# Throat Microphone Speech Enhancement Using Machine Learning Technique

**Chapter** · July 2020

DOI: 10.1007/978-3-030-65407-8_1

**4 authors:**

Subrata Kumer Paul
University of Rajshahi
**27** PUBLICATIONS **105** CITATIONS
SEE PROFILE

Rakhi Rani Paul
University of Rajshahi
**26** PUBLICATIONS **105** CITATIONS
SEE PROFILE

Masafumi Nishimura
Aichi Sangyo University
**137** PUBLICATIONS **778** CITATIONS
SEE PROFILE

Ekramul Hamid
University of Rajshahi
**49** PUBLICATIONS **183** CITATIONS
SEE PROFILE

Learning and Analytics in Intelligent Systems 21

Margarita N. Favorskaya ·
Sheng-Lung Peng · Milan Simic ·
Basim Alhadidi · Souvik Pal *Editors*

# Intelligent Computing Paradigm and Cutting-edge Technologies

## Proceedings of the Second International Conference on Innovative Computing and Cutting-edge Technologies (ICICCT 2020)

Springer

# Throat Microphone Speech Enhancement Using Machine Learning Technique

**Subrata Kumer Paul, Rakhi Rani Paul, Masafumi Nishimura, and Md. Ekramul Hamid**

**Abstract** Throat Microphone (TM) speech is a narrow bandwidth speech and it sounds unnatural, unlike acoustic microphone (AM) recording. Although the TM captured speech is not affected by the environmental noise but it suffers naturalness and intelligibility problems. In this paper, we focus on the problem of enhancing the perceptual quality of the TM speech using the machine learning technique by modifying the spectral envelope and vocal tract parameters. The Mel-frequency Cepstral Coefficients (MFCCs) feature extraction technique is carried out to extract speech features. Then mapping technique is used between the features of the TM and AM speech using Neural Network. This improves the perceptual quality of the TM speech with respect to AM speech by estimating and correcting the missing high-frequency components in between 4 and 8 kHz from the low-frequency band (0–4 kHz) of TM speech signal. Then the least-square estimation and Inverse Short-time Fourier Transform Magnitude methods are applied to measure the power spectrum is used to reconstruct the speech signal. The ATR503 dataset is used to test the proposed technique. The simulation results show a visible performance in the field of speech enhancement in adverse environments. The aim of this study is for natural human–machine interaction for vocal tract affected people.

**Keywords** Machine learning · Multi-layered feed forward neural network · Mel frequency cepstral coefficients · Speech spectra · Linear prediction coefficients

S. K. Paul (✉) · R. R. Paul · Md. E. Hamid
Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh
e-mail: sksubrata96@gmail.com

R. R. Paul
e-mail: rakhipaul.cse@gmail.com

Md. E. Hamid
e-mail: ekram_hamid@yahoo.com

M. Nishimura
Faculty of Informatics, Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu-shi, Shizuoka 432-8011, Japan
e-mail: nisimura@inf.shizuoka.ac.jp

M. N. Favorskaya et al. (eds.), *Intelligent Computing Paradigm and Cutting-edge Technologies*, Learning and Analytics in Intelligent Systems 21,
https://doi.org/10.1007/978-3-030-65407-8_1

# 1 Introduction

The throat microphone is comprised of a pair of units housing fixed on a neckband. This skin vibration transducer placed near the larynx is used to record the voice signal. Although this voice is perceptually intelligible, but the sound is not natural like AM speech sound. Again we know, the acoustic microphone speech undergoes from noise in a real environment [1]. The aim of this study is to enhance the speech quality of TM speech for better human–machine communication.

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. Shahina and Yegnanarayana [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 Murthy et al. [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. 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 the spectrogram in Fig. 1, for throat microphone speech, the frequency above 4000 Hz is totally missing. Moreover, speech energy in throat microphone in between 2000 and 4000 Hz 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 spectral 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 prepared as the following: Sect. 2 describes the system which is proposed, ATR503 Dataset, feature extraction process, designing the ANN model, and speech reconstruction methods. Section 3 is the experimental result and discussion and lastly, the concluding remarks include in Sect. 4.

