JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

Research and DSP Implementation of Speech Enhancement Technology Based on Dynamic Mixed Features and Adaptive Mask

[1]LAXMAN BAVANDLAPALLY ,M.Tech Assistant Professor, lax406@gmail.com
2Dr.THOTA SRAVANTI, M.Tech , Ph.D, Associate Professor, sravanti815@gmail.com
Department-ECE
Pallavi Engineering College Hyderabad, Telangana 501505.

ABSTRACT

In this research, we present a deep learning voice enhancement method based on dynamic hybrid features, adaptive masks, and DSP implementation to address the issue of feature loss and boost the performance of speech improvement. E dynamic characteristics consist of the log Mel power spectrum, Mel campestral coecients, and Multiresolution Auditory Cepstral. Coecients (MRACC), and extract derivatives to fully describe speech transient information. And minimize alterations to its nonlinear nature.

Introduction

The use of speech signal processing technology is growing in popularity as artificial intelligence technology advances, and voice enhancement is one of its many applications. A central focus of study is the importance placed on this element. Improvements to speech may be made using signal extraction techniques. Background noise, lessen interference, and keep that is resistant to distortion and has AI-related applications. Voice recognition, hearing aids, and other areas [1]. These days, speech amplification techniques may be broken down into unattended and supervised are the two options. Unsupervised Assumptions like smooth noise and uncorrelated speech are often used in voice improvement. Noise, resulting in insufficient capability to dampen down nonsmooth caused by the same source that distorts speech and creates noise. Representative Wiener filtering and spectral subtraction are two examples of algorithms. Noise is reduced using supervised speech enhancement ([2]) by studying the signal's statistical characteristics, which has clear benefits in situations with a poor signal-to-noise ratio, and noise that is not smooth, and may be classified in two ways: addressing both simple and complex versions of models. Models of the superficial layer often include this model combines Hidden Markov Analysis with Simple Neural Networks. Because of the high number of layers and other factors, learning ability and performance are constrained. Little number of nodes per layer, and the information is only needed for In addition, the training is minimal. One way in which deep-learning models may acquire knowledge analysis of the complex nonlinear interactions between languages, which significantly helps them function better in situations when they don't know what to expect place(s) with a lot of background noise [3]. About three distinct categories may be identified.

Improved Deep Neural Network Model

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

Recurrent Gated Unit (GRU). In Figure 1, we see a simplified representation of a single unit of a gated recurrent neural network, where in, bn, and bn1 represent the input, output, and output of the current and prior time periods, respectively. Respectively. Reset Gate (rn), Update Gate (kn), and Backup Gate (bn) and potential secret state candidates. (E neural network composed of gated recurrent units (GRUs) helps reduce network over fitting issue because to gating, to a lesser amount

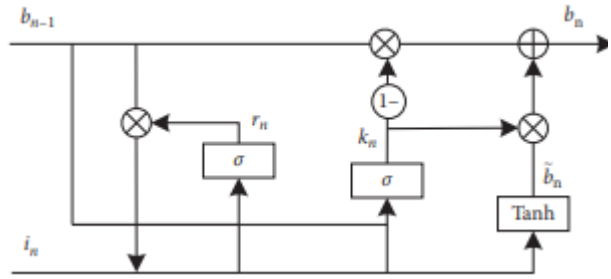


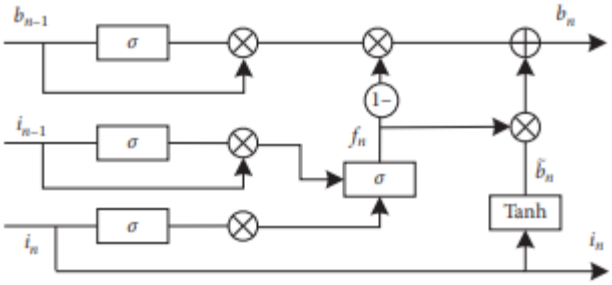
Figure 1: Gated recurrent unit

Network's method for understanding more extensive temporal connections. The neural network with gated recurrent units (GRUs), which helps with both short-term and long-term memory (LSTM) network, the complexity of networks has increased Depending on the network's topology, the amount of parameters, and other metrics. As demonstrated in Figure 1, compared to the three-gate network architecture of the long-term memory (LTM), The GRU neural network is limited to only a single kind of neuron. The RN gate must be adjusted and the KK gate must be updated. Unit update relation of the gated recurrent neural the following formula describes a network.

