INTERNATIONAL RESEARCH JOURNAL OF SCIENCE ENGINEERING AND TECHNOLOGY

ISSN 2454-3195

An Internationally Indexed Peer Reviewed & Refereed Journal

WWW.RJSET.COM
www.isarasolutions.com

Published by iSaRa Solutions

A REVIEW ON SPEECH ENHANCEMENT TECHNIQUES

Akash Redhu
DCRUST. University, Sonipat, HR 131039, INDIA

ABSTRACT

In a communication system, a signal is the information-containing component that has to be processed. However, noise is introduced into the signal during processing, making the signal noisy. The source of the noise, such as a noisy engine, pump, etc., adds noise into the radio communication equipment or telephone channel. While various speech improvement algorithms have been developed by researchers to reduce noise, nothing has been done to increase speech intelligibility. While it has been discovered that noise tracking or voice activity recognition algorithms work well in background noise that is constant, they do not work well in non-stationary noises like multi-talker chatter. In most adverse situations when hearing-impaired or normal listeners find it difficult to comprehend what is being said, most algorithms employ the soft gain function to suppress noise but fail to enhance speech quality and intelligibility. The gain function limitation in terms of increasing comprehensibility, stem from the fact that it is soft. This paper includes an in-depth analysis of existing types of speech enhancement techniques and recent advances in numerous modalities, most notably audio data.

Keywords: Speech Enhancement, Single Channel, Ideal Binary Mask, Power Signal to Noise Ratio

1. INTRODUCTION

1.1 SPEECH SIGNAL

Speech is the basic means of communication among human beings. It is a physical phenomenon in which variations in local acoustic pressure are caused by the actions of the human vocal system. Acoustic waves are created by these pressure changes and travel through the transmission medium, often air. Higher cortical regions of the brain at the receiving end process speech through the auditory system. The speech chain is represented in Figure 1.1.

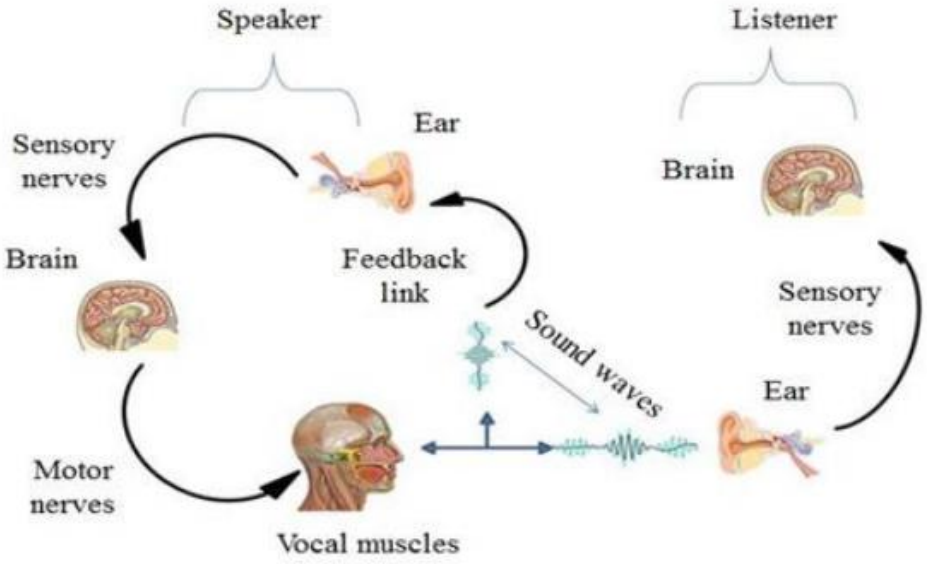

Figure 1.1 Speech Chain[1]

