International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

The Use of Throat Microphone Speech Signal to Find Pitch and Voiced/ Unvoiced Regions Using Empirical Mode Decomposition (EMD) in Noisy Environment

Rakhi Rani Paul*[1], Subrata Kumer Paul*[2]

Lecturer, Dept. of Computer Science and Engineering, Bangladesh Army university of Engineering and Technology(BAUET), Qadirabad Cantonment Natore, Bangladesh.

Abstract - *The paper presents that Throat Microphone Speech Signal can be used for finding pitch and voiced/ unvoiced regions in a heavy noisy environment. Pitch estimation is commonly employed in voice quality analysis. The accuracy of pitch extraction is lower due to the influence of noises in a noisy environment. The Speech Signal is recorded by the Throat Microphone mounted on the neck. The microphone transforms the vibrations into signals that it picks up into equivalent speech signals. Typically, the throat speech is a high amplitude signal compared to the Acoustic Microphone Speech signal. Since Throat Microphone Speech Signal is relatively more robust to environmental variations, it can be used to detect Voiced/Unvoiced regions of the Speech. Empirical mode decomposition (EMD) is employed for detecting voiced/unvoiced speech regions and also for finding pitch from Throat Microphone Speech Signal. The experimental results show that the performance of the proposed approach is prominent as compared to other reported approaches.*

Key Words: **Acoustic Microphone, Automatic Speech Recognition, Empirical Mode Decomposition, Intrinsic Mode Functions, Pitch, Throat Microphone, Voiced/ Unvoiced Regions.**

1. INTRODUCTION

In many speech processing applications, the voiced/ unvoiced (V/UV) detection of a speech signal is a crucial pre-processing step. It is essential in most synthesis and analysis systems. The principle of classification is to discover whether the speech production system involves the vibration of the vocal cords [1]. Speech recognizers achieve quite enormous recognition rates now a day. Noise robust-ness is one of the biggest problems that remain in the Automatic Speech Recognition (ASR) domain. Undesired back-ground signals corrupt the desired speech signal. It causes a mismatch between that speech signal and the training data of the acoustic models of the speech recognizer. This problem points to degraded recognition performance. The added undesired background signals also cause another problem that is more difficult to tell at what times exactly the user is speaking. This so-called Voice Activity Detection is a very important aspect of Automatic Speech Recognition which tells the recognizer when it has to listen to the input signal [2].

T. Dekens, W. Verhelst, F. Capman and F. Beaugendre [2] show that bone-conducted voice can be used in noisy environments to improve Automatic Speech Recognition. They conducted experiments where they used a throat micro-phone signal as a Voice Activity Detection (VAD) input signal and found .That recognition accuracies in non-stationary noise improve significantly compared to when VAD is executed on a conventional microphone signal.

M. K. I. Molla, K. Hirose, N. Minematsu and M. K. Hasan [3] show that a new approach for voiced/unvoiced discrimination based on the extraction of pitch period. EMD is employed for multi-band representation of speech signal in the time domain. The main fractional energy contributed by the oscillations with pitch period in different ACFs is used as the factor to classify a speech segment as V/UV.

In this paper we will try to find an efficient way to detect voiced/unvoiced regions as well as pitch of speech signal in heavy noisy environment. The use of throat microphone speech is very efficient in this case.

2. ALGORITHMS

Empirical mode decomposition (EMD) is formed to decompose a non-stationary and nonlinear signal into oscillating components that follow some basic properties [4, 5]. The main advantage of using EMD method is that it is an automatic decomposition and fully data-adaptive method. The principle of the EMD technique is to decompose a signal S(i) into a sum of the band-limited functions called intrinsic mode functions (IMFs). Each IMF meets two fundamental conditions. The first condition is the number of zero crossings and the number of extrema must be equal or differ at most by one. The second condition is the mean value of the envelope defined by the local maxima and the envelope de-fined by the local minima is zero at any point. For a stationary Gaussian process, the first condition is similar to the narrow-band requirement and the second condition is a local requirement induced from the global one and it is important to assure that the instantaneous frequency will not become redundant fluctuations as induced by asymmetric waveforms. There exist many approaches to computing EMD [5, 6]. The following algorithm is

employed here to de-compose signal S(i) into a set of IMF components. The below block diagram is given here.

