International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

LSTM - Aided Speech Enhancement with Wiener Filter Adaptation

Y. Sravanthi Student of ECE,
RVR&JC CE,
Guntur,A.p India,
sravanthiyaganti2@gmail.com

P. Sruthi Student of
ECE, RVR&JC CE,
Guntur,A.P India,
sruthipittala2025@gmail.com

Y. Chaitanya Student of
ECE, RVR&JC CE,
Guntur,A.P India,
chaitanyayendeti9696@gmail.com

P. Srilakshmi Student
of ECE, RVR&JC CE,
Guntur,A.P India,
potlurisiddardha@gmail.com

Abstract— Speech enhancement plays a pivotal role in various applications, from improving the intelligibility of spoken communication in noisy environments. With the assistance of deep learning, a novel approach speech signal enhancement model is introduced in this research. The proposed LSTM model estimates the tuning factor of the Wiener filter with the aid of extracted features to obtain the de-noised speech signal. This model is structured into two phases: Training and Testing. During the training phase, Non-negative Matrix Factorization (NMF) is employed to estimate both the noise and signal spectrum from the noisy input signal. Subsequently, Empirical Mean Decomposition (EMD) features are extracted from the Wiener filter and a de-noised speech signal is obtained via processing. Additionally, bark frequency information is evaluated. In the testing phase, the LSTM model has been trained by the extracted features (EMD) via a modified wiener filter. The combination of LSTM-based temporal modeling with trained features and the adaptive Wiener filter results in significantly improved speech quality and intelligibility.

Keywords— *Speech Enhancement, Non-negative Matrix Factorization, Empirical Mode Decomposition, Wiener Filter.*

I. Introduction

Speech is typically distorted in real-world environments by both room resonances and background noises. Speech enhancement aims to improve speech quality by using various algorithms. The objective of enhancement is improvement in intelligibility and/or overall perceptual quality of degraded speech signal using audio signal processing techniques.

Enhancing of speech degraded by noise, or noise reduction, is the most important field of speech enhancement. This is significant because speech signals are often degraded by various environmental factors such as background noise, echoes and distortions during transmission or recording. Noise profoundly affects the quality and intelligibility of speech signals in multiple ways. Firstly, it can mask crucial speech features, obscuring them amidst the background noise and hindering listeners' ability to discern speech from noise. Additionally, noise interference during transmission or recording introduces distortions and artifacts, compromising the clarity and fidelity of the speech signal. Moreover, noise diminishes the signal-to-noise ratio (SNR), making it challenging to extract the desired speech from the background noise. This reduction in SNR, coupled with noise-induced distortions, further exacerbates the difficulty in understanding the speech. Speech enhancement techniques play a pivotal role in mitigating the adverse effects of noise on speech signals. The major intention behind the speech enhancement is to suppress the noise and to boost up the SNR of noisy speech signals in challenging environments. By employing methods like noise reduction, adaptive filtering, and spectral subtraction, these techniques effectively suppress background noise while preserving the integrity and intelligibility of the speech signal. Through such enhancements, speech communication systems in applications spanning telecommunications, voice-controlled systems, hearing aids, and automatic speech recognition (ASR) can achieve improved performance and usability, even in noisy environments. Speech enhancement techniques have been studied for several decades with a variety of promising applications, such as telecommunications and hearing aid systems, to mitigate the harmful effects of background noise and interference. Acoustically added background noise to speech can degrade the performance of digital voice processors

The key aim of these speech enhancement methods is to enhance the Speech SNR. Techniques have been introduced regarding boosting up the speech quality and compacting speech bandwidth by suppressing the additive background noise. Recently, Generative Adversarial Network (GAN) based speech enhancement was utilized to overcome the traditional difficulties. Speech Enhancement GAN (SEGAN), conditional GAN (cGAN), Wasserstein GAN (WGAN) and Relativistic Standard GAN (RSGAN) techniques were introduced. Despite the success of GAN-based speech improvement techniques, two major difficulties were present : Training instability and a lack of consideration for varied speech characteristics.

Thus, to overcome the existing issues, an LSTM with trained speech features and an adaptive Wiener filter is introduced in this work. For decomposing the speech spectral signal, a modified wiener filter is introduced. addition, the LSTM model is introduced to properly estimate the tuning factor of the Wiener filter for all input signals. In a testing phase, the LSTM model has been trained by the extracted features (EMD) via a modified wiener filter.

