Non - Audible Murmur (NAM) Voice Conversion by Wavelet Transform

K. Kalaiselvi[1], M. S. Vishnupriya[2]

[1]PG student, M.E. Embedded system technologies,
Sri Muthukumaran Institute of Technology, Anna University, Chennai-600069, India

[2]Assistant Professor, Department of Electrical and Electronics Engineering,
Sri Muthukumaran Institute of Technology, Anna University, Chennai-600069, India

Abstract: ***In this paper, a statistical approach is presented to enhance the body-conducted unvoiced speech for silent speech communication. Body conducted unvoiced speech is difficult to use in human-to-human speech communication because it sounds unnatural and less intelligible owing to the acoustic change caused by body conduction. To address this issue, voice conversion (VC) method from NAM to normal speech (NAM-to-Speech) using Wavelet transform is proposed, where the acoustic features of body-conducted unvoiced speech are converted into those of natural voices. Wavelet analysis is capable to extract the data from the murmur and then it passes the data into a classifier for the recognition of isolated words. Simulation results show that NAM-to-Speech effectively improves intelligibility.***

Keywords: Nam microphone, Discrete Wavelet transform, voice conversion, Flat top window**.**

1. Introduction

Speech communication plays a very important role in our daily life. In recent decades, the style of speech communication has considerably changed with the advancement of information technology. For instance, the explosive spread of cell phones has enabled people to talk with each other whenever and wherever they want and has brought a more convenient style of speech communication to us. Although cell phones have made speech communication possible in various situations, there are actually some instances where we face difficulties in speech communication.

We would have trouble privately talking in a crowd; speaking itself would sometimes annoy others in quiet environments such as in a library; and we may lose our voice if subjected to surgery to remove speech organs such as the larynx due to laryngeal cancer. Many barriers still exist in speech communication. The development of technologies to overcome these inherent problems of speech communication is essential to make our speech communication more universal.

A near-time algorithm was used to recognize whole sentences from continuous tongue and lip movements. A set of words classified by the movements of sensors affixed to the tongue and lips [1]. Instead of air-conductive microphones, a silent speech interface can be implemented using tongue and lip imaging. Ultrasound data and video data recorded synchronously at their respective maximum frame rate together with the acoustic speech signal. The data streams are recorded, processed and stored digitally on a single PC using our stand-alone software *Ultraspeech* [2].

Speech support system retrieves speech via a transfer function from body-conducted speech, using a recognition method to determine sub-word sequences and duration time [3]. The modified speaker adaptive training (SAT) methods are used for silent speech interface, which is capable of using a larger amount of normal speech data by transforming them into NAM data [4]. The NAM speech also recognized by fusing the visual information extracted from the talkers' facial movements with NAM speech extracted by HMM method [5].

As one of the microphones to detect body-conducted speech, the Non-Audible Murmur (NAM) microphone has been developed by Nakajima. Inspired by a stethoscope, the NAM microphone was originally developed to detect extremely soft murmur called NAM, which is so quiet that people around the speaker barely hear its emitted sound. Although NAM is a truly different medium from natural voices, it can be used easily by anyone whose speech organs function reasonably well.

2. Non-Audible Murmur (NAM) Microphones

Non-audible murmur (NAM) is an unvoiced speech that can be received through the body tissue using special acoustic sensors (i.e., NAM microphones) attached behind the talkers ear. By attaching the NAM microphone behind the talker’s ear, the microphone can able to capture an inaudible murmur (NAM speech), which cannot be heard by listeners near the talker. Privacy, robustness to environmental noise, and a tool for sound-impaired people are the advantages of NAM microphone, when it is applied in a speech recognition system. Although the NAM signal is of poor quality, the signal envelope is similar to that of normal speech, and therefore speech recognition is possible.

