Throat to Acoustic Speech Mapping for Spectral Parameter Correction using Artificial Neural Network Approach

Subrata Kumer Paul
Department of Computer Science and Engineering, University of Rajshahi

Rakhi Rani Paul
Department of Computer Science and Engineering, University of Rajshahi

Nishimura, Masafumi
Faculty of Informatics, Shizuoka University

Hamid, Md. Ekramul
Department of Computer Science and Engineering, University of Rajshahi

https://doi.org/10.5109/4102497

出版情報：Proceedings of International Exchange and Innovation Conference on Engineering & Sciences (IEICES). 6, pp.238-242, 2020-10-22. Interdisciplinary Graduate School of Engineering Sciences, Kyushu University
バージョン：
権利関係：

Throat to Acoustic Speech Mapping for Spectral Parameter Correction using Artificial Neural Network Approach

Subrata Kumer Paul[1], Rakhi Rani Paul[1], Masafumi Nishimura[2], Md. Ekramul Hamid[1]

[1]Department of Computer Science and
Engineering, University of Rajshahi, Rajshahi,
Bangladesh.

[2]Faculty of Informatics,
Shizuoka University, 3-5-1
Johoku, Naka-ku, Hamamatsu-shi,
Shizuoka, 432-8011 Japan

Corresponding author email:
sksubrata96@gmail.com

Abstract: *In throat microphone (TM), two skin attached piezo-electric sensors can capture speech sound signals from the tissue vibration. Because of their small bandwidth, throat microphone recorded speech is robust to the surrounding noise but suffers from intelligibility and naturalness problems. This study addresses the issue of improving the perceptual quality of the throat microphone speech is based on the statistical mapping between the features of TM and AM speech using the Artificial Neural Network approach for correction of vocal tract parameters and spectral envelope. The target is for natural man machine communication especially for vocal tract affected people. This paper exploits the nonlinear mapping property of Multi-Layered Feed Forward Neural Network (MLFFNN) for estimation of high-frequency components (4-8kHz) from the low-frequency band (0-4kHz) of TM signal. The proposed algorithm is tested using ATR503 Dataset. The simulation results show a noticeable performance in the field of speech communication in adverse environments.*

Keywords: Multi-Layered Feed Forward Neural Network, Mel Frequency Cepstral Coefficient, Speech spectra, Linear Prediction coefficients.

1. INTRODUCTION

The throat microphone is used in this study comprises a pair of modules housing mounted on a neckband. The throat microphone speech signal recorded from a skin vibration transducer placed near the larynx. It is perceptually intelligible, but sounds unnatural. On the other hand, the acoustic microphone speech suffers from noise in an adverse environment [1]. The main objective of this study is to improve the perceptual quality of TM speech so as to make it sound natural.

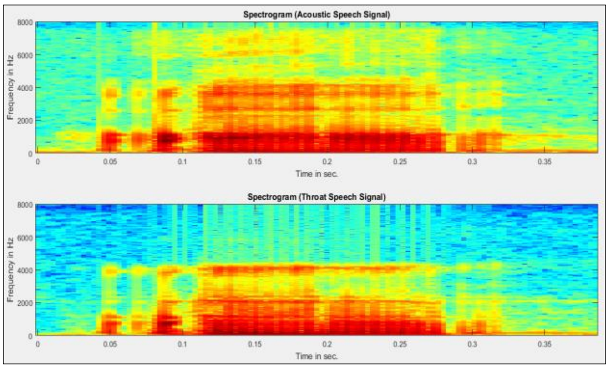


Fig. 1. Spectrograms of speech sound for vowel /a/ recorded simultaneously by using AM (top) and TM (bottom)

In the last few decades, so many studies focus on speech enhancement by suppression of additive background noise. In this section, we present a literature review of the various speech enhancement methods published to date for throat microphone speech enhancement. A. Shahina et al. (2007) in [2] shows that the throat microphone speech is stable to noise, but sound not natural. Their approach improves the perceptual quality of the throat microphone speech. In that study, ANN is used to map the speech features from the TM to the AM speech. The study also presents the mapping technique for bandwidth expansion of telephone speech. Another study by K. Sri Rama Murthy et al. (2008) in [3] presents a mapping technique from TM speech to AM speech to improve the speech quality of TM recording. To mapping, here pairwise vector quantization of spectral feature vectors that are obtained from every analysis frame of TM and AM speech. However, from the literature, we understand that the enhancement of throat microphone speech can be done in two different ways, one is based on the source-filter model and another is based on the use of neural networks deployed as mapping models.