$$
\begin{aligned}
k_n &= \sigma\left(M_k i_n + P_k b_{n-1} + h_k\right), \\
r_n &= \sigma\left(M_r i_n + P_r b_{n-1} + h_r\right), \\
g_n &= \tanh\left(M_g P_g \left(b_n - 1 \odot r_n\right) + h_g\right), \\
b_n &= \left(1 - k_n\right) \odot b_{b-1} + k_n \odot g_n.
\end{aligned} \qquad (1)
$$

1) The M and P matrices in formula (1) are the weight matrices. The biased word is h. They are a set of variables that may be altered by instruction. The Hadamard Symbol is Sigmoid activation function is the product of

Figure 2 illustrates how the current input I determine the CGRU neural network's output bn. in1 from the time before, and the output bn1 from past time together with the present input instant (is fully exploits the voice signal capabilities of previous picture. The inputs in, bn, in1, and bn1 are numbers, and the outputs are current output, as well as input and output of preceding instantaneous time, if you will. It was the argument that sparked the idea for the slogan.



In this research, we first use the in-and-out-of-focus attention mechanism and the gated linear unit (GLU) to calculate the weighted feature vectors of in and in1. Unit input of the CGRU neural network as and bn1 following

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

$$\begin{aligned}\hat{i}_n &= \sigma(M_i i_n) \odot i_n,\\ \hat{i}_{n-1} &= \sigma(M_{i-1} i_{n-1}) \odot i_{n-1},\\ \hat{b}_{n-1} &= \sigma(M_{b-1} b_{n-1}) \odot b_{n-1}.\end{aligned} \quad (2)$$

(2) Input in1 and output in from formula (2) are then utilized to find the function FN that represents the forgetting gate, which can be written as what follows

$$f_n = \sigma\left(M_n \hat{i}_n + M_{n-1} \hat{i}_{n-1} + h_f\right). \quad (3)$$

3) Unlike GRU, CGRU's candidate concealing status is solely based on the current input.

$$\tilde{b}_n = \tanh\left(M_b i_n + h_b\right). \quad (4)$$

Candidates for the hidden state bn and long-term memory play a role in the present network unit's output bn. the bn1 band-weighted characteristic of the prior gate fn current output, as in the following

$$b_n = f_n \odot \tilde{b}_n + (1 - f_n) \odot \hat{b}_{n-1}, \quad (5)$$

The Hadamard product is denoted by the expression 1, (5). Sigmoid activation function is denoted by. One forgetting gate FN is utilized across the network rather than several of them in an effort to simplify its layout. CGRU system. Meanwhile, in order to solve the issue of As a result of the speech enhancement's compression of feature data from an input voice stream In order to improve the quality of the author's voice, this study takes extensive use of characteristics of preceding frames' voice signals, input in 1 of the past combined with the present moment's input concurrently, the gated linear unit moment Using the Gluon Learning Unit (GLU) mechanism to regulate feature information transmission boosts the network.

Speech enhancement Algorithm in this paper

Dynamic Features, Section 3.1 Figure 3 illustrates how various linguistic elements are indicative of distinct aspects of the underlying audio stream. After a Mel-frequency analysis, the spectrum may be smoothed using LMPS. Bank of filters and nullifies its impact. Emphasizing the reverberant high points of the voice. A noisy speech power spectrum has several aspects, and MFCC shows how they all relate to one another. This new and enhanced Four cochleagram sparse representations of varying resolutions make up the MRCG feature, which may be used to represent global other regional details. To accurately portray the talk, nonlinear framework, these three characteristics are integrated and complimentary in order to get a rather full static feature. Derivatives of the first and second orders are then calculated for the grafted features to fully characterize short-term speech data (e., differential characteristics Identify the link between consecutive speech frames and you shouldn't depend just on the network to collect temporally linguistic data on possible future speeches. (E fusion of movement and static characteristics help fix the insufficiency and incompleteness of existing characteristics in speech structure representation, leading higher quality recreated speech with less distortion intelligibility.