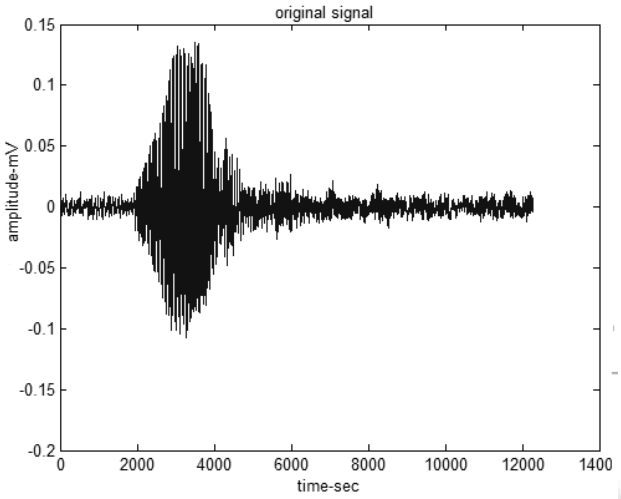

Figure 1: Normal Speech signal

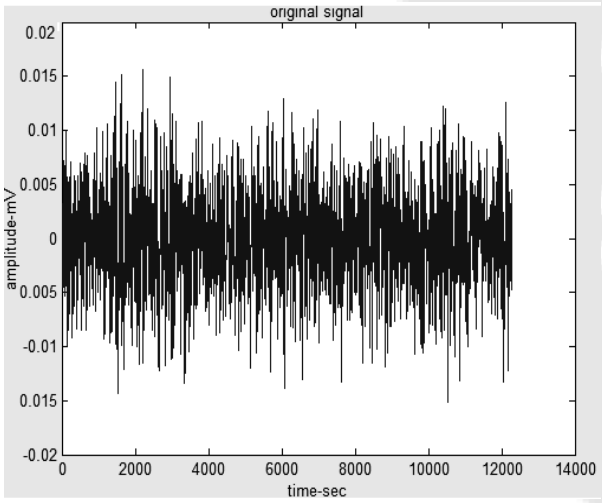

Figure 2: NAM speech signal

Figures 1 & 2 shows normal speech and NAM speech waveform respectively. For the recording of normal speech, a close-talking microphone was used. For the NAM speech, a NAM microphone was used. In the case of the NAM signal only the low frequency components are appeared and due to the body transmission some components are lost, therefore the NAM speech is of lower quality compared to the normal speech.

It is capable of detecting air vibrations in the vocal tract from the skin through only the soft tissues of the head. High-quality body-conductive recording of various types of speech, such as a very soft murmur as NAM, a whispered voice, soft voices, and normal speech, is possible from this position because the conduction through obstructions, Such as bones whose acoustic impedance is different from that of soft tissues, is avoided. It is also robust against external noise owing to its noise-proof structure like in other body conductive microphones.

Therefore, the NAM microphone allows us to talk using various types of body-conducted speech depending on the situation, e.g., NAM or a body-conducted whispered voice (BCW) for silent speech communication and body-conducted normal speech for noise-robust speech communication. Moreover, its usability is better than those of other devices such as EMG or ultrasound systems.

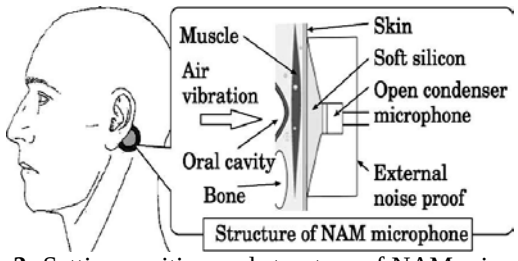

Figure 3: Setting position and structure of NAM microphone

3. Wavelet Transform

The Wavelet Transform provides a time-frequency representation of the signal. Wavelet transform has ability to analyse different speech quality problems simultaneously in both time and frequency domain. The wavelet transform is useful in extracting the features of various types of speech signal. A wave is an oscillating function of time or space and is periodic. In contrast, wavelets are localized waves. They have their energy concentrated in time or space and are suited to analysis of transient signals. While Fourier Transform and STFT use waves to analyse signals, the Wavelet Transform uses wavelets of finite energy.