The EMD method decomposes the input signal into i number of Intrinsic Mode functions (IMF) and a residue which is the final residue. The given equation will be as follows:

$$S(t) = \sum_{i-1}^{i} imf_i + res_i \ldots\ldots\ldots\ldots (1)$$

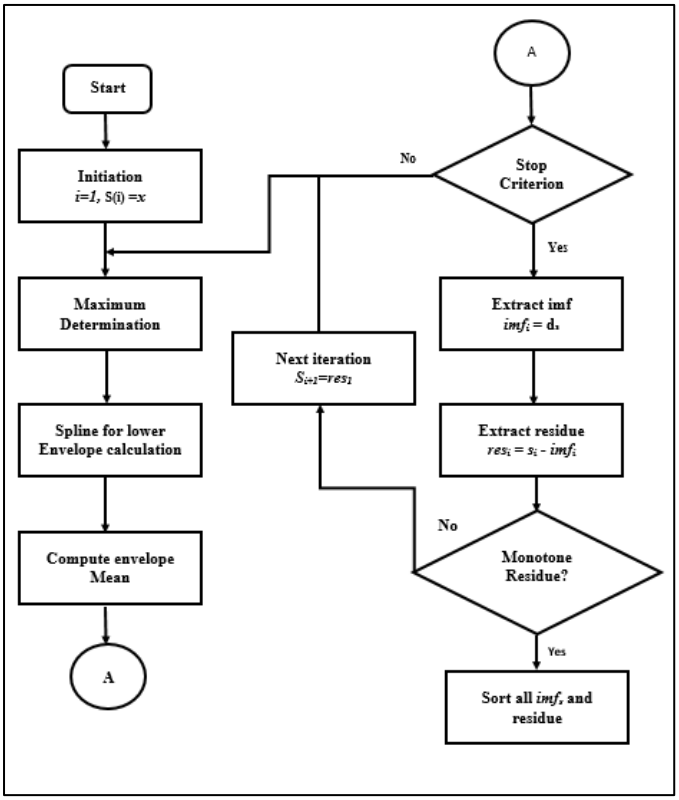

Fig 1: The block diagram of Empirical Mode Decomposition (EMD)

Another way to explain how EMD works is that it extracts out the highest frequency oscillation that remains in the signal. Thus locally, each IMF contains lower frequency oscillation than the one extracted just before. The IMFs are the basis for representing the time series data. Being data adaptive, the basis usually offers a physically meaningful representation of the underlying processes. There is no need of considering the signal as a stack of harmonics and, therefore, EMD is ideal for analyzing non-stationary and nonlinear data.

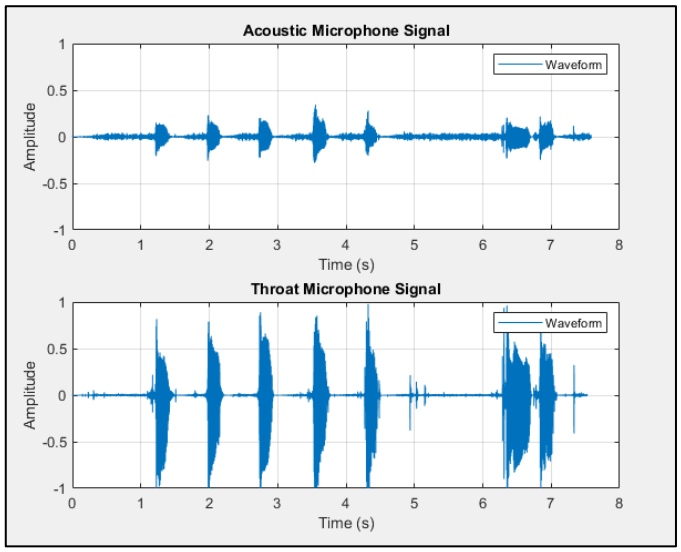

Fig 2: Time Domain Representation (Top: Acoustic Microphone Signal, Bottom: Throat Microphone Signal)

The EMD process can also be considered as dyadic filter-bank especially for white noise and the IMF components are all normally distributed [5,7]. Fig. 2 presents the time do-main representation of Acoustic and Throat Microphone Speech Signals.

