International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

A comparative study for throat microphone speech enhancement with different approaches

Md. Easir Arafat [1, 2, *], Indraneel Misra [1, 2] and Md. Ekramul Hamid [3]

[1] *Department of Computer Science & Engineering, Bangladesh.*
[2] *Pundra University of Science & Technology, Bogura, Bangladesh.*
[3] *University of Rajshahi, Rajshahi, Bangladesh.*

International Journal of Science and Research Archive, 2024, 13(01), 850–859

Publication history: Received on 23 July 2024; revised on 16 September 2024; accepted on 18 September 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.1.1631

Abstract

Throat microphones (TM) offer significant advantages in noisy environments by capturing speech signals directly from the throat, thus minimizing external noise. However, TM signals often lack clarity and intelligibility compared to conventional microphones. This paper presents a comparative study of three prominent feature extraction techniques—Mel-frequency cepstral coefficients (MFCC), Linear Predictive Cepstral coefficients (LPCC), Perceptual Linear Prediction (PLP) for enhancing speech captured by throat microphones. Each technique is evaluated based on its ability to enhance speech clarity and reduce noise interference. Experimental results on the ATR503 dataset, consisting of throat and close-talk microphone recordings, reveal that LPCC achieved an average Signal-to-Noise Ratio (SNR) improvement of 3dB and a Perceptual Evaluation of Speech Quality (PESQ) score increase of 1.3133 and 0.9553 compared to MFCC and PLP. In subjective evaluations the highest mean rating of 8.46 for LPCC indicates it was perceived as the most intelligible and clear. LPC spectra analysis demonstrates that Linear Predictive Cepstral Coefficients (LPCC) in retrieving missing frequencies in speech captured by throat microphones. These findings suggest that LPCC is a robust method for throat microphone speech enhancement, offering significant improvements in speech intelligibility and quality in noisy environments.

Keywords: Throat Microphone (TM); Mel-frequency cepstral coefficients (MFCC); Linear Predictive Cepstral coefficients (LPCC); Perceptual Linear Prediction (PLP); LPC Spectra; Perceptual Evaluation of Speech Quality (PESQ); Signal-to-Noise Ratio (SNR)

1. Introduction

Speech communication in noisy environments remains a persistent challenge across various domains, ranging from military operations to industrial settings. Throat microphones offer a promising solution by capturing speech directly from the larynx, thereby mitigating ambient noise and improving speech intelligibility. However, achieving optimal speech quality from throat microphone recordings necessitates effective speech enhancement techniques [1] [2].

Throat microphones, designed to pick up vocal vibrations directly from the neck, are renowned for their robustness in noisy conditions where conventional microphones fail to deliver clear signals. This capability makes them indispensable in applications such as speech communications, where reliable voice transmission in adverse acoustic environments is critical [1].

* Corresponding author: Easir Arafat

1.1. Problem Statement

Despite their advantages, throat microphones often capture speech signals contaminated with noise, posing significant challenges for intelligibility and reliability. Various techniques have been proposed to enhance speech quality in throat microphone recordings, including Mel-frequency cepstral coefficients (MFCC), Linear Predictive Cepstral coefficients (LPCC) and Perceptual Linear Prediction (PLP)[3][4][5]. Each technique addresses different aspects of noise reduction and speech clarity, yet their comparative effectiveness in the context of throat microphones remains underexplored.

Objectives

This study aims to conduct a comprehensive comparative analysis of MFCC, LPCC, and PLP techniques for enhancing speech captured by throat microphones. The primary objectives include evaluating the performance of these techniques in terms of noise reduction and speech quality enhancement. By identifying strengths and limitations of each method, this research seeks to provide valuable insights for optimizing speech enhancement strategies in throat microphone applications.

2. Literature Review

A throat microphone speech enhancement techniques aim to improve the quality and intelligibility of speech signals captured by throat microphones. This section provides an overview of well-known techniques—Mel-frequency cepstral coefficients (MFCC), Linear Predictive Coding coefficients (LPCC) and Perceptual Linear Prediction (PLP), and reviews previous studies on their application in enhancing speech from throat microphones.

2.1. Mel-frequency cepstral coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCC) [3] are a feature extraction technique in speech processing that captures the power spectrum of audio signals to mimic human auditory perception.

The process starts with a pre-emphasis filter $y[n]=x[n]-\alpha x[n-1]$, typically with $\alpha=0.97$, followed by framing, then windowing by hamming window windowing using $w[n]=0.54-0.46\cos(\frac{2\pi n}{N-1})$and then computing the FFT to obtain the power spectrum $P[k]=|X[k]|^2$.The power spectrum is passed through a Mel-scale filter bank, where the Mel scale is defined as $m=2595\log_{10}(1+\frac{f}{700})$.

