Role of Different Order Ranges of Autocorrelation Sequence on the Performance of Speech Recognition

POONAM BANSAL[1], AMITA DEV[2] and SHAIL BALA JAIN[3]

[1]Department of Computer Science and Engineering, Amity School of Engineering and Technology
[2]Department of Computer Science and Engineering, Ambedkar Institute of Technology
[3]Department of Electronics and Communication Engineering, IGIT
Guru Gobind Singh Indraprastha University, New Delhi
INDIA
E.mail: pbansal89@yahoo.co.in

Abstract:- In this paper, cepstral features derived from the Differentiated Relative Higher Order Autocorrelation Sequence Spectrum (DRHOASS) are proposed for improving the robustness of a speech recognizer in the presence of background noise. Proposed method is analyzed and compared in terms of the autocorrelation coefficients they employ with the traditional feature extraction methods based on Linear Pediction (LP) analysis. LP- based techniques used are Linear Predictive Cepstral Coefficients (LPCC), Short-Time Modified Coherence (SMC) and the One-Sided Autocorrelation Linear Prediction Coefficient (OSALPC). We evaluate the speech recognition performance of the proposed features on the Hindi isolated-word task and show that the proposed features show better recognition performance than the features derived from the robust liner prediction based methods for noisy speech.

Keywords: Autocorrelation coefficients, Feature vector set, LPCC, SMC, OSALPC.

1 Introduction

Speech signal carries information from many sources. But not all information is relevant or important for speech recognition. In speech recognition, the first crucial step is the feature extraction, where the speech signal of a given frame is converted to a set of acoustic features with the hope that these features will encapsulate the important information that is necessary for recognition. Once these features are computed, a backend classifier is used to recognize the input speech signal into a sequence of words in light of the extracted features and pre-trained models.

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) [1] and One-Sided Autocorrelation LPC (OSALPC) [2]. Pole preserving is an important property of the autocorrelation domain, i.e. if the original signal can be modeled 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 [3]. 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.

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 [4]. 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)[5], 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 [6]. In differential power spectrum (DPS) [7], 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.

In this paper we use different order ranges of the autocorrelation function separately for deriving speech recognition features, and investigate their role in speech recognition performance. This paper is organized as follows. In section 2, the specific algorithms for the five different speech feature sets are introduced. Following this in sections 3 and 4 the implementation and experimental framework is described, along with the results and discussion. Finally conclusions are given in section 5.

2 Autocorrelation Derived Features

To evaluate the effect of different order ranges of the autocorrelation function on a speech feature set's robustness to noise, five different feature sets are investigated. These include linear prediction cepstral coefficients (LPCC), short time modified coherence (SMC), one-sided autocorrelation linear prediction coefficient (OSALPC), mel frequency cepstral coefficients (MFCCs), and our newly proposed differentiated relative higher order autocorrelation coefficient sequence spectrum (DRHOASS). LPCC, in comparison, use the fewest coefficients of the features in the study, with only 13 lower-order autocorrelation coefficients (order 12 model). MFCCs use 256 unique autocorrelation coefficients for a 16 ms frame sampled at 16kHz system. Each of the feature sets used in the study are introduced next, along with the proposed DRHOASS features.

2.1 Linear Prediction Cepstral Coefficients (LPCC)

LPC features are generated in accordance with the vocal cord or human vocal tract. LPCC is based on LPC features. They model the speech as a linear but time varying system. Speech samples from previous time points are combined linearly to predict the current value. The conventional LPCC technique is very sensitive to the presence of background noise. Beginning with the speech signal, frames of 256 samples are formed with an overlap of 128 samples. A Hamming window is applied to each of these frames, before a biased autocorrelation estimate is made. Using these autocorrelation co-efficients, the Yule-Walker equations are solved using the Levinson-Durbin algorithm, then converted to cepstral coefficients using a recursion relation [8][9]. A block diagram for extracting LPCC feature vector set is shown in Fig. 1(a).

