SPEAKER VERIFICATION IN A NOISY ENVIRONMENT BY ENHANCING THE SPEECH SIGNAL USING VARIOUS APPROACHES OF SPECTRAL SUBTRACTION

Sriram.J, Umar Ali.S, Varadharajan.A, UG Students, Bharathi.B, Assistant Professor
Department of Computer Science and Engineering
SSN College of Engineering
Kalavakkam, Chennai-603110
E-mail: umar.ssn@gmail.com

Abstract— **This paper describes various noise reduction algorithms to analyse the performance of speaker verification in noisy environment. Speaker verification involves processing the speech signal and authenticating the speaker. The speaker verification process is affected by various factors such as noise, channel mismatch, health condition of the speaker, aging, emotion, fatigue etc. Speech signals from the uncontrolled environment may contain noise along with required speech components, speech signals are degraded by this noise and it renders the performance of speaker verification systems unacceptable. We propose to use noise reduction algorithms like Basic Spectral Subtraction, Spectral Subtraction with over subtraction, Non-Linear Spectral Subtraction and Multiband Spectral Subtraction and once the speech has been enhanced by each of the above methods, speaker models are trained for all speakers in the system and the performance of the verification system for the speech enhanced by each of the above methods is analysed.**

I. INTRODUCTION

Speech processing systems provide two major applications such as speaker identification and speaker verification. Speaker identification involves processing the voice signal and recognizing the speaker. Speaker verification involves authenticating the speaker. These two methods can be text dependent or text independent. Text dependent systems identifies the speaker based on a particular text while text independent systems identifies the speaker for a wide range of vocabulary. This speech signal is affected by various factors such as noise, channel mismatch, health condition of the speaker, long term variability in people's voice, emotion, fatigue etc. [9]. In this project, we focus only on noise as the major factor assuming all other factors meet their standards. Speech signals from the uncontrolled environment may contain noise along with required speech components. Speech signal degraded by additive noise, make the listening task difficult for a direct listener, giving poor performance in automatic speech processing tasks like speech recognition speaker verification, hearing aids, speech coders etc. [3]. The degraded speech therefore needs to be processed for the enhancement of speech components. The aim of speech enhancement is to improve the quality and intelligibility of degraded speech signal. Improving quality and intelligibility of speech signals reduces listeners fatigue; improve the performance of hearing aids, cockpit communication, videoconferencing, speech coders and many other speech systems. There are many noise reduction methods existing each method has its own advantages and disadvantages a literature survey is done to choose a method among the existing methods in Kalman Filtering method speech enhancement is done by recording the noisy speech and noise separately and then subtracting noise from the noisy speech but the problem with this method is that when the noise varies continuously in the environment this method is not feasible [8]. In Voice Activity Detection method speech enhancement is done by detecting the voiced and unvoiced parts of speech and the unvoiced part(i.e: silence parts) is considered as noise but always unvoiced part cannot be considered as noise because speech part may be affected by noise [19]. In this project, it is proposed to enhance the speech signal by using the principles of spectral subtraction algorithms, such as Basic Spectral Subtraction, Spectral Subtraction with over subtraction, Non linear Spectral Subtraction, Multiband Spectral Subtraction because they enhance the speech even if the noise is not constant in the environment and even if the noise affects the important spectral components of the speech [1]. Once the speech has been enhanced by each of the above methods , speaker models are trained for all speakers in the system and the performance of the verification system for the speech enhanced by each of the above methods is analysed.

The remainder of this paper is organized as follows. Section II describes the system architecture. Section III describes the methodologies. Section IV describes about training and testing. Section V presents performance analysis of speaker verification system. Finally, Section VI presents a conclusion of the paper.