In this research, we try to enhance the perceptual quality of the TM speech signal using the Artificial Neural Network (ANN) based mapping technique that maps the speech wise spectra from the TM to the AM speech. The frame-wise Multi-Layered Feed Forward Neural Network (MLFFNN) is used to obtain a smooth mapping without ‘spectral jumps’ between adjacent frames. However, speech features are estimated by using the MFCCs feature extraction method [4]. Now, the MLFF Neural Network technique is used to map between the features of the TM and AM Speech to improve the perceptual quality of the Throat Microphone speech with respect to acoustic microphone speech. As we see from the spectrogram in Fig.1, for throat microphone speech, the frequency above 4000Hz is totally missing. Moreover, speech energy in the throat microphone in between 2000Hz to 4000Hz is low compared to the acoustic microphone. It is perceived that the throat microphone and acoustic microphone are thoughtful to different features of the signal. Moreover, their spectra vary as a function of the speaker and the location of the transducers and as a function of the voicing of speech itself.

The paper is organized as follows- section II discusses the proposed system, feature extraction method, designing ANN model, speech reconstruction methods. Section III is the experimental result and discussion and lastly the concluding remarks included in section IV.

2. METHODS AND DATASET

2.1 Proposed Method

This speaker-specific correlation can be exploited to capture the relationship between the spectral features of the TM and AM speech of a speaker using a mapping technique. Since the information such as pitch and formant location can vary across sparkers. A speaker-independent mapping would result in distortions in the synthesized speech. The spectral features of the TM speech are mapped onto the corresponding spectral features of the AM speech using a Neural network based on mapping technique [5]. The mapping involves the following two stages (refer Fig.2).

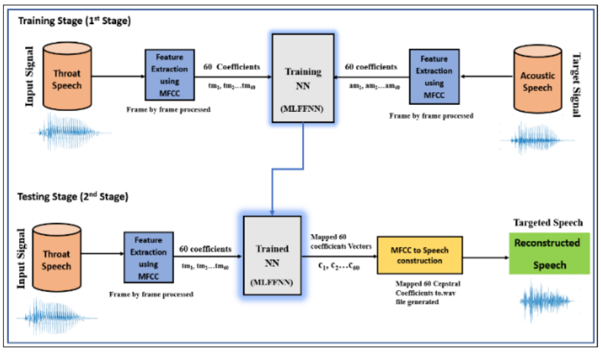


Fig. 2. Block diagram of the proposed model for training and testing of MLFFNN.

The first stage involves training a model to learn the mapping vectors. For training to be effective, speech is simultaneously recording using a TM and an AM from the same speaker. Simultaneous recording ensures that the model learns the mapping between the corresponding frames of the TM and AM speech. The Mel Frequency Cepstral Coefficients (MFCCs) feature extraction method is performed on the speech signals to extract the cepstral coefficients are derived [6].

For training, the Cepstral Coefficients are extracted from the TM Speech are mapped onto the Cepstral Coefficients extracted from the corresponding AM speech. That is, the Cepstral Coefficients derived from the TM data are used as input to the mapping network while the Cepstral Coefficients are obtained from the AM speech form the desired output. The mapping property of Multilayer FeedForward Neural Network (MLFFNN) is used to learn this mapping.