2.2. Short Time Modified Coherence (SMC)

SMC based technique of speech signals gives more robust estimation of the standard LPC parameters. This is based on LPCC features. SMC performs better in case of loud signals. They are more robust to additive white noise. SMC is an all-pole modeling of the autocorrrelation sequence with a spectral shaper. The spectral shaper is the square root operator in the frequency domain, which compensates for the inherent spectral distortion introduced by the autocorrelation operation on the autocorrelation sequence of the signal. For extracting the features by this method we split the speech signal into frames and apply a pre-emphasis filter, then autocorrelation is applied followed by windowing to remove discontinuities. Then FFT is applied to convert time domain to frequency domain followed by computing the absolute and the IFFT to convert back to time domain. A set of cepstral coefficients is then derived by applying Levinson Durbin Recursion Algorithm. Block diagram for SMC extraction is given in Fig. 1(b).

2.3 One-sided autocorrelation linear prediction coefficient (OSALPC)

OSALPC algorithm is based on LPCC features. OSALPC performs better in case of noisy signals. It is closely related to the short-time modified coherence (SMC) representation. SMC is also based on AR modeling in the autocorrelation domain. However, whereas in the OSALPC technique, the entries to the Levinson-Durbin algorithm are calculated from the OSA sequence using the conventional biased autocorrelation estimator, in the SMC representation, they are computed using a square root spectral shaper. OSALPC shows better speech recognition performance than conventional LPCC in severe conditions of additive noise. In this approach we split the speech signal into frames and apply a pre-emphasis filter, then

one-sided autocorrelation sequence is computed followed by windowing to remove discontinuities. Then, autocorrelation of order 12 is applied. A set of cepstral coefficients are then derived by applying Levinson Durbin Recursion Algorithm. Fig. 1(c) shows the OSALPC feature set extraction.

2.4 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 16 ms Hamming windowed time frames, which overlap by 8 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. 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.

2.5 Differentiated relative higher order autocorrelation coefficient sequence spectrum (DRHOASS)

The proposed DRHOASS method is 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-