The logarithm of the filter bank output is then transformed using the Discrete Cosine Transform (DCT) to produce the

$$\text{MFCCs: } \sum_{m-1}^{M} log(x_m)\cos[\frac{\pi n\,(0-0.5)}{M}] \qquad \ldots\ldots\ldots\ldots \text{ (i)}$$

These coefficients are widely used in speech recognition, speaker identification, and audio classification due to their ability to effectively capture the characteristics of speech signals.

2.2. Linear Predictive Cepstral coefficients (LPCC)

Linear Predictive Cepstral Coefficients (LPCC) [4] are derived from the Linear Predictive Coding (LPC) method, which models the speech signal as a linear combination of its past samples.

The LPC analysis yields the predictor coefficients a_i by minimizing the prediction error $e[n]=x[n]-\sum_{i=1}^{p} a_i\, x[n-i]$, where $x[n]$ is the speech signal and p is the order of the predictor. The LPCCs are then calculated from these LPC coefficients using the recursion

$$c_k = \alpha_k + \sum_{i=1}^{k-1} (\frac{i}{k})\, c_i a_{k-i} \text{ for } k \geq 1 \text{ and } c_0 = \log(E) \qquad \ldots\ldots\ldots\ldots\ldots\text{(ii)}$$

Where E is the prediction error energy. This transformation produces cepstral coefficients that are more robust for speech recognition tasks. LPCCs are widely used in speech and speaker recognition due to their ability to effectively represent the spectral properties of the speech signal.

2.3. Perceptual Linear Prediction (PLP)

Perceptual Linear Prediction (PLP) [5] is a feature extraction technique in speech processing that models the auditory system's perception of sound. The process starts with a pre-emphasis filter and then divides the speech signal into frames, followed by windowing and applying the Fast Fourier Transform (FFT) to obtain the power spectrum. The power spectrum is then warped to the Bark scale using

$$w_b = 6\sinh^{-1}\left(\frac{w}{600}\right) \qquad \text{..... (iii)}$$

Where ω is the frequency in Hz. The warped spectrum is smoothed by convolving it with a critical-band masking curve, and then downsampled to obtain the auditory spectrum. Finally, the Linear Predictive Coding (LPC) model is applied to the auditory spectrum, resulting in the PLP coefficients, which are more closely aligned with human hearing characteristics and improve the robustness of speech recognition systems.

2.4. Previous Research

Previous research has explored the effectiveness of these techniques in enhancing speech from throat microphones:

- **MFCC**: Studies have shown that MFCCs effectively reduce noise and enhance speech intelligibility in throat microphone recordings, making them suitable for applications in noisy environments [1] [3].
- **LPCC**: Research indicates that LPCCs can accurately model the vocal tract and improve speech quality by minimizing noise interference in throat microphone signals [1] [4].
- **PLP**: Studies have highlighted PLP's capability to enhance speech quality by emphasizing perceptually relevant features and suppressing noise components in throat microphone recordings [5].

These techniques have been applied in various contexts, demonstrating their potential to enhance speech quality in challenging acoustic environments. However, comparative studies evaluating their performance specifically in throat microphone applications are limited, underscoring the need for further investigation to identify the most effective technique under different noise conditions and signal characteristics.

3. Methodology

3.1. Data Collection

This research utilizes a dataset based on the ATR503 phoneme balance statements, designed for equal phoneme representation. The ATR503 dataset includes 10 sets (A to J) with a total of 503 sentences. Recordings were made using close-talk and throat microphones in a soundproof room, involving 14 speakers (8 males and 6 females). Initially recorded at 44 kHz, the audio was downsampled to 16 kHz for standardization.

