ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 6, No. 1, 2011, *pp*. 055-063

Normalized Autocorrelation based Features for Robust Speech Recognition in Context with Noisy Environment

Poonam Bansal[1], Amita Dev[2] and 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 3, 2009, accepted December 20, 2010*)

Abstract. This paper presents a robust approach for an automatic speech recognition system (ASR) when both additive and convolutional noises corrupt the speech signal. Robust features are derived by assuming that the corrupting noise is stationary and the channel effect is fixed during the utterance. In the proposed method the effect of additive and convolutional distortions are minimized by two stage filtering. The first filtering stage includes differential temporal filtering in the autocorrelation domain for reducing additive noise effects, followed by additional filtering in the logarithmic spectrum domain to reduce convolutional noise effects. Convolutional channel distortion is assumed to be linear and time invariant. A task of multispeaker isolated Hindi word recognition is conducted to demonstrate the effectiveness of using these robust features. The cases of channel filtered speech signal corrupted by white noise and different colored noises such as factory, babble and F16, which are further corrupted by channel distortion are tested. Experimental results show that the proposed method can significantly improve the performance of isolated Hindi word recognition system in noisy environment.

Keywords: ASR, Channel distortion, CMN, MFCC

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. One of the domains attracted attention in this regard is the autocorrelation domain. A number of feature extraction algorithms have been devised using this domain as the initial domain of choice. Considering a noisy speech, if the speech signal and noise are considered uncorrelated, then the autocorrelation of their sum is equal to the sum of their autocorrelations. Most important property of the autocorrelation sequence is considered to be the preservation of the original signal poles. 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 [1],[2]. Which means that the features extracted from the autocorrelation sequence could replace the features extracted from the original speech signal. Second property of the autocorrelation sequence is that as the autocorrelation of noise could in many cases be considered relatively constant over time, a high pass filtering of the autocorrelation sequence could lead to substantial reduction in its effect. Among the techniques used to exploit the autocorrelation properties are Short-time Modified Coherence [1], One-Sided Autocorrelation LPC (OSALPC) [3], Relative Autocorrelation Sequence (RAS) [4],[5] and Autocorrelation Me1 Frequency Cepstral Coefficient (AMFCC).

To overcome the problem of additive noise and the channel distortion, techniques proposed are parallel model combination (PMC) [6], Stochastic matching (SM) [7]-[9] and combining channel identification with power spectrum estimation [10],[11]. Recently, several novel techniques for handset and channel compensation are also proposed for speaker identification. [12] proposed a robust feature extraction using a

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

non-linear artificial neural network to optimize the speaker recognition performance. Some approaches have used magnitude spectrum of higher lag autocorrelation coefficients [13] and others have stressed upon preservation of spectral peaks [14]. Furthermore, for normalization various techniques are developed such as cepstral mean normalization (CMN), RelAtive SpecTrAl (RASTA) [15] and Blind Equalization (BE) [16]. Each of the above-mentioned autocorrelation-based methods has their own disadvantages. The PMC method needs a prior knowledge of the noise and SM takes time for iterative estimation of noise statistics. The RAS method, works well in low SNRs but does not perform as well in high SNRs. RAS and AMFCC methods considers distortion only due to additive noises, the convolutional noise in the form of channel distortion has not been considered. The AMFCC method works well for car and subway noises, but in babble and exhibition noises does not work as well. The reason could be that for the two later noise types, the noise properties are very similar to speech properties. Authors have already proposed a double fold additive noise suppression method for reducing the effect of additive noise. It is based upon the Differentiated Relative Autocorrelation Sequence Spectrum Mel Frequency Cepstral Coefficients (DRASS-MFCC) [17]. In this paper a novel dual filtering method on autocorrelation sequence is proposed to nullify the effects of both additive as well as convolutional noise. A temporal filtering method on autocorrelation domain is done to nullify the effect of additive noise plus an additional filtering has been suggested in the logarithmic spectrum domain by mean subtraction method to remove the channel effect. New derived parameters are called Channel Adaptive Relative Autocorrelation Sequence (CARAS). From the magnitude of CARAS the mel-scale frequency cepstral coefficients (MFCCs) of CARAS are derived. These MFCC are denoted as CARAS-MFCC. Comparisons in the recognition rate are made with DRASS-MFCC, RAS-MFCC and standard MFCC. It is well known that cepstral mean normalization (CMN) and delta cepstral coefficients are two effective methods for removing bias in traditional MFCC feature vector set. Here we have used CMN for that. CARAS-MFCC shows remarkable results in recognition accuracy for the signals corrupted by both additive as well as convolutional noises.

The remainder of the paper is organized as follows. Mathematical fundamentals for extracting RAS, DRASS and CARAS are derived in section 2. Block diagram of the proposed front end is described in section 3. In section 4 experiments conducted in clean and noisy environment with the proposed method are discussed. Finally a conclusion is given in section 5.

2. Robust features extraction

2.1. Extraction of RAS

Let m be the frame index and n be the time index within a frame. The clean speech x(m,n) corrupted by the additive noise u(m,n) result in a noisy speech expressed by

$$y(m,n) = x(m,n) + u(m,n),\ 0 \le m \le M-1, 0 \le n \le N-1 \qquad (1)$$

Where M denotes the number of frames in an utterance and N denotes the number of samples in a frame. 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.

$$r_{yy}(m,k) = r_{xx}(m,k\ ) + r_{uu}\ (m,k),\ 0 \le m \le M-1,\ 0 \le k \le N-1 \qquad (2)$$

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 (2) becomes

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

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

$$r_{yy}(m,k) = \sum_{i=0}^{N-1-k} y(m,i)\ y(m, i+k),\ 0 \le k \le N-1 \qquad (4)$$

Applying the temporal filtering on both sides of (3), it comes out

$$\Delta r_{yy}(m,k) = \Delta r_{xx}(m,k)\ ,\ 0 \le m \le M-1\ ,\ 0 \le k \le N-1 \qquad (5)$$

where

$$\Delta r_{yy}(m,k) = r_{yy}(m+1,k) - r_{yy}(m-1,k) \text{ and } \Delta r_{xx}(m,k) = r_{xx}(m+1,k) - r_{xx}(m-1,k)$$