Recognition and Enhancement of Speech under Noisy Conditions

Sai Rukmini Sahu, Janmejaya Rout  
Dept. of Electronics and Telecommunication Engineering  
Indira Gandhi Institute of Technology, Sarang  
Dhenkanal, Odisha  
sairukmanisahu12@gmail.com ,jrout.2006@gmail.com

Abstract—In this work we can see the influence of stress on speech signal and compare the features such as MFCC, LAR, LPC and recognize the speech using Vector Quantization (VQ) method under stressed or noisy conditions. The main objective of this paper is to robust the speech in diverse conditions and to get a robust speech we have to reduce the noise and enhance the speech. Here we have used Adaptive winner filter with DD, TSNR and HRNR methods for enhancement of noisy speech (car, babble, street noise) by considering SNR of 0 dB, 5 dB and 10 dB conditions and compared the method by considering CD (Cepstral Distance) and segmental SNR.

Keywords—speech recognition; MFCC; CD; speech enhancement; vector quantization; linear prediction coefficient;

I. INTRODUCTION

Speech enhancement problems in stress or noisy conditions is upcoming into picture while the speech corrupted with some environmental noise or when noisy is available in the speech. Now a day everywhere we required effective noise reduction technique, those required many approaches such as power spectral subtraction [1], winner filter [2], soft decision estimation [3]. In this paper first we recognize the signal under stressed condition such as Anger, Lombard, Neutral and question by using MFCC-VQ based recognition model under simulated and actual stress (SUSAS) database. The overall recognition model having two important operation such as signal modeling and pattern matching. First operation include signal preprocessing and feature extraction.

Fig. 1. Block diagram

A. signal preprocessing

End Point Detection  
Pre Emphasis  
Framing  
Windowing

Fig. 2. Preprocessing Structure

After voice signal sampled, the recognizer directly recognize the speech signal. As the speech signal is changing in nature and high redundancy of the samples we have to preprocess the signal to eliminate redundant information and extract useful information.  

II. SPEECH RECOGNITION

The speech characteristic data from a speaker’s sound is being extracted by a process called speech recognition[8]. By the growing advance technology speech recognition plays an important act in now a days. A perceptual characteristics by literal strains is called stress, some physiological acts from physiological effects is in respiratory variations such as rate increases, breathing discontinuity and Vocal cord problems etc [4]. Stressed speech recognition means recognition of speech under stress condition, here we have considered four stressed conditions such as Anger, Lombard, Neutral and Question and observed the recognition rate by using MFCC-VQ based recognition model under speech under simulated and actual stress (SUSAS) database.

The overall recognition model having two important operation such as signal modeling and pattern matching. First operation include signal preprocessing and feature extraction.
\[ E_n = \sum_{m=0}^{N-1} (W(m)x(n-m))^2 \]

(1)

Where, \( 0 \leq m \leq N - 1 \), \( x(n) \) denotes speech signal, the length of frame is \( N \), frame shift is \( m \), window function symbolize as \( W(m) \). Zero crossing rate is a different technique which has used in detection process. It is define as how many times a the signal waveform frame cross through horizontal axis. It is defined as

\[ z_n = \frac{1}{2} \sum_{m=0}^{N-1} \left| sgn[x(m)] - sgn[x(m+1)] \right| W(n-m) \]

(2)

Then pre emphasis is use to compensate the high frequency loss and it is also known as first order high pass filter [7]. After that as the speech signal is nonlinear in nature, that’s why for observation of the speech signal we can’t use LTIA method and so the original signal divided into small parts of continuous signal, because the speech signal of short time interval. After that to improve the frequency spectrum of the signal we have added the window function, which is use to select the desired part of the signal [8].

**B. Feature Extraction**

After preprocessing before going to recognition part we have to extract the characteristics features. Here we have taken 3 related features implemented through MATLAB programming such as LPC, MFCC and Log Area Ratio (LAR). In this paper we have used MFCC for feature extraction to recognize the signal [9].

