I J C T A, 9(28) 2016, pp. 165-174
© International Science Press

AN EFFECTIVE RECOGNITION OF PARTIAL SPEECH USING NON AUDIBLE MURMUR (NAM) FOR SPEECH IMPAIRED CHILDREN

Kumaresan. A* Mohan kumar. N Sureshan and. M*** and Suganya. J******

Abstract: In this paper, we present a method for recognizing partial speech with the help of non- audible murmur (NAM). NAM is a kind of soft murmur that is so weak that even people nearby the speaker cannot hear it. It can be detected only with the help of a special type of microphone termed as NAM microphone. We can use this approach for impaired people who can hear sound but can speak only partial words (Semi-mute). We can record and recognize partial speech using NAM microphone but after post processing the received signal we found it is weak. So, we propose the combined use of NAM and normal microphones for recognizing partial speech. We have experimented this scenario by deriving Energy spectrum, Periodogram graphs and we have done a comparative study where we can prove that the efficiency of stand-alone NAM microphone in recognizing partial speech can be improved with the application of the combined microphones.

Keywords: Non Audible Murmur, Acoustic signal, NAM, Partial Speech, Speech impaired, microphone

1. INTRODUCTION

The ordinary human sound is conducted through air. This sound is produced by the vocal cord vibration caused due to the air stream from lungs. Apart from ordinary voice, we use different types of voices in our daily life for enormous purposes such as singing, whispering, crying, and laughing etc. We are able to hear our own voice when it passes through bones while others' voices reach us through air. This is the reason our voice sounding different when we hear it from a playback. Voice can be classified in two categories. First one is the phonation kind of sound which is produced due to the air stream passing through the glottis. This process produces a random noise. Strong air stream is required for shouting or singing whereas a weak air stream is enough for whispering [1].

The second category is termed as conduction type of sound. The vocal tract wall gets vibrated due to the rapid vibration of air in the vocal tract. During this process of sound generation, certain amount of sound passes through the bones and tissues of the neck. We can use the body conductive microphone [2] here to recognize the sound that passes through the bone. This type of weakly murmured voice is not audible to the people nearby. This microphone is introduced by Nakajima inspired by the concept of stethoscope and also based on the fact that sound is not only propagated through air but it also passes through the muscles and our body. Such weak murmurs flowing through bone and muscles are termed as non-audible murmur (NAM). This concept can be used for a private conversation over mobiles and can be used for effective communication in noisy environments as this microphone is robust in noisy environment. As the soft silica layer touches the mastoid part of humans, it does not conduct external noise. We can use this NAM microphone to help impaired people who are hearing impaired but speak only partial words due to any vocal problem by birth. This also includes people who are affected by stammering. We are trying to

* Department of Computer Science and Engineering, SKP Engineering College, Tiruvannamalai, Tamil Nadu, India
Corresponding author, e-mail: 1kummaresan@gmail.com

** Department of Electronics and Communication Engineering, SKP Engineering College, Tiruvannamalai, Tamil Nadu, India

*** Department of Computer Science and Engineering, Sri Sairam Engineering College, Chennai, Tamil Nadu, India

**** Department of Electronics and Communication Engineering, Mailam Engineering College, Villupuram, Tamil Nadu, India

recognize the partial speech using a NAM microphone. In this paper, we have made use of both normal air conductive microphone and NAM microphone to record and recognize the partial speech and finally the results of both the microphones are compared and combined. Experiment is done for Tamil language using TSRE to record and identify the spoken words with the limited number of words stored in dataset. Figure 1 shows the optimal position of attaching the NAM microphone to the speaker.