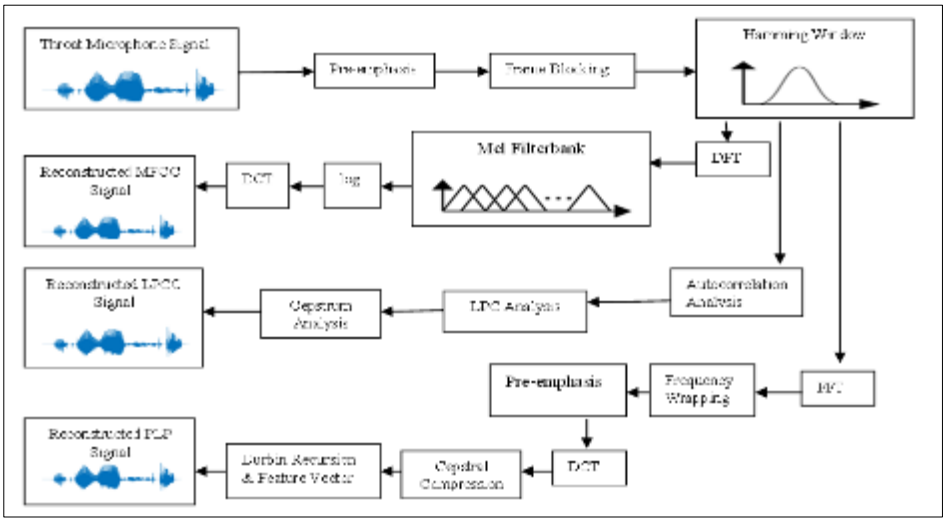

Figure 1 Internal diagram of various coefficient findings algoithms

In fig-1, show the internal diagram of various coefficient finding algorithm.

3.2. Feature Extraction

3.2.1. Mel-frequency cepstral coefficients (MFCC)

The throat microphone signal is loaded from the dataset, and the sampling rate is set to 16kHz. The setup includes 13 MFCC coefficients, a frame length of 25 ms, a frame shift of 10 ms, 26 Mel filter banks, and a pre-emphasis filter coefficient of 0.97. A pre-emphasis filter is applied to the signal to amplify high frequencies, thereby improving the signal-to-noise ratio. The speech signal is segmented into overlapping frames of 25 ms with a 10 ms shift. Each frame is multiplied by a Hamming window to reduce spectral leakage. The power spectrum for each frame is computed using the Fast Fourier Transform (FFT) with 512 points. A Mel filterbank is created to convert frequencies to the Mel scale, mapping them to FFT. The filterbank is then applied to the power spectrum, followed by taking the logarithm of the energies. The Discrete Cosine Transform (DCT) is performed to obtain the MFCCs, yielding 13 coefficients per frame. The process also attempts to reconstruct the audio signal by applying the inverse DCT and inverse filterbank to the MFCCs. The frames are overlapped and combined to reconstruct the signal, followed by normalization. The reconstructed speech signal is then saved as an audio file with the appropriate sampling rate for correct playback. The figure-2 shows three spectrograms, which are visual representations of sound signals. Each spectrogram displays the frequency content of a sound over time. The top spectrogram represents the sound captured by a close microphone, the middle one represents the sound captured by a throat microphone, and the bottom one represents the MFCC (Mel-frequency cepstral coefficients) signal derived from the throat microphone.

