Review On Speech Enhancement Techniques

[1]Thanushree S, [2]Sneha E, [3]Tejaswini L, [4]Rakshitha I M

[1]Student, [2]Student, [3]Student, [4]Student

Department of Electronics and Telecommunications

Dayananda sagar college of engineering, Bengaluru, India

Abstract: Speech is the most organic and efficient form ofcommunication. Throughout the speech When communicating, the signals carry some noise, so digital voice signal processing is acrucial phase in speech enhancement. A speech signal is typically corrupted by ambient noise, background noise, and reverberations. The reason for speech enhancement is to increase the understandability of speech and the speech signal's clarity. There are several techniques proposed using which speech signal enhancement is performed. We use MMSE, MAP, and DNN algorithms to enhance speech. The objective of this paper is to provide an overview of speech enhancement algorithms that areused for the enhancement of speech signals.

Index Terms: Background noise, Deep Neural Network, Speech communication, Speech signal enhancement, Speech signal processing.

I. INTRODUCTION

Speech has a vital part in social development and is the easiest and most efficient means to exchange information. In real life, noise is ubiquitous. Environmental noise frequently interferes with voice transmission, lowering voice quality. Voice enhancement technology has significant research and application value because it may enhance speech signal quality and lessen the influence of noise. Speech enhancement technology's goal is to eliminate or significantly lessen the effect of back ground noise on speech. It is frequently utilized in a number of different industries, including speech recognition, emotion identification, and endpoint detection.The study of artificial intelligence and its applications in the field of voice enhancement has gained increasing attention in recent years due to the rapid development of these technologies. Speech enhancement is the technique that improves the quality of speech signals. For applications like Automatic Speech Rec0gnition (ASR), mobile speech communication, speaker recognition, hearing aids, and speech coding, speech enhancement techniques are used to remove environmental noise that interferes with the target speech signal. One of the most crucial areas of research in the telecommunications and signal-processing industries is speech enhancement. Different enhancement techniques havebeen proposed as a result of the many uses of speech enhancement in voice communications, control systems, and the transmission of audio and speech signals. The quality of voice communications can be lowered when speech signals are contaminated with noise signals. Avoice enhancement system can be used to tackle this issue by reducing the impact of noise components on the noisy speech signal. In fact, a speech enhancement algorithm's goal is to extract the original, clear voice signal from the cluttered observations.

Numerous techniques have been put forth over the years to Improve noisy speech signals. These techniques can be divided into two groups: machine learning-based (ML) and non-MLbased or statistical approaches. For improving speech, amodified coherence-based dictionary learning method is proposed. To train the speech and noise dictionaries in the sparse domain, which considerably improves the voice quality in additive noisy settings, it is necessary to have data from both speech and noise. To estimate the complex non-linear mapping from observed noisy speech to desired cleansignals in the DNN learning approach, a large training set is necessary. Although machine learning (ML)-based methods show promising results in noisy environments where the system was initially trained, their ability to generalize to new noisy conditions is still under debate. Non-ML(statistical) approaches, on the other hand, are not plagued by this issue and are more effective when ML-based techniques are unsuccessful. Because of this, they are still applied in speech enhancement systems either as a stand alone enhancement procedure or in combination with ML-based ones. Due to their appropriate performance in various noisy c0nditions, MMSE-based methods continue to be widely used among statistical algorithms. The main problem with the traditional methods is that the improved speech contains musical noise. Minimum Mean Square Error (MMSE) provides better- quality improved speech with less musical noise when compared to other conventional approaches.

Over the years, many speech-improving technologies have been developed for various objectives. It hasn't been proven to be a successful speech enhancement technique over time for low SNR environments like an automobile, when in motion. Directional noise and diffuse noise both make up the noise in an automobile. It is well known that a MAP- based speech enhancer works well for diffuse noise, whereas a two-channel null beamformer works well for directional noise.To extract the target speech's previously suppressed low- frequency components, we added an additi0nal MAP-based speech enhancer technique. Despite improving suppressing performance on some criteria, the method's voice extraction precision was insufficient for practical usage.

