CHAPTER 2

LITERATURE SURVEY

2.1 INTRODUCTION

Humans use speech as their most common and preferred means of communication. Speech has several uses in human-machine interaction in addition to inter-human communication. When it comes to speech, it is always preferable for noise to have little or no impact so that it does not interfere with conversation. Speech is affected by several kinds of noises, so getting speech free from noises seems difficult. Hence, it is clear that the noise should be reduced to zero or minimum in a given speech signal. This is done through the process of speech enhancement. Single Channel Speech enhancement refers to improving the quality and intelligibility Phan et al., (2020) of a speech signal recorded using a single microphone in the presence of interferences.

Several issues, such as interference from background noise, reverberation, and environmental noise, easily impact a speech signal. This is why speech enhancement plays a significant role in processing speech signals. It is necessary to implement speech enhancement techniques to remove the noise that interferes with a speech signal Narwaria et al., (2012). This is essential for extracting clear speech for speech-based applications, such as mobile speech communication and Automatic Speech Recognition (ASR) Cohen, (2003); B. Li et al., (2013) speaker recognition J. Li et al., (2011) speech coding and hearing aids Chern et al., (2017); Levitt, (2001). It is possible to lessen noise distortion while enhancing the comprehensibility and aural quality of speech that unfavorable circumstances have compromised. There have been many techniques developed to enhance speech quality in recent years, including Log Spectral Amplitude Zhang & Wang, (2017), Spectral Subtraction Boll, (1979), Wiener Filtering Lim & Oppenheim (1979), and Minimum Mean Square Error (MMSE) Xu et al., (2015). Musical noise is one of the major drawbacks of these traditional techniques.

A wide range of noise conditions is considered by developing nonlinear DNN regression models using data from training sets, such as noisy speech, noise from speakers,

and noise types. In challenging circumstances and noisy real-time environments, the DNN's performance is constrained. The training set comprises hundreds of various noise kinds to get around this restriction and enhance the generalization capabilities for identifying varied inputs. This effort effectively managed the many categories of invisible noise and the non-stationary nature of noise. It is accomplished by balancing the global variance of the reference clear speech characteristics and the improved speech characteristics to lessen the over-smoothing issue. Dropout training Hinton et al. (2012) is used for TIMIT datasets when an overfitting issue occurs. By including noise information in the DNN inputs, Noise Aware Training (NAT) is accomplished by Seltzer et al. (2013) to enhance the performance and noise reliability of DNN-based voice augmentation systems. The Non-negative Matrix Factorization (NMF) is a crucial component when separating a target source from mixed data or noise from speech signals. There are many different applications for NMF, including speech enhancement, speech recognition in adverse environments, acoustic signal detection, and acoustic source separation, among others Dionelis & Brookes, (2018); He et al., (2017); Lan et al., (2020); Mowlaee et al., (2017). DNN estimates the encoding vectors to rebuild the desired source data vectors and improve the effectiveness of the source subspace overlapped NMF target data extraction technique Kang et al., (2015). The noise produced by the interfering sources and the clear speech are involved in the data sent to the DNN for training purposes. DNN modeling maps the data vectors to their corresponding encoding vectors. As an alternative to NMF for separating clean speech from noisy speech, DNN may be applied in two steps: first, to distinguish between noisy and clear speech, then subsequently, to improve the quality and intelligibility by enhancing the speech Grais et al., (2017b). Using NMF, exemplar-based speech enhancement can also be used to enhance speech. This technique trains on noisy data and clear speech using Time-Frequency (TF) representations Baby et al., (2015). It is well known that noise and speech differ in modulation frequency, and the modulation spectrogram characteristic is reliable in satisfactorily differentiating them. In addition to frequency masking, time-domain masking is used when environmental noise is a common problem Williamson et al., (2016). This approach enhances the phase and magnitude responses of noisy speech by analyzing the ideal ratio mask in the real and imaginary domains. DNN is used to learn the ideal ratio mask by mapping echoic speech to the complex ideal ratio mask (Williamson & Wang, 2017). In low signal-to-noise ratio

environments, improved Least Mean Square Adaptive Filtering (ILMSAF) by Lee et al., (2016) helps to overcome the disadvantages of inadequate performance and poor adaptability in different noise environments. The Deep Belief Network (DBN) Tamilselvan et al., (2012) and Xie & Li (2021) are capable of estimating adaptive filter coefficients that are crucial for efficiently removing noise. Filter parameters are selected depending on the noise classification performed by DNN to remove noise. Reduced performance in conditions of mismatched noise is the most frequent issue with DNN-based algorithms. For this problem to be solved, more noise types must be included in the training dataset. DNNs can also enhance speech to extract features based on linear frequency spectral mappings R. Li et al., (2016). It may be possible to recover clean speech features by proceeding with pre-enhancement to the spectral components of the input of the DNN.