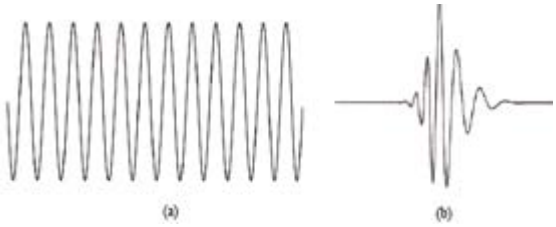

Figure 4: Demonstration of (a) a wave and (b) a wavelet

The fundamental idea of wavelet transforms is that the transformation should allow only changes in time extension, but not shape. Based on the uncertainty principle of signal processing,

$$\Delta T * \Delta W \geq 1/2 \quad (1)$$

Where, t represents time and ω angular velocity (ω = 2*Pi*frequency).

The higher the resolution in time is required, the lower resolution in frequency has to be. The larger the extension of the analysis windows is chosen, the larger is the value of Δt.

3.1 Types of Wavelet Transform

- Continuous wavelet transform
- Discrete wavelet transform

3.1.1 Continuous Wavelet Transform (CWT)

In order to analyze signals of very different sizes, it is necessary to use time-frequency atoms with different time supports. The wavelet transform decomposes signals over dilated and translated functions called wavelets, which transform a continuous function into a highly redundant function.

3.1.2 Discrete Wavelet Transform (DWT)

The Wavelet Series is just a sampled version of CWT and its computation may consume significant amount of time and resources, depending on the resolution required. Discrete wavelet transform overcomes the disadvantage of generating large amount of wavelet coefficients as of in continuous wavelet transform. The Discrete Wavelet Transform (DWT) is based on sub-band coding is found to yield a fast computation of Wavelet Transform. It is easy to implement and reduces the computation time and resources required. The DWT analyses the signal at different frequency bands with different resolutions by decomposing the signal into a coarse approximation and detail information. DWT employs two sets of functions, called scaling functions and wavelet functions, which are associated with low pass and high pass filters, respectively. Half band low pass filtering removes half of the frequencies, which can be interpreted as losing half of the information. Therefore, the resolution is halved after the filtering operation. However, the subsampling operation after filtering does not affect the resolution, since removing half of the spectral components from the signal makes half the number of samples redundant. Half the samples can be discarded without any loss of information. The low pass filtering halves the resolution, but leaves the scale unchanged. The signal is then subsampled by 2 since half of the numbers of samples are redundant. This doubles the scale.

This procedure can mathematically be expressed as

$$y[n] = \sum_{k=-\infty}^{\infty} h[k].x[2n-k] \quad (2)$$

The decomposition of the signal into different frequency bands is simply obtained by successive high pass and low pass filtering of the time domain signal. The original signal x[n] is first passed through a half band high pass filter g[n] and a low pass filter h[n]. This constitutes one level of decomposition and can mathematically be expressed as follows:

$$y_{high}[k] = x[n].g[2k-n] \quad (3)$$

$$y_{low}[k] = x[n].h[2k-n] \quad (4)$$

Where, y_{high} [k] and y_{low} [k] are the outputs of the high pass and low pass filters respectively, after subsampling by 2.

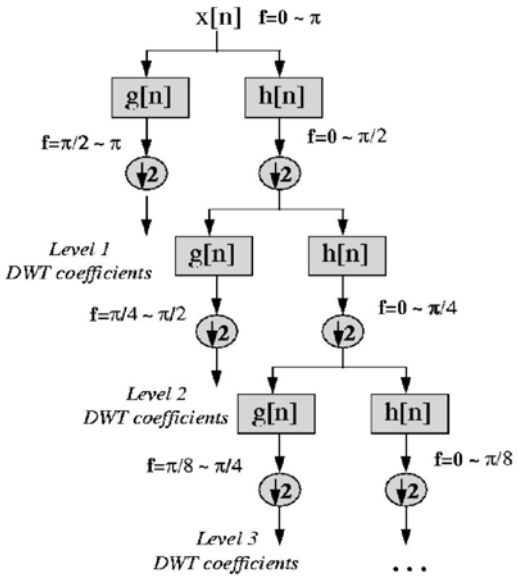

Figure 5: Illustration of decomposition procedure in DWT

Where x[n] is the original signal to be decomposed, and h[n] and g[n] are low pass and high pass filters, respectively. The bandwidth of the signal at every level is marked on the figure as "f".