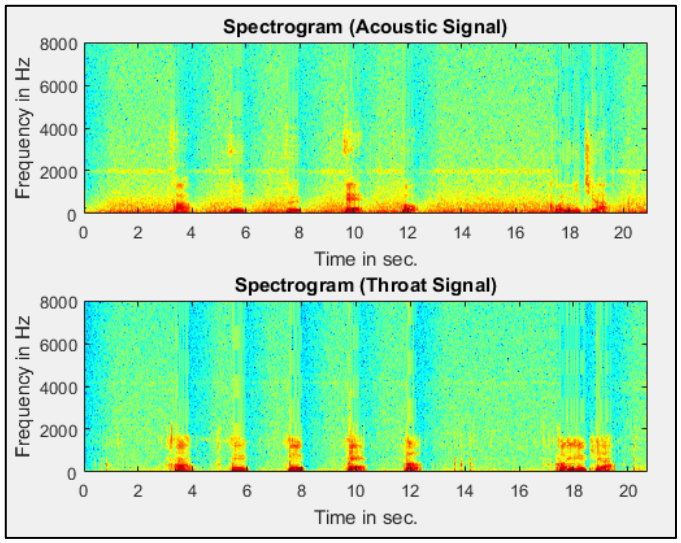

Fig 3: Spectrogram (Top: Acoustic Microphone Signal, Bottom: Throat Microphone Signal)

Fig 3 presents spectrogram of Acoustic and Throat Microphone Speech Signals. Five IMFs out of eight are shown in Fig 4 in the case of Acoustic and Fig 5 for Throat Microphone Speech Signal.

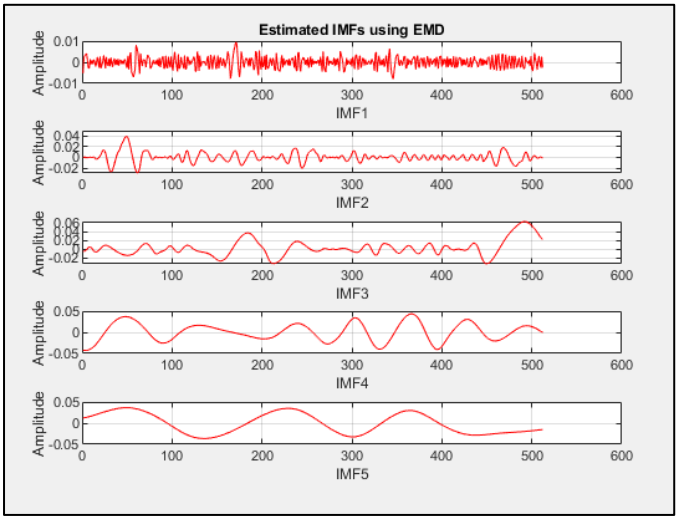

Fig 4: Estimated IMFs of Acoustic Microphone Signal (Here, first five IMFs out of eight IMFs).

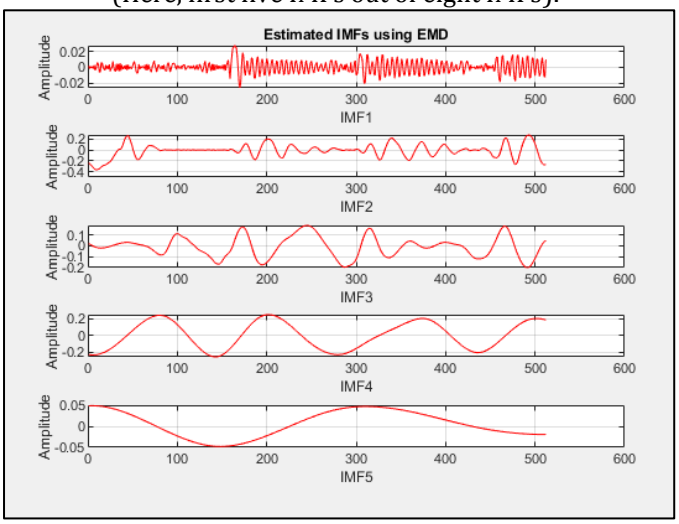

Fig 5: Estimated IMFs of Acoustic Microphone Signal (Here, first five IMFs out of eight IMFs).

It is observed that higher order IMFs contain lower frequency oscillations than that of lower order IMFs. For example, the 6 IMF contains a part of the speech signal of lower frequency than the signal contained by the 5 IMF. Each IMF is considered as a mono-component contribution such that the derivation of instantaneous amplitude and frequency provides a physical significance.