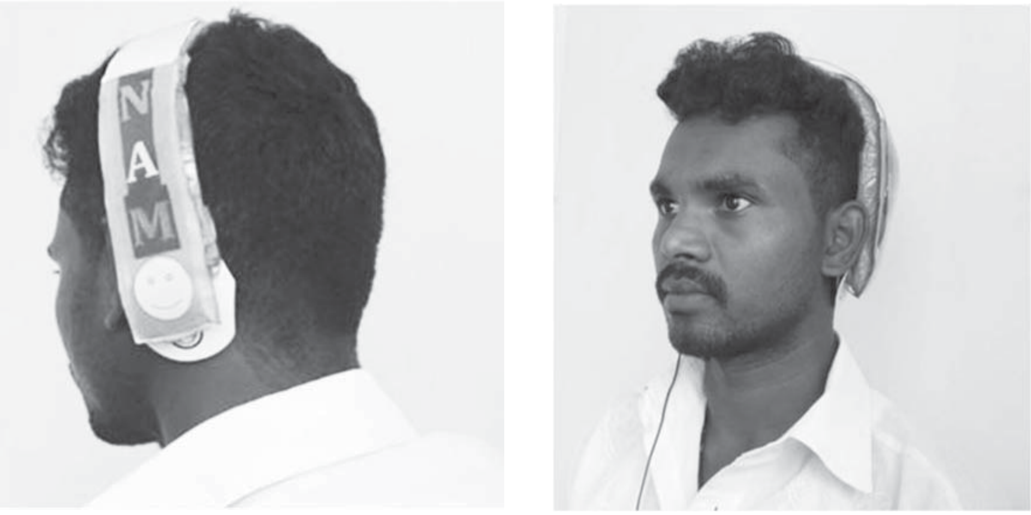


Figure 1. NAM microphone attached to the user

2. RELATED WORK

Speech recognition is a broad topic to deal with. Since we are concerned in recognizing partial speech, we are in need of being aware of the research done in speech recognition in order to understand the concepts used for efficient speech recognition. Scott Axelrod et al have studied the discriminative training of acoustic models for recognizing speech based on two major criteria such as maximum mutual information (MMI) and novel error weighted training technique. They have presented a proof saying that the standard MMI training technique is reasonable only for a general class of acoustic models. They have also reported their implementation results for subspace constrained Gaussian mixture models (SCGMM), in which they required the exponential model weights of all Gaussians to belong to the same common tied subspace. Subspace precision and mean (SPAM) models impose separate subspace constraints on precision matrices and mean. They showed that SPAM and SCGMMs models have yielded significant improvements in error rate when compared with the previously considered models such as diagonal models, extended maximum likelihood linear transformation (EMLLT) models [3]. An SCGMM requires all of the in the equation (1) should belong to the common F-dimensional affine subspace of all parameters

$$\mathrm{N}(\mathrm{x}; \mu, \Sigma) = \mathrm{e}^{\theta\mathrm{T}} f(x) + \mathrm{K}(\mathrm{P} + \varphi) \tag{1}$$

If we consider baseline of affine space to be $b_0 \in \mathrm{R}^{d(d+3)/2}$ and also considering B to be a matrix of size $d(d + 3)/2 \times \mathrm{F}$ and their columns forming the basis of the subspace, we can say,

$$\theta_g = b_0 + \mathrm{B}\lambda_g \tag{2}$$

The parameters B and b_0 in equation (2) are termed as tied parameters and they are shared across gaussians whereas λ_g is an untied parameter.

Donghyun Kim et al have presented Linear Spectral Transformation (LST) technique for robust speech recognition. LST is a transformation based rapid adaptation technique. The authors have also presented a maximum mutual information (MMI) criterion. This method requires only one word from the adaptation data. Their proposed algorithm was named as MMI-LST, which was implemented using phonetic lattices and they have successfully reduced the speech recognition error rate [4]. An adaptation method for speech recognition based on non-linear transformation of feature space has been proposed by Mukund Padmanabhan et al. Since most of the existing adaptation methods assume the affine transform of either the acoustic model or feature vectors that model the feature vectors, they have proposed a general nonlinear transformation based on two transformations, of which one is based on affine transformation which combines the original feature space dimension and the other method is nonlinear transformation applied independently to each dimension of previously transformed feature space. These methods lead to a general multidimensional nonlinear transformation of the original feature space [5]. The authors have conducted experiments with a functional form including the linear function as a special case. They have applied the inverse of the transformation in the recognition stage and it is referred as MLNLT-Q. Since piecewise quadratic seems to be a better choice for inverse mapping, the pieces in g can be defined as follows in equation (3) and (4):

MLNLT – Q:

$$z_{t,d} = g_d\,(y_{t,d}) = Z_{d,k} + \frac{-b_{d,k} + \sqrt{b_{d,k}^2 + 4a_{d,k}(y_{t,d} - Y_{d,k})}}{2a_{d,k}} \quad \forall z_{t,d} \in [z_{d,k}, z_{d,k+1}) \tag{3}$$

and the corresponding transformation g_d^{-1} is,

$$y_{t,d} = g_d^{-1}(z_{t,d}) = Y_{d,k} + a_{d,k}\,(z_{t,d} - Z_{d,k})^2 + b_{d,k}\,(z_{t,d} - Z_{d,k}) \forall z_{t,d} \in [Z_{d,k}, Z_{d,k+1}] \tag{4}$$

Hui Ye et al have worked on a different scenario regarding speech which is termed as voice morphing. It is a technique by which a source speaker's speech can be modified into a sound spoken by some other target speaker. The process of voice morphing depends upon certain major factors such as the spectral envelope transformation of source speaker to target speaker. Other essential factors include linear transformation of parallel training data aligned with time. They have studied about the naïve linear transformation and conducted investigation on two issues. First, they have proposed a general maximum likelihood framework in order to transform estimation for the sake of avoiding parallel training data which was followed in conventional approaches. Secondly, they have found the cause of artifacts to be glottal coupling, unnatural phase dispersion and spectral variance of unvoiced sounds. They have proposed a compensation technique in order to mitigate this. They have found that the results were compromising enough for effective transformation of speech with high quality [6].

