2026 IEEE International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE 2026) 29–31 January 2026, Rajshahi-6204,

Speech Enhancement Using Neural Network-Based Residual Approach

Tanvir Mahtab Navid
Department of EEE
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
navid1903002@stud.kuet.ac.bd

Tithi Das
Department of EEE
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
tithidas2k18@gmail.com

Mohaimenul Islam
Department of EEE
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
mohaimenul.work.2001@gmail.com

Md Rakibul Hasan
School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University
Bentley, WA 6102, Australia
rakibul.hasan@curtin.edu.au

Md Mahbub Hasan
Department of EEE
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
mahbub01@eee.kuet.ac.bd

***Abstract*—Speech enhancement in noisy environments remains a challenging problem, especially when noise interferes with the excitation component of speech. Because of its interpretability and low latency, Linear Predictive Coding (LPC) is widely utilized in the speech processing domain. However, LPC performance becomes less reliable when noise corrupts the excitation signal. In this work, we proposed a hybrid speech enhancement approach that combined LPC with a fully connected neural network to refine the excitation residual while keeping the LPC spectral envelope unchanged. For each noisy speech frame, we extracted LPC coefficients and residuals. Then we fed a short context of residual frames into the neural network to predict a cleaner excitation signal. The enhanced residual was then passed through the original LPC filter to reconstruct the speech. We used an existing dataset of isolated Bengali vowels and words, which was then corrupted with 15 dB additive white Gaussian noise to create clean–noisy pairs for each vowel and word. Evaluation metrics from the system showed significant and consistent improvement in spectral SNR across LPC orders from 12 to 24. We observed peak improvement of about +9.4 dB for vowels using LPC order 16 and about +20.4 dB for words using LPC order 20. The spectrograms illustrated clear reductions of background noise and improved harmonic structure. This suggested that improving the excitation signal created an important step towards a middle ground between traditional methods of LPC and modern enhancement methods using neural networks.**

Index Terms—Speech enhancement, Linear predictive coding (LPC), Residual denoising, Neural networks, Excitation modeling, Additive white Gaussian noise, Low latency, Bengali speech corpus.

I. INTRODUCTION

Speech is an essential part of human interaction. In today's era, most of the interactions are based on technology. The mechanism involves recording the original speech, transmitting the speech signal, and then receiving it. In the process of recording, some background noise is added to it, which masks the clarity of speech information and makes communication difficult. In the field of speech processing, numerous noise reduction methods have been established over time, including spectral subtraction, Wiener filtering, and MMSE-based techniques [1]–[4]. But a major drawback of these is that they work well only in constant noise environments. Yet the performance often degrades when the noise signal fluctuates [5]. This can lead to undesirable effects such as musical noise or missing important speech components.

Even with all the advancements in noise reduction techniques, Linear Predictive Coding (LPC) continues to be a popular and effective tool in the area of speech processing. It is highly efficient due to its relatively simple computation and representation of the speech structure [6], [7]. LPC allows us to distinguish between the vocal tract's spectral envelope and the source of the sound by representing the signal as the output of an all-pole filter driven by a residual signal. The LPC model has a key advantage over other methods, which is that its coefficients tend to stay stable even with moderate background noise. To improve the speech quality effectively, we can enhance the excitation residual, particularly where much of the noise is concentrated.

Throughout the past few years, the speech enhancement field has seen observable growth by deep neural networks (DNNs), often producing very strong results [8]–[10]. However, these models require a lot of training data and substantial computing power, hence it can result in latency. Thus, these are less practical for real-time devices that need to operate rapidly. Because of these limitations, researchers have started

979-8-3315-6135-2/26/$31.00 ©2026 IEEE

Available Online at https://ieeexplore.ieee.org/document/11429315

to look toward hybrid strategies that take advantage of the clear structure offered by LPC modeling while using neural networks only where they provide meaningful improvement, particularly in refining the excitation signal.