2.2 REVIEWS ON TRADITIONAL ALGORITHMS FOR SINGLE CHANNEL SPEECH ENHANCEMENT

He et al., (2017) proposed a multiplicative update method for calculating linear predictive components of noise and speech in the codebook-driven Wiener filtering speech enhancement method. In contrast to the current codebook-driven approaches, which estimate speech and noise linear predictive gains using the Maximum-Likelihood method, the suggested method uses a multiplicative updating algorithm to analyze the linear predictive gains. It has the advantage of preserving more parts of the speech from the improved speech. The multiplicative update method reduces the noise between the harmonics of noisy speech due to the development of an upgraded codebook-driven Wiener filter and the likelihood of speech presence. Additionally, the method addresses the problem of the standard codebook-driven Wiener filter method's coarse fitting of the speech spectrum. Between the harmonics of the spoken utterance, the coarse fitting will leave a noise residue.

The Double Spectrum (DS) is used to analyze single-channel speech augmentation Mowlaee et al., (2017), comprising modulation and pitch-synchronous transformations. First, the theoretical underpinnings of the proposed DS domain are investigated, along with its favourable features for pitch estimation and speech presence probability estimation. Wiener filtering and adaptive weighting-based speech enhancement techniques have been

suggested for the DS domain. The results demonstrate that the suggested DS-based approaches are more effective than conventional modulation and short-time Fourier transform techniques. It is shown that the suggested technique achieves a considerable tradeoff between enhanced perceived efficiency and a moderate decrease in speech intelligibility under various signal-to-noise ratios and noise kinds.

Lavanya et al., (2020) proposed the phase compensation algorithm to enhance the intelligibility and quality of distorted speech by using temporal and spectral changes. In existing work, this was done by modifying the magnitude spectrum by holding the phase spectrum constant. The first level of speech enhancement is achieved by changing the phase spectrum alone. Using second-level enhancement, weak speech and non-speech areas are extremely contrastive by performing energy redistribution in the distorted speech signal. Using the adaptive power law transformation (APLT) approach, energy is redistributed from the weak unvoiced areas to the energy-rich voiced by calculating the parameters with a total amount of energy using the particle swarm optimization algorithm. The third degree of speech enhancement is accomplished by combining the Log MMSE approach with a new SPU estimate method. The enhanced speech signal is further modified using the corrected magnitude and phase spectrum calculated using log MMSE. The proposed method increases the intelligibility and signal quality under non-stationary and stationary noise conditions. The suggested algorithm is compared with previous methods, considering the magnitude and phase spectra for better signal estimation. Evaluation is done for different SNR values in non-stationary and stationary noise environments.

Dionelis & Brookes (2018b) proposed a technique for improving speech using circular statistics, where modulation-domain Kalman filtering is used to monitor the phase and spectral log amplitudes of speech and noise. The proposed method provides a better speech phase spectrum for speech signal reconstruction by using the posteriori of the speech phase. While the Kalman filter updates its step of the models on the assumption that the speech and noise accumulate in the challenging short-time Fourier transform domain, the prediction step separately models the temporal inter-frame correlation of the speech and noise spectral log amplitude. Utilizing various noises and SNRs, the phase-sensitive enhancement method is assessed using speech quality and intelligibility measures. According to instrumental measurements, modulation-domain Kalman filtering

outperforms the traditional enhancement algorithms in terms of speech quality than the traditional enhancement algorithms. Several suggested enhancement techniques employ Kalman filtering in the time domain. They differ, however, in the choice of Kalman filtering update, Kalman filtering state, and Kalman filtering prediction for the time-frequency domain, with the Kalman filtering state having a log-spectral, a power-spectral, or a speech amplitude spectral dimension. The signal model utilized to include noise and speech impacts the Kalman filtering update. Speech and noise can be distinguished using the complicated Short-Time Fourier Transform (STFT) domain. However, noise and speech accumulate in the power spectrum or amplitude spectral domain.

Krawczyk & Gerkmann (2014) suggested that speech enhanced in the short-time discrete Fourier transform domain can be corrupted by noise. By using a single microphone signal, the spectral amplitude is modified. Various studies have shown that an improved spectral phase can be well utilized for speech enhancement. In this literature, a method has been presented from the fundamental frequency and noisy observation in the spectral phase of the speech. The advantages of the spectral phase and the reason for reducing noise through modifying the phase are discussed. It is also mentioned that the enhancement of the noisy phase is used in the proposed reconstruction method. Instrumental measures predict speech quality through various signal-to-noise ratios without amplitude enhancement. The proposed phase reconstruction method can be combined with spectral amplitude estimators to increase speech enhancement performance. Those methods can overwhelm the conventional amplitude enhancement methods.

Mowlaee & Stahl (2020) proposed a signal channel speech algorithm that operates in a Short-Time Fourier Transform along with dependencies concerning frequency. A Minimum Mean Square Error (MMSE) is obtained due to inter-frequency dependencies. Generally, the optimal value for estimating the Short-Time Fourier Transform (STFT) expansion coefficients maximizes the covariance matrix's complexity. From the observed data, covariance matrices need to be estimated. Based on the analysis made on the single channel speech enhancement method, a linear multidimensional short-time spectral amplitude estimator is derived from the complex-valued second-order statistics. For single-channel speech enhancement, inter-frequency dependencies have been evaluated, and then the estimator resulting from the statistical model is compared. The findings demonstrate

that inter-frequency dependencies can improve speech quality and intelligibility more than the conventional methods.