Panikos Heracleous et al have conducted experiments for detecting NAM speech in Japanese vocabulary. NAM microphone is normally attached behind the ear exactly at the mastoid so that the non-audible murmur speech passing through the bone can be recognized easily. The authors have previously conducted experiments for recognizing NAM speech using a stethoscopic and silicon microphone and they have attained a high level of accuracy for around 20k Japanese vocabulary task. They have also done further research on NAM using hidden markov models (HMM). The HMM distance is reduced during the experiment due to the decreased spectral space of NAM speech compared to that of the normal speech. The recognition accuracy have reduced drastically due to the significant difference that occur during the recognition of Japanese plosives. Due to this degradation, they have conducted experiments using speaker dependent phoneme recognition. Due to the spectral reduction and unvoiced nature of NAM, NAM becomes similar and confusing with the normal speech. But in order to increase the accuracy of the results, they have considered the lip movements along with facial expressions of the speaker [7]. Chafic Mokbel have introduced a robust speech recognition system. The author's work includes a framework clustered for online adaptation of hidden markov models (HMM) for real-time scenarios. The convergence of adaptation in unsupervised models can also be controlled with the help of the developed methods. HMM parameters can be adapted to new conditions based on two approaches which are spectral transformation and Bayesian adaptation [8]. The author proposed a unified framework in this paper in which both of these approaches were used as particular cases. They have implemented the general adaptation algorithm within the speech recognition system and the whole system was evaluated and found to perform better.

3. SPEECH RECOGNITION USING NORMAL MICROPHONE

Human computer interaction (HCI) is found to be an essential need these days. For an effective HCI system to go live and be user friendly, we are in need of providing voice input to these systems. The reason why we go for speech based inputs is that the use of hand and eyes may put the user in critical situations. When we are in need of providing voice inputs, we could rely upon a microphone for achieving that. Inspite of the availability of enormous conventional speech recognition methodologies, all of them suffer from various problems such as microphone mismatch, environmental complexity and noise [9]. In order to overcome this problem several enhancements techniques using microphone array have been proposed by Jwu-Sheng Hu et al in [9]. They have tried microphone array setup in cars. The Eigen vector, which is the correlation matrix of the microphone array, was achieved by Balan et al, [10] by separating speech signal and noise signal into two separate orthogonal spaces. They also have estimated the direction of arrival (DOA) by projecting the manifold vectors on the noise subspace. Even after the availability of powerful algorithms such as MUSIC [11], [12] along with smoothing [13], they are not well suitable to the situation when SNR is low. So we can make use of single microphone linear array. But it would face a serious drawback when there are barriers between the microphone and the speaker. Due to the lack of efficiency of single microphone array in case of non-line-of-sight conditions, multiple microphone arrays were used. The relationship between the sound source and the receiver is a complicated relationship in a real-time environment and so it is very difficult to characterize the nature of this relationship. With respect to the observation on room acoustics according to [14], the count of Eigen-frequencies whose upper limit is $fs/2$ kHz can be determined by using the following result:

$$\mathrm{L} = 4\frac{\pi}{3}\mathrm{B}\left(\frac{fs}{2v}\right)^3 \tag{5}$$

The above equation (5) was used to indicate that the poles are too high when the frequency is high where represents the sampling frequency, v representing the velocity of sound and B is geometrical volume. We should always consider the geometry of the microphone array in order to maximize the phase difference of each frequency band for determining the accurate location of the speaker. For the purpose of dealing with different frequency bands, the microphone array could be decoupled into several pairs with varying distance between the microphones. The overall probability information is integrated by the location detector from frequency bands for the purpose of detecting the speaker's location. As the distance between the microphone increases, the phase difference of the signal becomes more significant. When the distance exceeds the half of the maximum wavelength of the received signal, the aliasing problem occurs [15]. The distance between the microphones is chosen with respect to the selected frequency band in order to obtain a clear phase difference data for enhancing the accuracy of detecting the location of the speaker and to prevent aliasing.

Even after the availability of various approaches for determining the location of the speaker with the help of multiple microphone arrays, we cannot afford a normal microphone for the purpose of recognizing partial speech. The term partial speech refers to the speech of impaired people who are not dumb but are able to speak only partial words due to any vocal defect by birth or due to the removal of partial vocal cord who are termed partial laryngectomies [16].