1.2 INTRODUCTION TO SPEECH PROCESSING

In human physiology, communication is one of the most essential functions. There are various ways available for human beings for information retrieval and communication. The three main forms of information communication are speech, images & written text. However, speech is predominately used for efficient and convenient communication. This is due to the fact that communication by speech helps in conveying the linguistic contents and also the mood of the speaker. If the speaker and listener are closer to one another in a quiet setting, the accuracy and ease of verbal communication is further enhanced. If the speaker is far away or the environment is noisy, the listener's capacity to understand speech decreases. Hence, the quality and unambiguity of speech holds greater importance for ensuring easy and accurate exchange of information during speech communication [3]

The majority of speech processing systems that aid in conveying or storing speech are typically created for an environment with minimal background noise. The speech signal is severely degraded when interference, such as additive noise and channel noise, is present in the real-world environment. Another type of noise that degrades the speech signal is Multiplicative Noise or Convolutional noise. In practice, It is possible to transform convolutional noise into additive noise such that it may be eliminated by standard additive noise removal algorithms. As a result of substantial study on speech processing, a variety of techniques have been created for recovering the desired speech from the degraded speech. The intricacy of the speech signal, however, makes it difficult to separate the required speech signal from the background noise and speech mixing. Therefore, noise suppression is mandatory for maintaining the signal quality [4]

The process of communication in speech communication systems may introduce a narrow band additive disruption. The intelligibility of the speech would be reduced due to this degradation by means of using the channel. An example of this is when the pilot of an aeroplane speaks with the air traffic control tower. Speech in such a setting is typically impaired by background noise that is added on top of other noise. For this kind of communication, intelligibility is the most important perceptual property of speech. Therefore, it is vital to increase the speech signal's understandability [5]

1.3 SPEECH ENHANCEMENT SYSTEM

The primary objective of the speech enhancement process is to improve the signal's quality and clarity. Speech enhancement occurs when damaged speech is returned to its original speech signal. However, few important differences lie between enhancement and restoration. While speech enhancement seeks to make the processed signal sound better than the unprocessed signal, speech restoration seeks to bring the processed signal as close to the original as possible. In actual practice, even though it is known that further restoration of the degraded signal is not possible, still can be enhanced by improving its clarity

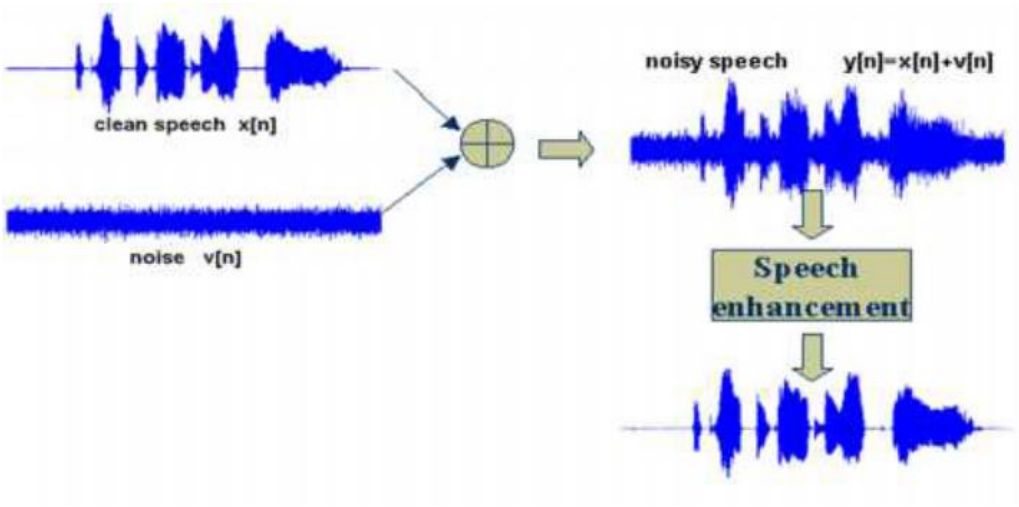

Figure 1.2 Speech Enhancement process[5]

Figure 1.3 shows the speech enhancement process. The Speech enhancement approach varies Depending on the circumstances surrounding the issue. For instance, speech that has been distorted by additive noise requires a very different kind of processing than speech that has been distorted by echoes. The connection between speech signal quality and speech bandwidth compression system is another important application of speech enhancement. Speech bandwidth compression system is known to be a key factor in speech communication system owing to the increased use of digital communication channels for speech signal decoding and placing more of an emphasis on voice/data networks that are integrated.