![Feature extraction model](image)

Fig.3. Feature extraction model

The above model is MFCC or Mel frequency cepstral coefficient model, here the calculations is based on the Mel frequency scale rage. Here for calculating the MFCC coefficient, it is necessary to divided the whole speech signal into number of frames called framing, then to select the desired part of the signal Hamming window has used. Then FFT has used and by means of a filter-bank the spectrum is divided in to number of band. filter bank having overlapping triangular filters which can receive the relating frequency resolution nearer to human ear frequency resolution. Then by applying log and discrete cosine transform to filter bank signal we have got the MFCC features. And MFCC have much better identification then LPC and LAR.

**C. Pattern Matching**

After doing feature extraction we have to go for matching process, it is also known as classification and it is applicable for many speech processing techniques. There are various type of techniques or algorithm named as VQ, HMM, GMM, SVM etc. But here we have used VQ or Vector Quantization method. VQ is a technique which is use for modeling of PDF by splitting of sample vectors. First of all dividing a huge content of data into clusters or groups having approximately the same counts of data approximately same to them. The each content represented by Centroid point. The density matching property of this quantization technique is very powerful. Data points are shown by their closest centroid indexing, commonly stirring information contain little blunder, and unusual information soaring fault. Hence, this method is also appropriate for lossy data compression.

**III. ENHANCEMENT OF SPEECH**

Speech enhancement is define as improvement of speech which is corrupted by preservative surroundings noise. The surroundings noise having bad crash on our capability to contrary without proficiently in very noisy surroundings, such as busy streets, in a car and babble noise. Noise reduction is helpful in numerous applications such as tone message and ASR where well-organized technique are used for noise reduction [10]. As SNR is an important characteristic of a signal which gives information about background noise w. r. t signal. Here we estimate the a priori and posteriori SNR. From the activities of this estimator in [11] and verified that a priori SNR follow the outline of a posteriori SNR by means of a interruption of one fame. This unfairness is owed to the apply of the speech range predictable to calculate the present a priori SNR at prior frame. It reduces the presentation of enhancement process or noise reduction technique. This paper represents the enhancement of noisy speech using adaptive winner filter with DD, TSNR and HRNR method. These methods are mostly appropriate for estimation of priori SNR.
Fig. 4. Block Diagram
Figure 4 shows that the Adaptive Wiener filter using Decision Directed, Two Step Noise Reduction, and Harmonic Regeneration Noise Reduction techniques [12]. Let \( x(n) = s(n) + d(n) \) represents the noisy signal where \( s(n) \) and \( d(n) \) indicates speech signal and noise respectively. First noisy signal transformed by STFT to time domain using a Hamming window. \( X(p, k) \) where, \( p = \) frame index, \( k = \) sample index. In STFT, first noisy speech is decomposed into small frames. Frames are overlapped with previous frame. Each frame is passed through a smoothing window such as Hamming window. Then FFT is taken for signal spectrum, conversion from time to frequency domain. Magnitude and phase of noisy speech are taken from FFT. Noise parameters are estimated from noisy speech. Modification of Noisy speech magnitude is calculated using DD, TSNR and Harmonic Regeneration Noise Reduction (HRNR) techniques. After that we get a new magnitude of noisy speech. The new magnitude and phase is multiplied for Inverse Fast Fourier Transform (IFFT). 'Overlap and add' IFFT of each frame is used to generate enhanced speech [12].

Though, the estimated SNR is an important parameter which determines the speech enhancement efficiency under a known power spectral density of noise. Estimation of SNR parameter such as priori SNR and posteriori SNR requires by many noise reduction techniques.

\[
\text{SNR}_\text{post}(p, k) = \frac{|X(p, k)|^2}{E[|D(p, k)|^2]}
\]

(3)

\[
\text{SNR}_\text{prio}(p, k) = \frac{E[|S(p, k)|^2]}{E[|D(p, k)|^2]}
\]

(4)

Where \( s(p, k), D(p, k) \) and \( X(p, k) \) and represent the \( r^{th} \) spectral components of the short time frame \( p \) of the speech signal \( x(t) \), noise \( d(t) \), and noisy speech \( x(t) \). Again we have considered new parameter called as instantaneous SNR.