4. PARTIAL SPEECH RECOGNITION USING NAM MICROPHONE

Experiments have been done already by Panikos Heracleous et al for recognizing normal speech with the help of NAM microphone [17] where separate GMMs (gaussan mixture model) were allotted for recognizing the normal and NAM speeches respectively since they have used a single NAM microphone for receiving both the inputs. Due to the above scenario, the authors have used separate HMMs (hidden markov model) for recognizing NAM and normal speeches individually. But we propose the use of two microphones, a NAM microphone and a normal microphone. Both of these microphones would record the partial speech of the speaker which can be then combined and fed into the HMM (hidden markov model) whose architecture is shown in Figure 2. HMM considers our dataset for recognizing the word spoken by the speaker which is based on the concept of correlation.

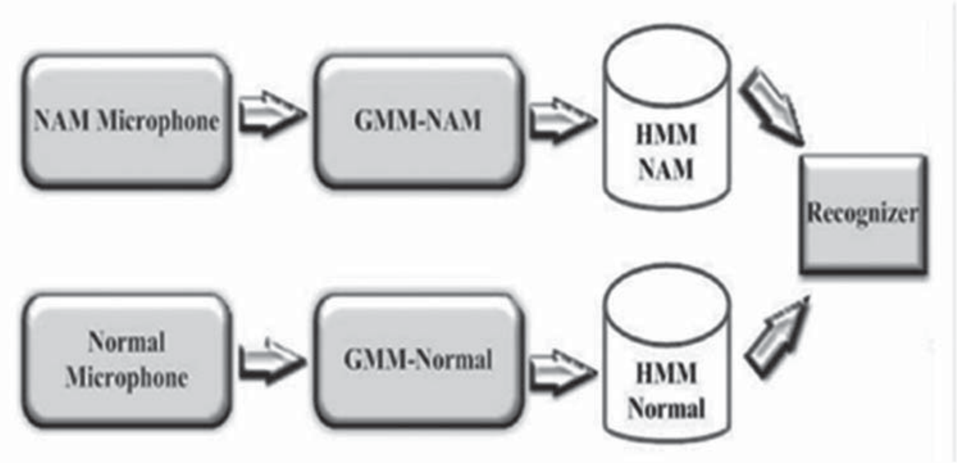


Figure 2. Architecture for integrating the outputs of normal and NAM microphone

The partial speech waveform recorded using a normal microphone for set of words is shown in Figure 3.

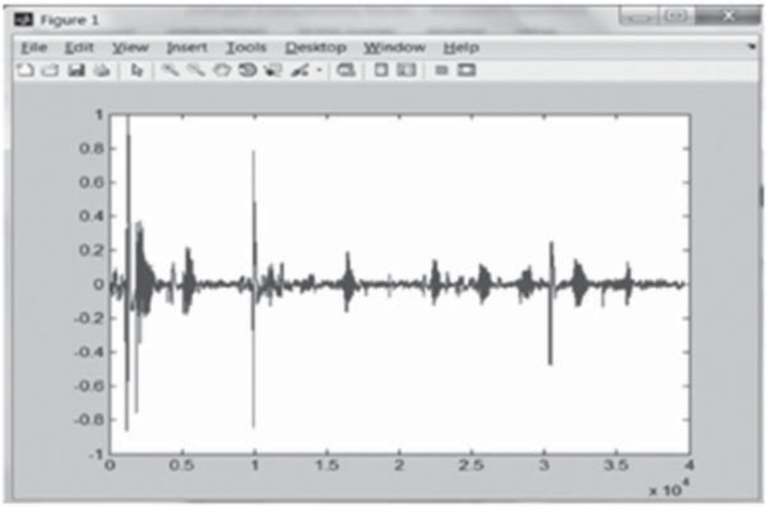


Figure 3. Partial speech waveform recorded using normal microphone

The partial speech waveform recorded using a NAM microphone is shown in Figure 4. We can verify from Figure 4 that the NAM signal is very minimum but is observed at a constant rate. Since we are detecting the signals passing through the bones and muscles, external noise would have no effect since the microphone would be in close contact with speaker's skin but the signal would be a weak soft murmur which can be enhanced during recognition.