3. EXPERIMENTAL RESULTS

Throat Microphone Speech Signal can be used for finding pitch and voiced/ unvoiced detection in a heavy noisy environment. The estimated pitch of the Throat microphone signal is clear than the acoustic microphone signal.

For some systems in an adverse environment, an extremely accurate and reliable pitch estimation as well as voiced/unvoiced classification of speech is required. Pitch analysis tries to capture the fundamental or dominant frequency of the sound source by analyzing the speech utterance. However, pitch estimation with high accuracy is difficult, because it does not have periodicity with the complete vibration of the vocal cord.

It is difficult to extract pitch in the speech transition parts ex. voiced to unvoiced, silent to speech and in the edge parts ex. beginning and ending of speech. Moreover, pitch estimation in a noisy acoustic environment is a challenging issue be-cause a very complicated procedure is required due to the effect of noise

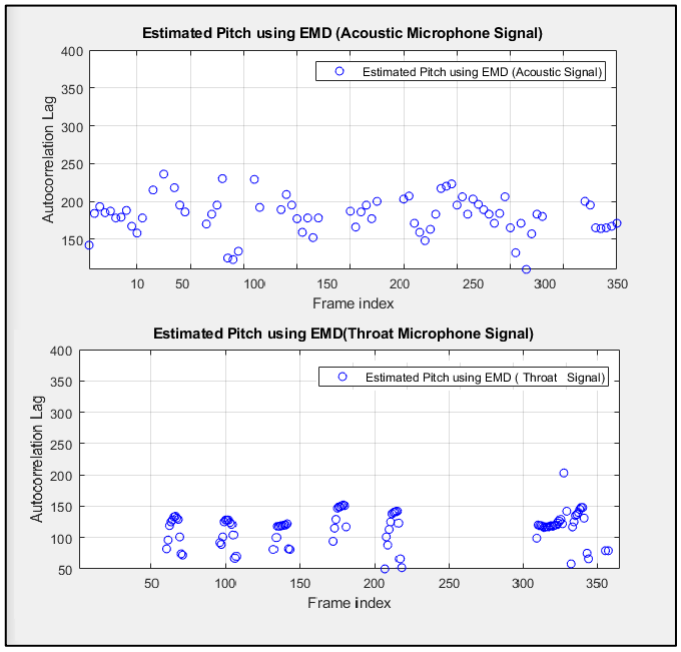

Fig 6: Estimated Pitch using EMD (Top: Acoustic Microphone Signal, Bottom: Throat Microphone Signal).

In this situation Throat Microphone Signal is helpful. In the figure 7, we can see that the pitch estimated from Throat Microphone Signal is clearer in noisy environment than Acoustic Microphone Signal. The Pitch estimated from Acoustic Microphone Signal is influenced by the environmental noises. Noise is also considered as speech signal in the case of Acoustic Microphone Signal.

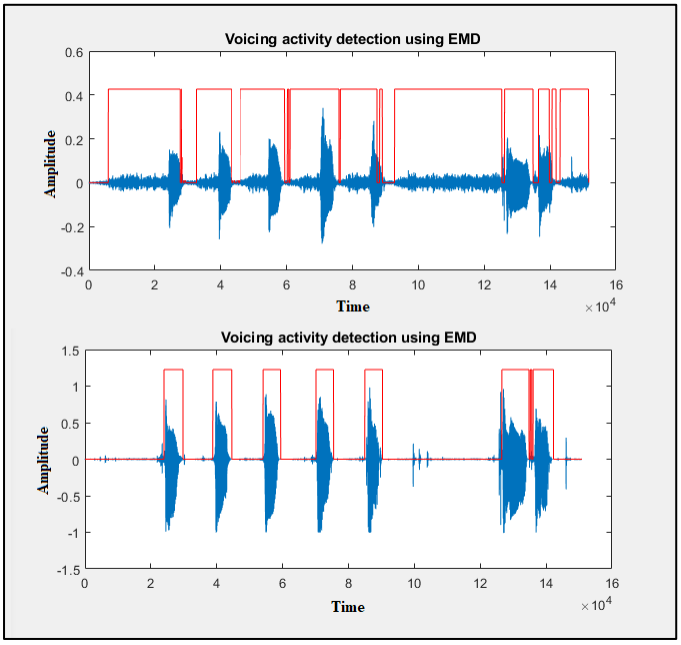

Fig 7: Voicing Activity detection using EMD (Top: Acoustic Microphone Signal, Bottom: Throat Microphone Signal)