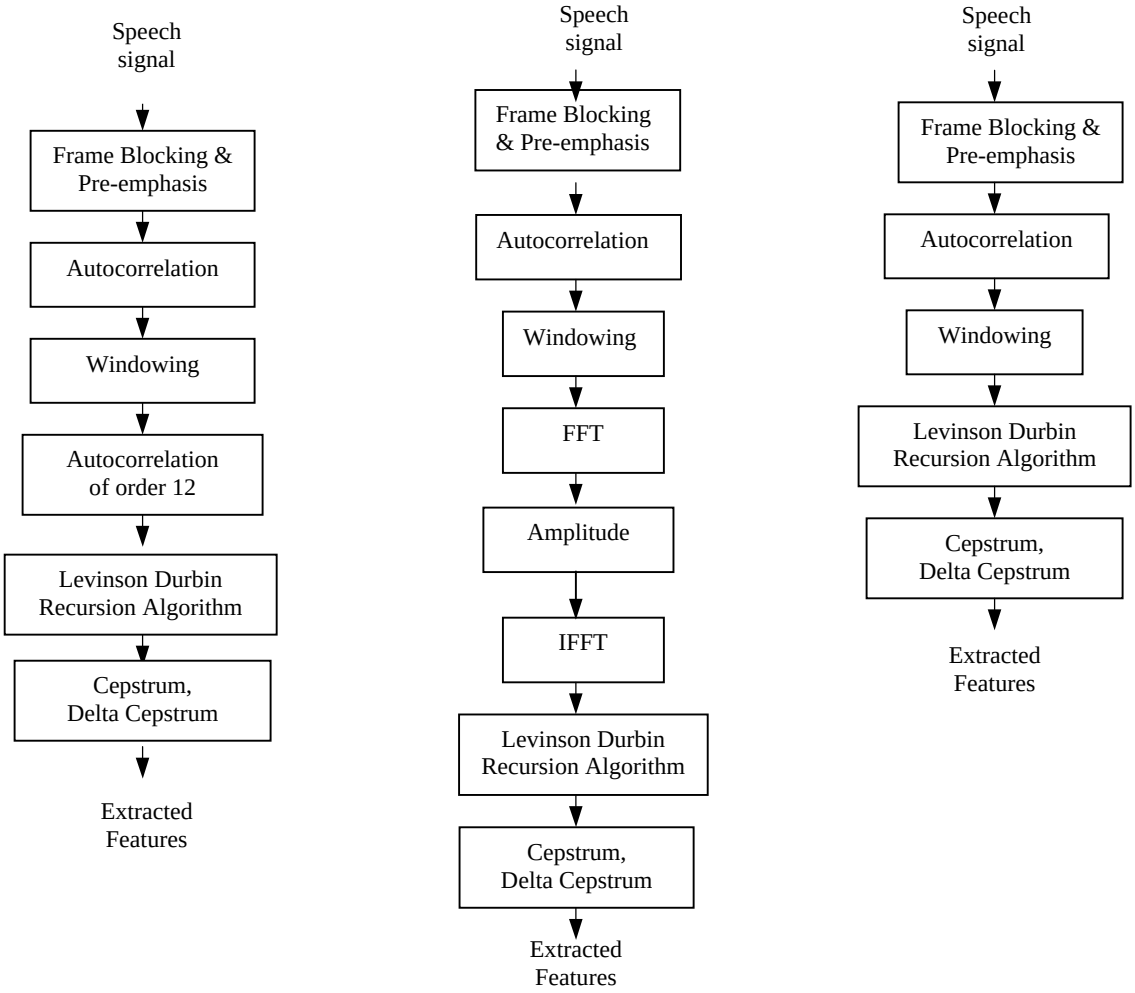

Fig. 1 (a) Block diagram of LPCC (b) Block diagram of SMC (c) Block diagram of OSALPC

lag autocorrelation coefficients are zero (or, very small). Theoretically (and asymptotically), the autocorrelation function should be zero for all the orders except for the zeroth order. To verify this fact 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. 2 (a) (b) and (c), we show the waveform of the frame at 0.5 s, its power spectra and its autocorrelation spectra. Similarly Fig. 2 (d) (e) and (f) and Fig. 2(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-order autocorrelation coefficients are smaller in magnitude than the zeroth autocorrelation coefficient, but they have non-zero values due to short-time analysis. The extraction of features based on DRHOASS is as follows:

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 :

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

Where M denotes the number of frames in an utterance and N denotes the number of samples in a frame and $\otimes$ 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 :

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

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.

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

where $r_{yy}(m, k)$, $r_{xx}(m, k)$ and $r_{uu}(m, k)$ 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

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

The N-point $r_{yy}(m,k)$ is computed from N-point y(m,n) using the following equation,

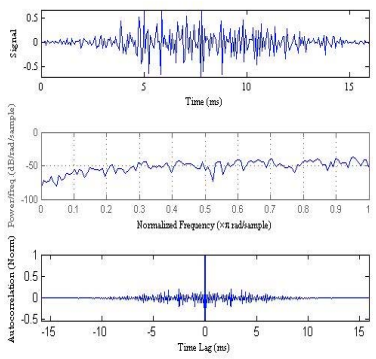

Fig. 2. Short-time analysis of a noisy signal using 16 ms frame. (a) Waveform of noise frame taken at 0.5 sec.(b) Power spectrum estimate of given frame (c) Autocorrelation spectrum corresponding to power

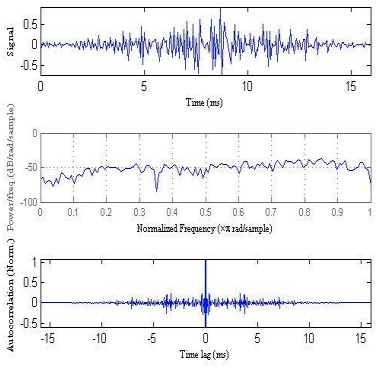

Fig. 2. Short-time analysis of a noisy signal using 16 ms frame. (d) Waveform of noise frame taken at 1.0 sec.(e) Power spectrum estimate of given frame (f) Autocorrelation spectrum corresponding to power