For example, fully convolutional models that operate directly on the raw speech waveform can improve how the enhanced signal sounds, but they often have difficulty recovering accurate phase information and subtle high-frequency components [11]. Likewise, methods that combine LPC with spectral subtraction and voice activity detection (SS-VAD) perform reasonably well when the background noise stays fairly constant; however, once the noise begins to change or fluctuate, their performance typically drops [12]. More recently, LPCSE has attempted to integrate neural modeling directly into the LPC structure, but it alters both the spectral envelope and the residual and still faces challenges in real-time deployment [13]. Together, these studies suggest that LPC-guided neural enhancement is promising, but there is still a need for methods that enhance only the residual, maintain the LPC envelope, and remain efficient enough for real-world use.

Beyond model architecture, perceptual and intelligibility-oriented metrics such as PESQ, STOI, and spectral SNR are widely used to evaluate enhancement quality beyond simple waveform similarity [14], [15]. Prior research in residual-domain enhancement also shows that excitation signals contain structured temporal patterns that compact neural models can learn to refine, reducing noise without altering the spectral envelope [16]–[18].

Motivated by these insights, this work introduces a residual-focused enhancement method that pairs LPC analysis with a lightweight Fully Connected Neural Network (FCNN). The LPC envelope estimated from the noisy speech is kept unchanged, and only the excitation residual is refined. Context-stacked residual frames are standardized and mapped to clean excitation predictions using the FCNN, after which inverse LPC filtering and overlap–add synthesis are used to reconstruct the enhanced speech signal. This approach improves the speech signal in a way that is easy to interpret and keeps the LPC envelope intact using a lightweight model that can run in real time.

The key contribution of this work is demonstrating that enhancing only the LPC residual, while keeping the spectral envelope fixed, is sufficient to produce higher-quality speech without increasing computational complexity.

We evaluated the proposed system upon a publicly available dataset that includes isolated Bengali vowels and words, in which 15 dB of additive white Gaussian noise was added to each clean signal to create clean-noisy pairs [19]. Across LPC orders from 12 to 24, the system consistently increased spectral SNR and generated spectrograms with more defined harmonic structures. The results suggest that the LPC residual maintains meaningful excitation structure that can even be refined meaningfully by a small neural network. Since this method simply adds a lightweight neural layer to the standard LPC pipeline, it is computationally efficient and appropriate for real-time and embedded processing.

II. METHODOLOGY

A. Front-End and LPC Analysis

To implement our framework for improvement, we start with paired clean and noisy versions of the same speech utterance. Noisy signals are generated by adding additive white Gaussian noise to the clean recordings at an overall signal-to-noise ratio (SNR) of 15 dB, while maintaining comparable loudness to create a realistic noisy environment. Both signals contain the same speech content, differing only in the added noise, allowing the model to learn how noise perturbs the excitation signal while preserving the underlying speech articulation. After applying a pre-emphasis filter to both clean and noisy waveforms to improve spectral energy balance and prediction filter stability, we perform linear predictive coding (LPC) analysis. The speech signal is then framed using a 20 ms Hann window (320 samples at 16 kHz) with a 5 ms frame shift (80 samples), allowing it to be treated as locally stationary within each frame. Before framing, pre-emphasis was applied with $\alpha = 0.97$ to flatten the spectrum:

$$y[n] = x[n] - 0.97x[n-1] \tag{1}$$

Following this, the pre-emphasized speech signal is segmented into overlapping Hann-windowed frames. For every frame, we then carry out Linear Predictive Coding (LPC) analysis. LPC represents the speech signal as the output of an all-pole vocal-tract filter driven by an underlying excitation signal $e[n]$:

$$x[n] = \sum_{k=1}^{p} a_k \, x[n-k] + e[n] \tag{2}$$

Here, $\{a_k\}_{k=1}^{p}$ are the LPC coefficients of order p. The corresponding analysis filter is defined as

$$A(z) = 1 - \sum_{k=1}^{p} a_k \, z^{-k} \tag{3}$$

The residual signal, representing the excitation, is computed as

$$e[n] = x[n] - \sum_{k=1}^{p} a_k \, x[n-k] \tag{4}$$

This decomposition separates the smooth vocal-tract envelope $A(z)$ from the rapidly varying excitation component $e[n]$. We evaluate LPC orders $p \in \{12,16,20,24\}$ to balance envelope fidelity with robustness against narrowband noise.