The second stage consists of testing, where the Cepstral Coefficients are derived from a test TM utterance are given as input to the trained Neural Network. The output produced by the network is the estimated Cepstral Coefficients of the TM is called mapped Cepstral Coefficients corresponding to the AM speech as a test input speech. The MFCC Cepstral Coefficients are derived from these estimated Cepstral Coefficients in order to reconstruct the speech signal. The method of working acoustic microphone. Like the way, it can reduce the chances of a lack of naturalness like throat microphone speech. The way it can enhance the speech signal obtained from the throat microphone and improve the perceptual quality.

2.2 ATR503 Dataset

ATR503 Data Set is s Japanese phoneme balance statements. This data comes from Nishimura Laboratory, Shizuka University, Japan. It contains two types of data: Acoustic Microphone (AM) and Throat Microphone (TM) audio files. Acoustic Microphone contains 5-vowels [a, e, i, o, u]. Each vowel contains 100 audio files. Hence, total 100x5 = 500 audio data files. Similarly, Throat Microphone contains 500 audio data files. So, total data files are 1000 in our dataset. ATR503 created a data set by recording reading voice using 2 channel microphones, acoustic microphone and a throat microphone in a soundproof room. This data set contains 5 male speakers. Audio data was recorded at 44 kHz and then down sampled to 16 kHz.

2.3 Features Extraction

Figure 3 presents the block diagram of the computation steps of Mel-frequency Cepstral Coefficients estimation for speech feature extraction. The MFCC calculation includes the following steps: it starts with preprocessing of signal, framing, and windowing, then Fast Fourier Transformation (FFT), after that the Mel Filter Bank and logarithm and lastly the Discrete Cosine Transformation (DCT) [7].

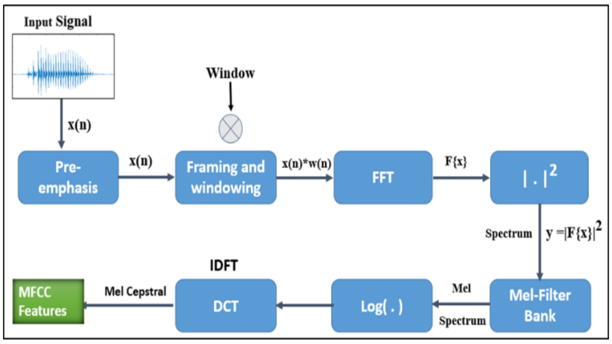


Fig. 3. Speech feature extraction process using MFCC method

Depending on the number of Mel Filter bank, the number of MFCCs are computed. The number of MFCCs size are 10,20,30,40,50,60,70,80,90,100. From a single speech signal generates 10 reconstruction files. Calculating the formant distance between the original signal and reconstructed signal is used by format frequency for 10

MFCCs. We can calculate the other MFCCs [20,30…100 in the same way. Correspondingly, we can calculate the others format distance for 200 signals (Refer Fig.4).

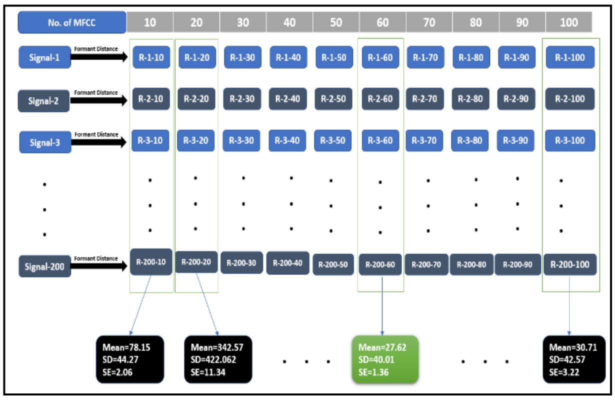


Fig .4. Analysis the reconstruction speech signals

Now, calculating the mean, standard deviation and standard error for each of the filter banks. The goal is to find out which is so similar between original and reconstructed speech signals. And hence find the error rate between them. From the Fig.5, we see that: if we use 60 CC per frame, it will give better performance. Because, Standard deviation and Standard error rate is minimum.