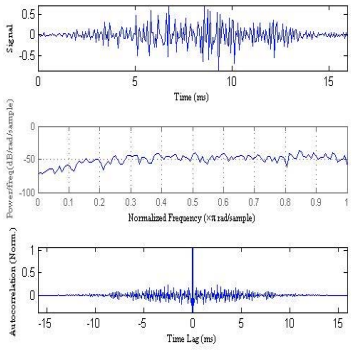

Fig. 2. Short-time analysis of a noisy signal using 16 ms frame. (g) Waveform of noise frame taken at 1.5 sec.(h) Power spectrum estimate of given frame (i) Autocorrelation spectrum corresponding to power

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

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

$$r_{yy}(m,k) = r_{yy}(m,k), \quad D \le m \le M-1$$

$$r_{yy}(m,k) = 0, \quad 0 \le m < D \quad (6)$$

Where D is the Elimination threshold (found experimentally). Differentiating the resultant autocorrelation sequence with respect to m, will remove the noise autocorrelation and gives:

$$\frac{\partial r_{yy}(m,k)}{\partial m} = \frac{\partial r_{xx}(m,k)}{\partial m} + \frac{\partial r_{uu}(k)}{\partial m} \cong \frac{\partial r_{xx}(m,k)}{\partial m}$$

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

The sequence, $\left\{ \partial r_{yy}(m,k) \right\}_{k=0}^{N-1}$ 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

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

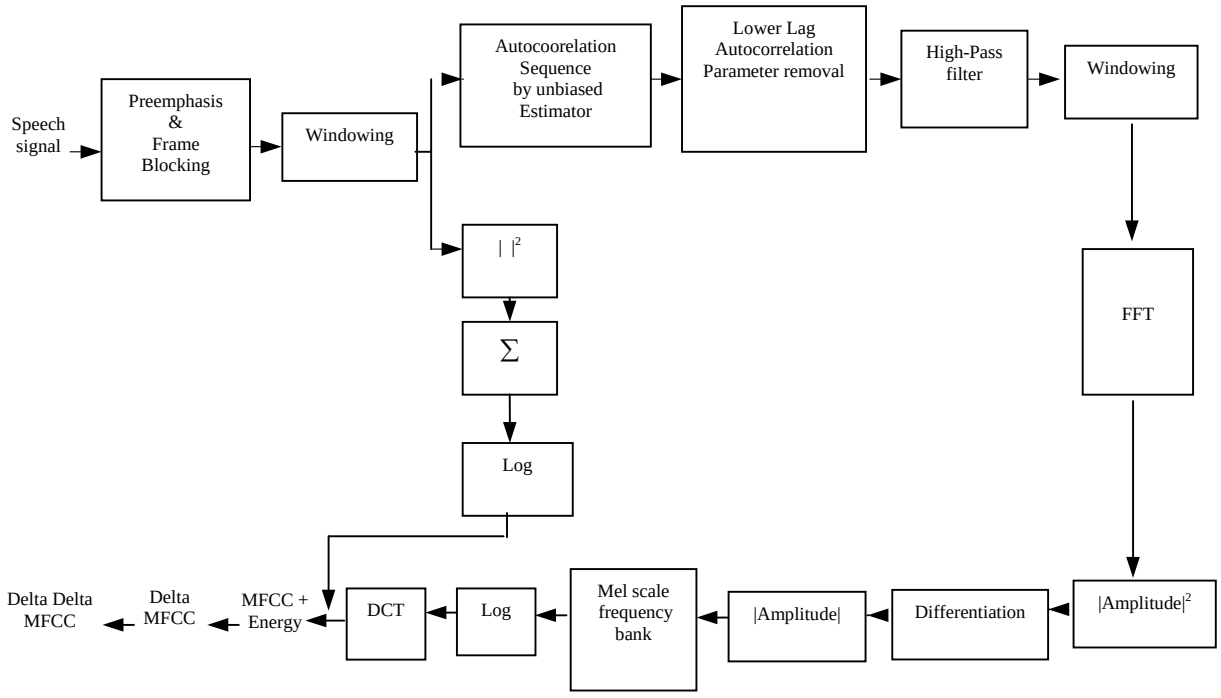

Fig. 3. Block diagram of proposed front end

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

3 Implementation

In this section we describe the implementation of the proposed method (DRHOASS) 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-emphasized speech signal, we perform frame blocking with a frame size of 16ms and a frame shift of 8 ms and the signal is then analyzed 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 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.