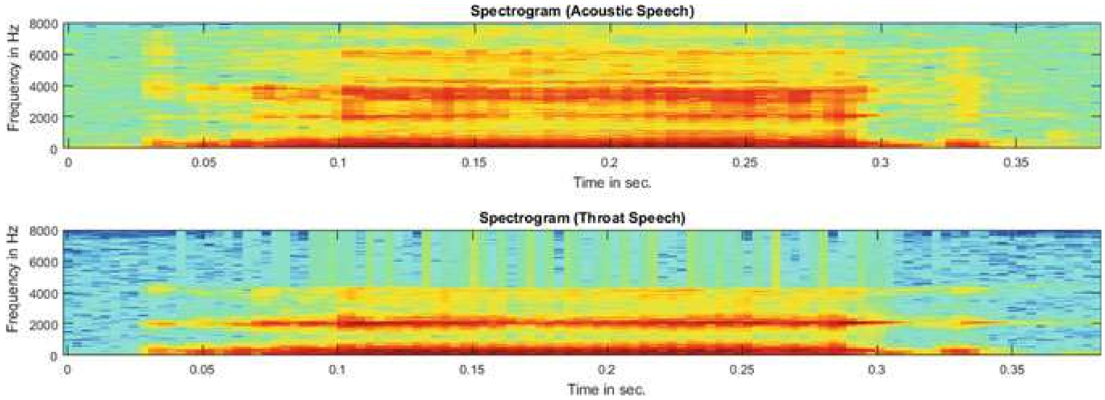

**Fig. 1** Spectrograms of speech sound for vowel /e/ recorded simultaneously by using AM (top) and TM (bottom)

# 2 Methods and Dataset

## *2.1 Proposed Machine Learning Based Method*

This study addresses an improvement of the perceptual quality of the TM speech signal using the Artificial Neural Network (ANN) based mapping technique that maps the speech wise spectra from TM speech signal to AM speech signal [4]. The study consists of two stages. In the first stage, we trained the system to learn the mapping vectors. For training, the speakers are used to utter the sample speech and record simultaneously using a TM and an AM placed near the larynx and near the mouth properly. Concurrent recording confirms that the model learns the mapping between the parallel frames of both microphone speech. Then the Cepstral coefficients are estimated using the MFCC feature extraction technique [5]. Figure 2 illustrated the two stages of mapping techniques.

In the figure it shows during the training phase the Cepstral Coefficients are extracted from the throat Speech. These coefficients are used to map into the Cepstral Coefficients (CCs) extracted from the corresponding AM speech. That is, the Cepstral Coefficients resulting from the TM data are the input of the Multilayer Feed Forward Neural Network (MLFFNN) mapping system while the Cepstral Coefficients are extracted from the AM speech form the wanted output.

However, testing the second stage, which is a trained neural network, where the Cepstral Coefficients derived from a test TM utterance are given as input to the system. The output formed by the network is the estimated Cepstral Coefficients of the throat microphone speech is called mapped Cepstral Coefficients. This corresponds to the AM speech as a test input speech. The MFCC Cepstral Coefficients are estimated from these derived Cepstral Coefficients for speech reconstruction.