Tantibundhit et al., (2010) proposed the joint TF approach, which divides the speech signal into non-transient and transient components. This separation is done using a wavelet packet of the examined speech signal. The transient component is selectively increased and mixed with the original speech to create the modified speech, which has energy equivalent to the original. This method could create a new dynamic speech enhancement approach by multiplying each wavelet packet coefficient by a factor based on the features detected by the tiling ratio. As a result, the joint time-frequency method is altered in both the time and frequency domains.

Mowlaee et al., (2017) suggested a single-channel speech enhancement employing the Double Spectrum (DS), which combines modulation and pitch-synchronous transformations. The effectiveness of the single-channel speech enhancement method is compared with the conventional method and the Short-Time Fourier Transform method. The recommended method successfully strikes a balance between improved perceived quality and a slight loss of speech understanding across various noises and SNR. Using previous estimates of the underlying probability density function of DS magnitude, it is possible to build statistically based noise suppression techniques, such as a DS-based maximum posteriori (MAP) estimator.

Wood et al., (2019) present a universal codebook-based speech amplification architecture in which all speech and noise components are encoded using the same codebook. The possibility that a specific codebook atom is encoded at any given time is known as the Atomic Speech Presence Probability (ASPP). The hybrid version and the Interaural Level Difference (ILD) exploit the interaural transfer mechanism and the Interaural Coherence Magnitude (ICM). The effectiveness of the resultant ASPP-based speech enhancement algorithms using binaural mixes of reverberant speech and background noise has been noted. The methodology surpasses two benchmark approaches for binaural speech enhancement using objective speech efficiency and intelligibility over a wide range of input SNRs based on PESQ and binaural STOI measures. The ASPP system successfully combines binaural cue preservation and binaural noise reduction. It

focuses on adding an auditory model to a subtractive improvement procedure. Changing the subtraction parameters can alter the degree of noise reduction, speech interference, and musical residual noise associated with single-channel subtractive algorithms.

2.3 REVIEWS ON SINGLE CHANNEL SPEECH ENHANCEMENT TECHNIQUES BASED ON MACHINE LEARNING AND DEEP LEARNING

Shimada et al. (2019) proposed a concept to improve ASR in noisy surroundings. This concept has the benefit of directing the speech vector and spatial covariance matrix. The unsupervised speech enhancement method based on Multichannel Nonnegative Matrix factorization (MNMF) guided beamforming is performed in this literature. In the suggested technique, the Spatial Covariance Matrices (SCMs) of speech and noise are unsupervised and detected using MNMF. It generates an improved speech signal with beam forming. A web-based version of the MNMF has been added, and Independent Low-Rank Matrix Analysis (ILRMA) is used to initialize the MNMF. Beamforming of many kinds under several different circumstances has been assessed. The experimental findings of the actual recording show that ASR-based approaches were more robust in an unknown environment than the conventional beamforming strategy using DNN-based mask estimation. This study has suggested integrating blind source separation (BSS) with a DNN-based spatial covariance matrix (SCM) to enhance the performance of ASR. It is believed that understanding a matrix from a clean speech database is used to boost speech enhancement performance. The findings of DNN-based mask detection might be utilized to initialize the MNMF and filter out the SCMs of speech and noise.

Saleem et al. (2018) suggested a supervised learning method to enhance a speech-babble-degraded signal. The less aggressive wiener combines wiener filtering, and DNN is used to improve speech. The DNN simultaneously calculates the amplitude spectrum of noise-free speech and the noise signals from the speech characteristics masked by the input noise during the training phase. Then, an additional layer of the less aggressive Wiener filter is integrated with the DNN to construct the enhanced magnitude spectrum. The phase of a noisy speech signal can then be used to recover the approximated clear speech signal. The trained DNN uses noise-masked speech signals during testing to produce an improved speech signal. Deep learning approaches recently considered in speech enhancement

applications use learning methods with multi-layer representation. The nonlinear model transforms the representation from one higher order to another at each layer.

M. Tu & Zhang (2017) suggested a new Deep Neural Network-based architecture to enhance speech. This has been employed in contrast to standard feedforward network architecture by skipping connections between network inputs and outputs to indirectly force the DNNs to learn the ideal mask ratio (IRM). This work shows that the performance could be improved by stacking multiple network blocks. The study results prove that the architecture performs better and is more satisfying than the existing methods in three commonly used objective measurements under two real noise conditions. Also, the network architecture should be further improved for its learning ability, achieved by using the stacked network.

P. Wang et al. (2020) suggested a monaural speech enhancement method to achieve better quality and intelligibility. The production of ASR systems has remained the same as anticipated when trained with noisy speech. When there is a discrepancy between speech recognition and monaural speech augmentation, it is typically due to speech interference generated during the speech enhancement process. The results show that the distortion problem may be resolved by utilizing distortion-independent acoustic modeling. A different speech enhancement model during the training stage is also suitable for the acoustic model. On the CHiME-2 database, the models examined in this research performed better than the previous research. Reducing or eliminating speech distortions in enhancement front ends is one technique to solve the distortion issue. After comparing the acoustic models, it has been found that the interferences-independent acoustic model, trained using a wide variety of improved speech, will be used to solve the interference issue. The usefulness of extensive training for acoustic modeling is suggested by its capacity to generalize untrained noises.