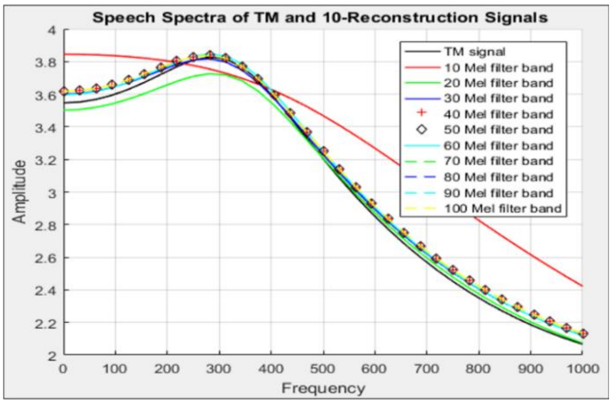


Fig. 5. LPC speech spectra of TM and its 10-Reconstruction Signals

2.4 Design a MLFFNN

In this section, we discuss MLFFNN which is a part of Deep Learning. It uses one or two hidden layers that recognize more complex features. The function of the hidden layer fits weights to the inputs and directs them through an activation function as the output. A fully connected multi-layer neural network is called a Multilayer Perception (MLP). The main advantage of the ANN is that it can be used to solve difficult and most complex problems. However, it needs a long training time sometimes. In this study, the proposed MLFFNNs provide the least mean absolute errors at a given SVD value.

An MLFF neural network has three layers: input, hidden and output layers (Refer to Fig.6) [8]. The network is activated by the hidden neurons to recognize more complicated tasks by takeout gradually the features from the input vector pattern. Fig.6 illustrates the architecture, it shows the input layer forwards data to hidden layers and to the output layer. This step is forward propagation. On the other hand, for backward propagation, a method is used to adjust the weights to minimize the difference between the estimated output and the original. The parameters and their values are used to train MLFFNN are illustrated in Table 1.

In this experiment, two hidden layers are considered and each hidden layer contains 100 neurons which give better results with less distortion. The sigmoid function is used in this experiment because it occurs between 0 to 1. It is mainly used for artificial neural networks where we estimate the probability as an output. For that, we use the sigmoid function for estimation.

Table 1: MLFFNN parameters

No.	Parameters Name	Values
1.	Input layer: (No. of Neurons)	60
2.	Hidden Layer-1: (No. of Neurons)	100
3.	Hidden Layer-2: (No. of Neurons)	100
4.	output layer: (No. of Neurons)	60
5.	Activation function	Sigmoid
6.	Learning rate	0.01
7.	No. of epochs	500
8.	No. of training goal	1e-25
9.	Batch size	10

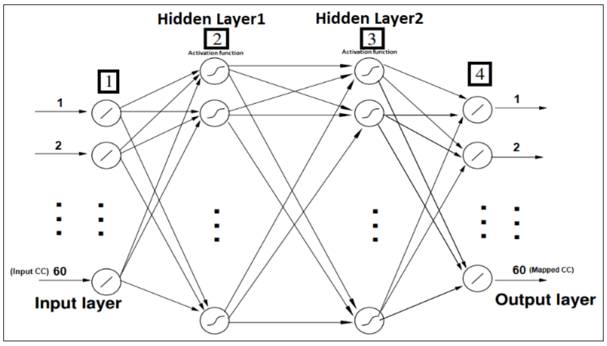


Fig. 6. A MLFFNN architecture of the proposed method

2.5 Speech Reconstruction

The speech power spectrum is obtained from the Cepstral Coefficient by using Moore-Penrose Pseudo-inverse techniques. Using the Least-square Estimate, Inverse Short-time Fourier Transform Magnitude (LSE-ISTFTM) algorithm is applied to estimate the power spectrum and recover the speech waveform [9].

3. Experimental Result and Discussion

3.1 Speech Spectrogram Comparison

We plot spectrograms of both the AM speech and enhanced speech to visually compare the speech enhancement performances of TM speech [10]. Fig.7 shows the speech spectrograms of AM speech, TM speech, and enhanced speech using the proposed method. It is observed that the enhanced speech is much similar to the AM speech by acquiring the missing frequencies in high bands using the proposed MLFFNN model.