In the case of voiced and unvoiced region detection, the Throat Microphone Speech signal is very helpful and from the Signal the voiced and unvoiced region detection is easy. The above figure shows the Voicing activity detection using Empirical mode decomposition (EMD). The top subplot in the figure is Acoustic Microphone Signal and the bottom is Throat Microphone Signal. In the case of Acoustic Micro-phone Signal, the voiced and unvoiced regions are not clearly identified because of the influence of environmental noises as well as within a crowd.

Acoustic Microphone picks up the vibrations through air pressure and also captures interventions. But in the case of Throat Microphone Signal, the voiced and unvoiced regions are clearly detected because the Throat Microphone is a body-attached transducer. It captures the tissue vibrations from the skin on where it is attached. Due to its skin contact, it is more robust to the environmental noise compared to Acoustic Microphone. From the above figure it is noticeable that Throat Microphone signal is free from noises. That's why Throat Microphone Speech is very helpful for the purpose of finding pitch as well as voiced and unvoiced regions.

4. CONCLUSIONS

A new technique for finding pitch as well as voiced and un-voiced regions is developed. The proposed technique is based on the data-adaptive multi-band decomposition using EMD. The paper shows that the effective use of Throat Microphone Signal in the case of pitch as well as voiced and unvoiced regions detection. The use of throat Microphone can be efficient for patients who have lost their voices due to injury or illness. It also provides excellent communication in high noisy environments such as on street crowd, industrial noisy environment etc. The future work of this research is the Emotions recognition from Throat Microphone Speech Signal based on pitch and intensity in noisy environment which can be used for natural Human-Communication especially for vocal tract affected people.

ACKNOWLEDGEMENT

Foremost, I would like to express my sincere gratitude to my coauthor Subrata Kumer Paul for the continuous support of my research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my research study. This work was supported in part by a grant from co-author.

Last but not the least, I would like to thank my family: my parents for giving birth to me at the first place and supporting me spiritually throughout my life.

REFERENCES

[1] J. K. Shah et. al., "Robust voiced/unvoiced classification using novel features and Gaussian mixture model", in proceedings of ICASSP, 2004

[2] T. Dekens, W. Verhelst, F. Capman and F. Beaugendre, "Improved speech recognition in noisy environments by using a throat microphone for accurate voicing detection" 2010 18th European Signal Processing Conference, Aalborg, 2010, pp. 1978-1982.

[3] M. K. I. Molla, K. Hirose, N. Minematsu and M. K. Hasan, "Voiced/Unvoiced Detection of Speech Signals Using Empirical Mode Decomposition Model," 2007 International Conference on Information and Communication Technology, Dhaka, 2007, pp. 311-314.

[4] N. E. Huang et. al., "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis", Proc. Roy. Soc. London A, Vol. 454, pp. 903-995, 1998.

[5] P. Flandrin, G. Rilling and P. GonqalvBs, "Empirical mode decomposition as a filter bank", IEEE signal processing letters, 2003.

[6] G. Rilling, P. Flandrin and P. Gonqalves, "On empirical mode decomposition and its algorithms", in the proceedings of IEEE-EURASIP Workshop on nonlinear signal and image processing (NSIP), 2003

[7] B. Z. Wu, and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method", in the Proc. Roy. Soc. Lond. A (460), pp: 1597-1611, 2004

BIOGRAPHIES

Rakhi Rani Paul was born in Bangladesh in 1996. She graduated in 2017 from Rajshahi University, in Computer Science and Engineering. Her M.Sc. Engineering (degree) is pursuing. Now, she is working as a Lecturer at Bangladesh Army University of Engineering and Technology (BAUET), Natore, Bangladesh. Her research field is Speech Signal Processing, Data Mining, and Machine Learning.

Subrata Kumer Paul was born in Bangladesh in 1993. He completed his B.Sc. and M.Sc. Engineering from Rajshahi University, in Computer Science and Engineering 2016 and 2018 respectively. Now, he is working as a Lecturer at Bangladesh Army University of Engineering and Technology (BAUET) Natore, Bangladesh. His research field is Speech Signal Processing, Data Mining, and Machine Learning.