DNN's applications span a variety of fields, including speech enhancement, audio recognition, and others. To increase the performance and robustness to noise in DNN-based voice enhancement systems, noise aware training(NAT) is performed by adding noise information to the DNN inputs. Both the noise produced by the interfering sources and the clear speech are included in the mixture data sent to the DNN for training. By translating the data vectors to the relevant encoding vectors, DNN modelling is carried out. The DNN is designed to learn the mapping between the complex ideal ratio mask and the reverberant speech. DNN functions as a noise classifier, and the filter parameters are selected to remove noise.

II .REVIEW OF PREVIOUS WORK

Literature on speech enhancement has been reviewed and discussed in this session based on different algorithms.

V K Gupta and Anirban Bhowmick[1] have used MMSE for speech enhancement. They performed an experiment for fifty speakers, and a database of ten Hindi digits was created. The noisy database was created using Speech and F16 noises, both of which have varying Signal-to-Noise Rati0 (SNR) levels (-5dB, 0dB,5dB, 10dB). Prior to feature extraction, the voice was de-noised using spectral estimation techniques such as Spectral Subtraction (SS) and Minimum Mean Square Error(MMSE) estimation-based methods. With respect to both noises, MMSE Cohen and MMSE Cohen Log performed better for speech rec0gnition at lower and higher SNR , respectively. Figure 1 shows that, among all MMSE approaches, the recognition results with the MMSE Cohen technique are best between -5dB and 0dB. Additionally, it is clear from Fig. 1 that, out of all MMSE procedures, the MMSE Cohen Log methodology produces the greatest recognition results at 5dB and 10dB.

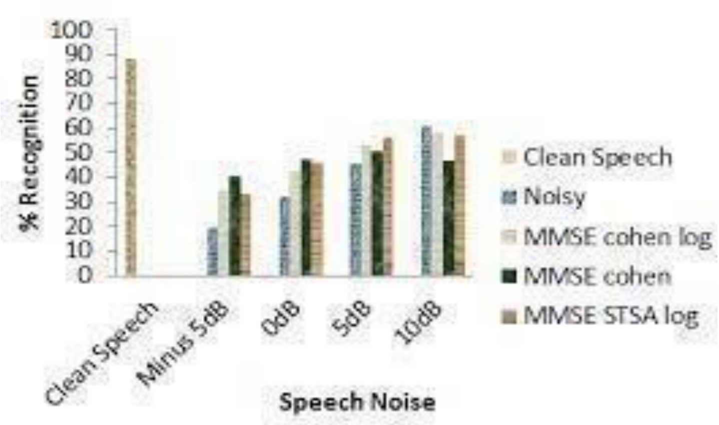


Fig 1: Speech enhancement with MMSE estimation methods with speech noise.

A different approach has been taken by Muhammad H. Santriaji, Human Mauludin, Dian Surgawiwaha, Trio Adiono[2] proposed a sophisticated method for estimating n0ise for various audio noise reduction applications is called Maximum a Posteriori (MAP). The variable speech distribution MAP approach requires complex and time- consuming computing. This work offered a strategy for software implementation of the MAP algorithm with variable voice dispersion.Some arithmetic functions have been optimized by switching them out for approximation functions, and the compiler now includes an optimization option. The simulation outcomes demonstrate that the improved MAP method, when subjected to a time budget constraint, results in a linear improvement in SNR and a shorter calculation time. To do benchmarking, they have used four different speech types, each with unique characteristics. Figure 2 displays each speech's profile. They altered each speech's SNR ratio by inserting noise, causing the SNR to fluctuate between 1 and 10. The outcomes of the MAP speech enhancement are then recorded for those speeches.