II. SYSTEM ARCHITECTURE

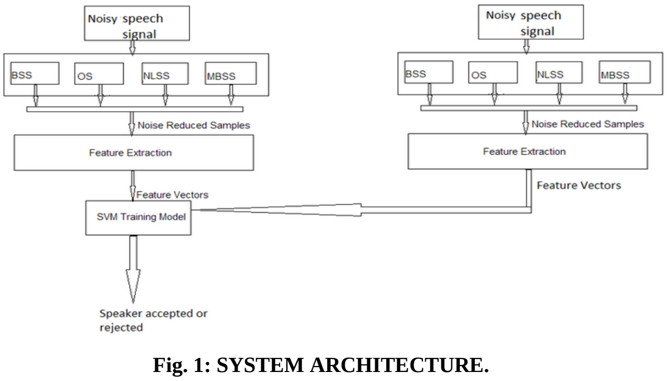

Fig. 1: SYSTEM ARCHITECTURE.

Speech samples(text dependent) of the speakers in the system are collected in a noisy environment and noise reduction algorithms such as Basic spectral subtraction(BSS), Spectral subtraction with over subtraction(OS), Non-linear spectral subtraction(NLSS) and Multiband spectral subtraction(MBSS) are applied to the speech samples. Features of the speech are extracted from noisy and noise reduced samples obtained from each of the above methods and it is trained with SVM(support vector machine) to get training model. Same process is done in testing part till feature extraction and the features are tested with the training model already created to verify that speaker is accepted or rejected.

III. METHODOLOGIES

A. Basic Spectral Subtraction(BSS)

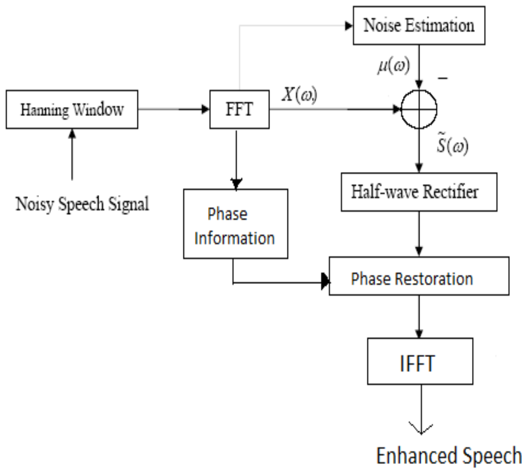

Fig. 2: BASIC SPECTRAL SUBTRACTION.

Speech is non-stationary signal where properties change quite rapidly over time.This is fully natural it makes the use of DFT impossible. For most phonemes the properties of the speech remain invariant for a short period of time (5-100ms). Thus for a short window of time, signal processing methods can be applied successfully. Most of speech processing is done in this way by taking short windows and processing them. The short window of signal like this is called frame. In speech recognition the windows are usually overlapping windows, which are analyzed in order to make hypothesis of the current phoneme. The hypotheses are combined over several frames and finally the decision is made to maximize the joint probability. It is a good idea to use overlapping windows summing approximately to 1. Hanning-window has the property of summing to 1 when the time difference between successive windows is half of the length of the window. So hanning window has been used for windowing in this system. After windowing process Fast Fourier Transform(FFT) by which the noisy signal in the time domain is converted to frequency domain to analyze the amplitude, frequencies and phase cleary because in the time domain the values cannot be analyzed clearly. The magnitude of the FFT of the signal is found which can be used further. The noisy signal is converted to frequency domain in which phase and magnitude present. Magnitude is affected by noise and phase is not affected by noise hence the phase is stored which is used in later stage for restoration of phase. The estimation of noise is done using Adaptive noise spectral estimation. In which the noise is estimated in frequency domain, the values of alpha and beta are set to 0.9 and 2 respectively and noise is estimated for each frequency bin in each frame based on condition

```
noimag = sigmag;
if the sigmag(l, k) > beta_noimag(l, k-1)
then
noimag(l, k) = noimag(l, k-1)
else
noiest = (1-alpha)*sigmag(l, k) + alpha*noimag(l, k-1)
noimag(l, k) = noiest;
```

where sigmag- magnitude of signal.
noimag- magnitude of noise.
noiest- estimate of noise.