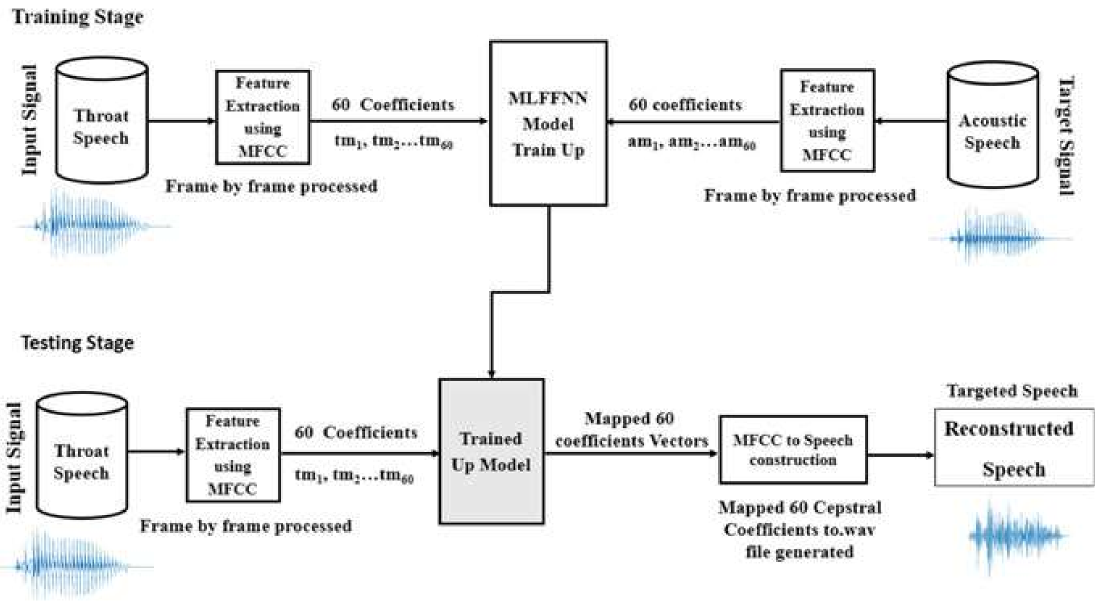

**Fig. 2** The artificial neural network model for training and testing

## 2.2 Experimental Dataset Description

We use ATR503 Data set to evaluate the proposed study. This data set is collected from Nishimura Laboratory, Shizuoka University, Japan. It has Acoustic Microphone (AM) and Throat Microphone (TM) simultaneous recording audio files of 5-vowel sounds [a, e, i, o, u]. Each vowel is uttered 100 times. So, the dataset contains total $100 \times 5 = 500$ audio files for both TM and AM speech. The recording is done by reading voice using 2 channel microphones, acoustic and throat microphones in a soundproof room. This data set contains 5 male speakers at 44 kHz.

## 2.3 Speech Features Extraction Using MFCC Method

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) [6].

The number of Mel Filter bank ensures the number of MFCCs to be computed. Figure 4 illustrates the number of MFCCs size is 10, 20, 30, …, 90, 100. Here in this study, for a single speech signal, ten reconstruction files are generated. Then for each of these ten MFCCs, we calculate the formant distance between the original signal and reconstructed signal. A Similar way, calculates formant distances for all MFCCs. However, in this study, consider formant distances for all 200 signals in the database as shown in Fig. 4. All are the MFCC parameters describe in Table 1.