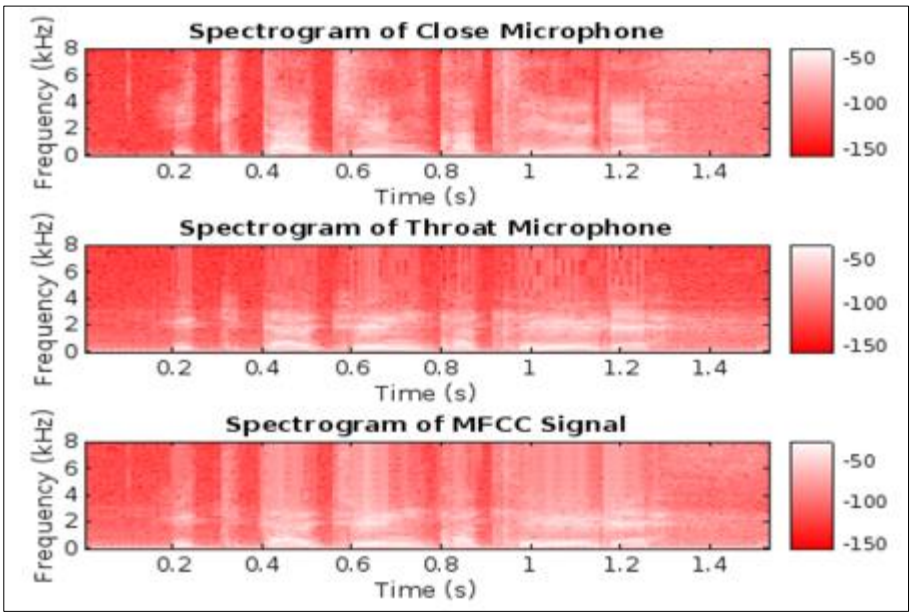

Figure 2 Spectrogram comparison among three different speech signal

3.2.2. Linear Predictive Coding coefficients (LPCC)

The throat microphone signal is loaded from the dataset, and the sampling rate is set to 16 kHz. The signal is normalized by dividing each sample by the maximum absolute value, ensuring the amplitude ranges from -1 to 1 for consistent signal processing. The frame length is defined as 256 samples, with an overlap of 128 samples between consecutive frames. A hamming window of length 256 is applied to each frame, reducing spectral leakage and improving frequency analysis. The order of the Linear Predictive Coding (LPC) model is set to 12. LPC coefficients are computed 13 for each frame using the autocorrelation method and then converted to Linear Predictive Cepstral Coefficients (LPCC). This transformation helps in efficiently representing and analyzing speech signals by converting LPC coefficients to cepstral coefficients, which represent the spectral envelope of the signal in the cepstral domain. A vector is initialized to store the reconstructed signal. For each frame, the LPCC coefficients are extracted and converted back to LPC coefficients. A synthesis filter is applied using these LPC coefficients to reconstruct the frame. The starting and ending indices determine where the reconstructed frame is placed within the overall reconstructed signal. The reconstructed frame is added to the signal, completing the reconstruction process by implementing the inverse transformation from the cepstral to the LPC domain, which is essential for speech signal reconstruction. The reconstructed speech signal is then saved as an audio file with the appropriate sampling rate for correct playback. The figure-3 shows three spectrograms, which are visual representations of sound signals. Each spectrogram displays the frequency content of a sound over

time. The top spectrogram represents the sound captured by a close microphone, the middle one represents throat microphone, and the bottom one represents the LPCC signal derived from the throat microphone.