II. METHODOLOGY

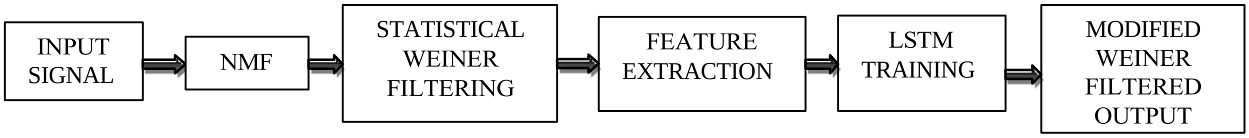

Fig 1: Flowchart of the proposed methodology

A. Data Acquisition and Preprocessing

Gathered a dataset of noisy speech recordings from Kaggle. It is a dependable source for online datasets, facilitating research and analyses across various domains. Noisy recordings include various types of environmental noise, such as traffic noise, babble noise, or background music. It's important to have a diverse range of noise types and signal-to-noise ratios (SNRs) to train a robust model. In the training phase, the noisy signal W(t) is incorporated into the clear speech signal S(t). The formulated noisy speech signal is shown in Eq. (1)

$$R(t) = S(t)+W(t) \qquad (1)$$

Then, for this R(t), the NMF estimates the noise spectrum $Spe_N(n)$ and signal spectrum $Spe_S(n)$ respectively. The obtained spectrum (noise and signal) is given as input to the statistical wiener filter, from which the filtered signal F(n) is generated. The Wiener filter's purpose is to determine a statistical estimate of an unknown signal by taking a similar signal as an input and filtering it to create the estimate as an output. Thus, the approximation is done by reducing the MSE between the target signal and the noise distorted signal. $Spec_S$ has a power spectral density of $pdf_S(\omega)$, while $Spec_N$ has a power spectral density of $pdf_W(\omega)$. The filter transfer function shown in Eq. (2). Moreover, Eq. (3) shows the statistical formula for SNR, and Eq. (4) shows how the SNR formula can be used in the filter conversion function where $G_W(\omega)$ represents the approximate signal magnitude range.

$$F(\omega) = \frac{pdf_S(\omega)}{pdf_S(\omega)+pd\ _W(\omega)} \qquad (2)$$

$$SNR = \frac{pdf_S(\omega)}{G_W(\omega)} \qquad (3)$$

$$F(\omega) = 1 + \left[\frac{1}{SNR}\right]^{-1} \qquad (4)$$

B. Feature Extraction

From these filtered signals, the features like EMD, bark frequency are extracted.

1. Empirical Mode Decomposition

EMD is a signal processing technique that aims to decompose a signal into a finite number of oscillatory components that are adaptively derived from the data. The EMD features are extracted from F(n). Huang proposed EMD as an adaptive strategy in which a limited number of Intrinsic Mode Functions (IMF) were applied to reflect complex data. IMFs ye(n) and residue q(n) are decomposed from the data set F(n).

Intrinsic Mode Functions (IMFs) are fundamental building blocks obtained through a data analysis technique called Empirical Mode Decomposition (EMD). They are essential components in understanding the time-frequency characteristics of non-stationary and nonlinear signals, particularly in fields like signal processing, audio analysis, and vibration analysis.

Key characteristics of IMFs:

Number: EMD decomposes a signal into a set of IMFs and a residual component. The number of IMFs depends on the data's complexity, but it typically ranges from a few to several dozen.

Frequency: Each IMF represents a single oscillation mode, meaning it has a specific frequency and amplitude that may vary with time. This allows for capturing the dynamic nature of non-stationary signals.

Local extrema: An IMF must satisfy two criteria: The number of extrema (local maxima and minima) and the number of zero-crossings must be either equal or differ by at most one.

The mean value of the envelope defined by the maxima and the envelope defined by the minima must be zero at any point.

Interpretation: IMFs can be interpreted as individual components that contribute to the overall signal. For example, in speech signals, IMFs might represent the fundamental voice frequency, formants, and noise components.

2. Bark Frequency

Bark is a unit on the Bark scale. The Bark scale is a psychoacoustic scale, it relates frequencies to how we perceive them rather than their objective measurement in Hertz (Hz). It does this by dividing the audible spectrum into 24 critical bands that represent how our ears process sound.

By applying the Bark scale to the resulting IMFs from EMD, you create a representation that aligns with human hearing. Calculate the Bark scale representation of each IMF.

The frequency conversion from Hz to the Bark scale uses the following formula shown in Eq. (5):

$$\text{bark} = \frac{(26.81)(hz)}{1960+hz} - 0.53 \qquad (5)$$

if : bark < 2 , bark=bark+(0.15)(2−bark)

if: bark > 20.1→bark=bark+(0.22)(bark−20.1)

C. LSTM Training

For speech enhancement, the extracted features i.e. IMFs and bark frequency are subjected to LSTM. The architecture of an LSTM model consists of multiple LSTM layers. The output of one layer becomes the input for the other. Each LSTM cell is made up of three multiplicative units, which represent the “forget gate, input gate, and output gate”.

Forget Gate: Decides which information to discard from the cell state.

