ARTICLE

Analysis of statistical estimators and neural network approaches for speech enhancement

Ravi Kumar Kandagatla*[1], V. Jayachandra Naidu[2], P.S. Sreenivasa Reddy[3], Gayathri M.[1], Jahnavi A.[1], and Rajeswari K.[1]

[1]Department of Electronics and Communication Engineering, Lakireddy Bali Reddy College of Engineering (Autonomous), Mylavaram-521230, Andhra Pradesh, India
[2]Department of Electronics and Communication Engineering, Sri Venkateswara College of Engineering & Technology (Autonomous), Chittoor, India
[3]Department of Electronics and Communication Engineering, Nalla Narasimha Reddy Education Society's group of Institutions, Telangana, India

ABSTRACT

Speech communicated is adversely affected by environmental noise. It is important to process the speech and reduce noise for better understanding. Speech enhancement or noise reduction is useful to provide comfort for human or machine listening. Traditional algorithms provide better noise reduction and better-quality speech. Due to the non-stationary nature of noise and the quasi-stationary nature of speech, the traditional methods are proven inadequate in achieving high-quality speech. Later statistical estimators based on Gaussian, and super-Gaussian Probability Density Function (PDF) assumption further improved the enhancement performance. But still, non-stationary noise nature introduces artifacts in processed signal and results in decreased performance. It is observed that neural network approaches and the factorization approach provide better performance even under non-stationary noises by proper training and large database. Different features result in variations in output performance under unseen noise and speaker conditions. It is important to understand the importance and advantages of traditional methods, statistical estimators, and neural network approaches performances. To select the suitable method for a required application, it is essential to consider the trade-off between quality and distortion. In this work, the importance of speech enhancement methods is discussed. Performance measures used for understanding the speech enhancement like Signal to Noise Ratio (SNR), Segmental SNR, Log-Likelihood Ratio (LLR), Weighted Spectral Slope (WSS), Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), Signal to Distortion Ratio (SDR) and Mean Opinion Score (MOS), are given. Highlights of important results are discussed for analyzing better speech enhancement methods for the required application. In this work, performance is compared using objective and subjective performance measures. Simulation results show superior performance when neural network is employed in statistical estimators.

*Corresponding author
Email Address: 2k6ravi@gmail.com
Date received: May 6, 2023
Date revised: December 21, 2023
Date accepted: January 11, 2024
DOI: https://doi.org/10.54645/202417SupXBB-31

KEYWORDS

Speech Enhancement, Neural Networks, Statistical Estimators

Nomenclature
n –Sample number
l- Time frame
k- frequency bin
t- time
$x[n]$- Clean Speech Signal
$v[n]$- Noise Signal
y[n]- Noisy Speech Signal
$h(t)$ – Convolutive Noise
s(t)- Clean Speech
$Y(k,l)$- Short time fourier transform coefficients of Noisy Speech Signal
$X(k,l)$- Short time fourier transform coefficients of Clean Speech Signal
$N(k,l)$ -Short time fourier transform coefficients of Noise Signal.
$\widehat{S}(\omega)$ – Enhanced clean speech signal in frequency domain
$\hat{H}(\omega)$ – Transfer Function/ Gain
ML- Maximum Likelihood
AR- Auto-Regressive
MISO- Multiple Input Single Output
SNR- Signal to noise ratio
DNN- Deep Neural Network
RNN- Recurrent Neural Network
LSTM- Long Short Term Memory
CNN- Convolutional Neural Network
GAN- Generative Adversrial Network
TF-Time Frequency
STFT- Short Time Fourier Transform
NAT- Noise Aware Training
MFCC- Mel Frequency Cepstral Coefficients
ResNet- Residual Network
SDR- Signal to Distortion Ratio
LLR -Log-Likelihood Ratio
WSS -Weighted Spectral Slope
PESQ -Perceptual Evaluation of Speech Quality
STOI - Short-Time objective Intelligibility

INTRODUCTION

Speech is one of the most organic way of human communication, and a powerful way for people to share ideas or express their wants and emotions. Speech communication is no longer limited to face-to-face interactions. Instead, it can now be performed over large distances via telecommunications and is even employed as a natural method of human-machine connection. Humans rely mostly on voice as one of their primary modes of communication. Speech communication has recently emerged as an essential component of many applications that involve interaction between humans and machines. However, the speech that is conveyed from human to human to machine is distorted by the noise in the environment (the babble, the train, the automobile, the street, and the restaurant) (interfering speakers and so on). Noise reduction / Speech Enhancement primarily aims to increase the quality or standard of degraded speech while simultaneously lowering the amount of background noise. Neural network techniques provide better speech enhancement results. In this work different speech enhancement methods performance is analyzed. The basic speech enhancement process in a speech communication system is shown in (Figure 1). Here, x[n] represent the clean speech and v[n] denote the noise. The effect of noise can be modeled as either additive or correlated. In this work, additive noise is considered, and the performance of the methods is discussed accordingly.

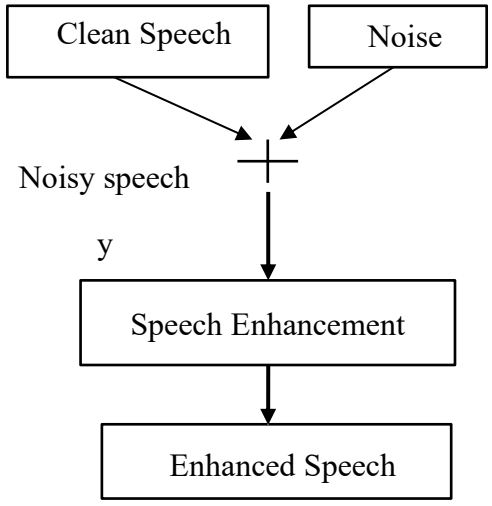

Figure 1: Speech Enhancement in Communication System

SIGNAL MODEL

Noise can easily tamper with speech signals in real-world settings. Reverberations can be classified as either stationary (not changing with time) or non-stationary (changing as time is shifted) noise. Street noise, railway noise, babble noise (other speaker's voices), and instrumental sounds are examples of background noise that fails into the non-stationary category (Zhao Y et al. 2018). The relationship between speech and background noise may be described as follows in the time domain:

$$y(t) = s(t) * h(t) + n(t) \quad (1)$$

Here, $y(t)$ is the noisy speech, $s(t)$ is the clean speech signal, $h(t)$ is the convolutional noise (also known as impulse response of noise), and $n(t)$ is the additive noise.