$$W(x,w) = \left[W_{\mathrm{LMPS}}(x,w); W_{\mathrm{MFCC}}(x,w); W_{\mathrm{MRACC}}(x,w)\right], \quad (6)$$

Where x is the frame count and w is the index of the feature dimension. The LMPS, MFCC, and MRACC

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

are respectively denoted as the characteristics of MRACC, in that order.

$$\Delta W(x,w) = \frac{\sum_{z=1}^{2} z\,(W(x+z,w) - W(x-z,w))}{\left(2\sum_{z=1}^{2} z^2\right)^{1/2}},$$

$$\Delta(\Delta W(x,w)) = \frac{\sum_{z=1}^{2} z\,(\Delta W(x+z,w) - \Delta W(x-z,w))}{\left(2\sum_{z=1}^{2} z^2\right)^{1/2}}, \quad (7$$

The first two frames of the current frame are represented by the index z.

$$\Omega(x,w) = [W(x,w); \Delta W(x,w); \Delta(\Delta W(x,w))]. \quad (8)$$

Modifiable Soft Mask. The effectiveness of a speech-enhancement system based on a deep neural network is proportional to the performance of the learning goal. Speech amplification, the level of which is correlated to the amount of distortion and the quantity of remaining background noise in the improved pronunciation. Implementing IRM is one of several educational goals. regarded as the primary training subject for speech improvement efficient; it's worth is determined by how it compares to unadulterated speech activity and commotion energy at each discrete interval of time and frequency, may significantly boost the clarity of improved voice, reduce the ambient noise. The usage of IRM is a caveat, however, to eliminate background noise at various signal-to-background noise ratios in the same way and can't be done mechanically modified based on signal-to-noise ratio data, issue of omitting necessary parts of speech when The retention of noise components is common.

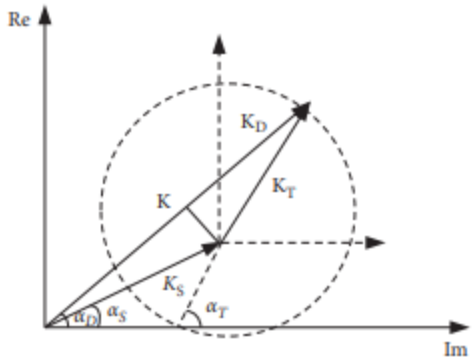


Figure-4 Phase geometric Relationship Diagrammed

The phases' geometric connection is seen in Figure 4. The letter KD indicates a person's voice is very distracting. The letter KS represents truthful communication. The KT symbol represents the volume of garbled words. In this case, D, S, and T are the Figure 4 illustrates the transitions between the three speech noise levels (noisy, pure, and silent).

$$K_S^2 = K_D^2 + K_T^2 - 2K_D K_T \cos(\alpha_T - \alpha_D). \quad (9)$$

The given formula of allows for the extraction of phase difference information from noisy speech and noisy speech. Ratios between a priori S/N () and post hoc S/N (c)

$$\cos(\alpha_{TD}) = \cos(\alpha_T - \alpha_D) = \frac{K_D^2 + K_T^2 - K_S^2}{2K_D K_T} = \frac{\gamma + 1 - \xi}{2\sqrt{\gamma}}. \quad (10)$$

The figure's geometric connection allows us to draw the following conclusion:

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

$$\cos(\alpha_T - \alpha_D) = \frac{(K_D - K)}{K_T}, \qquad \cos(\alpha_{DS}) = \cos(\alpha_D - \alpha_S) = \frac{K}{K_S}. \tag{11}$$

So, knowing the phase difference between it's possible to distinguish between clear and distorted speech.

$$\begin{aligned}\cos(\alpha_{DS}) &= \frac{K}{K_S} \\ &= \frac{K_D - K_T\cos(\alpha_T - \alpha_D)}{K_S} \\ &= \frac{K_D/K_T - \cos(\alpha_T - \alpha_D)}{K_S/K_T} \\ &= \frac{\sqrt{\gamma} - \gamma + 1 - \xi/2\sqrt{\gamma}}{\sqrt{\xi}} \\ &= \frac{\gamma + \xi - 1}{2\sqrt{\gamma\xi}}.\end{aligned} \tag{12}$$