At every level of decomposition, the filtering and subsampling will result in half the number of samples (and hence half the time resolution) and half the frequency band spanned (and hence doubles the frequency resolution). While computing the DWT, discard all values in the DWT coefficients that are less than a certain threshold and save only those DWT coefficients that are above the threshold for each frame.

The reconstruction in this case is very easy since half band filters form orthonormal bases. The above procedure is followed in reverse order for the reconstruction. The signals at every level are up sampled by two, passed through the synthesis filters g[n], and h[n] (high pass and low pass, respectively), and then added. The interesting point here is that the analysis and synthesis filters are identical to each other, except for a time reversal.

4. System Design

4.1 Algorithm

Thus the algorithm can be summarized as

- Pre-processing the input signal
- Framing input speech signal
- Filtering the frames using windowing techniques.
- DWT of a frame
- Thresholding wavelet coefficients
- Inverse DWT

4.2 Block Diagram

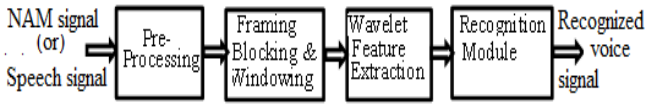

Figure 6: Block diagram for NAM to normal speech signal conversion

4.3 Pre- Processing

Pre-processing of Speech Signal serves various purposes in any speech processing application. It includes Noise Removal, Endpoint Detection, Pre-emphasis, Framing, Windowing, Echo Cancelling etc. Out of these, silence/unvoiced portion removal along with endpoint detection is the fundamental step for applications like Speech and Speaker Recognition.

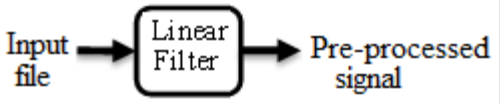

Figure 7: Block diagram for pre-processing

Pre-Processing of speech signals, i.e. segregating the voiced region from the silence/unvoiced portion of the captured signal is usually advocated as a crucial step in the development of a reliable speech or speaker recognition system. This is because most of the speech or speaker specific attributes are present in the voiced part of the speech signals; moreover, extraction of the voiced part of the speech signal by marking and removing the silence and unvoiced region leads to substantial reduction in computational complexity.

4.4 Frame & Windowing Blocking

The input speech signal is segmented into frames of 20~30 ms with optional overlap of 1/3~1/2 of the frame size. Usually the frame size (in terms of sample points) is equal to power of two in order to facilitate the use of FFT. If this is not the case, we need to do zero padding to the nearest length of power of two. If the sample rate is 16 kHz and the frame size is 320 sample points, then the frame duration is 320/16000 = 0.02 sec = 20 ms. Additional, if the overlap is 160 points, then the frame rate is 16000/(320-160) = 100 frames per second.

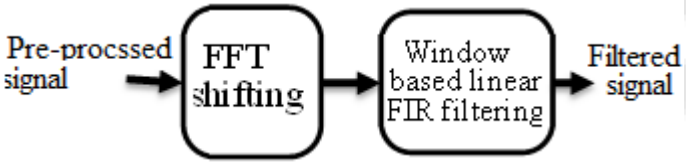

Figure 8: Block diagram for framing and windowing

4.5 Windowing

Windowing is the process of taking a small subset of a larger dataset, for processing and analysis.
Consider the system *H(z)*, with input *X(z)* and output *Y(z)*.