Using $x(t) = s(t) * h(t)$ as target speech, we can rewrite equation as (2)

$$y(t) = x(t) + n(t) \quad (2)$$

Suppose t is the index of time. The signal can be written as $y = [y(1), y(2), \ldots\ldots, y(T)]$, whereas t is the utterance's duration. We may describe the acoustic signal model of Eq. (2) in time-frequency (TF) domain by using the short-time Fourier transform (STFT).

$$Y(k,l) = X(k,l) + N(k,l) \quad (3)$$

Here, $Y(k,l)$,$X(k,l)$, and$N(k,l)$are the STFT coefficients for the noisy speech signal, clean speech and noise respectively, and k signifies the frequency bin index, l is the time frame index. The aforementioned definitions apply to single channel microphones (Gao T et al. 2016). In this instance, the aim of the SE task is to extract the target speech signal x from the cluttered speech signal y. When it comes to multichannel SE, the signal is stated as Eq. (4)

$$y_m(t) = x_m(t) + n_m(t), \quad m = 1,2,3, \ldots\ldots, M. \quad (4)$$

Here, M represents the total number of microphones in the array.

IMPORTANCE OF TRADITIONAL SPEECH ENHANCEMENT METHODS

The primary objective of speech enhancement is to increase the intelligibility or general perception quality of a distorted or degraded speech signal by applying a variety of algorithms and

audio signal processing techniques (Karjol P et al. 2016). An important area of speech enhancement is the improvement of speech that has been weakened by noise (noise reduction). It is utilized in products like Voice Over Internet Protocol (VOIP), teleconferencing systems, speech recognition software, and mobile phones. The speech enhancement can be achieved by various methods (Xu Y et al. 2015). The basic approach to speech enhancement varies based on the type of noise in the acquired speech signal.

The algorithms of speech enhancement for noise reduction or enhancing the speech by noise can be categorized single channel enhancing techniques and multi-channel enhancing techniques as shown in (Figure 2) .The techniques are categorized as filtering techniques, spectral restoration, and model-based methods as shown in (Figure 2).

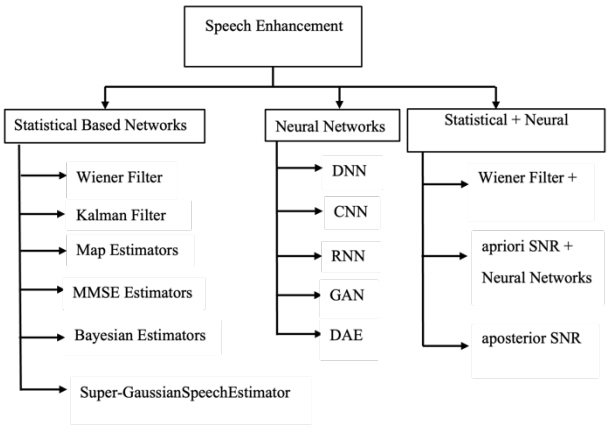

Figure 2: Block Diagram of Neural and Statistical Based Networks

Single Channel Enhancement Techniques

Speech that has been corrupted by noise and captured by a single microphone is the only signal available to single-input speech improvement systems. Basic idea of reducing noise is subtracting noise from noisy speech (noise+clean speech) signal (Zhao Y et al. 2018). But the problem here is estimating the noise to be subtracted. Later speech enhancement systems (Bagchi D et al. 2018) that employ estimation of the signal-to-noise Ratios (SNR) of the corrupted input speech are developed. These methods include wiener filters and statistical approaches (Gao T et al. 2016) which suppress the noise rather than cancel it out. It can be done by providing high attenuation in low SNR regions and low attenuation at high SNR regions.

Wiener Filter

The Wiener or iterative Filter is the speech enhancement algorithm that is most frequently used (Bagchi D et al. 2018) traditional method. In this method, a gain equation based on apriori SNR is obtained. If both the signal and noise estimations are 100% accurate, this method will estimate the enhanced speech by minimizing the mean squared error (MMSE) between the estimated speech signals and clean speech signals (Figure 3). The wiener gain equation given in Eq.5 can achieve a clean or enhanced signal.

$$H(\omega) = \left[\frac{|\hat{S}(\omega)|^2}{|\hat{S}(\omega)|^2 + |N(\omega)|^2}\right] \quad (5)$$

$$\hat{S}(\omega) = H(\omega)S(\omega)$$

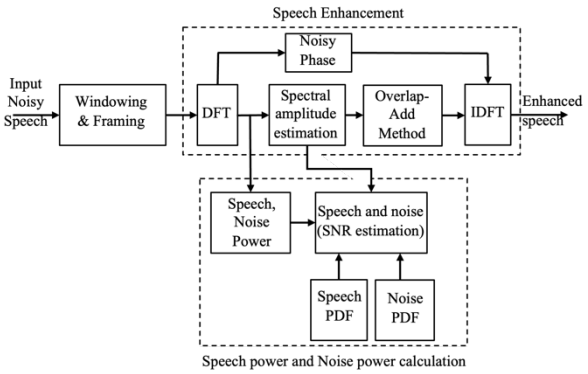

Figure 3: Single Input Speech Enhancement System Block Diagram

Where, S and N be the noise spectra and noise-corrupted speech spectra, respectively, Hdenotes the Wiener or iterative filter, because the spectrum of the Wiener filter has zero phase. The output phase for the calculation of clean signals PDS is the phase from the noise signal. The spectral subtraction algorithms are comparable to this. The spectral subtraction algorithms are comparable to this. The Wiener filter makes the assumption that noise and the desired signal are independent, ergodic, stationary random processes. Speech signals can be divided up into frames to make them appear stationary to tolerate, or hold, their non-stationary, in speech signal processing.

Kalman Filter

Kalman Filter is Linear Quadratic Estimation (LQE) which estimates the required component. It is an algorithm that makes use of series observations made overtime, including statistical noise when a direct measurement is additionally, it is used to aggregate data from several sensors when noise is present in order to identify the best estimate states (Figure 4). Equations are used to explain both the observation model and the state process model,

$$x(m) = Ax(m-1) + Bu(m) + e(m) \quad (6)$$
$$y(m) = Hx(m) + n(m) \quad (7)$$

In control applications, the control vector u(m) is utilized, whereas x(m) is the P-dimensional signal at time m, A is a P*P dimensional state processes at times m and m-1, B is the Matrix of control.

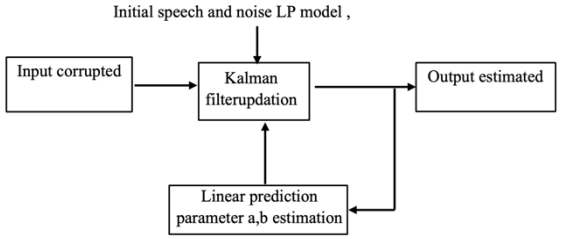

Figure 4: Speech Enhancement using Kalman Filter

Minimum Mean-Square-Error (MMSE) Estimation Methods

Short-time Spectral Amplitude MMSE Estimate (STSA) is implemented in MMSE estimation approaches. It estimates real or complex spectral amplitudes for optimized enhancement. To provide the best performance, two approaches are employed to estimate SNR for each frequency: Maximum likelihood and decision-directed approach. The ratios of noisy signal to instantaneous SNR (power of noise) and half-wave rectification are used in the maximum likelihood (ML)

approach to estimate SNR, and the result is non-negative. By weighing the average of this highest probability estimate, decision-directed techniques calculate SNR. Both methods presumptively know the mean noise power spectrum beforehand. Cohen suggested changes to the decision-directed technique can increase the performance and demonstrate the speech's delayed response.