Cos(DS) or cos(TD) must be positive for Speech recognition, the minimum threshold is set at 0. Once the phase difference data was included, To extract the new mask R from the time-frequency mask, we write

$$\overline{R} = \frac{S^2(n,f)\max(\cos(\alpha_{DS}(n,f)),0)}{S^2(n,f)\max(\cos(\alpha_{DS}(n,f)),0) + T^2(n,f)\max(\cos}$$

$$R(n,f) = \alpha\overline{R}(n,f) + (1-\alpha)\sqrt{\overline{R}(n,f)}. \tag{14}$$

Experiments show that the optimal effect is achieved at a value of 0.7, indicating a value of 0.7 for the ratio mask R. combines the speech phase information with the benefits of variation

Speech Enhancement Hardware and Software Design

Implementing a Digital Signal Processor-Based Speech Enhancement System Hardware Design. These peripheral communication connections are available on the TI Production TMS320F281x series DSP with other features. Analogue to digital converter (ADC), serial peripheral interface (SPI), serial communication interface Serial Peripheral Interface (SPI) and Serial Buffered I/O (SBIOS) (McBSP). McBSP is able to offer full-duplex because it uses a double-buffering method for communications and a registers for sending and receiving data that are triple-buffer transmitting data in a constant stream. (E length of data is designed to establish a Serial Link with Commercial Devices analogue interface chips with standard decoders (CODECs) (AIC), to continue with the same pattern. The AIC23 is a sigma-delta high-efficiency analog-to-digital and digital-to-analog audio converter chip internal converters for direct interface with digital signal processors McBSP. The DSP allows the sampling rate to be set to successfully receive and send voice signals at quick and swift. Concurrently, DSP's lightning-fast processing power benefits such as frame processing capabilities, adaptability, low power consumption, and more, it has become the de facto standard in recent years. favoured method of digital voice processing. So, the purpose of this article is to: choose the TMS320F2812 microcontroller as the heart of the system the hardware design of the system is finished with the

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

addition of AIC23 and the related peripheral circuit. To counteract this, the DSP is using a little amount of storage for both programmers and data not easy to accommodate speech processing requirements. Off-chip data storage that can hold an additional 256 kilobytes of 16-bit SRAM storage space, and 512 KB of 16-bit FLASH for auxiliary software data storage, as seen in the block diagram of the system's architecture in Diagram 6.

In order to clean up the distorted audio signal, the microphone's input is routed via the A/D converter and anti-alias filter in the AIC23 before being sent on to the DSP chip. A system for reducing ambient noise. Meanwhile, AIC23 receives the processed data through McBSP for D/A. a filtering system that converts and rearranges data. Usual AIC23 have a circuit for a headphone driver built in, therefore they're not going to require for processing by the vehicle's external driver, but the voice signal after the results of the noise reduction operation are sent directly into the headphone.

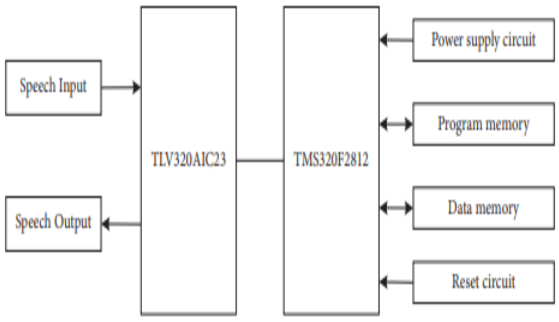


Figure 6: DSP speech enhancement system hardware structure design.

Software Architecture for a Digital Signal Processor-Based Voice Enhancement Algorithm (e algorithm's performance is confirmed by developing a speech noise reduction algorithm based on skip using MATLAB's linear and nonlinear decoupling functions is rebuilt from scratch using C and assembly language in CCS downloaded IDE and integrated development environment DSP for real-time simulation bug testing. System software Figure 7 depicts the implementation process flow.