The estimated noise is subtracted from the magnitude of the fft of the signal. If the estimated noise is greater than actual signal then subtraction lead to negative value in the magnitude so if the estimated noise is greater than actual signal then that part is replaced by zero this process is called as half wave rectification. After the subtraction of noise and half wave rectification, the phase which is already stored is restored to maintain phase information. After restoration of phase, inverse fourier transform is applied to convert frequency domain to time domain to get the enhanced speech.

B. Spectral subtraction with over subtraction(OS)

In basic spectral subtraction due to half-wave rectification process, small isolated peaks in the spectrum occurs at random frequency locations in each frame. Converted in the timedomain, these peaks sound like tones with frequencies that change randomly from frame to frame. This new type of noise introduced by the half-wave rectification process has been described as warbling and of tonal quality, and is commonly referred as musical noise. Hence over subtraction is a method to reduce musical noise. In this method if the estimated noise is greater than actual noise then the following formula is applied

if $|Y\ j(w)|^2 > (a+b)\ |De(w)|^2$

$|Xe\ j(w)|^2 = |Y\ j(w)|^2 - |De(w)|^2$

else

$b|De(w)|^2$

With $a >= 1$ and $0 < b <= 1$

$|Xe\ j(w)|$ denotes the enhanced spectrum

$|De(w)|$ is the noise spectrum
Where
a is over subtraction factor
b is the spetral floor parameter
If b is too large, then the residual noise will be audible but the musical issues related to spectral subtraction reduces. Parameter ’a’ affects the amount of speech spectral distortion. If a is too large then resulting signal will be severely distorted and intelligibility may suffer. If a is too small noise remains in enhanced speech signal. When a > 1, the subtraction can remove all of the broadband noise by eliminating most of wide peaks. So in this system, the value of is a is set as 1 and value of b is set as 0.3

C. Non-linear spectral subtraction(NLSS)

NLSS is a modification of the method which suggests to make the over subtraction factor frequency dependent and the subtraction process non-linear. In case of NLSS assumption is that noise does not affects all spectral components equally. Certain types of noise may affect the low frequency region of the spectrum more than high frequency region. This suggests the use of a frequency dependent subtraction factor for different types of noise. Due to frequency dependent subtraction factor, subtraction process becomes nonlinear. Larger values are subtracted at frequencies with low SNR levels and smaller values are subtracted at frequencies with high SNR levels. The subtraction rule used in the NLSS algorithm has the following form.

if $|Y(w)| > a(w)N(w) + b|De(w)|$
$|Xe(w)| = |Y(w)| - a(w)N(w)$
else
$b|Y(w)|$

Where
b is the spectral floor set to 0.1
$|Y(w)|$ and $|De(w)|$ are the smoothed estimates of noisy speech and noise respectively.
$a(w)$ is a frequency dependent subtraction factor
$N(w)$ is a non-linear function of the noise spectrum

where
$N(w) = Max (|De(w)|)$
The $N(w)$ term is obtained by computing the maximum of the noise magnitude spectra $|De(w)|$ over the frames.

The $a(w)$ is given as $a(w) = 1/r + p(w)$

where r is a scaling factor and $p(w)$ is the square root of the posteriori SNR estimate given as $P(w) = |Y(w)| / |De(w)|$

D. Multiband spectral subtraction(MBSS)

In MBSS approach the speech spectrum is divided into N overlapping bands and spectral subtraction is performed independently in each band. The process of splitting the speech signal into different bands can be performed either in the time domain by using band pass filters or in the frequency domain by using appropriate windows. The estimate of the clean speech spectrum in the ith band is obtained by

$|Xe_i(w_k)|^2 = |Y_i(w_k)|^2 – a_i d_i |D_i(w_k)|^2$
$b_i < (w_k) < e_i$
Where
$w_k = 2(pi)k/N$, k = 0, 1, ……N are the discrete frequencies
$|De_i(w_k)|^2$ is the estimated noise power spectrum obtained during speech absent segment
a_i is the over subtraction factor of ith band.
d_i is an additional band.
Subtraction factor can be individually set for each frequency band to customize the noise removal process
b_i and e_i are the beginning and ending frequency bins of the ith frequency band.The band specific over subtraction factor is a function of the segmented SNR_i of the ith frequency band and is computed as follows