\[
\text{SNR}_{\text{inst}} = \frac{|X(p, k)|^2 - E[|D(p, k)|^2]}{E[|D(p, k)|^2]}
\]

(5)

A. Decision Directed method

In Decision Directed method, Estimation of DD Priori SNR of current frame follow the outline of the predictable priori SNR of DD with some delay i.e. previous frame and instantaneous SNR of current frame. Estimation of Priori SNR of current frame has given by

\[
\hat{S}N R_{\text{DD prio}}(p, k) = \beta \frac{|S(p-1, k)|^2}{P_n(p, k)} (1 - \beta) P[SN R_{\text{post}}(p, k) - 1]
\]

(6)

This estimation of priori SNR of current frame is called DD approach. Here \( \beta \) is control parameter i.e. 0.98. Now the DD Gain of current frame has given as

\[
G_{DD}(p, k) = \frac{SN R_{\text{DD prio}}(p, k)}{1 + SN R_{\text{DD prio}}(p, k)}
\]

(7)

DD Estimated magnitude of noisy speech current frame is

\[
\hat{S}_{\text{DD}}(p, k) = G_{DD}(p, k) X(p, k)
\]

(8)

B. Two Step Noise Reduction

The Two Step Noise Reduction method is capable to restrain the disadvantage of DD approach. TSNR is second step for noise reduction. Hence, it is called as Two Step Noise Reduction method. Estimation of TSNR Priori SNR of current frame follows “(6)” [14].

\[
SN R_{\text{TSNR prio}}(p, k) = \frac{|G_{\text{DD}}(p, k) + X(p, k)|^2}{E[|D(p, k)|^2]}
\]

(9)

TSNR Gain of current frame is given by

\[
G_{TSNR}(p, k) = \frac{SN R_{\text{TSNR prio}}(p, k)}{1 + SN R_{\text{TSNR prio}}(p, k)}
\]

(10)

TSNR Estimated magnitude of noisy speech current frame is

\[
\hat{S}_{\text{TSNR}}(p, k) = G_{TSNR}(p, k) X(p, k)
\]

(11)

C. Harmonic Regeneration Noise Reduction

As the output of signal \( \hat{x}_{\text{TSNR}}(p, k) \) is still having some distortion or degradation of signal. Each of the harmonic components of speech are lower in level or amplitude than the fundamental, here some harmonics are measured as covered up and noisy signal. Here to avoid this distortion Harmonic composition has used. The distorted signal is processed to regenerate a completely Harmonic signal. Thus, it is named as harmonic Regionalization Noise Reduction (HRNR) method [13]. HRNR approach is applicable to re-establish harmonics at the most wanted frequency range which is same as

\[\text{ISSN (Online) 2278-5841 ISSN (Print) 2320-5156} \]
noise free speech. And for that a nonlinear function \( NL \) is use to TSNR Enhanced speech. Then, the synthetically regeneratated speech is given by

\[
S_{\text{harmonic}}(t) = NL\hat{S}(t)
\]

(12)

\[
\text{SNR}_{\text{HRNR}}^\text{prior} (p,k) = \frac{g_{\text{TSNR}}(p,k)\cdot|\hat{S}(p,k)|^2 - (1 - g_{\text{TSNR}}(p,k))\cdot|S_{\text{harmonic}}(p,k)|^2}{E[|D(p,k)|^2]}
\]

(13)

HRNR gain of current frame is a function of HRNR estimated priori SNR

\[
G_{\text{HRNR}}(p,k) = \frac{\text{SNR}_{\text{HRNR}}^\text{prior} (p,k)}{1 + \text{SNR}_{\text{HRNR}}^\text{prior} (p,k)}
\]

(14)

HRNR Estimated magnitude of noisy speech current frame is

\[
\hat{S}(p,k) = G_{\text{HRNR}}(p,k) \cdot X(p,k)
\]

(15)

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this work, the database used is a stressed speech database which is popularly known as speech under simulated and actual stress (SUSAS) database. Here we will see how the speech is changing under stressed conditions. From this database four stress classes or emotions i.e. Anger, Lombard, Neutral and Question are used. It has an unique benefit. The database spoken by 9 American speakers. In this database each speaker has uttered each word under each stress. All speech signals are sampled at 8kHz. here we have considered one word (‘break1.sph’) of one speaker (‘nyc3’) under four emotions (Anger, Lombard, Neutral and Question) for recognition process.