Experimental Results and Analysis

Details of an Experimental Procedure and Related Data. For the purpose of testing the efficacy of the suggested strategy, Pure Speech takes 2000 samples of audio from the TIMIT Speech Dataset was used as the training set for this study. Put 500 pieces of audio data into a random set are used for testing purposes. An example of training set noise is ambient noise from a database of 100 species was used to choose the

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

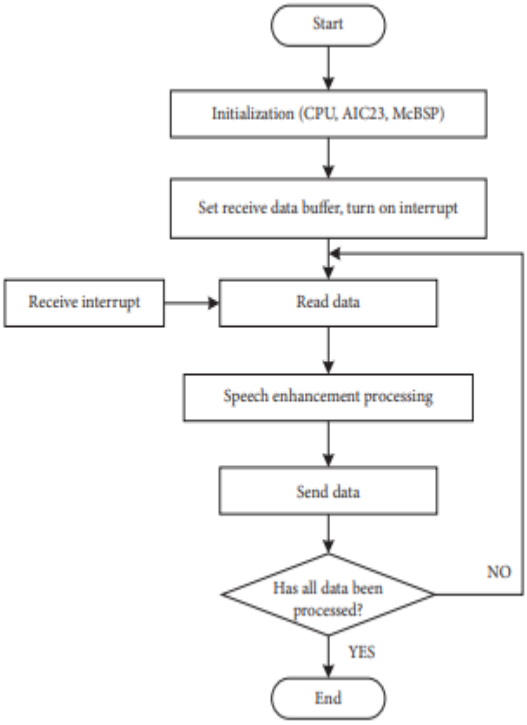


Figure 7: System main program flow chart.

References in scholarly works, with NOISE-92's 15-species noise serving as the test set's background noise. A Random Sample of 2000 fragments of TIMIT's training set audio, plus a control set of 100 three signal-to-noise ratios were generated from a randomly mixed set of sounds. Of 5, 0, and 5 to provide an 8,000-song training dataset. 500 randomly chosen voice data strips test data from the TIMIT database plus 15 samples from the NOISE-92 noise database collection are shuffled with three signal-to-noise ratios using the values (5), (0.0), and (5.0) as ratios, we can construct a dataset of 2000 strips to test. Of distorted voice recordings. When doing feature extraction, the default value for the sampling frequency of both pure speech and background noise is Frames are 256 in length (or around 31 ms) and the frequency is 8000 hertz. A 128-frame-per-second frame shift.

An Examination of the Mode Tables 1 and 2 reveal that the simple recurrent neural network (SRNN) has the lowest Perceptual Evaluation of Speech Quality (PESQ) and mean Short-Term Objective Intelligibility of Speech (STOI) Comparatively improved speech, with a streamlined in both the standard recurrent neural network (SRU) and the gated recurrent neural network's general-receiving-unit (GRU) may provide acceptable speech-enhancement results. This long-term knowledge cannot be learned by a basic recurrent neural network, which dependence. The GRU/SRU e-gating method notably improves the network's capacity for learning. The causal speech network is the most improve connection In comparison to conventional network architectures; the CGRU described in this research achieves better results. Superiority in understanding what is being said in a short amount of time. In Moreover, the CGRU network's cell structure follows the structure of gated recurrent neural networks in Priority must be given to the most advanced input features if they are to be used effectively. Information. The computation of the current's output feature network, CGRU does not only factor in input at the node level. Now with the output bn1 from the previous instant, but it also includes the input in1 from the previous this very time. (It capitalizes extensively on the existing N-frame structure. Figure 8: Data from the Speech Signal. In this study, we evaluate our proposed method to many others already in use in order to ensure its efficacy. Algorithms. The MRACC component and IRM, which are used in the first

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

algorithm, have the most influence, and are most often seen among the three species. Prepare the neural network. Algorithm 2 is a combined LMPS, and it goes like this:

Table 1: Average Perceptual Evaluation of Speech Quality (PESQ)