In this regard, the signal degradations can be classified into three groups based on the way in which the required speech signal is altered. The needed speech signal may contain irrelevant

additive noise that is added to it in either the acoustic or electrical domains. The additive noise degrades the listening ability and intelligibility. In extreme cases, It might entirely conceal the necessary signal. In the case of some additive noise, the spectral properties are stationary or change gradually over time. Hum, amplifier noise, and other environmental acoustic noise sources are examples of this additive noise. Spectral subtraction and single-channel adaptive filtering techniques are used to lower the perceived level of such stationary noise sources. [6].

Intermittent or very non-stationary noise is the second type of additive noise. Media interference, unwanted co-talkers, and some types of electrical interference are some examples of such non-stationary noise sources. Due to complex effects caused by the third type of noise being significantly connected with the desired signal, it is thought to be reverberation. [7] The acoustic reflections are responsible for arising of reverberation and echo which in turn leads to degradation of intelligibility.

The main constraints on communication and measuring systems are noise and distortion. Therefore, the core of communications theory and practice is the modeling and elimination of the effects of noise and distortion. [8] Due to the assorted qualities of sources of noise and their influence upon various applications, speech enhancement is treated as a challenging issue.

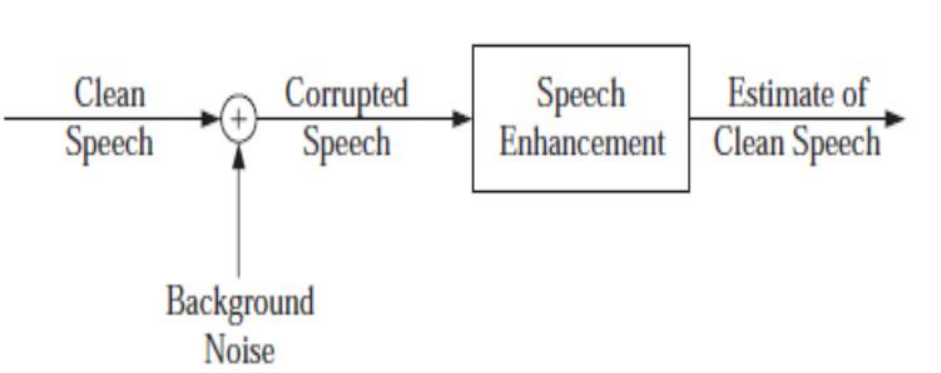

Figure 1.3 Basic speech enhancement method[6]

The three main objectives of any speech enhancement system are:
(i) to improve the clarity and quality of the processed speech
(ii) to increase the speech coders' reliability and
(iii) to increase the accuracy of speech recognition systems

1.4 TYPES OF SPEECH ENHANCEMENT METHODS

Different criteria can be used to categories speech enhancement algorithms:

- Number of channels: Depending on the number of sensors being taken into consideration, either single- or multi-channel.
- Algorithm output: If the algorithm acquires the target speech signal after being cleaned, or whether it obtains all the sources in the mixture.
- Mixture type: Anechoic or echoic, immediate.
- Algorithmic Strategy: channels and sources, the human auditory model, a spatial filter, a

time or frequency domain, etc.

The algorithms are separated into two distinct but related sections in this overview of the state of the art: Techniques for reducing noise and SSS algorithms.

Both are divided into single-channel and multichannel techniques as well. Assuming that all sound sources are noise, noise reduction algorithms resolve the problem of speech signal prediction from a noise-corrupted version of itself. The issue of determining each original source present in an audio mixture is connected to SSS. The most significant solutions that have been put out to address these issues are reviewed in this part.It is also examined whether it is feasible to use these algorithms for speech enhancement in hearing aids.

