Lee. Improved robust speech recognition considering signal correlation approximated by Taylor series. *in Proc. ICSLP'98.*

[10] D.Mansour and B.H. Juang. The short-time modified coherence and noisy speech recognition. *IEEE Trans. Acoustics and signal processing.* 1989, **37** (6): 795-804.

[11] D.Mansour and B.H. Juang. A family of distortion measures based upon projection operation for robust speech recognition. *IEEE Trans. Speech and Audio Processing.* 1989, **37**(11): 1659-1671.

[12] J. Hernando and C. Nadeu. Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition. *IEEE Trans. Speech Audio Processing.* 1997, **5**(1): 80-84.

[13] D.P. McGinn and D.H. Johnson. Estimation of all-pole model parameters from noise-corrupted sequence. *IEEE Trans. Acoustics Speech and Signal Processing.* 1989, 37(3): 433-436.

[14] B.J. Shannon and K.K. Paliwal. Feature extraction from higher-lag autocorrelation coefficients for robust speech

recognition. *Speech Communication*, 2006, **48**(11): 1458-1485.

[15] Kuo-Hwei Yuo and Hsiao-Chum Wang. Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences. *Speech Communication.* 1999, **28**: 13-24.

[16] M. Padmanabhan. Spectral peak tracking and its use in speech recognition. *in Proc. ICSLP 2000*.

[17] B. Strope and A. Alwan. Robust word recognition using threaded spectral peaks. *in Proc. ICASSP'98*. pp. 625-628.

[18] J.Sujatha, K.R. Prasanna Kumar, K.R. Ramakrishnan and N. Balakrishnan. Spectral maxima representation for robust automatic speech recognition. *in Proc. Eurospeech'03.* pp. 3077-3080.

[19] Q. Zhu, M. Iseli, X. Cui and A. Alwan. Noise robust feature extraction for ASR using the AURORA2 database. *in Proc. Eurospeech*, 2001.

[20] B. Strope and A. Alwan. A model of dynamic auditory perception and its application to robust word recognition. IEEE Trans. *on Speech and Audio Processing*. **5** (5), 451-464.

[21] J. Chen, K.K. Paliwal and S. Nakamura. Cepstrum derived from differentiated power spectrum for robust speech recognition. *Speech Communication.* 2003, **41**: 469-484.

[22] S. Kay. Modern Spectral Analysis. 1988, Prentice Hall.

[23] S.M. Kay. The effects of noise on the autoregressive spectral estimator. *IEEE Trans. Acoust. Speech Signal Process*. 1979, **27** (5), 478-485.

[24] A. Varga and H. J. M. Steeneken. 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.* 1993, **12**: 247-251.