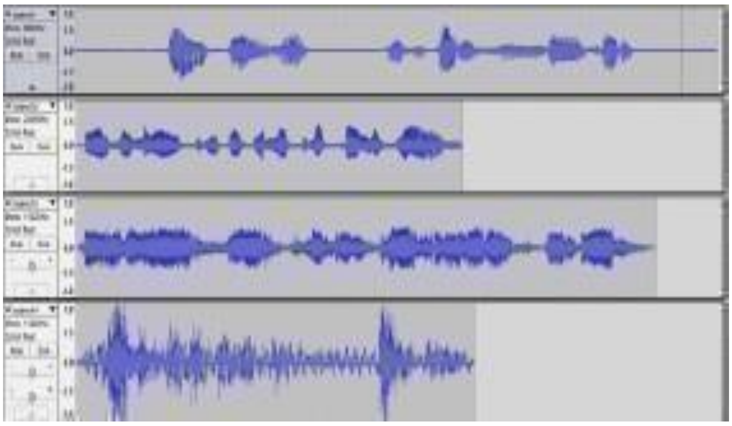
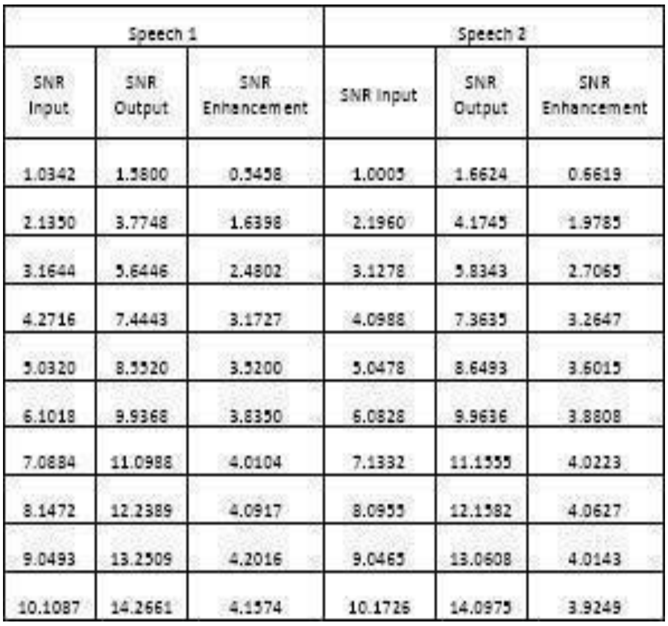
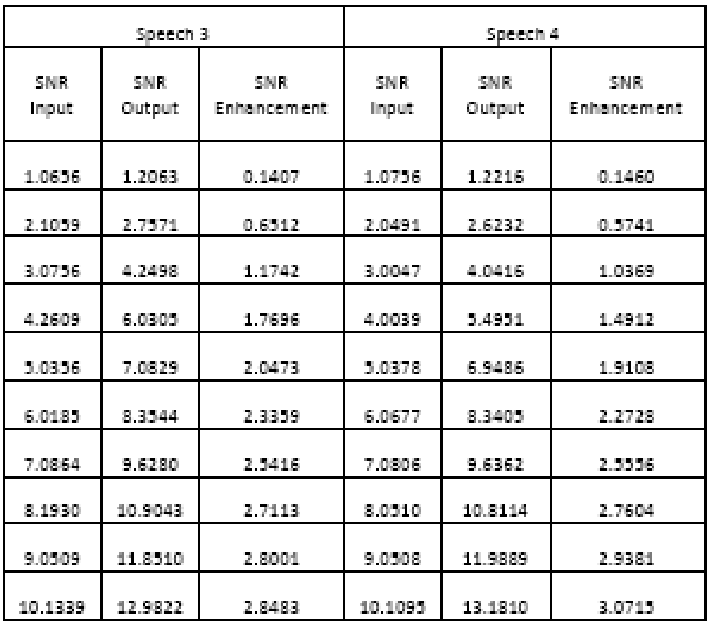


Fig 2: Four different speech’s profile.