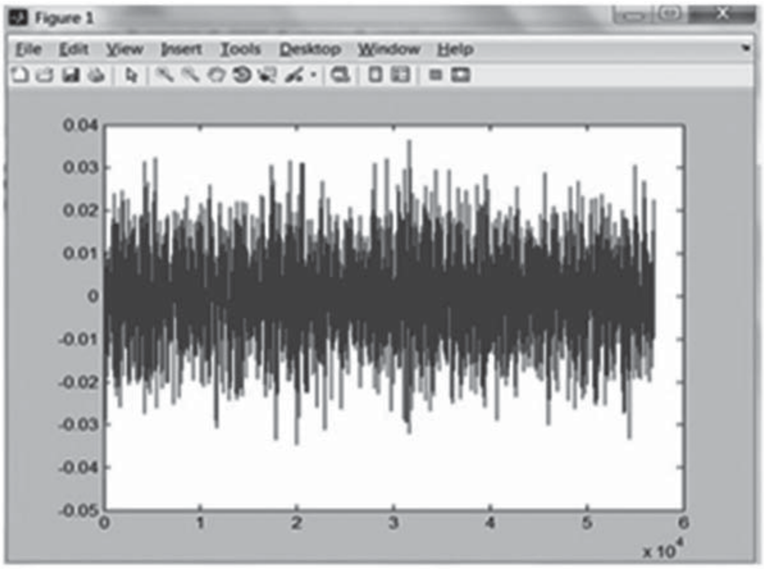


Figure 4. Partial speech waveform recorded using NAM microphone

We are proposing an effort to recognize partial speech with the help of NAM microphone along with the normal microphone for the purpose of increasing the rate of efficiency and accuracy. The partial speech waveform obtained by combining the normal microphone along with the NAM microphone is shown in Figure 5.

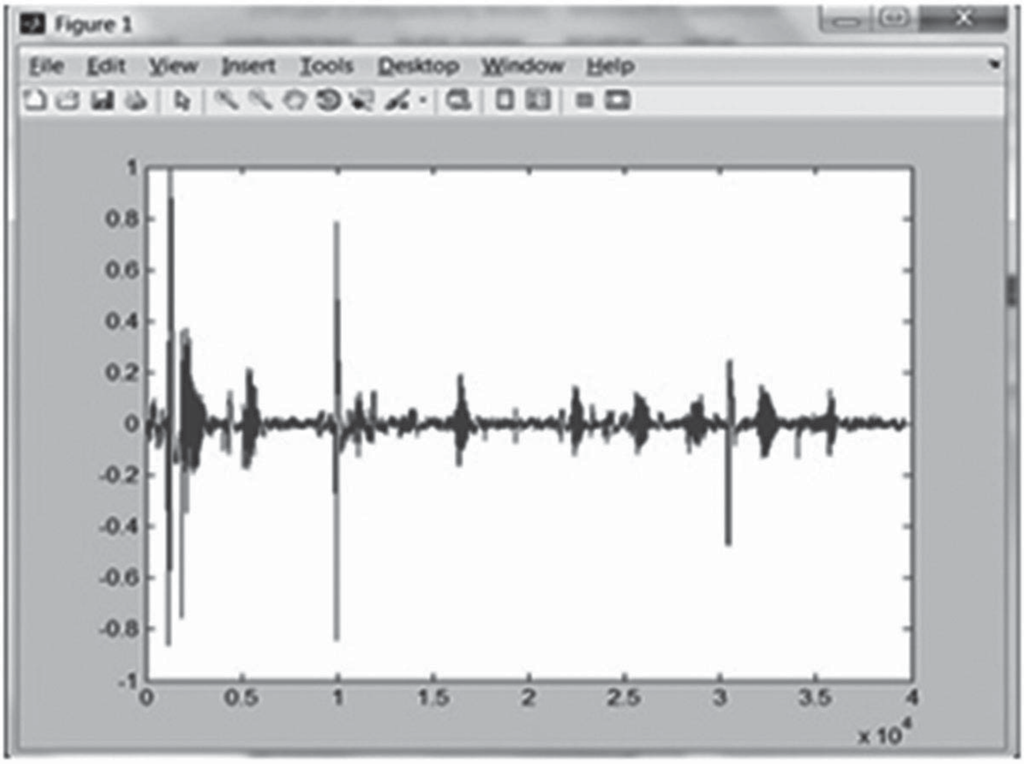


Figure 5. Partial speech waveform obtained by combining the NAM and normal microphone