We model this as:

$$Y(z) = X(z).H(Z) \quad (5)$$

If we have a window with transfer function *W(z)*, we can mathematically apply the window to our signal, *X(z)* as such:

$$\hat{X}(z) = X(z).W(z) \quad (6)$$

Then, we can pass our windowed signal into our system, *H(z)* as usual:

$$\hat{Y}(z) = \hat{X}(z).H(z) \quad (7)$$

4.5.1 Hamming Windowing

Each frame has to be multiplied with a hamming window in order to keep the continuity of the first and the last points in the frame. If the signal in a frame is denoted by s(n), n = 0,…N-1, then the signal after Hamming windowing is s(n)*w(n), where w(n) is the Hamming window defined by:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad (8)$$

4.5.2 Rectangular Windowing

The rectangular window (sometimes known as the **boxcar** or **Dirichlet window**) is the simplest window, equivalent to replacing all but *N* values of a data sequence by zeros, making it appear as though the waveform suddenly turns on and off:

W(n)=1 for n=0,1,…….N-1

4.5.3 Flattop Windowing

Flat Top windows have very low pass band ripple (< 0.01 dB) and are used primarily for calibration purposes. Their bandwidth is approximately 2.5 times wider than a Hann window. A flat top window is a partially negative-valued window that has a flat top in the frequency domain. Such windows have been made available in spectrum analyzers for the measurement of amplitudes of sinusoidal frequency components. They have a low amplitude measurement error suitable for this purpose, achieved by the spreading of the energy of a sine wave over multiple bins in the spectrum. This ensures that the unattenuated amplitude of the sinusoid can be found on at least one of the neighbouring bins. The drawback of the broad bandwidth is poor frequency resolution. To compensate, a longer window length may be chosen.

Flat top windows can be designed using low-pass filter design methods or they may be of the usual sum-of-cosine-terms variety. An example of the latter is the flat top window available in the Stanford Research Systems (SRS) SR785 spectrum analyzer

$$w(n) = a0 - a1\cos\left(\frac{2\pi n}{N-1}\right) + a2\cos\left(\frac{4\pi n}{N-1}\right) - a3\cos\left(\frac{2\pi n}{N-1}\right) + a4\cos\left(\frac{8\pi n}{N-1}\right)$$

where a0=1; a1=1.93; a2=1.29; a3=0.388; a4=0.028

4.6 Feature Extraction

In speaker independent speech recognition, a premium is placed on extracting features that are somewhat invariant to changes in the speaker. So feature extraction involves analysis of speech signal. Broadly the feature extraction techniques are classified as temporal analysis and spectral analysis technique. In temporal analysis the speech waveform itself is used for analysis. In spectral analysis spectral representation of speech signal is used for analysis. Then the word can be recognized using the above extracted feature from which the isolated word can be obtained.

5. Steps to Convert Nam Speech Into Normal Speech

- Initially Cut-off frequency, analyzing time and filter order is fixed to particular value and the Input signal is filtered by using linear filter. The output of filter signal is termed as pre-processed signal.
- The Pre processed signal is segmented into frames using FFT shifting.
- Apply the windowing process for each frame for analysis purpose. Three different types of windowing techniques are applied.
- Apply Discrete Wavelet Transform on each windowing output to extract the wavelet features extraction.
- Signal is decomposed into many level by entering the decomposition level, the output signal is obtained.
- The signal is reconstructed from the decomposed signal to give the normal speech signal.

6. Simulation Results

The normal speech signals and NAM signals obtained by the sigview is analyzed using MATLAB.