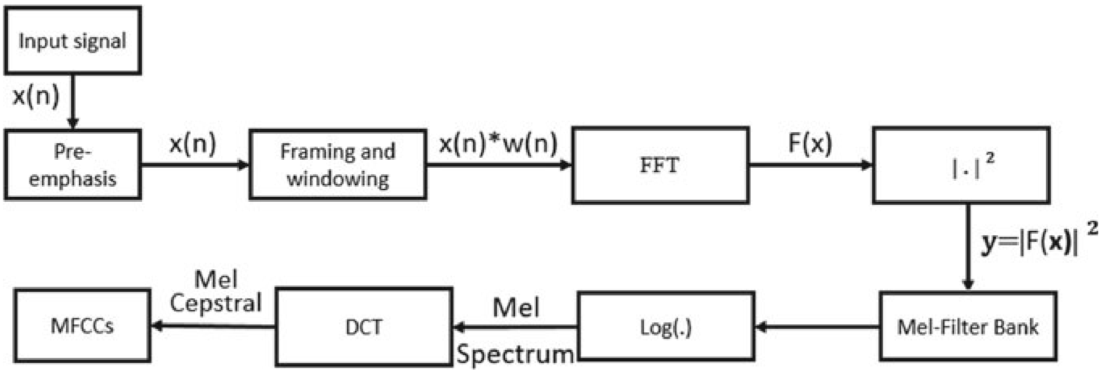

**Fig. 3** MFCC steps for speech feature extraction

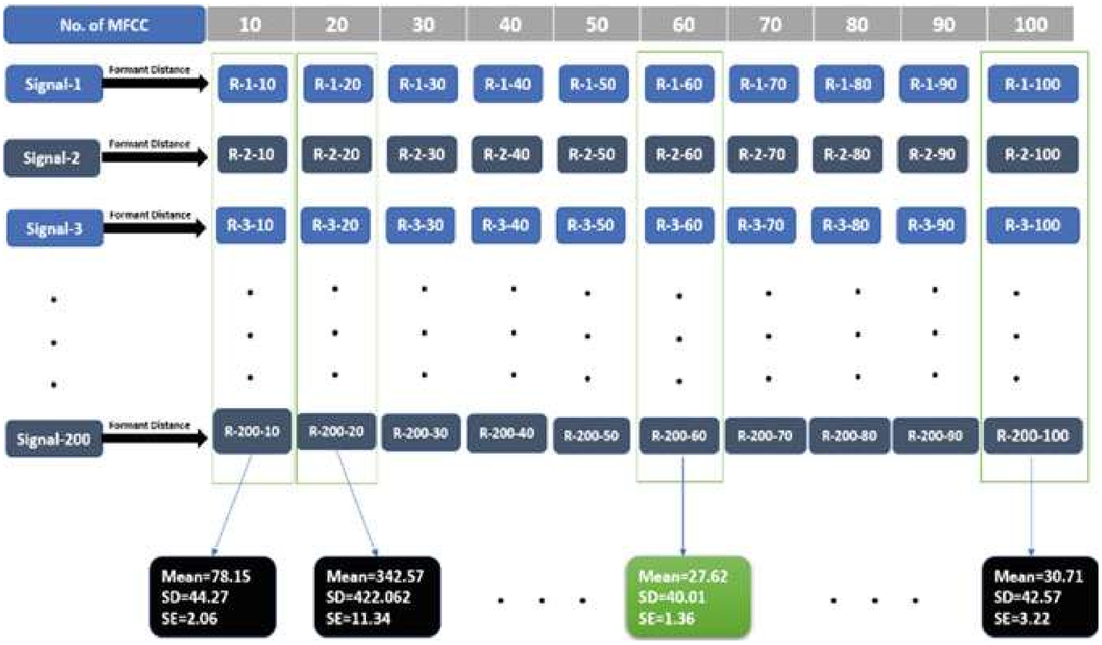

**Fig. 4** Analysis of the reconstruction speech signals

**Table 1** MLFFNN parameters and their values

| No | Parameters | Values |
|---|---|---|
| 1 | Neurons (Input layer) | 60 |
| 2 | Neurons (Hidden layer 1) | 100 |
| 3 | Neurons (Hidden layer 2) | 100 |
| 4 | Neurons (Output layer) | 60 |
| 5 | Activation function | Sigmoid |
| 6 | Learning rate | 0.01 |
| 7 | Epochs number | 500 |
| 8 | Number of training target | 1e-25 |
| 9 | Size of batch | 10 |

Finally, we calculate band wise the mean, standard deviation, and standard error. The target is to find out which is much similar between original and reconstructed speech signals. Then find out the error rate between them. It is observed that when 60 Cepstral coefficients per frame are taken the standard deviation and the standard error rate is minimum that shows the best performance.

### 2.4 Design a MLFF Neural Network

In this section, we discuss MLFFNN which is a part of Deep Learning. It uses one or two hidden layers that recognizing 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. 5) [7]. The network is activated by the hidden neurons to recognize more complicated tasks by takeout gradually the features from the input vector pattern. Figure 5 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.