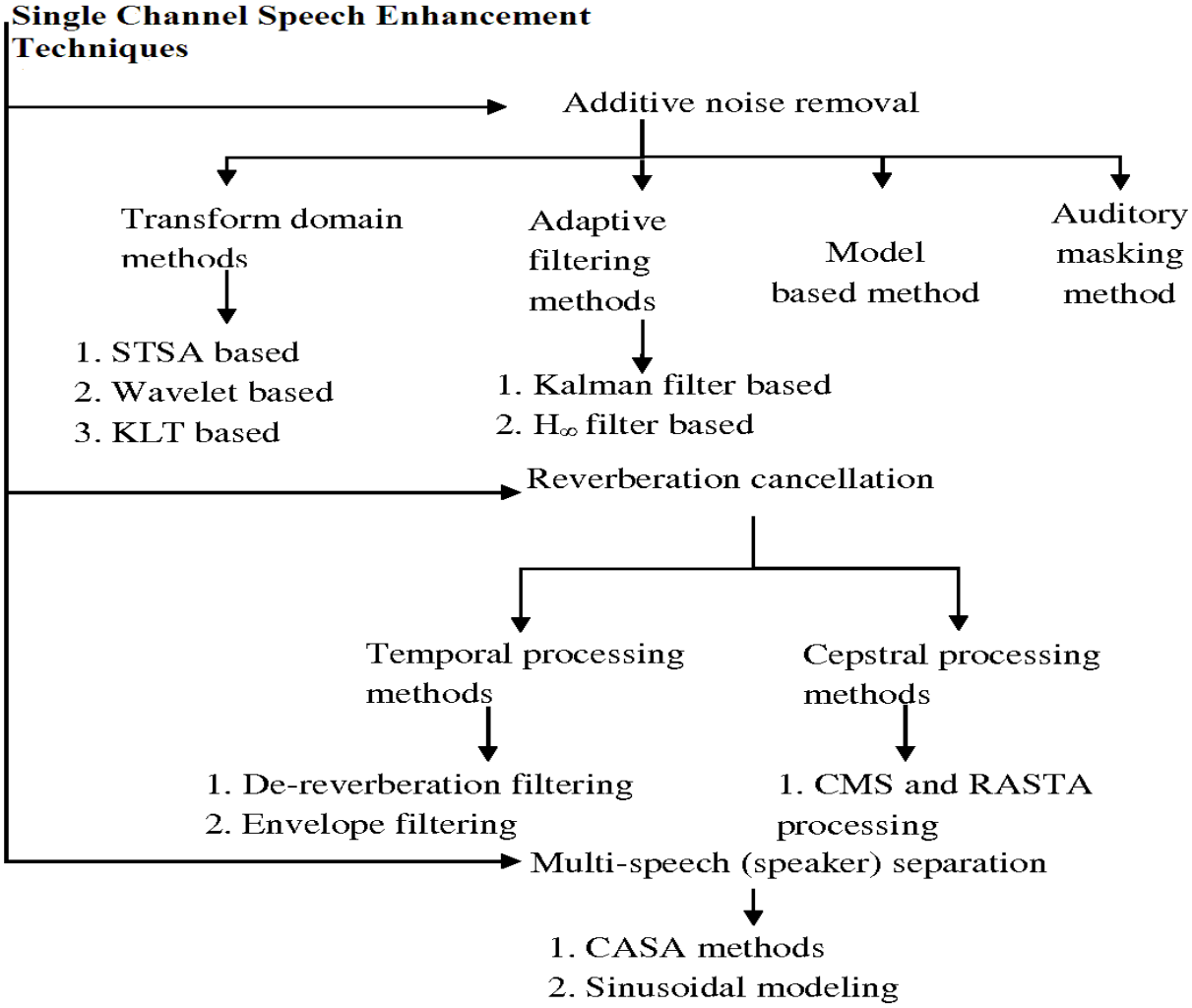

Figure 1.4 The chart of Single Channel Speech Enhancement techniques[9]