During reconstruction, we synthesize the enhanced waveform by filtering the predicted clean residual $\hat{e}[n]$ through the inverse LPC filter constructed from the noisy coefficients:

$$\hat{s}[n] = \text{filter}\left(\frac{1}{A(z)}, \hat{e}[n]\right) \tag{5}$$

We check the LPC coefficient stability frame-by-frame and replace unstable filters with the most recent stable set. By keeping the envelope fixed from the noisy input and modifying only the residual, we ensure low-latency, causal operation and attribute perceptual improvements to the residual enhancement stage.

B. Neural Residual Reconstruction

We design a lightweight fully connected neural network (FCNN) to predict a clean excitation residual from its noisy counterpart. Each input to the network is formed by temporally stacking $2K + 1$ consecutive residual frames centered on the target frame,

$$\mathrm{r}_i = [\, e_{i-K}, \ldots, e_i, \ldots, e_{i+K} \,] \tag{6}$$

where K defines the temporal context radius. We choose $K = 2$, resulting in five frames (two past, one current, two future). With each residual frame containing 320 samples, the input dimensionality becomes $320 \times 5 = 1600$, allowing short-term temporal dependencies to be captured while maintaining low latency.

Before being fed into the network, residual frames are standardized using the mean and variance computed from the training set only, ensuring consistent scaling and preventing data leakage. The FCNN maps the standardized input r_i to a clean residual estimate:

$$\hat{e}_i^{\text{clean}} = f_\theta(\mathrm{r}_i) \tag{7}$$

where f_θ refers to the nonlinear mapping the network has learned throughout the training process.

Our FCNN consists of two hidden layers with 512 and 1024 neurons, respectively, followed by ReLU activations and a linear output layer with 320 units corresponding to the residual frame size. The network is trained to minimize the mean squared error (MSE) between predicted and reference residuals,

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N} \| \hat{e}_i^{\text{clean}} - e_i^{\text{clean}} \|_2^2 \tag{8}$$

where N is the number of training frames. Optimization uses Adam (learning rate 10^{-3}, batch size 64); we train for 30 epochs without early stopping.

During inference, the model processes each noisy residual frame within its local context to produce a denoised estimate. The predicted residuals are de-standardized, overlap–added, and used to reconstruct the enhanced waveform through inverse filtering with the LPC envelope, as in (5). By restricting learning to the residual domain, the LPC envelope remains unchanged, allowing the FCNN to focus on transient and harmonic refinement while preserving causal, interpretable, and low-latency operation.

C. Training, Inference, and Evaluation

We trained the proposed residual enhancement model using the paired clean–noisy dataset described earlier. All preprocessing and normalization parameters were computed from the training split and reused unchanged during inference to avoid information leakage. A random 70/30 file-level split (seed = 42) was applied without enforcing speaker-based partitioning. Training was conducted for 30 epochs without early stopping, where each mini-batch consisted of context-stacked noisy residual frames as inputs and corresponding clean residuals as targets, minimizing the mean-squared error (MSE) defined in (8).

During inference, the network processed a short symmetric context of standardized noisy residual frames (five stacked frames) to predict the denoised center frame. The predicted residuals were de-standardized, overlap–added, and synthesized into the time-domain waveform using the inverse LPC filter $1/A(z)$ constructed from the noisy coefficients, preserving the spectral envelope while replacing the excitation.

In our implementation, the FCNN consists of two fully connected hidden layers with 512 and 1024 neurons, respectively (ReLU activations), followed by a linear output layer with 320 units corresponding to the residual frame size. This results in an input dimensionality of 1600 (= 320 × 5) for a context size of five frames. To evaluate enhancement quality, we employed a no-reference spectral SNR estimator based on short-time Fourier analysis and minimum-statistics noise power estimation. Let the observed noisy speech be:

$$x[n] = s[n] + n[n] \tag{9}$$

where $s[n]$is clean speech and $n[n]$is additive noise.