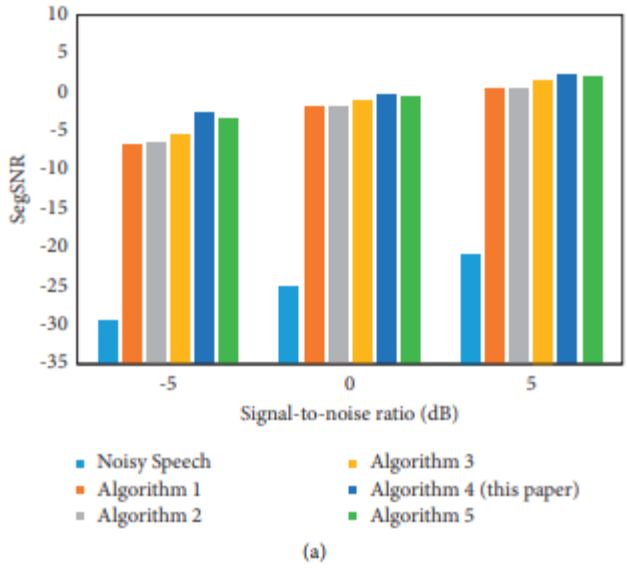
Noise	Signal-to-noise ratio (dB)	SRNN	SRU	GRU	CGF
factory2	−5	2.346	2.515	2.426	2.49
	0	2.651	2.806	2.729	2.84
	5	2.877	3.102	3.065	3.12
buccaneer1	−5	1.861	1.952	1.763	1.95
	0	2.240	2.333	2.201	2.34
	5	2.533	2.654	2.587	2.66
destroyerengine	−5	1.903	2.030	2.106	2.22
	0	2.212	2.296	2.240	2.32
	5	2.521	2.578	2.526	2.60
hfchannel	−5	1.801	1.952	1.706	1.83
	0	2.157	2.177	2.111	2.28
	5	2.469	2.513	2.502	2.59

Table 2: Average Short-Term Objective Intelligibility of Speech (STOI).

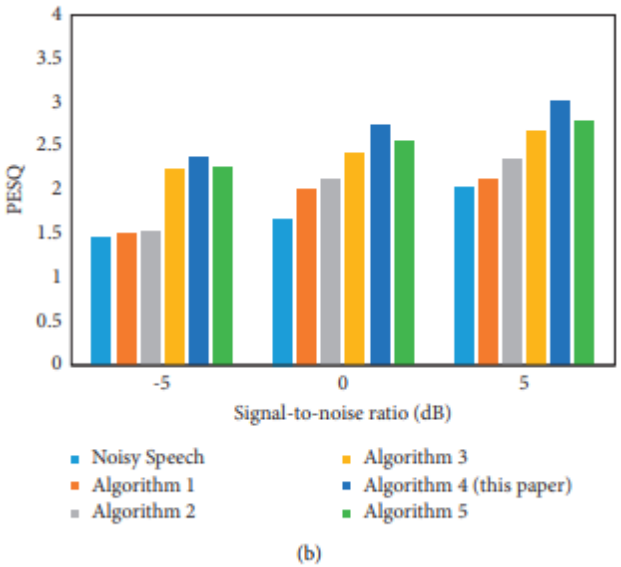
Noise	Signal-to-noise ratio (dB)	SRNN	SRU	GRU	CGR
factory2	−5	0.773	0.789	0.786	0.80
	0	0.852	0.868	0.870	0.88
	5	0.901	0.916	0.912	0.92
buccaneer1	−5	0.623	0.625	0.584	0.63
	0	0.745	0.748	0.720	0.76
	5	0.833	0.836	0.826	0.85
destroyerengine	−5	0.624	0.616	0.595	0.63
	0	0.749	0.752	0.734	0.76
	5	0.841	0.850	0.841	0.86
hfchannel	−5	0.655	0.655	0.645	0.67
	0	0.768	0.774	0.773	0.79
	5	0.847	0.862	0.861	0.87

The neural network was trained using MFCC, MRACC, and IRM. In the third algorithm, the neural network is trained using a combination of the dynamic features and the adaptive mask r. In Algorithm 4: A Technique for Improving Speech

Based on the in this research, a CGRU depth model was trained using a combination of static features, dynamic features, and an adaptive mask. The 5th Algorithm consists of: networked voice processing that works from beginning to finish; emphasizes minimal channel load.