Kim et al. (2019) suggested that speech enhancement methods facilitate distinguishing between noisy and clear speeches. Several limitations hinder the enhancement methods from obtaining clear speech targets. In this literature, an Acoustic Adversarial and Supervision (AAS) learning algorithm is implemented. In AAS, each supervision enhances generalization on an unseen noisy speech by maximizing the likelihood of transcription on the pre-trained acoustic model and ensuring general properties of clean speech in the augmented output. This approach is tested by using

CHiME-4, DEMAND, and Librispeech datasets. The authors illustrated each supervision's function with the enhanced feature's visualization. AAS demonstrated a lower WER than speech enhancement techniques utilizing a clean target. Any noisy speech and clean speech model and transcription can be combined with the proposed AAS. As a result, the AAS algorithm's improved output has enhanced generalization to previously unvoiced signals and maximizes the likelihood of transcribing using the pre-trained acoustic model. The supervision functionality of the AAS is observed through the prediction and visualization of improved speech with various loss combinations.

Kawase et al. (2020) suggested that speech-improving algorithms based on the gradient approach deal with additive noise in some devices, such as headsets and earphones. These techniques are developed to improve the signal-to-distortion ratio and not for ASR. Therefore, the front-end speech enhancement is adapted and adjusted to each acoustic model. An increase in ASR engine accuracy is observed in this literature when the front-end speech enhancement technique is changed. When ambient noise fluctuates, the suggested solution makes it easier for consumers to utilize ASR through their devices. This study uses a genetic algorithm to generate parameter values for the front-end speech enhancement for specific situations. The initial grouping of environments based on noise characteristics assigns the computed values to input speech signals. In post-evaluations, the parametric values predetermined exceed in performance compared to the existing ones done by human experts. The system adjusts the front-end speech enhancement parameter-set values systematically rather than experimentally. Once appropriate parameter-set values are established in advance for each noise environment, the best among them is automatically picked depending on the noise environment. A real-coded genetic algorithm (GA) is used to maximize ASR accuracy. ASR performance is improved by assigning optimal parameter-set values to some speech communication devices. Fully Convolutional Neural Networks (FCNNs) enhance speech intelligibility using simulated electric and acoustic stimulation.

Xiang & Bao (2020) suggested that the DNN-based speech enhancement method requires a large dataset corpus of clean and noisy speech to train the DNN network and enhance the network's efficiency. However, not all noisy signals are utilized in the training process due to the lack of clear speech. As noise varies with time, parallelly speech cannot be obtained due to limited speech and noise data. A novel parallel-data-free speech

enhancement technique that combines multi-objective learning and the CycleGAN, a cycle-consistent generative adversarial network, is used to solve this issue. During the training process, two encoders were used to encode noisy and clean speech characteristics. Next, the optimal TF mask and Log-Power Spectrum (LPS) of clear speech are predicted using dual forward generators. The LPS and magnitude spectrum of noisy speech are mapped using two inverse generators. This algorithm also proves its good efficiency in speech intelligibility and speech quality. Using the parallel-data-free speech enhancement technique maximizes the advantages of CycleGAN and multi-objective learning. During the test stage, encoders are directly connected with the alternators to get enhanced speech.

Yuan (2020) conveyed that the network designs utilized in the current DNN-based speech augmentation approaches are not explicitly created to separate local aspects of loud speech in a non-causal manner. The feature calculation method based on the TF correlation in the Improved Minima Controlled recursive averaging (IMCRA) by using Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) is used to model the correlation in the time and frequency dimensions, respectively. A TF smoothing neural network is recommended for enhancing speech. Numerous speech improvement systems created depend on diverse networks, and comprehensive tests are conducted on speech quality and intelligibility to confirm the suggested network's effectiveness for improving speech. First, the IMCRA frequency and time smoothing are implemented. Second, the IMCRA's feature transfer between time and frequency smoothing is considered while creating the TF smoothing neural network. Local characteristics of unclear speech are transferred between LSTM and CNN. Therefore, the TF smoothing neural network employs convolutional smoothing in the frequency dimension, whereas it uses recursive averaging in the time dimension. Thirdly, a time-frequency smoothing neural network (TFSNN) approach for improving speech is put forth to establish the best TFSNN structure. Finally, by evaluating the speech enhancement performance of TFSNN, rigorous, objective tests are conducted using speech intelligibility and quality.

The speech enhancement methods based on DNN have been compared by Gelderblom et al. (2019). Generally, a speech enhancement network targets to enhance the quality of noisy speech. A speech recognition test assesses both systems' speech comprehensibility and quality. Both were compared with the STOI and Perceptual Objective Listening Quality Analysis (POLQA). However, POLQA and STOI could only

accurately forecast the necessary subjective outcomes. However, results of STOI show degradation in speech intelligibility.

Lan et al., (2020) proposed an Attention-based Redundant Convolutional Network (ARCN) that includes an attention mechanism and a Redundant Convolutional Encoder-Decoder (RCED) for speech improvement. The performance of speech enhancement has been evaluated using attention weight assignment. The basic speech enhancement technique provides weights to channels based on information at large. The authors suggest using Spatial Speech Enhancement (SSE) to weight each time-frequency point after considering spatial information. The performance is boosted by combining SSE and Channel-wise SE (CSE) advantages in four ways. In this literature, the authors have proved that the suggested SE techniques improve the performance of the model by avoiding heavy computational processes, and it has been generalized well to train SNRs, speakers, and untrained noises.