4 Recognition Experiment

To evaluate the performance of different feature vector sets TIFR Hindi speech database base of 200 Hindi words (Table1) 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 feature extraction methods. We compared it with the LPCC, SMC, OSALPC and MFCC methods. With the extracted 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 methods and analysed.

(a) Testing on clean speech

This experiment is to evaluate the performance of LPCC, SMC, OSALPC, MFCC 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 LPCC, SMC, OSALPC, MFCC, 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.

(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 [10]. The noises are added to the clean speech signal at 20, 15, 10 5 and 0 dB SNRs. Feature sets by LPCC, SMC, OSALPC 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 LPCC, SMC, OSALPC, MFCC and DRHOASS front-ends for different noises at various SNR levels. For the case of white noise corruption, i.e., in Fig. 4(a), the performance of LPCC, SMC and OSALPC degrades

most significantly among all features, its performance is worse than MFCC 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 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.

5 Conclusion

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 produce features that give high recognition accuracy in clean conditions. It is shown that features that are derived from the magnitude of the differentiated high order relative autocorrelation power spectrum are always more robust to noises. The experiments also show that speech recognition features that are derived from higher order range autocorrelation coefficients are more robust than features that use all orders (MFCC) for all tested noise types. They also show that higher order range derived features are more robust than lower order ranges (LPCC, SMC, OSALPC) for most of the tested noise types.

Table1. Speech Database

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, 10 female and 5 children) 5 repetitions each
5. Audio Recording	Recording on a casette tape in studio S/N > 40
6. Digitization	16 kHz., Sampling 16 bit quantization.

Table2. Recognition rates (%) under clean enrollment

Feature Extraction Method	Recognition Rate (%)
LPC	96.55
OSALPC	98.27
SMC	93.00
MFCC	98.24
DROHASS	99.67

References

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

[2] J. Hernando and C. Nadeu, "Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition," *IEEE Trans. Speech Audio Processing*, vol.5, no.1, pp.80-84, 1997.

[3] D.P. McGinn and D.H. Johnson, "Estimation of all-pole model parameters from noise-corrupted sequence," *IEEE Trans. Acoustics Speech and Signal Processing*, Vol 37, no. 3, pp. 433-436, 1989.

[4] B.J. Shannon and K.K. Paliwal, "Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition," *Speech Communication*, Vol. 48, No. 11, pp. 1458-1485, Nov. 2006.

[5] 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*, vol. 28, pp.13-24, 1999.

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

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

[8] J.Makhoul, " Spectral analysis of speech by linear prediction," *IEEE Transactions on Audio*