2 LITERATUREREVIEW

Hari, A., *et al*. (2020) suggested that researchers and engineers have developed a number of speech improvement algorithms over the past 40 years to reduce background noise, but nothing has been done to increase speech intelligibility. This study's major objective has been to explore

how binary mask can be used to improve speech quality and comprehension in the most difficult listening situations, such as when hearing-impaired or non-impaired listeners find it difficult to grasp what is being said.[1]

In January 2021, Fu, Y., *et al.* presented a multi-channel network for concurrent voice de-reverberation, augmentation, and separation (DESNet).The attentional selection method of the multi-channel features that was initially given in the end-to-end unmixing, the authors implemented gradient propagation and cooperative optimization using a fixed-beamforming and extraction (E2E-UFE) framework.[2]

In many applications, speech augmentation is an essential and difficult undertaking. In this research, Cui, X., *et al*. (2020). provided a novel simple recurrent unit (SRU)-based technique for voice augmentation Following that, the mapping relationship between the noisy and clean speech spectra is learned using a multi-layer stacked SRU network[3]

In many situations, improving speech in loud, reverberant environments is an important and difficult undertaking. Cui, X., *et al.* (2021) In order to combat noise and reverberation, The BiLSTM network is used in this study's proposal for a multi-objective based, multi-channel speech enhancement system. Each channel of the microphone array is originally furnished with the log-power spectra of loud speech so that the BiLSTM network can anticipate the LPS and IRM of clean speech.[4]

Masking-based deep neural networks (DNN) voice augmentation cannot establish the time-frequency masking value exactly because any potential speech structure information is ignored. Jia, H., *et al*.(2021) suggest a way for improving speech in this work. To generate a combined dictionary, the sparse NMF (SNMF) will next independently learn the speech and noise cochleagrams.[5]

Despite the long history of single channel speech enhancement research, there are still two crucial practical problems that need to be overcome. First of all, it is difficult to balance quality improvement with computational efficiency, and low-latency is always a trade-off for quality. Second, one of the complex issues with conventional methods is improvement in certain situations, like singing and emotional speech[6]

For the pre-processing of a speech recognizer, Tu, Y. H., *et al*.(2019) suggested a novel teacher-student learning approach. A teaching model with deep architectures is first created using simulated training pairs of clean and noisy voice samples in order to learn the target of the IRMs. After that, a student model is trained using the IMCRA method and the instructor model's estimated IRMs to learn an enhanced speech presence probability.[7]

S. Chakrabarty *et al.* (2019). introduced a method for online multi-channel voice augmentation based on time-frequency (T-F) masking that estimates the mask using a CRNN. It is explained

how to estimate two different masks, the ideal binary mask and the IRM, as well as how to use the mask in two different ways to create the desired signal (IBM). In the second technique, a beam shaper will apply recursive updating of PSD matrices, and the masks are employed as an activity indication, as contrast to the first method, when a reference microphone signal is instantly subjected to the mask in a real-valued gain manner.[8]

Here, Ingale, P. P., and Nalbalwar, S. L.(2019) presented a voice augmentation technique based on DNN that employs a mono channel mask The proposed technique finds the first binary mask using a cochleagram. Prior to producing the mono channel mask, DNN selects the appropriate spectral structure in the target speech-related frames. The mono channel mask is used to reconstruct the spoken stream. Mono channel mask reduces unwanted interference from noisy time-frequency (T-F) devices[9]

3.PROBLEMS IDENTIFIED

After analyzing the facts in the previous sections, the existing speech enhancement approaches have considerable limitations. These limitations are as follows:

- **Noise Sources that Impair Speech :**Noise is a phrase used to describe an unwanted signal that obstructs another signal's communication or measurement.Noise can taint speech at any point before it reaches the recipient.
- **High Complexity:** A high complex algorithm always utilizes extra computational time and resource power, whereas a low complex algorithm is prone to the intruder
- **Low Detection Convergence, Perceptual and Audio Quality**: It is also noticed and analyzed from the previous sections that speech enhancement process has low PSNR. This reflects that existing speech enhancement algorithms suffer from Low Detection Convergence, Perceptual and Audio Quality.
- **Longer Execution Time**: The time taken by most of the existing proposed methods is considerably high. Computational time is directly connected to the algorithm's complexity. As discussed earlier, complexity has to be reduced to a moderate level to reduce the computational time.

4. CONCLUSION

Following an examination of speech enhancement methods, it is clear that each method clearly has its own set of disadvantages. Our review focuses on evaluations of various speech enhancement methodologies andclassifications, as well as an in-depth examination of the literature and numerous research works in this subject. As discussed above that the existing approaches have limitations, such as Noise Effects, high complexity, low- lower perceptual and audio quality, longer execution time and there is a strong need to overcome these limitations, although in past many authors or researchers have already worked to overcome these limitations based on the various significant speech enhancement algorithms. In my Future work, I will propose a modified approach that is based on ideal binary masking (IBM) for speech enhancement.

REFERENCES

[1]. Hari, A., Afnan, A., Abbas, J., Shariff, F. A., & Nuthakki, R. ,Speech Enhancement Using Ideal Binary Mask Based on Channel Selection Criteria. *International Journal of Research in Engineering, Science and Management*, Volume-3 (Issue-1), pp. 111-114 , 2020

[2]. Fu, Y., Wu, J., Hu, Y., Xing, M., & Xie, L. Desnet, A multi-channel network for simultaneous speech dereverberation, enhancement and separation. *IEEE Spoken Language Technology Workshop (SLT)*, pp. 857-864, 2021.

[3]. Cui, X., Chen, Z., & Yin, F., Speech enhancement based on simple recurrent unit network. *Applied Acoustics*, Volume *157,pp.*107019-107028, 2020.

[4]. Cui, X., Chen, Z., & Yin, F. , Multi-objective based multi-channel speech enhancement with BiLSTM network. *Applied Acoustics*, Volume *177*, pp. 107927-107939, 2021.

[5].Jia, H., Wang, W., & Mei, S. , Combining adaptive sparse NMF feature extraction and soft mask to optimize DNN for speech enhancement. *Applied Acoustics*, Volume *171*, pp. 107666-107672,2021.

[6]. Li, J., Luo, D., Liu, Y., Zhu, Y., Li, Z., Cui, G. & Chen, W., Densely Connected Multi-Stage Model with Channel Wise Subband Feature for Real-Time Speech Enhancement. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,pp. 6638-6642, 2021.

[7]. Tu, Y. H., Du, J., & Lee, C. H., Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition. *IEEE/ ACM Transactions on Audio, Speech, and Language Processing*, Volume *27*(12), pp. 2080-2091, 2019.

[8]. Chakrabarty, S., & Habets, E. A., Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, Volume *13*(4), pp. 787-799, 2019.

[9]. Ingale, P. P., & Nalbalwar, S. L., Deep neural network based speech enhancement using mono channel mask. *International Journal of Speech Technology*, Volume *22*(3), pp.841-850, 2019.

[10]. Bao, F., & Abdulla, W. H., A new ratio mask representation for CASA-based speech enhancement. *IEEE/ ACMTransactions on Audio, Speech, and Language Processing*, Volume *27*(1), pp. 7-19, 2018.

[11].Lee, G. W., & Kim, H. K., Multi-task learning u-net for single-channel speech enhancement and mask-based voice activity detection. *Applied Sciences*, Volume No *10*(9), issue(3230), pp. 1-15 2020.

[12]. Chakrabarty, S., Wang, D., &Habets, E. A., Time-frequency masking based online speech enhancement with multi-channel data using convolutional neural networks. *International Workshop on Acoustic Signal Enhancement (IWAENC)* ,pp. 476-480, 2018.