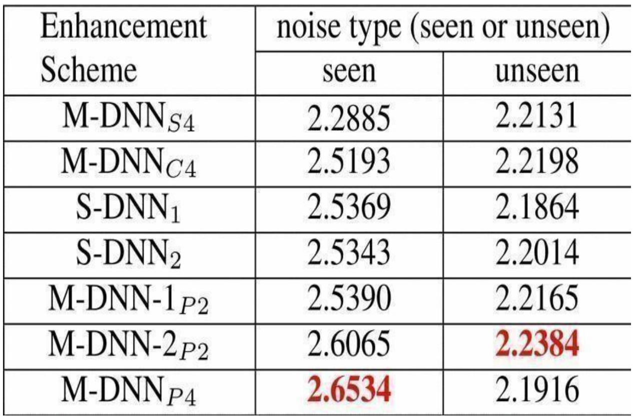
Speech 1			Speech 2		
SNR Input	SNR Output	SNR Enhancement	SNR Input	SNR Output	SNR Enhancement
1.0342	1.5800	0.5458	1.0005	1.6624	0.6619
2.1350	3.7748	1.6398	2.1960	4.1745	1.9785
3.1644	5.6446	2.4802	3.1278	5.8343	2.7065
4.2716	7.4443	3.1727	4.0988	7.3635	3.2647
5.0320	8.5520	3.5200	5.0478	8.6493	3.6015
6.1018	9.9368	3.8350	6.0828	9.9636	3.8808
7.0884	11.0988	4.0104	7.1332	11.1555	4.0223
8.1472	12.2389	4.0917	8.0955	12.1582	4.0627
9.0493	13.2509	4.2016	9.0465	13.0608	4.0143
10.1087	14.2661	4.1574	10.1726	14.0975	3.9249



Speech 3			Speech 4		
SNR Input	SNR Output	SNR Enhancement	SNR Input	SNR Output	SNR Enhancement
1.0656	1.2063	0.1407	1.0756	1.2216	0.1460
2.1059	2.7571	0.6512	2.0491	2.6232	0.5741
3.0756	4.2498	1.1742	3.0047	4.0416	1.0369
4.2609	6.0305	1.7696	4.0039	5.4951	1.4912
5.0356	7.0829	2.0473	5.0378	6.9486	1.9108
6.0185	8.3544	2.3359	6.0677	8.3405	2.2728
7.0864	9.6280	2.5416	7.0806	9.6362	2.5556
8.1930	10.9043	2.7113	8.0510	10.8114	2.7604
9.0509	11.8510	2.8001	9.0508	11.9889	2.9381
10.1339	12.9822	2.8483	10.1095	13.1810	3.0715

Table I: The outcomes of the MAP speech enhancement

Pavan Karjol, M Ajay Kumar, Prashanth Kumar Ghosh[3]published a method of DNN algorithm. For many real-time speech applications, voice enhancement is essential, but it can be difficult when using single-channel because there is essentially only one data channel accessible. There are an infinite number of solutions that might be used to tackle this issue without any restrictions. They describe a supervised learning method in this research to improve speech that has been damaged by speech-babble noise, the most difficult kindof noise for speech enhancement systems. Deep neural networktechnology underlies the suggested approach (DNNs). Using segmental SNR, perceptual evaluation of speech quality (PESQ), and short-term objective intelligibility (STOI) as the evaluation metrics, they also compare the proposed method with a single DNN-based speech enhancement scheme and existing multiple DNN methods. These comparisons demonstrate the proposed method's advantages over baseline approaches in both seen and unseen noise. When the PESQ measure is averaged over all noises and SNRs for the cases of seen and unseen noise,they find an absolute improvement of 0.07 and 0.04 compared to a single DNN.



Enhancement Scheme	noise type (seen or unseen)	
	seen	unseen
M-DNN$_{S4}$	2.2885	2.2131
M-DNN$_{C4}$	2.5193	2.2198
S-DNN$_1$	2.5369	2.1864
S-DNN$_2$	2.5343	2.2014
M-DNN-1$_{P2}$	2.5390	2.2165
M-DNN-2$_{P2}$	2.6065	**2.2384**
M-DNN$_{P4}$	**2.6534**	2.1916

Table II: PESQ results averaged over SNRs and noises

III.SPEECH ENHANCEMENT

Speech enhancement is a phase in digital voice signal processing that aims to improve the quality of the speech signal by improving its clarity, intelligibility, understandability, and comprehensibility with the aid of an algorithm or filter. There are many factors, including reverberation, babbling, and other background noises that are recorded during the recording, that might cause the speech signal to degrade. Clean and noise- free speech signals are necessary for certain types of speech- enabledapplications, such as speaker recognition, mobile applications, hearing aids, VoIP, etc. Different techniques can be used to improve speech. The method for speech enhancement differs depending on the kind of deterioration and noise in the acquired speech signal. Fig 3 shows theBasic steps of the speech enhancement system.