The short-time Fourier transform (STFT) of the noisy signal is:

$$X(k,m) = \mathrm{STFT}\{x[n]\} \tag{10}$$

with power spectral density:

$$P_X(k,m) = |\, X(k,m) \,|^2 \tag{11}$$

Following minimum-statistics noise estimation [2], the noise floor at each frequency bin is obtained as:

$$\hat{P}_N(k) = \min_m P_X(k,m) \tag{12}$$

The mean spectral energy across time is computed as:

$$\bar{P}_X(k) = \frac{1}{M}\sum_{m=1}^{M} P_X(k,m) \tag{13}$$

Thus, the estimated speech power spectrum is:

$$\hat{P}_S(k) = \max(\bar{P}_X(k) - \hat{P}_N(k), 0) \tag{14}$$

The broadband signal and noise powers are then obtained by summation over frequency:

$$P_{\text{signal}} = \sum_k \hat{P}_S(k) \quad , \quad P_{\text{noise}} = \sum_k \hat{P}_N(k) \tag{15}$$

Finally, the spectral SNR in decibels is defined as:

$$\mathrm{SNR}_{\text{dB}} = 10\log_{10}\left(\frac{P_{\text{signal}}}{P_{\text{noise}}}\right). \tag{16}$$

This no-reference, frequency-domain estimator provides a consistent and speech-structure-aware measure of enhancement quality without requiring access to clean speech. At test time, the system operates using only the noisy input, ensuring that the enhancement stage remains fully deployable in real-world conditions.

D. System Overview

Fig. 1 summarizes the proposed LPC + FCNN pipeline. The noisy input signal $x[n]$ is first analyzed using Linear Predictive Coding (LPC), which decomposes the signal into the all-pole spectral envelope coefficients a_k and the excitation residual $e[n]$. The residual signal is segmented into frames, temporally context-stacked, and then passed to a lightweight fully connected neural network (FCNN) that estimates a denoised residual $\hat{e}[n]$ for the center frame. The enhanced speech signal $y[n]$ is subsequently reconstructed by filtering the predicted residual through the inverse LPC filter $1/A(z)$ derived from the noisy LPC coefficients. Reusing the noise-stable LPC envelope while learning only the residual enables a simple, interpretable, and low-latency design with consistent gains on vowel and word data.

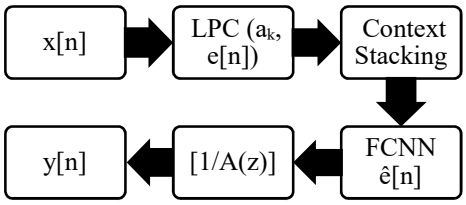

Fig. 1. Block diagram of the proposed LPC + FCNN residual enhancement framework.

We evaluate the proposed LPC + FCNN residual enhancement model on vowel and word datasets using both quantitative and qualitative measures. The next section presents spectral SNR results and spectrogram-based evidence illustrating the effectiveness of residual-domain denoising.

III. RESULTS AND DISCUSSION

A. Objective and Evaluation Setup

We evaluated our method using spectral SNR as the main objective measure. The experiments were carried out separately on vowel and word recordings so that we could observe how the model behaves with shorter vs. longer speech sounds.

To study the effect of prediction detail, we tested four LPC orders: 12, 16, 20, and 24. A MIN-statistics, no-reference SNR estimator was used. The power spectrum was obtained using the STFT, and the noise floor was estimated from the lowest spectral energy over time. This allows us to measure how well noise is separated from speech without needing a clean reference during evaluation.

The dataset contained 280 vowel and 280 word samples. We then employed a speaker-independent 70/30 split to produce 84 test samples in each category. The same files were used to create the clean, noisy, and enhanced sets to maintain consistency and allow for direct comparisons. All recordings were captured using the same sampling rate and framing settings. Noisy signals were made by adding 15 dB AWGN to the clean recordings. As different LPC orders were tested and two separate types of speech units were used, we are able to evaluate how the model performed with variations in the pathways of articulation, and signal-related increased complexity in residuals under an additive level of noise.

B. SNR Performance Across LPC Orders

To evaluate the effectiveness of the enhancement, we calculated the spectral SNR for the clean, noisy, and enhanced signals across different LPC orders. For both vowels and words, Tables I and II report the spectral SNR values as mean ± standard deviation (in dB), along with the SNR improvement (Δ), defined as the difference between enhanced and noisy SNR under identical 15 dB additive white Gaussian noise conditions. These results quantify the level of noise reduction and the corresponding improvement in speech quality. The clean SNR values remain constant across LPC orders because the same clean recordings were used for all configurations.