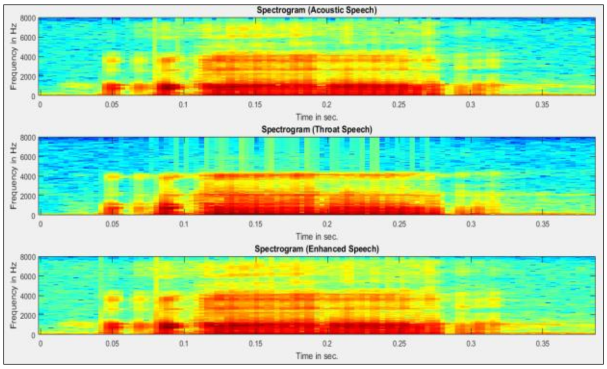


Fig. 7. Speech spectrogram comparison between the acoustic speech, throat speech and enhanced speech signal. (/a/ Vowel)

3.2 LPC power spectrum comparison

Linear prediction coefficients (LPCs) is a form of linear regression. We can compute the spectral envelope magnitude from the LPC parameter by evaluating the transfer function [11]. Fig.8 presents the speech spectra of the acoustic (blue line), throat (red line), and enhanced speech signal (green line).

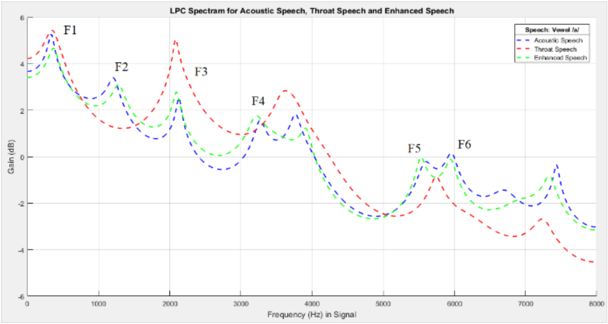


Fig.8. LPC Power spectra comparison between the acoustic speech, throat speech and enhanced speech signal. ‘F’ represents ‘Formant Frequency’. F1, F2… for 1st, 2nd … Formant Frequency respectively. (/a/ Vowel)

All-pole filter in the presence speech signal corresponding to the frequency response, where the LPCs are obtained from these speech samples separately. It is seen that the acoustic spectra (blue line) are closer to the enhanced speech spectra (green line). It is clear from the figure that the AM speech spectra are much closer to the enhanced speech spectra. So, noted that the improved spectra by this proposed method are a close approximation to the AM speech spectra.

3.3 Perceptual Evaluation of Speech Quality (PESQ) measure

The speech quality of enhanced speech is also tested with PESQ. The PESQ score as defined by the ITU recommendation P.862 ranges from 0.5 (worst) up to 4.5 (best). PESQ provides an accurate result for speech quality and is widely considered and used as the best algorithm as an estimation of a subjective measure [12]. Table-2 represents the acoustic signal, throat Microphone signal and enhanced signal PESQ score values. More score values indicate the best speech quality. PESQ algorithm is a useful tool for verification of voice quality performance, competitive analysis and system optimization. It is not accurate enough to specify speech quality requirements in Service Level agreements (SLAs).

3.4 Log Spectral Distance (LSD) measure

The LSD measures the quality of the estimated speech signal concerning the original speech wide-band counterpart. Table 2 shows the average Log Spectral Distance (LSD) scores between the original acoustic speech and the throat speech signal. Moreover, the average PESQ scores between the AM and enhanced TM speech signal. Notice that, for increasing the PESQ, the LSD scores decrease in a consistent way [13]. LSD is defined as:

$$\boldsymbol{d_{LSD}} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\,log_{10}\frac{\boldsymbol{p(w)}}{\boldsymbol{\hat{p}(w)}}\right]^{2}}\boldsymbol{dw}\ldots\ldots\ldots\ (1)$$

where, $\boldsymbol{p(w)}$ and $\boldsymbol{\hat{p}(w)}$ are power spectra.