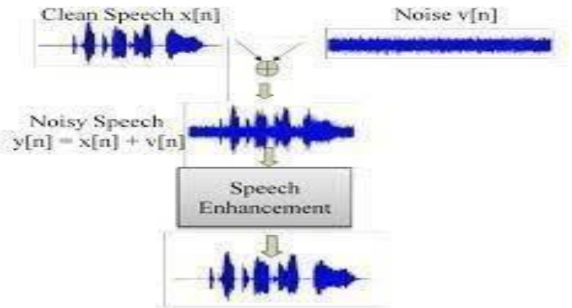


Fig 3: basic steps of speech enhancement

IV.DATABASE

For the purposes of the research findings, the term "database" refers to both the clean speech data and the noisy speech data. Data from the NOIZEUS corpus was used to create the clean speech. Clean speech utterances are available in eight distinct languages in the NTT database (English, American English, Japanese, German, Chinese, Spanish, French, and Italian). The Aurora-2 and are used to collect noisy data. Common noise sources used to train and test he DNN include babble, restaurants, cars, exhibition halls, airports, train stations, streets, cafeterias, machine guns, Volvo factories, and others.

V. SPEECH ENHANCEMENT TECHNIQUES

(1) MINIMUM MEAN SQUARE ESTIMATOR(MMSE)

The following examples illustrate how noise can degrade clearspeech:

y(n) = x(n) + v(n)--------------(1)

The undesirable additive noise (v(n)) in equation

(1) is thought to be a zero mean random process that is uncorrelated with x(n). In a noisy environment, it is necessary toestimate x(n) from the noise-corrupted signal y(n). There are numerous ways to calculate x(n) or lower v(n).

The mean square error (MSE), a typical parameter of estimator quality, is minimized using a minimum mean square error (MMSE) estimator. According to equation (1), x(n) is a random variable with a uniform distribution between (0, 1),and v(n) is a random variable with a Gaussian distribution and a mean and variance of 0. Noise v(n) and the signal x(n) are uncorrelated. Finding an estimator that reduces Mean Square Error is the primary goal (MSE). MSE is represented as :

MSE= $E[(X-\hat{X})2]$ ————————(2)

Here, X stands for the pure signal x(n), and $\hat{X}$ for the estimatedclean signal. The estimator achieving the lowest MSE is known ∂ as the MMSE estimator. Here, the noisy signal y(n), which is the random variable Y, is known but the random variable X is unknown. Given that Y= y and assuming that the distribution of X|Y is known, X can be approximated by identifying a function

$\hat{X}$= g(y),such that

J= MSE= E[(X-g(y))2|Y=y]------(3)

An estimator is any function that depends on y. The MMSE estimator (g(y)) is a type of estimator that minimizes J. Now for minimization,

$$\frac{\partial}{\partial\, g(y)} = \text{-------------}(4)$$

So from equations (3) & (4), we can conclude equation (5)

g(y) = E[X|Y = y]------------- (5)

The amplitude of each Fourier expansion coefficient of the speech signal is computed from the noisy speech signal in MMSE log STSA (Short Time Spectral Amplitude). The speech signal's and the noise's Fourier expansion coefficients are treated as statistically independent Gaussian random variables.

(2) MAXIMUM A POSTERIOR (MAP)

Figure 4 shows an overview of the MAP based noise reduction algorithm implemented in this work
Noisy input signal x(n) is consist of clean speech s(n) and noised(n).

x(n)= s(n)+d(n) ----------------(6)

Frame after frame, the spectral-based voice augmentation is applied. As a result, the input signal is multiplied by a window w(n).

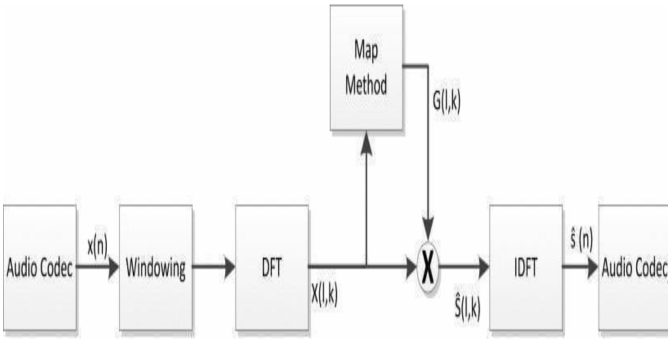


Figure 4: MAP Speech enhancement system.