Restoration

Restoration or model based speech enhancement methods makes use of a speech explicit stochastic model, as well as some prior knowledge circumstances, interfering noise(i.e. it is a tool for estimating probability distributions of outputs for random fluctuations or variations in one or more inputs overtime). There are numerous speech models that combine hidden Markov models, models of coefficient , autoregression (AR) models, and pitch track models. Speech enhancement techniques that use an AR model of speech often impose no restrictions on the estimated set off AR coefficients other than stability.

Multi-Channel Enhancement Techniques

Multiple microphones are used in multiple-input speech enhancement or augmentation systems to pick up the signals containing both the speech and noise signals. Examples include beam-forming, multiple-input multiple-output (MIMO), adaptive noise cancellation systems. For optimal performance, the microphones in multiple-input systems can be setup symmetrically. This is helpful for applications such as teleconference systems, in-car communication devices, hearing aids, and automated speech recognition. Several input noise reduction systems eliminate distortions brought on by interfering voice, echo, noise, and room reverberations by filtering multiple noisy signals utilized with microphones with desired signals.

Adaptive Noise Cancellation

It is a different method of accessing signals that have been distorted by additive noise or other types of signal interference (Figure 5). The key benefit is that degrees of noise rejection that other signal processing would be very difficult or impossible to acquire by other noise removal techniques used in signal processing are attainable or reachable with no priori estimates of noise or signals.

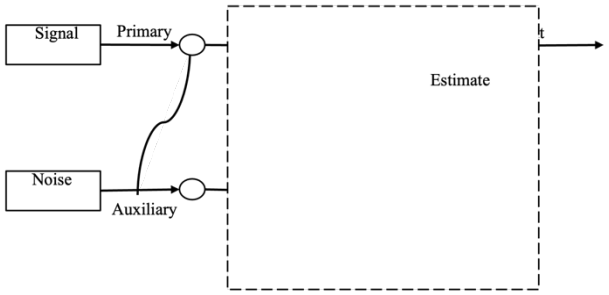

Figure 5: Adaptive Noise Cancellation

Beam-forming

A multiple-input and single-output (MISO) is an application of beam-forming. It consists of multiple channels employing recent multidimensional or algorithms for space-time domain filtering, which increases the desired signal while concurrently reducing noise signals. When beam-forming, two or more microphones are arranged in a geometric array. Additionally, based on the signals directions of arrival, it attenuates the signals and filters the sensor outputs (DOA). The underlying concept of this method is based on the presumption that offering reflections is small, making it possible to combine it with acoustic feedback cancellation. Additionally, if the desired signal's direction is known, it can be combined with acoustic feedback cancellation to form the correct alignment of the phase function present in each sensor. Beam-forming has uses in teleconferencing, in-car communication, strong speech recognition and voice communication over personal computers, among other hand-free communications scenarios.

Multi-input multi-output (MIMO)

The adaptive noise cancellation, adaptive beamforming, teleconferencing systems with multi-input multi-output (MIMO), stereophonic echo cancellation, and in-car MIMO communication systems are examples of speech enhancement systems using multiple inputs.

There are several microphones present in the speech enhancement system. Each microphone's output consists of a combination of voice signals, loudspeaker feedback, wall reflections, and noise (Figure 6). When there are M microphones and N sets of signal and noise sources, there are N×M unique acoustic channels between the microphones and the signal sources. The correlation of the signals emitted from the various sources $x_i(m)$, and those detected by the microphones $y_j(m)$is described by linear equations as

$$y_j(m) = \sum_{i=1}^{N} \quad \sum_{k=0}^{P} \quad h_{ij}(k)x_{ji}(m-k)j = 1, \ldots . M. \quad (8)$$

using finite impulse response (FIR) linear filter, the channel response from source i to microphone j is represented by $h_{ij}(k).x_i(m)$ represents the signal and noise sources, and m is the discrete-time index.

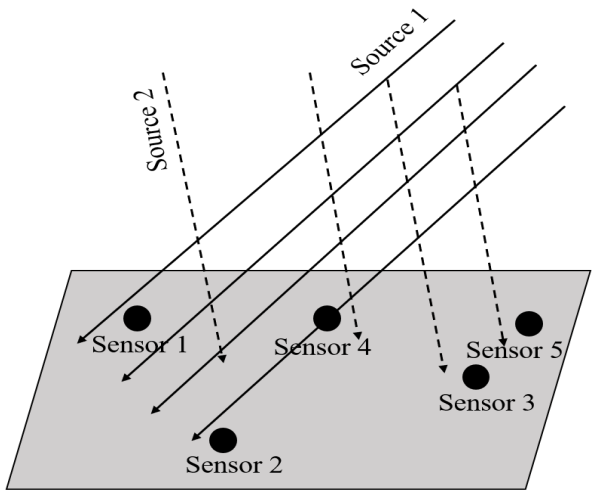

Figure 6: Array of sensors

Neural Network Approaches for Speech Enhancement

An artificial intelligence method known as a neural network tells computers or other machines how to interpret data in a way that is comparable to how human brains do it. Deep learning is a type of machine learning that imitates the human brain by using interconnected neurons or nodes in a layered framework. It creates an adaptable technique that enables a computer or other machine to continuously learn from its mistakes. Therefore, this artificial neural network makes an effort to find solutions to the challenging issues. These neural networks are used for the purposes of applications like medical diagnosis, financial predictions by processing the historical data, chemical compound identification, and also for process and quality control.

Acoustic Features:

Feature	Methodology
Time domain Speech waveforms (.wav)	20 ms frame duration and 10 ms frame shift.
Mel Frequency Cepstral Coefficients- MFCC and log MFCC	Representation as short term power spectrum
NAT (Noise Aware Training)	Noisy periodogram output is used to obtain features.
SNR-NAT based features (Noise Aware Training)	Logarithm of a-priori and a-posteriori SNRs are considered as features
Pitch	Signal is passed through Gammatone filter bank and each sub signal is applied to PEFAC algorithm for pitch tracking.
Log-Magnitude Spectrum	Features that are obtained after operation log (Magnitude(FFT(Noisy Speech)) operation.
Gabor filter bank feature	Filter bank separates into different bands and process

Different features are used in a variety of speech processing contexts are given below. Each feature is briefly described in this section.

Waveform signal (WAV)
In automated speech recognition, waveform signals may be utilised directly without any extraction of features, as demonstrated by (Muhammed Shifas PV et al. 2021) (ASR). We simply employ 320 signal samples with a 160 sample shift, or 20ms time frames with a 10ms frame shift, to evaluate this feature in our system.

Pitch-based feature (PITCH)
In several research on speech separation, pitch has been used as a crucial cue for acoustic scene interpretation. The noisy signal is sent through a 64-channel gammatone filterbank to determine the pitch-based feature. Following that, each sub-band signal is subjected to the PEFAC pitch tracking algorithm. We extract 6 dimensional characteristics using the observed pitch in the manner reported in (Wany Y et al.2013). We combine all 64 channel's features in the end.