TABLE I
AGGREGATE SNR RESULTS FOR THE VOWEL DATASET UNDER DIFFERENT LPC ORDERS

LPC Order	Clean SNR (dB)	Noisy SNR (dB)	Enhanced SNR (dB)	Δ vs Noisy (dB)
12	17.74 ± 5.75	17.37 ± 5.27	26.46 ± 7.25	+9.08
16	17.74 ± 5.75	17.37 ± 5.27	26.82 ± 7.43	+9.45
20	17.74 ± 5.75	17.37 ± 5.27	26.33 ± 7.11	+8.96
24	17.74 ± 5.75	17.37 ± 5.27	26.44 ± 7.55	+9.07

TABLE II
AGGREGATE SNR RESULTS FOR THE WORD DATASET UNDER DIFFERENT LPC ORDERS

LPC Order	Clean SNR (dB)	Noisy SNR (dB)	Enhanced SNR (dB)	Δ vs Noisy (dB)
12	36.63 ± 8.13	28.64 ± 3.31	45.32 ± 7.29	+16.68
16	36.63 ± 8.13	28.64 ± 3.31	47.44 ± 8.13	+18.80
20	36.63 ± 8.13	28.64 ± 3.31	49.09 ± 8.02	+20.45
24	36.63 ± 8.13	28.64 ± 3.31	48.54 ± 8.50	+19.90

The SNR values for the noisy signals are slightly lower than the nominal 15 dB because the energy profile for each of the 84 recordings is variable. The results in Table I show our method consistently improves the SNR for vowel recordings with LPC-16, resulting in the highest gain of +9.4 dB for a mean SNR of 26.8 dB. This suggests such a method provides a good compromise between detail and stability for steady vowel sounds.

Table II demonstrates even larger improvements for the word recordings, showing an increase in SNR of +20.4 dB with LPC-20, resulting in an average SNR of 49.1 dB. This shows that the higher-order LPC is better suited to model the time-varying spectral envelopes in words.

C. Residual-Domain Spectrogram Analysis

To analyze how the modification affected the excitation signal, we reviewed a few vowel and word recordings from the perspective of the residual spectrogram. We took two figures, Fig. 2 as a vowel processed with LPC order 20, and Fig. 3 as a word processed with LPC order 24, for display. The upper panels in the figures show the noisy residuals, whilst the lower panels show the enhanced residuals. Importantly, in producing the residual spectrograms, we kept the STFT parameters and color scale consistent to permit valid comparisons among them. For the reconstruction of the enhanced residual, we kept the original noisy LPC coefficients, so any differences observed are due entirely to modifications to the residual, and not due to modifications to the vocal-tract envelope.

Both examples show a clear improvement. In the vowel case, the enhanced residual has harmonic lines that are more regular and well-defined, with less diffuse background energy in between, signifying a more powerful and stable periodic structure. The word example has even more improvement: the noisy residual contains a clear layer of broad background noise, while the enhanced version has much less noise, and the voiced regions appear sharper and more continuous. The harmonic tracks are easier to track across time, which signifies generally improved clarity in the reconstructed speech.

The enhancement generally produces a cleaner residual, but still maintains the natural harmonic organization of the speech. Also, the spectrograms provide evidence that noise reduction can be achieved without modifying the underlying vocal-tract organization. This demonstrates the usefulness of addressing the residual domain to improve the quality of speech, but maintain the original spectral envelope.

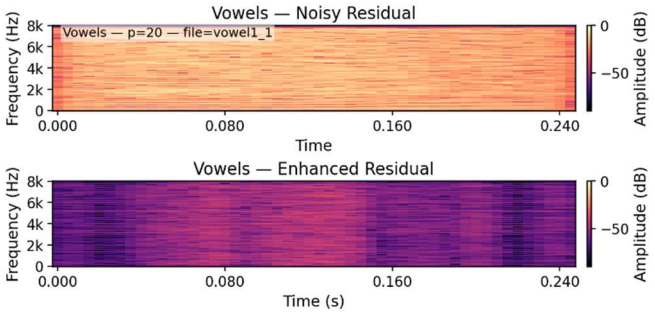

Fig. 2. Residual spectrograms for a vowel (LPC-20): noisy (top) and enhanced (bottom) residuals with identical STFT and shared color scale.