6.1 Analysis of Normal Speech

Vowel A

Filter order = 12
Cut-off freq =500 Hz
Sampling rate = 11025
Number of bits = 16 bps
Analysis time =1.1146 sec

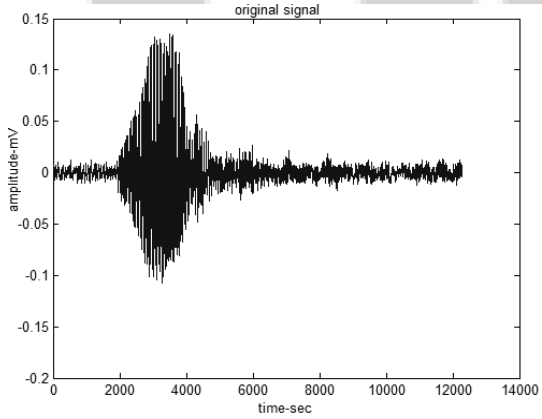

Figure 9: Original signal waveform for 'normal speech a'

The above magnitude plot shows the original signal obtained from the person using microphone

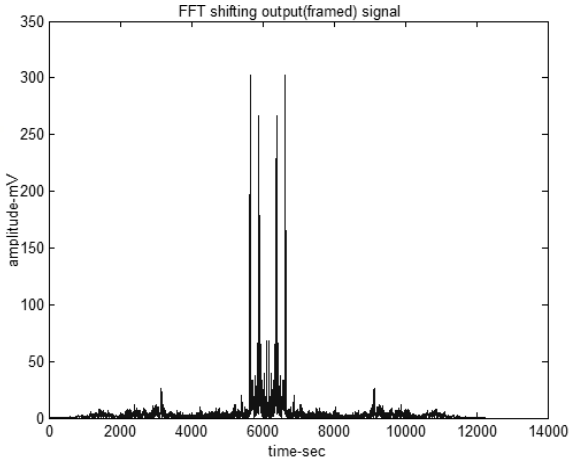

Figure 10: FFT shifting (framed) signal waveform for 'normal speech a'

The input signal is divided into many frames by using FFT shifting method for the analysis purpose.

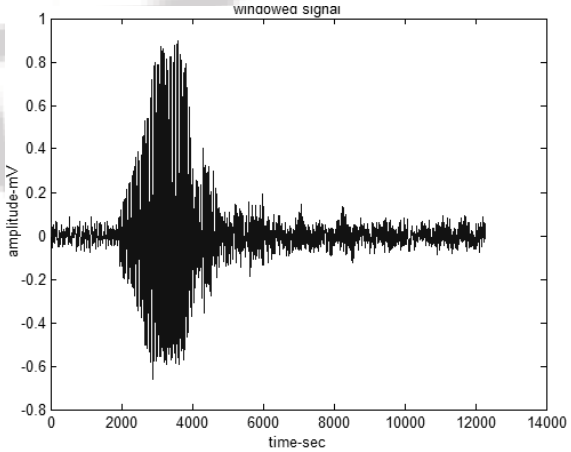

Figure 11: Flat top windowed based FIR filter output signal in amplitude plot for normal speech 'a'

Each frames of signal is filtered by using flat top window based FIR filter

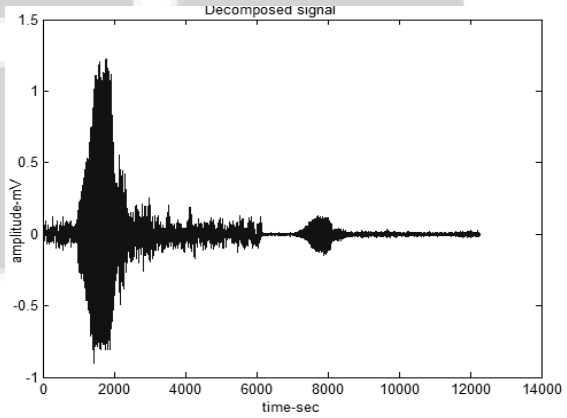

Figure 12: Decomposed signal waveform for 'normal speech a'