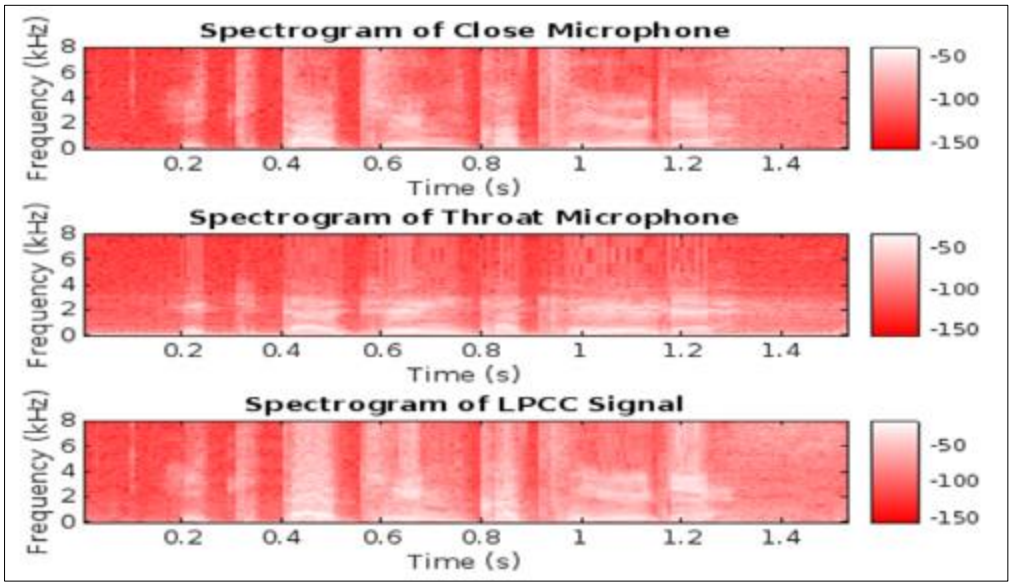

Figure 3 Spectrogram comparison among three different speech signal

3.2.3. Perceptual Linear Prediction (PLP)

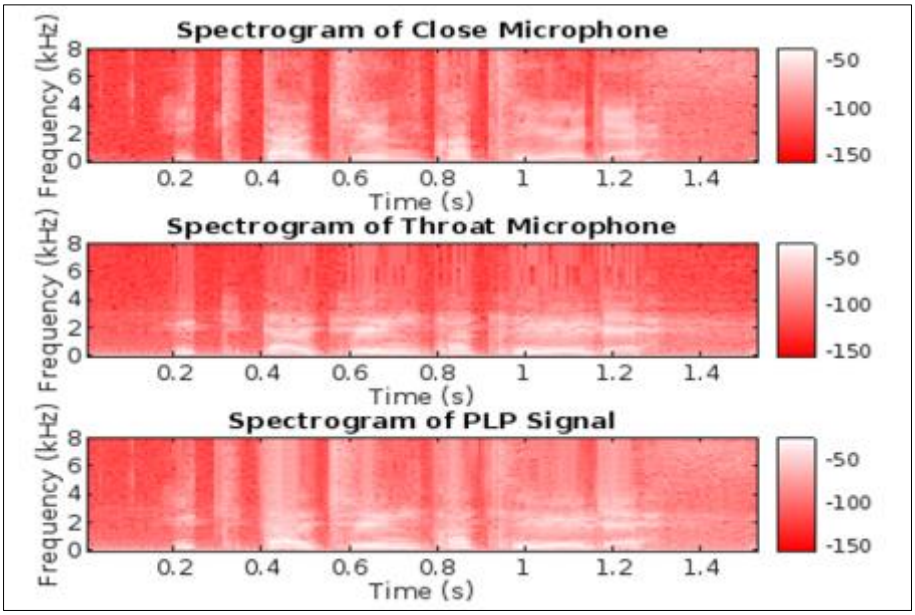

Figure 4 Spectrogram comparison among three different speech signal

The throat microphone signal is loaded from the dataset, and the sampling rate is set to 16 kHz. Parameters for LPC analysis are defined to the frame length is set to 0.025 seconds, and the frame shift to 0.010 seconds. The LPC order is specified as 12, determining the number of coefficients computed for each frame of the speech signal. A high-pass filter is applied to enhance high frequencies, improving the stability of LPC analysis by balancing the speech signal spectrum. The pre-emphasized speech signal is divided into overlapping frames with a specified overlap between consecutive frames. LPC coefficients 13 computed for each frame using the autocorrelation method and recursion. The autocorrelation of each frame is determined, followed by applying recursion to calculate LPC coefficients. These coefficients are used with random noise excitation to synthesize speech frames. Random noise is filtered through each LPC filter to produce synthesized speech frames. The overlapping synthesized frames are combined to reconstruct the full synthesized speech signal, enhancing continuity and smoothness. The synthesized speech signal is then saved as an

audio file with the appropriate sampling rate for correct playback. The figure-3 shows three spectrograms, which are visual representations of sound signals. Each spectrogram displays the frequency content of a sound over time. The top spectrogram represents the sound captured by a close microphone, the middle one represents throat microphone, and the bottom one represents the PLP signal derived from the throat microphone.

It is observed from fig 2-4 that in throat microphone speech high frequencies are totally missing which is significantly restored in the enhanced spectrogram by MFCC, LPCC and PLP coefficients.

4. Experimental Setup

4.1. Tools and Software

- Programming Languages: The experiments are implemented using MATLAB. MATLAB is utilized for its extensive signal processing toolboxes
- Speech Processing Toolboxes: MATLAB's Signal Processing Toolbox for feature extraction (MFCC, LPCC, PLP) and speech enhancement algorithm development.
- Development Environment: Experimentation and analysis are conducted on high-performance computing systems equipped with multicore processors to ensure efficient processing of large datasets and complex algorithms.

5. Results and discussion

5.1. Signal-to-Noise Ratio (SNR)

Quantifies the ratio of signal power to noise power before and after speech enhancement. Higher SNR values indicate better noise reduction and improved speech clarity [7]. During this experiment, the environmental noise is picked up by the CM which is around 5 dB .The figure-5 shows the signal to noise ratio (SNR) in decibels (dB) for CM (at 5dB SNR), TM speech and enhanced speech by MFCC, LPCC and PLP coefficients. The highest SNR is achieved by the LPCC signal, followed by the TM signal. The lowest SNR is achieved by the MFCC signal. This indicates that the LPCC signal has the best performance in terms of noise reduction. The other signals have lower SNRs, which suggests that they may have more noise present in the signal.

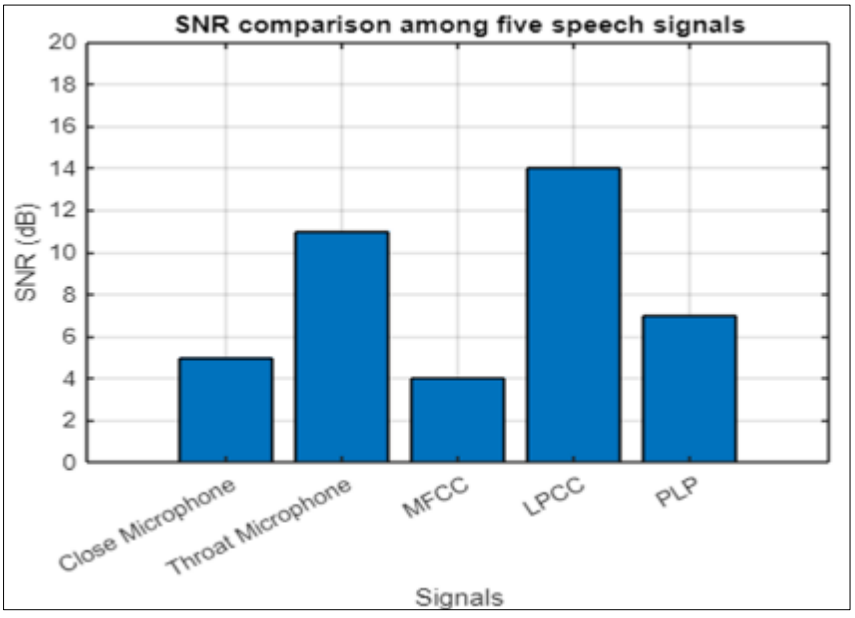

Figure 5 SNR Comparison among CM, TM speech and enhanced speech by MFCC, LPCC and PLP coefficients