Hsieh et al. (2020) proposed a Wave Convolution Recurrent Neural Network (WaveCRN) E2E (end-to-end) SE model. Generally, WaveCRN uses a bi-directional network to design the sequential correlation of extracted features. It has been proved that WaveCRN achieves denoising capability and computational efficiency compared to other similar works of experts. The authors have proposed four major methods in this literature: WaveCRN, which combines Simple Recurrent Unit (SRU) and CNN to perform E2E SE. Secondly, they have proposed a Restricted Feature Mask (RFM) method derived from transforming the noisy features into enhanced ones. Third, they experimented with the Stacked SRU model, which resulted in excellent performance compared to other SE models. Finally, they have designed a new practical application based on SRU, and its performance has been tested.

Azarang & Kehtarnavaz (2020) presented an overview of multi-objective deep learning methods, which have been widely used for speech denoising. This literature reviews the mathematical structure of the multi-objective deep learning approaches for speech denoising and an overview of conventional, single-objective deep learning or integrated conventional deep learning methods. A sample speech denoising approach is applied using publicly available corpora and four commonly used objective metrics. The results obtained after the comparison show the effectiveness of the multi-objective method when the SNR is low compared to other similar methods. The authors have also suggested

that future improvement could be achieved by designing deep neural networks depending on the noise characteristics encountered.

Liu et al. (2017) proposed a unique perceptually-weighted goal function inside a feedforward DNN architecture that reduces the perceived difference between the enhanced and target speech. The suggested objective function incorporates a perceptual weight tested on ideal ratio masks and spectra as two different output characteristics. The psychoacoustic properties of the perceptual weight model are taken into consideration, and it is also evaluated on a feedforward regression DNN. Authors have conducted unbiased assessments of speech intelligibility and quality. The suggested unique strategy improved objective speech intelligibility and signal quality compared to standard approaches using uniform weights.

Bao & Abdulla, (2019) proposed a ratio mask based on Computational Auditory Scene Analysis (CASA) to enhance speech signals. They utilized Inter-Channel Correlation (ICC) between noise and speech to adjustably reallocate the speech power ratio and noise power ratio when the ratio mask was being constructed. As a result, more speech components are retained, where masking is done more efficiently and accurately for noise components. The authors recommended a channel-weight contour that might be used across the Gammatone channels to enhance the ratio mask, depending on the equal loudness of the hearing characteristic. The updated ratio mask was successfully applied to train a five-layer structured deep neural network by a five-layer structured deep neural network. According to the authors' experiments, by employing six different types of noise to demonstrate speech quality, intelligibility, and spectrum distortion, the suggested ratio mask works better than the traditional ratio mask representation.

Pandey & Wang (2019) proposed a learning mechanism for a full CNN to improve speech enhancement in the time domain. The time frames of noisy utterances have been taken as input for training the network. An additional operation that changes the time domain into the frequency domain has been implemented throughout the training period. This transformation is differentiable since it is equivalent to a straightforward matrix multiplication, which means that training in the time domain might be done using a frequency domain loss. It is recommended that a mean absolute error loss between the clean STFT magnitude and the enhanced STFT magnitude be employed to train the CNN. Thus, the model might use its domain expertise to analyze signals by translating them to

the frequency domain. It has been demonstrated that the experimental results of the learning mechanism outperform the other speech enhancement techniques.

R. Li et al., (2020) proposed a method for enhancing speech performance using post-processing and network model. First, the authors worked on improving a power function compression Mel-Frequency Cepstral Coefficients (MFCC). Second, the authors suggested extracting abstract data from high-level and comprehensive sequence modeling. A stacked Convolutional Neural Network (SCNN) has local connection properties on the 2D plane and a strong modeling capacity to extract local information. However, the Temporal Convolution Neural Network (TCNN) is more effective when dealing with temporal modeling due to its dilated causal convolution and residual block. The STCNN's layered convolution and dilation blocks are advantageous in speech enhancement tasks.

Y.-H. Tu et al., (2020) proposed a fully CNN-based regression model that directly achieves the 2D noisy Log-Power Spectra (LPS) input to 2D TF mask output mapping, denoted as 2D-RFCNN. In order to ensure that each convolutional filter observes more contextual information, the initial input is a 2D noisy LPS. The deep convolutional layers with a modest filter size can be employed for the regression problem since the network input and output are the entire utterance feature map. In order to guarantee that the end dimension of the frequency bin has a value of 1 that can map the frequency dimension, the pooling operation on the frequency bin is utilized. The deep convolutional layers with a tiny size filter, commonly used in speech recognition and enhancement, are then employed. Research demonstrates that the 2D-RFCNN model enhances speech quality and intelligibility and lowers the recognition error rate on actual test data.