Log-magnitude spectral feature (LOG-MAG)
The noisy voice spectrogram serves as the basis for computing the LOG-Mag feature. Particularly, the magnitude responses of the STFT are subjected to a log operation.

Perceptual linear prediction feature (PLP)
PLP focuses on suppressing individual speaker-specific characteristics within a spectrum. The PLP is derived by utilizing the power spectrum, which is calculated, translated to the bark scale, downsampled, and filtered using a critical band masking curve. Subsequently, the downscaled spectrum undergoes pre-emphasis, followed by compression through an intensity-loudness power law and a curve operation involving cubic roots. The resulting spectrum is subjected to a Discrete Inverse Fourier Transform (DIFT). The final cepstrum is obtained by solving for the Twelfth-order autoregressive coefficients, which are subsequently converted to the cepstral all-pole model coefficients.

Amplitude modulation spectrogram (AMS)
AMS's primary aim to degrade the full wave corrected envelop of the noisy signals by a factor of 4. The signal is then divided into 32 ms frames and given a 10 ms frame shift. A 256-point FFT is used after the signal in each frame is windowed using a Hann function. The modulation responses are compounded by 15 evenly centred triangular windows between 15.6 and 400 Hz. The AMS feature is made up of the 15 replies that followed (Morten Kolkek et al. 2016).

Gabor filter bank feature (GFB)
A bank of 41 spectro temporal Gabor filters are used to process each of the sub-band signals in the log-mel-spectrogram of the mixed signal. The correlations between the feature components are then reduced by carefully choosing a subset of the channels.

Mel-frequency cepstral coefficients (MFCC)
A common aspect in voice processing is MFCC. The spectrogram of the input signal is computed in order to calculate MFCC. The power spectrum is then compressed and transformed to the mel scale. Eventually, use DCT and the MFCC feature is represented by the first 31 cepstral coefficients.

Log-mel-filter bank feature (LOG-MEL)
The LOG-MEL characteristic is commonly employed in ASR and speech separation. A 40-channel mel filter bank processes the spectrogram of the mixed signal. The outcome of the LOG-GEL feature is log operation.

Relative autocorrelation sequence MFCC (RAS-MFCC)
RAS-MFCC computes autocorrelation sequences for each time frame to give a noise-resistant feature. A high-pass filter is then used. The MFCC extraction procedure receives the filtered sequences as input, producing the RAS-MFCC feature.

Phase autocorrelation (PAC-MFCC)
The phase trajectory of the signal over time is the foundation of PAC-MFCC. The MFCC method is used to calculate the PAC-MFCC by calculating the phase angle between the noisy signal and its shifted variants.

SNR-NAT features
SNR NAT features are developed based on apriori SNR and a posteriori SNR. The utilization of these features in a neural network provides better results.

Deep learning methods
We examine the latest deep learning approaches designed to address the speech enhancement (SE) model problem, such as DNN, DAE, RNN-LSTM, CNN, and GAN.

Based on Deep Neural Network (DNN)
One of the most popular models for SE is the Deep Neural Network (DNN), also known as the Feedforward Fully Connected Layer or the Multilayer Perceptron (MLP) with Multiple Hidden Layers (Karjol P et al. 2016). Because every node in the layer has a connection to every other node in the layer above, the network is known as a fully connected network. DNN has relatively huge parameters as a result. A voice improvement method employing several DNN-based systems was developed in (Karjol P et al. 2016). Each of the DNNs employs a gating network, which assigns weights to aggregate the outputs of the n DNN and contribute to the final improved speech (Figure 7). The model $n = 4$ layers with a depth of three each. On the TIMIT corpus, an average SNR of -5 to 10 dB yields a visible noise PESQ of 2.65 and an unseen noise PESQ of 2.19. In contrast, the DNN masking-based techniques in (Xu Y et al. 2015) might achieve a PESQ of 2.705 or higher. To further develop DNN, a speech intelligibility metric was included in (Zhao Y et al. 2018). The results revealed an average PESQ of 1.99 and an SNR that was under matched or mismatched. Another publication by (Bagchi D et al. 2018) also additionally tried to improve the model by

training the speech enhancer using both mimic loss and the conventional criterion. The mean square deviation of the results of two spectral classifiers is known as the mimic loss. The enhanced speech that DNN generates typically deteriorates in low SNR scenarios, despite the fact that it has been effectively utilised as a regression model foe SE. A framework for progressive learning for DNN-based SE was developed by (Gao et al. in 2017). The WSJ corpus was used to train the model on multi-SNR and single-SNR settings, and it was tested with heard and unseen noises including babbling, factory, and destroy engine. PESQ scores of 1.93 for training with a single-SNR and 1.82 for training with multi-SNR were averaged as a result (SNR -5 and 5dB).

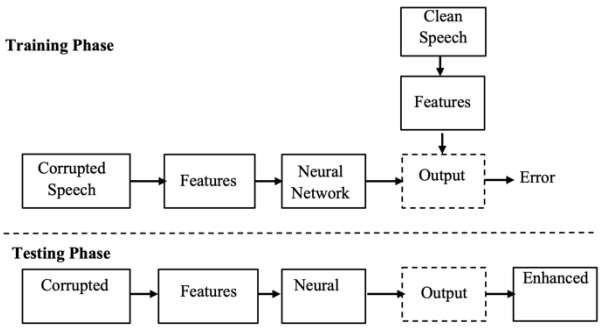

Figure 7: Block Diagram of DNN Based SE system

Based on Denoising Autoencoder (DAE)

In the majority of work (Feng X et al. 2014), DNN-based deep autoencoder with same input and output dimensions. For representation learning, deep autoencoders are frequently employed. The Denoising autoencoder is a denoising criteria used in the spectral mapping approach (DAE), was first presented by (Lu et al. 2014) then (Feng X et al. 2014) extended this model to include deep DAE . The model performed a mapping from noise to cleanness. Mel-spectral speech is spoken. DAE is initially trained to convert the damaged input y into a concealed representation $z$created with the encoder specified in Eq. (9).

$$z = \sigma(Wy + b), \tag{9}$$

Here, σ represents the non- linear function of activation. Weighted matrix and bias vector, respectively are denoted by W and b. The output representation z is then changed back into an input y that has been rebuilt using the decoder described in Eq. (10)

$$\acute{y} = \sigma(Wz + \acute{b}) \tag{10}$$

Here,W and $\acute{b}$ are, respectively, the suitable scaled parameters of W and b. The loss of MSE between the input y and its recreated input y is used to train the DAE. However, the DAE network is limited in its ability to learn short-term information. Consequently, it is common practice for the network's training with a limited condition window. The temporal issue has recently been addressed using DAE and convolutional layers.