4.75 $SNR_i < -5$

$a_i = 3/20(SNR_i) - 5 < SNR_i < 20$

1 $SNR_i > 20$

The values for d_i are set to

1 $f_i < 1$ KHz

$d_i = 2.5$ 1KHz < $f_i < (F_s/2)$2KHz

1.5 $f_i > (F_s/2)$2KHz

Where f_i is the upper frequency of the ith band and F_s is the sampling frequency in Hz.

E. Feature Extraction

MFCC(mel frequency cepstral coefficients) which gives the features of the speech are extracted from the speech signal.The process followed in extracting mfcc is

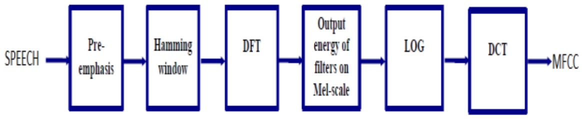

Fig. 3: FEATURE EXTRACTION.

Speech signal is segmented in to frames of 30 ms and overlapping factor is set to 256 and hamming window is applied. Then DFT is calculated for each frame. Filterbank Amplitudes is obtained by setting the setting number of filters as 20 in the filterbank and giving the length of fft and sampling frequency of each sample. Then magnitude of fft is multiplied with filterbank amplitudes.A set of 20 MFCC coefficients is extracted for each and every single frame after taking log and direct cosine transform.

IV. TRAINING AND TESTING

SVM(support vector machine) is used for training and testing. support vector machines are supervised learning models with associated learning algorithms that analyze the data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

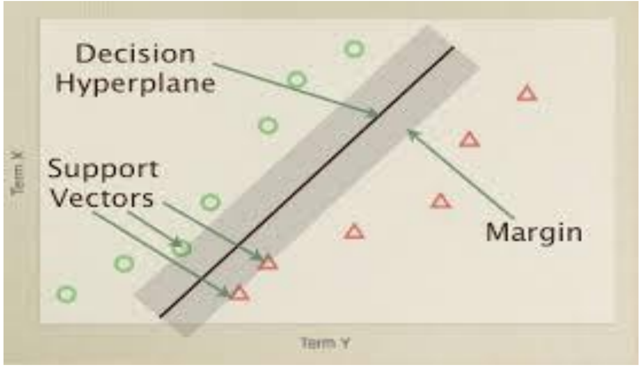

Fig. 4: SUPPORT VECTOR MACHINE.

Training and testing are done using the software for SVM. From the features of all the samples training and testing files are created for all the speakers in the system. Training for a speaker in svm gives svm-model as output. Testing is done in svm using the testing file and the svm-model obtained from training which gives the speaker verification accuracy and a file which contains true and false scores of verification.The true and false scores are the probability values of the speaker being actual or imposter i.e true or false.

V. PERFORMANCE ANALYSIS

A. Speech Data Collection

There are three speakers in the system for each speaker 50 samples are collected each sample with a duration of 3.5 seconds to 3.8 seconds. Samples are collected in the environment with **white noise** and **babble noise.** White noise is the fan noise and babble noise is the group of speakers talking at the background.

B. Metrics used

Two metrics are used for the performance analysis they are Signal to noise ratio(SNR) and Equal Error Rate(EER).

Signal to noise ratio(SNR) : SNR is used to analyse the amount of noise present in the speech signal

SNR= [Power of signal]/[Power of noise]

If the SNR value is low, amount of noise present in the signal is high as they are inversly proportional and vice versa. Average of SNR values of speech samples of all the speakers in the system for noisy and noise enhanced speech are calculated and they are compared in the tables 1and 2.

Speech	SNR
Noisy	-1.0865304
BSS	-1.069016933
OS	-1.086080467
NLSS	-1.081474867
MBSS	-1.082062133

TABLE 1: SNR COMPARISON IN WHITE NOISE ENVIRONMENT.