Ming & Crookes (2017) proposed a speech enhancement approach that depends on multi-frame and segment estimation. It is feasible to estimate speech from noise accurately by having the Zero-mean Normalized Correlation Coefficient (ZNCC). The proposed method will apply to predictable and unpredictable noise without using proper training data. A novel realization that combines clear speech recognition with full-sentence speech correlation is employed to overcome the problem of data sparsity. Various noises, including extremely non-stationary noise, were tested with the system on two separate datasets. It is demonstrated that the multi-frame and segment estimation method significantly outperforms established techniques, which employ optimal noise tracking on various objective metrics, including automated speech recognition.

Fu et al. (2018) suggested a Fully Convolutional Neural Network (FCN) that uses an end-to-end utterance-based speech enhancement framework to bridge the gap between model optimization and assessment criteria. Each utterance is treated as a whole to achieve the optimal global solution of the speech. However, the congruence between the training, assessment objectives and experimental findings demonstrate that the STOI of speech processed using the proposed technique is superior to speech optimized using traditional MSE. The frame boundary problem caused by zero-padding has been solved using utterance-based optimization. Also, it has the advantage of an objective function used throughout the utterance. The best goal of the literature is to demonstrate the known characteristics of human audio perception and minimize the MSE between clean and augmented features.

Kameoka et al., (2020) proposed a Sequence-to-Sequence (seq2seq or S2S) learning algorithm. This voice conversion technique adaptably translates the voice characteristics, pitch contour, and length of input speech. The model it employs has a completely convolutional architecture. This has the added benefit of being appropriate for GPU-based concurrent calculations. Additionally, it is advantageous since it permits batch normalization and other efficient normalizing methods for all the networks' hidden layers. Secondly, instead of learning mappings between each speaker pair independently using a distinct model, it simultaneously learns mappings between numerous speakers using only one model to enable many-to-many conversion. Identifying common latent properties that may be shared between speakers enables the model to use the training data gathered from many speakers fully. As a result of its structure, this model can perform many-to-many conversions even without knowing the source speaker. Finally, the conditional batch normalization technique alters the layers of batch normalization based on the target speaker. For the many-to-many conversion paradigm, it has been observed that this specific approach is incredibly successful.

Tuzla (2011) proposed a mono-channel speech-music separation technique based on spectral masks and Nonnegative Matrix Factorization (NMF). The mono-channel speech-music separation method divides the mixed signal using NMF, and masking is done after training the data of speech and music signals. NMF trains a set of basis vectors for each source using the training data during the training phase. The magnitude spectrum domain of NMF is used to train these bases. In order to deconstruct the magnitude spectra

of the mixed signal into a linear mixture of the learned bases for both sources, NMF is employed to observe the mixed signal. The findings demonstrate that applying masks after NMF improves separation performance even when NMF is computed with fewer iterations, resulting in a quicker separation procedure. A set of masks with a parameter for controlling saturation level is adopted. The recommended strategy speeds up the separation process and enhances performance.

Grais et al., (2017a) suggested that the Single Channel Source Separation (SCSS) problem is frequently addressed using DNN, which forecasts TF masks. The sources are further differentiated from the mixed signal using the masks. Different kinds of masks create distinct sources with varying degrees of interference and distortion. Some types of masks produce low interference between the separated sources, whereas others produce low distortion between the separated sources. For SCSS, a mixture of DNNs' prediction masks is employed rather than each DNN to enhance the quality of the separated sources. The four distinct masks are anticipated by training four separate DNNs that minimize four different cost functions. Reference binary and soft masks are trained into the first and second DNNs. A mask is directly predicted by the third DNN using the reference sources as training data. A discriminative constraint is added to the last DNN's training, as with the third DNN, to enhance the differences between the estimated sources. According to the findings of experiments, combining the predictions of various DNNs resulted in separated sources with greater quality than utilizing each DNN separately.

Huang et al. (2018) discuss an innovative method for improving speech, such as DNN with Leaky ReLU and DNN with sigmoid activation functioning in a Multi-Band Excitation (MBE) structure built on a DNN. The learning level and the stage of development make up the two main tiers of the system. Throughout the training phase, two DNNs, such as DNN with Leaky ReLU and DNN with sigmoid activation function, are trained along with various targets. The training is focused on identifying how the clean speech signal fits within harmonic magnitude and band gap. Two DNNs get their input from the LPS of the noisy speech stream. A better speech signal may be produced through MBE speech synthesis by integrating the output of DNNs with the expected online pitch period. The recommended approach makes it feasible to accurately approximate the MBE model's parameters to synthesize a better speech signal with minimum noise and better signal quality. As a result, harmonic distortion is practically eliminated.

Chung et al. (2018) discussed coping with distinct noise characteristics. Based on the results of a separate noise classification DNN, the noise-dependent adaption vectors are identified, and the weights and biases of the spectrum mapping DNN are modified. During the training phase, the vectors and the DNN spectral mapping parameters are concurrently computed for adaptability. The authors combine the suggested DNN-based method with a conventional unsupervised speech enhancement algorithm during the enhancement step. The spectral mapping DNN weights and biases are evaluated using the noise-dependent adaptation vectors to enhance the speech signal quality. The settings for the noise detection DNN, adaption vectors, and spectrum mapping DNN are initialized during the training phase. A well-known non-supervised speech enhancement algorithm with the DNN-based approach has been integrated to improve speech quality further.