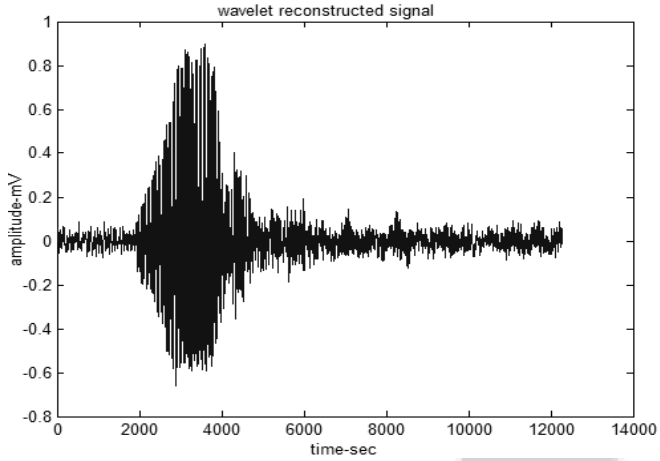

Figure 13: Reconstructed signal waveform for 'normal speech a'

Mean: m = 1.41E-04
Standard deviation: st .dev=0.1439

6.2 ANALYSIS OF NAM SPEECH

Vowel A

Filter order =12
Cut-off freq =500 Hz
Sampling rate = 11025
Number of bits = 16 bps
Analysis time =1.1146 sec

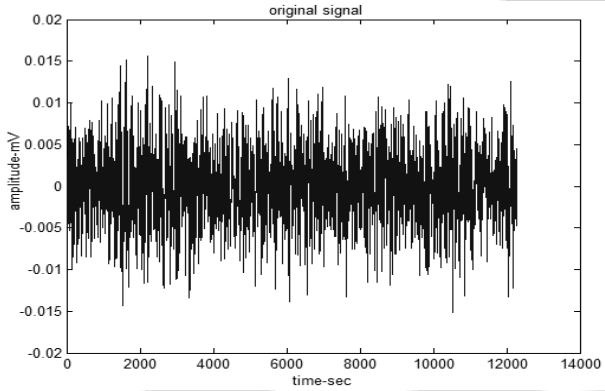

Figure 14: Original signal waveform for 'NAM speech a'

The above magnitude plot shows the original signal obtained from the person using NAM microphone.

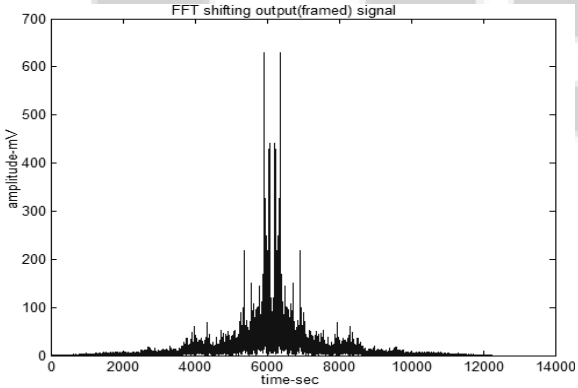

Figure 15: FFT shifting (framed) signal waveform for 'NAM speech a'

The input signal is divided into many frames by using FFT shifting method for the analysis purpose.

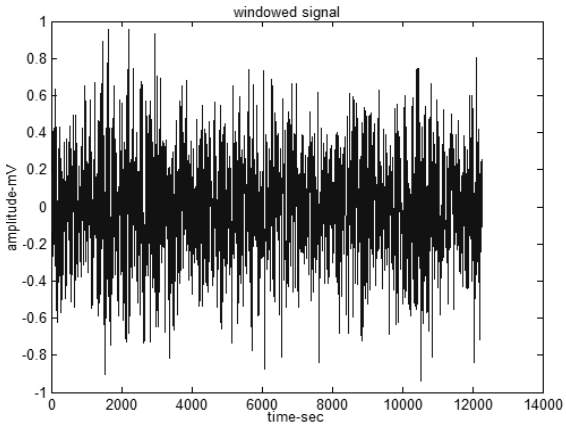

Figure 16: Flat top windowed based FIR filter output signal in amplitude plot for 'NAM speech a'