From table 1, Analysis of SNR values of all methods shows that BSS has the highest SNR value and hence speech samples enhanced by BSS will have less noise prevailing in white noise environment.

Speech	SNR
Noisy	-1.0937567
BSS	-1.0968016
OS	-1.0968015
NLSS	-1.0915377
MBSS	-1.0932961

TABLE 2: SNR COMPARISON IN BABBLE NOISE ENVIRONMENT.

From table 2, after comparing SNR values of all 4 methods, it was found that NLSS method has the highest SNR value in babble noise environment. This means that NLSS method will have less noise prevailing in babble noise environment.

Equal Error Rate(EER): False acceptance rate is the rate at which impostor speaker sample getting accepted and false rejection rate is the rate at which actual speaker sample getting rejected. A biometric security system predetermines the threshold values for its false acceptance rate and its false rejection rate and when the rates are equal the common value is referred to as the equal error rate. The value indicates that the proportion of false acceptances is equal to the proportion of false rejections. The lower the equal error rate value the higher the accuracy of the biometric system. Equal error rate is found by using the true scores and impostor scores of all the speakers in the system.

Equal Error Rate values in white noise environment:

Noisy Model	EER(%)
Noisy	17.08
BSS	11.04
OS	17.1
NLSS	15.62
MBSS	15.52

TABLE 3: EER VALUES FOR NOISY TRAINING MODEL AND NOISELESS TEST FILE.

Noisy Model	EER(%)
BSS	5.0
OS	15.62
NLSS	15.4
MBSS	15.4

TABLE 4: EER VALUES FOR NOISELESS TRAINING MODEL AND NOISELESS TEST FILE.

Table 3 and Table 4 shows that basic spectral subtraction method has the lowest EER value. This shows that BSS method is effective in speaker verification process in white noise environment.

Equal Error Rate values in babble noise environment:

Noisy Model	EER(%)
Noisy	5.0
BSS	2.71
OS	2.71
NLSS	2.64
MBSS	2.71

TABLE 5: EER VALUES FOR NOISY TRAINING MODEL AND NOISELESS TEST FILE.

Noisy Model	EER(%)
BSS	2.7
OS	2.7
NLSS	2.6
MBSS	2.7

TABLE 6: EER VALUES FOR NOISELESS TRAINING MODEL AND NOISELESS TEST FILE.

Table 5 and Table 6 shows that basic spectral subtraction method has the lowest EER value. This shows that BSS method is effective in speaker verification process in babble noise environment.

VI. CONCLUSION

Speaker verification process is affected by many factors such as noise, channel mismatch, aging etc. But we are considering only noise as a major factor in our project. Hence we use four methods such as basic spectral subtraction, over subtraction, non-linear spectral subtraction, multiband spectral subtraction for noise reduction. Analysis of all the four methods produced the following conclusions: For a environment with white noise BASIC SPECTRAL SUBTRACTION(BSS) has lowest EER thus BSS has the higher accuracy in environment with white noise.

For a environment babble noise NON LINEAR SPECTRAL SUBTRACTION(NLSS) has lowest EER thus NLSS has the higher accuracy in environment with babble noise and the remaining three methods has same EER in this environment. Hence the remaining three methods has same accuracy in babble noise environment.

REFERENCES

[1] Anuradha R. Fukane, Shashikant L. Sahare, 2011 “Different Approaches of Spectral Subtraction method for Enhancing the Speech Signal in Noisy Environments”, International Journal of Scientific & Engineering Research, Volume 2, Issue 5.ISSN 2229-5518.

[2] W. Alkhaldi, W. Fakhr and N. Hamdy, 2002 “Automatic Speech/Speaker Recognition In Noisy Environments Using Wavelet Transform”, IEEE Conference on Circuits and Systems. , vol. 1, pp 463-466.

[3] Andrzej Drygajlo and Mounir El-Muliki, 1998 “Speaker Verification in Noisy Environments with Combined Spectral Subtraction and Missing Feature Theory”, IEEE Trans. Acoustics,Speech and Signal Processing.,vol. 1, pp. 121-124.