(a)



(b)

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

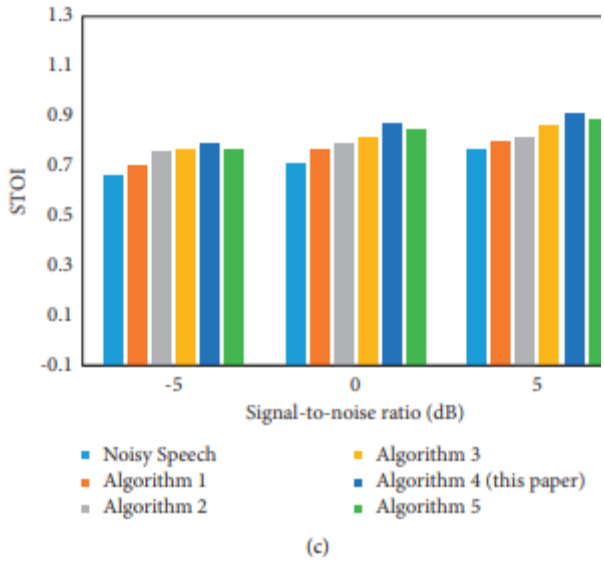


Figure 8: Comparison of three results of SegSNR, PESQ, and STOI under different algorithms in white-noise environment. (a) SegSNR; (b) PESQ; (c) STOI.

Conclusion

This article proposes a neural network-based approach for enhancing voice, including dynamic features and a combined adaptive mask optimization. Specifically, in this article, we present the causal gated recurrent unit (CGRU) neural architecture, wherein network that addresses the issue of real-time speech transmission system of improvement (e) with dynamic functions and adaptive mask are sent into a deep neural network in order to Figure out the intricate mapping dynamics between distracting speech, and unadulterated speech under close supervision. The new features and mask more properly reflect the time-frequency masking value of each time-frequency unit, allowing the neural network to more accurately estimate the pure speech spectrum. In conclusion, the hardware is provided in this document. systemic use of the algorithm for enhancing voice The method has been shown to be effective in reducing the Enhanced speech distortion in a variety of noise environments and situations of varying signal-to-noise ratios, with vocal output remaining discernible superiority in terms of clarity and intelligibility enhanced performance to a higher standard. As a further step, we will one of the main focuses of current research is to increase the speech augmentation system's real-time performance in an effort to lower the It is planned to look at the complexity of networks and ways to speed up DSP voice enhancement algorithms.

References

[1] N. Das, S. Chakraborty, J. Chaki, N. Padhy, and N. Dey, "Fundamentals, present and future perspectives of speech enhancement," International Journal of Speech Technology, vol. 24, no. 4, pp. 883–901, 2021.

[2] R. Jaiswal and D. Romero, "Implicit wiener filtering for speech enhancement in non-stationary noise," in Proceedings Of :e 2021 11th International Conference On Information Science And Technology (icist), pp. 39–47, IEEE, Chengdu, China, May 2021.

[3] Y. Hu, Y. Liu, S. Lv et al., "DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement," 2020, https://arxiv.org/abs/2008.00264.

[4] Y. Wang, H. Jia, and H. Ji, "Feature joint optimization of deep belief network for speech enhancement," Computer Engineering and Applications, vol. 55, no. 9, pp. 38–42, 2019.

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

[5] Z. Xu, S. Elshamy, and T. Fingscheidt, "Using Separate Losses for Speech and Noise in Mask-Based Speech enhancement," in Proceedings Of :e Icassp 2020-2020 Ieee International Conference On Acoustics, Speech And Signal Processing (icassp), pp. 7519–7523, IEEE, Barcelona, Spain, May 2020.

[6] N. Saleem, M. I. Khattak, M. Al-Hasan, and A. B. Qazi, "On learning spectral masking for single channel speech enhancement using f and recurrent neural networks," IEEE Access, vol. 8, pp. 160581–160595, 2020.

[7] N. Saleem and M. I. Khattak, "Multi-scale decomposition based supervised single channel deep speech enhancement," Applied Soft Computing, vol. 95, Article ID 106666, 2020.

[8] S. Chakrabarty and E. A. P. Habets, "Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 4, pp. 787–799, 2019.