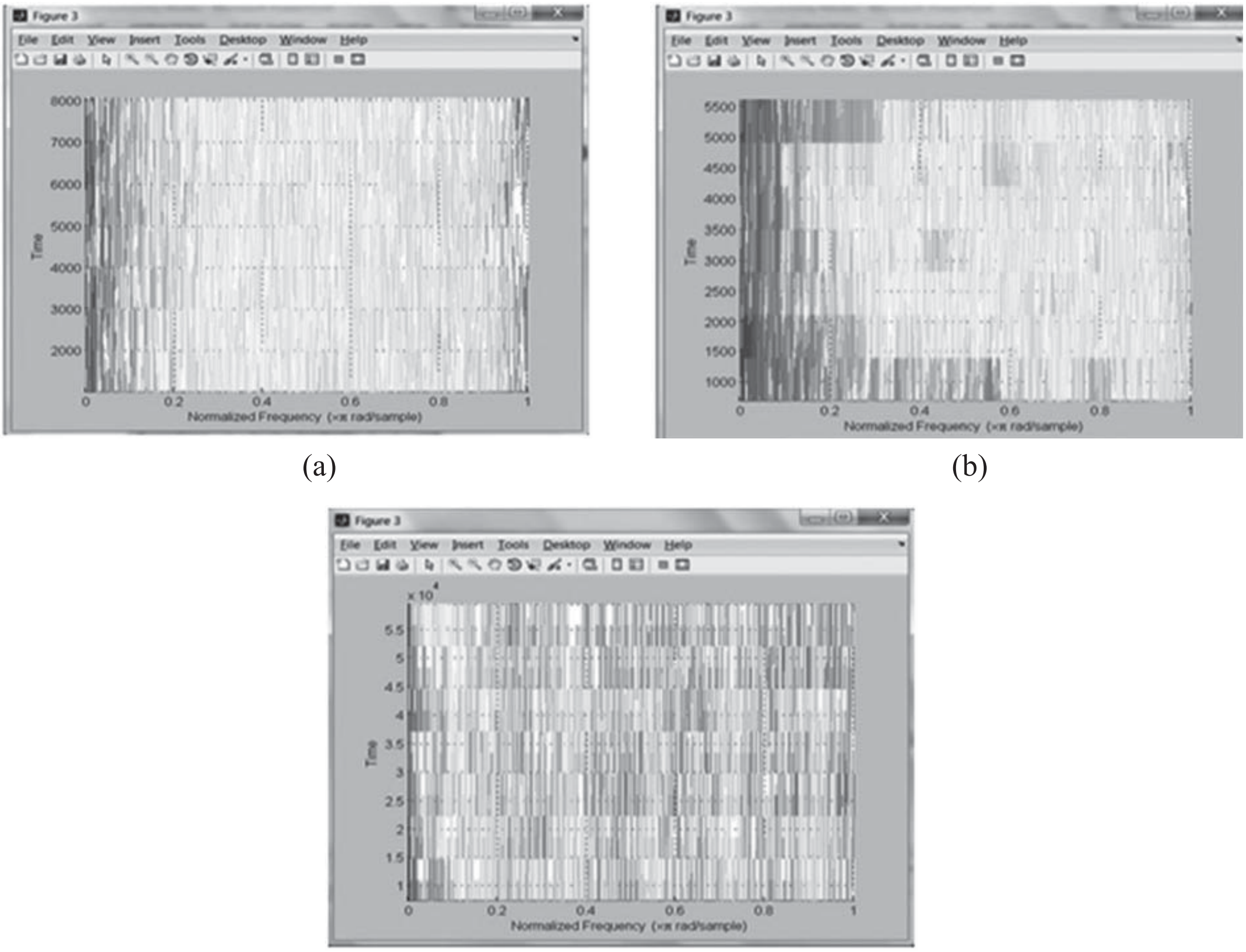
Comparing the above waveform with Figure 3, we can observe that the partial words were not recorded effectively using a normal microphone alone whereas Figure 4 reveals the minute NAM signals recorded using a NAM microphone. But in Figure 5, a compromising level of efficiency has been achieved by the combined effort of both the normal and NAM microphones for recognizing partial speech.

5. EXPERIMENTAL RESULTS

We have conducted our experiment in a sound proof environment in order to calculate the exact efficiency of the architecture shown in Figure 2 without any interruption by external noise. We have experimented using 20 partial utterance of 15 words recorded using normal microphone along with 20 partial utterance of those words uttered with the help of NAM microphone which acts as our dataset.

- We have conducted various comparisons between the different factors of the following waveforms:
- Partial speech recorded using NAM microphone.
- Partial speech recorded using normal microphone and
- Partial speech recorded using both the microphones.

Energy spectrogram is a term used for representing the energy levels of different types of sounds present in a waveform. This spectrogram represents the normalized frequency against time where normalized frequency represents the frequency which is normalized to a standard value which is done for the sake of making the calculations accurate and better. There are different colors used in this representation such as BLUE color represents the clear words without any sort of disturbances and RED color represents the words spoken in high pitch and other colors represents the intermediate pronunciations. The comparison of Energy spectrogram is shown in Figure 6.