[13].Tu, Y. H., Tashev, I., Zarar, S., & Lee, C. H., A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , pp. 2531-2535, 2018.

[14]. Hao, X., Shan, C., Xu, Y., Sun, S., & Xie, L., An attention-based neural network approach

for single channel speech enhancement. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* ,pp. 6895-6899, 2019.

[15].Bulut, A. E., &Koishida, K., Low- latency single channel speech enhancement using u-net convolutional neural networks.*IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , pp. 6214-6218, 2020.

[16].Lee, J.,Skoglund, J., Shabestary, T. & Kang, H. G., Phase-sensitive joint learning algorithms for deep learning-based speech enhancement. *IEEE Signal Processing Letters*, Volume *25*(8), pp. 1276-1280, 2018.

[17].Taherian, H., Wang, Z. Q., Chang, J., & Wang, D., Robust speaker recognition based on single-channel and multi-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Volume *28*, pp. 1293-1302, 2020.

[18].Lee, J., & Kang, H. G., A joint learning algorithm for complex-valued tf masks in deep learning-based single-channel speech enhancement systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Volume *27*(6), PP. 1098-1108, 2019.

[19]. Li, C., Shi, J., Zhang, W., Subramanian, A. S., Chang, X., Kamo, N., & Watanabe, S. , ESPnet-SE: end-to-end speech enhancement and separation toolkit designed for ASR integration. *IEEE Spoken Language Technology Workshop (SLT)* , pp. 785-792,2021.

[20]. Luo, Y., Chen, Z., & Yoshioka, T., Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 46-50,2020.

[21]. Das, N., Chakraborty, S., Chaki, J., Padhy, N., & Dey, N., Fundamentals, present and future perspectives of speech enhancement. *International Journal of Speech Technology*, Volume *24*(4), pp. 883-901, 2021.

[22].Wang, Z. Q., Wang, P., & Wang, D., Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR. *IEEE/ACM transactions on audio, speech, and language processing*, Volume *28*, pp. 1778-1787, 2020.

[23]. Roy, S. K., Nicolson, A., & Paliwal, K. K., Deep learning with augmented Kalman filter for single-channel speech enhancement. *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1-5, 2020.

[24]. Zhao, Y., & Wang, D., Noisy-Reverberant Speech Enhancement Using DenseUNet with Time-Frequency Attention. In *INTERSPEECH*, pp. 3261-3265,2020.

[25].Morrone, G., Bergamaschi, S., Pasa, L., Fadiga, L., Tikhanoff, V., &Badino, L., Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , pp. 6900-6904, 2019.

भारतीय भाषा, शिक्षा, साहित्य एवं शोध

ISSN 2321 – 9726

WWW.BHARTIYASHODH.COM

INTERNATIONAL RESEARCH JOURNAL OF MANAGEMENT SCIENCE & TECHNOLOGY

ISSN – 2250 – 1959 (0) 2348 – 9367 (P)

WWW.IRJMST.COM

INTERNATIONAL RESEARCH JOURNAL OF COMMERCE, ARTS AND SCIENCE

ISSN 2319 – 9202

WWW.CASIRJ.COM

INTERNATIONAL RESEARCH JOURNAL OF MANAGEMENT SOCIOLOGY & HUMANITIES

ISSN 2277 – 9809 (0) 2348 - 9359 (P)

WWW.IRJMSH.COM

INTERNATIONAL RESEARCH JOURNAL OF SCIENCE ENGINEERING AND TECHNOLOGY

ISSN 2454-3195 (online)

WWW.RJSET.COM

INTEGRATED RESEARCH JOURNAL OF MANAGEMENT, SCIENCE AND INNOVATION

ISSN 2582-5445

WWW.IRJMSI.COM

JOURNAL OF LEGAL STUDIES,POLITICS AND ECONOMICS RESEARCH

WWW.JLPER.COM