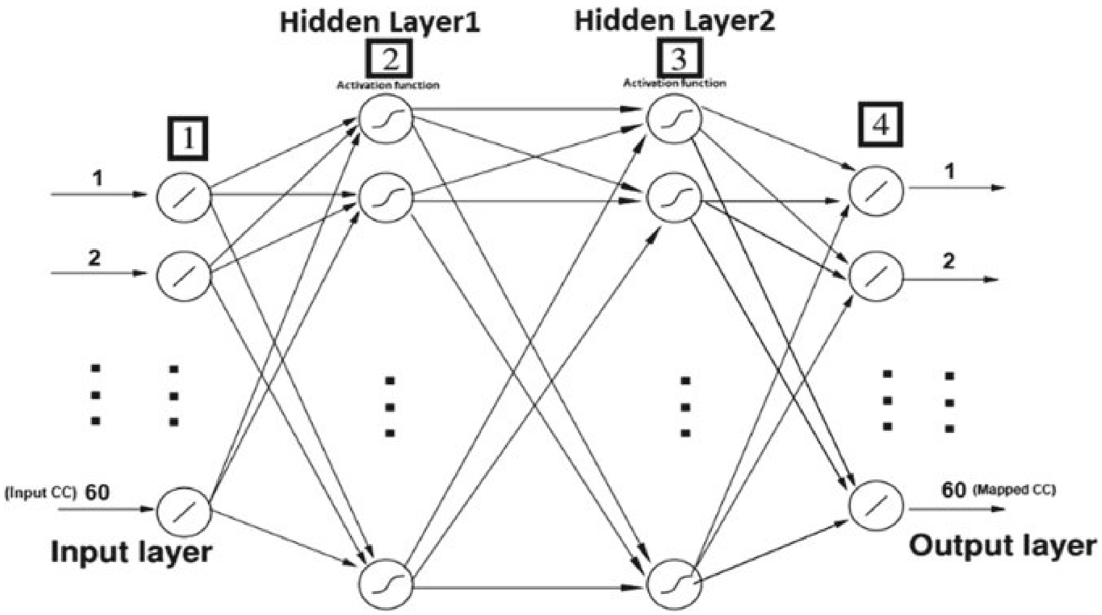

**Fig. 5** A MLFFNN model architecture of the proposed method

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 and 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.

### 2.5 Speech Reconstruction

The speech power spectrum is obtained from the Cepstral Coefficient by using Moore–Penrose pseudo-inverse techniques. Then by using the Least-square estimation technique followed by Inverse Short-time Fourier Transform Magnitude (LSE-ISTFTM) method to measure the power spectrum. However, the power spectrum is used to reconstruct the speech waveform [8].

## 3 Experimental Results and Discussions

### 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 [9]. Figure 6 shows the speech spectrograms of AM speech, TM speech, and enhanced speech using the