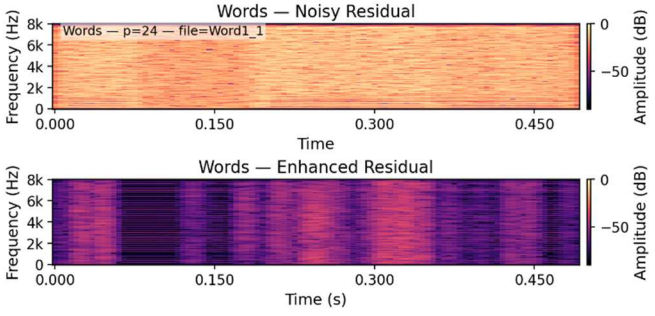

Fig. 3. Residual spectrograms for a word (LPC-24): noisy (top) and enhanced (bottom) residuals showing reduced noise and clearer harmonics.

D. Waveform-Domain Spectrogram Analysis

Furthermore, we analyzed the spectrograms of one vowel sample and one word sample from the test set to compare clean, noisy, and enhanced speech signals. Here, we intentionally chose the spectrogram recordings that are imperfect to represent what the model usually does. We kept the STFT setting and color code the same for all plots in order to focus exclusively on the enhancement. This way, we ensured that the comparisons were based on the actual improvements made by the model.

In the noisy spectrogram of the vowel (Fig. 4), we can notice a "haze" or background noise across the frequencies. And this makes it hard to identify the exact beginning and end of each harmonic. But after enhancement, the "haze" is nearly eliminated. Now the harmonic lines can be identified clearly, and the formant shapes are more accessible. Aurally, The vowel sounds the same, but the noise that was covering it up is significantly reduced.

The improvement is especially obvious for the word sample owing to the additional complexities of speech dynamics. The word spectrogram (Fig. 5) shows similar behavior, but in a more dynamic context. In words, there are both vowels and consonants, so the level of fluctuation is higher in this example. We can observe the same sort of "haze" in the noisy version. In contrast, the enhanced version has more crispness and coherence of voiced segments, but there is still noise in the form of consonant segments, such as stops and fricatives. This indicates, this model is not capable of entirely removing noise-like events. Rather, it is simply removing any noise elements that impede the speech while appropriately leaving the details of "natural" speech unaffected.

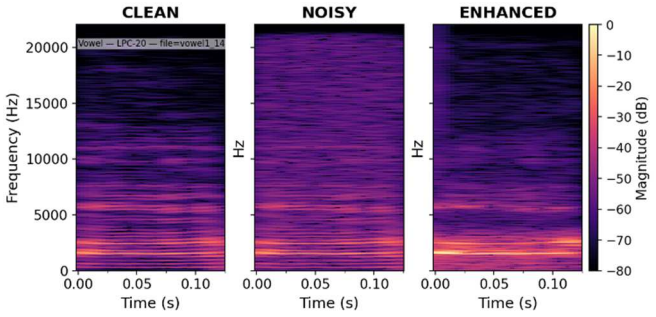

Fig. 4. Vowel spectrograms—Clean, Noisy, and Enhanced (shared STFT/dB scale). Enhanced lowers the floor and sharpens harmonic structure while preserving formants.

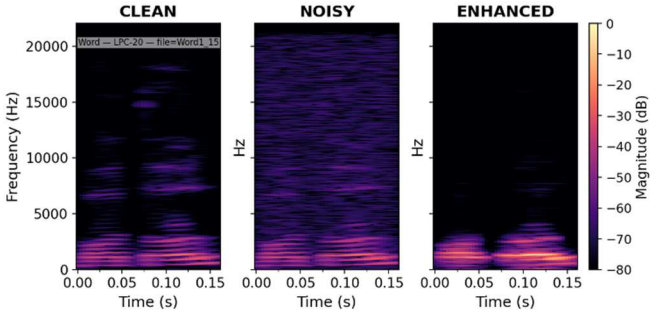

Fig. 5. Word spectrograms—Clean, Noisy, and Enhanced (shared STFT/dB scale). Enhanced suppresses background in voiced regions and preserves aperiodic consonants without artifacts.