Yang et al. (2018) employed a soft-decision method based on the likelihood of Speech Presence Probability (SPP) to decrease noise PSD effectively. The threshold parameters must be calculated using several heuristics to discover noise PSD using SPP estimation. Because of this, a typical PSD noise-based statistical model requires rule-based experiments to accurately compute several parameters, which fails to ensure excellent performance. Deep learning architectures have demonstrated notable performance lately, and they have a solid reputation for excelling at modeling nonlinear input and target output data. A deep learning-based noise estimator PSD is exceptionally reliable with a sufficient count of noise categories and database size. However, deep learning structures are inappropriate for real-time communication systems as they need a lot of memory and processing power. Adopting a deep SPP estimation method for training aims to enhance the performance of the PSD noise estimation technique. The SPP gives statistically significant efficiency improvements with minimal complexity.

N. Y.-H. Wang et al. (2021) investigated the performance of deep learning-based Speech Enhancement based on Electric and Acoustic Stimulation (EAS) simulated speech. The Fully Convolutional Neural Networks (FCN) approach has been compared with the conventional MMSE speech enhancement approach and deep learning-based Deep Denoising Autoencoders (DDAE) approach in this literature. According to the results, the FCNs outperformed the other two SE techniques under all test settings. The results of these tests demonstrate that the FCN speech enhancement technique is superior to the MMSE and DDAE SE approaches. The benefits of the FCN speech enhancement in EAS have

been demonstrated. This work shows that it provides more gain in speech intelligibility than the other two speech enhancement models under the test conditions. The FCN model shows efficacy in enhancing normal and EAS-vocoded and has applications in EAS speech processors to enhance speech intelligibility.

Saleem et al. (2020) proposed speech enhancement using supervised learning of spectral masking. RNN and DNN are trained together to learn spectral masking of the degraded speech. Iterations are carried out by the post-processing step to deal with the noisy phase. Key band significance function weights are used for iterative intelligibility enhancement, with larger weights contributing more. Various experiments with the supervised method successfully attenuated background noise to an excellent level. The proposed method improves speech quality, intelligibility, and automatic speech recognition.

Takeuchi et al. (2020) proposed an application of DNN for speech enhancement. An Equilibrated Recurrent Neural Network (ERNN) is proposed to avoid or remove the vanishing or exploding gradient problem without increasing parameter numbers. A causal method that requires past information in real-time is noticed. The ERNN uses only a few parameters to achieve results similar to uni-directional and bi-directional LSTM networks. When the parameters are reduced, the experimental observations prove that the suggested method outperforms the other LSTM methods.

Chen & Liang (2018) initiated a speech enhancement technique based on combining Ensemble Empirical Mode Decomposition (EEMD) and DNN. The original signal is pre-processed using EEMD to match better the time-variation criterion, which involves breaking down a series of TF information from the IMF component. Using DNN, the IMF component's weight has been modified to improve the speech. The variations in speech enhancement performance between utilizing EEMD and employing EEMD have been compared. The outcomes demonstrate that the improved EEMD method as a pre-processing technique increases the PESQ and STOI scores and significantly enhances speech quality and intelligibility.

Wijayasingha & Stankovicare (2021) have proposed a method that develops and analyses various distinct noise mitigation strategies for CNN-based Speech Emotion Recognition (SER). The Berlin Database of emotional speech SER models are tested using

clean data, and data blended with ten distinct types of noise at various noise levels. Many approaches have been demonstrated to improve the noise robustness of SER models, including synthetic noise, modified group delay spectrogram, and the magnitude spectrogram together by adding an attention mechanism. The models trained using combined input on noisy data outperformed individual input models. It also shows that the noise-resilient algorithms could be applied to other datasets and evaluated on the RAVDESS dataset.

Xiao & Nickel (2010) proposed a targeted speaker enrolment and device training technique for typical background noise. Speech efficiency is increased using this speech enhancement algorithm. However, speech intelligibility cannot be increased by using this algorithm. In this research, the authors provided a theoretical framework that may be utilized to explore potential influences on the comprehension of processed speech. This method focuses on the in-depth analysis of the distortions explicitly caused by speech enhancement algorithms. If these distortions are adequately controlled, it is hypothesized that improvements in intelligibility can be made. These tests aim to identify the perceived effects of the numerous distortions that speech enhancement algorithms may induce on speech intelligibility. Results from three distinct augmentation algorithms show that some distortions are more detrimental to the degradation of speech intelligibility than others. The approach is intended for feasible situations to enroll specific speakers and train systems in a typical noisy environment.

Pandey & Wang (2021) suggested that speech enhancement in the time domain has become more common recently due to its capacity to enhance the phase and volume of speech simultaneously. Self-attention with a Dense Convolutional Network (DCN) enhances speech in the temporal domain. The DCN design uses skip connections and is based on encoders and decoders. There are dense blocks and attention modules on each layer of the decoder and encoder. In order to extract features, attention modules, and dense blocks combine maximal context aggregation and deeper networks. The SNR improvement and artifact issues are lost based on the spectrum magnitude of amplified speech, which are previously unknown. Based on the expected noise and augmented speech volume, it provides STFT magnitude loss for solving the SNR improvement issue. The noise prediction constraint ensures that the loss improves both the magnitude and phase, even if the proposed loss is only based on magnitudes. Experimental findings show

that DCN trained with the loss performs significantly better than other cutting-edge non-causal and causal speech enhancement methods.