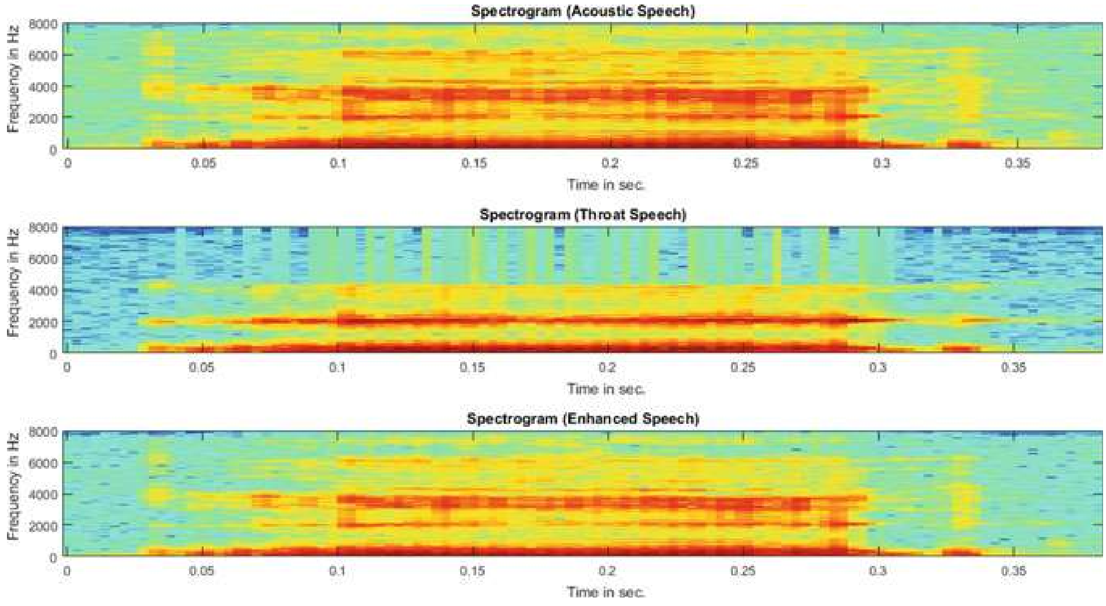

**Fig. 6** Speech spectrograms of the acoustic microphone speech, throat microphone speech and the enhanced speech by the proposed method for vowel sound /e/

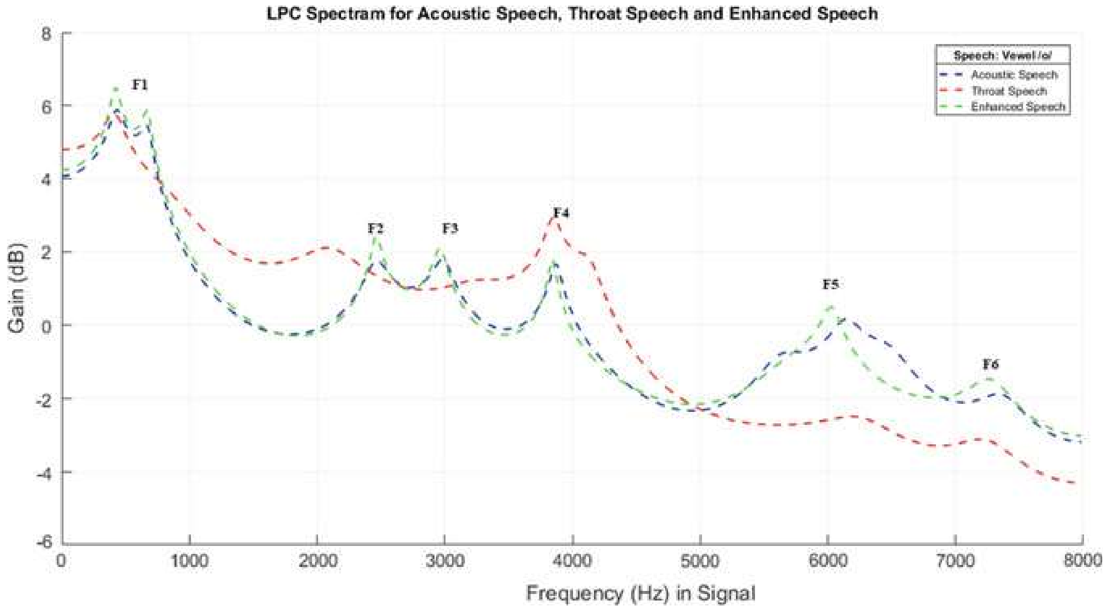

**Fig. 7** LPC Power spectra of the AM speech (Blue), TM speech (Red), and the enhanced speech (Green) by the proposed method for vowel sound /e/. In figure F1, F2, … are the formant frequencies

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.