E. Overall Performance Comparison and Discussion

To understand the overall impact of this enhancement model reviewed both the SNR results in Tables I–II and the spectrogram examples shown in Figures 2–5. The model steadily elevated the SNR in the enhanced signal compared to the noisy signal for all vowels and word data.

The greatest enhancement we got was for the vowel when we used LPC order 16. Vowels are likely to exhibit steady, smooth harmonic patterns, and an LPC order 16 is effective at detecting this structure accurately without adding instability. On the contrary, word recording demonstrated the best improvement with LPC order 20. Words contain sound transitions and extended voice segments, so the model gets more aspects to work with if a slightly higher LPC order is used. Hence, we can say that the complexity of the signal suggests the LPC order.

These results are supported by the spectrograms. From the residual domain, we can see that the enhanced signals have a cleaner harmonic coding and less background noise. The spectrogram waveforms indicate that the actual structure of the vowel or word, actual articulation, and presence of consonant bursts or fricatives remain unchanged. Importantly, the enhanced speech does not introduce any artificial tones or "musical noise" that accompany aggressive denoising approaches. It simply just sounds the same but clearer.

So these observations suggest that the method generalizes well across different vowel and word categories, while maintaining low computational cost. It preserves the original timbre and articulation, and it reduces noise by only enhancing the excitation portion of the signal, with the LPC envelope kept unchanged. Therefore, this method is suitable for a real-time or embedded speech enhancement system.

IV. CONCLUSION

In this study, we attempted to enhance the quality of speech by solely concentrating on improving the excitation residual, while leaving the LPC envelope unchanged. This approach combined LPC-based vocal-tract modeling with a lightweight neural network to improve perceived speech quality without modifying the speaker's voice. Experiments conducted on isolated Bengali vowels and words corrupted with additive white Gaussian noise demonstrated significant increases in spectral SNR and clearer harmonic structures. The best performance was observed at LPC order 16 for vowels and LPC order 20 for words, which reflected the greater temporal variability present in word-level speech. The improved speech retained a natural sound, without the presence of metallic artifacts or musical noise. These results indicate that the proposed residual-focused approach can effectively enhance speech clarity while preserving natural voice characteristics under controlled noise conditions. In the real world, noise will be more complex due to combinations of background noise, reverberation, and varying speaking styles, and the model will be evaluated under more realistic acoustic conditions in future work. Standard perceptual evaluation metrics such as PESQ and STOI will also be considered, along with extensions to continuous speech and adaptive LPC modeling. Overall, the method was simple and practical, making it suitable for real-time and embedded speech enhancement applications, while preserving the original voice characteristics.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2013.

[5] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[6] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[7] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.

[8] Y. Xu, J. Du, L. R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[9] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[10] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proceedings of Interspeech*, 2018, pp. 3229–3233.

[11] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proceedings of APSIPA ASC*, 2017, pp. 1–6.

[12] Y. G. Thimmaraja, B. G. Nagaraja, and H. S. Jayanna, "Speech enhancement and encoding by combining SS-VAD and LPC," *International Journal of Speech Technology*, vol. 24, no. 6, pp. 165–172, 2021.

[13] Y. Liu, N. Tang, X. Chu, Y. Yang, and J. Wang, "LPCSE: Neural speech enhancement through linear predictive coding," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, 2022, pp. 5335–5341.

[14] ITU-T Recommendation P.862, *Perceptual Evaluation of Speech Quality (PESQ)*, International Telecommunication Union, 2001.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[16] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice source estimation using harmonic plus noise models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 138–150, 2014.

[17] A. Paliwal, K. Wojcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

[18] R. Kumar, S. Singh, and S. R. M. Prasanna, "Excitation enhancement for robust speech processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1668–1680, 2020.

[19] M. M. Hasan and M. R. Hasan, "Isolated Bengali vowel and word speech sounds," Mendeley Data, V1, 2021. doi: 10.17632/2h6975kdsx.