(a) (b)

(c)

Figure 6(a), (b), (c) Energy spectrogram of Partial speech obtained by normal and NAM microphones

From the above comparison of Figures 6 a, 6b and 6c, we can come to a conclusion that the BLUE color representing the clearly spoken or clearly recorded words is more in 6c which is obtained as a result of combining the normal and NAM microphones for recording partial speech.

We have conducted experiments by comparing the Gain of the waveforms recorded using normal, NAM and combined efforts. Gain is the ratio of a signal output of the system to the signal input of the system which is represented in the form of dB. The value of Gain would be normally in negative numbers in case of a system with a passive circuit whereas the Gain value would be greater than 1 if the circuit of a system happens to be active. Since we had a passive circuitry, our obtained Gain values would be negative as shown in Figure 7. This can be termed as Periodogram which is the estimation of the spectral density of the signal.

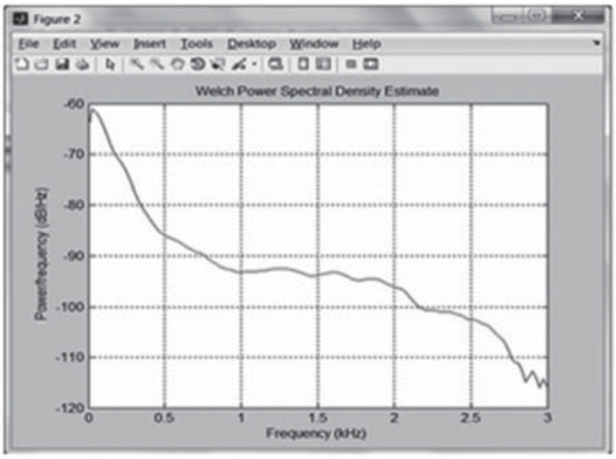


Figure 7(a). Periodogram of Partial speech obtained by NAM microphone.

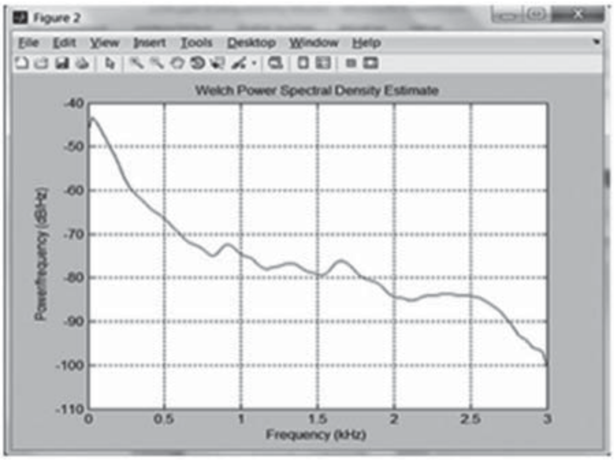
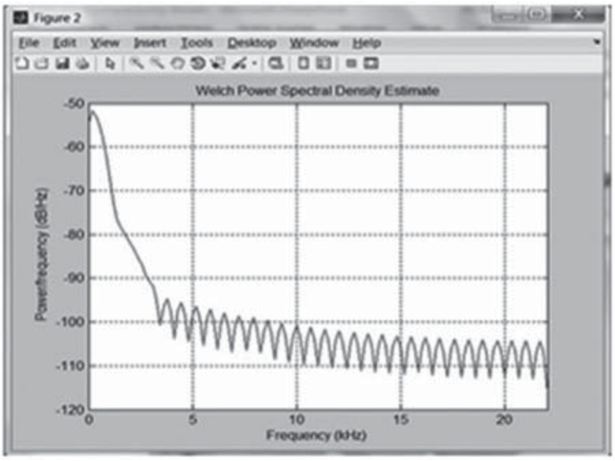


Figure 7 (b). Periodogram of Partial speech obtained by normal microphone



(c)

Figure 7(c). Periodogram of Partial speech obtained by normal and NAM microphones.

From the periodograms compared in Figure 7a, 7b and 7c, though the gain of 7b representing the normal microphone output has a higher gain and the output of the combined microphones is in a moderate level, the recognition of uniformity is consistent at the combined microphones output. So we can conclude that the effort of NAM and normal microphones in recognizing partial speech together has more efficiency than their individual recognition results.

With reference to the architecture of plug and play toolkit implemented in [18] with the help of Julius speech recognition engine for Japanese language has been considered and modifications are done in the training set for the purpose of identifying words in this paper. Based on the dataset created, words can be recognized based on the concept of correlation which is actually done in order to recognize the partial words spoken by the speaker no matter how the word is pronounced which could be louder, inverted or in any other means.