$X_W(n) = s_W(n)+d_W(n)$----------(7)

The Discrete Fourier Transform (DFT) is applied to the observed signal $x_W(n)$ to eliminate noise in the frequency domain (n). Next, the observed signal specter $X_W(l,k)$ in the lframekis stated as follows:

$X_W(l,k) = S_W(l,k)+D_W(l,k)$------ (8)

when the speech spectrum is denoted by $S_W(l,k)$ and the noise spectrum is denoted by $D_W(l,k)$. On the basis of empirical data, a MAP estimator will produce a pointestimate of an unobserved quantity. The observed signal spectre $X_W(l,k)$ expresses the estimate value of speech $S_W(l,k)$ as

$\hat{S}_W(l,k)= G_{LV,MAP}(l,k)X_W(l,k)$ ----------(9)

A spectre gain from the Lotter and Vary approach is $G_{lv,MAP}(l,k)$. The observed signal spectre X(l,k) is multiplied by the appropriate spectre gain $G_{lv,MAP}(l,k)$ to get the0estimate value-specter.

$$G_{LV,MAP}(l,k)= v(l,k)+ \sqrt{v^2(l,k)+\frac{v(l,k)}{2\gamma(l,k)}}\text{------}(10)$$

γ is A posteriori SNR and ξ is a priori SNR, given by

$$\gamma=\frac{|x|^2}{\sigma^{2d}}\text{------------}(11)$$

$$\xi= \frac{\sigma^{2s}}{\sigma^{2d}}\text{----------}(12)$$

The algorithm for determining v (l, k) is given by-

$$V(l,k) = \frac{1}{2} - \frac{\mu(l,k)}{4}\sqrt{\frac{1}{\gamma\xi(l,k)}}\text{-----------}(13)$$

It is possible to obtain the spectral gain based on the variable speech distribution by allowing the spectral gain parameter to vary.

$$V(l,k)= \begin{cases} 2.0, \tilde{v}(l, k) > 2.0 \\ \tilde{v}(l, k), 0.0 \le \tilde{v}(l, k) \le 2.0 \text{-----------}\ (14) \\ 0 \quad 0.\ \tilde{v}(l, k) \le 0.0 \end{cases}$$

$$\tilde{v}(l,k) = \alpha(10 \log10 \frac{\sum_{K=0}^{N-1}|X(l,k)|^2}{\sum_{K=0}^{N-1}|D(l,k)|^2})\text{----------}(15)$$

N is the number of DFT spectrum and α is parameter to adjust ṽ.

(3) DEEP NEURAL NETWORK

DNN's applications encompass a variety of fields, including speech enhancement, audio recognition, and others. Because DNN is a feedforward network, it may simulate non-linear relationships. With a set of data that includes both clean and noisy speech, the DNN is trained. To model the DNN, the Log Power Spectra (LPS) feature is taken from the speech signal. The DNN is trained to understand the mapping function and relationship between noisy and clean speech during the training stage, where noisy data with various Signal-to-Noise Ratios (SNR) are taken into account. DNN is commonly used to classify noise, with adaptive filter coefficients chosen according on the level of noise. By reducing interference and distortion, DNN significantly aids in separating the source0signal from the mixed signal.

Figure 5 depicts the system's block diagram, which basically consists of four modules: feature extraction ,training, DNN decoding, and wave- form reconstruction. In the training-stage, stereo data consisting of pairs of noisy and clean speech represented by the log power spectral characteristics are used to rain a regression DNN model. The well-trained DNN model is fed with the characteristics of noisy speech during the improvement stage to0produce the improved log- power spectral features. From the original noisy speech, additional phase information is calculated. The estimated clean speech waveform is finally synthesized using the overlap-add technique.