Based on Recurrent Neural Network (RNN)-Long short-term memory (LSTM)

There are recurrent neural networks (RNN) and Long short-term memories (LSTM) that can manage context information when working with a sequence- based data, such as a speech signal. This network uses information from the preceding hidden layer in addition to the current input. The inspiration for more recent work by (Gao T et al. 2018) came from curricular learning. To enhance the performance of DNN-based speech in low SNR environments, they suggested a progressive learning architecture using the LSTM network. The last layer of each target layer is designed to learn smoother transition in speech, with improved SNR. Additionally, LSTM- RNN has been used to solve the issues with reverberation, multichannel noisy speech (Kinoshita K et al. 2020), and extremely non-stationary additive noise (Wollmer M et al. 2013). In (Wollmer M et al. 2013) bottleneck features produced by the bidirectional LSTM network surpass manually created features like MFCC (BN-BLSTM). While employing MFCC, using BN-BLSTM, the WA is 43.55% compared to the average WA of 38.13%. The use of LSTM-RNN has significantly enhanced speech processing systems. However, it is well known that learning the RNN parameters is challenging and time-consuming or takes a lot of resources.

Based on Convolutional Neural Network (CNN)

Convolutional neural network (CNN) has drawn a lot of interest from speech researches (Pandey A et al. 2019). CNN can use a set of local connections to detect patterns in the adjacent frames using spectrogram characteristics. Additionally, it has reportedly been shown to be more efficient than a typical feed-forward neural network and more effective than RNN. Additionally, here are more recent CNN-based SE for ASR application works (Rownicka J et al. 2020), (Kinoshita K et al. 2020). CNN- Based voice denoising based on masking estimation is proposed (Kinoshita K et al. 2020). This approach draws inspiration from the efficiency of temporal convolutional networks for voice separation (Conv-TasNet). They modified the network design for the Denoising TasNet task, which is carried on in the TF and temporal domains. This study also looked into multi-task loss, which predicted speech and noise as two outputs. The network with multi-task loss performs best in the time domain. Additionally, a suggested extension of CNN that makes use of the residual network (ResNet) has been proposed (Kinoshita K et al. 2020). Since ReNet's architecture is compatible with the SE task of reconstructing the input signal by removing the residual noisy signal, an enhanced result can be obtained.

Additionally, (Rethage D et al. 2018) suggested an end-to-end learning method for voice denoising that makes advantage of direct waveform processing. In (Rethage D et al. 2018) the new WaveNet network structure served as the model's foundation. The network is made up of a number of dilated, non-causal convolutional layers that learn under supervision by minimising regression loss. The receptive field created by the dilation parameters can greatly reduce the computational complexity. The total outcomes demonstrate that, with a 23% relative improvement in MOS quality, CNN is superior to a traditional Wiener filter.

Based on Generative Adversarial Network (GAN)

To further boost the performance of model enhancement, Generative adversarial networks (GAN) have attracted increasing attention. GAN consists of a generating network (G) plus a discriminator network (D). GAN training, convolutional layers are frequently used (Li X and Horaud R 2019),(Kinoshita K et al. 2020) or fully connected layers (Donahue C et al. 2018). (Rownicka J et al. 2020) the one who initially presented speech improvement based on GAN training (SEGAN). The generator network gains the ability to translate characteristics of loud speech into clean speech. Following that, the binary classifier discriminator network decides whether the samples come from the clean voice (real) or the improved speech (fake).

The generator tries to modify the distribution to produce better outputs based on the results of the discriminator until the discriminator is unable to tell whether the outputs are real or fake. GAN training, however, is challenging and unstable (Figure 8). Numerous more efforts were made to enhance

SEGAN's performance (Li X and Horaud R 2019), (Kinoshita K et al. 2020), (Donahue C et al. 2018). By (Baby D et al. 2018) implemented a gradient-penalized relativistic loss function at the discriminator network. This research demonstrated that a cleaner speech might be produced by a better discriminator. The approach additionally used gradient penalties to stabilise the training. By (Phan H et al. 2020) suggested to utilise more than one generator rather than just one. The main objective was to do multi-stage improvement mapping gradually. In terms of PESQ, CSIG, CBAK, COVL, and SSNR, the suggested approach out performed SEGAN. While (Xu Z et al. 2020) made an effort to change the SEGAN architecture in order to provide features. Because it is designed for ASR applications, unlike SEGAN, the work uses log-Mel characteristics rather than waveforms in its implementation.

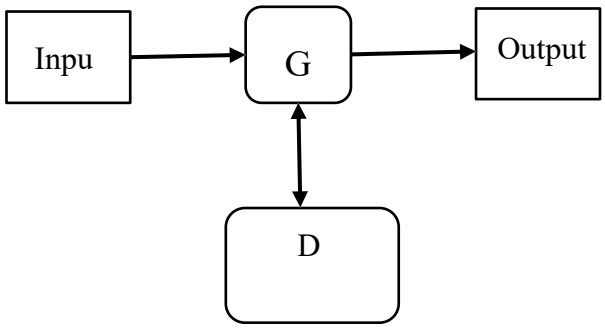

Figure 8: Architecture of a GAN for Enhancement of speech

Cross guided CNN based neural network is developed in (Zhang W et al. 2023) for reduced loss. To reduce the mismatch between testing and training the generalized networks are developed (Gonzalez et al. 2023). Composite cost function based wiener filter is developed to reduce noise (Thaleiser S and Enzer G 2023). To focus on wide band speech signals and its enhancement neural network structure is proposed (XU Z et al.2023).

Some works (Phan H et al. 2020), (Soni MH et al. 2018) attempted to apply GAN in the masking-Based technique, even though it is becoming more and more popular in the mapping-based method. GAN is used in (Phan H et al. 2020) to anticipate the mask. The vanilla GAN is enhanced with a regularised objective MSE function. The outcomes demonstrate that PESQ and STOI outperform a recent GAN-based voice enhancement. Various approaches and the outcomes are listed in Table 1.

Table 1: Summary of Neural and Statistical Based Networks methods for Speech Enhancement