Input Gate: Determines which new information to store in the cell state.

Cell State: Represents the memory of the cell, which can be updated and passed along the sequence.

Output Gate: Determines the output of the cell based on the current input and the cell state.

These memory cells allow the network to remember important information over longer sequences, preventing the vanishing problem. The Formulae for parameters in the architecture of LSTM:

$$I_t = \sigma\,(J_I X_t + K_I M_{t-1} + B_I) \qquad (6)$$

$$F_t = \sigma(J_F X_t + K_F M_{t-1} + B_F) \qquad (7)$$

$$O_t = \sigma(J_0 X_t + K_0 M_{t-1} + B_0) \qquad (8)$$

$$C_t = F_t C_{t-1} + I_t G_t \qquad (9)$$

$$G_t = \tanh(J_G X_t + K_G M_{t-1} + B_G) \qquad (10)$$

$$M_t = O_t \tanh(C_t) \qquad (11)$$

Here, I_t , F_t *and* O_t are the input, forget, and output gates at a time t. The weight matrices that map the hidden layer output to gates are denoted by K_I, K_F *and* K_0. The weights which map the hidden layer input to the input, forget as well as output gates are represented as J_I , J_F *and* J_0. B_I, B_F, B_0 *and* B_G are the bias vectors. The sigmoid function σ is used to represent the gate activation function. Furthermore, the cell outcome and layer outcome is denoted by G_t, M_t respectively

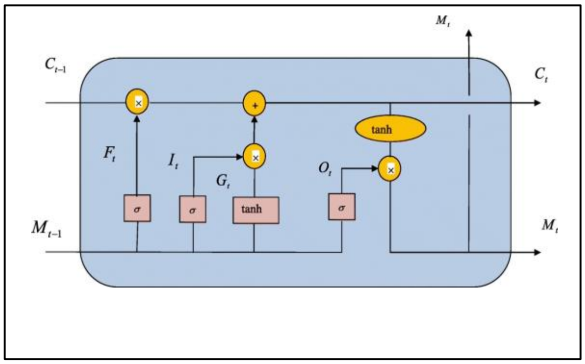

Fig 2: The architecture of LSTM

For tuning η in a much precise manner, modified wiener filter is introduced, which overcomes the drawbacks of the existing wiener filter. The existing wiener filter couldn’t estimate the power spectra efficiently; it is challenging for the existing wiener filtering to acquire the perfect restoration for the random nature of the noise. The estimated tuning ratio of the new modified Wiener filter is fine-tuned by LSTM. The properly estimated tuning ratio η_{tuned} acquired from LSTM is fed as input to the modified Wiener filter. The appropriate tuning factor for diverse noises is identified and with this, the LSTM is trained. The difference between the predicted and actual values is measured using a loss function. It is then used to update the model's weights, allowing it to learn from its mistakes and improve its predictions. Training an LSTM model involves feeding batches of data through the network, calculating the loss, and updating the weights through back propagation. LSTM based speech enhancement is again trained and implemented with learned loss function.

III. COMPUTATIONAL RESULTS

The following figure is input speech signal :

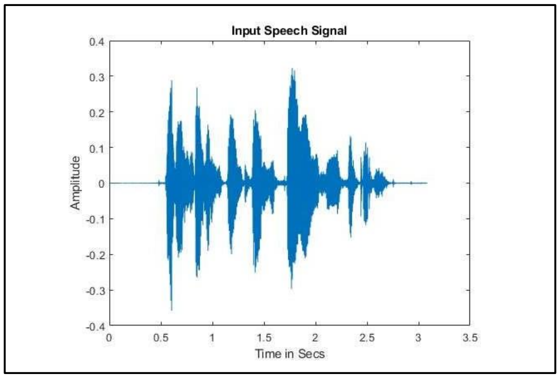

Fig 3: Input Speech Signal

Next, the noise is added to speech signal at differing SNR levels (-10dB, -5dB, 0dB, 5 dB and 10 dB respectively) to measure the efficacy of the suggested work for speech enhancement.

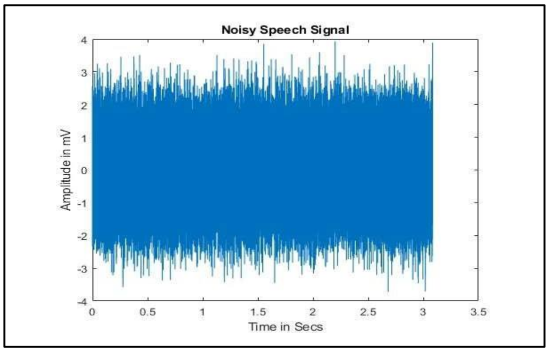

Fig 4: Noisy Speech Signal

Intrinsic mode functions (IMFs) are obtained from the wiener filtered output signal. Each IMF represents a specific frequency or mode within the signal.