Liang et al. (2020) initiated a speech enhancement technique, a single channel based on attention-gated LSTM. The program separates the frequency spectrum according to the Bark scale to imitate human auditory perception qualities. The derivative features of pitch-based features and Bark Frequency Cepstral Coefficients (BFCC) features are derived from these bands. The computational complexity is decreased by using these band gains as training objectives rather than the full band gain. Additionally, as various noises have varying effects on clean speech, the attention mechanism is employed to sort out the less polluted information by noise, which helps reconstruct clean speech. IRM with ICC is selected as the learning objective to adaptively reallocate the power ratio between speech and noise. The technique uses a speech activity detector to optimize the networks using a multi-objective learning strategy, enhancing the network's performance. According to subjective and objective studies, this algorithm performs better than previous baseline algorithms. The recommended method maintains high real-time performance and improved convergence speed in a real-time experiment. The method's robustness and speech enhancement performance are improved and compared with the other algorithms, demonstrating the attention model's higher Speech Enhancement performance. This approach reduces noise, enhances speech quality, and is well-generalizable to non-matching samples while maintaining low complexity.

Hasannezhad et al., (2021) initiated an RNN that effectively generalizes the model to represent strong temporal dynamics of speech using the sequential information from auditory frames. Additionally, a model's performance may be enhanced using the CNN's capacity to extract speech components automatically. An innovative low-complexity LSTM network and FCNN are combined in a hybrid neural network model that is presented to estimate a phase-sensitive mask for speech enhancement. While maintaining a reasonable level of model complexity, the model is built to effectively harness the temporal dependencies and spectral correlations in the input speech signal. CNN-derived features are changed by applying an attention mechanism. To show CNN's superiority in feature extraction, certain high-quality conventional acoustic characteristics are compared to CNN. The model also considers the most popular RNN versions for the mapping portion, with the LSTM proving to be the optimum trade-off through computational time,

number of model parameters, and performance. According to extensive comparison experiments, the model significantly outperforms several well-known neural network-based speech enhancement algorithms when there are several non-stationary noises.

Strake et al. (2020) designed Fully Convolutional Recurrent Networks (FCRN) for speech enhancement. FCRN integrates convolutional layers for feature extraction and recurrent layers for temporal modeling, thereby addressing the challenges posed by noisy and degraded speech signals. The FCRN architecture integrates convolutional and recurrent layers to capture both local and temporal dependencies in noisy speech signals. This architecture can handle long-range temporal information, making it well-suited for speech enhancement. FCRN employs convolutional layers to extract meaningful spectral and spatial features from the input spectrogram of the noisy speech signal. This feature extraction step is crucial for distinguishing between the desired speech and background noise. The recurrent layers in FCRN capture the temporal dynamics of speech. These recurrent connections allow the network to learn and exploit long-term dependencies in the audio signal, which is essential for effective speech enhancement. This research contributes significantly to the field of speech enhancement and can potentially improve the quality of speech communication in various real-world applications.

Y.-H develops a teacher-student learning system. Tu et al. (2019) integrate deep learning techniques and unsupervised speech improvement methods, such as IMCRA, at the training stage. The trained student model may be applied to speech enhancement in ASR systems as the pre-processor before the recognition step. Through experimental research, it is shown that under challenging conditions, the regression model fails to learn the nonlinear connection between the noisy LPS characteristics and the desired IRM, where it cannot perform in terms of unprocessed noisy speech data in terms of ASR performance. This algorithm is used to build the recommended student model, which has the IRM as the learning target that could increase recognition accuracy. Even with a more powerful recognition system, the authors suggested that speech enhancement techniques might still enhance ASR performance.

Das & Hansen, (2012) discuss Auto-Line Spectral Pair (Auto-LSP), a constrained iterative speech enhancement method used to address selectively boosting speech based on broad phoneme classes. Altering the Auto-LSP parameters creates multiple enhanced utterances for each noisy utterance. After the noisy utterance has been divided into parts

based on general phoneme classes, each segment is subjected to restrictions using a hard decision solution. A solution depending on a Gaussian mixture model (GMM) and maximum likelihood (ML) soft decision is also implemented to eliminate the effects of noise.

Grais et al. (2017b) suggest a two-stage strategy for separating the sources. The first stage is the separation of the sources from the combined signal. The second stage employs DNNs to reduce distortions and interference between the dispersed sources (DNNs). Instead of using a distinct trained DNN for each separated source as in the first strategy, each source is improved independently in the second technique. In the second stage, discriminative training is done on the DNNs suggested to reduce interference and distortions further, enhancing the separated sources' quality. According to the experimental findings, utilizing two stages of separation instead of only one improves the quality of the separated signals by reducing interference and distortions between the separated sources.

2.4 SUMMARY OF THE LITERATURE SURVEY

This chapter discusses the literature on single-channel speech enhancement. In the next chapter, traditional algorithms for speech enhancement have been discussed, and the performance metrics of traditional algorithms like Wiener and hybrid algorithms (Wiener Filter, Wavelet Transform, and LMS algorithm) are compared.