Method	Features	Dataset	Evaluation Metrics	Results
DNN	CEGM	CHiME-4	WER	WER: 19.39%(Noisy), 25.70(OM-LSA), 24.46(DNN)
	Dropout, Global Variance Equalization	Aurora2, TIMIT, NOISEX-92 corpus	PESQ, LSD, SSNR	PESQ: 3.39(RBM), 3.40(Random) LSD: 2.90(RBM), 2.27(Random) SSNR: 8.42(RBM), 8.44(Random)
	MFMPDR, SPP, IFC	WSJ0 corpus, NOISEX92, Aurora	PESQ	PESQ: 0.22(MFMPDR), 0.28(WG)
	SE Algorithm	Akustiske Database for Dansk, TIMIT	STOI, PESQ, BBL, SNR	STOI: 0.023, PESQ: 0.878, BBL: 0.033, SNR: 0.00(0.968)
	IMCRA, BGRU	CHiME-4	WER	WER: 19.98%
	(Log/power) mag.	NOISEX + IEEE Corpus	PESQ, SDR, STOI	Average results with mismatched SNR (-3 to 3 dB) PESQ is 1.99, SDR is 11.35, and STOI is 90.61%.
	(Log/power)	TIMIT + noises from Aurora	Seg SNR, STOI, PESQ	Average best PESQ scores for seen noise are 2.65 and unseen noise are 2.19.
	LPS	WSJ+ musical noises	SSNR, STOI, PESQ	PESQ 1.93 was evaluated on unseen noise, while PESQ 1.82 was used for training with multiple SNRs.
DAE	MFCC	CHiME-2	WER	34% Error rate.
RNN-LSTM	LPS (log-power spectral)	WSJ + surrounding and musical noises	SDR, STOI	SDR of average results: 9.46 and STOI: 0.86
	MFCC, Bottleneck features (BN)	Spontaneous speech + CHiME noises	WA	BN-BLSTM: 43.55%, average WA using MFCC: 38.13%.
	Raw signal	SSN + TIMIT +NOISEX	SI-SDR, PESQ, STOI	Results indicate that Autoencoder CNN better performed SEGAN in terms of performance.
	(Log) Mel	Aurora-4, AMI	WER	8.31%(WER) on Aurora-4
	Raw Signal	DEMAND + Voice Bank	MOS, SIG, BAK, OVL	3.60 MOS is attained. When compared to the Wiener filter, overall outputs are superior.
CNN	LPS, Log Mel-filterbank	HINT, TIMIT, LibriSpeech corpus, DCASE 2017	PESQ, LSD, SegSNR	PESQ: 1.6024 LSD: 12.0219 SegSNR: 5.8420
	(Log/power)mag, raw signal	CHiME-4,Aurora-4	WER, SDR	WER 8.3% (actual data) and 10.8% (simulated), SDR: 14.24, and 6.3% for

				CHiME-4.
	(Log/power) mag	Grid Corpus +CHiME-3 noises	PESQ, STOI	For heard noises, PESQ is 2.60 and STOI is 0.70, and for unknown noises , it is 2.63 and 0.74, respectively.
CNN- E2E	SSDRC, FFTNet	Raw Speech Samples	SIIB	SIIB: 12.81(Unprocessed), 18.65(MBSSDRC), 23.48(DnsFFTNet+SSDRC), 26.56(Casual FFTNet), 34.51(non-casual FFTNet)
GAN	CGMM, MT-GAN, IRM	TIMIT	SSNR, PESQ, STOI	SSNR: -11.6167PESQ: 0.1369 STOI: 0.1046
	Raw Signal	Voice Bank + DEMAND	PESQ, CSIG, CBAK, COVL, STOI, SSNR	2.39 (PESQ), 3.55 (CSIG), 3.11 (CBAK), 2.93 (COVL), 8.72 (SSNR)
	(Log) Mel	WSJ + surrounding and musical noises	WER	17.6% Error rate
	Raw Signal	DEMAND + Voice Bank	Seg SNR, STOI, PESQ	Seg SNR:17.68, STOI: 0.942, PESQ:2.62
HNN	CNN, LTSM, PSM	TIMIT, IEEE Corpus, NOISEX-92	PESQ, SSNR	PESQ: 1.59SSNR: 1.03
ERNN	Speech PSD Uncertainty, posteriori SPP	DEMAND + Voice Bank	PESQ, CBAK, CSIG, COVL	PESQ: 2.49, CBAK: 3.02, CSIG: 3.63, COVL: 3.03
MMSE	MOSIE, Super-Gaussian Speech Estimator	TIMIT	PESQ	PESQ: 0.43(NON-MLSE), 0.41(DNN), 0.48(NMF)

Challenges of Speech Enhancement

Recent advancements in artificial intelligence (AI) and machine learning (ML) have yielded excellent results in addressing speech-related problems. These developments demonstrate the capability to effectively eliminate various types of background noise, such as dog barking, kitchen noise, environmental sounds, music, babbling, and traffic, showcasing the enhanced performance of speech-related applications. Compared to conventional statistical signal processing methods, which typically only effectively attenuate quasi-stationary noise, this is an interesting novelty. However, ML-based speech improvement still has a long way to go before it is sufficiently developed to be commercialized, and it must overcome the following challenges:

1. Speech Quality: Although AI-powered voice improvement has excellent suppressing capabilities, speech quality frequently suffers as a result. More study is focused on enhancing data collection and expansion, investigating optimization objectives, and enhancing network models in order to improve voice quality. For example, we employed convolutional-recurrent network topologies for voice augmentation in one of our early experiments.

2. Inference efficiency: Very massive neural network models, which have prohibitively high inference complexity and occasionally include processing latency, are frequently used to produce great audio quality. In order to execute these models on edge devices with limited resources, research is actively being done to lower the model size, complexity, memory footprint, and processing time. For enhancement of speech and voice activity identification, we have previously investigated bit-precision scaling as a method of improving model efficiency. We also looked into tiny recurrent networks to see whether they might be improved to meet real-time interference restrictions.

3. Unsupervised learning: Supervised learning produces the most effective results in machine learning i.e. In order to train a speech enhancement model, a dataset comprising both clean and noisy target speech must be prepared. The drawback of this is that a robust dataset for all scenarios experienced in reality would need a lot of work. However, situations for which you are not prepared. As a ground truth is not necessary and theoretically, a model may be developed to adjust to invisible noise on-the-fly, unsupervised learning could assist to solve this issue. Recurrent networks were used in our initial effort to adapt a voice enhancement algorithm to the input data via reinforcement learning.

EVALUATION METRICS

Subjective and Objective measures can be used to evaluate the impact of Speech Enhancement (SE) systems.

Subjective Measures: Objective assessments include unbiased measurements of signal distortion, mean opinion score (MOS), and background noise intrusiveness (BAK). A scale from 1 to 5 is used for SIG, BAK, and MOS, with a higher number being preferable.

Objective Measures: The segmental Signal-to-noise ratio (Seg-SNR), Signal-to-Noise ratio (SNR), distances measures, source-to-distortion ratio (SDR), perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), Log-Likelihood Ratio (LLR) and Weighted Spectral Slope (WSS) are examples of common objective metrics. The discuss of all the performed measures indicates follows:

Signal-to-Noise Ratio (SNR):

The ratio of signal to noise power is Known as the signal-to-noise (SNR) ratio and is measured in decibels (dB).

Segmental Signal-to-Noise Ratio (Seg SNR):

Average value of each SNR ratio or the average SNR values of brief parts are calculated rather than the entire signal (15-20ms) is known as Segmental SNR Ratio.

In order to calculate the segmental SNR_{seg} measure,

$$SNR_{seg}(dB) = \frac{10}{M}\sum_{m=0}^{M-1} \frac{\sum_{n=Lm}^{Lm+L-1} x^2(n)}{\sum_{n=Lm}^{Lm+L-1} [\hat{x}(n)-x(n)]^2} \quad (11)$$

Where $x(n)$ is the clean speech and $\hat{x}(n)$ is enhanced speech i.e. Processing signal or approximate clean speech, M is the total number of frames in the signal and L is frame length.