Table I

<table>
<thead>
<tr>
<th>Emotions</th>
<th>Recognition rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>86</td>
</tr>
<tr>
<td>Lombard</td>
<td>91</td>
</tr>
<tr>
<td>Neutral</td>
<td>34</td>
</tr>
<tr>
<td>Question</td>
<td>27</td>
</tr>
</tbody>
</table>

Table. I show the recognition rate using MFCC-VQ Method; here we can see the highest recognition rate with Lombard i.e. 91% as compared to others. Why we have taken MFCC rather than LPC or LAR? Now we will see by comparing 3 features by considering recognition rate.
From above comparison we observed that under stressed condition the MFCC feature is most suitable to extract the features and gives high recognition rate or robust the speech signal under various noisy conditions.

Then we will go for enhancement of noisy speech signal. This has implemented through MATLAB programming. Here the signals used having sampling frequency 8khz. consequently, the subsequent parameters have been chosen: window length is 20ms, windows overlap 50%, and the size of FFT is 320. In the TSNR technique, the parameters are $\beta = 0.98$ and $\beta = 1$.

We have a clean speech and 3 different types of noise (car, babble and street) at different SNR such as 0 dB, 5 dB, and 10 dB. Here, we have considered 0 dB car noise for speech enhancement using Adaptive wiener filter. Fig 6(a) shows the noisy speech with 0 dB SNR car noise, (b) shows TSNR enhanced speech signal and (c) shows HRNR enhanced speech signal.

Table II shows the comparison results of TSNR and HRNR. Here cepstral distance has measured between clean speech and enhanced speech by TSNR and HRNR method. Here noisy signal under various SNR conditions have taken wit different noisy area such as car, babbleand street. From above compression we will confirm that HRNR method gives better enhancement signal in diverse conditions.
environment then TSNR which we can indicate by bolt value.

Table III

<table>
<thead>
<tr>
<th>Types of noise</th>
<th>given SNR (dB)</th>
<th>Sag. SNR</th>
<th>TSNR</th>
<th>HRNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td>0</td>
<td>3.44</td>
<td>3.67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>7.16</td>
<td>7.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10.33</td>
<td>11.12</td>
<td></td>
</tr>
<tr>
<td>Babble</td>
<td>0</td>
<td>4.95</td>
<td>5.41</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8.92</td>
<td>9.53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>13.85</td>
<td>14.53</td>
<td></td>
</tr>
<tr>
<td>Street</td>
<td>0</td>
<td>2.85</td>
<td>2.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>4.82</td>
<td>5.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>7.30</td>
<td>7.91</td>
<td></td>
</tr>
</tbody>
</table>

Table III presents the comparison results of taken methods such as TSNR and HRNR by considering segmental SNR. Here also see the better performance of HRNR then TSNR.

V. CONCLUSION

Here we observed that the effectiveness of noise robust features for recognition of speech under stress conditions
And also we have compared three traditional features such as LAR, MFCC and LPC and from the results we could see that feature extraction using MFCC is most appropriate to recognize the stressed speech signal using VQ model. But noise robust features are not necessarily reducing the external noise due to variability. The basic problem of noise reduction is to reduce the external noise without disturbing the unvoiced and low intensity noise like compensates of speech signal itself .Then we have reduce the noise to enhance of speech by adaptive winner filter with DD,TSNR and HRNR methods are used. TSNR methods are used to enhance the speech signal. Then HRNR is used to regenerate the harmonics which is lost from the original signal. Experimental result shows that HRNR method gave the best results as compared to TSNR by considering CD and segmental SNR.

ACKNOWLEDGEMENTS

This work was supported in part by a grant from the National Science Foundation.

REFERENCES