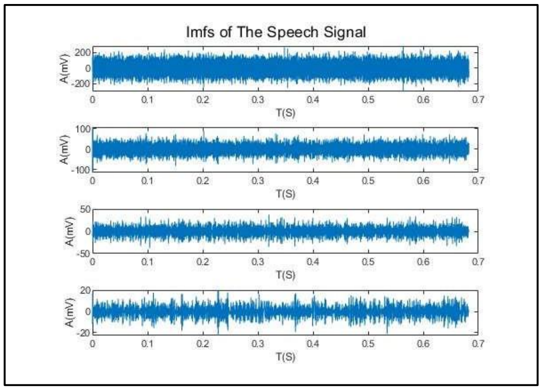

Fig 5: IMFs of the speech signal

The enhanced output signal is obtained after the training the LSTM model with IMFs and bark frequency. Again LSTM is trained with the learned loss function.

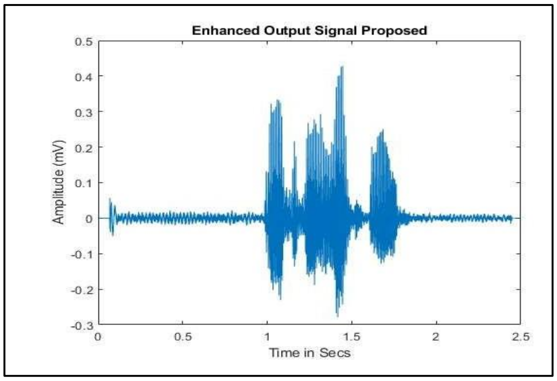

Fig 6: Output Signal

After the implementation of LSTM based speech enhancement with learned loss function, we can finally conclude that, the implementation improved the PESQ score and enhancement results. The implementation is also simple and increases enhancement at a deeper levels. The proposed implementation even produced better PESQ an SNR values compared with existing method.

TABLE I OUTPUT PARAMETERS

PARAMETERS	Existing Method	Proposed Method
Signal-to-Noise-Ratio(SNR)	13.8 dB	14.8 dB
Perceptual Evaluation Of Speech Quality (PESQ)	2.8845	4.1537

IV. CONCLUSION

Most studies have shown that reducing signal noise without distorting speech is a difficult challenge, which is one of the main reasons why perfect enhancement systems aren't available. The major goal of this work is to enhance the speech signals with various noise sources. In this paper, we focus on the issues such as not suitable for complex noise conditions, lower SNR, lower PESQ. Compared to the existing models, the proposed work introduces a wiener filter-assisted deep learning LSTM model. The LSTM model estimates the tuning factor of the Wiener filter with the aid of extracted features to obtain the de-noised speech signal. After the implementation of LSTM based speech enhancement with learned loss function, the implementation improved the PESQ and SNR values even better as compared with existing method. It can be used in complex noisy environments as we have taken noise level ranging from
-10 dB to 10dB. The main potential applications of the proposed model are given below: Hearing aids Automatic speech recognition , Mobile communications, video captioning for teleconferences ,Voice over Internet protocol , Hand-free communications. This method provides better outcomes and it suits many potential application fields.

REFERENCES

[1]. Garg A (2020) Enhancement of speech signal using diminished empirical mean curve decomposition- based adaptive wiener filtering. in comm

[2]. E. Darren Ellis Department of Computer and Electrical Engineering – University of Tennessee, Knoxville Tennessee 37996 topic on “Design of a Speaker Recognition Code using MATLAB”

[3]. Topic on "Controlling of Device through Voice Recognition Using Matlab1" ISSN NO: 2250-3536 VOLUME 2, ISSUE

[4]. Topic on "Extraction of Pitch and Formants and its Analysis to identify 3 different emotional states of a person" ijcsi.org/papers/IJCSI-9-4-1-296-299.pdf

[5]. Topic on "Speech Recognition using Digital Signal Processing" ISSN: 2277-9477, Volume 2, Issue 6.

[6]. Jolicoeur-Martineau A (2018) The relativistic discriminator: a key element missing from standard GAN. arXiv preprint arXiv:1807.00734

[7]. Arul VH, Sivakumar VG, Marimuthu R, Chakraborty B (2019) An approach for speech enhancement using deep convolutional neural networks. Multimedia Res 2(1):37–44

[8]. Cuiv X, Chen Z, Yin F (2021) Multi-objective based multi-channel speech enhancement with BiLSTM network. Appl Acoust.

[9]. Chai L, Du J, Liu Q-F, Lee C-H (2021) A cross- entropy-guided measure (CEGM) for assessing speech recognition performance and optimizing DNN-based speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 29:106–117