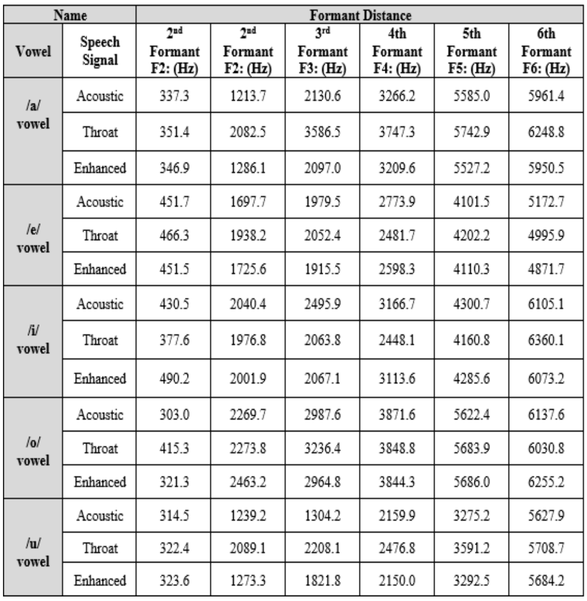
Table 2. LSD and PESQ measurement comparison between AM and TM Speech Signal.

Vowel	**Signal**	**LSD** (dB)	**PESQ** (dB)
/a/ vowel	Acoustic & Throat Speech	1.28	3.09
	Acoustic & Enhanced Speech	1.22	3.17
/e/ vowel	Acoustic & Throat Speech	1.43	3.26
	Acoustic & Enhanced Speech	1.31	3.32
/i/ vowel	Acoustic & Throat Speech	1.85	3.74
	Acoustic & Enhanced Speech	1.77	3.85
/o/ vowel	Acoustic & Throat Speech	1.02	4.03
	Acoustic & Enhanced Speech	1.09	4.11
/u/ vowel	Acoustic & Throat Speech	1.18	3.88
	Acoustic & Enhanced Speech	1.22	3.97

3.5 Speech Formant Analysis measure

Linear prediction coefficients can be used to represent a signal. Formants are resonance frequencies of the vocal tract and observed by the characteristic amplitude peaks in the spectrum [14]. Table 3 illustrates the AM speech formant distances are very much close to the corresponding distances of the enhanced speech by the proposed method. Fig.8 illustrates the graphical representation of formant frequency distances for vowel sound /a/. From the figure, it is clearly observed that the formant structure of the reconstructed signal (enhanced) is much close to the desired AM signal. In Fig 8 graphical representation presents the comparison of Speech formant distance for vowel sound /a/ which is more interpretable than Table 3.

Table 3. Speech Formant Distance Measurement



Name		Formant Distance					
Vowel	Speech Signal	2nd Formant F2: (Hz)	2nd Formant F2: (Hz)	3rd Formant F3: (Hz)	4th Formant F4: (Hz)	5th Formant F5: (Hz)	6th Formant F6: (Hz)
/a/ vowel	Acoustic	337.3	1213.7	2130.6	3266.2	5585.0	5961.4
	Throat	351.4	2082.5	3586.5	3747.3	5742.9	6248.8
	Enhanced	346.9	1286.1	2097.0	3209.6	5527.2	5950.5
/e/ vowel	Acoustic	451.7	1697.7	1979.5	2773.9	4101.5	5172.7
	Throat	466.3	1938.2	2052.4	2481.7	4202.2	4995.9
	Enhanced	451.5	1725.6	1915.5	2598.3	4110.3	4871.7
/i/ vowel	Acoustic	430.5	2040.4	2495.9	3166.7	4300.7	6105.1
	Throat	377.6	1976.8	2063.8	2448.1	4160.8	6360.1
	Enhanced	490.2	2001.9	2067.1	3113.6	4285.6	6073.2
/o/ vowel	Acoustic	303.0	2269.7	2987.6	3871.6	5622.4	6137.6
	Throat	415.3	2273.8	3236.4	3848.8	5683.9	6030.8
	Enhanced	321.3	2463.2	2964.8	3844.3	5686.0	6255.2
/u/ vowel	Acoustic	314.5	1239.2	1304.2	2159.9	3275.2	5627.9
	Throat	322.4	2089.1	2208.1	2476.8	3591.2	5708.7
	Enhanced	323.6	1273.3	1821.8	2150.0	3292.5	5684.2