Perceptual Evaluation of Speech Quality (PESQ):
PESQ is the series of standards that includes a test process for an automated evaluation of the speech quality as perceived by a telephone system user. PESQ is used by telecom operators, phone manufacturers for objective voice quality testing. The higher the value of PESQ, which ranges from -0.5 to 4.5, the better the quality of the speech.

Short-time Objective Intelligibility (STOI):
A measure of intelligence that has a strong correlation with the intelligibility of speech signals that have been damaged by additive noise, single or multi-channel noise reduction, binary masking. The STOI-measure, which is a function of the clean and degraded speech signals, is invasive.

Signal-to-Distortion Ratio (SDR):
When determining how much distortion is present in a signal, the Signal-to-Distortion ratio is employed as a metric of the signal's quality. It is a measure of how much distortion is present in relation to the original, undisturbed signal power. Typically, the SDR formula compute as:

$$SNR(dB) = 10\left(\frac{P\_signal}{P\_distortion}\right) \quad (12)$$

Here, Power of the original signal, and undistorted signal represents the P_signal and Power of the distortion existing in the signal represents the P_distortion. It is expressed as in decibels (dB), and the greater the SDR, the higher the quality of the signal. A greater SDR represents that the original signal is being precisely retained and that there is less distortion of the signal and a lower SDR represents that there is more distortion in the signal and that the original signal is being more distortion.

Log-likelihood Ratio Measure (LLR):
Log-likelihood Ratio objective measures are used to assess the effectiveness of the proposed technique for each speech frame.

$$d_{LLR} = \left(\frac{a_y R_s a_y^T}{a_s R_s a_s^T}\right) \quad (13)$$

Where is the Linear Prediction Coefficient (LPC) vector of the unaltered (original) speech frame, represents the autocorrelation matrix, and is the Linear Prediction Coefficient (LPC) vector of the improved (estimated) speech frame. The individual frame LLR data were averaged to provide the final average LLR value.

Weighted Spectral Slope (WSS):
Weighted spectral slope is computed as

$$d_{wss} = \frac{1}{M}\sum_{m=0}^{M-1} \frac{\sum_{j=1}^{k} w(j,k)\left(s_c(j,k) - s_p(j,k)\right)^2}{\sum_{j=1}^{k} w(j,k)} \quad (14)$$

Where, $w(j,k)$ represents the corresponding weights,j denotes the frequency, k denotes the number of frames. $s_c(j,k)$ and $s_p(j,k)$ represents the spectral slopes for j^{th} frequency at the k number of original speech signal and processing or enhanced speech signal respectively.

All of these actions are taken to analyze the clarity and quality of the speech word. Aside from that, word accuracy (WA) or word error rate (WER) is a common metric to specifically assess the effectiveness of ASR systems.

CHALLENGES AND OPPORTUNITIES

The meaning of voice or speech enhancement is continuously becoming more general as the applications are expanded. (Figure 9) indicates the categorization of speech enhancement methods and (Figure 2) shows statistically and neural networks based techniques are the fundamental foundations for the categorization. It should obviously incorporate the signals reinforcement from the deterioration of competing speech or even from the degradation of the signals filtered from. More difficult than the traditional noise reduction challenge are the signal separation and dereverberation problems. The voice that a talker emits and that a microphone picks up in a room and in a hands-free situation comprises signal in both the direct and delayed paths copies and once the source signal has been attenuated due to reflections from the objects and room limitations. Reverberation, also known as spectral distortion and echo, will be introduced into the observation signal as a result of these multipath channels. The ability of people who have normal hearing to recognize and comprehend one speaker among many other speakers, or in the middle of a cacophony of background noise and discussion, must be determined. The cocktail party effect is the capacity to sort and choose one speaker's speech patterns from among many others. Hence, the impact of a cocktail party. Although it is widely known that listening in these circumstances with one ear is uncomfortable and that it is challenging to focus on one specific signal when several voice signals are present all around at once, it will be done. When doing blind source separation with several microphones, we have a tendency to try to distinguish between sounds that are originating from entirely opposite directions at the same time. The cocktail party effect may be able to distinguish the relevant signal from the background noise, which is not exactly what blind source separation techniques achieve.

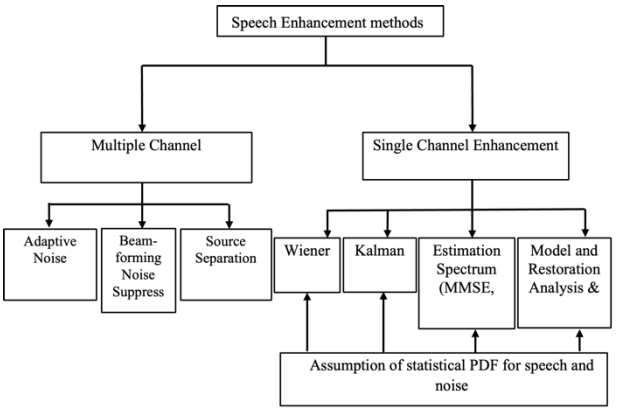

Figure 9: Traditional Speech Enhancement Methods

CONCLUSION

Outline of speech enhancement techniques based on statistical approaches and neural network approaches are discussed, together with their advantages and disadvantages. It is discussed how classical, statistical estimators and neural networks advanced the speech processing and quality improvement. In this work the choice of features, neural network models and performance measures for required application is discussed.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

CONTRIBUTION OF INDIVIDUAL AUTHORS

R.K. Kandagatla, V.J. Naidu, P.S. Sreenivasa Reddy were involved in design of study analysis and interpretation of data and manuscript revision. Gayathri M, Jahnavi A and Rajeswari K were involved in conception, design of study, acquisition, analysis of data and writing and drafting of manuscript.

REFERENCES

Bagchi D, Plantinga P, StiffA, Fosler-lussier E. Spectral feature mapping with mimic loss for robust speech recognition. 2018 IEEE Int. Conf. Acoustics, Speech and signal Processing Proc. 2018.

Baby D, Verhulst S. Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty. IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc. 2018.

Chai Li, Jun Du, Qing-Feng Liu, and Chin-Hui Lee. A Cross-Entropy-Guided Measure (CEGM) for Assessing Speech Recognition Performance and Optimizing DNN-Based Speech Enhancement. IEEE/ACM Trans. Audio Speech Lang. Process 2020;29(1):106-117.

Donahue C, Li B, Pranhavalkar R. Exploring speech enhancement with generative adversarial networks for robust speech recognition. IEEE Int. Conf. Acoustics Speech and Signal Processing Proc., 2018.

Feng X, Zhang Y, Glass J. Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. IEEE Int. Conf. Acoustics Speech and Signal Processing Proc., 2014.

Gonzalez P, Alstrøm TS, May T. Assessing the Generalization Gap of Learning-Based Speech Enhancement Systems in Noisy and Reverberant Environments. IEEE/ACM Transactions on Audio, Speech, and Language Processing 2023;31(1):390-3403.