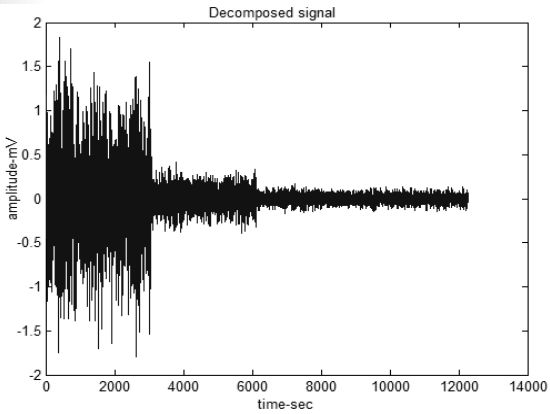

Figure 17: Decomposed signal waveform for 'NAM speech a'

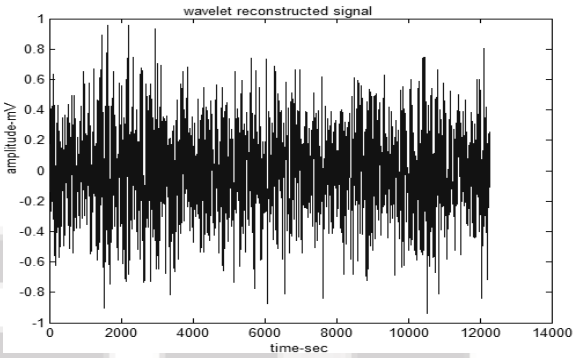

Figure 18: Reconstructed signal waveform for 'NAM speech a'

Mean: m = -8.39E-04
Standard deviation: std .dev=0.2667

6.3 Comparison of NAM and Normal speech using standard deviation

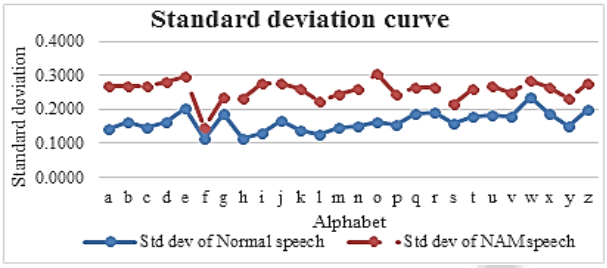

Figure 19: Standard deviation curve for NAM and Normal speech

Using the above comparison curve, the pronounced alphabet can be recognized from the NAM signal.

7 Conclusion

Thus the speech signal is recognized from the NAM speech by using wavelet transform, using MATLAB. The normal speech signal and NAM signals were analyzed by using wavelet transform in different decomposition level and the accurate speech signal is obtained by comparing the standard deviation values. Literature and simulated outputs are verified.

References

[1] Jun Wang, Ashok Samal, Jordan R. Green, Frank Rudzicz,"Sentence recognition from articulatory movements for silent speech interfaces",2012.
[2] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M.Stone," Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips". Speech Communication, Vol.52, No. 4, pp. 288–300, 2010.
[3] Shunsuke Ishimitsu, Kouhei Oda and Masashi Nakayama,"Body-conducted speech recognition in speech support system for disorders" August, 2011.
[4] Denis Babani,Tomoki Toda, Hiroshi Saruwatari,Kiyohiro Shikano," Acoustic model training for non-audible murmur recognition using transformed normal speech data",2011.
[5] P. Heracleous, V.-A. Tran, T. Nagai, and K. Shikano, "Analysis and recognition of NAMspeech using HMM distances and visual information," IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 6, pp. 1528–1538, Aug. 2010.

Author Profile

K. Kalaiselvi earned her B.E degree in Electronics and Communication Engineering, from Anna University, Chennai, in 2011 and she pursing her M.E degree in Embedded System Technologies in Anna University.

M. S. Vishnupriya received her B.E degree in Electrical and Electronics Engineering from the University of Madras, in 2001 and M.E degree in Applied Electronics from Anna University, Chennai in 2007. During 2007-2011, she worked as a lecture in Meenakshi College of Engineering. She is currently as an Assistant professor in Sri Muthukumaran Institute of Technology, Anna University, Chennai. Her field of interest is Neural Network, Digital Image Processing and Drives and control in Power electronics.