6. CONCLUSION

We have proposed the integrated effort of NAM microphone along with normal microphone for recognizing partial speech in a closed environment (clean environment without noise). We have recorded the partial speech using normal and NAM microphones. In addition to this, normal and NAM microphones are combined to record the partial speech. We have extracted Energiogram and Periodogram from the recorded waveforms. Our experimental comparison have proved that the ccombined effort of normal and NAM microphone is far effective in recognizing partial speech achieving around 66% accuracy. We are planning for future work to enable wheelchair guidance with partial speech inputs in a noisy environment.

7. ACKNOWLEDGEMENTS

We would like to acknowledge Mr. yoshikazu nishiura, Mr. Nawaki along with Mr. Nakajima for providing us great support from NARA Institute of Science and Technology, Japan for replying our queries and guiding us in our research.

References

1. Tatsuya Hirahara, Shota Shimizu And Makoto Otani."*ACOUSTICCHARACTERISTICS OF NON-AUDIBLE MURMUR*". Japan-China Joint Conference Of Acoustics.**2007**
2. Nakajima.Y.et al," *NON-AUDIBLE MURMUR RECOGNITION INPUT INTERFACE USING STETHASCOPIC MICROPHONE ATTACHED TO THE SKIN*", IN PROC.icassp.708-711(2003)
3. Scott Axelrod, Vaibhava Goel, Member, IEEE, Ramesh Gopinath, Senior Member, IEEE, Peder Olsen, Member, IEEE, And Karthik Visweswariah, "discriminative estimation of subspace constrained gaussian mixture models for speech recognition". In ieee transactions on audio, speech, and language processing, vol. 15, no. 1, january 2007.
4. Donghyun Kim, Student Member, IEEE, And Dongsuk Yook, Member, IEEE, "linear spectral transformation for robust speech recognition using maximum mutual information" in ieee signal processing letters, vol. 14, no. 7, july 2007
5. Mukund Padmanabhan, Senior Member, IEEE, And Satya Dharanipragada,"maximum-likelihood nonlinear transformation for acoustic adaptation" in ieee transactions on speech and audio processing, vol. 12, no. 6, november 2004
6. Hui Ye, Student Member, IEEE, And Steve Young, Member, IEEE,"quality-enhanced voice morphing using maximum likelihood transformations" in ieee transactions on audio, speech, and language processing, vol. 14, no. 4, july 2006.
7. Panikos Heracleous, Viet-Anh Tran, Takayuki Nagai, And Kiyohiro Shikano, Fellow, IEEE,"analysis and recognition of nam speech using hmm distances and visual information" in ieee transactions on audio, speech, and language processing, vol. 18, no. 6, august 2010
8. Chafic Mokbel,"online adaptation of hmms to real-life conditions: a unified framework" in ieee transactions on speech and audio processing, vol. 9, no. 4, may 2001.
9. Jwu-Sheng Hu, Member, IEEE, Chieh-Cheng Cheng, And Wei-Han Liu," robust speaker's location detection in a vehicle environment using gmm models", in ieee transactions on systems, man, and cybernetics—part b: cybernetics, vol. 36, no. 2, april 2006.
10. R. V. Balan And J. Rosca, "apparatus and method for estimating the direction of arrival of a source signal using a microphone array," european patent us2 004 013 275, 2004.
11. R. O. Schmidt, "multiple emitter location and signal parameter estimation," ieee trans. Antennas propag., vol. Ap-34, no. 3, pp. 276–280,mar. 1986.

12. M. Wax, T. Shan, And T. Kailath, “spatio-temporal spectral analysis by eigenstructure methods,” ieee trans. Acoust., speech, signal process., vol. 32, no. 4, pp. 817–827, aug. 1984.

13. H. Wang And M. Kaveh, “coherent signal-subspace processing for detection and estimation of angles of arrival of multiple wide-band sources,” ieee trans. Acoust., speech, signal process., vol. Assp-33, no. 4, pp. 823–831, aug. 1985.

14. H. Kuttruf, room acoustics. London, u.k.: elsevier, 1991, ch. 3, p.56.

15. M. Brandstein And D. Ward, microphone arrays: signal processing techniques and applications. New york: springer-verlag, 2001, ch. 2, p. 26.

16. Tomoki Toda, Keigo Nakamura, Takayuki Nagai,Tomomi Kaino, Yoshitaka Nakajima, Kiyohiro Shikano “technologies for processing body-conducted speech detected with non-audible murmur microphone”, in interspeech 2009 brighton.

17. Panikos Heracleous,”audible (normal) speech and inaudible murmur recognition using nam microphone”

18. Tatsuya Kawaharay,”free software toolkit for japanese large vocabulary continuous speech recognition