Gautam Bhat S, Nikhil Shankar, Chandan Reddy AA, Issa Panahi MS. A Real-Time Convolutional Neural Network Based Speech Enhancement for Hearing Impaired Listeners Using Smartphone. IEEE Access 2019;7:78421-78433.

Gao T, Du J, Dai LR, Lee CH. Densely connected progressive learning for LSTM- based speech enhancement. IEEE Int. Conf. Acoustics Speech and Signal Processing Proc., 2018; 5054-5058.

Gao T, Du J, Dai LR, Lee CH. SNR-based progressive learning of deep neural network for speech enhancement. Proc. Annu. Conf. Int. Speech Communication Association Interspeech. 2016.

Jing Yuan, Changchun Bao. Multi-Channel Speech Enhancement with Multiple-target GANs. 2020 IEEE Int. Conf. on Signal Processing, Communication and Computing. 2020.

Karjol P, Kumar MA, Ghosh PK. Speech enhancement using multiple deep neural networks. 2018 IEEE Int. Conf. Acoustics, Speech and signal Processing Proc. 2018.

Kim G, Lu Y, Hu Y, and Loizou PC. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. J. Acoust. Soc. America. 2009; 126(3):1486-1494

Kinoshita K, Ochiai T, M. Delcroix, and T. Nakatani. Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc. 2020.

Li X and Horaud R. Multichannel speech enhancement Based on time-frequency masking using subband long short-term memory. IEEE Workshop Applications Signal Processing to Audio and Acoustics 2019.

Mojtaba Hasannezhad, Zhiheng Ouyang, Wei-Ping Zhu, Benoit Champange. Speech Enhancement with Phase Sensitive Mask Estimation Using a Novel Hybrid Neural Network. IEEE Open Journal of Signal Processing 2021;2:136-150.

Marvin Tammen, Dorte Fischer, Bernd Meyer T, Simon Doclo. DNN-Based Speech Presence Probability Estimation for Multi-Frame Single-Microphone Speech Enhancement. 2020 IEEE Int. Conf. Acoustics Speech and Signal Processing Proc. 2020.

Muhammed Shifas PV, Catalin Zorila, Yannis Stylianou. End-to-End Neural Based Modification of Noisy Speech for Speech-in-Noise Intelligibility Improvement. IEEE/ACM Trans. Audio Speech Lang. Process 2021; 30(1):162-173

Minseung Kim, Jong Won Shin. Improved Speech Enhancement Considering Speech PSD Uncertainty. IEEE/ACM Trans. Audio Speech Lang. Process 2022;30(1):1939-1951

Morten Kolkek, Zheng-Hua Tan, Jesper Jensen. Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems. IEEE/ACM Trans. Audio Speech Lang. Process 2016;.25(1):153-167.

Panagiotis Tzirakis, Anurag Kumar. Multi-Channel Speech Enhancement Using Graph Neural Networks. 2021 IEEE Int. Conf. Acoustics Speech and Signal Processing Proc. 2021.

Phan H, Ian McLoughlin V, Lam Pham, Oliver Chen Y, Philip Koch. Improving GANs for speech enhancement. IEEE Signal Process. Lett. 2020;27(1):1700-1704.

Pandey A, Wang D. On adversarial training and loss functions for speech enhancement. IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc 2018.

Plantinga P, Bagchi D, Fosler-Lussier. An exploration of mimic architectures for residual network based spectral mapping. IEEE Workshop Spoken Language Proc. 2019.

Pandey A, Wang D. A new framework for CNN-Based speech enhancement in the time domain. IEEE/ACM Trans. Audio Speech Lang. Process 2019; 27(7):1179-1188.

Rownicka J, Bell P, Renals S. Multi-Scale octave convolutions for robust speech recognition. IEEE Int. Conf. Acoustics Speech and Signal Processing Proc., 2020.

Rethage D, Pons J, Serra X. A wavenet for speech denoising. IEEE Int. Conf. Acoustics Speech and Signal Processing Proc. 2018.

Robert Rehr, Timo Gerkmann. On the Importance of Super-Gaussian Speech Priors for Machine-Learning Based Speech Enhancement. IEEE/ACM Trans. Audio Speech Lang. Process 2017; 26(2):357-366.

Ravi Kumar K, Subbaiah PV. A Survey on Speech Enhancement Methodologies. I. J. Intelligent Systems and Applications.2016; 8(12):37-45.

Soni MH, Shah N, Patil HA. Time-Frequency Masking- Based speech enhancement using generative adversarial network. IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc 2018.

Sunnydayal V, Sivaprasad N, Kishore Kumar T. A survey on Statistical Based Single Channel Speech Enhancement Techniques. I. J. Intelligent Systems and Applications. 2014;6(12):69-85

Sainath TN, Weiss RJ, senior AW, Wilson KW, Vinyals O. Learning the speech front-end with raw waveform CLDNNS. proc. Interspeech, 2015; 1-5.

Thaleiser S, Enzner G. Binaural-Projection Multichannel Wiener Filter for Cue-Preserving Binaural Speech Enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing 2023; 31(1):3730-3745.

Wang Y, Han K, Wang DL. Exploring monaural features for classification- based speech segregation. IEEE Trans. Audio, Speech, Lang. Process. 2013, 21(2): 270-279

Wollmer M, Zhang Z, Weninger F, Schuller B, Rigoll G. Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise. IEEE Int. Conf. Acoustics Speech and Signal Processing Proc., 2013; 6822-6826.

Xu Z, Zhao Z, Fingscheidt T. Coded Speech Quality Measurement by a Non-Intrusive PESQ-DNN. IEEE/ACM Transactions on Audio, Speech, and Language Processing 2023; 31(1):3404-3417

Xu Y, Du J, Huang Z, Dai LR, C. H. Lee. Multi-Objective learning and mask-based post-processing for deep neural network based speech enhancement. Proc. Annu. Conf. Int. Speech Communication Association Interspeech 2015

Xu Z, Elshamy S, T. Fingscheidt. Using separate losses for speech and noise in mask-based speech enhancement. IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc., 2020.

Yan-Hui Tu, Jnu Du, Chin-Hui Lee. Speech Enhancement Based on Teacher-Student Deep Learning Using Improved Speech Presence Probability for Noise-Robust speech Recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 2019; 27(12):2080-2091.

Yong Xu, Jun Du, Li-Rong Dai, Chin-Hui Lee. A Regression Approach to Speech Enhancement Based on Deep Neural Networks. IEEE/ACM Trans. Audio Speech Lang. Process 2014; 23(1):7-19.

Zhao Y, Xu B, Giri R, Zhang T. Perceptually guided speech enhancement using deep neural networks. 2018 IEEE Int, Conf. Acoustics Speech and Signal Processing Proc. 2018; 5074-5078.

Zhang W, Du H, Liu Z, Zhang Q, Gao J. Cross-Representation Loss-Guided Complex Convolutional Network for Speech Enhancement of VHF Audio. IEEE Transactions on Instrumentation and Measurement 2023;72(1): 1-10.