[4] Bachu R.G., Kopparthi S., Adapa B., Barkana B.D., “Seperation of Voiced and Unvoiced using Zero Crossing rate and Energy of the Speech signal”, Electrical Engineering Department School of Engineering, University of Bridgeport.

[5] F. Capman, J. Boudy, P. Lockwood, “Acoustic Echo Cancellation and Noise Reduction in the Frequency-Domain: A Global Optimisation”, Matra Communication, Speech Processing Department.

[6] Hanspeter Schmid, 2012 “How to use the FFT and Matlab’s pwelch function for signal and noise simulations and measurements”, University of applied sciences NorthWestern Swizerland School of Engineering.

[7] H.T. Hu and C. Yu, 2007, “Adaptive noise spectral estimation for spectralsubtraction speech enhancement”.

[8] Bc. Jan Kybic, 1998 “Kalman Filtering and Speech Enhancement”, Diploma work.

[9] Ji Ming, Timothy J. Hazen, James R. Glass, Douglas A. Reynolds, 2007 “Robust Speaker Recognition in Noisy Conditions”, IEEE Trans. Audio, Speech and Language Processing., vol.15, no.5, pp 1711-1723.

[10] Marwa A. Abd El-Fattah ·Moawad I. Dessouky ·Alaa M. Abbas ·Salaheldin M. Diab ·El-Sayed M. El-Rabaie ·Waleed Al-Nuaimy ·Saleh A. Alshebeili ·Fathi E. Abd El-samie, 2013 “Speech enhancement with an adaptive Wiener filter”, Journal from Faculty of Electronic Engineering, Menoufia University,Egypt.

[11] Mohammad Sadgh Talebi, “Frequency Domain Signal Processing Using MATLAB”, Sharif University of Technology.

[12] Poonam Sharma and Abha Kiran Rajpoot, “Automatic Identification of Silence,Unvoiced and Voiced Chunks in Speech”, Computer Science Engineering Department Sharda University, Greater Noida.

[13] Rafael C. Gonzalez Richard E. Woods, “AN INTRODUCTION TO MATRIX MANIPULATION IN MATLAB”, Digital Image Processing, Second Edition.

[14] Ramin Halavati, Saeed Bagheri Shouraki,Mina Razaghpour, Hossein Tajik, Arpineh Cholakian, 2006 “A Novel Noise Immune, Fuzzy Approach To Speaker Independent, Isolated Word Speech Recognition” published in World Automation Congress(WAC), Conference in Budapest.

[15] Ricardo A. Losada, 2008 “Digital Filters with MATLAB”, The MathWorks,inc.

[16] Saeed V. Vaseghi, “Advanced Digital Signal Processing and Noise Reduction”, Second Edition. , ISBNs: 0-471-62692-9.

[17] Sahare Anuradha R. Fukane, Shashikant L. Sahare, 2011 “Noise estimation Algorithms for Speech Enhancement in highlynon-stationary Environments” , International Journal of Computer Science Issues, vol. 8, Issue 2, ISSN 1694-0814.

[18] Serajul Haque, Roberto Togneri, 2010 “A Psychoacoustic Spectral Subtraction Method For Noise Suppression In Automatic Speech Recognition”, IEEE International Conference on Acoustics Speech And Signal Processing. ,pp 1618-1621.

[19] E. Verteletskaya, K. Sakhnov, 2010 “Voice Activity Detection for Speech Enhancement Applications”, Acta Polytechnica Vol. 50 No. 4.

[20] .www.mathworks.in

[21] .www.stackoverflow.com

[22] .http://www.uni-oldenburg.de/en/medicine/departments/mediphysics-acoustics/speech-signal-processing/publications/noise-power-estimation/

[23] .https://home.cse.ust.hk/~tomko/useful_technic/wav2mfcc.txt

[24] http://www.dsprelated.com/showmessage/37269/1.php