4. CONCLUSION

In this paper, we focused on processing the speech signals obtained from the throat microphone (TM) for improving its perceptual quality. The perceived lack of naturalness in the TM speech was reduced by exploiting the characteristics of the high-quality acoustic microphone (AM) speech. The intelligibility of the TM speech in noisy ambiance was exploited for speech applications in noisy conditions. Techniques used to improve the perceptual quality of the TM speech signal. The perceptual quality of a speech signal depends on the acoustic characteristics. The task of enhancing the TM speech exploded the characteristics of the high perceptual quality of the AM speech. To compensate for the acoustic differences, the task was divided into two subtasks. The first subtask involved extracting the speech features using the MFCC method for the TM speech and AM speech signal. The second subtask of the nonlinear mapping property of the multi-layered Feedforward Neural Network (MLFFNN) is used for the spectral mapping. After that we get enhanced speech. Speech Spectrogram, LPC power spectrum comparison, LSD, PESQ, formant distance and Itakura–Saito (I-S) distance were the methods used to evaluate speech quality Measurement. The result shows a noticeable performance in the field of speech communication in adverse environments.

5. REFERENCES

[1] Dafydd Gibbon, Prosody: "Rhythms and Melodies of Speech", Bielefeld University, Germany, 2001, pp.1-35

[2] A. Shahina and B, Yegnanarayana, "Mapping Speech Spectra from Throat Microphone to Close-Speaking Microphone A Neural Network Approach", International Institute of Information Technology, Gachibowli, Hyderabad 500032, India, March 2007, pp. 1-10.

[3] K. Sri Rama Murty, "Efficient representation of throat microphone speech", Conference Paper, Jan 2008, pp. 2610-2613.

[4] Yong Xu, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks", Volume: 21, Jan. 2014, Page(s): 65 — 68 [17]

[5] K. M. Keenaghan, "A Novel Non-Acoustic Voiced Sensor: Experimental Results and Characterization". MS thesis, Electrical and Computer Engg., Worcester Polytechnic Institute, Feb. 2004, pp. 318-362.

[6] D. E. Rumellhart, G. E. Hinton, and R J. Williams, "Learning Internal Representation by Error Backpropagation," Vol. pp 318- 362, MIT Press, Cambridge, MA, 1986, pp.19-22

[7] Karthika Vijayan, "Comparative study of spectral mapping techniques for enhancement of throat microphone speech", Conference: March 2014, pp.3-10

[8] Gokay Disken, Lutfu Saribulut, A Review on Feature Extraction for Speaker Recognition under Degraded Conditions", Pages 321-332, 08 Jun 2016

[9] A. Shahina and B, Yegnanarayana, "Artificial neural networks for pattern recognition ", Vol. 19, Part 2, April 2014, pp. 189-23

[10] Gang Min, Xiongwei Zhang, Jibin Yang, Xia Zou. "Speech reconstruction from melfrequency cepstral coefficients via ℓ1-norm minimization", IEEE 17th International Workshop on Multimedia Signal Processing (MMSP), 2015

[11] Tassadaq Hussain, "Experimental Study on Extreme Learning Machine Applications for Speech Enhancement", October 2017IEEE Access PP (99):1-1

[12] Mehmet Ali Tu˘gtekin Turan, "Enhancement of Throat Microphone Recordings Using Gaussian Mixture Model Probabilistic Estimator", Koc University August, 2013

[13] Rabiner, Lawrence R; Juang, Biing-Hwang (1993). "Fundamentals of speech recognition". PTR Prentice Hall.

[14] A. Gray and J. Markel, "Distance measures for speech processing", IEEE Transactions on Acoustics, Speech, and Signal Processing (Volume: 24, Issue: 5, Oct 2006)