### 3.2 LPCs Power Spectrum Comparison

Linear prediction coefficients (LPCs) are a form of linear regression. We can compute the spectral envelope magnitude from the LPC parameter by evaluating the transfer function [10]. Figure 7 shows the speech spectra of the AM speech, TM speech, and the enhanced speech by the proposed method. LPC is obtained from the All-pole filter in the presence speech signal corresponding to the frequency response. 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 PESQ measure is carried out to see the quality of speech enhancement. The PESQ score varies from 0.50 (worst) up to 4.50 (best) as determined by ITU. It is a widely used method and one of the best algorithms for an estimation of a subjective

**Table 2** LSD and PESQ scores comparison

| Vowel | Signal | LSD (in dB) | PESQ (in dB) |
|---|---|---|---|
| /a/ | AM and TM speech | 1.3 | 3.1 |
| | AM and enhanced speech | 1.2 | 3.2 |
| /e/ | AM and throat speech | 1.4 | 3.3 |
| | AM and enhanced speech | 1.3 | 3.3 |
| /i/ | AM and TM speech | 1.9 | 3.7 |
| | AM and enhanced speech | 1.8 | 3.9 |
| /o/ | AM and TM speech | 1.0 | 4.0 |
| | AM and enhanced speech | 1.1 | 4.1 |
| /u/ | AM and TM speech | 1.2 | 3.9 |
| | AM and enhanced speech | 1.2 | 4.0 |

measure [11]. Table 2 presents the PESQ score values of AM speech, TM speech, and the enhanced speech by the proposed method. More score values specify the best speech quality.

## 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 [11].

## 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. Table 3 illustrates the AM speech formant distances are very much close to the corresponding distances of the enhanced speech by the proposed method. Figure 7 illustrates the graphical representation of formant frequency distances for vowel sound /e/. 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 /e/ which is more interpretable than Table 3.

**Table 3** Speech formant frequencies measurement

| Name | | Formant frequencies | | | | | |
|---|---|---|---|---|---|---|---|
| Vowel | Speech | F1 | F2 | F3 | F4 | F5 | F6 |
| /a/ | AM | 337 | 1214 | 2131 | 3266 | 5585 | 5961 |
| | TM | 351 | 2083 | 3587 | 3747 | 5743 | 6249 |
| | Enhanced | 347 | 1286 | 2097 | 3210 | 5527 | 5951 |
| /e/ | AM | 431 | 2040 | 2496 | 3167 | 6105 | 7365 |
| | TM | 378 | 1977 | 2064 | 2448 | 6360 | 7322 |
| | Enhanced | 490 | 2002 | 2389 | 3114 | 6073 | 7312 |
| /i/ | AM | 452 | 1698 | 1980 | 2774 | 4102 | 5173 |
| | TM | 466 | 1938 | 2052 | 2482 | 4202 | 4996 |
| | Enhanced | 452 | 1726 | 1916 | 2598 | 4110 | 4872 |
| /o/ | AM | 303 | 2270 | 2988 | 3872 | 5622 | 6138 |
| | TM | 415 | 2274 | 3236 | 3849 | 5684 | 6031 |
| | Enhanced | 321 | 2463 | 2965 | 3844 | 5686 | 6255 |
| /u/ | AM | 315 | 1239 | 1304 | 2160 | 3275 | 5628 |
| | TM | 322 | 2089 | 2208 | 2477 | 3591 | 5709 |
| | Enhanced | 324 | 1273 | 1822 | 2150 | 3293 | 5684 |