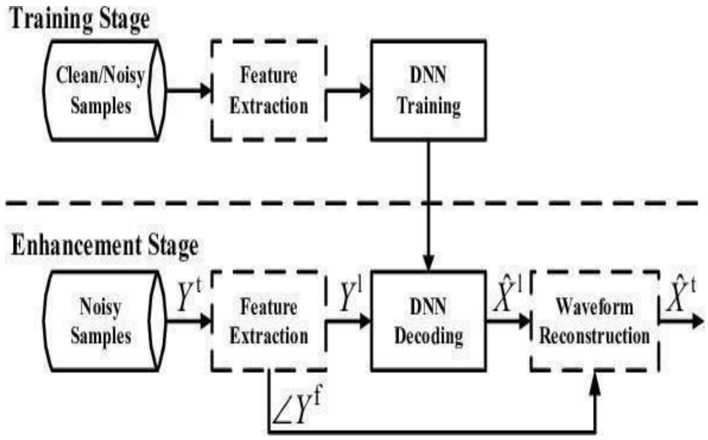


Figure 5 : A block diagram of the DNN-based speech enhancement system

VI. CONCLUSION

A technique called speech enhancement aims to improve the quality of the speech signal. This paper discusses various speech enhancing techniques. We studied several noise sources and noise- removal methods. The research potential for deep neural networks, an emerging method for improving speech, is enormous. The main disadvantage of other0 conventional approaches is the existence of musical noise in improved speech. As evidence that DNNs have the capacity to remember and evaluate the complex properties of the wireless channels, the deep learning approach provides advantages when wireless channels are complicated by significant distortion and interference. According to our research on previous papers for speech enhancement, audio separation, and noise classification, the Deep Neural Network is crucial. In comparison the DNN system will produce better results than the standard MMSE and MAP approaches for speech enhancement.

VII. REFERENCES

[1] Speech enhancement using MMSE estimation and spectral subtraction method by V. K.Gupta, Anirban Bhowmick, 2011.

[2] Real time implementation of Maximum A Posteriori (MAP) based noise reductions using Leon 3 system on chip Muhammad H. Santriaji, Human Mauludin, Dian Surgawiwaha, Trio Adiono, 2014

[3] Speech enhancement using multiple deep neural network Pavan Karjol,M Ajay Kumar, Prashanth Kumar Ghosh, 2018

[4] R. Rehr and T. Gerkmann, "Improving the generalizability of deep neural network based speech enhancement", submitted to IEEE Trans. On Audio Speech and Language Processing, 2017.

[5] S. Mavaddaty, S.M. Ahadi and S. Seyedin, "Modified coherence-based dictionary learning method for speech enhancement", IET Signal Processing, vol. 9, no. 7, pp. 537545, 2015.

[6] Y. Xu, J. Du, L.-R. Dai and C.-H. Lee, "An experimental study on speech enhancement ased on deep neural networks", IEEE Signal processing letters, vol. 21, no. 1, pp. 65-68, 2014.

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

[8] Patrick Schaumont, "Hardware/Software Codesign is a starting point in Embedded Systems Architecture Education" in , ACM, 2004.

[9] S.Pedre, T.Krajnik, E. Todorovich and P. Borensztejn, "A co-design methodology for processor-centric embedded systems with hardware acceleration using FPGA", Programmable Logic (SPL) 2012 VIII Southern Conference on, pp. 1-8, 20– March 2012.

[10] L. Khan, T. T. Jeong, G. Park and A. P.Ambler, "A hw/sw Co-design Methodology: An Accurate Power Efficiency Model and Design Metrics for Embedded System", Software Engineering Artificial Intelligences Networking and Parallel/Distributed Computing 2009. snpd '09. 10th acis International Conference on, pp. 3-7, 27–29 May 2009.

[11] M. Mizumachi and M. Akagi, "Noise reduction by paired microphones using spectral subtraction", Acoustics Speech and Signal Processing 1998.Proceedings of the 1998 IEEE

[12] Y. Kaneda and J. Ohga, "Adaptive microphone array system for noise reduction", Acoustics Speech and Signal Processing IEEE Transactions on ,vol. 34, no. 6, pp. 13911400, Dec 1986.

[13] Benesty, J., Makino, S., Chen, J.D.: Speech Enhancement. Springer, New York, NY (2005).

[14] Loizou, P.C.: Speech Enhancement: Theory and Practice. CRC Press,Boca Raton, FL(2013).

[15] Xu, Y., Du, J., Dai, L.-R., Lee, C.-H.: An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process. Lett. 21(1), 65–68 (2014)

[16] Boll, S.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. ASSP-27(2), 113–120 (1979).