5.2. Perceptual Evaluation of Speech Quality (PESQ)

Assesses the perceived quality of enhanced speech compared to the original signal. PESQ scores range from 1 (poor) to 5 (excellent), providing a subjective measure of speech enhancement effectiveness [6].

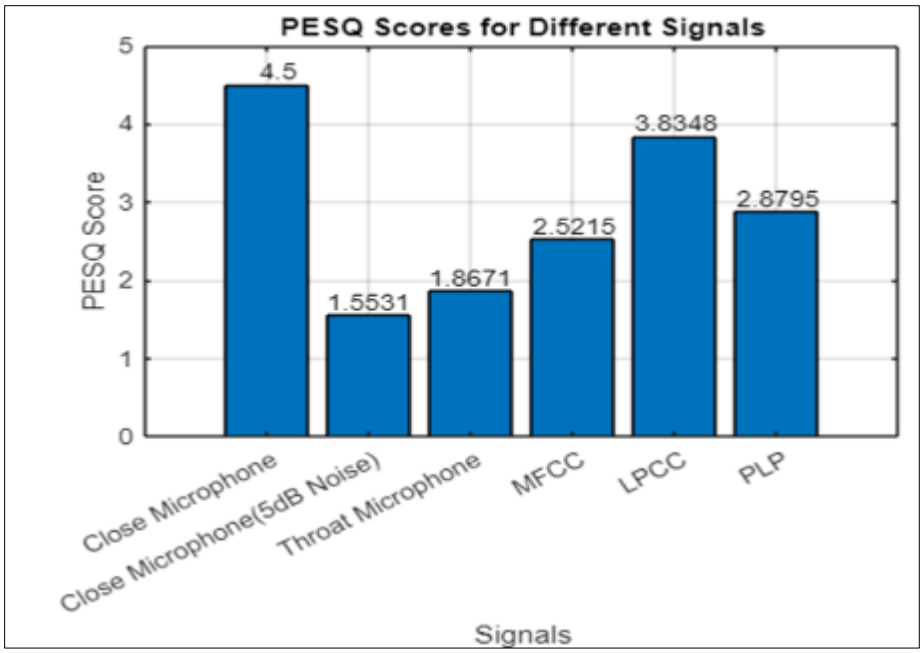

Figure 6 PESQ Score comparison among CM, TM speech and enhanced speech by MFCC, LPCC and PLP coefficients

In the above figure 6, shows the PESQ scores for CM (at 5dB SNR), TM speech and enhanced speech by MFCC, LPCC and PLP coefficients. The PESQ score is a measure of speech quality, with higher scores indicating better quality.

It is observed from fig 6 that when environmental noise is at 5 dB SNR, the Close Microphone speech signal, it's PESQ score becomes the lowest compared to the other speech signals. The Throat Microphone signal has a PESQ score of 1.8671, indicating that it has significantly lower speech quality than the Close Microphone.

The MFCC signal has a PESQ score of 2.5215, indicating that it has slightly better speech quality than the Throat Microphone. The LPCC signal has a PESQ score of 3.8348, indicating that it has slightly lower speech quality than the Close Microphone.

Finally, the PLP signal has a PESQ score of 2.8795, indicating that it has lower speech quality than both the MFCC and LPCC signals. Overall, the results suggest that the LPCC signal offers the best speech quality, while the Close Microphone signal with noise added provides the worst quality.

5.3. LPC Spectra

In the following figure-7, The LPC spectra [9] comparison reveals significant differences in the frequency characteristics and magnitude responses of the various signals:

In the Close Microphone (CM) signal, the blue curve exhibits the highest peaks in the low-frequency range (below 2000 Hz), indicating that the close microphone captures more energy in these lower frequencies compared to the other signals. However, when 5 dB of noise is added to the close microphone signal, its performance significantly degrades.

In Throat Microphone (TM), the orange curve exhibits a peak around 1000 Hz, which is typical for throat microphones as they emphasize certain lower-frequency components and attenuate higher frequencies at 5000 Hz.