‘F’ represents Formant Frequency in Hz

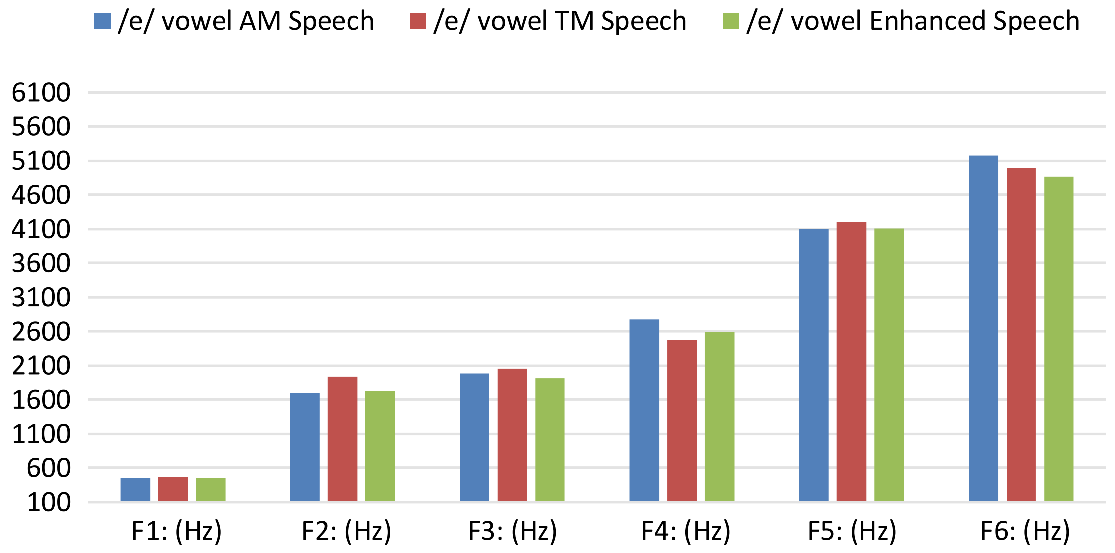

**Fig. 8** Comparison of speech formant distance for /e/ vowel

## 4 Conclusion

In this research, we focused on improving the perceptual quality of the TM speech using a machine learning technique. We used the Multi-Layered Feed Forward Neural

Network (MLFFNN) to model the proposed system. The perceptual quality of a speech signal depends on the acoustic characteristics. The method was divided into two subtasks. The first subtask involved find out the speech features using the MFCC method for both the AM speech and TM speech. The second subtask was spectral mapping using the multi-layered Feedforward Neural Network (MLFFNN). We used ATR503 Dataset for both the training and testing process. The output of the neural network was the enhanced speech that corrected the missed and degraded frequencies of the TM speech. Various subjective and objective measures were taken to evaluate the performance of the proposed method. The result shows a noticeable performance in the field of speech communication in adverse environments.

## References

1. Gibbon, D. (2001). *Prosody: Rhythms and melodies of speech* (pp. 1–35). Germany: Bielefeld University.
2. Shahina, A., & Yegnanarayana, B. (2007). Mapping speech spectra from throat microphone to close-speaking microphone: A neural network approach. *EURASIP Journal on Advances in Signal Processing, 2007*, 1–10.
3. Murty, K. S. R., Khurana, S., Itankar, Y. U., Kesheorey, M. R., & Yegnanarayana, B. (2008). Efficient representation of throat microphone speech. In *INTERSPEECH, 9th Annual Conference of the International Speech Communication Association* (pp. 2610–2613).
4. Rumellhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Learning internal representation by error backpropagation* (pp. 318–362). Cambridge, MA: MIT Press.
5. Vijayan, K., & Murty, K. S. R. (2014). Comparative study of spectral mapping techniques for enhancement of throat microphone speech. In *Twentieth National Conference on Communications (NCC)*, Kanpur (pp. 1–5).
6. Sremath, S., Reza, S., Singh, A., & Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications, 90,* 250–271.
7. Shahina, A. & Yegnanarayana, B. (2014). Artificial neural networks for pattern recognition. *Sadhana, 19*(Part 2), 189–238.
8. Min, G., Zhang, X., Yang, J., & Zou, X. (2015). Speech reconstruction from Mel frequency cepstral coefficients via $\ell 1$-norm minimization. In *IEEE 17th International Workshop on Multimedia Signal Processing (MMSP),* Xiamen (pp. 1–5).
9. Hussain, T. (2017). Experimental study on extreme learning machine applications for speech enhancement. *IEEE Access, 5,* 1–13.
10. Roy, S. K. (2016). *Single channel speech enhancement using Kalman filter* (Master's thesis), Concordia University.
11. Ali, M., & Turan, T. (2013). *Enhancement of throat microphone recordings using gaussian mixture model probabilistic estimator* (Master's thesis), Koc University.