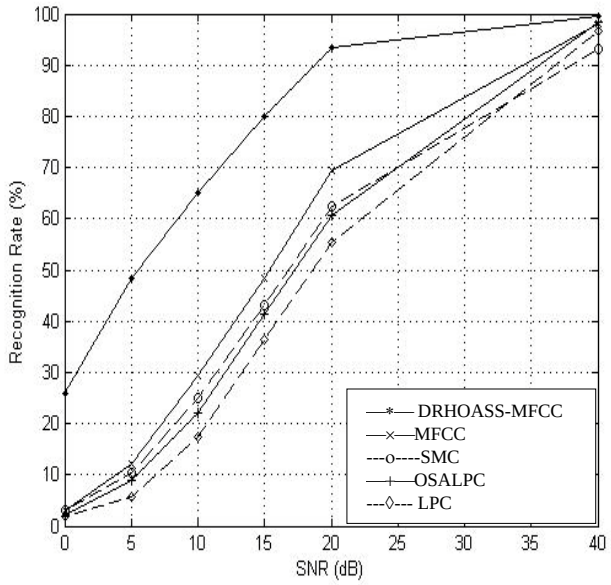

Fig. 4 (a) Recognition rate (%) for testing speech corrupted by white noise

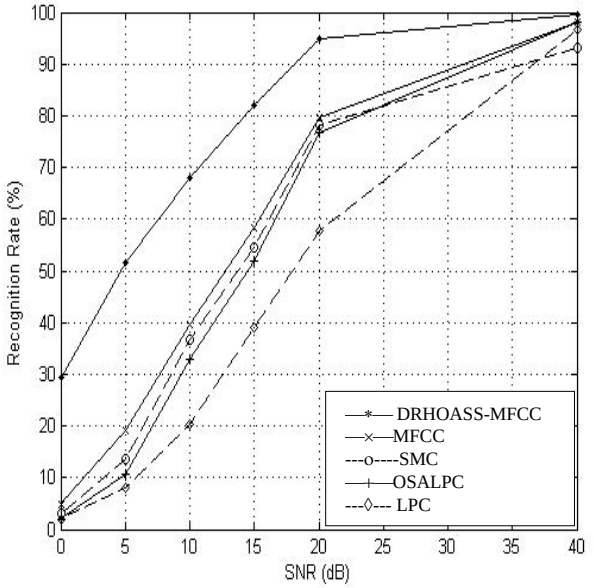

Fig. 4 (b) Recognition rate (%) for testing speech corrupted by factory noise

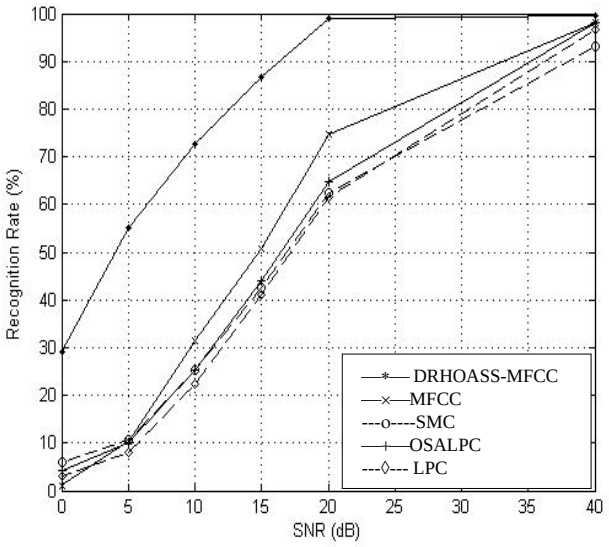

Fig. 4 (c) Recognition rate (%) for testing speech corrupted by F16 noise

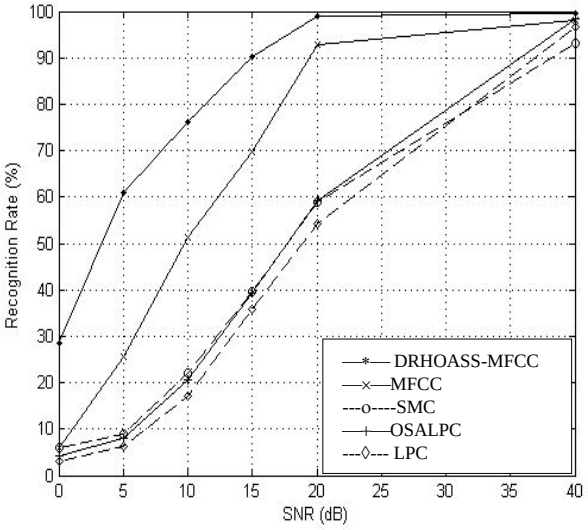

Fig. 4 (d) Recognition rate (%) for testing speech corrupted by babble noise

Electroacoust.,vol.21,pp.140-148, June 1973.

[9] J. Makhoul, " Spectral linear prediction, properties and applications," IEEE Trans. Acoustics, Speech and Signal processing, pp.283-296, June1975.

[10] 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.*, vol. 12, pp. 247–251, 1993.