In the MFCC and PLP signals, the spectral curves (purple and green) exhibit similar patterns, indicating smoother spectral characteristics. However, MFCC fails to capture high-frequency components beyond 3000 Hz, while PLP retrieves frequencies only up to 7000 Hz.

In the LPC spectra comparison reveals that the LPCC (Linear Predictive Cepstral Coefficients) curve demonstrates superior performance in retaining important frequency components, particularly in the lower frequency range. The LPCC spectrum (blue curve) not only closely aligns with the Close Mic (CM) signal but also successfully captures the missing frequency components of the Throat Mic (TM) signal, which are otherwise attenuated in the MFCC and PLP spectra.

The LPCC (Linear Predictive Cepstral Coefficients) spectrum demonstrates superior performance by effectively retaining critical frequency components, particularly in the lower frequency range. LPCC successfully retrieves frequency components that are attenuated in the Throat Microphone (TM) signal. In contrast, MFCC fails to capture high-frequency components beyond 3000 Hz, and PLP only retrieves frequencies up to 7000 Hz. Overall, LPCC outperforms other methods by effectively preserving and recovering important frequency details that are otherwise lost.

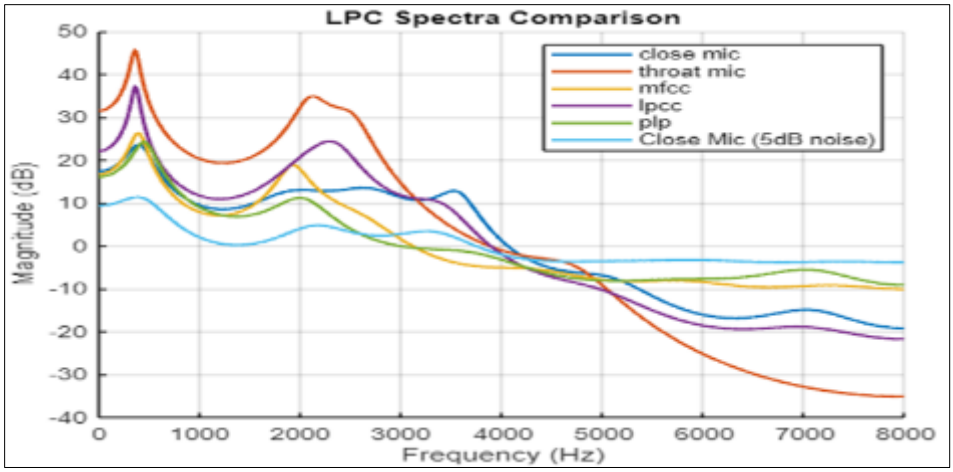

Figure 7 LPC Spectra comparison among six speech signals

5.3.1. Subjective Study of Speech Intelligibility

To assess speech intelligibility, a subjective study is conducted with a diverse group of 15 participants where 9 males and 6 females. Each participant listen to a set of speech samples and rated the intelligibility on a scale from 1 to 10, with 1 being completely unintelligible and 10 being perfectly intelligible. The ratings are collected through a structured questionnaire.

The following table-1 summarizes the ratings provided by each participant:

Table 1 Subjective Study of Speech Intelligibility

No. of Speech	Participant Name	Average Rating (1-10)			
		Throat Microphone	MFCC Signal	LPCC Signal	PLP Signal
25 Different Speech	Sabah	8	9	8	7
	Mokrema	9	7	9	6
	Labony	8	9	8	7
	Sabuj	7	8	9	8
	Munna	9	9	8	7
	Sabina	8	7	8	9
	Tarikul	9	8	9	8
	Kabir	9	9	9	7
	Riyad	7	8	8	6

	Rekha	8	7	8	9
	Mehedi	8	9	9	7
	Mashrafee	7	7	9	9
	Raiyan	8	6	8	8
	Jaima	9	9	10	7
	Zalal	8	9	7	9

The average rating across all participants was calculated to provide an overall measure of speech intelligibility. The mean [8] rating of Throat Microphone:

$$\text{Mean Rating (TM)} = \frac{8+9+8+7+9+8+9+9+7+8+8+7+8+9+8}{15} = 8.134$$

$$\text{Mean Rating (MFCC)} = \frac{9+7+9+8+9+7+8+9+8+7+9+7+6+9+9}{15} = 8.06$$

$$\text{Mean Rating (LPCC)} = \frac{8+9+8+9+8+8+9+9+8+8+9+9+8+9+8}{15} = 8.46$$

$$\text{Mean Rating (PLP)} = \frac{7+6+7+8+7+9+8+7+6+9+7+9+8+7+9}{15} = 7.60$$

The average ratings from the subjective study for Throat Microphone (8.134), MFCC (8.06), LPCC (8.46), and PLP (7.60) suggest that LPCC-enhanced speech is generally perceived as highly intelligible. The ratings, which range from 6 to 10, reflect some variability in how different participants perceive speech intelligibility. This variability could be influenced by individual differences in hearing ability, familiarity with the accent, or the quality of the listening environment. Notably, participant Zalal awarded a perfect rating of 10 to the LPCC speech signal, likely due to its superior clarity, naturalness, and overall intelligibility. In contrast, participant Raiyan gave the lowest rating of 6 to the PLP speech signal, possibly due to perceived issues with clarity or increased distortion.

These findings highlight the importance of considering individual differences when evaluating speech intelligibility. The data suggests that LPCC consistently achieves higher average ratings compared to the other methods.

6. Conclusion

In noisy environments, throat microphones (TM) are advantageous for capturing speech directly from the throat, but they often struggle with clarity and intelligibility compared to traditional microphones. This study addressed the challenge of enhancing TM speech quality by comparing three prominent feature extraction techniques: Mel-frequency cepstral coefficients (MFCC), Linear Predictive Cepstral Coefficients (LPCC), and Perceptual Linear Prediction (PLP). We evaluated these techniques using the ATR503 dataset, which includes recordings from both throat and close-talk microphones.

Our results demonstrated that LPCC significantly outperforms MFCC and PLP in enhancing speech captured by throat microphones. LPCC achieved an average Signal-to-Noise Ratio (SNR) improvement of 3 dB and led to substantial increases in Perceptual Evaluation of Speech Quality (PESQ) scores, indicating better speech clarity and intelligibility. Subjective evaluations further support these findings, with LPCC receiving the highest mean rating for clarity and intelligibility.

Despite these advancements, our study has limitations. The performance of LPCC might vary with different types of noise and acoustic conditions not covered in this study. Future research could explore these variations and investigate other feature extraction techniques or hybrid methods to further enhance speech quality. Additionally, applying these techniques to real-world applications and larger datasets could provide more comprehensive insights into their effectiveness.

Overall, this research provides a robust foundation for optimizing speech enhancement techniques for throat microphones and highlights LPCC as a particularly effective method. Further exploration and refinement of these

techniques could lead to even greater improvements in speech intelligibility and quality in challenging acoustic environments

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Md. Easir Arafat ,Masafumi Nishimura ,Md. Ekramul Hamid , (2020) " Improvement of Throat Microphone Speech by Enhance Spectral Envelope using GMR-LPC based Method " , International Journal of Advance Computational Engineering and Networking (IJACEN) , pp. 10-14, Volume-8,Issue-5.

[2] Throat Subrata Kumar Paul, Rakhi Rani Paul, ,Masafumi Nishimura ,Md. Ekramul Hamid (2020) "Microphone Speech Enhancement Using Machine Learning Technique" Chapter: Learning and Analytics in Intelligent Systems

[3] Amritha Vijayan, Bipil Mary Mathai, Karthik Valsalan, Riyanka Raji Johnson, L. Mathew, K. Gopakumar (2017) "Throat microphone speech recognition using mfcc" International Conference on Networks & Advances in Computational Technologies (NetACT).

[4] Harshita Gupta, Divya Gupta (2016) "LPC and LPCC method of feature extraction in Speech Recognition System" Conference: 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)

[5] Mahmud, Nahyan Al and Munni, Shahfida Amjad, Qualitative Analysis of PLP in LSTM for Bangla Speech Recognition (2020). The International Journal of Multimedia & Its Applications (IJMA) Vol.12, No. 5, October 2020, Available at SSRN: https://ssrn.com/abstract=3727781

[6] Antony W. Rix, John G. Beerends, Michael P. Hollier (2024) "Perceptual Evaluation of Speech Quality (PESQ): A New Method for Speech Quality Assessment of Telephone Networks and Codecs" February 2001 ,Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on 2:749-752 vol.2 ,2:749-752 vol.2

[7] Naser Elkum, Mohamed M Shoukri (2008), "Signal-to-noise ratio (SNR) as a measure of reproducibility: Design, estimation, and application" Health Services and Outcomes Research Methodology 8(3):119-133, 8(3):119-133.

[8] Lee DK, In J, Lee S (2015). Standard deviation and standard error of the mean. Korean J Anesthesiol. Jun;68(3):220-3. [PMC free article] [PubMed]

[9] Hugo Tito Cordeiro, José Manuel Fonsecaa , Carlos Meneses Ribeirob (2013) International Conference on Project MANagement / HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies