#Instructions and Guidance in this template

You will find guidance and information in this template.

Instructions are highlighted in turquoise and inside hash symbols, #like this#. All such information is designed to aid you in complying with the University's submission and completion guidance on producing a research thesis. This should be deleted when finalising your thesis.

Text placeholders, where you need to enter data, are highlighted in yellow and surrounded by straight brackets. Remove both the yellow highlighting and the square brackets when finalising the document.

The Table of Contents, Table of Tables and Figures will need to be updated and links on this are provided above each.

The template meets all these requirements; however, you are heavily advised to use the support materials to find out more about using Word to ensure your document continues to comply as you add your own content.

There are demonstration seminars available as Doctoral College Workshop and Events and web support on the Producing your thesis in Word site

Copyright Notice
The copyright notice below should be left unaltered. It should be included in your e-thesis but is not required in your print document.#

University of Southampton

Faculty of Engineering and Physical Sciences

Institute of Sound and Vibration Research

Speech Enhancement by Using Deep Learning Algorithms

by

Jianqiao Cui

ORCID ID: 0000-0002-6016-5574

Thesis for the degree of Doctor of Philosophy

15 July 2024

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences

Institute of Sound and Vibration Research

Thesis for the degree of Doctor of Philosophy

Thesis Title Speech Enhancement by Using Deep Learning Algorithms

by

Jianqiao Cui

Speech signals are often degraded by ambient noise, which significantly hampers speech intelligibility and quality, posing challenges for both human communication and speech-related technologies. Over the past decade, the advent of deep learning has catalysed remarkable progress in the field of speech enhancement. With the proliferation of smart devices demanding real-time processing capabilities, the development of real-time deep learning-based speech enhancement systems has become increasingly pertinent.

The primary objective of this thesis is to advance the state-of-the-art in real-time speech enhancement algorithms, with a focus on improving the intelligibility and quality of speech in noisy environments. Our research commences with an exploration into the intricacies of auditory perception and the impact of hearing loss on speech comprehension, setting the stage for the development of sophisticated speech enhancement techniques.

Traditional speech enhancement methods are reviewed in chapter 2, leading to an in-depth discussion on the selection of features critical for distinguishing speech from noise. The work transitions to deep learning neural networks, detailing architectures like LSTM-RNNs and CNNs, and their implementation in speech enhancement, emphasizing the importance of quantitative evaluations.

Chapter 3 delves into the application of Generative Adversarial Neural Networks (GANs) in the domain of speech enhancement, building upon existing research to further refine the use of these models. The chapter focuses on the innovative integration of the magnitude spectrum as an input feature, which significantly contributes to the performance enhancement of GANs. Additionally, the exploration of various deep learning architectures as potential generators within the GAN framework is presented, showcasing the adaptability and continuous improvement potential of GANs in speech enhancement.

Attention mechanisms are presented as a driving force for innovation in speech enhancement, with the novel 'Mask First, Compensation Last' topology aiming to reduce speech distortion and residual noise. Motived by them, the chapter 4 further explores a new cascaded architecture on raw waveform input against the complexity of auditory perception.

Chapter 5 brings a new combination method in speech enhancement, contrasting mapping-based and masking-based methods, and proposing a parallel dual-module system, the Compensation for Complex Domain Network (CCDN), that unifies the magnitude spectrum with complex domain details.

The final chapter addresses the challenge of data mismatch in traditional supervised methods. We proposed an innovative strategy that combines unsupervised pre-training with supervised fine-tuning. This approach not only enhances speech quality in complex noise environments but also simulates the advantages of supervised learning without requiring paired data. Our model's adaptability to real-world noise conditions and its effectiveness in various speech enhancement tasks are validated through rigorous experimental evaluations and subjective listening tests. This chapter culminates in showcasing a robust, and practical speech enhancement model fit for real-world application, highlighted by its adaptability to real-world noise conditions and the integration of unsupervised learning strategies for enhanced model robustness and versatility. By enhancing the quality of human communication and addressing challenges faced by individuals with hearing impairments or in noisy environments.

Table of Contents

Table of Tables

Table of Figures

List of Accompanying Materials

1. Librispeech dataset: openslr.org
2. DNS-Challenge 2023: GitHub - microsoft/DNS-Challenge: This repo contains the scripts, models, and required files for the Deep Noise Suppression (DNS) Challenge.
3. NoiseX-92: Saeedi, Jamal & Ahadi, Seyed Mohammad & Faez, Karim. (2015). NOISEX-92 8KHz.
4. Musan (A corpus of music, speech, and noise): openslr.org

Research Thesis: Declaration of Authorship

Print name: Jianqiao Cui

Title of thesis: Speech Enhancement by Using Deep Learning Algorithms

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:
   - BDAI conference 2023.
   - 53rd International Congress and Exposition on Noise Control Engineering

Signature: .... Date: ....

Acknowledgements

First and foremost, I extend my deepest appreciation to my advisor, Professor Stefan Bleeck. His guidance has been instrumental in the completion of this dissertation. Reflecting on my journey, it is a tapestry of growth and learning, starting from my initial days as a Ph.D. student with a limited research background. Over the past five years, Professor Bleeck has been a beacon of inspiration, instilling in me the values of rigor, persistence, and adaptability—hallmarks of a successful researcher. I am privileged to have been his student and to have collaborated with him on a myriad of fascinating research problems. His mentorship, insights, and encouragement have been pivotal in sculpting me into the researcher I am today. The wisdom I have gleaned from him will undoubtedly enrich my professional journey.

I am profoundly grateful to Professor Thomas Blumensath and Professor Jon Barker for their dedicated service on my final examination committee.

Additionally, I extend my thanks to my junior lab colleagues, including Xiaoxue Wang, Kubra Kumrular, Esma Akis, Aimee Zhang, Michael Chesnaye, and Yi Zhuang, for the camaraderie and shared experiences that have enriched our collective journey.

I am also thankful for the circle of friends who have added depth and joy to my life. While it is impractical to enumerate all their names, I must acknowledge Professor Ying Zheng, Dr. Jun Wu, Dr. Shuyue Li, Yaping Zou, and Yitian Meng, among others.

My most profound gratitude is reserved for my family, whose unwavering support has been the bedrock of my success. I am eternally grateful to my parents, whose boundless love and tireless sacrifices have underpinned my every endeavor. Without their steadfast encouragement and companionship, my academic odyssey would not have been possible. To them, my gratitude is boundless.

Lastly, I owe a special debt of gratitude to my beloved wife, Lingqing Han, who has been a constant source of joy, solace, and inspiration. She is not merely my life partner but also a mentor, from whom I continually learn the art of personal growth. Her companionship, support, trust, and love are the cornerstones of my existence.

I also wish to express my sincere thanks to Sci-hub, Google Scholar, CSDN, Github, Hugging Face, and arXiv for making a huge amount of literature accessible to me. Without these fantastic and very helpful websites, I would have been unable to learn many interesting and important ideas from research papers written by others.

Definitions and Abbreviations

SE	Speech Enhancement
STFT	Short-time Fourier Transform
ISTFT	Inverse Short-time Fourier Transform
IBM	Ideal Binary Mask
TBM	Target Binary Mask
IRM	Ideal Ratio Mask
cIRM	Complex Ideal Ratio Mask
SMM	Spectral Magnitude Mask
AMS	Amplitude Modulation Spectrogram
Mag	Magnitude
GFCC	Gammatone Frequency Cepstral Coefficient
MFCC	Mel-Frequency Cepstral Coefficient
DNN	Deep Neural Network
ReLU	Rectified Linear Unit Activation Function
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
LSTM	Long-short Time Memory
BN	Batch Normalization
GCN	Gated Control Neural Network

Chapter 1 Background

The realm of speech enhancement has garnered significant attention in the past decades due to its profound impact on a myriad of applications, ranging from telecommunications to assistive technologies for the hearing impaired. The core objective of speech enhancement is to improve the quality and clarity of speech in the presence of noise, a challenge that has been continuously addressed through various signal processing techniques.

Historically, speech enhancement was approached with methods that were relatively simplistic in nature, often limited by the computational tools available at the time. These traditional methods, while effective to a certain extent, could not fully capture the complexities inherent in human speech and the variability of noise. With the advent of more powerful computing capabilities and the explosion of data availability, the field has witnessed a paradigm shift towards utilizing machine learning and, more recently, deep learning algorithms to tackle these challenges.

The intricate interplay between speech and noise presents a formidable challenge; it involves understanding not just the acoustic properties but also the perceptual aspects of how humans interpret sounds. This necessitates a dual focus on improving speech intelligibility and quality—a task that becomes significantly more complex in noisy environments. The background of this work is hence established on the foundation of auditory perception, exploring how the human auditory system perceives and processes sound waves and how this understanding can guide the development of advanced speech enhancement algorithms.

The deep learning techniques employed in this thesis represent the cutting edge in computational auditory scene analysis. Deep learning's ability to learn hierarchical representations makes it uniquely suited to model the complex structures of speech and noise, thereby enabling the development of more effective enhancement strategies. This thesis takes a novel approach by leveraging Generative Adversarial Networks (GANs) and advanced neural network architectures to innovate upon traditional enhancement methods. The focus is on not only the magnitude but also the phase spectrum—a component often neglected in previous research, yet crucial for preserving the naturalness of speech.

As the abstract of the thesis indicates, a considerable portion of this work is dedicated to pushing the envelope in neural network design and implementation for speech enhancement. This includes the application of LSTM-RNNs and CNNs, which have demonstrated great promise in modeling temporal and spatial dependencies in data, respectively. The thesis evaluates these methods against various performance metrics, showcasing the quantitative improvements over traditional models.

The background context set by this thesis provides a comprehensive survey of past and current methodologies, setting the stage for the innovative strategies introduced in subsequent chapters. It recognizes the limitations of prior work and addresses them by developing a multi-faceted approach that not only enhances speech but does so in a way that is robust to the diverse noise environments encountered in real-world scenarios.

In summary, this thesis stands at the intersection of technology and human-centric design. It is grounded in a deep understanding of the challenges posed by the field of speech enhancement and seeks to address these with novel, evidence-based deep learning methods. The background provided herein elucidates the complexity of speech in noisy environments and the subsequent need for innovative computational approaches to improve speech intelligibility and quality, ultimately contributing to the field's advancement and the betterment of communication.

Chapter 2 Advancements in Speech Enhancement: From Traditional Techniques to Deep Learning Innovations

2.1 Introduction

Chapter 2 of this thesis presents a comprehensive exploration of speech enhancement techniques, a critical component in improving the intelligibility and quality of speech signals within noisy environments. Beginning with an overview of traditional methods, the chapter progresses to discuss the evolution and implementation of classical deep learning methods, highlighting their significance in advancing the field of speech enhancement.

The advent of deep learning has opened new avenues for robust speech enhancement, providing sophisticated tools capable of learning complex patterns in data. This chapter elucidates such advancements, starting with the foundational concepts of autoencoders and advancing through various neural network architectures including Convolutional Neural Networks (CNNs), Wave-U-Net, and Deep Complex U-Net (DCUnet). Each method is examined for its unique approach and contribution to enhancing speech signals.

In the pursuit of optimal speech enhancement, feature selection stands as a pivotal step, and this is thoroughly discussed, followed by the necessary considerations in determining targets and labels for the enhancement process. The chapter then delves into the intricate process of waveform resynthesis, an essential step to regenerate clear and enhanced speech from processed signals.

The latter sections of the chapter are dedicated to a detailed analysis of deep learning neural networks, tracing their origins, structure, and the specialized training methods that equip these networks to effectively tackle speech enhancement challenges. Special attention is given to Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) and CNNs, underlining their particular applications in this context.

An examination of the evaluation metrics offers insights into the assessment of speech enhancement methods, ensuring the reader understands the criteria for success within this domain.

Furthermore, the chapter introduces the innovative application of Generative Adversarial Networks (GANs) in speech enhancement. This includes an in-depth look at the roles of the generator and discriminator within GANs and an evaluation of existing GAN variations tailored for speech signal improvement.

As this chapter unfolds, it lays down a structured pathway of understanding, from basic principles to complex implementations, providing a solid foundation for further investigation and innovation in speech enhancement strategies.

2.2 Traditional speech enhancement methods

In the field of single-channel speech enhancement, spectral subtraction and Wiener filtering are considered as the more traditional methods. When the background noise is stationary, these methods can produce an ideal denoising effect [1]. The Wiener filter is a filter that is commonly used in signal processing to generate an estimate of a desired or target random process by linear time-invariant (LTI) filtering of an observed noisy process, assuming known stationary signal and noise spectra, and additive noise. The Wiener filter works by minimizing the mean square error between the estimated random process and the desired process.

Assume that the noisy speech sequence collected by the microphone is $y(n)$, clean speech is $x(n)$, and the noise is $d(n)$. Then the noisy speech sequence is the sum of the noiseless speech sequence and the noise sequence. Both the original speech signal and noise can be regarded as random signals.

$$y(n) = x(n) + d(n) \tag{2.1}$$

Commonly used speech enhancement methods are in the frequency domain, and it is necessary to frame, window, and short-time Fourier transform (STFT) the noisy signal $y(n)$ to obtain the frequency domain signal of each frame, where X, Y and D are clean speech, noisy signal and noise frequency domain signal, respectively.

$$Y(w_k) = X(w_k) + D(w_k) \tag{2.2}$$

And the cleaned speech $X(wk)$ can be estimated by $H(wk)$:

$$X\ (w_k) = H(w_k)Y(w_k) \tag{2.3}$$

Where:

$$H(w_k) = \frac{P_{xx}}{P_{yy}}$$

Where Pxx is the clean speech magnitude and Pyy is the noisy speech magnitude. Meanwhile, $H(wk)$ can be expressed by the priori SNR εk and the posteriori SNR γk respectively.

$$H(w_k) \ = \frac{P_{xx}}{P_{yy}} = \frac{P_{xx}}{P_{xx} + P_{dd}} = \frac{\varepsilon_k}{1 + \varepsilon_k} \tag{2.4}$$

$$H(w_k) \ = \frac{P_{xx}}{P_{yy}} = \frac{P_{yy} - P_{dd}}{P_{yy}} = \frac{1 - \gamma_k}{\gamma_k} \tag{2.5}$$

Note that $H(wk)$ is real, nonnegative, and even [2]. By minimizing the error between clean and cleaned speech, combining the phase information, we can get the estimate of the clean speech.

Advantages and disadvantages

The Wiener filter has carved a niche for itself as a cornerstone in the realm of signal processing. Its prominence stems from its adaptability to a diverse array of signals - be they continuous or discrete, scalar or vector, and especially stationary random processes[3]. In specific scenarios, one can derive an explicit solution for the filter's transfer function, paving the way for Wiener filters to be realized through an assembly of basic physical components [4].

However, the Wiener filter isn't without its constraints. Key challenges encompass its requirement for observation data over a semi-infinite time interval and its unsuitability for vector applications or non-stationary random processes with nois [5]. These limitations cast a shadow on its applicability in real-world scenarios. Central to the implementation of the Wiener filter is the prerequisite that the input process be generalized stationary, and the statistical intricacies of said process be well-understood [6]. Filters following analogous optimal criteria often echo these requirements. The external signal and interference environment dictate the input process, rendering its statistical traits unpredictable and often transient. This dynamic nature often contravenes the Wiener filter's stipulations[6]. Consequently, in practical scenarios, where the input process is non-stationary or its statistical nuances remain elusive, alternatives like spectral subtraction and time-frequency masking gain traction [7].

2.3 Existing classic deep learning methods for speech enhancement

In the world of speech enhancement with deep learning, we often talk about the difference between systems that work in real-time and those that don’t. This really shines a light on how we have to balance how fast the system runs with how complex the algorithms are. In real-time systems, which don’t use future information much or at all, it’s really important to make sure things happen as they come. It underscores the importance of designing lightweight models that do not compromise on performance despite the constraints of hardware capabilities.

What’s really interesting is that there’s more focus on making sure models can handle immediate data rather than just being fast， which is super important for things like hearing aids or phones. It’s about understanding that “real-time” can mean different things depending on what you’re using it for, and that’s something we need to think about more.

The conversation also brings up a point that’s been a bit of a hot topic: how we’ve been testing these models. There’s a push for being clearer and fairer when we compare models, which is great because it means we can really see which innovations are worth it. This is not just about keeping things honest; it’s also about helping people who are new to this area in navigating the field with better judgment and understanding.

2.3.1 Autoencoders for Speech Enhancement

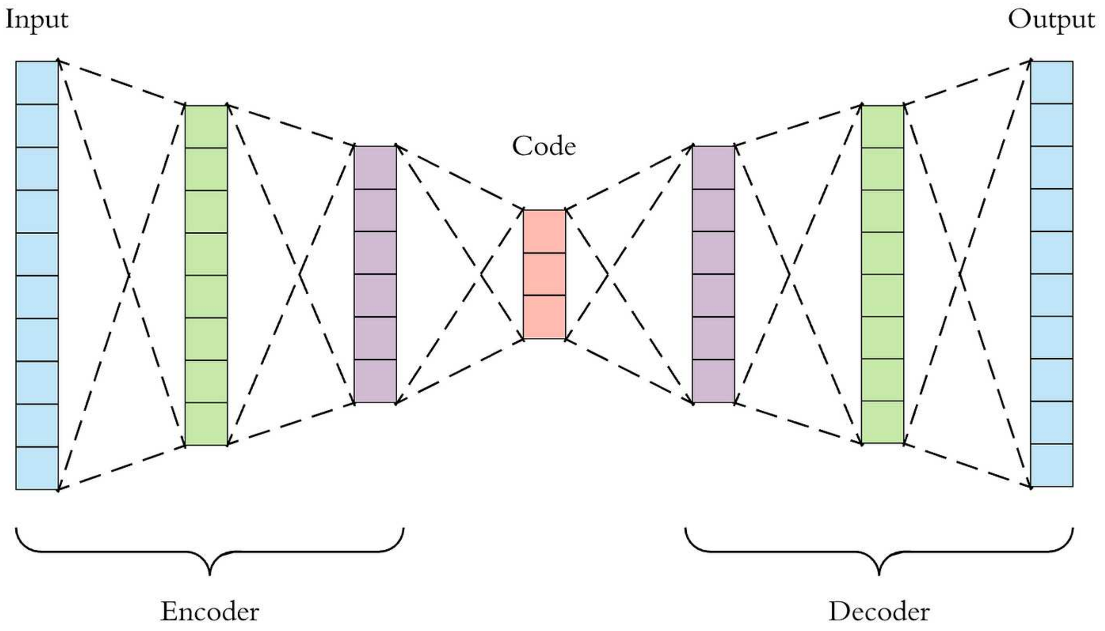


Figure 2.1 The autoencoder structure.

Autoencoders (AEs), shown in Figure 2.1, are an artificial neural network utilized for learning efficient data codings. Comprised of an encoder for input compression and a decoder for reconstructing the original format, AEs capture the salient features of the input data. For speech enhancement, Denoising Autoencoders (DAEs), like [9][8], are often employed, designed to reconstruct clean speech from noisy inputs. They are appealing for their ability to model complex, non-linear relationships and unsupervised training. This methodology assumes that the clean and noisy speech shares a common low-dimensional representation, with noise seen as a corruption of this representation.

2.3.2 Convolutional Neural Networks (CNN) for Speech Enhancement

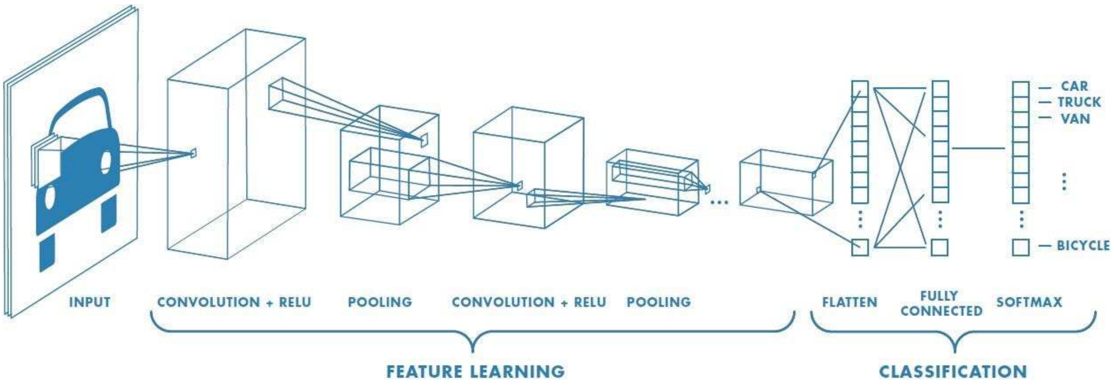


Figure 2.2 The classic CNN for a classification task [10]

Convolutional Neural Networks (CNNs) [11], a class of deep neural networks, have demonstrated significant success in image and speech processing domains. Originating from their prowess in image classification tasks, the inherent architecture of CNNs makes them adept at capturing local and hierarchical patterns in data. This property is particularly beneficial for speech signals, which contain hierarchical structures, such as phonemes, words, and sentences, as well as intricate temporal and spectral patterns.

In the realm of speech enhancement, CNNs, shown in Figure 2.2, are employed to extract robust features from noisy speech signals and then utilize these features to suppress or eliminate the underlying noise [11]. Their capability to process data in their native form, without the necessity for manual feature extraction, provides them an edge. By leveraging multiple convolutional layers, CNNs can automatically learn discriminative features from the raw waveform or spectrogram representations, facilitating the distinction between the desired speech and unwanted noise [12].

Several studies have confirmed the efficacy of CNNs in speech enhancement tasks. For instance, a CNN-based model was proposed to work directly on the raw waveform, bypassing traditional time-frequency representations, and achieved notable improvements in speech quality and intelligibility [13]. Another work integrated CNNs with traditional spectral subtraction methods, leading to enhanced performance under various noise conditions [14].

In conclusion, the incorporation of CNNs into speech enhancement algorithms offers a promising direction, potentially paving the way for more sophisticated and robust systems in the future.

2.3.3 Wave-U-Net for Speech Enhancement

The Wave-U-Net [15] emerged as a pioneering adaptation of the standard U-Net, tailored for the one-dimensional time domain, to facilitate end-to-end audio source separation. Distinctively, this network operates directly on the raw audio waveform, negating the need for traditional time-frequency representations. By doing so, Wave-U-Net is endowed with the ability to intrinsically learn the optimal time-frequency trade-off, a vital aspect in audio processing.

Empirical studies have underscored the merit of Wave-U-Net in source separation tasks, like Figure 2.3. Specifically, in the realm of speech enhancement, this architecture has displayed remarkable prowess. When trained to distinguish clean speech from noisy environments, Wave-U-Net has consistently achieved state-of-the-art performance, surpassing many contemporary models [16][17]. However, it's worth noting that while the Wave-U-Net showcases impressive results in several scenarios, like any model, it is not devoid of limitations and may encounter challenges in extremely noisy or unpredictable environments [18].

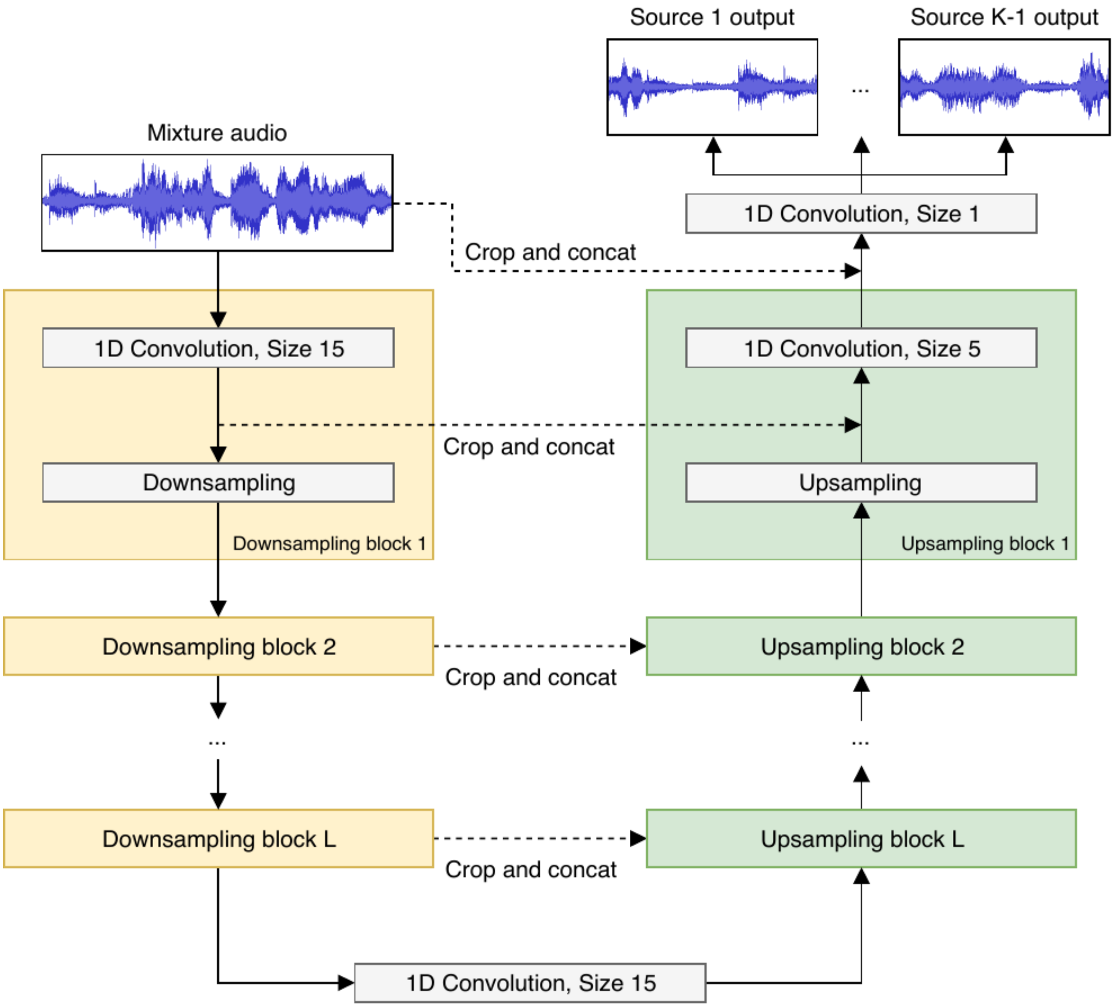


Figure 2.3 The Wave-U-Net for speech enhancement.

2.3.4 Deep Complex U-Net (DCUnet) for Speech Enhancement

The Deep Complex U-Net (DCUnet) stands as a notable advancement in the realm of speech enhancement models [20]. This sophisticated architecture (Figure 2.4) is an innovative extension of the original real-valued U-Net, transitioning into the complex domain. Such a design choice is pivotal in accurately modeling both the spectral magnitude and phase of speech signals. By doing so, the DCUnet harnesses the full potential of the information encapsulated within the short-time Fourier transform of noisy speech.

A salient feature of the DCUnet is its incorporation of complex convolution layers. These layers are adept at preserving the inherent complex nature of the input data. Empirical evaluations have reinforced the efficacy of DCUnet, highlighting its superior performance across a plethora of speech enhancement benchmarks [13]. For instance, comparative studies with traditional speech enhancement methods have frequently underscored the DCUnet's ability to deliver clearer and more intelligible speech outputs [21].

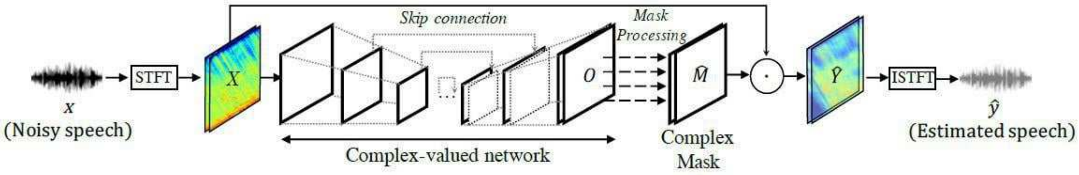


Figure 2.4 The DCUNet for speech enhancement.

2.3.5 Conv-TasNet for Speech Enhancement

Conv-TasNet [19] is a deep learning-based approach for speech separation. Conv-TasNet uses a fully convolutional network to estimate ideal time-domain masks, which are used to separate individual sources in a mixture. The core idea of Conv-TasNet is to operate directly on the raw time-domain signal, which can model the full complexity of acoustic mixtures.

From the Figure 2.3.5, the Conv-TasNet architecture consists of three main components:

1. The encoder: which transforms the input time-domain signal into a non-linear representation.
2. The separator: a temporal convolution network (TCN), which infers a mask on the encoded representation.
3. The decoder: which applies the mask and transforms the representation back into the time-domain.

One key advantage of Conv-TasNet is its superior performance in real-time low-latency applications, such as hearing aids and telecommunications, where the delay caused by the transformation in time-frequency domain methods might be unacceptable.

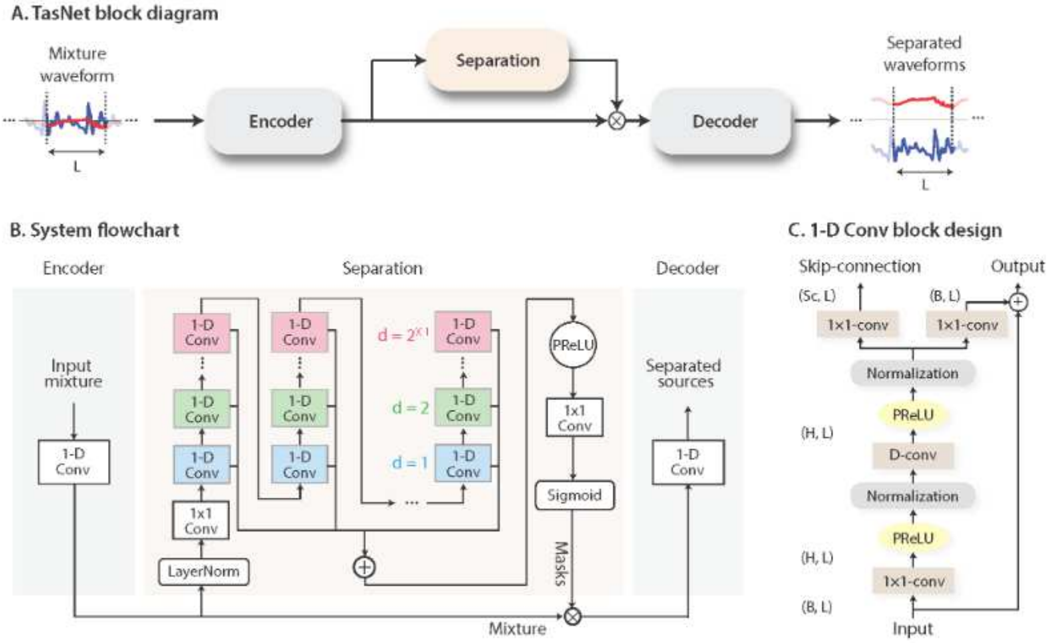


Figure 2.5 The structure of Conv-Tasnet.

2.4 Feature selection

Speech signal processing plays a critical role in many applications, such as speech recognition, speaker identification, and speech enhancement. To accurately extract and analyze the speech features, both time domain analysis and frequency domain analysis are commonly used. However, relying solely on either approach has its limitations. For instance, time-domain waveforms and characteristic parameters cannot fully capture the frequency information contained in speech signals, while frequency domain analysis can only provide a static representation of the speech signal, failing to capture its dynamic temporal characteristics. To address these issues, a dynamic frequency spectrum called a spectrogram can be used to represent speech features. Spectrograms are three-dimensional plots of time, frequency, and amplitude, where the amplitude of the speech signal is indicated by the color intensity or gray value of each point on the plot. This representation allows for a comprehensive analysis of the time-frequency characteristics of the speech signal, making it a valuable tool for speech processing and analysis.

Spectrograms serve as a vital tool in speech signal analysis, offering a comprehensive visualization of how the spectral density of signals evolves over time. Essentially, they portray the intensity of various frequencies in a speech signal throughout its duration, effectively capturing both its spectral and temporal characteristics. This dual representation is crucial for many speech processing tasks, as it helps in distinguishing different phonemes, understanding speech rhythms, and identifying various speech artifacts.

A crucial component of the spectrogram is the power spectrum, which represents the magnitude of the signal's frequencies at a particular time. The power spectrum is derived by computing the magnitude of the Short-Time Fourier Transform (STFT) of the signal for that specific time window. Mathematically, the power spectrum $S(f, t)$ of signal $x(t)$ can be represented as:

$$S(f, t) = \left| \int x(\tau) w(t - \tau) e^{-j2\pi f \tau} d\tau \right|^2 \tag{2.6}$$

Where $w(t)$ is the window function.

Power spectra are pivotal in numerous speech processing applications:

Speech Recognition: The spectral details captured by the power spectrum are often used as features in automatic speech recognition systems, aiding in the differentiation of phonemes and other linguistic units [22].

Speaker Identification: The unique spectral patterns of individuals can be extracted from the power spectrum, providing distinctive features for speaker recognition [23].

Noise Reduction: By analyzing the power spectrum, unwanted noise components in a speech signal can be identified and suppressed, leading to enhanced speech clarity [24].

Given the significance of spectrograms and power spectra in speech processing, they serve as foundational elements in the development of effective speech enhancement algorithms, especially those grounded in deep learning.

The choice of input features plays a crucial role in determining the effectiveness and precision of the supervised speech enhancement system. In speech enhancement, frequency domain features are commonly obtained by applying time-frequency decomposition techniques to mixed signals. Two of the most popular time-frequency decomposition techniques are the short-time Fourier transform (STFT) [25] and the Gammatone auditory filtering model [26]. The corresponding spectral features obtained from these techniques are known as the Fourier domain features [27][28][29] and the Gammatone domain features [30][31], respectively. These features can capture the spectral characteristics of the speech signal and are therefore commonly used as input features for speech enhancement algorithms. However, it is important to note that the selection of the appropriate time-frequency decomposition technique and spectral feature extraction method can significantly impact the performance of the speech enhancement system.

The choice of input features undoubtedly shapes the efficacy of speech enhancement systems. While both the short-time Fourier transform (STFT) and the Gammatone auditory filtering model are prevalent time-frequency decomposition techniques, their suitability varies depending on the specific requirements and challenges posed by different speech enhancement tasks.

STFT (Short-Time Fourier Transform)

Advantages: The STFT offers a linear frequency resolution, making it suitable for capturing harmonic structures in speech, especially in lower frequency regions [32]. It's computationally more straightforward and is widely adopted in many speech processing tasks due to its ease of implementation and interpretability.

Drawbacks: However, the fixed resolution of STFT across all frequencies can be a limitation, especially when capturing details in higher frequency regions where human hearing is more logarithmically spaced [33].

Gammatone Auditory Filtering Model

Advantages: The Gammatone filter simulates the human auditory system's frequency selectivity, providing a non-linear frequency resolution that is more in line with human auditory perception [34]. This makes it particularly adept at capturing speech nuances that are more perceptually relevant. In scenarios where perceptual quality is paramount, such as in hearing aids, the Gammatone filter often outperforms linear frequency representations [35].

Drawbacks: However, its non-linear representation might not be optimal for all applications, and its computational complexity is higher than STFT [36].

Comparative Insights: Several studies have shown that while STFT-based features offer a robust and general representation, Gammatone-based features tend to provide a more perceptually-aligned representation, which can be advantageous in specific contexts. For instance, in noise-robust automatic speech recognition, the Gammatone filter has been shown to outperform STFT, especially in highly non-stationary noise environments [37]. However, for tasks like speech synthesis or compression, the linear frequency resolution of STFT might be preferred due to its computational efficiency and simplicity [38].

In conclusion, the decision between STFT and the Gammatone auditory filtering model depends heavily on the specific goals of the speech enhancement task. While STFT offers a universal representation suitable for a wide range of tasks, the Gammatone model provides a more perceptually-relevant representation, especially beneficial in scenarios where the perceptual quality of the enhanced speech is of utmost importance.

The selection of appropriate spectral features plays a crucial role in the performance and accuracy of a supervised speech enhancement system. Spectral features can be distinguished based on the size of the modelling unit, which can be either time-frequency (T-F) unit-level or frame-level. Due to the computational constraints, most of the existing studies on speech enhancement have focused on modelling based on T-F units, where the spectral features of each unit are extracted using time-frequency decomposition techniques [39][40][41]. However, recent research has shown that frame-level features can better capture the correlation information between adjacent time-frequency units and improve the performance of speech enhancement systems [42][43].

In traditional speech enhancement techniques, processing is often conducted on individual time-frequency units. This isolated approach can sometimes overlook the intricate interdependencies between these units. Recent advances in the field of deep learning have provided insights into the benefits of considering frame-level features, which represent entire frames or windows of time-frequency units, as opposed to isolated points.

Frame-level features offer a more holistic view of the speech signal. By capturing the correlation information between adjacent time-frequency units, these features allow a model to account for the contextual dependencies that exist in natural speech signals. For instance, in speech signals, the energy and frequency characteristics of a given unit are often influenced by its neighboring units. By processing on a frame level, systems can leverage this inherent structure of speech to make more informed enhancement decisions.

Several studies have demonstrated the superiority of frame-level features over traditional approaches. In a study by Zhang et al. [44], a deep neural network utilizing frame-level features achieved significant improvements in both objective and subjective speech quality measures compared to systems that only considered individual time-frequency points. Another research by Li and Wang [45] revealed that frame-level features could help in capturing long-term temporal structures in speech, leading to enhanced clarity and intelligibility.

These findings suggest that incorporating frame-level features in speech enhancement systems, especially those based on deep learning architectures, can lead to more accurate and natural-sounding enhanced speech. Such an approach not only captures the nuances of the speech signal but also aligns better with the way human auditory perception works, wherein context plays a pivotal role in deciphering sound.

In feature extraction, the input features of each time-frequency unit are processed based on ideal target masking, which determines whether the unit is dominated by speech or noise based on the estimated binary mask output. In frame-level feature extraction, a feature window is set, which includes the center frame and several adjacent frames, and all the time-frequency unit targets of the corresponding center frame are used as the output of the model. The system then learns a nonlinear regression model to estimate the ideal binary mask. By using the context information between speech frequency bands, frame-level features can better exploit the spatiotemporal structure information in the spectrogram. Therefore, frame-level features are becoming increasingly popular in the research community.

Amplitude Modulation Spectrogram (AMS)

One of the key advantages of the amplitude modulation spectrogram (AMS) is that it enables the analysis of the time trajectories of each frequency band of the non-logarithmic energy spectrogram [46]. This helps to capture important temporal information about the speech signal that is not easily captured by other feature extraction methods. To extract AMS features, the input signal is first subjected to full-wave rectification. Then, a quarter of the signal is selected for windowing and frame pre-processing, typically using a Hanning window for convolution. The Fourier transform is then applied to obtain a two-dimensional time-frequency signal, and the amplitude spectrum is calculated. Finally, a 15-dimensional AMS feature vector is obtained by convolution with 15 triangular windows of varying center frequencies between 15.6 Hz and 400 Hz. This approach has been shown to be effective in improving the performance of speech enhancement systems, particularly in noisy environments where it is important to capture the temporal structure of the speech signal in order to accurately separate speech from noise.

Advantages:

- Captures both spectral and temporal dynamics of the speech signal.
- Useful in distinguishing between speech and noise components [49].

Disadvantages:

- Requires additional preprocessing steps.
- May introduce computational overhead in real-time scenarios.

FFT-Magnitude [47]

The Fourier log amplitude spectrum (FFT-Log-Magnitude) is a feature that emphasizes the high-frequency components of the speech signal, and it is obtained by performing the logarithm operation on the Fourier amplitude spectrum. The process of obtaining FFT-Log-Magnitude features usually involves framing and windowing of the speech signal, followed by the application of the short-time Fourier transform (STFT) to each frame signal. The resulting STFT coefficient is then subjected to modulo calculation to obtain the energy, which is then converted to the Fourier amplitude spectrum. To emphasize the high-frequency components of the signal, the logarithm operation is performed on the Fourier amplitude spectrum, and the resulting feature is called the FFT-Log-Magnitude feature.

The FFT-Log-Magnitude feature is a commonly used feature in speech signal processing, as it captures important frequency information in the speech signal. It is particularly useful for tasks such as speech recognition and speaker identification, where the high-frequency components of the speech signal can contain important cues for identifying speakers or recognizing speech sounds. However, it should be noted that the FFT-Log-Magnitude feature may not be suitable for all applications, as it may not capture certain aspects of the speech signal that are important for other tasks, such as speech enhancement or speech synthesis.

Advantages:

- Offers a detailed frequency perspective of the signal.
- Computationally efficient and straightforward to implement.
- Has shown consistent results in preliminary tests and various deep learning architectures [48].

Disadvantages:

- Phase information is disregarded, which can be crucial in certain enhancement scenarios.
- Potential issues with spectral leakage.

Magnitude and Phase in Speech Enhancement

In speech enhancement, the selection of appropriate features is pivotal. Two primary features commonly used in deep learning-based speech enhancement are magnitude and phase.

Magnitude, often referred to as amplitude, represents the intensity or size of the speech signal in its time-frequency representation. In speech processing, the magnitude is typically derived through the Short-Time Fourier Transform (STFT) [50], providing us with the spectral characteristics of the signal. For deep learning models, magnitude offers a clear, continuous representation, allowing the model to easily learn from it and carry out the enhancement task [51].

Phase conveys the position or relative delay of the signal in its time-frequency representation [52]. In speech enhancement, while the magnitude contains most of the information related to the content of speech, phase remains vital as it provides structural information of the signal. In deep learning-driven speech enhancement, the phase of the noisy speech is typically used alongside the enhanced magnitude for waveform reconstruction [53].

Why Choose Magnitude as a Feature? The magnitude offers models a relatively stable and continuous representation, enabling the capturing of more characteristics of speech without interference from unnecessary variations [54]. Additionally, compared to phase, magnitude is easier for models to handle and learn since it isn't affected by periodic variations [55].

In conclusion, in deep learning-driven speech enhancement, employing magnitude as the input feature and using the phase of noisy speech for waveform reconstruction has proven to be effective [56]. This strategy merges the continuity of magnitude with the structural information of phase, providing a robust framework for improving speech quality and intelligibility.

Gammatone Frequency Cepstral Coefficient (GFCC)

In the realm of academic research, the gammatone frequency cepstral coefficients (GFCC) are a popular feature extraction method used for speech signal processing. The GFCC features are obtained through a series of processing steps. Firstly, a 64-channel gammatone filter is applied to the input speech signal to filter out the noise and enhance the useful speech information. Then, the filtered signal is resampled at a lower sampling rate of 100 Hz, and the amplitude is compressed by taking the cube root to improve the robustness of the features. Finally, the discrete cosine transform (DCT) method is utilized to extract the GFCC features, which are characterized by a set of coefficients representing the spectral envelope of the speech signal in the cepstral domain. Typically, a 31-dimensional GFCC feature vector is used for speech analysis applications [13]. This method has been widely studied and has demonstrated good performance in various speech enhancement tasks, such as speech recognition, speaker recognition, and speech quality evaluation.

Advantages:

- Aligns well with the human auditory system's frequency perception.
- Proven to be effective in noisy conditions, especially when traditional features fail [58].

Disadvantages:

- Computationally more intensive compared to traditional features.
- Extraction process can be intricate, requiring precise tuning.

Mel-Frequency Cepstral Coefficient (MFCC)

Mel frequency cepstral coefficients (MFCCs) are widely used in speech processing as they provide a compact representation of the spectral envelope of a sound. The calculation of MFCC features begins by framing and windowing the input signal, typically using a 20ms frame length with a 10ms frame shift. Different window functions, such as Hamming, triangular, or cosine overlapping windows, can be used [57]. Subsequently, the short-time Fourier transform (STFT) is used to obtain the energy spectrum of each frame. The resulting energy spectrum is then transformed into the Mel domain through a logarithmic operation and a Discrete Cosine Transform (DCT) to obtain the MFCC feature.

The Mel-Frequency Cepstral Coefficients (MFCCs) have been a cornerstone in the world of speech and audio processing for decades. The significance of MFCCs lies in their ability to represent the short-term power spectrum of sound, capturing the phonetically relevant characteristics of speech [60].

Sensitivity to Additive Noise MFCCs, however, have an inherent sensitivity to additive noise, which can degrade their effectiveness, especially in noisy environments. The presence of noise, especially non-stationary noise, can introduce significant perturbations in the MFCC feature space. This can result in a mismatch between the clean training data and noisy test data in speech recognition systems [61].

Normalization Techniques: To mitigate the influence of noise on MFCCs, various normalization techniques are employed. One widely used technique is Cepstral Mean and Variance Normalization (CMVN) [62]. CMVN compensates for channel and noise distortions by normalizing the mean and variance of the MFCCs over a given utterance or time window. Another method is RASTA (Relative Spectra) filtering, which emphasizes the modulation frequencies relevant to speech and de-emphasizes the slower modulation frequencies associated with channel and noise variations [63].

Relevance in Deep Learning Speech Enhancement

In the realm of deep learning for speech enhancement, the vulnerability of MFCCs to noise becomes especially pertinent. While deep learning models are robust in many ways, the quality of input features remains paramount. Some researchers argue that utilizing raw spectrogram or alternative features might offer better resilience against noise, especially when using deep neural networks [64]. However, the compact and phonetically relevant nature of MFCCs still makes them a valuable asset, provided they are processed correctly to counteract noise.

While MFCCs hold a pivotal position in speech processing, their sensitivity to additive noise remains a concern. Through normalization and other preprocessing techniques, their robustness can be enhanced, making them suitable for various applications, including deep learning-driven speech enhancement.

Researchers have proposed various modifications to the basic MFCC algorithm to enhance its robustness, such as raising the log-mel-amplitudes to a suitable power (e.g., around 2 or 3) before applying the DCT, which reduces the impact of low-energy components [59]. Examples of the speech waveform, corresponding MFCC spectrogram, and scaled MFCC spectrogram are shown in Figure 2.6, 2.7, and 2.8, respectively.

the speech waveform, corresponding MFCC spectrogram, and scaled MFCC spectrogram are shown in Figure 2.6, 2.7, and 2.8, respectively.

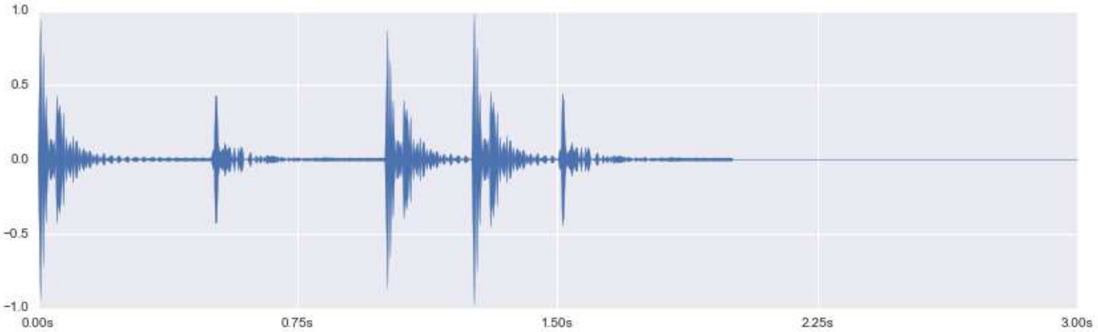


Figure 2.6 Speech waveform example. The horizontal axis represents time, and the vertical axis represents amplitude normalized from -1 to 1.

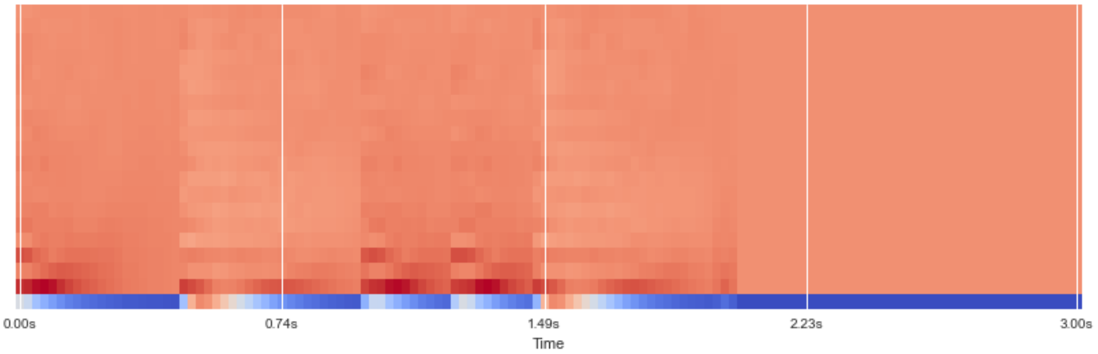


Figure 2.7 MFCC spectrogram. The horizontal axis represents time, and the vertical axis represents Mel frequency.

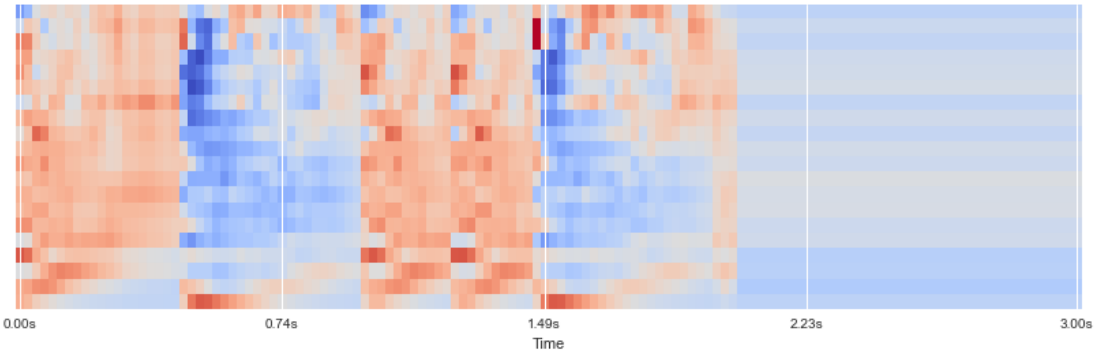


Figure 2.8 Scaled speech MFCC spectrogram.

Pitch Based Feature

Pitch features have been shown to be effective for speech recognition and analysis, particularly in noisy signals with or without reverberation. The pitch is a fundamental characteristic of speech that is determined by the periodic vibration of the vocal cords [65]. In this project, the magnitude is chosen as the input features for the deep learning model. This involves using the Fourier transform to break down the sound wave into its component tones of different frequencies, each represented by a sine wave of a different amplitude and phase. The magnitude of each component is used as an input feature, while the phase information is discarded.

However, it has been recognized that the phase information is important for accurately reconstructing the waveform, and early studies only focused on magnitude-related training targets. This approach limited the upper bound of performance, as the phase of the estimated speech deviated significantly from the original signal in the presence of interferences [66][67][68]. Recent approaches have been proposed for phase reconstruction, but the neural network models remained real-valued.

After evaluating the various features, for the purposes of this research, **Magnitude & Phase** was chosen as the primary input feature for the deep learning model. Its computational efficiency, combined with its detailed frequency representation, offered an optimal balance between complexity and performance. Preliminary tests also validated its compatibility and effectiveness with the proposed deep learning architectures. In future plans, the phase information will be taken into consideration, like [69][70]. One possible approach is to use raw waveform and complex-valued features as input features. This will allow the neural network to capture both magnitude and phase information, leading to more accurate speech reconstruction. The consideration of phase information will help to further improve the performance of the deep learning model in speech enhancement.

Frame-level processing

As we know that, there are various speech features we can use. Of course, some models like GRN [8] mainly adopt the magnitude as the input, and then reconstruct wave by combining with noisy phase. But the information of phase is also required to be taken into the consideration while training the model. In this case, the complex-valued feature [9] and time-domain frame-level feature will be introduced into the proposed models. Furthermore, speech is a non-stationary signal, consequently its statistical properties are not constant over time. Therefore, its spectral features and other characteristic properties (for example: short-time energy, MFCC etc.) should be extracted from small blocks of the signal. This is based on the assumption that is the signal is stationary (i.e., its statistical properties are constant within this region) in this small frame. On top of it all, frame blocking is often used in real-time systems as it maximizes the efficiency of the system by distributing the fixed process overhead across many samples.

Framing is a fundamental signal processing technique that consists of dividing the original signal into blocks often called frames with length N_f an overlap M and a framing hop $H (H = N_f - M)$.

Overlapping the frames help avoiding information loss between adjacent frames. The framing procedure is shown in Figure 2.9.

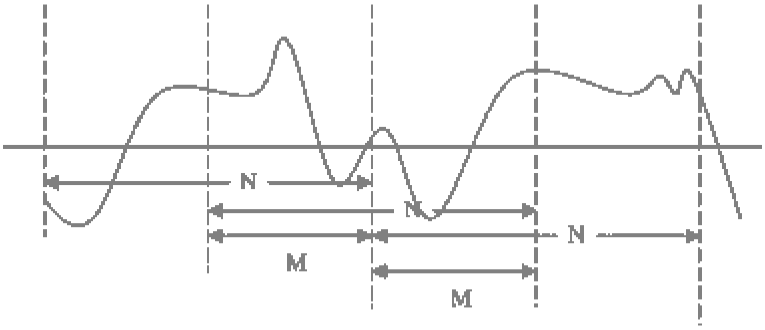


Figure 2.9 Framming.

Assuming a signal $S = \sum_{n=0}^{N_S-1} x[n]$, this can be mathematically formulated as follows:

$$S = \sum_{n=0}^{N_S} X[n] = \sum_{i=0}^{\#F-1} F[i] \tag{2.7}$$

Where S: discrete signal, $x[n]$: signal samples in time domain, N_s: signal length in samples, $f[i]$: signal frame, $\#F$: number of frames. Figure 2.10 intuitively shows the details of frame-level processing procedure: the original waveform contains 5 frames (showed by green blocks) and the overlap (hop length, showed by yellow blocks) is 1/2 frame length, finally get the framing features that will act as input features into models.

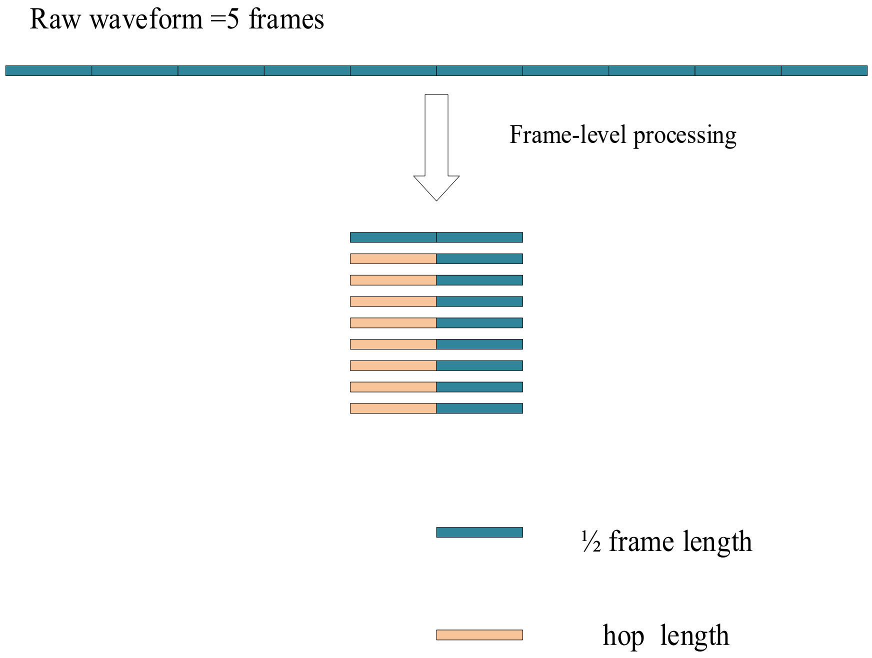


Figure 2.10 The details of frame-level processing procedure.

2.5 Target and labels

In some speech enhancement systems, the choice of target plays a critical role in determining the model's learning ability and system performance. Typically, the target is computed using pure speech and background noise signals. There are two commonly used types of speech enhancement targets: time-frequency masked targets and targets estimated based on the amplitude spectrum of clean speech. The former type reflects the energy relationship between the speech signal and the interference noise in the mixed signal, while the latter type represents the amplitude spectrum characteristics of clean speech. Over the years, researchers have made significant progress in improving speech enhancement targets and have proposed several new enhancement goals that have led to substantial improvements in performance.

It is worth noting that the choice of target should be made based on the specific application requirements and the characteristics of the target environment. For example, if the goal is to enhance speech in noisy environments, a target that emphasizes the reduction of interference noise may be more suitable. On the other hand, if the goal is to improve speech quality in teleconferencing or broadcasting applications, targets that are more closely related to perceptual quality may be preferred. Here is a brief summary and introduction of the main speech enhancement goals:

Ideal Binary Mask

The Ideal Binary Mask (IBM) has been widely used as a primary target in speech enhancement systems due to its effectiveness in improving the intelligibility of separated speech [71][72][73][75]. The IBM target is essentially a binary function that classifies each time-frequency unit of a signal as either speech or noise. To achieve this, the signal is first decomposed into a two-dimensional time-frequency domain representation. Then, for each unit in this representation, the IBM target is determined based on the ratio of speech energy to noise energy in that unit. If this ratio exceeds a certain threshold, the unit is classified as speech-dominant and assigned a masking value of 1. Otherwise, it is classified as noise-dominant and assigned a masking value of 0. The IBM target function can be expressed as follows:

$$IBM(t,f) = \begin{cases} 1, & if\ SNR(t,f) > LC \\ 0, & else \end{cases} \tag{2.8}$$

Where $SNR(t,f)$ represents the local signal-to-noise ratio of the time-frequency unit corresponding to the time frame t and frequency f; $\mathrm{IBM}(t,f)$ represents the ideal target of the time-frequency unit corresponding to $SNR(t,f)$ Masking value; LC stands for the set local threshold (in this paper, LC = 0). The set value of this parameter is generally smaller than the signal-to-noise ratio of the mixed signal. The purpose is to obtain a more adequate target speech energy spectrum.

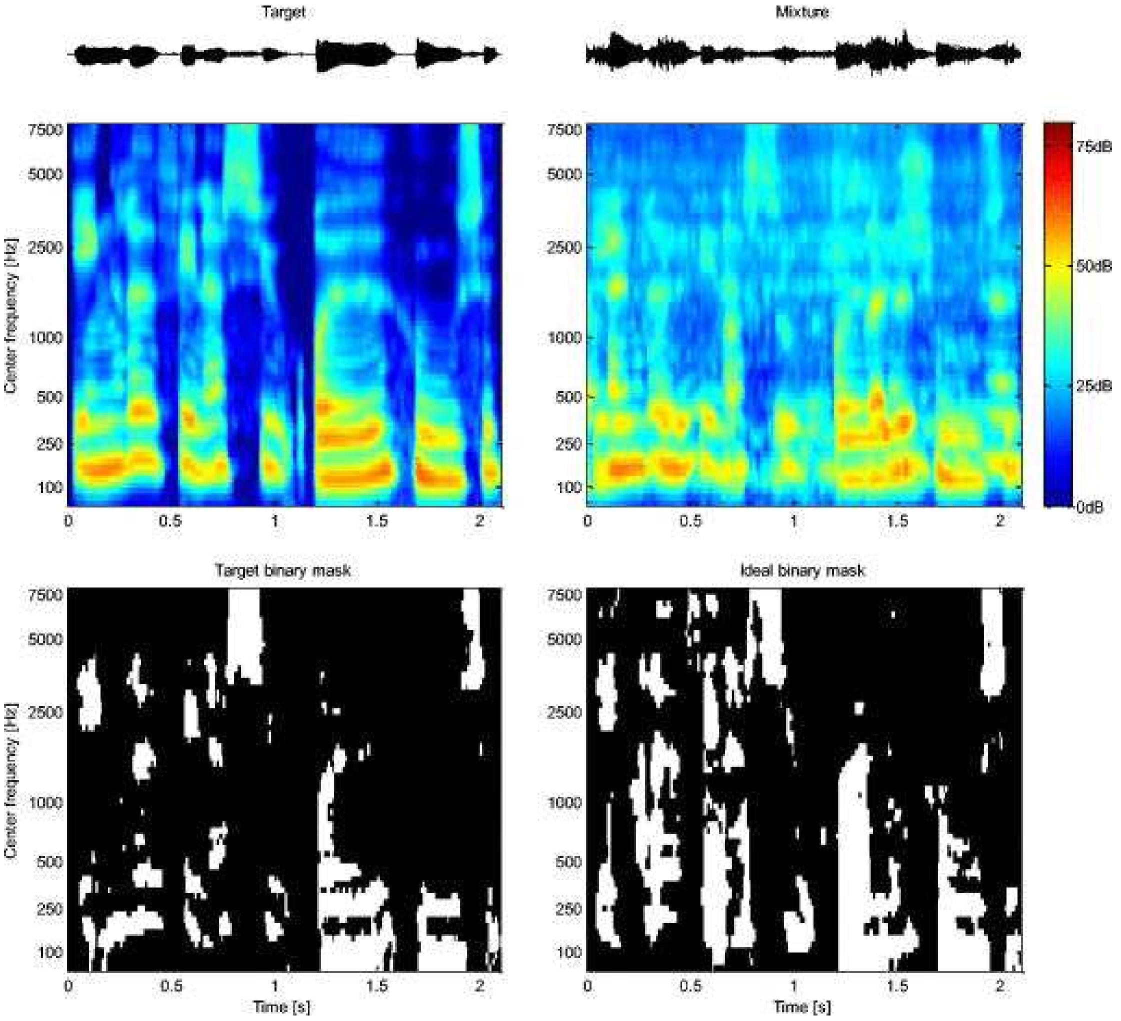


Figure 2.11 IBM used in speech enhancement.

While IBM has shown remarkable performance in speech enhancement, it has certain limitations. For instance, it assumes that speech and noise are independent in the time-frequency domain, which is not always the case in real-world scenarios. As a result, various modifications have been proposed to improve the effectiveness of IBM, including variations that use soft rather than hard decision boundaries to determine the masking values. In addition, other targets, such as the Minima Controlled Recursive Averaging (MCRA) and Weighted Prediction Error (WPE) targets, have been proposed to address some of the limitations of IBM and achieve better speech enhancement performance in different contexts. Figure 2.11 shows the IBM used in speech enhancement.[76]

Applications of IBM:

Speech Separation: IBM has been employed as a foundational technique in computational auditory scene analysis to segregate speech from complex auditory scenes [77].

Noise Suppression: In speech enhancement tasks, IBM can provide a binary guide to suppress time-frequency regions dominated by noise [80].

Limitations of IBM:

Oracle Information: The construction of an accurate IBM requires a priori knowledge of both the clean speech and the noise, which is usually unavailable in real-world applications [81].

Binary Nature: The binary approach of IBM can sometimes lead to the removal of speech components, especially in regions where speech and noise overlap [82].

Temporal Smearing: Due to the time-frequency analysis, IBM can introduce artifacts and temporal smearing, affecting the quality of the enhanced speech [83].

Performance with Non-stationary Noise: While IBM shows good performance with stationary noise, its efficiency reduces with non-stationary noises like babble or traffic noise [84].

Given these limitations, while IBM serves as a fundamental benchmark, newer methodologies and techniques have been proposed to overcome its constraints. In the context of deep learning, for instance, soft masks or ratio masks, which provide a continuous value instead of a binary decision, have been explored to offer more refined and nuanced speech enhancement [85].

Target Binary Mask (TBM)

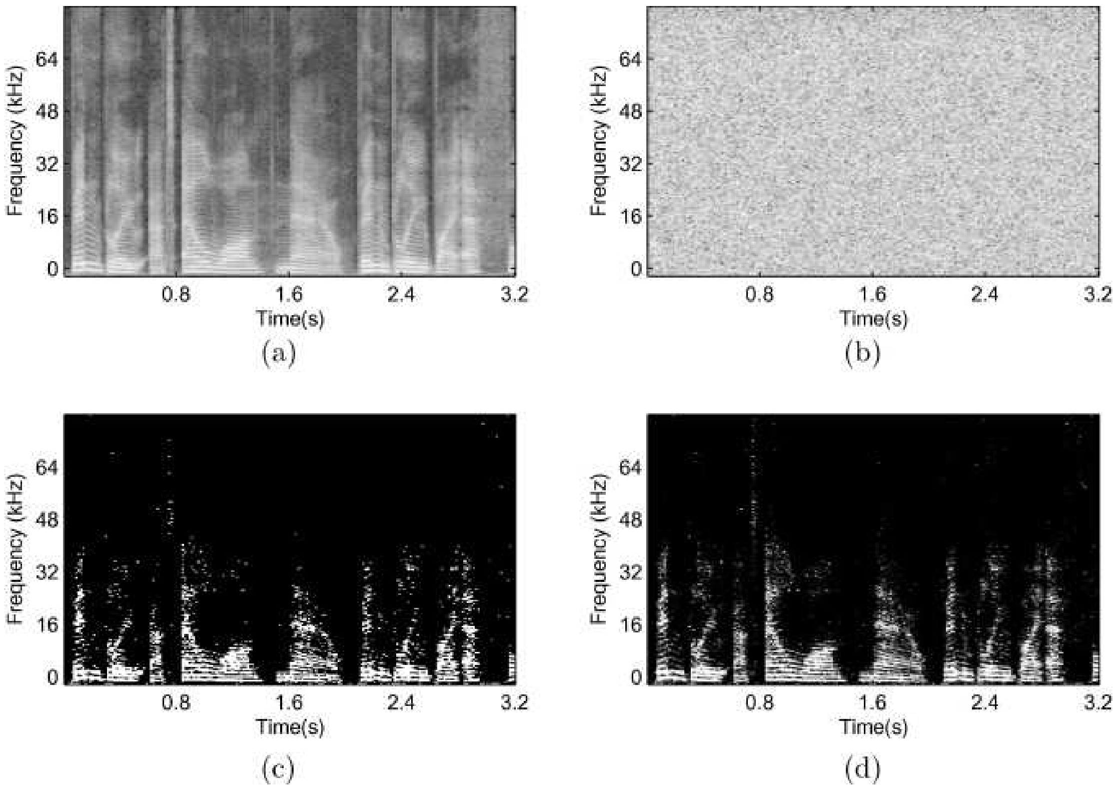
In speech enhancement problems, various targets have been proposed to improve the quality of the output speech. Time-frequency masking targets, such as IBM and TBM, have been proved to be effective in improving the intelligibility of separated speech. The TBM target is similar to IBM in that it is a binary matrix obtained by calculation. However, unlike IBM, TBM uses a fixed reference noise (speech-shaped noise) instead of the actual noise to calculate the binary matrix. This allows for independent calculation of TBM for interference noise. Retaining spatio-temporal structure information in the time-frequency domain of speech signals, TBM is conducive to improving speech intelligibility [79].

Ideal Ratio Mask

Wang et al. proposed the ideal ratio mask (IRM) as a soft function in the frequency domain signal for the first time, which is sometimes referred to as the ideal proportional mask [78]. Unlike the binary masking targets, IRM represents the desired ratio of the clean speech amplitude spectrum to the noisy speech amplitude spectrum. The IRM target value is calculated by dividing the clean speech spectrum by the sum of the clean speech and noise spectrum. The resulting IRM values range between 0 and 1, where 0 indicates that the frequency bin should be entirely masked (i.e., noise-dominated), and 1 indicates that the frequency bin should be preserved (i.e., speech-dominated). The IRM is commonly used in speech enhancement problems, and it has been shown to be effective in improving the quality and intelligibility of the enhanced speech [39][40][41]. Additionally, the IRM can be extended to a multichannel version, such as the ideal binary mask (IBM) and ideal amplitude ratio mask (IARM), which have been proven to be effective in enhancing speech signals corrupted by multiple interfering sources [42][43][25].

$$IRM(t,f) = \left(\frac{S^2(t,f)}{S^2(t,f) + N^2(t,f)}\right)^{\beta} = \left(\frac{SNR(t,f)}{SNR(t,f) + 1}\right)^{\beta} \tag{2.9}$$

Where $\mathrm{IRM}(\mathrm{t},\mathrm{f})$ represents the mask value of the time-frequency unit corresponding to time frame t and frequency f; $S^2(t,f)$ and $N^2(t,f)$ respectively represent the speech energy value and noise energy of the corresponding time-frequency unit value. β is an adjustable parameter used to control the range of target values. Wang et al. [74] have found that setting the threshold to 0.5 results in the best prediction performance for the IRM model, which is similar to the RMS Wiener filter. Compared with IBM, the two have their own strengths. The main difference is that the IRM target value is continuous, with a value range generally within the interval [0, 1], thus minimizing the loss of target speech energy. Figure 2.12 shows the IRM used in speech enhancement.[76]



(a)

(b)

(c)

(d)

Figure 2.12 IRM used in speech enhancement. The x-axis typically represents time, while the y-axis represents frequency. Each point (or pixel) in the image reflects a ratio value ranging from 0 to 1.

Complex Ideal Ratio Mask (cIRM)

The cIRM target [86] is an improvement of the IRM target that takes into account the phase information in addition to the energy in the amplitude domain. By considering the phase relationship between speech and noise, cIRM can effectively suppress noise and preserve more useful speech information, which is particularly important for speech enhancement in complex environments with high noise levels. The cIRM target is calculated as the ratio of the clean speech spectrogram to the noisy spectrogram in the complex domain. Unlike the IRM target, cIRM is a complex function with both magnitude and phase information, which can provide more accurate separation of speech and noise. Therefore, using the cIRM target as a training goal can significantly improve the performance of speech enhancement systems in terms of speech quality and intelligibility, especially in noisy and reverberant environments.

$$S = cIRM * M \tag{2.10}$$

Among them, S and M represent the STFT coefficients of pure speech and noisy speech, respectively, "*" represents the multiplication operation in the complex domain. According to the above formula we can get:

$$\mathrm{cIRM} = \frac{M_r S_r + M_i S_i}{Y_r^2 + Y_i^2} + j\frac{M_r S_i + M_i S_r}{Y_r^2 + Y_i^2} \tag{2.11}$$

Among them, Mr and Mi represent the real and imaginary parts of the noisy speech M respectively, Sr and Si represent the real and imaginary parts of the target speech S respectively, and j represents the imaginary unit. Therefore, the obtained cIRM is also a complex number with real part cIRMr and imaginary part cIRMi, namely:

$$cIRM_r = \frac{M_r S_r + M_i S_i}{Y_r^2 + Y_i^2} \tag{2.12}$$

$$cIRM_i = \frac{M_r S_i + M_i S_r}{Y_r^2 + Y_i^2} \tag{2.13}$$

In the complex domain, the cIRMr and cIRMi estimated by calculation usually exceed the range [0, 1], which makes the calculation more difficult. Therefore, it is usually necessary to use the sigmoid function or the hyperbolic tangent function for amplitude suppression. Figure 2.13 shows the cIRM used in speech enhancement [86].

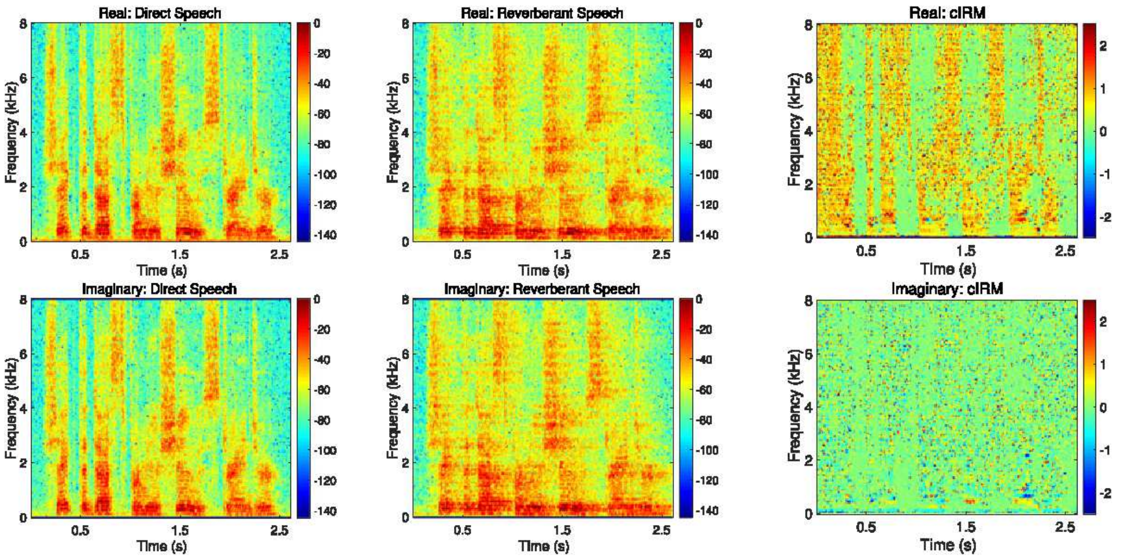


Figure 2.13 cIRM used in speech enhancement.

Spectral Magnitude Mask (SMM)

The SMM target, also known as FFT-MASK, is another widely used target in speech enhancement systems, which is calculated based on the short-time Fourier transform (STFT) coefficients of the noisy speech and target speech [78]. It is a soft target that reflects the amplitude spectrum ratio between the target speech and the noisy speech. The calculation of the SMM target is relatively simple, and its formula is as follows:

$$SMM(t,f) = max\left\{0, 1 - \frac{|S(t,f)|}{|M(t,f)|}\right\} \tag{2.14}$$

where $S(t,f)$ and $M(t,f)$ represent the amplitude of the target speech and noisy speech amplitude spectrum, respectively. Compared with the IBM target, SMM considers both the speech and noise energy, and is a continuous function. However, it does not consider the phase information, which may affect the performance of the speech enhancement system. Therefore, researchers have proposed more advanced targets, such as cIRM, to further improve the system performance.

Short-Time Fourier Transform Spectral Magnitude, FFT-Magnitude

FFT-Magnitude, also known as Target Magnitude Spectrum (TMS), is an amplitude spectrum estimation-based target that directly uses the amplitude spectrum of pure speech as the ideal target for speech separation. Unlike cIRM, it does not consider the phase information. In supervised speech separation, the time-domain signal is first windowed and then converted into a two-dimensional time-frequency signal using the Short-Time Fourier Transform (STFT) method. The amplitude spectrum of the noisy speech is then used to estimate the amplitude spectrum of the target speech through the model output. Finally, the phase information is combined using inverse transform technology to obtain the time-domain waveform [87].

Since FFT-Magnitude only considers the amplitude spectrum, it cannot guarantee perfect separation in complex noise environments, where the phase information plays a crucial role. However, it has been shown to be effective in simple noise environments, such as white noise or babble noise.

The Fast Fourier Transform (FFT) is a foundational algorithm in signal processing, and its magnitude representation offers a compact and computationally efficient method for describing a signal's spectral characteristics. In the realm of speech enhancement, various target representations have been proposed, including the Ideal Ratio Mask (IRM), complex Ideal Ratio Mask (cIRM), and FFT-Magnitude.

FFT-Magnitude's appeal lies in its simplicity and computational efficiency. While IRM and cIRM provide a more detailed mask for separating speech from noise, they necessitate additional computational steps and resources for mask estimation and application [89]. FFT-Magnitude, on the other hand, directly represents the spectral amplitude of the speech signal, sidestepping the need for intricate mask computations. This directness translates to faster processing times, which is especially valuable in real-time applications where computational speed is paramount [90].

Furthermore, FFT-Magnitude has been employed in various deep learning architectures for speech enhancement and has demonstrated promising results in terms of both objective metrics and perceptual evaluations [78]. Its widespread adoption in real-time systems can be attributed to its balance between computational efficiency and enhancement quality.

2.6 Waveform resynthesis

After obtaining the estimated target speech amplitude spectrum, it needs to be combined with the phase information of the noisy speech to generate the one-dimensional time-domain waveform signal of the separated speech. This is usually achieved through the inverse STFT (ISTFT) technique. The ISTFT takes the estimated target speech amplitude spectrum and the phase information of the noisy speech to generate the separated speech in the time domain.

The waveform synthesis process may also involve additional post-processing techniques, such as waveform gain adjustment and smoothing, to improve the overall quality of the separated speech. Taking IRM as an example, the calculation formula is as follows:

$$S(t,f) = M(t,f) \times IRM(t,f) \tag{2.15}$$

Where $M(t,f)$ represents the target speech amplitude spectrum and IRM(t, f) represents the characteristics of the mixed signal mask. If the estimated target is IBM, then the result of the above calculation is to retain the corresponding feature unit with a mask value of 1 for the time-frequency unit.

Then combined with the phase information in the mixed signal, the spectrum estimate of the target speech is calculated according to the following formula:

$$\hat{S}(t,f) = S(t,f) \times ej\angle M(t,f) \tag{2.16}$$

Where $\hat{S}(t, f)$ is the reconstructed target speech spectrum and $j\angle M(t, f)$ is the phase information of the mixed signal.

Finally, after obtaining the estimated target speech amplitude spectrum using the time-frequency mask or speech amplitude spectrum, the last step is to use a waveform synthesis technique to obtain the one-dimensional time-domain waveform signal of the target speech. The inverse transform method is commonly used for this purpose.

The inverse transform technique used depends on the specific target and processing method. The most commonly used inverse transform techniques include inverse Fourier transform and inverse gammatone filtering. The inverse Fourier transform is used for targets such as IBM, IRM, and NFFT-Magnitude, while inverse gammatone filtering is used for targets such as GF-POW.

In the inverse Fourier transform method, the estimated target speech amplitude spectrum is multiplied with the phase information from the mixed signal's STFT to obtain the complex-valued STFT of the target speech, which is then transformed back to the time domain using the inverse Fourier transform. In the inverse gammatone filtering method, the estimated gammatone domain amplitude spectrum of the target speech is converted back to the time domain using inverse gammatone filtering, which involves filtering the estimated spectrum with the inverse gammatone filterbank to obtain the time-domain waveform.

Overall, the choice of inverse transform technique is crucial in determining the quality of the separated speech, and it is important to select an appropriate technique based on the target and processing method used.

2.7 Deep learning neural network

Deep Neural Networks (DNNs) have undeniably revolutionized many domains of artificial intelligence, particularly in the realm of auditory signal processing. Their ability to model intricate patterns and relationships in data has made them particularly appealing for speech enhancement tasks, where the goal often involves isolating clear speech signals from noisy backgrounds.

Table 2.1 Overview of Research in Neural Networks for Speech Enhancement

Neural Network Type	Key Findings	Comments
Autoencoder [88]	Demonstrated efficacy in noise reduction	One of the early adaptations of deep learning for speech enhancement
RNN [114]	Successful handling of temporal speech	Highlighted the importance of sequential data modeling
CNN [151]	Effective in identifying noise patterns	Extended the application of CNNs beyond image processing
GAN [13]	Generated high-quality clean speech signals	Introduced a novel approach for speech enhancement using adversarial networks

2.7.1 The origin and introduction of the deep learning algorithm

Traditional DNNs consist of multiple layers of interconnected nodes and have demonstrated efficacy in various speech tasks. They are particularly adept at handling straightforward mappings from noisy to clean speech signals. However, their major limitation lies in their inability to handle sequential data, which is essential for capturing temporal dependencies in speech [1].

The high fault tolerance and non-linear capabilities of neural networks have made them increasingly popular among researchers and widely used in various domains, particularly in classification problems. However, it is important to note that all classification tasks rely heavily on the quality of the input features. Hence, the ability to extract highly informative and representative features from the data plays a critical role in determining the performance of artificial neural networks in achieving the desired results. In order to obtain accurate and reliable classification results, researchers need to devote significant efforts towards feature extraction and selection techniques that can effectively capture the relevant patterns and structures in the data. Only by carefully considering the feature extraction process can the full potential of artificial neural networks be realized in solving complex classification problems.

From an academic perspective, it is important to note that artificial neural networks are highly popular among researchers due to their high fault tolerance and nonlinear descriptive ability, particularly in classification problems. However, the success of artificial neural networks in achieving desired results is highly dependent on the extraction of highly representative data features, as classification problems are fundamentally based on features.

In our daily lives, we are presented with vast amounts of data, and extracting the main features from this data is crucial in achieving our goals. This ability to extract features from data is also a capability of the brain. Neurological scientists have found that the human brain is layered in its approach to receiving external information [78]. For example, in the visual system, the brain can recognize external object information not directly dependent on the retina, but rather through the information received and processed by the brain. This hierarchical processing of sensory signals in the brain greatly reduces the amount of information the nervous system has to handle while retaining the most useful information of the original signal, ensuring the brain's ability to quickly respond to the external world [91].

Inspired by the hierarchical processing of sensory signals in the human brain, it is feasible to design neural networks that process raw data hierarchically to extract useful information. In 2006 [92], Geoffrey Hinton and colleagues proposed a deep learning algorithm that simulates the learning process of the human brain through a multi-level neural network, mapping the original data and integrating the extracted features into a specific learning framework for classification or fitting purposes. The deep neural network originated from a simple perceptron algorithm, which is the basic neural unit of the artificial neural network, as shown in Figure 2.14.

The basic perceptron contains multiple inputs, an adder, an activation function and a output, and its operation express is:

$$\mathrm{ak} = \sum \mathrm{wi} * \mathrm{xi} + \mathrm{w0} = \sum \mathrm{wi} * \mathrm{xi} \tag{2.17}$$

$$\mathrm{y_k} = \emptyset(\mathrm{a_k}) = \emptyset(\mathrm{wTx}) \tag{2.18}$$

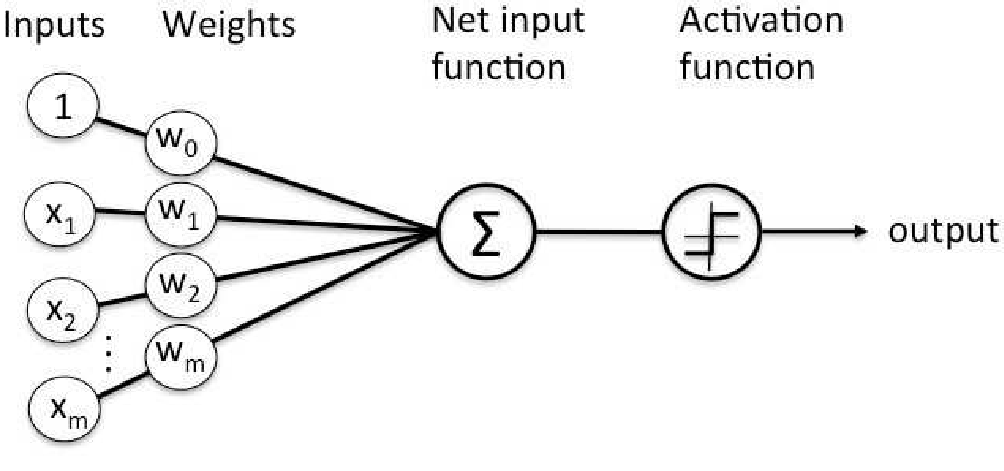


Figure 2.14 The perceptron model.

Where x denotes input feature vector, w denotes the weight vector, $w0$ denotes bias.

Activation functions play a crucial role in the neural network as they introduce non-linearity to the network, which helps in modeling complex relationships between the input and output data. There are several commonly used activation functions, such as the threshold function, sigmoid function, tanh function, softmax function, rectified linear units function (ReLU), and more. The choice of activation function depends on the specific requirements of the application and the task.

For example, in multi-class classification tasks, the softmax function is preferable as it outputs the probability distribution of the classes, while in regression tasks, such as speech enhancement, the sigmoid function is commonly used as it can map the input to a continuous output range between 0 and 1. The selection of an appropriate activation function is crucial to ensure the neural network can achieve satisfactory results for specific tasks.

2.7.2 Deep learning network structure and training method

The structure of the deep learning algorithm system is a crucial aspect of deep learning and is shown in Figure 2.15. This system comprises an input layer, multiple hidden layers, and an output layer. The hidden layers are denoted by S_i (i = 1,2, ..., N), and their number N can be set according to the complexity of the problem. The deep learning system can be represented as $I \geq S_1 \geq S_2 \geq \cdots \geq S_n \geq O$, where I is the input data, and O is the output data. This means that after deep network learning, the final output data is consistent with the original input data I.

During the deep learning process, each layer performs a feature mapping of the upper layer, and there is no loss of primary information. The output of any layer of the network is the input feature map in a multi-level learning process. This hierarchical learning process provides another representation of the original data, allowing for a series of hierarchical features to be extracted.

The deep learning algorithm's core idea is based on the extreme learning machine algorithm, which uses the hierarchical features of the input data to learn the mapping relationship between the input and output data. This way, we can obtain a series of hierarchical features of the original input data, which are then used to learn the complex mappings between input and output. The deep learning-based extreme learning machine algorithm is widely used in various fields due to its ability to learn complex non-linear mappings from high-dimensional data.

The deep learning network is known for its strong feature expression mapping ability [93], however, its initial development was slow mainly because deep learning training is extremely difficult. The traditional shallow network training methods such as gradient descent algorithms are no longer suitable for deep learning networks due to various reasons. Firstly, the gradient descent algorithm relies on labelled datasets, which are difficult to obtain in real-life scenarios. Without a sufficient sample size, the powerful feature expression ability that deep learning networks aim to achieve cannot be realized. Secondly, supervised learning methods such as the gradient descent algorithm are easy to train and obtain reasonable parameters for shallow networks with only one or two layers, but for deep networks, it is prone to falling into local extremum, leading to a poor network effect. Thirdly, when the network's depth is very high, using the back-propagation algorithm will make the network weights of the previous layers change slowly, resulting in ineffective learning and dispersion of the gradient. Therefore, traditional shallow network training methods are not applicable to deep learning networks, and instead, deep learning networks employ a new network training method called layer-by-layer greedy training method.

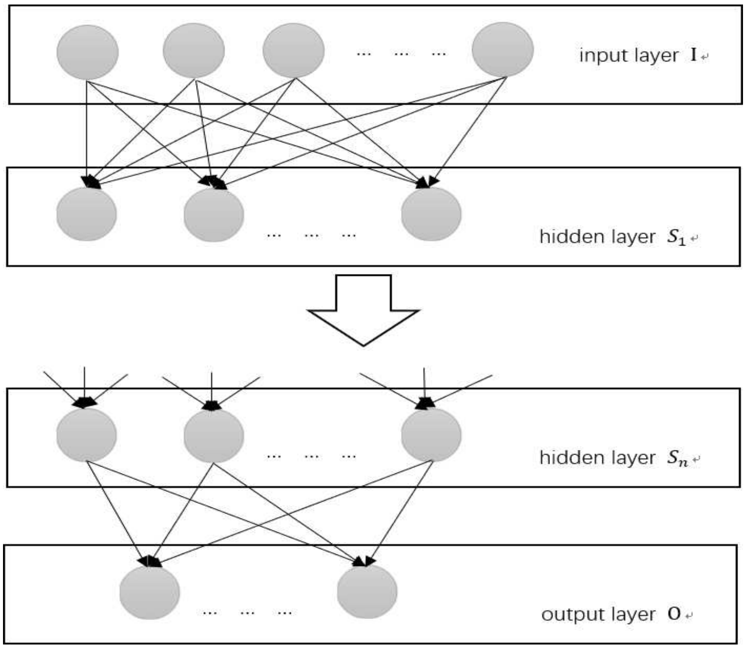


Figure 2.15 Deep learning system diagram.

The idea behind this training method is to train only one layer of the network at a time instead of training the entire network simultaneously. This means that the first layer of the hidden layer of the network is trained first, followed by the second layer, and so on. After training each layer, the previously trained network remains fixed. Each layer of training can use supervised or unsupervised training, but in most cases, unsupervised training is preferred. The weight parameters obtained by these individual training networks are then used as the initial values of the weights of the final network, which is fine-tuned to obtain the final weights. This layer-by-layer greedy training method has been proven effective in training deep learning networks and has greatly promoted the development of deep learning technology.

2.7.3 Deep learning algorithms applied in speech enhancement.

Long short-term memory recurrent neural networks (LSTM-RNNs) for SE

Recent studies, such as those by Hefei and Atlanta [94], have underscored the effectiveness of Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) in speech enhancement. Their findings suggest that LSTM-RNNs can yield significant improvements in objective measures related to both speech quality and intelligibility [96]. This is particularly notable considering the inherent limitations of deep neural networks (DNNs) in some contexts of speech enhancement. Specifically, while DNNs offer robust capabilities, they often fall short in capturing the interdependencies among adjacent frames in the long-term acoustic context [97].

Recurrent Neural etworks (RNNs) address this limitation by leveraging connections between prior and current units, effectively capturing long-term contextual information. Consequently, RNNs often outperform traditional DNNs in tasks that require the modeling of temporal sequences [98]. Figure 2.16 illustrates the fundamental structure of RNNs [99].

LSTM-RNNs, a variant of RNNs, have garnered attention in speech enhancement due to their adeptness at modeling temporal dependencies within sequential data [99]. Such capabilities render them highly effective for tasks like speech recognition, wherein input data comprises a time-series of acoustic features. The inherent architecture of RNNs, especially LSTM-RNNs, facilitates the processing and retention of temporal information, making them exceptionally suited for speech signal processing [100]. Moreover, recent innovations, including bidirectional RNNs and attention mechanisms, have further amplified their efficacy in speech enhancement endeavors[95].

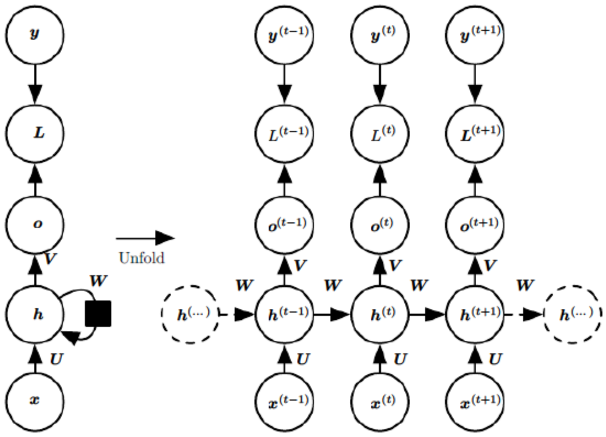


Figure 2.16 The basic structure of RNN.

However, despite its ability to capture long-term context, RNNs are not without their limitations. One major challenge arises when updating their hyperparameters with the backpropagation algorithm, which can lead to vanishing or exploding gradients [101]. To address this issue, the

Long Short-Term Memory (LSTM) RNN architecture was proposed, which incorporates a memory cell and various gates (shown in Figure 2.17 [103]) to control information flow. These additions allow the LSTM-RNN to effectively mitigate the vanishing and exploding gradient problems associated with traditional RNNs. In fact, studies such as Pascans and Mikolov [94] have demonstrated that the LSTM-RNN consistently and significantly improves both speech quality and intelligibility, particularly in noise reduction at low signal-to-noise ratios (SNRs), outperforming traditional DNNs.

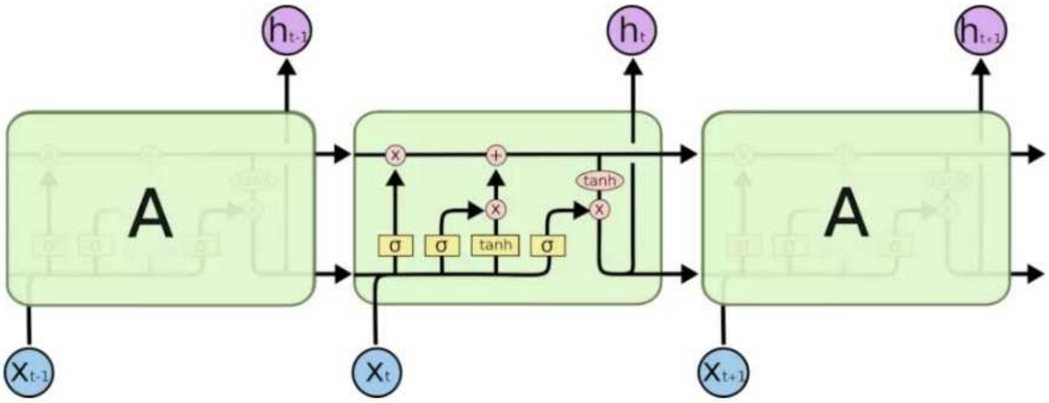


Figure 2.17 The diagram of LSTM.

Convolutional neural networks (CNNs) for speech enhancement

In recent years, deep neural networks have been widely used for identifying and removing noise from noisy speech. There are two main approaches for achieving this: (1) estimating masks, such as ideal binary mask (IBM) and ideal ratio mask (IRM), which can be used to obtain the speech spectrum, and (2) directly estimating the clean speech. CNNs have primarily been championed for image processing but have found applications in speech enhancement, especially when treating spectrogram images. They can capture local patterns and hierarchies in data, making them adept at identifying noise patterns in audio signals. Their limitation, however, is their insensitivity to global patterns or long-range dependencies. Convolutional models have been found to be effective in noise reduction due to their ability to capture the strong temporal correlations in speech and their invariance to translational variance caused by different speaking styles [102].

(1) Convolutional layer: Existing DNN-based enhancement methods typically add a convolution layer and a pooling layer between the input layer and the hidden layer to automatically learn the deep features hidden in the data. The convolutional layer is responsible for feature extraction, and the "incomplete connection, parameter sharing" feature greatly reduces the number of network parameters, ensures the sparsity of the network, and prevents overfitting. The reason why "parameter sharing" is possible is that the samples have locally related characteristics. Suppose that each feature map in the input layer $Fi(i = 1,2, \dots, I)$ is connected to multiple feature maps in the convolutional layer $Ci(i = 1,2, \dots, J)$. I and J are the number of

feature maps in the input layer and the convolutional layer, respectively. Therefore, the size of local weight matrices wij(i = 1,2, … , I, j = 1,2, … , J) should be *I* × *J*. This mapping relationship is what we often say about convolution operations, and it has been widely used in the field of signal processing. The calculation formula of each node in the convolutional layer is:

$$c_{j,m} = \sigma(b_{j,m} + \sum_{i=1}^{I}\sum_{n=1}^{K}\left(f_{i,m+n-1}w_{i,j,n}\right) \qquad (2.19)$$

Where *j = 1,2, …, J, m = 1,2, …, M*. In the above equation, σ represents the activation function. cj,m is the *m-th* unit of the *J-th* feature map in the convolution layer C_j . $b_{j,m}$ is the corresponding bias of the feature map; M is the number of nodes in the convolution layer C_j and $f_{i,k}$ is the *k-th* node in the *i-th* input feature map; $w_{i,j,n}$ represents the *n-th* element of the weight matrix $w_{i,j}$ which is the weight matrix connecting the *i-th* input feature map and the *j-th* input feature map. Whist K is the size of the convolutional kernel and represents the number of connections between each feature node in the convolutional layer and the nodes in the input feature graph.

(2) Pooling layer: The pooling layer is an essential component of deep neural networks used in speech enhancement. It comprises functions such as MaxPooling and AveragePooling, but MaxPooling is more commonly used due to its ability to reduce the size of the convolution kernel while retaining corresponding features. It achieves dimensionality reduction by operating on several nodes in a local area of the convolutional layer feature map to extract main features. Consequently, the pooling layer has the same number of feature maps as the convolutional layer but with smaller feature dimensions. Its role in reducing feature resolution allows it to generalize the upper layer feature map. This operation is invariant to displacement in the time-frequency domain, thereby mitigating the effects of positional movement and change. This property is critical in handling different speakers and unstable noise environments in speech enhancement.

The CNN model has demonstrated its capability in applying one-dimensional convolution pre-processing operation on frame-level features and utilizing the interdependencies among adjacent frequency bands in each time frame. The full-time sharing of convolutional layers within the CNN network also provides advantages in reducing the number of parameters when compared to a DNN network with the same depth. Consequently, it can effectively improve the training efficiency of the model. This feature makes the CNN model an ideal candidate for speech enhancement tasks where computational efficiency is a key factor.

In [102], the CNNs consists of 5 layers, 2 convolutional with a max-pooling layer in-between and 2 fully connected layers on the top. The corresponding convolutional layer's size is 5*1 and the size of pooling layer is 3*1. The Figure 2.7.5 shows the basic structure of the CNNs [10].

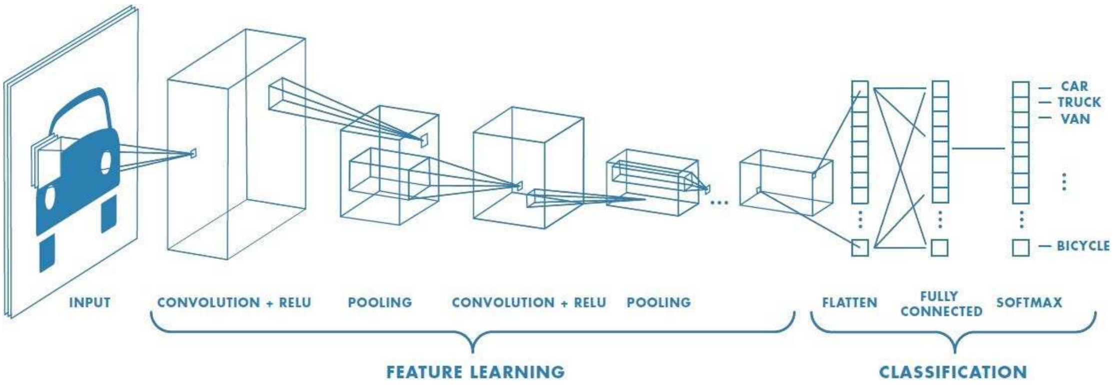


Figure 2.18 The basic diagram of the CNN.

2.7.4 Evaluation metrics

Speech enhancement systems are widely used to improve speech intelligibility and quality in various applications, such as teleconferencing, hearing aids, and voice-controlled devices. Evaluating the performance of such systems is essential to ensure their effectiveness in real-world scenarios. In general, two types of evaluation methods are used: subjective and objective.

Subjective evaluation methods, centered around human perception, remain an indispensable tool in assessing the quality of enhanced speech. Often, these evaluations involve a carefully selected group of listeners, sometimes including those with hearing impairments, who are asked to subjectively rate or judge the quality, clarity, and naturalness of the enhanced speech [104].

While it's true that such methods can be time-consuming and may be influenced by varying subjective factors, their significance in the speech enhancement domain is undeniable. In fact, subjective evaluations are frequently considered the gold standard in speech quality assessment. The International Conference on Acoustics, Speech, and Signal Processing (ICASSP) has even incorporated them into their challenges, underscoring their importance [105].

The strength of subjective evaluations lies in their ability to capture the intricacies of human auditory perception. Objective metrics, while valuable, cannot always fully represent the nuances of how humans perceive and interpret sound. By integrating feedback directly from listeners, researchers and engineers can obtain a richer understanding of the real-world applicability and effectiveness of their enhancement techniques.

Standards for conducting subjective evaluations, such as the Mean Opinion Score (MOS), have been established to ensure consistency and reliability [106]. MOS, in particular, offers a numerical indication of the perceived quality of speech, usually on a scale from 1 (bad) to 5 (excellent). Such standards help in mitigating the potential biases and variations that can arise in subjective evaluations, making them a crucial component of comprehensive speech enhancement assessment.

Objective evaluation methods use computer-based algorithms to measure the performance of the voice enhancement system. This method compares the original pure signal with the separated estimated signal to determine the system's performance using metrics such as spectral distance and signal-to-noise ratio. This method is flexible, cost-effective, and widely used in practice.

Objective evaluation methods provide a quantitative measure of the system's performance based on mathematical criteria. Common metrics include Signal-to-Noise Ratio (SNR), Segmental SNR, Short-Time Objective Intelligibility (STOI) [109] and Perceptual Evaluation of Speech Quality (PESQ) [87]. These metrics offer a consistent and reproducible means of evaluation, devoid of the variability and biases inherent in human perception. Moreover, objective methods can be rapidly deployed, enabling quick comparisons among various enhancement techniques.

The PESQ index is a widely used objective evaluation method that simulates the subjective perceived quality of speech. It is designed to model subjective tests commonly used in telecommunications and employs true voice samples as test signals. The value range of the PESQ indicator is generally between -0.5 and 4.5, with higher values indicating better voice quality [107]. To ensure proper application of voice test samples, guidelines are defined in the PESQ application guide contained in Recommendation ITU-T P.862.3 [110].

Short-Time Objective Intelligibility (STOI) is another commonly used objective evaluation method for evaluating speech intelligibility. This popular state-of-the-art speech intelligibility estimator relies on the linear correlation of speech temporal envelopes. The algorithm is based on pure speech and noise-reduced speech to construct a function for both. The literature [108] has shown that the STOI index is highly correlated with the actual human intelligibility of speech. The output of the algorithm is a scalar value with a value range of [0,1]. A higher value indicates better enhancement and higher intelligibility of the resulting speech. Compared with the subjective evaluation method of the human ear, using objective intelligibility to evaluate the quality of noise-reduced speech can significantly reduce computation time and calculation cost.

The selection of multiple objective evaluation indices provides a comprehensive and reliable assessment of the performance of the speech enhancement system, which is essential for further improvement and optimization of the system. However, while objective measures are valuable, they don't always correlate perfectly with human auditory perception. That is, a method that scores highly on an objective scale may not necessarily sound better to a human listener. This is where subjective evaluation methods, like Mean Opinion Scores (MOS), come into play [104]. They incorporate human listeners' judgments, capturing the nuances and intricacies of human perception.

In summary, while objective evaluation methods provide a consistent and bias-free measure of a speech enhancement system's performance, they should be used in tandem with subjective evaluations to ensure the enhanced speech not only meets mathematical criteria but also aligns with human auditory perception.

2.8 Deep learning theories in speech enhancement

Speech enhancement is a crucial process that involves removing background noise, isolating target speech, and minimizing signal distortion. There are two primary application scenarios for speech enhancement: the first is to enhance the voice of multiple speakers, while the second is to reduce background noise while preserving the clarity of speech [112][113]. This paper focuses on the latter scenario, which is to separate clean target speech from mixed signals with interference noise.

Speech enhancement can be viewed as a classification problem, which has led to the widespread use of data-driven methods in this field. In recent years, with the advancement of computing power, the performance of voice enhancement systems has been significantly improved by increasing the volume of training data. Based on the dependence of the model on prior information of the sound source signal, speech enhancement can be divided into two types: supervised learning and unsupervised learning. In the unsupervised learning model, the target speech signal is unknown, as is the case with blind source enhancement. Enhancement of signal sources is mainly based on acoustic characteristics or linear transformation. On the other hand, supervised speech enhancement is achieved by learning the nonlinear mapping relationship from mixed signals to clean speech, with known target signal sources. Figure 2.19 describes the general flow chart of a supervised speech enhancement system.

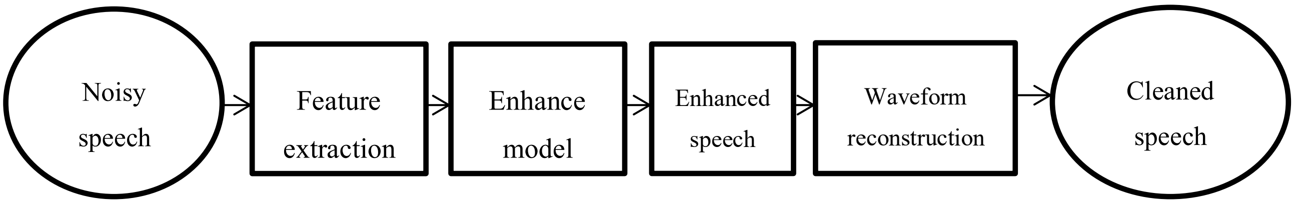


Figure 2.19 Supervised speech enhancement system.

Figure 2.20 depicts the structural block diagram of a supervised speech enhancement system, which consists of several main functions such as speech data synthesis, speech feature extraction, model construction, training, and waveform reconstruction. In the data pre-processing stage, a sufficient amount of pure speech and various types of noise need to be collected, and then mixed speech signals are synthesized with a predetermined signal-to-noise ratio (SNR). Next, the mixed speech signals are converted into their corresponding time-frequency representations. Various feature extraction methods are then used to extract the necessary feature information from the mixed time-frequency signals, which are used to construct the training and testing datasets for the machine learning model.

The primary output of the machine learning model is the ideal binary masking matrix (IBM) or other types of masking matrices. These matrices are mainly calculated from the energy of pure speech and noise. Finally, the model is continuously optimized by training and updating the model using the predicted masking matrix and the corresponding mixed signal. Mathematical operations are then performed to obtain the final processed speech, and the time-domain waveform signal is reconstructed using waveform reconstruction methods.

The proposed supervised speech enhancement system [114] leverages machine learning algorithms to learn the nonlinear mapping relationship between mixed signals and clean speech. By using a large volume of training data and optimizing the model continuously, the performance of the speech enhancement system can be significantly improved. Such systems find applications in various scenarios where clean target speech needs to be separated from mixed signals with interference noise, including telecommunication, speech recognition, and hearing aid devices.

In their study [111], Mao and Wang proposed a speech enhancement model that directly maps from noisy to clean speech in the log-power spectral (LPS) domain. To address the problem of vanishing and exploding gradient, the model employs Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) with input gate, forget gate, output gate, and peepholes. These components enable the network to dynamically control the information flow, allowing for effective learning and utilization of temporal information in the acoustic context. In their approach, the Ideal Ratio Mask (IRM) was selected as the target and Mean Square Error (MMSE) was used as the objective function.

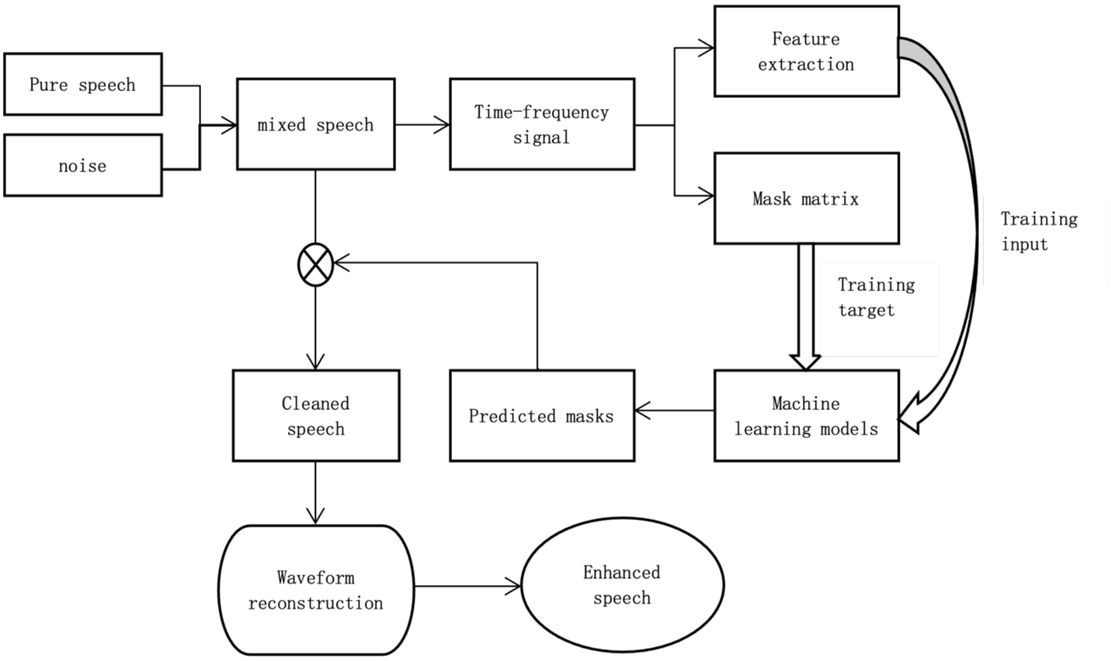


Figure 2.20 The structural block diagram of the supervised speech enhancement system.

2.9 Apply Generative Adversarial Network (GAN) to Speech enhancement.

Generative Adversarial Network (GAN) is a type of generative model that was introduced by Goodfellow et al. in 2014 [115]. GAN consists of two main components: the generator G and the discriminator D. The generator G aims to generate synthetic samples that are similar to the real data samples, while the discriminator D tries to distinguish between real samples and generated samples. In other words, the generator and discriminator are trained simultaneously, with the discriminator providing feedback to the generator in terms of how realistic the generated samples are.

The GAN framework has been widely used in various fields, such as image processing, natural language processing, and speech enhancement, due to its ability to generate high-quality synthetic data. One of the main advantages of GAN is its flexibility, as it allows for the generation of complex and diverse data. However, one of the challenges of GAN is to find the Nash equilibrium between the generator and discriminator, as the optimization process can become unstable. Despite this, GAN remains a popular choice for many generative modeling tasks due to its powerful and flexible nature.

Unlike the aforementioned methods for enhancing speech from noisy to clean/perturbed speech/mapping functions to signal-to-noise ratio, generative models for speech enhancement focus on learning the prior distribution of clean speech in order to capture temporal or spectral features of the speech. The successful application of these algorithms in the field of speech synthesis has prompted researchers to explore the potential of generative models in speech enhancement tasks. After all, speech enhancement can also be seen as a generative problem, in addition to being viewed as a filtering (separation) problem.

The Generative Adversarial Network (GAN) undoubtedly stands out as one of the most remarkable generative models in recent years, drawing attention from the entire deep learning community for its remarkable ability to produce highly realistic outputs in the field of computer vision. The introduction of GANs into the domain of speech enhancement was pioneered by Pascual et al. with their SEGAN [13] model. The motivation behind using GANs can be attributed to the limitations of non-GAN models in speech enhancement, as the choice of loss functions fails to effectively measure perceptual similarity, which remains a challenge even today. This inevitably constrains speech enhancement models within certain boundaries. However, utilizing a discriminator to quantitatively discern the similarity between enhanced speech and clean speech based on data-driven methods offers compensation for the shortcomings introduced by current loss functions. This compensatory approach avoids the dilemma encountered when optimizing networks solely based on speech enhancement evaluation metrics, which often only result in noticeable improvements in those specific metrics.

What is the Nash equilibrium?

Within the framework of game theory, the Nash equilibrium stands as a pivotal concept, offering a lens through which the strategic interactions between multiple players in a non-cooperative game can be scrutinized. This equilibrium is christened after John Forbes Nash Jr., whose seminal contributions to game theory have been universally acknowledged [116]. At the point of Nash equilibrium, every player, considering the anticipated strategies of their counterparts, solidifies their own strategy. Notably, once this equilibrium is achieved, no player stands to gain by unilaterally deviating from their chosen strategy, holding the strategies of others constant [7].

In the context of GAN, the generator and discriminator play a non-cooperative game in which the generator seeks to generate synthetic data that closely resemble the real data while the discriminator tries to distinguish between the real and synthetic data. The two processes are optimized iteratively in a way that the generator learns to generate more realistic samples, while the discriminator improves its ability to distinguish between the real and synthetic samples. Through this iterative optimization process, the GAN seeks to find a Nash equilibrium between the generator and the discriminator, where neither process can improve its performance by unilaterally changing its strategy.

In this minimum- maximum optimization, the optimization objective function of the generator and discriminator is:

$$\min maxE_{GAN}(G,D) = \ E_{x\sim P_x}[\log D(x)] + E_{z\sim P_{G(z)}}[\log(1 - G(z))] \qquad (2.20)$$

Where, x is the real data sample obeying $Pdata(x)$, and G(z) represents the data sample of the distribution $PG(z)$. In this project, it is specifically expressed as a real voice signal and a voice signal generated by a generator. Assuming that there is an additional constraint vector y as auxiliary information, the generator G(z, y) generates the speech signal under the constraint of y; In the same way, the discriminator D(x, y) can also discriminate the voice signal under the constraint of y [117], the objective function is converted into:

$$min_G max_D E_{GAN}(G,D) = E_{x,y\sim P_{data(x,y)}}[logD(x,y)] + E_{z,y\sim P_{G(x,y)}}[1 - D(G(z),y))] \qquad (2.21)$$

Actually, the goal of the generator is to ‘deceive’ the discriminator and enable the discriminator cannot distinguish between real data and generated data. Therefore, when training the generator, we want to minimize this error, and at the same time we try to maximize it for the discriminator. The Figure 2.21 shows the general structure of GANs, like [13], applied for the speech enhancement.

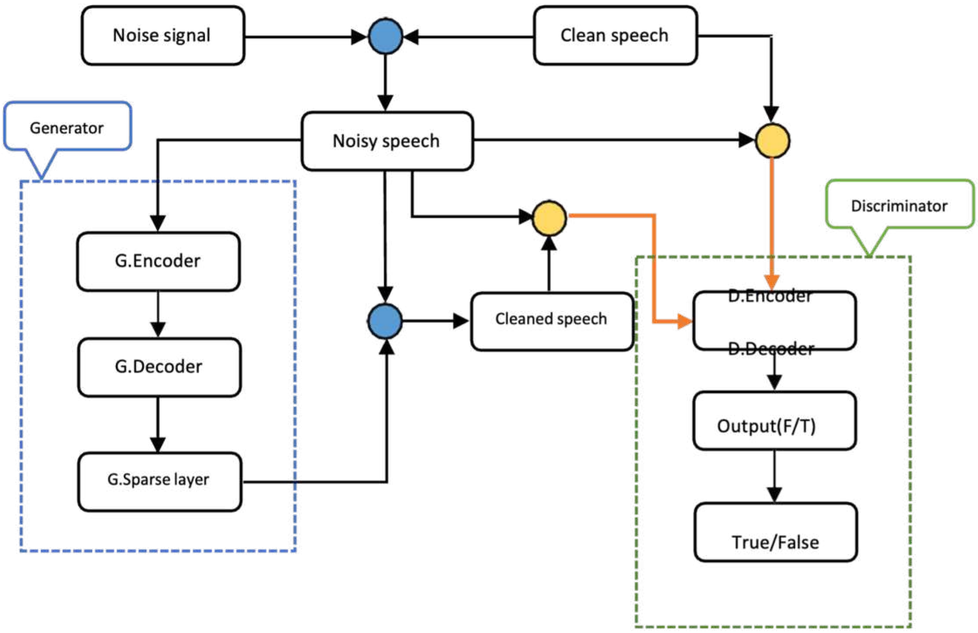


Figure 2.21 The diagram of the GANs applied for the speech enhancement.

2.9.1 The generator

The left part of Figure 2.21 shows the architecture of the generator used in GAN, which mainly comprises an encoder, a decoder, and a sparse layer. As discussed in Chapter 2, convolutional neural networks (CNNs) are better suited for learning abstract features in speech compared to deep neural networks (DNNs). Hence, CNNs are used as the building blocks for the encoder and decoder in the proposed generator. Although increasing the number of CNN layers, such as convolutional and pooling layers, can improve performance, it can lead to overfitting and an increase in the number of model parameters, resulting in higher training costs. To address this issue, a residual network can be introduced, which adds the input and output of a specific layer as the input to the next layer of the network (as shown in Figure 2.22). This technique helps to speed up the training process [118]. Additionally, the Inception network (shown in Figure 2.23, sourced from Inception v2 [120]) can be used to control the total number of parameters in the network. This architecture has been shown to improve the utilization of computing resources inside the network and has achieved good performance in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC 2014) [119]. Furthermore, in many cases, it is challenging to determine which type of CNN kernel to use. However, the Inception network can combine all kernels and concatenate their corresponding results to obtain the final output, leading to better performance.

In our speech enhancement project, the input of the generator is processed/extracted speech signal, and the generator will output enhanced speech signal, fed into the followed discriminator, which will be introduced next, to make a discrimination. In the past, training deep neural networks has been challenging due to high training costs, overfitting, and accuracy issues. However, the proposed method addresses these problems by utilizing an inception network and stacking various convolutional layers to better capture information across frame levels, leading to improved speech intelligibility.

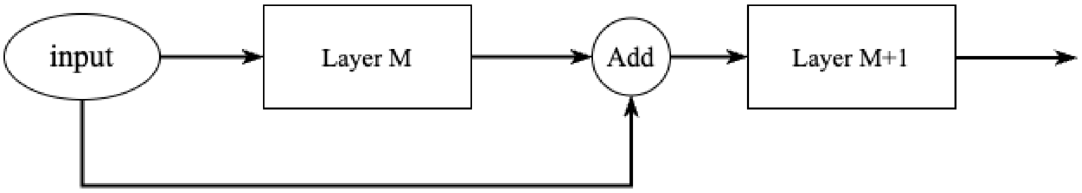


Figure 2.22 The residual network basic structure.

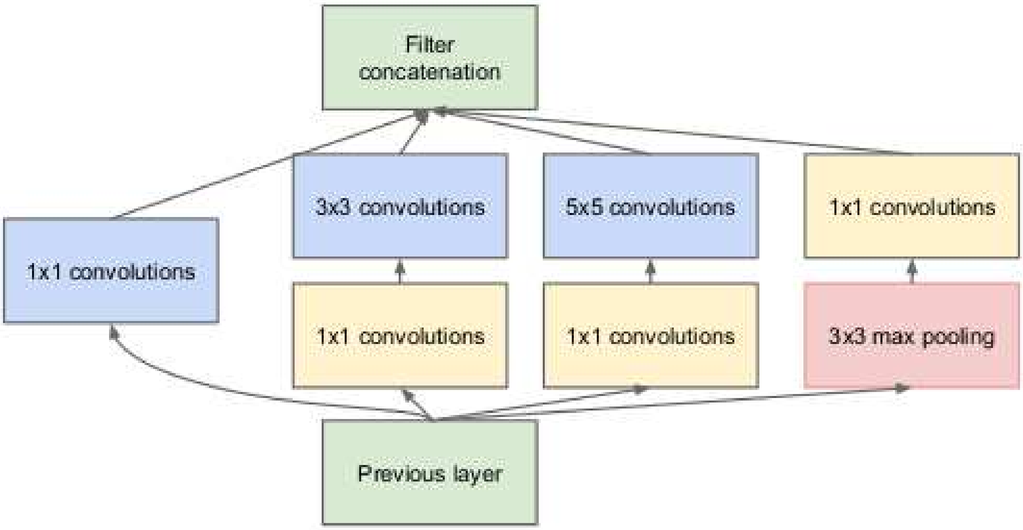


Figure 2.23 The basic unit of the Inception network.

2.9.2 The discriminator

In this GAN applied in speech enhancement, the generator produces time-frequency features, which represent the enhanced speech signal. However, these features are corrupted by the noise signal. The discriminator plays a crucial role in distinguishing between true and false speech signals, allowing the generator to learn and enhance the target speech signal by reducing noise.

To achieve optimal performance, the discriminator outputs a value within the range [0,1]. The true and false speech signals are determined by a threshold set beforehand. Two types of inputs are used: one is the combination of the generator's speech and noisy speech, while the other is the combination of the clean speech and the noisy speech.

To obtain the optimal time-frequency features, the generator and discriminator are iteratively optimized, resulting in an estimate of the magnitude spectrum of the target speech signal. By combining the phase spectrum of the mixed signal and applying the Inverse Short-Time Fourier Transform (ISTFT), the time-domain signal can be reconstructed.

This method [13] can effectively reduce the negative impact of noise on speech signals, improving speech intelligibility in various applications such as speech recognition and hearing aids. However, it is important to note that the choice of hyperparameters, such as the threshold for the discriminator, can have a significant impact on the performance of the method and should be carefully selected and tuned.

2.9.3 Existing Generative Adversarial Networks (GANs) Variations for Speech Enhancement

Generative Adversarial Networks (GANs) and their variants have marked a significant breakthrough in the domain of speech enhancement (as shown in Figure 2.9.4 [122]). Among these, CycleGAN [121] and SEGAN [13] are particularly noteworthy for their innovative applications to this field.

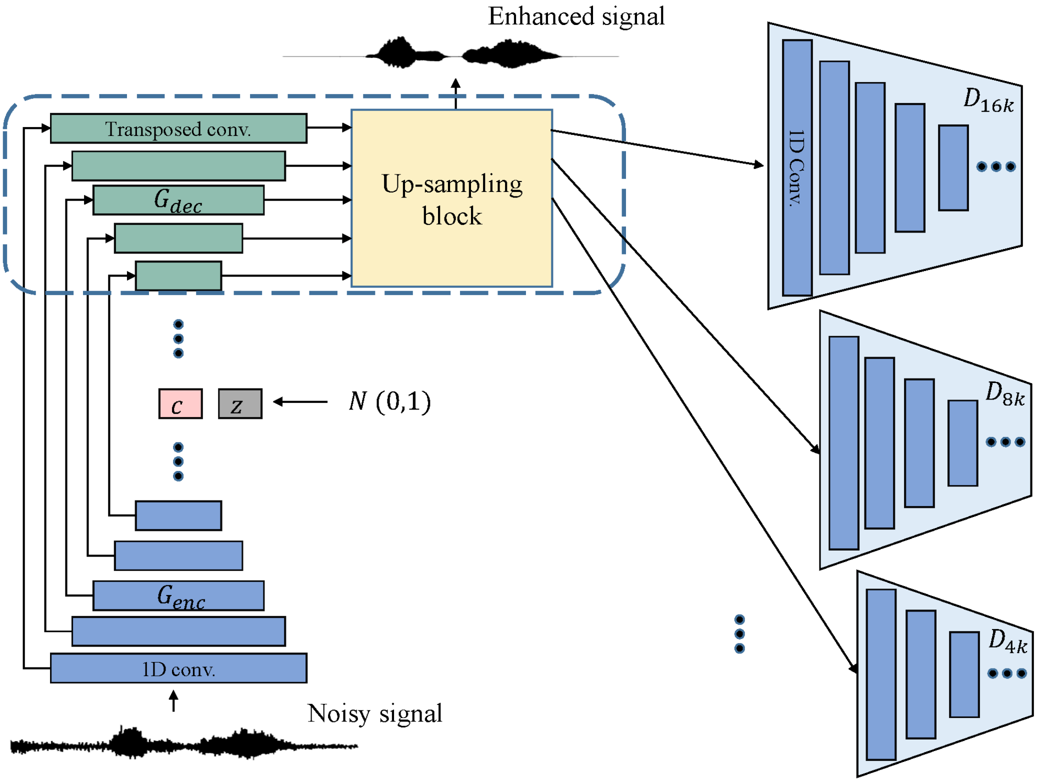


Figure 2.24 The GAN for speech enhancement.

CycleGAN has been heralded for its utility in unsupervised speech enhancement. Unlike traditional models that rely on paired examples of clean and noisy speech, CycleGAN ingeniously learns a bijective mapping between the clean and noisy speech distributions. This alleviates the need for paired examples, making it an attractive option when such paired datasets are scarce or unavailable [123].

StarGAN, in contrast, offers a multi-domain speech enhancement solution. It possesses the unique capability to disentangle domain-specific attributes from domain-independent ones. By subsequently recombining these factors, SEGAN can seamlessly transition a speech signal across various domains. This versatility is especially crucial when dealing with diverse noise environments or when aiming for multifaceted enhancement objectives [124].

The empirical success of both CycleGAN and SEGAN in the realm of speech enhancement is well-documented. Comparative analyses have underscored their ability to outperform traditional methods, particularly in terms of speech clarity and reduction of background noise [125].

Apart from SEGAN and CycleGAN, another notable model in the realm of speech enhancement utilizing Generative Adversarial Networks (GANs) is the Wave-U-Net architecture proposed by Stoller et al. (2018) [126]. Wave-U-Net is a GAN-based neural network designed for speech enhancement tasks, specifically focusing on source separation and denoising of audio signals. It combines the U-Net architecture with GAN components to enhance the quality of speech signals efficiently.

Additionally, the Parallel WaveGAN model introduced by Yamamoto et al. (2020) [127] is another significant advancement in speech enhancement using GANs. Parallel WaveGAN leverages GANs for high-quality speech waveform generation while addressing issues such as mode collapse and instability often encountered in GAN training. This model has shown exceptional performance in generating high-fidelity and natural-sounding speech.

These models, including Wave-U-Net and Parallel WaveGAN, represent further advancements in utilizing GANs for speech enhancement tasks, showcasing their efficacy in improving speech quality and addressing noise issues in audio signals.

Chapter 3 Generative Adversarial Network in Speech Enhancement

3.1 Background

In the dynamic field of speech enhancement, the integration of Generative Adversarial Networks (GANs) has been a significant milestone, with the SEGAN model by Pascual et al. [47] leading the charge. The appeal of GANs arises from the limitations of non-GAN architectures, particularly their reliance on loss functions that inadequately capture perceptual fidelity. Traditional loss metrics fail to holistically gauge the qualitative aspects of speech, a gap that GANs aim to fill.

Traditional speech enhancement models are limited by their dependency on loss functions that do not effectively measure perceptual similarity— a persistent challenge in the field. This creates a measurable discrepancy in performance, as these models are often optimized for specific evaluation metrics that do not necessarily translate to perceptibly improved speech quality.

Against this backdrop, GANs have emerged as a revolutionary approach. By employing a discriminator to gauge the likeness between enhanced and clean speech, GANs offer a data-driven strategy to circumvent the limitations posed by conventional loss metrics. This discriminator serves as a judge, ensuring that the enhanced speech not only scores high on traditional metrics but also possesses the qualitative attributes of clean, natural speech.

This chapter examines the application of GANs to speech enhancement tasks, adopting alternative deep learning models as generators in place of those used in the SEGAN framework. At the same time, we also explored how much performance improvement different deep learning models can bring as generators and compared each corresponding model as a baseline model. This exploration focuses on using magnitude as opposed to waveform for input and output features, which confers the benefits of reduced model complexity and computational demand. Such a shift paves the way for more efficient speech enhancement without compromising on performance.

Utilizing magnitudes over waveforms implies a strategic move towards simplicity and efficiency. By refining the input and output feature space to magnitudes, researchers have effectively cut down on the parameter overhead and computational costs typically involved in handling raw waveform data.

This chapter underscores the novel trajectory that speech enhancement has embarked upon with the introduction of GAN-based models. The upcoming experimental chapters will delve into the practical applications of these theoretical advancements and investigate the specific research questions pivotal to this exploration. The efficacy, efficiency, and perceptual quality improvements offered by magnitude-based GAN models over their waveform-based predecessors will be thoroughly analyzed, aiming to set new benchmarks for speech enhancement technology. The experiment was carried out using Librispeech (clean speech) in conjunction with NoiseX-92 (background noise) datasets, covering a Signal-to-Noise Ratio (SNR) range from -2 dB to 6 dB.

3.2 Methods and structure

In the realm of audio signal processing, the challenge of enhancing speech, especially amidst noisy environments, continues to be an area of profound interest and active research. This chapter delves deeply into the potential of Generative Adversarial Networks (GANs) as a solution to this challenge. Starting with an introduction to the foundational mechanics of GANs, we explore the dynamic interplay between the generator and the discriminator components.

Through the evolution of GANs, several refined architectures have surfaced, such as DCGAN, WGAN, and LSGAN. Each of these variants offers specific advantages tailored to address unique challenges in speech enhancement. But our exploration isn't confined to traditional GAN architectures. We introduce the Gated Control Neural Network (GCN) into our discourse—a model that introduces an added layer of complexity and control to the enhancement process.

The heart of this chapter is our rigorous experimental process. Here, we evaluate the performance of a variety of GAN models, both standalone and in combination with the GCN. Through these evaluations, we aim to discern the most effective strategies and configurations for speech enhancement. Our results, both in terms of numbers and qualitative observations, accentuate the promise and efficacy of our chosen methodologies in the broader context of audio signal enhancement.

3.2.1 Generator

In Chapter 2, we introduced the use of Generative Adversarial Networks (GANs) in speech enhancement, and SEGAN [13] has made significant progress with the advancement of Convolutional Neural Networks (CNNs) [129]. GANs have become a new paradigm in speech enhancement, with the generator aiming to produce convincing speech by sanitizing the content of the original noisy speech, and the discriminator minimizing the difference between clean and enhanced speech signals [130][13][65].

However, some GAN-based approaches, such as SEGAN [13], used architectures originally designed for computer vision tasks without any modifications for speech features [131]. These approaches may not fully capture the global context information of each sample, which is crucial for effective noise elimination. In this chapter, we address this issue by incorporating more global context information into the generator using a Gated Control Neural Network (GCN). The GCN was selected because of its ability to capture long-range dependencies and model interactions among features. Our goal is to improve speech enhancement by exploiting the global utterance-level context, which provides a more comprehensive description of overall speech interpretation, while also leveraging local patch-level features, which are more sensitive to noise interference.

Introduction for Gated Control neural network (GCN)

In this chapter, motivated by GRN [8], we adopt the Convolutional Encoder-Decoder (CED) network as the baseline network, which is illustrated in Figure 3.1. The input feature of the network is the magnitude spectrum of the noisy speech, which is characterized by two dimensions of time (frame) and frequency. The network outputs the enhanced speech magnitude spectrum, and by combining it with the phase of the noisy speech, we can reconstruct the enhanced speech waveform. In general, the network can be roughly divided into three layers, namely the encoding layer, the middle layer and the decoding layer. The structure and function of each layer are as follows:

Encoder: As for the encoder layer, it comprises of 5 two-dimensional convolutional layers, as depicted in Figure 3.1. Each layer is made up of a 2D-Convolution operation, batch normalization (BN) layer, and exponential linear unit (ELU) activation function. Specifically, the feature map undergoes 2D convolution, followed by batch normalization, and finally activation via the ELU function. The BN layer performs operations such as data normalization to satisfy the assumption of independent and identical distribution, which facilitates faster convergence and prevents gradient explosion, as confirmed by research [132]. Moreover, ELUs are applied to all convolutional and deconvolutional layers, except for the output layer, as they have been demonstrated to lead to faster convergence and better generalization than ReLUs [128]. By employing 5 layers of 2D-Convolution, the encoder layer extracts the magnitude spectrum features of speech in a hierarchical manner. After each convolution operation, the feature map remains unchanged in the time dimension, halved in the frequency dimension, and doubled in the number of channels. Importantly, the feature map outputted by the encoder layer is consistent with the input feature in the time dimension, enabling the model to handle speech of arbitrary length (frame length) with better real-time processing capabilities.

Middle layer: In this section, we will discuss the middle layer of the network architecture used in the study. The middle layer contains two one-dimensional convolutional layers, each consisting of a one-dimensional convolution operation, a BN layer, and an activation function called LeakyRelu (LRelu). The convolution kernel used in the one-dimensional convolution has a stride of 1 for all dimensions. Before the convolution operation, the two-dimensional output generated by the encoder is reshaped to reduce the dimensionality so that it meets the input requirements for the one-dimensional convolution. Similarly, the output of the middle layer needs to be adjusted to restore its dimension.

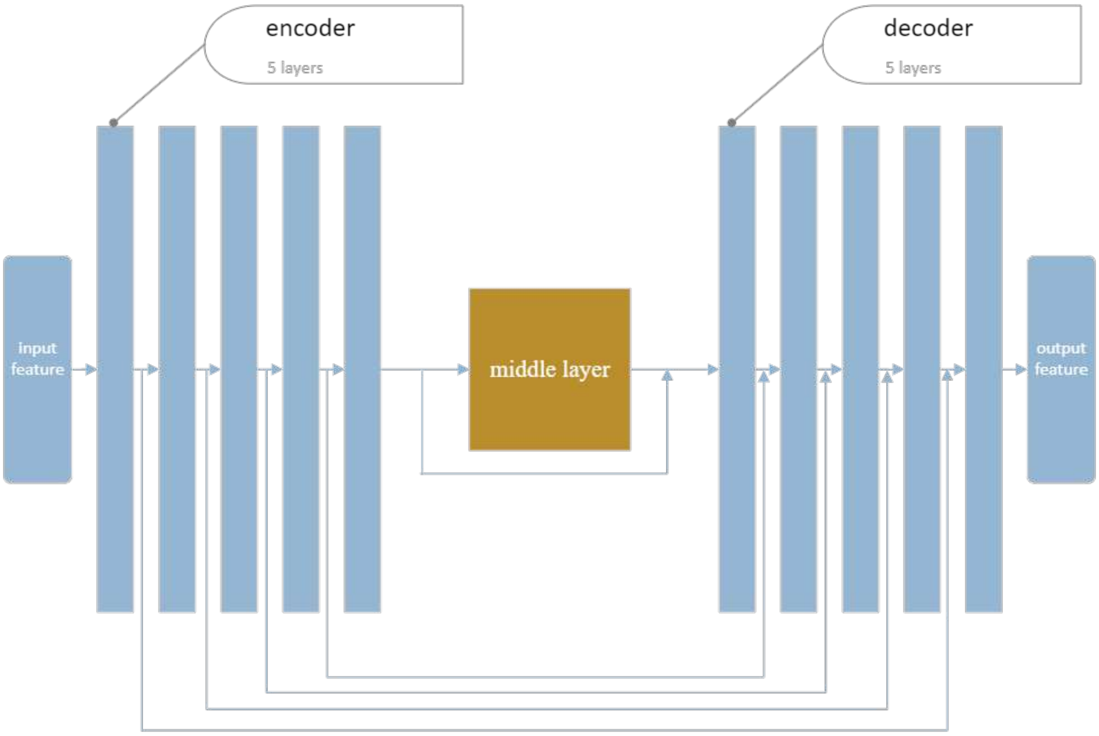


Figure 3.1 Speech enhancement flow chart based on CED network.

The main role of the middle layer is featuring transfer. In previous studies, such as[114], the middle layer was implemented using LSTM (long short-term memory) units. LSTM [103] is a specific type of recurrent neural network (RNN) that incorporates a memory cell and has been shown to be successful in modeling temporal dependencies in various applications such as acoustic modeling and video classification. However, in this study, the one-dimensional convolutional layers with BN and Relu activation functions were found to be effective in transferring features between the encoder and decoder layers.

Decoder layer: The decoder layer comprises of five two-dimensional deconvolution layers. Each layer includes two-dimensional deconvolution (2D-Deconv), a batch normalization (BN) layer, and an exponential linear units (ELU) activation function. The 2D-Deconv layer can be viewed as the inverse process of 2D-Conv, and it restores the position information of the feature map by adjusting the convolution step size. Moreover, we feed the feature map of the encoding layer to the decoding layer of the same dimension and double the number of channels of the feature map of the corresponding decoding layer through channel splicing. This operation helps to restore fine-grained feature information during the decoding process.

Gate control unit

The primary role of the one-dimensional convolutional layer in the middle is to transfer features. This means that the encoder layers' output is passed as input to the decoder layers, resulting in limited processing capability of sequence information. Therefore, we employ a gating mechanism to handle the one-dimensional information flow. Previous studies used RNN-based gating mechanisms to process sequence information, such as LSTM and Gate Recurrent Unit. However, these models faced the challenges of poor parallelism and a significant amount of computation.

To address the issues mentioned above, we utilize a fully convolutional gate control unit (GCU) [133]. As shown in Figure 3.2.2, the GCU gating unit significantly reduces network training parameters, possesses excellent parallelism, and effectively transmits information. It is crucial to note that the GCU's middle layer has two activation functions, namely linear activation and Sigmoid. The linear activation function provides a linear path for gradient backpropagation to mitigate the problem of gradient vanishing. Sigmoid is utilized to maintain the nonlinear characteristics of the network, with a value range of 0 to 1. Through this activation function, we can focus on the necessary speech features and ignore irrelevant ones. For instance, if the input feature is inp, the activation value multiplies with inp, resulting in the output. Since the activation function value ranges from 0 to 1, the input selectively remains as the output (expressed by equation 3.1).

$$out = sigmoid(inp) \tag{3.1}$$

In summary, the one-dimensional convolutional layer in the middle of the model is mainly responsible for feature transfer, leading to limited processing capability of sequence information. To overcome this limitation, we adopt a gating mechanism, the GCU. Compared to previous RNN-based gating mechanisms (shown in Figure 3.2), the GCU significantly reduces network training parameters, possesses excellent parallelism, and effectively transmits information. The GCU's middle layer contains two activation functions, linear activation, and Sigmoid, which work together to maintain the nonlinear characteristics of the network and address the problem of gradient vanishing. Overall, the GCU can selectively retain the essential speech features and ignore irrelevant ones.

$$S_{L+1} = \sigma(W_L * S_L + B_L) \odot (V_L * S_L + H_L) \tag{3.2}$$

$$sigmoid(\sigma) = \frac{1}{1 + e^x} \tag{3.3}$$

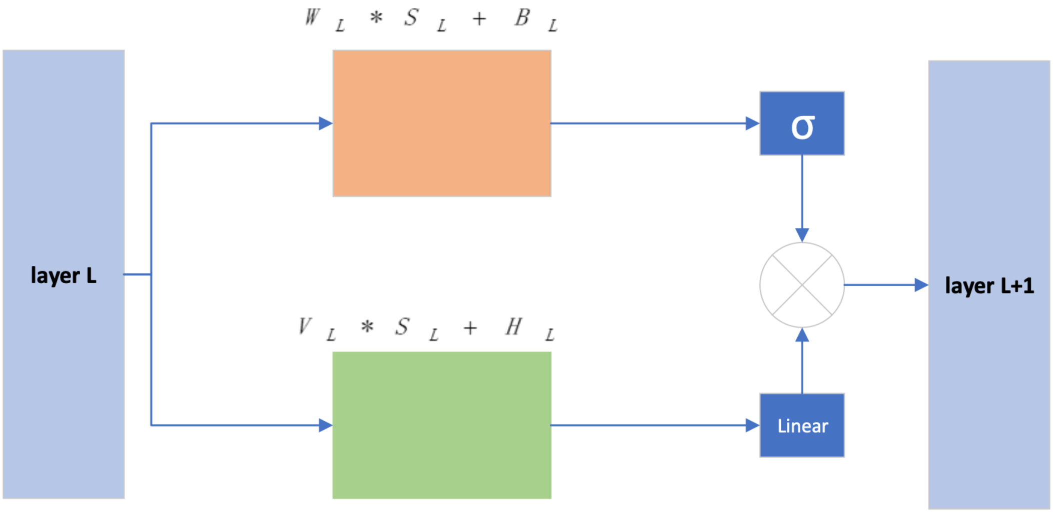


Figure 3.2 Gate linear unit.

Where S_{L+1} and S_L represent the output feature of $L+1_{th}$ and L_{th} layer, respectively. W_L, V_L, B_L, H_L represent weights and bias of each layer. σ, * are sigmoid activation function and convolution function. And then get S_{L+1} by element-wise dot product.

In the middle layer of the model, a one-dimensional dilated convolution is applied to process the speech input. One-dimensional dilated convolution offers advantages over the one-dimensional ordinary convolution due to its ability to capture a larger receptive field [134], as illustrated in Figure 3.3. The receptive field typically grows exponentially, which allows for the convolution process to extract more contextual information from the speech input. As a result, the model can better capture the dependencies between different elements in the input sequence, enabling it to mine more informative speech features. This feature extraction mechanism is crucial for the model's ability to accurately capture speech patterns and make predictions.

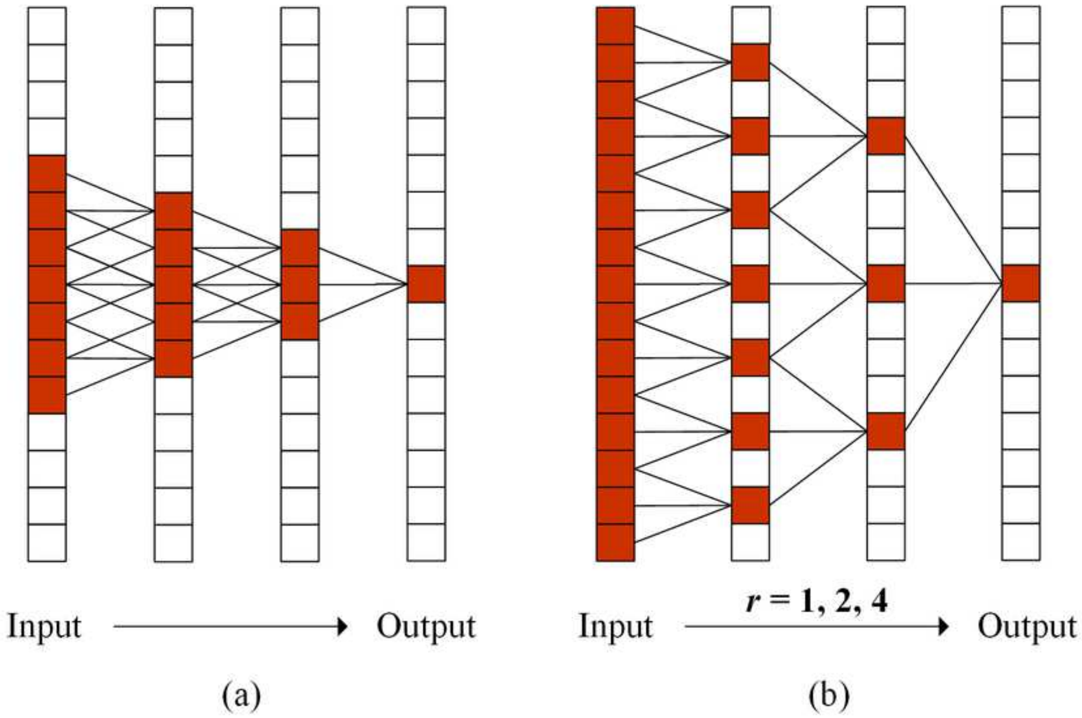


Figure 3.3 (a) traditional 1-D convolution; (b)1-D dilated convolution.

Gate control residual block

The use of deep neural networks has brought significant improvements to various applications, including speech enhancement. However, deep neural networks often suffer from overfitting and gradient vanishing or explosion, leading to poor performance. To address these issues, researchers have proposed using residual networks, which can improve the accuracy of the model and prevent overfitting, as well as gated control units (GCUs), which can better capture dependencies between features.

To further enhance the performance of speech enhancement models, we combine these techniques and introduce a dimensionally dilated convolution to create a gated residual unit (GRU). The structure of the GRU is shown in Figure 3.4. Different from the gated unit in GRN [8], shown in Figure 3.5, our proposed gated residual module contains a parallel path: Residual output as well as skip connection, which can extract more speech granularity. The gated residual module contains four convolutional layers, with the upper half replacing the two one-dimensional convolutions in the GCU with one-dimensional dilated convolutions. The convolution kernel size, stride, and number of output channels are set to 5, 1, and 128, respectively. The lower part of the module consists of two parallel one-dimensional ordinary convolutional layers with a convolution kernel size, stride, and number of output channels of 1, 1, and 128, respectively. The GRU produces both a residual output and a skip connection output, allowing for the integration of the extracted features from corresponding layers into the final prediction.

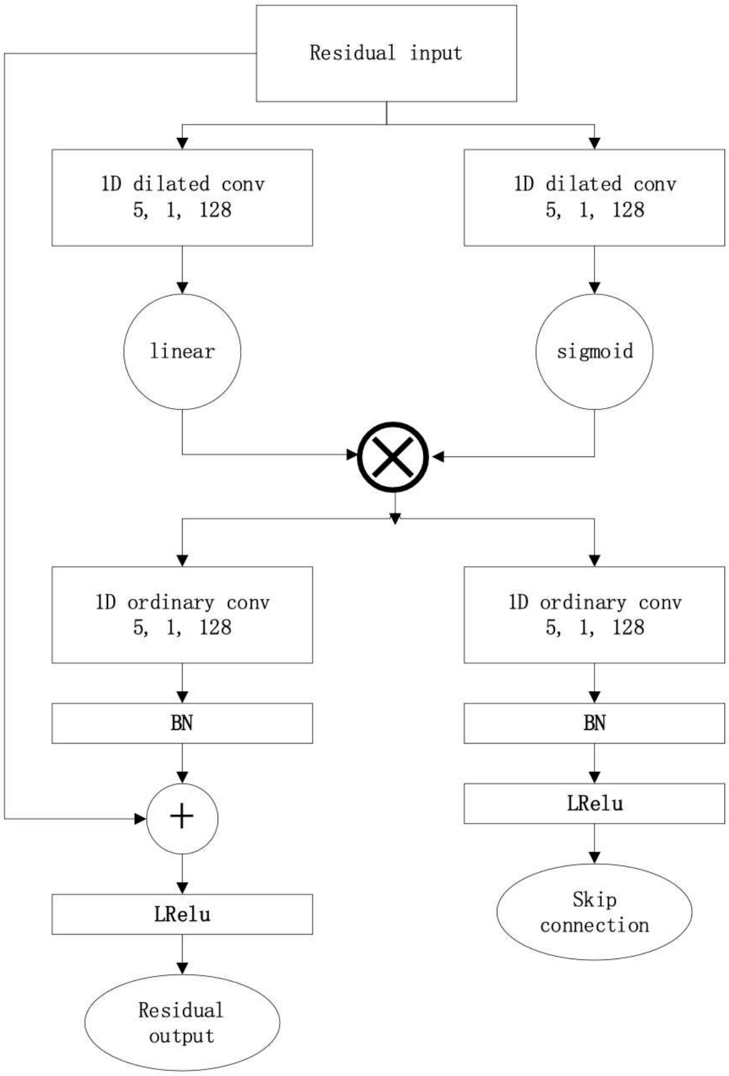


Figure 3.4 The proposed gate control residual block.

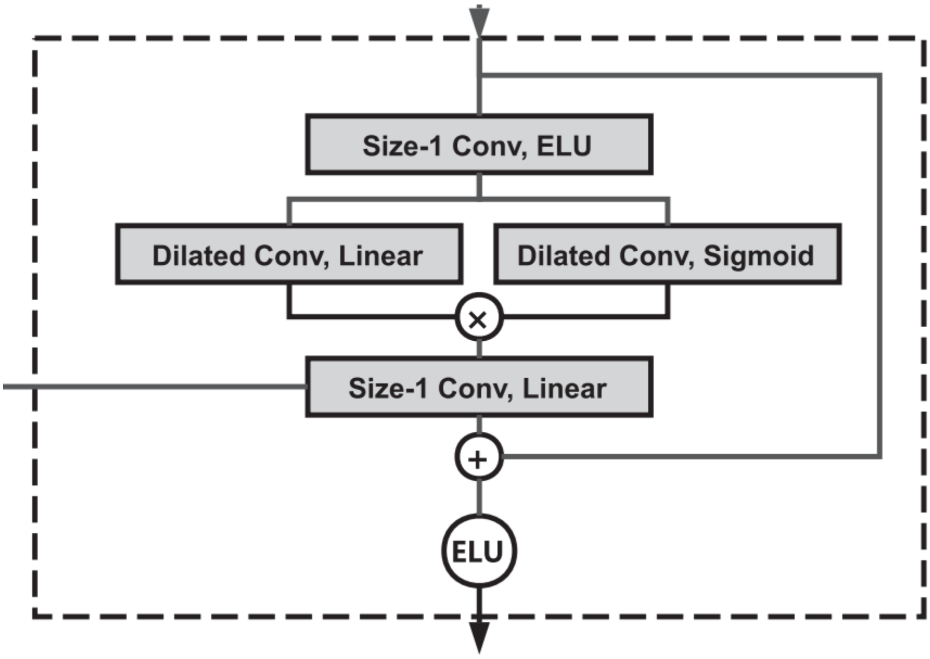


Figure 3.5 The middle gate unit in GRN.

Since the number of channels of the input and output of the module is the same. Suppose s is the input of the network, and the output $F(s)$ is the output through multiple hidden layers, then the residual output is $s + F(s)$. After the two one-dimensional ordinary convolution layers, the BN layer and the activation function LeakyRelu (LRelu) are also added to ensure that the output features of the module still satisfy the independent and identical distribution assumption and maintain the nonlinear characteristics of the network. It is worth noting that the two parallel dilated convolutional layers in the module use the same dilation rate. By stacking gated residual modules and gradually increasing the dilation rate, the purpose of expanding the receptive field can be achieved.

Based on the baseline convolutional encoder and decoder network, we introduce the gated residual module constructed above into the middle layer of the network and obtain a gated residual convolutional encoder and decoder network, whose network structure is shown in Figure 3.6. In the middle layer of the network, a gated residual network is formed by stacking a set of gated residual modules with an expansion rate r up to a certain maximum coefficient of 2^n. This structure can significantly improve the receptive field with a small number of parameters, and at the same time improve the processing ability of the model to sequence information and pay more attention to the temporal features. In addition, skip connections are used in the gated residual network, which allows the network to incorporate (Add) features extracted from the corresponding layers into the final prediction. It is worth noting that, inspired by [135], we implement ISTFT through convolutional layers, so that time-domain enhanced speech can participate in network training. Moreover, the phase of the original time-domain speech can compensate for the noisy speech phase and reconstruct a more accurate time-domain enhanced speech.

In Table 3.1, we give the specific parameter settings of the network model. On the whole, the network input and output feature dimensions are $T * 256$, where T represents the time frame, 256 represents the frequency dimension, and the third dimension represents the number of feature map channels. Among the hyperparameters, k is the size of the convolution kernel, s is the convolution stride, c is the number of output channels, and R is the dilation rate. Since the input and output dimensions of the gated residual module remain unchanged, it has good portability, and the number of modules can be added according to the model requirements. In this experiment, the gated residual network in this paper is composed of 20 gated residual modules, which are obtained by superimposing 4 groups of gated residual modules with expansion rates R of 1, 2, 4, 8, and 16, respectively.

To obtain the optimal network model, an appropriate loss function is selected for training. The enhanced speech magnitude spectrum is then obtained through network mapping. By taking advantage of the human ear's insensitivity to phase information, the time-domain enhanced speech is obtained through phase reconstruction of the noisy speech.

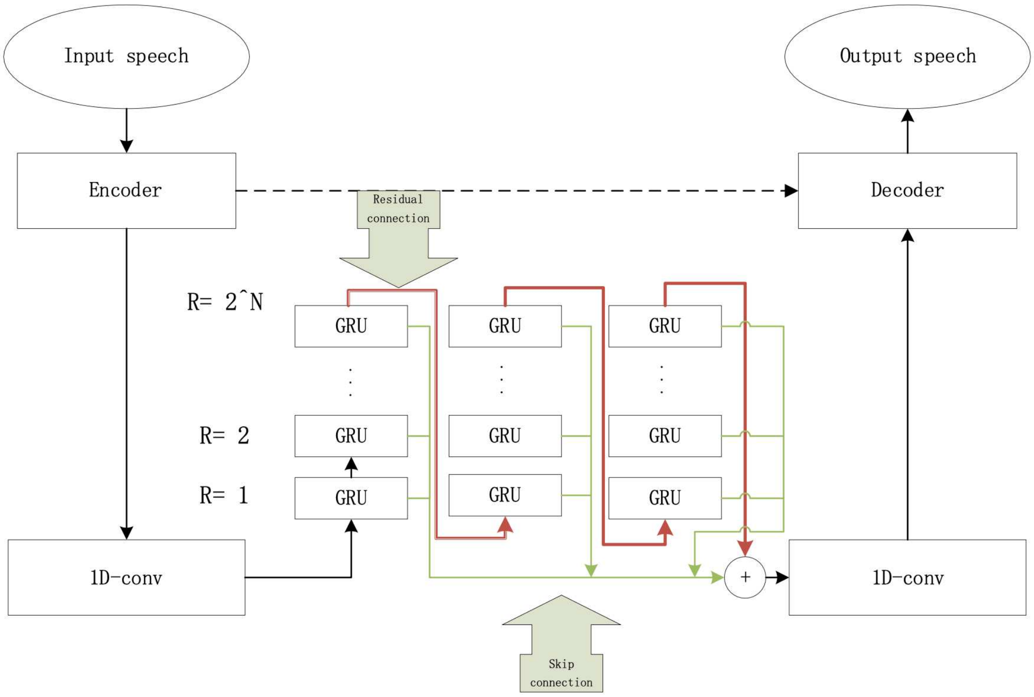


Figure 3.6 Speech enhancement flow chart based on gated control network.

Table 3.1 Network parameters.

Layer name	Input size	Output size	Hyperparameters
Reshape1	T*256	T*256*1	
Encoder layer1	T*256*1	T*128*4	k = 1*3; s = 1*2; c = 4
Encoder layer2	T*128*4	T*64*8	k = 1*3; s = 1*2; c = 8
Encoder layer3	T*64*8	T*32*16	k = 1*3; s = 1*2; c = 16
Encoder layer4	T*32*16	T*16*32	k = 1*3; s = 1*2; c =32
Encoder layer5	T*16*32	T*8*64	k = 1*3; s = 1*2; c = 64
Reshape2	T*8*64	T*512	
1D_conv1	T*512	T*256	k = 1; s = 1; c = 256

Gate control residual block (Add)	T*256	T*256	R = 1, 2, 4, 8, 16;
			k_residual = 5; k_skip = 1; s = 1;
			c = 256

1D_conv2	T*256	T*512	k = 1; s = 1; c = 512
Reshape3	T*512	T*8*64	
Decoder layer5 (concat×2)	T*8*128	T*16*32	k = 1*3; s = 1*2; c = 32
Decoder layer4 (concat×2)	T*16*64	T*32*16	k = 1*3; s = 1*2; c = 16
Decoder layer3 (concat×2)	T*32*32	T*64*8	k = 1*3; s = 1*2; c = 8
Decoder layer2 (concat×2)	T*64*16	T*128*4	k = 1*3; s = 1*2; c = 4
Decoder layer1 (concat×2)	T*128*8	T*256*1	k = 1*3; s = 1*2; c = 1
Reshape4	T*256*1	T*256	

Long-term memory unit applied into the GCN

This paper also proposes a novel topology called Gated Control U-Network (GCU-N) to model compressed complex features, which is inspired by the U-Net architecture [136]. GCU-N, as shown in Figure 3.7, consists of two four-layer U-net units, serving as an encoder and a decoder, respectively. This design allows the network to capture dynamic long-term context information, enabling it to extract both global and local information more effectively. The middle layer of GCU-N employs the same GCN structure mentioned earlier.

GCU-N's multi-scale feature representation from each layer makes it possible to have a deep structure while still being computationally efficient and memory-cost-effective. Each layer in the U-net units includes 2D-convolution, batch-normalization, the activation function (ELU), and 2D-deconvolution. The first layer of Conv2D and the last layer of DeConv2D have a kernel size of (2, 5), while the rest of the layers use a kernel size of (2, 3). The number of channels is set to 64, and the stride is (1, 2) in all layers.

To train the GCU-N model, an appropriate loss function is selected to obtain the optimal model. Once the model is trained, the enhanced speech amplitude spectrum is obtained through network mapping. As human ears are insensitive to phase information, the time-domain enhanced speech is obtained by using phase reconstruction on the noisy speech.

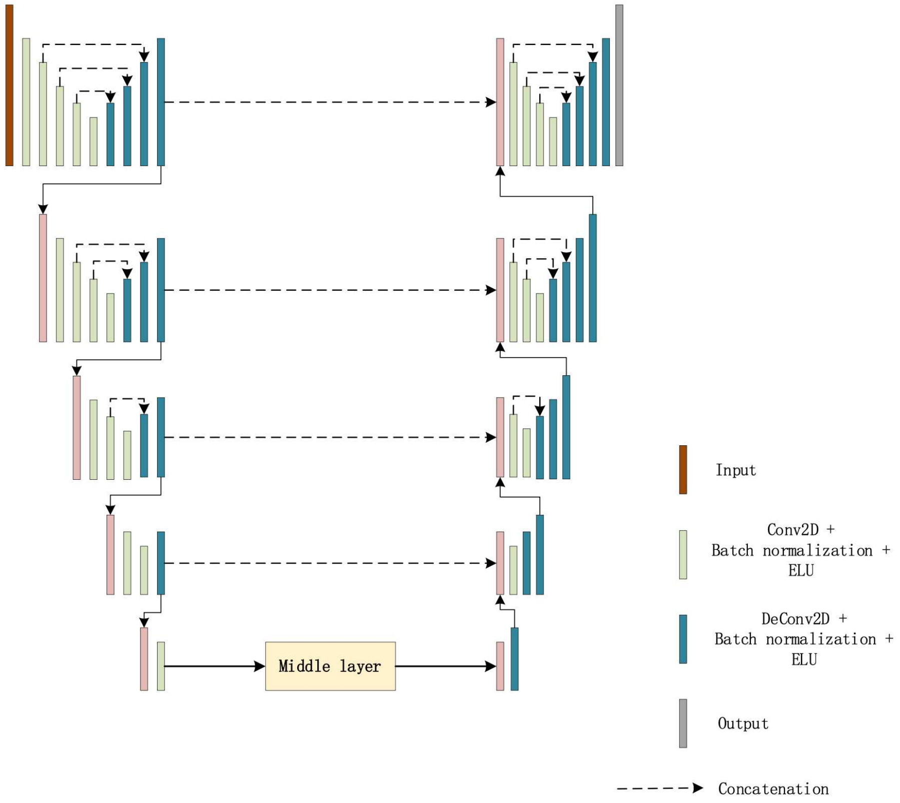


Figure 3.7 Structure diagram of the proposed GCU-N. The main architecture of this model like Encoder-Decoder, where each layer consists of residual units.

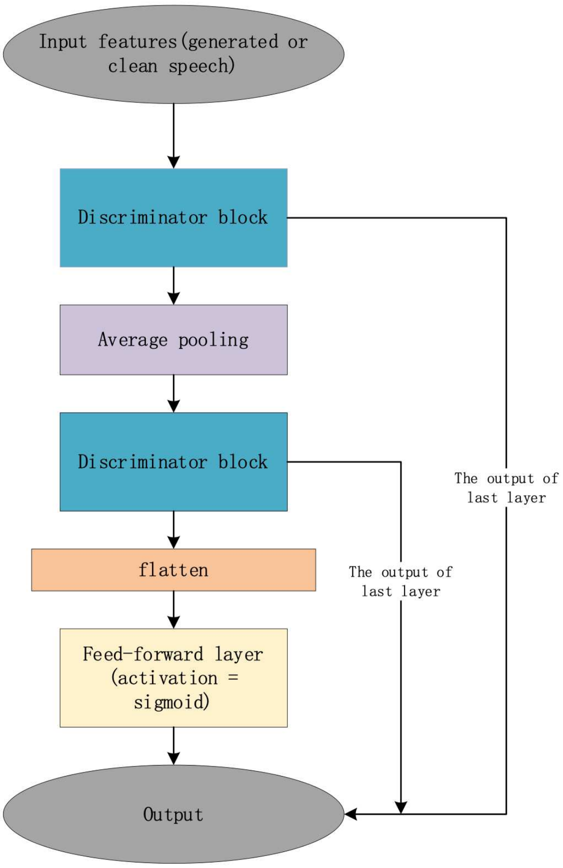
3.2.2 Discriminator

The discriminator is a crucial component of the proposed model, and its structure is composed of two discriminator blocks, as illustrated in Figure 3.8. Each discriminator block comprises of five layers that utilize 1-dimensional convolution (Conv1D), batch normalization to prevent gradient vanishing or exploding, and the activation function LeakRelu (LRelu), as shown in Figure 3.9. The kernel size, number of channels, and the strides of each layer are specified in Table 3.2.

The discriminator's primary purpose is to distinguish between real and fake speech signals, thereby providing feedback to the generator network. Specifically, the discriminator generates an output by computing the final layer output of each discriminator block and the feed-forward layer output. By comparing the output with the ground truth, the generator network is updated to produce more realistic speech signals.

Table 3.2 The hyperparameter setup of each discriminator block

Layer name	hyperparameters
Layer1	kernel_size = 8, strides = 1, channels = 16
Layer2	kernel_size = 8, strides = 2, channels = 64
Layer3	kernel_size = 8, strides = 2, channels = 256
Layer4	kernel_size = 4, strides = 2, channels = 64
Layer5	kernel_size = 2, strides = 2, channels = 1



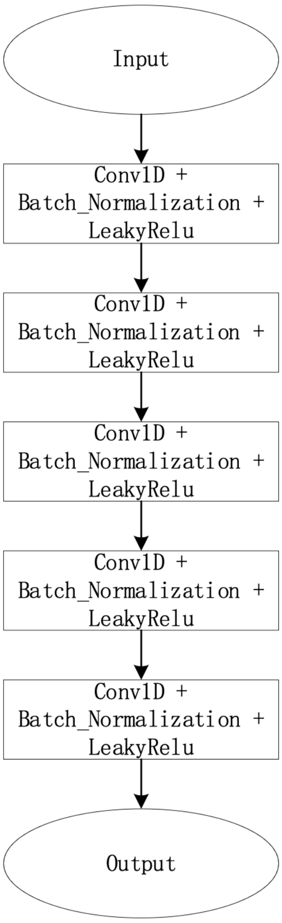

Input
Conv1D + Batch_Normalization + LeakyRelu
Conv1D + Batch_Normalization + LeakyRelu
Conv1D + Batch_Normalization + LeakyRelu
Conv1D + Batch_Normalization + LeakyRelu
Conv1D + Batch_Normalization + LeakyRelu
Output

Figure 3.8 The diagram of the discriminator block. Figure 3.9 The discriminator.

3.2.3 Loss function

To optimize the network model, the mini-batch gradient descent method is utilized in this study. The choice of loss function is a critical aspect of training the network to obtain the optimal results. In this work, the time domain loss function is employed as it can better emphasize the time-frequency related features of speech. The time domain loss is calculated at the output of the decoding layer, and the mean absolute error function (MAE) is used as it has been shown to improve performance in previous studies [112]. The MAE-based expression for the time domain loss function is presented below:

$$L_{MAE} = \frac{1}{M}\sum_{n=1}^{M}\left\|\widehat{S_n} - S_n\right\| \tag{3.7}$$

Where $\|.\|$ denotes the absolute value of vector, S_n and $\widehat{S_n}$ represent the amplitude spectrum vectors of the original speech and the enhanced speech of the n_{th} frame, respectively. M is the batch size in the training period. Since the calculation formula of the speech objective evaluation index is different from the network training loss function, there may be a problem of mismatch between the loss function and the evaluation index, that is, when the loss function drops to a certain level, some evaluation indexes may not continue to change [113]. To address the issues mentioned above, this study proposes the use of a speech evaluation index as the network loss function, which can further enhance the performance of speech enhancement. Specifically, based on the findings in literature [137], the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) is chosen as the training function of the network. This objective index has been shown to significantly improve the quality of enhanced speech. The SI-SDR is a measure of the ratio between the energy of the target speech signal and the energy of the distortion, which includes noise and artifacts. It is scale-invariant, which means that it is insensitive to changes in the amplitude of the signals. By optimizing the network to maximize the SI-SDR, the model can effectively reduce the distortion in the output speech signal, leading to improved speech quality and intelligibility., and the formula for calculating SDR is:

$$SI - SDR = 10\log_{10}\left(\frac{\|aS_n\|^2}{\left\|\alpha S_n - \widehat{S_n}\right\|^2}\right) \tag{3.8}$$

S_n and $\widehat{S_n}$ represent the original speech and the enhanced speech of the n_{th} frame, respectively. α is the weighting factor of pure speech, the calculation formula is:

$$\alpha = \arg min_{\alpha} \left\|\alpha S_n - \widehat{S_n}\right\|^2 = \frac{\widehat{S_n}^T S_n}{\|S_n\|_2^2} \tag{3.9}$$

So, the optimized SI-SDR function is:

$$L_{SI-SDR} = -SI - SDR = -\frac{1}{M}\sum_{n=1}^{M} 10\log_{10}\left(\frac{{S_n}^T\widehat{S_n}}{{S_n}^T\widehat{S_n}\widehat{S_n}^T S_n - {S_n}^T\widehat{S_n}}\right) \tag{3.10}$$

L_{SI-SDR} is calculated using frequency domain signals. At the same time, we jointly optimize MAE and SI-SDR, and the final network optimization function is (Joint):

$$L_{joint} = L_{MAE} + L_{SI-SDR} \tag{3.11}$$

3.3 The network gradient is updated by minimizing the optimization function and the error is passed to each layer of the Data processing

In real-world scenarios, sound signals are often contaminated by various types of noises during the transmission process. As a result, the signals captured by recording equipment often contain mixed signals from multiple sources, including ambient noise, other speakers' sounds, echoes, and reverberations. To simulate such scenarios, speech and noise samples are first selected from a corpus and a noise database, respectively, and are mixed at different signal-to-noise ratios

(SNRs) to create a pre-mixed signal dataset. Before mixing, a fixed sampling rate is used to resample the signals, ensuring that the sampling frequency of speech and noise remains consistent. Typically, the sample rate is chosen as 8 kHz or 16 kHz. The 8 kHz sample rate is suitable for telephone and encrypted walkie-talkie, wireless intercom, and wireless microphone transmission, providing adequate quality for human speech without sibilance. The 16 kHz sample rate, on the other hand, provides wideband frequency extension over standard telephone narrowband 8,000 Hz and is commonly used in modern VoIP and VVoIP communication products **错误!未找到引用源。**. Figure 3.10 displays time-domain waveform diagrams of pure speech, noise, and mixed signals obtained when noise and speech are superimposed at a signal-to-noise ratio of 0 dB.

Speech signals are usually collected as digitized time series, which exhibit time-varying fluctuations. Speech features and related parameters also vary over time due to this time-varying characteristic. However, the generation mechanism of speech involves human oral movements that cause changes in the shape of the channel. The frequency of these movements is much lower compared to the vibration frequency of speech, resulting in the speech signal being almost stationary and unchanged over short periods. Therefore, analysis and processing of speech signals can be carried out by treating them as a combination of multiple short-term stationary signals.

The pre-processing of speech signals involves three main steps, namely pre-emphasis, framing, and windowing. Pre-emphasis aims to increase the high-frequency spectrum in the speech signal, which corresponds to smaller components, so that it becomes relatively flat. This facilitates the use of the same bandwidth from low to high frequency for spectrum analysis and acoustic feature extraction. After pre-emphasis, the voice signal is windowed and framed with a frame length of 20-40ms [139]. In this study, a frame length of 32ms is chosen.

Besides that, we also need to compress input speech signals, like complex-value domain or magnitude.

$$Y = |Y_o|^c e^{jY_p} = Y_m e^{Y_p} = Y_r + jY_i \tag{3.12}$$

Where $Y_m, Y_p, Y_r\ and\ Y_i$ denote the magnitude, phase, real and imaginary components of the compressed spectrogram, respectively. C is the compression exponent ranging from 0 to q, here we follow [142] to set $c = 0.3$. The power-law compression of the magnitude equalizes the importance of quieter sounds relative to loud ones, which is closer to human perception of sound [143][144].

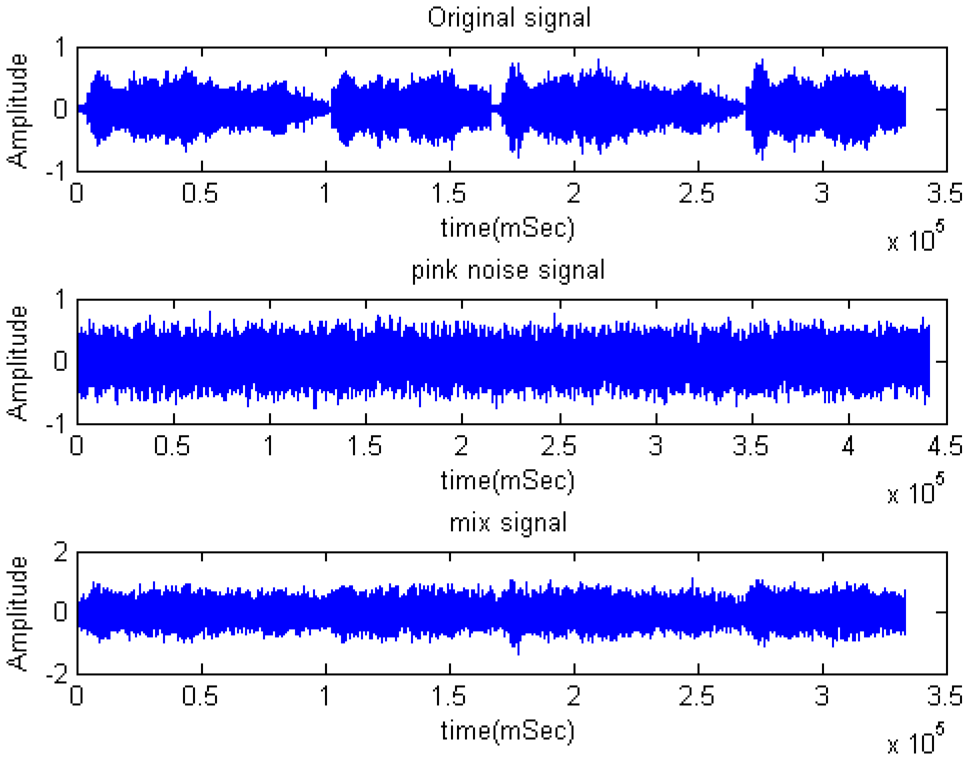


Figure 3.10 Waveforms of original signal, pink noise signal and the mix signal.

The overlap between two adjacent frames is called frame shift. In general, the ratio between frame shift and frame length is set about 0.5 [140][141][9]. Framing is implemented by a method of moving weighting with a window function of limited length (shows in Figure 3.11), which can be expressed as a convolution symbol as follows:

$$S_W(n) = S(n) * w(n) \tag{3.13}$$

In the above formula, $S(n)$ represents the original speech signal, $w(n)$ is the window function, and $Sw(n)$ is the windowed speech signal. Common window functions of speech signal processing are mainly rectangular windows and Hanning windows, where the expression of the rectangular window is as follows (where N is the frame length):

$$w(n) = \begin{cases} 1, & 0 \leq \mathrm{n} \leq (\mathrm{N}-1) \\ 0, & else \end{cases} \tag{3.14}$$

The expression of the Hanning window is as follows:

$$w(n) = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), & 0 \leq \mathrm{n} \leq (\mathrm{N}-1) \\ 0, & else \end{cases} \tag{3.15}$$

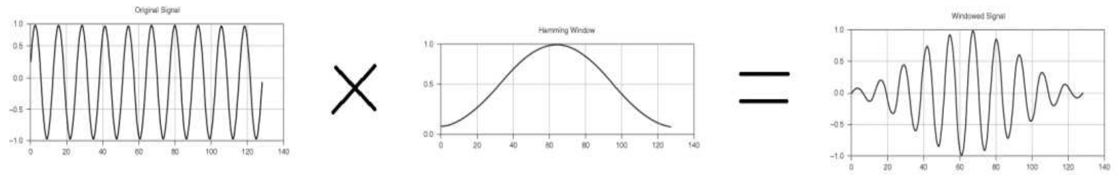


Figure 3.11 The diagram of window function on signal.

3.4 Experiment implement

The LibriSpeech dataset [147], which is a corpus of more than 100 hours of read speech, based on LibriVox's public domain audio books, is used to evaluate all models. For training, three types of noise from NoiseX-92 [145], i.e., 'pink,' 'volvo,' and 'babble,' are utilized to generate 100,000 noisy utterances at different SNRs uniformly sampled from {-2dB, 0 dB, 2 dB, 4 dB, 6 dB}. For testing, 1000 utterances for 'pink,' 'babble,' and 'volvo' are generated at SNRs of -2dB, 0 dB, 2 dB, 4 dB, 6 dB for both types of noise. For training dataset, we use train-clean-100 as pure speech. While the dev-clean was used for testing dataset.

In order to ensure consistency and comparability, all utterances in this experiment were resampled at 16 kHz. To segment the utterances, a sliding window of 32 milliseconds (512 samples) with a 16 milliseconds (256 samples) overlapping was used. The input feature in this experiment is magnitude formatted as (Batch × Time_steps × Feature_maps) to facilitate analysis. The initial learning rate was set to 0.001, and was adjusted as the epoch increased, for instance, it was set to 0.0001 after the 5th epoch.

To train the models, the Adam optimizer [146] was utilized with a batch size of 8, and the network was trained over 5 epochs on 2-second-long segments. Tensorflow2 was employed to develop all models, and its default settings for initialization were adopted. To carry out the training process, one NVIDIA RTX 2060 super 6GB GPU was used, and the training process took approximately 6 hours.

In this experiment, three different models were trained, namely GAN, GCN and GCU-N. Afterwards, comparisons were made among the three models, and relevant analyses were conducted.

3.4.1 Results

In this study, two widely used metrics, short-time objective intelligibility (STOI) [109] and perceptual evaluation of speech quality (PESQ) [87], are employed to evaluate the performance of the proposed speech enhancement models. STOI measures the similarity between the enhanced speech and the clean speech in terms of their short-time segments, with a typical range between 0 and 1, where a higher score indicates better speech quality. PESQ, on the other hand, is designed to evaluate the overall perceptual quality of the enhanced speech, which varies between -0.5 and 4.5, with a higher score indicating better quality.

To assess the performance of the proposed models, we present the PESQ and STOI evaluation results of the GAN, GCN and GCU-N models under matched noise conditions. As shown in the figures below, we evaluate the models at each signal-to-noise ratio (SNR). The results demonstrate that the proposed models can effectively enhance the speech quality, as indicated by the increased STOI and PESQ scores.

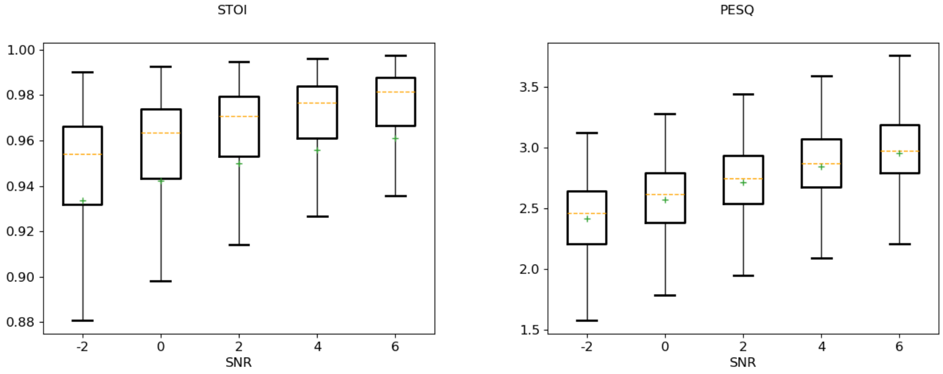


Figure 3.12 The experimental results of proposed GAN under ‘pink’ noise conditions for STOI and PESQ.

The performance of GAN, GCN, and GCU-N models were evaluated under the ‘pink’ noise at SNRs of -2dB, 0 dB, 2 dB, 4 dB, and 6 dB. The median STOI values for GAN at these SNRs were 0.95, 0.96, 0.97, 0.98, and 0.98, respectively (as indicated by the yellow line in each box of the box plots).

Similarly, the median PESQ values for GAN under the ‘pink’ noise at the same SNRs were 2.46, 2.61, 2.75, 2.87, and 2.97, respectively (as indicated by the yellow line in each box of the box plots).

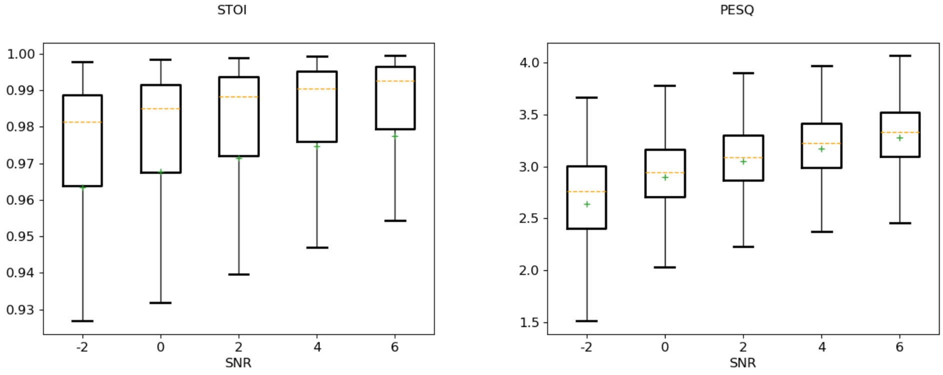


Figure 3.13 The experimental results of proposed GAN under ‘volvo’ noise conditions for STOI and PESQ.

The GAN model was evaluated under the ‘volvo’ noise at various SNRs (-2dB, 0 dB, 2 dB, 4 dB, 6 dB) using the short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) as the final evaluation metrics. The median STOI values (represented by the yellow line in each box) for GAN were found to be 0.98, 0.99, 0.99, 0.99, and 0.99 respectively for SNRs of -2dB, 0 dB, 2 dB, 4 dB, and 6 dB.

Similarly, the median PESQ values (represented by the yellow line in each box) of GAN were found to be 2.61, 2.76, 2.88, 2.99, and 3.09 respectively for SNRs of -2dB, 0 dB, 2 dB, 4 dB, and 6 dB under the ‘volvo’ noise. These results demonstrate the effectiveness of the GAN model in improving the quality of speech under different levels of noise.

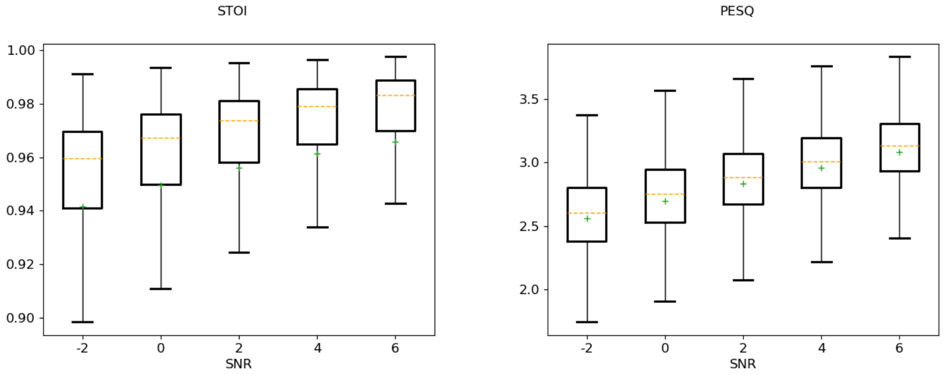


Figure 3.14 The experimental results of proposed GCN under ‘pink’ noise conditions for STOI and PESQ.

The median STOI values (the yellow line in each box) of GCN under the ‘pink’ noise at SNRs of -2dB, 0 dB, 2 dB, 4 dB, 6 dB are 0.96, 0.97, 0.97, 0.98, 0.98 respectively. The median PESQ values (the yellow line in each box) of GCN under the ‘pink’ noise at SNRs of -2dB, 0 dB, 2 dB, 4 dB, 6 dB are 2.61, 2.75, 2.88, 3.00, 3.13 respectively.

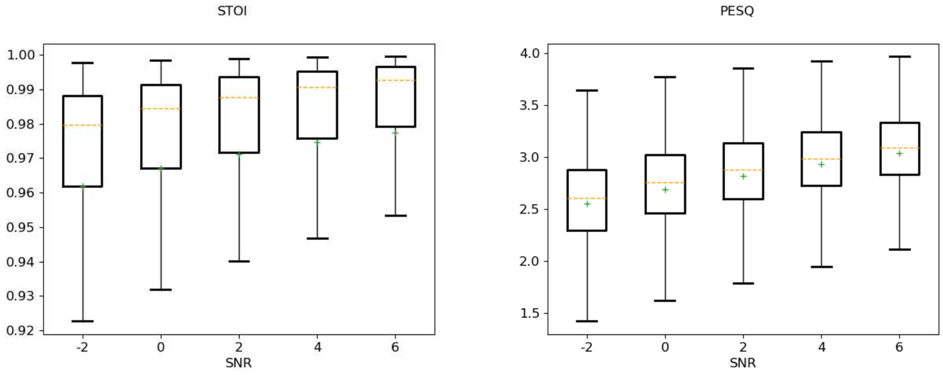


Figure 3.15 The experimental results of proposed GCN under ‘volvo’ noise conditions for STOI and PESQ.

The median STOI values (the yellow line in each box) of GCN under the ‘volvo’ noise at SNRs of -2dB, 0 dB, 2 dB, 4 dB, 6 dB are 0.98, 0.98, 0.99, 0.99, 0.99 respectively. The median PESQ values (the yellow line in each box) of GCN under the 'volvo' noise at SNRs of -2dB, 0 dB, 2 dB, 4 dB, 6 dB are 2.76, 2.94, 3.09, 3.22, 3.33 respectively.

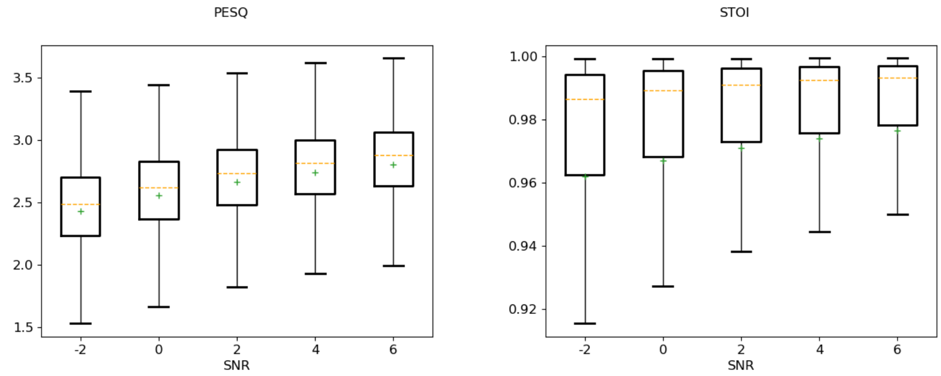


Figure 3.16 The experimental results of proposed GCN under 'babble' noise conditions for PESQ and STOI.

The median PESQ values (the yellow line in each box) of GCN under the 'babble' noise at SNRs of -2dB, 0 dB, 2 dB, 4 dB, 6 dB are 2.49, 2.61, 2.73, 2.82, 2.88 respectively. The median STOI values (the yellow line in each box) of GCN under the 'babble' noise at SNRs of -2dB, 0 dB, 2 dB, 4 dB, 6 dB are 0.99, 0.99, 0.99, 0.99, 0.99 respectively.

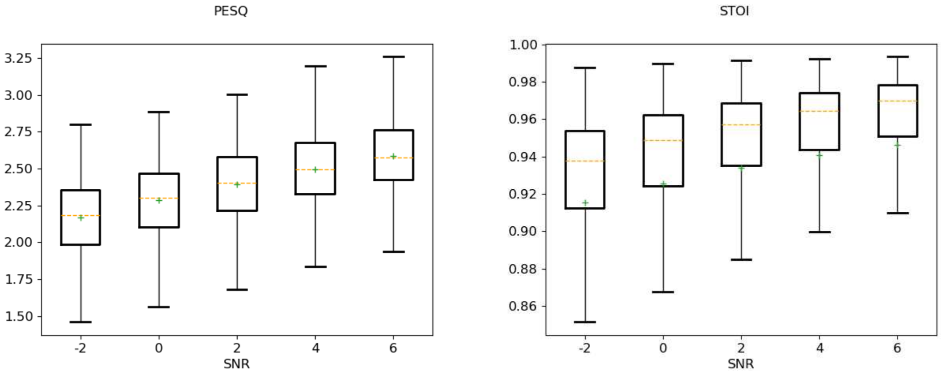


Figure 3.17 The experimental results of proposed GCU-N under 'pink' noise conditions for PESQ and STOI.

The median PESQ values (the yellow line in each box) of GCU-N under the 'pink' noise at SNRs of -2dB, 0 dB, 2 dB, 4 dB, 6 dB are 2.18, 2.30, 2.40, 2.50, 2.58 respectively. The median STOI values (the yellow line in each box) of GCU-N under the 'pink' noise at SNRs of -2dB, 0 dB, 2 dB, 4 dB, 6 dB are 0.94, 0.95, 0.96, 0.96, 0.97 respectively.

3.4.2 Discussion

Emphasizing the Speech-to-Noise Ratio (SNR) range from -2 dB to 6 dB (shown in Table 3.3 & 3.4), the GCN consistently outperformed the proposed GAN model and GCU-N across the board, achieving higher scores in both the Short-Time Objective Intelligibility (STOI) and the Perceptual

Evaluation of Speech Quality (PESQ). Notably, where the GCN excelled, the proposed GAN faltered, indicating a critical area of focus for discriminator improvement.

Delving into the specifics, the GCN has better performance in STOI and PESQ over its GAN counterpart, an advancement that signals a pivotal shift towards a more effective use of convolutional layers. In stark contrast, the GCU-N's performance lagged, suggesting that an excessive stack of convolution or deconvolution layers might lead to information distortion rather than enhancement. This revelation prompts a strategic re-evaluation of network depth and model complexity in speech enhancement applications.

The shortcomings of the proposed GAN's discriminator—unable to discern between the generator's output and the original pure speech—illuminated the nuanced challenge of distinguishing enhanced speech from its clean counterpart. This pitfall was notably evident in the discriminator's output, where abstract features extracted by convolution functions might have obscured critical decision-making factors. Future iterations could benefit from structural alterations, such as the elimination of the pooling layer, the incorporation of LSTM blocks for contextual focus, dimensionality reduction of inputs, or the integration of the Attention mechanism, which promises improved context analysis and decision accuracy.

Table 3.3 The summary of proposed models' results.

Evaluation matrix	STOI (AVG)			PESQ (AVG)		
noise	Pink	Volvo	babble	Pink	Volvo	babble
noisy	91.2	96.76	75.13	1.295	1.746	1.105
GCN	**97.2**	98.6	**99**	**2.874**	**3.068**	**2.706**
GAN	96.8	**98.8**	98	2.732	2.866	2.601
GCU-N	95.6	97.6	95	2.392	2.890	2.537

Table 3.4 The improvement ratio of each model comparing with noisy speech.

Evaluation matrix	STOI (AVG)			PESQ (AVG)		
noise	Pink	Volvo	babble	Pink	Volvo	babble
GCN	**6.58%**	1.9%	**31.77%**	**121.93%**	**75.71%**	**144.89%**
GAN	6.14%	**2.1%**	30.44%	110.97%	64.15%	135.38%
GCU-N	4.82%	0.86%	26.45%	84.71%	65.52%	129.59%

Figure 3.18, 3.19, 3.20, 3.21 respectively show the comparison between proposed GCN and GAN results under different noise in STOI and PESQ, where y axis shows the score, and x axis shows the values of corresponding SNR.

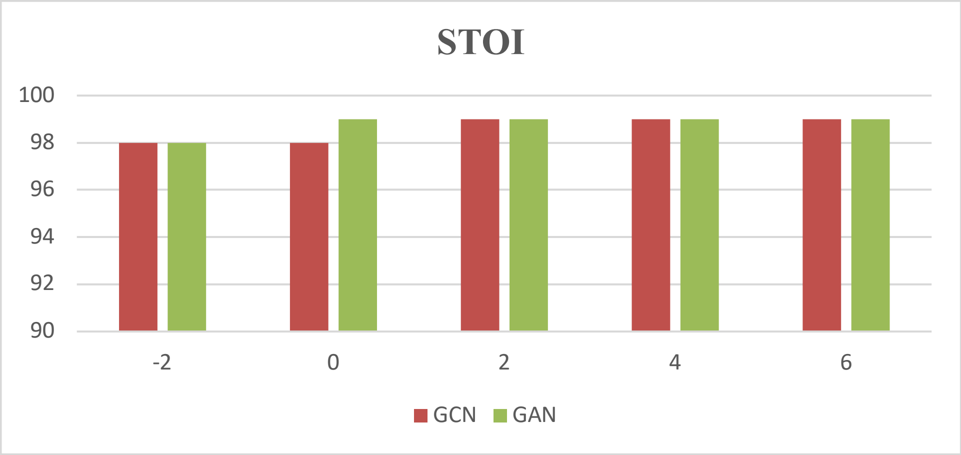


Figure 3.18 The STOI score comparison under volvo noise.

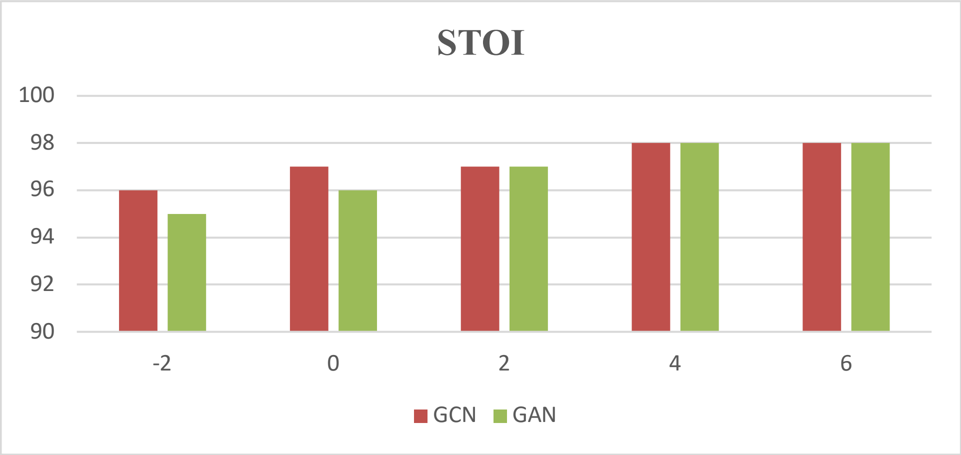


Figure 3.19 The STOI score comparison under pink noise.

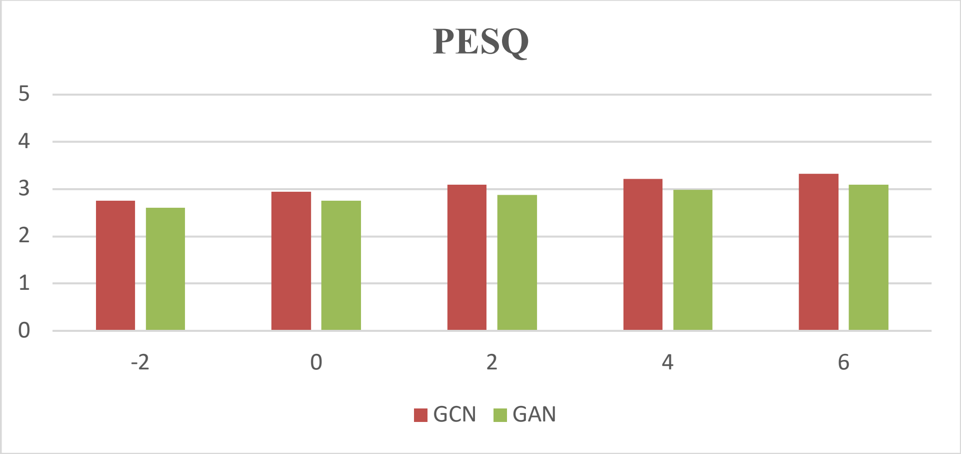


Figure 3.20 The PESQ score comparison under volvo noise.

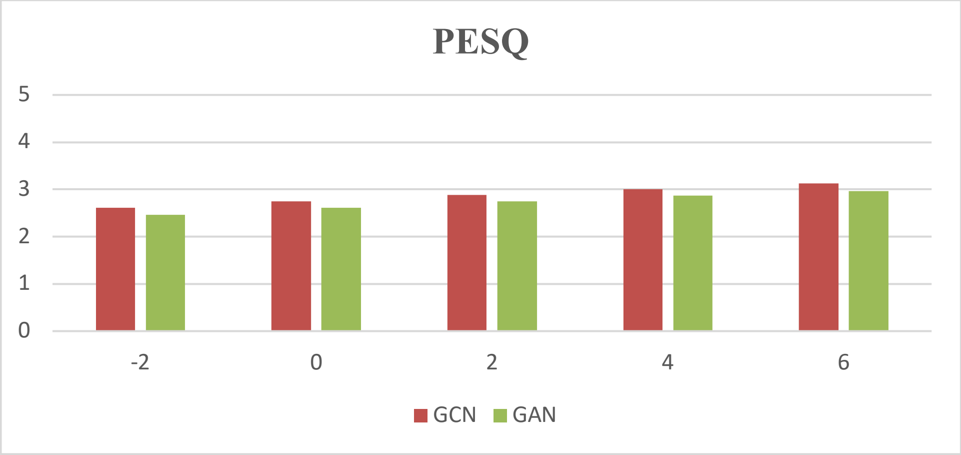


Figure 3.21 The PESQ score comparison under pink noise.

The present study investigates the performance of two models, GCN and GAN, in terms of speech enhancement in the time domain. The results show that both models achieve better scores in STOI and PESQ as the value of SNR increases. Surprisingly, GCN achieves a remarkable performance in babble noise, as evidenced by the improvement ratio. This finding suggests that GCN can learn the features of target speakers effectively, even when the noise is not static, which enhances its robustness. In contrast, GCU-N produces more residual noise and waveform distortion in the speech-enhanced signal, which reduces its performance.

The evaluation results show that GCN outperforms GAN in terms of speech quality improvement, as demonstrated by the PESQ scores. Although the two models achieve similar STOI scores, GCN demonstrates a superior ability to remove background noise and produce minor waveform distortion, which results in higher speech quality. This finding is supported by the magnitude spectra of the proposed models, as shown in Figures 3.22to 3.26, which reveal that the speech enhanced by GCU-N has more residual noise and waveform distortion problems than that enhanced by GCN and proposed GAN.

It is noteworthy that the proposed GCN achieves its superior performance by using an encoder-decoder structure with skip connections and 1D and 2D convolution modules that enhance feature extraction. Moreover, the proposed GCN uses a gated residual unit in conjunction with a normal convolution with a flexible receptive field, which is helpful for time-domain enhancement. It is also possible to develop causal and non-causal GCN, which can modify existing approaches to talker- and noise-independent speech enhancement.

Although the proposed GAN's generator part uses the same structure as GCN, its performance is worse than that of GCN in terms of STOI and PESQ. This finding can be attributed to the fact that the discriminator cannot clearly distinguish between the output of the generator and the original pure speech, which leads to information loss. To address this issue, the discriminator structure can be modified by removing the pooling layer, introducing more discriminator blocks containing LSTM, suppressing the dimension of input fed into the discriminator, or introducing the Attention mechanism. These modifications can help the discriminator to analyze context information and make better decisions.

In conclusion, the present study demonstrates that the proposed GCN outperforms GAN and GCU-N in terms of speech quality improvement, especially in the presence of background noise. the insights gleaned from these experiments not only bolster the GCN's methodology but also forge a path forward for convolutional neural networks in speech enhancement. The encoder-decoder framework with skip connections, paired with 1D and 2D convolution modules enhanced by dilated convolution, positions the GCN as a formidable approach for both causal and non-causal speech enhancement techniques. Moving forward, the development of talker- and noise-independent models stands as the next frontier, underscoring the dynamic evolution of this field and the relentless pursuit of perceptual fidelity in speech enhancement technology.

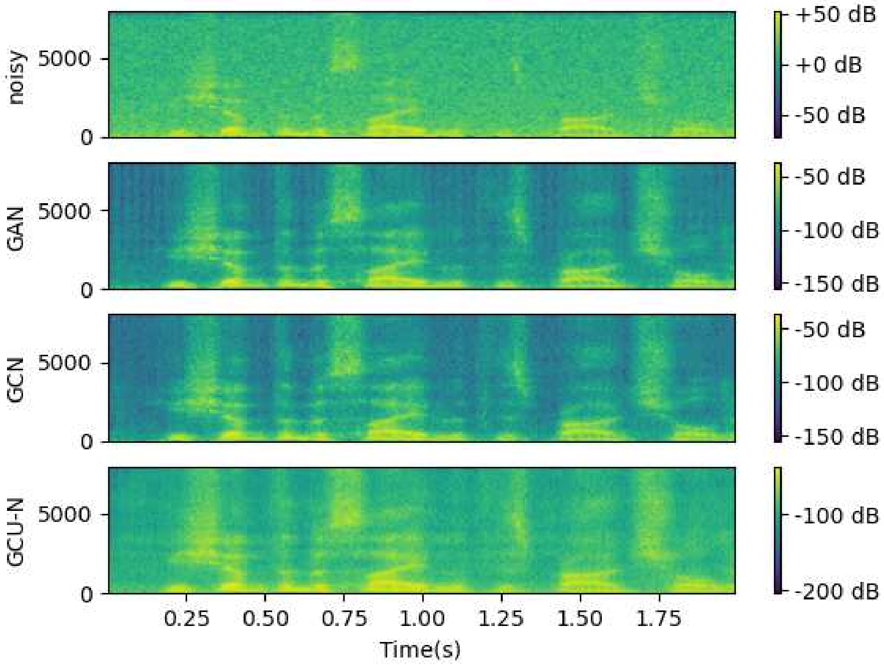


Figure 3.22 The spectrum of noisy speech and enhanced speeches at SNR of -2dB.

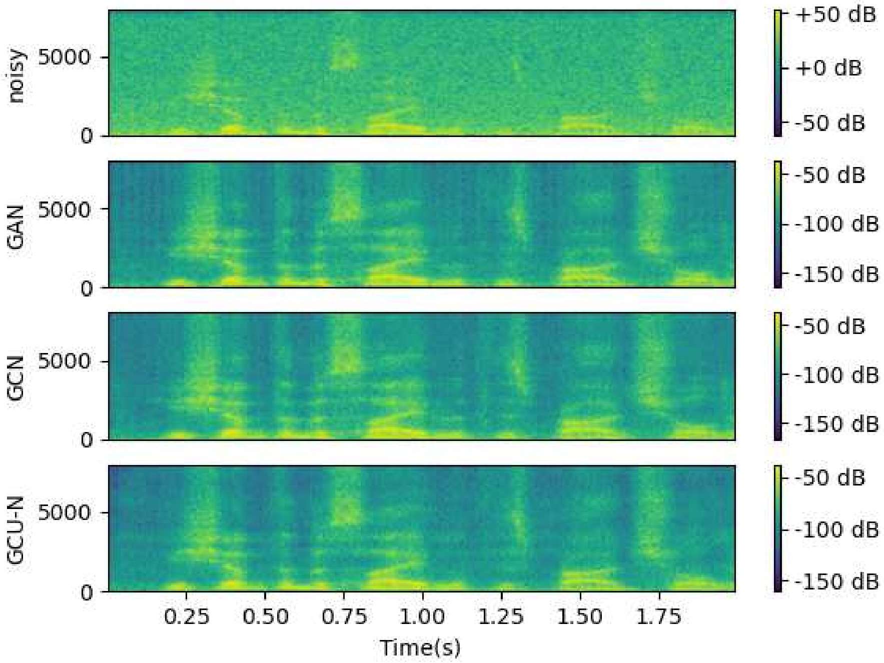


Figure 3.23 The spectrum of noisy speech and enhanced speeches at SNR of 0dB.

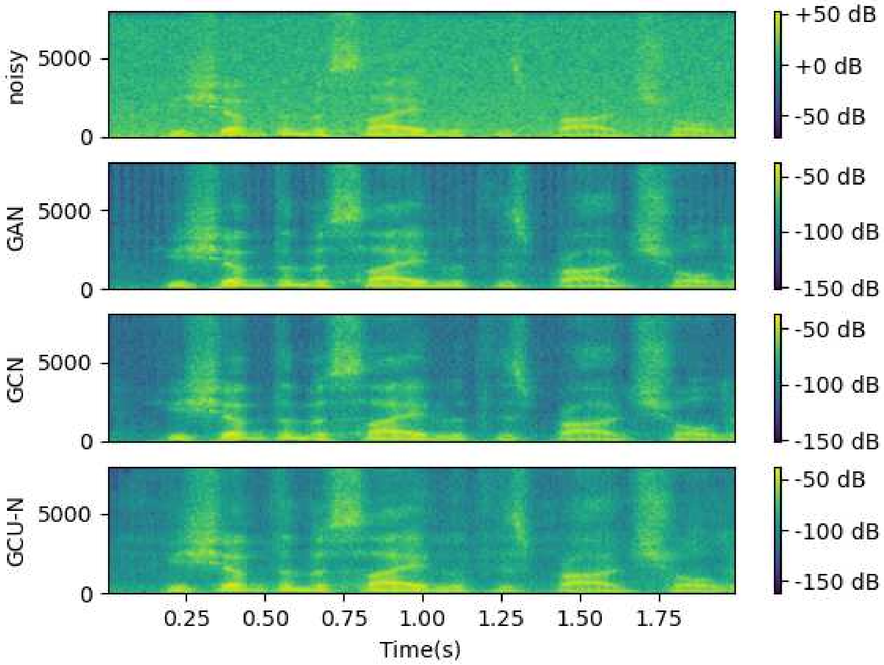


Figure 3.24 The spectrum of noisy speech and enhanced speeches at SNR of 2dB.

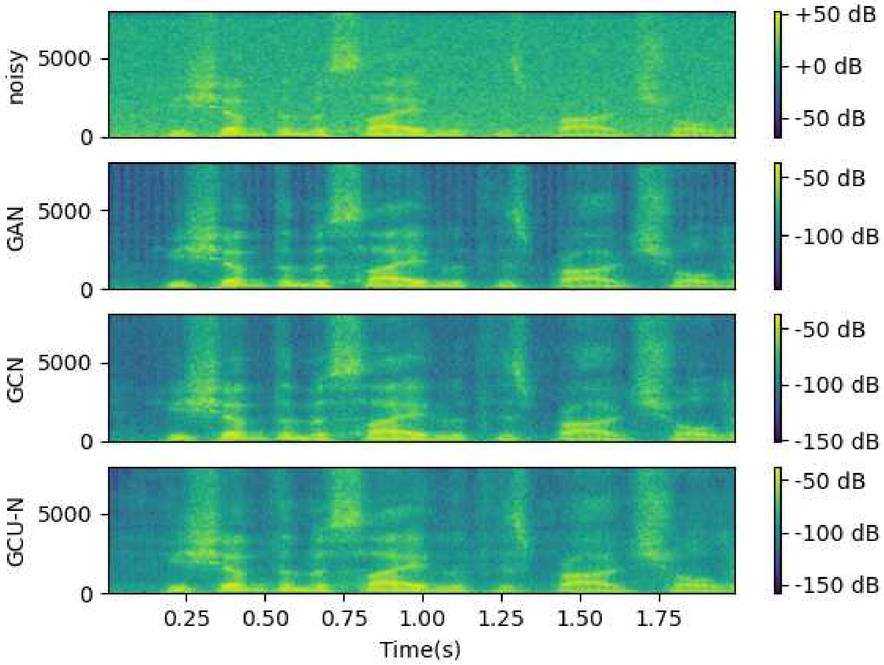


Figure 3.25 The spectrum of noisy speech and enhanced speeches at SNR of 4dB.

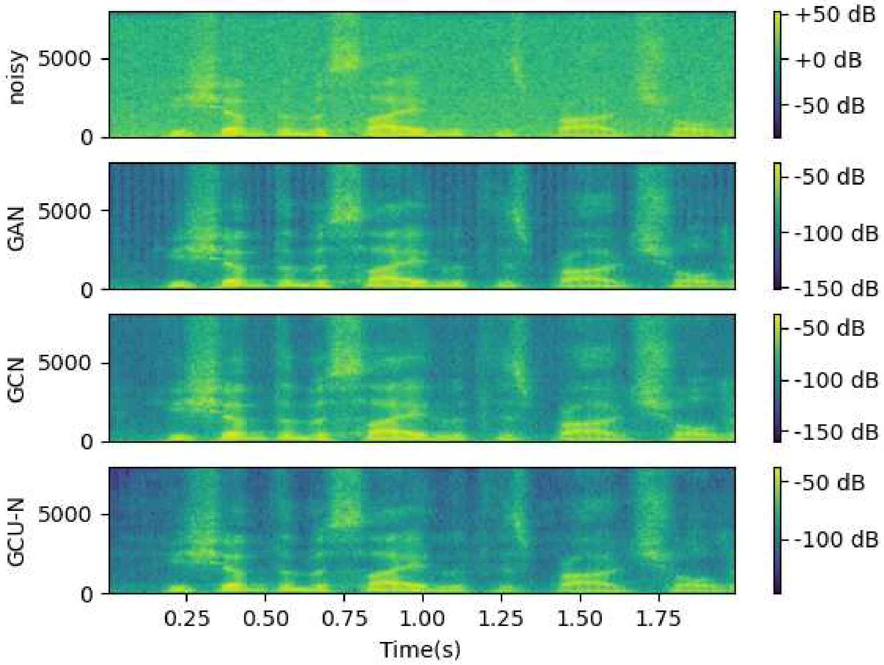


Figure 3.26 The spectrum of noisy speech and enhanced speeches at SNR of 6dB.

As a comparison, we take the test results of state-of-the-art methods as a reference to make a clear comparison in Table 3.5. We use the same background noise 'babble' frome NoiseX-92 [145] as competitors, but we only adopt librispeech as the pure speech while others, like [152], used DNS-2020 [153], DNS-2021 dataset [154].

State-of-the-art methods:

DCN-causal [149]:

Dense CNN: The network architecture is inspired by DenseNet, a state-of-the-art image classification model. In the context of speech enhancement, the Dense CNN allows for the direct flow of information and gradients between layers, aiding in the training process and resulting in enhanced feature propagation. This architecture helps in preventing vanishing gradient issues and promotes feature reuse, which can be vital for modeling the intricacies of speech signals.

Self-Attention Mechanism: The self-attention mechanism enables the model to weigh the significance of different parts of the input signal, allowing it to focus on more relevant portions when making enhancement decisions. This is particularly crucial for speech signals where certain segments (like voiced parts) might be more critical than others (like silent pauses).

TCNN [150]:

Temporal Convolution Layers: The TCNN employs temporal convolution layers to extract sequential features directly from raw time-domain speech waveforms. This approach eliminates the necessity of converting signals into the frequency domain, leading to real-time enhancement with lower latency.

Residual Blocks: The architecture incorporates residual blocks, inspired by ResNet, which facilitate deeper networks by alleviating the vanishing gradient problem. Each block contains convolutional layers, batch normalization, and ReLU activation functions.

Dilated Convolutions: To capture long-term dependencies in speech signals without significantly increasing computational demands, the model uses dilated convolutions in its layers. This allows the network to have a broader receptive field, capturing more temporal context.

Input-Output Features:

Input: The model accepts raw time-domain speech waveforms. This direct use of time-domain signals bypasses the need for spectrogram computation or other time-frequency transformations.

Output: The network produces an enhanced time-domain speech waveform. Post-processing or phase estimation isn't required, simplifying the enhancement pipeline.

Feature Reuse: To retain crucial temporal information, the model utilizes strided convolutions to downsample the feature maps and then upsample them back to the original resolution. This process helps in reusing features across different scales and resolutions.

GCRN [151]

Gated Convolutional Layers: At the heart of the model are its gated convolutional layers, inspired by the gating mechanisms in recurrent networks like GRU and LSTM. These layers adaptively control the flow of information, ensuring that relevant features are emphasized during processing.

Recurrent Layers: Following the convolutional layers, the model uses recurrent layers to capture temporal dependencies in the audio signal, ensuring that temporal patterns and sequences are effectively recognized.

Complex Spectral Mapping: Unlike traditional models that often operate on the magnitude spectrum alone, the GCRN is designed to handle both magnitude and phase, leveraging complex-valued convolutions. This allows the model to maintain and enhance both amplitude and phase information, leading to better speech quality in the enhanced output.

Input-Output Features:

Input: The model's input is the complex Short-Time Fourier Transform (STFT) of the noisy speech, encapsulating both magnitude and phase information in a time-frequency representation.

Output: The GCRN produces an enhanced complex spectrum (both magnitude and phase). This can be transformed back to the time domain via the inverse STFT to yield the enhanced speech signal.

Advantages & Notable Observations: The GCRN's architecture allows it to effectively model and enhance speech in noisy conditions. Its ability to operate on complex spectra (both magnitude and phase) sets it apart from traditional models, and preliminary evaluations indicate significant improvements over existing state-of-the-art methods in various noise scenarios.

GRN [8]

Gated Residual Blocks: Central to the GRN are its gated residual blocks. These blocks combine the advantages of residual connections (from ResNet) and gating mechanisms (from LSTM). The gating mechanism helps in adaptive feature selection, ensuring that only relevant features are propagated through the network layers.

Dilated Convolutions: The model employs dilated convolutions to expand the receptive field without increasing the number of parameters or computational complexity. This enables the network to capture long-range dependencies within the audio signal.

Skip Connections: To facilitate better gradient flow and feature fusion, the architecture incorporates skip connections. These connections allow the model to combine low-level details with high-level semantic features, improving its enhancement capability.

Activation Functions: The model employs the Rectified Linear Unit (ReLU) for non-linearity, ensuring that the network can capture complex relationships in the data.

Input-Output Features:

Input: The GRN model directly takes the noisy speech's Short-Time Fourier Transform (STFT) magnitude spectra as its input. This offers a time-frequency representation, allowing the model to work on both spectral and temporal characteristics of the signal.

Output: The network outputs an enhanced magnitude spectrum. This enhanced spectrum can then be combined with the phase of the noisy speech to reconstruct the time-domain signal using the inverse STFT.

CRN [148]

Convolutional Neural Network (CNN): The core of the model is a CNN that has been trained to map from noisy complex spectrograms to clean ones. The architecture is designed to capture both local and global features from the input spectrogram.

Multi-metrics Learning: To train the CNN, a multi-metrics learning strategy is adopted. This involves three loss functions: mean squared error (MSE) on magnitude, phase, and a cosine similarity term. The combination of these losses ensures that the model learns to reconstruct both the magnitude and phase of the clean speech effectively.

Input-Output Features:

Input: The primary input to the model is the complex spectrogram of the noisy speech. This is obtained using the Short-Time Fourier Transform (STFT), which provides a time-frequency representation comprising both magnitude and phase information.

Output: The CNN produces an enhanced complex spectrogram, which represents the cleaned-up version of the input. This enhanced spectrogram can be converted back to the time-domain signal using the inverse STFT.

Notable Observations:

The proposed approach of directly enhancing the complex spectrogram (both magnitude and phase) aims to overcome the limitations of traditional methods that often neglect phase information. Preliminary evaluations indicate that the proposed model, with its multi-metrics learning strategy, achieves superior performance in terms of both objective and subjective measures when compared to several baseline methods.

SEGAN-T [13]

Generative Adversarial Network (GAN): SEGAN employs the GAN structure, which comprises two primary components:

Generator (G): This component takes in noisy speech and attempts to generate a clean version of it. It uses a fully convolutional network, where the encoder captures a compressed representation of the noisy speech, and the decoder then tries to reconstruct a clean signal from this representation.

Discriminator (D): The discriminator's role is to differentiate between real clean speech and the speech generated by the generator. It provides feedback to the generator, guiding it to produce better-enhanced outputs.

Skip Connections: To ensure that the model can capture both low-level details and high-level features, skip connections (similar to those in U-Net) are used. They bypass certain layers, enabling the network to retain more information.

Adversarial Training: This training strategy pits the generator and discriminator against each other in a game. The generator tries to produce clean speech that the discriminator can't distinguish from real clean speech, while the discriminator tries to get better at telling the difference.

Input-Output Features:

Input: The primary input to SEGAN is a one-dimensional raw waveform of noisy speech. This approach avoids the need for time-frequency domain transformations, such as spectrograms, which are commonly used in other speech enhancement methods.

Output: SEGAN outputs a one-dimensional waveform of the enhanced (or denoised) speech. This direct end-to-end mapping from noisy to clean waveforms allows for a more holistic understanding and manipulation of the speech signal.

In selecting the baselines for comparing our proposed models, we aimed to encompass a broad spectrum of state-of-the-art approaches in the field of speech enhancement. The choice of baselines is dictated by their relevance to the key attributes we seek to measure and improve with our models: intelligibility, quality, and computational efficiency.

Dense CNN (DCN-causal) was chosen due to its foundational architecture, DenseNet, which has shown remarkable success in image classification tasks. We expect that the dense connectivity will aid in capturing the intricate details within speech signals, which is hypothesized to enhance both STOI and PESQ scores by preserving important temporal information.

Temporal Convolution Network (TCNN) serves as a baseline to evaluate the effectiveness of extracting sequential features from raw waveforms without transforming them into the frequency domain. This real-time enhancement capability sets a benchmark for latency, which is a crucial factor in many real-world applications.

Gated Convolutional Recurrent Network (GCRN) [60] incorporates both gated convolutional and recurrent layers, and thus, it is posited as a strong competitor against which the performance of other models can be measured, especially in terms of modeling temporal dependencies and handling complex spectral mappings.

Gated Residual Network (GRN) and Convolutional Recurrent Network (CRN) are included as baselines due to their unique combination of residual and gating mechanisms, and the multi-metrics learning approach, respectively. These characteristics are anticipated to contribute to improved speech enhancement, especially in noisy conditions, and thus provide a comparison for the effectiveness of the proposed methods in similar scenarios.

Lastly, Speech Enhancement Generative Adversarial Network (SEGAN-T) is chosen due to its novel end-to-end waveform enhancement capabilities. It presents a different paradigm by using adversarial training, which has been less explored in speech enhancement. By including SEGAN-T, we aim to assess the efficacy of GANs in generating clear speech from noisy inputs and their impact on both objective and subjective quality metrics.

In summary, these baselines were selected to represent a comprehensive range of approaches in the current literature, providing a robust platform for comparison to demonstrate the advancements our proposed models offer. By establishing these benchmarks, we aim to validate our contributions to the field in terms of innovation, performance, and practical applicability.

Table 3.5 STOI and PESQ comparisons between proposed models in noise 'babble' at -5 dB.

method	STOI (%)	PESQ
noise	babble	babble
Test SNR	-5 dB	-5 dB
DCN-causal [149]	85.3	2.34
TCNN [150]	82.8	2.18
GCRN [151]	82.4	2.17
GRN [8]	80.2	2.16
CRN [148]	79.71	2.15
SEGAN-T [13]	81.5	2.11
Proposed GCN	**99**	**2.24**
Proposed GAN	**98**	**2.19**

Notable Observations:

SEGAN's end-to-end waveform-to-waveform approach is a significant departure from traditional methods that operate in the time-frequency domain. This direct approach aims to capture and utilize the intricate relationships within the raw waveform.

Experimental results demonstrate that SEGAN achieves competitive speech enhancement performance, especially when considering non-intrusive quality measures. Moreover, its capability to generalize well to unseen noise types makes it a promising solution for real-world applications.

After comparing Table 3.5, it is evident that the proposed GCN and GAN models outperform most recent models in terms of PESQ and STOI, even under matched noise conditions. Despite our proposed model topology is like GRN and GCRN, but the middle layer brings a better performance due to is a parallel structure enabling the model can caputure global and local speech information. However, it is important to note that due to computational resource limitations, the training time and dataset size were restricted in this project. For instance, the total training time was approximately 6 hours, and the training data comprised approximately 10 hours. In contrast, DCCRN [152] utilized over 5000 hours of data, while the system setup in [155] required two NVIDIA Volta V100 16GB GPUs and one week of training. It is reasonable to assume that with sufficient resources for training, the proposed models can achieve greater robustness and avoid overfitting [156].

Furthermore, the performance of GCN is satisfactory when the speaker characteristics and background noise are similar. The 'babble' noise used in this study for training and testing the model has demonstrated that GCN is capable of effectively enhancing speech, as evidenced by the STOI and PESQ values in Figure 3.4.16. Nevertheless, to improve the model's robustness, it is necessary to record more multi-speaker audio as background noise in the future. Additionally, diverse training data can aid the model in avoiding gradient vanishing or exploding and overfitting to some extent.

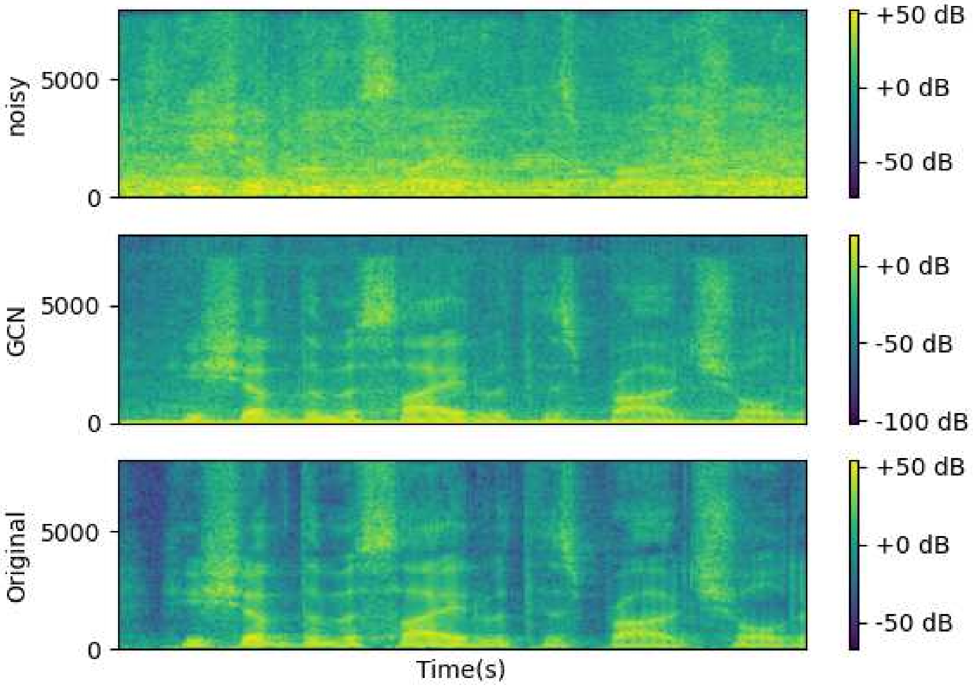


Figure 3.27 The spectrum of noisy speech, enhanced speeches at SNR of -2dB and the original speech.

3.5 Conclusions of this chapter

GANs are commonly viewed as a distinct network architecture, in the context of speech enhancement, they serve as a training method known as adversarial training. GANs consist of a generator and discriminator. The generator can be replaced with various network architectures, and different loss functions can be used as the generator's loss or equivalent. Thus, GANs in speech enhancement act as a pseudo loss function to emphasize noise components and speech artifacts that conventional methods struggle to address.

In this study, our exploration into the realm of speech enhancement is marked by the introduction of a novel convolutional neural network (GCN) with gated residual units as the generator part of the GAN, tailored for time-domain processing. Our GCN architecture, anchored in an encoder-decoder structure augmented with skip connections, has been engineered to harness the power of dilated convolution functions, thereby expanding depth and maximizing context aggregation for superior feature extraction.

Results from our rigorous testing validates the supremacy of the sole GCN over the proposed GAN using the GCN as its generator part and GCU-N models under varied conditions, particularly in terms of STOI and PESQ metrics, where the GCN demonstrates a consistent outperformance.

This suggests that beyond a certain threshold, additional layers may not contribute to enhancement, and could, in fact, distort the information.

This research also concentrates on the exploration of both causal and non-causal CNNs, which hold the potential to transform the approach to speech enhancement independent of specific talkers and noise conditions. Despite the promising results, we recognize the constraints of our study, including limited training data—merely 11 hours—which may impede the depth of learning required for a robust deep learning model. Additionally, the computational limitations of our resources, restricted to personal laptop capacities, have imposed a ceiling on the training duration and, by extension, the potential of our models.

As we move forward, enriching our dataset and extending training times are pivotal steps that could dramatically bolster the robustness and generalization capabilities of our GCN. At the same time, it is notable that the proposed GAN performance is not satisfactory, comparing with the GCN, even though the setup of the generator part is the same as that in GCN. So, the architecture of the discriminator is also worthwhile to focus on. In the future, we should diverse the targets of the discriminator, for example, it not only makes discrimination on fake or true pure speeches, but also identity them with their objective evaluation metrics, such as PESQ and STOI etc.

Chapter 4 A new version GCN with attention mechanism

4.1 Background

Speech enhancement has traditionally revolved around the primary goal of noise suppression, wherein the task is to segregate and remove noise components from noisy signals and retain speech elements. This process, although seemingly straightforward, is nuanced and reveals two distinct strategies: noise suppression and speech enhancement.

The traditional approach of noise suppression involves identifying and nulling time-frequency points dominated by noise. Conversely, speech enhancement involves extracting and preserving time-frequency points where speech is predominant. These two views, while operationally similar, leads to a problematic intersection; attempting the former may induce speech distortion, while the latter may leave residual noise. Common techniques like spectral subtraction, Wiener filtering, and statistical model-based methods estimate a gain function, filtered by the signal-to-noise ratio, that inevitably results in artifacts or distortion at low signal-to-noise frequency points [163]. On the other hand, the method of harmonic regeneration addresses the harmonic distortion problem caused by nonlinear functions [164].

In my previous research, dilated convolution has been effectively used to capture long-term contextual information in speech signals. The effectiveness of this method showcases the potential for enhanced model performance in capturing speech dynamics. However, the human ability to process complex auditory scenes, epitomized by the 'cocktail party effect', suggests that auditory attention is critical. Mirroring this phenomenon, the integration of attention mechanisms—particularly self-attention in sequence-to-sequence tasks—promises to advance the domain of speech enhancement.

The introduction of self-attention is hypothesized to bridge the gap between human and machine auditory processing capabilities. By enabling selective focus on sounds of interest, similar to the cognitive filtering humans demonstrate in noisy environments, the quality and intelligibility of the enhanced speech are expected to significantly improve.

This chapter will explore the application of attention mechanisms juxtaposed with dilated convolutions to enhance the speech enhancement models. Therein, the exploration will seek to redefine and refine the process, taking cues from both, the traditional and the novel, to arrive at a methodology that minimally distorts speech and adeptly suppresses noise. Another contribution of this chapter is the introduction of the ‘Masking first and Mapping second’ method. In this cascade topology, the second-stage model will enough tolerance to rectify previous stages' errors, which can better restore high-fidelity enhanced speech.

4.2 Methodology

Self-attention has proven to be a powerful tool in various fields, such as image generation, machine translation, and automatic speech recognition. Specifically, self-attention can better capture the context of spoken utterances that contain many repeating phones, especially in low SNR conditions where phones can be present in both high and low SNR regions of the utterance. By attending over phones in high SNR regions, a speech enhancement system based on self-attention can better reconstruct phones in low SNR regions. Although the self-attention mechanism is effective in embedding context, in short-speech experiments, the performances of CNN, RNN, and self-attention are almost the same. Therefore, in future experiments, we plan to set each input data to more than 2 seconds to fully exploit the advantages of the attention mechanism.

Self-attention mechanism

The attention mechanism has 3 parts: query (Q) whose size is $\mathrm{Time\_Q} * \mathrm{Q\_dim}$, key (K) whose size is $\mathrm{Time\_V} * \mathrm{Q\_dim}$ and value (V) whose size is $\mathrm{Time\_V} * \mathrm{V\_dim}$. Firstly, the correlation scores are from the following equation:

$$W = QK^T \tag{4.1}$$

Where K^T is the transpose of K and the size of W is $\mathrm{Time\_Q} * \mathrm{Time\_V}$. Then, the softmax is introduced to calculate the probability values P. Finally, the attention output A is obtained by P and V.

$$P = Softmax(W) \tag{4.2}$$

$$A = P * V \tag{4.3}$$

In recent years, attention mechanisms have gained considerable traction in the realm of speech enhancement, presenting a paradigm shift in how temporal dependencies within speech signals are modeled [159]. Originating from the success in tasks like machine translation, the attention mechanism [95] in speech enhancement operates by computing similarities between the current frame and a range of past frames, dynamically assigning weights to these past frames based on their relevance [158].

Compared to traditional LSTM and RNN approaches, methods employing attention mechanisms have demonstrated advancements in both the quality and intelligibility of enhanced speech [159]. However, while their efficacy is noteworthy, it's crucial to recognize their limitations. For instance, attention-based models can sometimes overemphasize certain frames leading to unintended artifacts in the enhanced speech [160]. Additionally, their performance can be contingent on the choice of similarity metric used, which presents challenges in diverse noise environments [161].

Furthermore, while attention mechanisms have shown improved performance metrics in several benchmarks, their computational complexity, especially in deeper architectures, can be a deterrent for real-time applications [162]. As research in this domain progresses, it will be imperative to balance the benefits brought by attention mechanisms with the constraints of practical implementation.

However, since input models contain both clean speech information and noise information on each time-frequency unit, this method amplifies both clean speech information and noise information when weighting and does not significantly suppress noise. Therefore, suppressing noise in attention mechanism operation is a breakthrough for improving performance of speech enhancement that needs to be solved at present.

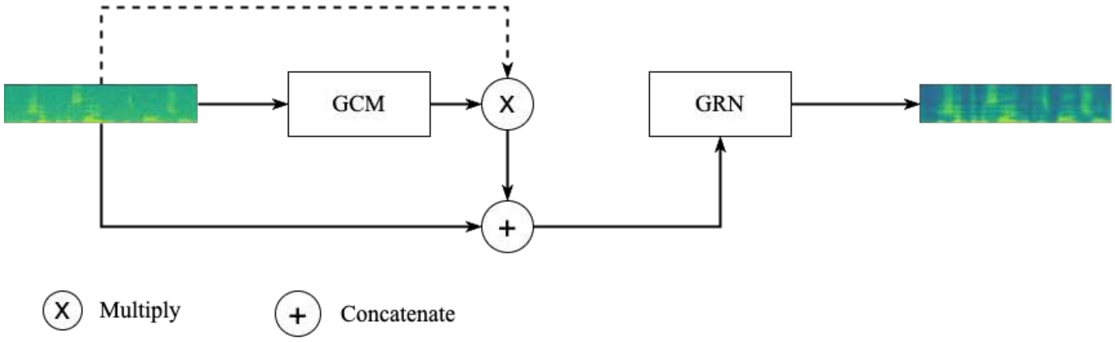


Figure 4.1 The architecture of proposed model.

The proposed method (as shown in Figure 4.1) utilizes a Gated Residual Network (GRN) consisting of an encoder, decoder, and stacked Gated Residual Units (GRUs), as shown in Figure 4.2, to generate a mask that suppresses noise in its corresponding domain, thereby filtering coarse features towards the overall spectrum. The encoder and decoder consist of 5 sub-layers each, where the former uses a 1-D convolutional layer with Batch Normalization and an ELU activation function, while the latter replaces the convolution layer with a transposed convolution layer. The stacked GRUs incorporate multi-head self-Attention (MHSA) in the frequency and time dimension to improve the model's acoustic receptive field and processing ability for sequential temporal information. The network also incorporates skip connections to add features extracted from corresponding layers into the final prediction. The Inverse Short-Time Fourier Transform (ISTFT) is implemented through convolutional layers inspired by prior work [157] to allow the use of time-domain enhanced speech for further training. Finally, the original time-domain speech's phase is utilized to compensate for noisy speech phase and reconstruct more accurate time-domain enhanced speech.

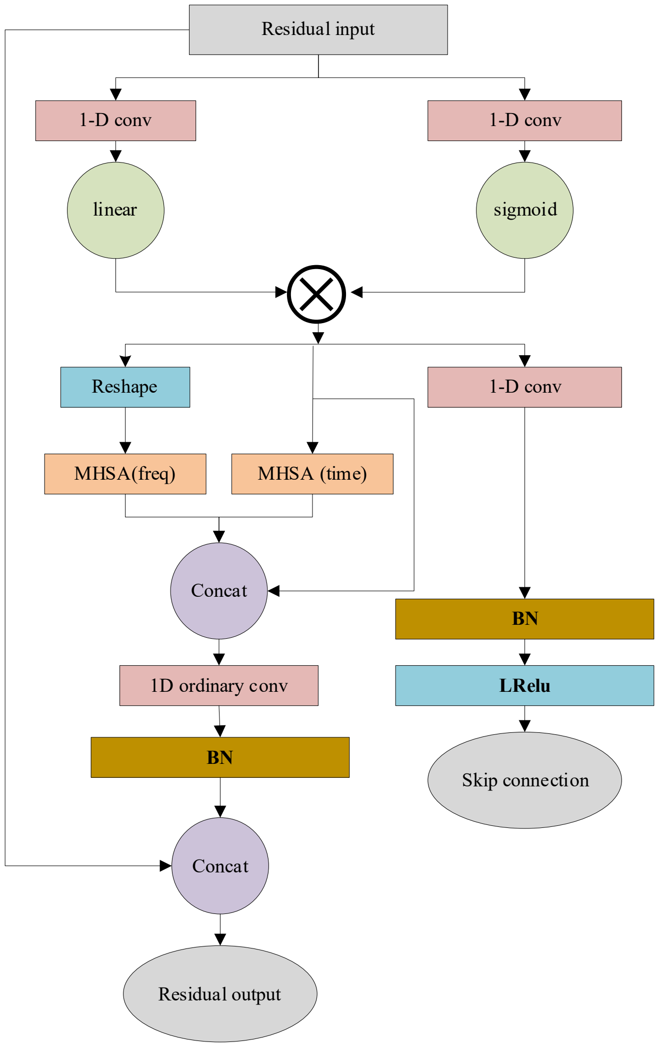


Figure 4.2 The architecture of proposed GRU in GCM.

MHSA is often used to extract long term sequence information [165]. MHSA takes as input an L-length sequence feature and produces an output sequence of the same size. This attention mechanism, which is also introduced in Chapter 6, can be described by

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{L}}\right)V \tag{4.4}$$

Where T represents the transpose symbol. The input shape is [Time_frames * Frequency_bins] from encoder, then the proposed MHSA for frequency reshapes the input into [Frequency_bins * Time_frames], as shown in Fig 4.2. Each sub-MHSA can map information along their own specific axis. At the same time, 2 sub-MHSA mechanisms are constructed in parallel and finally their outputs are concatenated together with the original input as the input for the next step. The MHSA for the time frame is described as

$$multi_{head(Q^t,K^t,V^t)} = concat(h_1, h_2, \dots, h_t)W_t^{out} \tag{4.5}$$

Where

$$h_i = Attention\left(Q^t W_t^Q, K^t W_t^K, V^t W_t^V\right) \tag{4.6}$$

Finally, the attention maps are concatenated with the original input and processed by a 1-D convolutional layer to obtain the residual output as the next GRU's input, and is described by:

$$Residual\ output = Conv\left(MHSA_{time} + MHSA_{freq}\right) \tag{4.7}$$

$$Residual\ output = Conv\left(MHSA_{time} + MHSA_{freq}\right) \tag{4.7}$$

4.3 Experiment setup

The loss function and dataset used in the experiments for Gated Convolutional Networks (GCN) were also used for the proposed Gated Convolutional Masking (GCM) method. The GCM method utilizes a Gated Residual Network (GRN) architecture with 5 encoder and decoder layers and 20 Gated Residual Units (GRUs) grouped into 4 GRU groups. In the GRU layers, the left 1D-conv kernel size was 5 and the right 1D-conv kernel size was 1, with a channel length of 256. Multi-Head Self-Attention (MHSA) was incorporated with 2 heads. The Gated Residual Masking (GRM) architecture had the same setup of kernel size, stride, and channel in the GRU layers as in the GCM method. The encoder and decoder layers had a kernel size of 8, channel length of 512, and stride of (1, 2) in the time and frequency axes. The middle part of the GRM comprised of 4 GRU groups, with each group containing 5 GRUs with dilation rates of [1, 2, 4, 8, 16].

For the dataset setup, we utilized clean speech segments from the ICASSP DNS3 dataset [167], which we then combined with noise sources from NoiseX-92 [145] for training dataset (about 100 hours). While the testing dataset was composed of clean audio extracts from the 5-hour Librispeech corpus [147] mixed with NoiseX-92 noises.

For our training set, we cut the utterances into 2-second segments. But for the test set, we didn't make any cuts, so the lengths vary. We used a Hamming window with a 25 ms window length (equivalent to 400-point FFT) and a hop size of 200 points, which means there's a 50% overlap. The input feature of the proposed model is the magnitude spectrum.

All audio data was sampled at 16kHz and extracted by using frames of 512 with frame shift 256, and the model is directly fed by raw waveform. All models are optimized using the Adam-algorithm [166] with a learning rate of 0.001, decaying by half after each epoch.

In this experiment, to carry out the training process, two GTX 1080Ti 10GB GPUs were used, and the training process took approximately 30 hours for 10 epochs. The computation resources were generously offered by the High-Performance Computation (HPC) at the University of Southampton.

To evaluate the quality of the denoised speech, we picked a range of standard metrics. We used PESQ, which has a score range from -0.5 to 4.5. For judging how clear the speech sounds, we used ESTOI, which scores between 0 and 1. For all these metrics, a higher score means better speech quality.

4.4 Results and analysis

4.4.1 Baseline introduction

The findings of our study are presented in Table 4.1, which includes a comparison of our proposed model with five state-of-the-art (SOTA) models that were selected based on their similarity of methodology and recentness: Gated Convolutional Recurrent Network (GCRN) [168], Deep Complex Convolutional Recurrent Network (DCCRN)[152], Phase-Only Speech Enhancement Network (PHASEN) [169], Audio Enhancement Convolutional Neural Network (AECNN) [170], and Convolutional Time-domain Audio Separation Network (ConvTasNet) [19]. GCRN, DCCRN, and PHASEN use complex-domain features as input features and involve magnitude spectrum and phase recovery, while AECNN and ConvTasNet use raw time-domain waveform as input and output. The reason why we choose these competitors is they did not adopt attention mechanism and all of them take the encoder-decoder structure as baseline same as our proposed model. Their details go as follow:

Deep Complex Convolutional Recurrent Network (DCCRN)

Encoder: The encoder comprises eight complex convolutional layers. This encoder extracts hierarchical features from the noisy input and down-samples the feature maps by half after each convolution.

Recurrent Enhancement Blocks (REB): After the encoder, there are four REB blocks. Each REB block contains two Complex Gated Recurrent Units (CGRUs) and one complex convolutional layer. These blocks capture long-term dependencies and refine the hierarchical features.

Decoder: The decoder, similar to the encoder, has eight transposed complex convolutional layers. It up-samples the feature maps and finally reconstructs the enhanced magnitude spectrogram.

Active Perception: This is a significant contribution of the paper. Active perception is inspired by the human auditory system, where humans can actively focus on a specific sound source. In DCCRN+, this concept is realized by introducing a weighted summation operation after each REB block. This operation helps the network emphasize certain temporal and spectral regions during speech enhancement.

Input Feature: The input to the DCCRN+ model is the magnitude spectrogram of the noisy speech. This is derived by applying the Short-Time Fourier Transform (STFT) to the raw noisy speech waveform. The phase of the noisy speech is kept aside and used later during the waveform reconstruction.

Output Feature: The model outputs an enhanced magnitude spectrogram. This is then combined with the phase of the noisy speech (stored from the input stage) to perform an inverse STFT and obtain the enhanced speech waveform.

Phase-Only Speech Enhancement Network (PHASEN)

The **PHASEN** model is introduced as a novel approach to speech enhancement, focusing on both the phase and magnitude components of speech signals. Its architecture and design principles are tailored to address the unique characteristics and importance of phase in speech enhancement tasks.

Magnitude Subnetwork: This segment of the network handles the enhancement of the magnitude spectrogram. It comprises a U-Net-like architecture with convolutional layers, aiming to suppress noise while preserving the speech magnitude.

Phase Subnetwork: This segment explicitly deals with the phase component of the speech signal. It utilizes a similar U-Net-like structure, but with a different design to specifically cater to the phase's cyclic nature.

Harmonics Block: Recognizing that speech signals have harmonic structures, this block is introduced to capture and leverage such harmonic relations, further enhancing the model's performance.

Input Feature: The PHASEN model takes in both the noisy magnitude and phase spectrograms derived from the Short-Time Fourier Transform (STFT) of the noisy speech signal.

Output Feature: The model outputs enhanced magnitude and phase spectrograms. These are combined to perform an inverse STFT to derive the enhanced speech waveform.

Harmonics and Phase Preservation: One of the standout features of PHASEN is its explicit emphasis on preserving the harmonics and phase of the speech signal, understanding their critical role in speech intelligibility and quality.

Audio Enhancement Convolutional Neural Network (AECNN)

The **AECNN** model is proposed as a solution to both denoising and dereverberation of speech signals, aiming to enhance speech quality and intelligibility. The model is structured to adaptively learn the combination of denoising and dereverberation filters, making it particularly effective in scenarios with unknown noise and reverberation conditions.

Adaptive Enhancement Block: This core block employs two enhancement filters - a denoising filter and a dereverberation filter. The block learns to adaptively combine these filters based on the characteristics of the input speech, ensuring optimal enhancement for varied conditions.

Enhancement Filter: This filter is designed using a convolutional neural network (CNN) architecture. It processes the input noisy and reverberant speech to produce an enhanced output.

Adaptive Combination: A sigmoid activation function is applied to the output of the enhancement filter to adaptively determine the combination of denoising and dereverberation, ensuring the most suitable enhancement is applied.

Input Feature: The AECNN model takes in a spectrogram derived from the Short-Time Fourier Transform (STFT) of the noisy and reverberant speech signal.

Output Feature: The model outputs an enhanced spectrogram, which, after applying an inverse STFT, results in the enhanced speech signal.

Performance: The AECNN model, with its adaptive enhancement strategy, achieves notable improvements in speech quality and intelligibility over several benchmark methods. It is particularly effective in scenarios with unknown noise and reverberation conditions, demonstrating its versatility.

Deep Complex Convolutional Recurrent Network (DCCRN) has been selected as a baseline for its innovative integration of complex convolutional and recurrent structures. The hierarchical feature extraction and long-term dependency modeling capabilities of DCCRN make it an exemplary candidate to evaluate against, especially with its focus on enhancing magnitude spectrograms. Its incorporation of active perception mechanisms, inspired by the human auditory system, also provides a unique angle on speech enhancement that we aim to explore in comparison to our methods.

Phase-Only Speech Enhancement Network (PHASEN) represents a novel direction in the field by concentrating on the enhancement of both phase and magnitude components of speech signals. Given the critical role of phase in speech intelligibility and quality, PHASEN's unique harmonics block and phase subnetwork make it a valuable comparison point for assessing the performance of models that aim to comprehensively enhance speech signals.

Audio Enhancement Convolutional Neural Network (AECNN) is included as a baseline due to its dual focus on denoising and dereverberation of speech signals. Its adaptive enhancement block and the capability to learn the optimal combination of filters for varied acoustic scenarios set a performance standard, especially in conditions with unknown noise and reverberation characteristics. AECNN's approach to adaptively managing enhancement strategies provides a distinct contrast to other methods, highlighting its potential for robust real-world applications.

Each of these models brings a unique perspective to speech signal processing, addressing different aspects of the enhancement task. By comparing our proposed methods to these baselines, we can showcase the specific advantages our models may have in terms of spectral-temporal feature processing, complex signal modeling, and adaptive enhancement in varied noise conditions. The selection of these baselines also facilitates a comprehensive evaluation over multiple dimensions of speech enhancement, including quality, intelligibility, and computational efficiency, which are crucial for the deployment of these models in practical settings.

Table 4.1 Result of all tested models show PESQ, ESTOI values for SNRs between -3 and 6 dB. ‘noisy’ is the original noisy speech (Babble). ‘Proposed model’ is the new version GCN. Higher is better.

Metrics	PESQ					ESTOI(%)				
Test SNR (dB)	-3	0	3	6	Avg.	-3	0	3	6	Avg.
Noisy	1.61	1.77	1.99	2.19	1.89	31.59	40.23	49.61	58.64	45.02
GCRN	2.32	2.62	2.87	3.08	2.72	59.57	68.83	75.72	80.78	71.23
DCCRN	2.31	2.61	2.88	3.11	2.72	59.57	68.11	75.86	81.56	71
PHASEN	2.36	2.7	2.99	3.21	2.82	61.78	71.25	78.32	83.31	73.67
AECNN	2.32	2.64	2.91	3.11	2.74	62.13	71.57	78.25	83.03	73.74
ConvTasNet	2.26	2.52	2.76	2.96	2.63	63.88	72.21	78.49	83.22	74.45
Proposed model	**2.91**	**3.01**	**3.03**	**3.11**	**3.015**	**68.67**	**76.53**	**82.16**	**85.31**	**77.92**

Table 4.1 summarizes the results of all experiments, including baseline models and comparison with our proposed model. The table presents the results in terms of average Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) for the original noisy speech. Our proposed model outperforms all other models in terms of both average PESQ and STOI. However, it should be noted that the improvement in speech enhancement also depends on the Signal-to-Noise Ratio (SNR), which is not captured by the average values presented in the table.

This experiment sought to address the critical challenges highlighted in the introductory portion of this chapter: the distortion of speech signals during noise suppression and the residual noise commonly found after speech enhancement processes. By implementing mixing Gated Convolutional Masking (GCM) with Gated Residual Networks (GRM), our model surpasses the DCCRN benchmarks with an average improvement of 10.8% in Perceptual Evaluation of Speech Quality (PESQ) and 9.74% in Extended Short-Time Objective Intelligibility (ESTOI) scores. These enhancements are attributed to our model's superior handling of waveform information throughout the training process—a feature not present in DCCRN.

The introduction of Multi-Head Self-Attention (MHSA) within GCM has significantly elevated the model's ability to parse through time-frequency dimensions, mirroring human cognitive capabilities to focus selectively on auditory information. Meanwhile, the introduction of MHSA contributes to performance gains when compared to existing models like ConvTasNet, with our model achieving average improvements of 14.6% in PESQ and 4.7% in ESTOI.

Our model's architecture has a good performance at various signal-to-noise ratios (SNRs) conditions. However, it also performs well under the stringent conditions of lower SNRs, where it consistently outperforms its counterparts. This robustness in challenging scenarios underscores the efficacy of our dual architectural strategy, combining the detail oriented GCM for waveform fidelity with the GRM's potent capacity for targeted reconstruction.

The application of MHSA in both time and frequency dimensions further enables our model's advanced capability to synthesize global and contextual information, thereby more popular to replace traditional dilated convolution techniques in speech enhancement tasks.

4.5 Conclusions of this chapter

In our initial investigations, we centered our attention on a feature-mapping topology, uniquely designed to process noisy speech signals. This approach directly produced enhanced speech, sidestepping the traditional use of a spectrogram mask. Notably, we confined our input parameters to primarily consider the magnitude as the primary distinguishing feature. Although the enhanced speech output was of satisfactory quality and intelligibility, we faced an inherent limitation: the reliance on phase information derived from the noisy speech during the resynthesis phase.

Further, the experiment detailed in Chapter 4, which utilized raw waveform as the model's input, brought forth another challenge. This approach struggled to adequately capture and represent the nuances of the target speaker's characteristics, especially when benchmarked against the intricate facets of human auditory perception.

Such limitations offer significant avenues for refinement. The upcoming chapter will take a deeper consideration into these considerations, charting out potential enhancements and future research directions to bolster the efficacy of our model.

Chapter 5 Parallel GCN fed by both magnitude spectrum and complex value domain

5.1 Background

Although the first two chapters have brought many performance improvements, the above algorithms still remain focused on modeling speech patterns (or signal-to-noise ratio). Since it comes to noise suppression, the assistance of explicit noise information in speech enhancement should inevitably be discussed. Historically, noise spectrum estimation has been a cornerstone task in speech enhancement, exemplified by methods like spectral subtraction. Due to the diversity of noise types, accurately estimating noise spectra for complete removal from noisy speech signals is a big challenge.

In 2020, Xu et al. [181] explicitly estimated noise spectra, integrating them alongside noisy speech spectra as network inputs to derive denoised speech. This approach can be likened to a "masking and completion" process, drawing inspiration from traditional noise spectrum estimation methods. Liu et al. [182] in 2021 combined amplitude-phase compensation and noise suppression-speech restoration concepts, employing a multitask learning approach to estimate noise and speech amplitude spectra simultaneously, thereby avoiding the noise spectrum estimation challenges.

Liu et al. adopted a dual-stage framework that jointly enhanced amplitude and noisy speech spectra, avoiding the issues on noise spectrum estimation. Zheng et al. [183] introduced a dual-branch structure to jointly estimate speech and noise spectra, facilitating information exchange to enhance spectral estimation accuracy.

The outlined methodologies represent significant strides in explicitly incorporating noise information in speech enhancement strategies. These advancements present novel avenues for refining speech processing models by integrating explicit noise insights. The upcoming experimental sections will delve into the practical implementations and evaluative frameworks to ascertain the effectiveness and efficiency of these cutting-edge approaches in enhancing speech quality and robustness in noisy environments.

In the field of Speech Enhancement (SE), there are two main approaches: mapping-based and masking-based methods. Mapping-based methods utilize spectral magnitude or complex-valued features as input to restore the clean speech [171]. Masking-based methods employ either ideal binary mask (IBM) or ideal ratio mask (IRM) [172][173]. For IBM, the magnitude and phase information are individually used in the complex domain to estimate the clean speech, whereas for IRM, the original phase information is directly utilized to reconstruct the output. Typically, mean square error (MSE) and scale-invariant SNR (SI-SDR) [174] are employed as loss functions for deep neural networks (DNNs). However, speech quality estimation is challenging, as it has weak correlations with human ratings [175].

In recent times, cascaded or multi-stage concepts have been proposed for SE [176]. These approaches leverage intermediate priors to improve optimization by decomposing the original task into several sub-tasks. However, each sub-model's performance is constrained, as they incrementally improve the SNR. A two-pipeline structure was suggested in [176], consisting of a coarse spectrum method followed by a compensating and polishing method. Nonetheless, the performance of the second-stage model is highly reliant on the previous output, thus in a cascade topology, the second-stage model should have enough tolerance to rectify previous stages' errors.

In recent advancements in the field of speech enhancement using deep learning, the work of Li et al. [177] stands out, particularly in the context of optimizing spectral components of speech. In their pivotal study, Li and colleagues proposed an innovative two-stage complex spectral mapping approach. They emphasized the pivotal role of separately optimizing the magnitude and phase components of speech signals. Through their experiments, they established that by decoupling the optimization of these two components, one can achieve notable improvements in the clarity and quality of enhanced speech. This decoupling methodology resonates with our following proposed compensation path, underscoring its relevance and potential in contemporary speech enhancement techniques. Such findings, as presented by Li et al., set a foundational benchmark and provide valuable insights for future research in this domain [177].

In this chapter, we propose a novel parallel structure comprising two modules to perform coarse and refined estimation. The first module, named Compensation for Complex Domain Network (CCDN), is responsible for calculating masked features that compensate for complex components from the second module. To achieve this, we use a parallel-path structure where one path takes the magnitude spectrum as input and estimates a mask, while the second path outputs the complex domain details. However, the mask path only deals with the magnitude information, which leads to the loss of some spectral details. To address this issue, we introduce the compensation path, which aims to remove distortion and compensate for lost details. Furthermore, in our proposed model, we employ a module that extracts more abstract feature details to facilitate the next estimation. Another contribution of this chapter is that comparing with the sole raw waveform, we use both of magnitude and complex-valued domain as input features. It can better help to capture and represent the nuances of target speakers' speech characteristics.

5.2 Signal model formulation

Single-channel speech enhancement aims to remove the background noise n from the single-channel noisy speech y, and the corresponding original clean speech denotes x, which is expressed by equation 5.1.

$$y[t] = x[t] + n[t] \tag{5.1}$$

Where t represents the time sample index. Meanwhile, we use the short-time Fourier transformation (STFT) to convert the time domain speech signals into time-frequency (TF) domain, that is:

$$Y_{t,f} = X_{t,f} + N_{t,f} \tag{5.2}$$

Where y, x, n are transformed as Y, X, N by STFT, respectively. t is the corresponding time index and f is the frequency bin. Eq. 5.2 can also be written as

$$Y_{r\,(t,f)} + jY_{i\,(t,f)} = \left(X_{r\,(t,f)} + N_{i\,(t,f)}\right) + j\left(X_{r\,(t,f)} + N_{i\,(t,f)}\right) \tag{5.3}$$

where subscripts r, i respectively represent the real and imaginary part of the complex-valued feature. In the rest content, the (t, f) will be dropped.

5.3 Methodology

The diagram of the proposed model is presented in Figure 5.1, which is composed of four blocks: Feature Extraction Block (FEB), Mask Block (MB), Complex-valued Enhancement Block (ComEB), and Compensation Block (CB). The input to the model is the noisy complex spectrum, denoted as $X = Cat(X_r, X_i) \in R^{T*F*2}$, where T and F represent the time frames and frequency bins, respectively, and X_r and $X_i$are the real and imaginary parts of the complex spectrum. The target output is denoted as $S = Cat(S_r, S_i) \in R^{T*F*2}$, where subscripts r, i are the real and imaginary parts of the target output, respectively. The concatenation operation is represented by Cat.

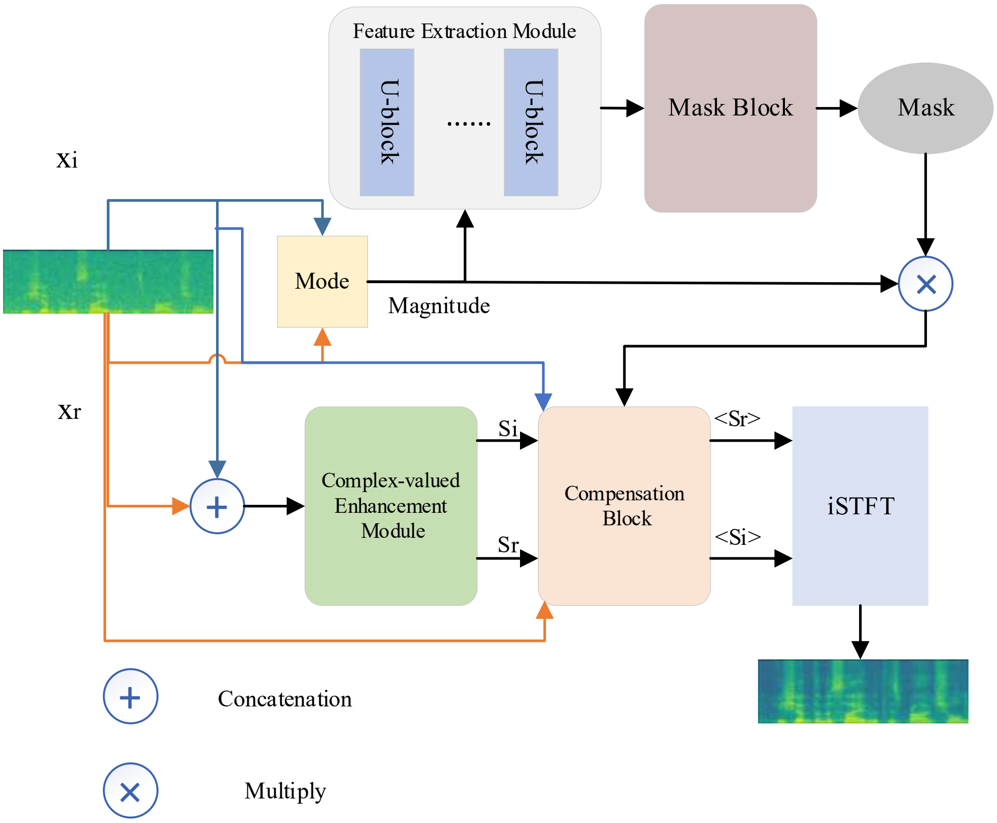


Figure 5.1 The architecture of the proposed Compensation for Complex Domain Network.

5.3.1 Feature extraction block

Acoustic feature extraction using U-nets has been successful [178], but it suffers from the loss of spectral information due to consecutive up and down sampling. For instance, the power spectral density of harmonic structure from low to high frequency regions gradually attenuates. Additionally, the correlation between adjacent frames is crucial in speech processing, making it necessary to obtain both local and global information of each speech sample. In 2020, U2net [179] was proposed, which employed sub-Unet as an embedding layer with residual learning to effectively learn multi-scale features. This study was motivated by the success of U2net and proposes replacing the traditional 2-D convolutional layer with a U-block module, where LSTM is used as the middle layer to mitigate information loss, as shown in Fig 5.2. The FEB, as shown in Fig 5.3, is comprised of Gated Linear Unit (GLU), Layer Normalization (LN), ELU activation function and U-block with residual connection. This structure has two advantages. Firstly, the U-block can capture multi-scale information between frames, resulting in better abilities to capture contextual features. Secondly, the 2-D GLU can filter out disturbing information and keep useful details. The progress of FEB can be represented as follows:

$$y = GLU(x) + U_{block(GLU(x))} \tag{5.4}$$

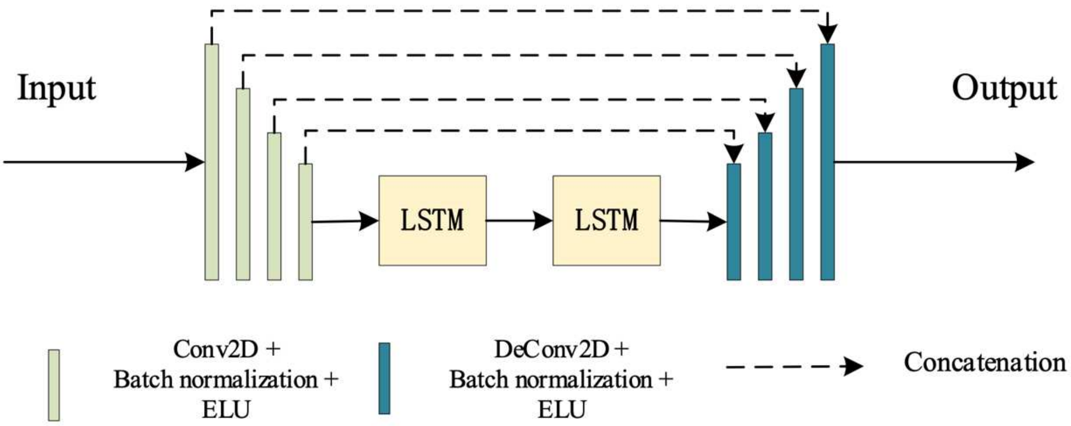


Figure 5.2 The U-net structure.

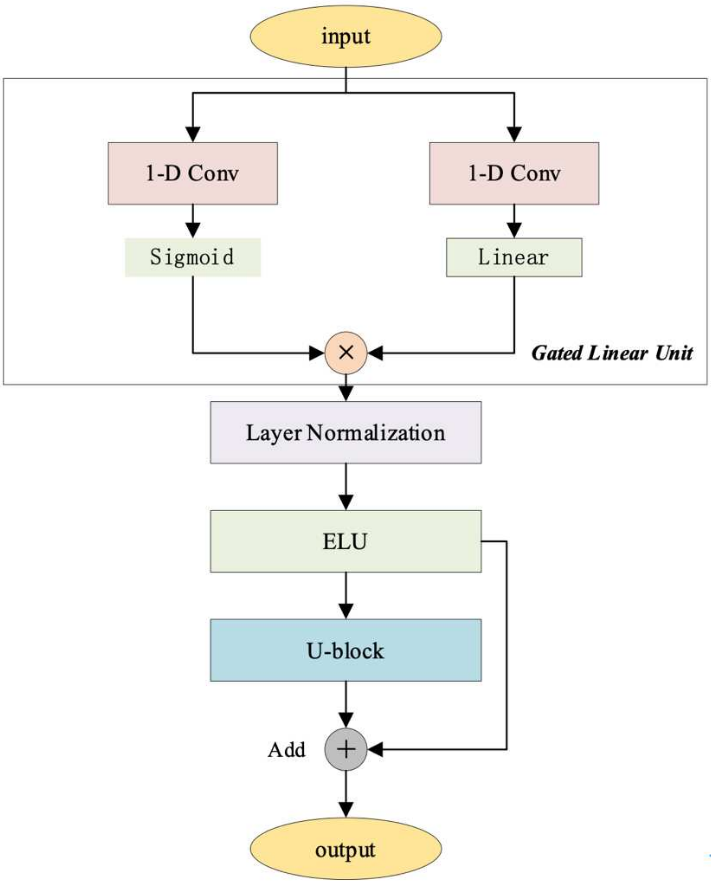


Figure 5.3 The architecture of proposed FEB module.

5.3.2 Mask Block (MB) and Complex-valued Enhancement Block (ComEB)

In this study, the Mask Block (MB) plays a crucial role in enhancing speech signals by suppressing noise in the magnitude domain. The output of MB is a mask that filters out the noise in the input signal. The filtered signal is then used as coarse features that are further processed to obtain the overall spectrum. To achieve this, we adopted the same structure as the new version of Graph

Convolutional Networks (GCN) presented in Chapter 3, along with the same configuration as the MB section.

The Complex-valued Enhancement Block (ComEB) is responsible for enhancing the speech signal by compensating for the lost details in the magnitude domain. We adopted a similar structure as the GCN in Chapter 3, but with some modifications to handle complex domain features. Specifically, we replaced the attention mechanism in the GRU with dilation convolution to effectively capture the long-range dependencies between adjacent frames.

The ComEB's primary objective is to recover the lost information from the coarse features by compensating for the complex-valued features. This module consists of a series of layers that include dilation convolution, normalization, and activation functions. We chose the dilation convolution to improve the ComEB's performance by enabling it to capture the long-term temporal dependencies between the speech signal's frames. The normalization and activation functions are added to enhance the ComEB's stability and learning capacity.

Overall, the proposed ComEB provides an effective means of enhancing the speech signal in the complex domain by compensating for lost details from the coarse features, which significantly improves the quality of the reconstructed signal. The ComEB's architecture was designed to handle the complexities inherent in speech signals, making it a valuable tool for various applications in speech processing.

5.3.3 Compensation Block (CB)

Let M and $X_{com} = \{X_r, X_i\}$ denote the output of MB and ComEB, respectively. Both of them will be fed into CB together with original complex-valued input. As the complement, the RI component as the input and some significant information may be lost by propagation. To update the RI components of the whole model in a collaborative manner, we propose the compensation block acting an important role in this project, shown in Fig 5.4.

To be specific, the input feature is RI spectrum $\{R_{n-1}, I_{n-1}\}$ firstly decoupled into magnitude spectrum Mag_{n-1}, given by:

$$Mag_{n-1} = \sqrt{|R_{n-1}|^2 + |I_{n-1}|^2} \tag{5.5}$$

$$\theta_{n-1} = arctan2(R_{n-1}, I_{n-1}) \tag{5.6}$$

Motivated by the spectrogram masks that can effectively and coarsely suppress noise, the Mag will be multiplied by its corresponding mask from MB. However, by focusing solely on the magnitude feature and neglecting the phase information, there's a potential for speech mismatches. Besides, values in masks range from 0 to 1 for training stability, both the residual noise and speech distortion will happen accordingly [180]. So, in this case, we design the CB to focus on and compensate the lost detail from the complex domain perspective. The whole procedure is given by equation 5.7:

$$Smag_n = Mag_{n-1} * Mask$$

$$Comp_r = Smag_n * \cos(\theta_{n-1})$$

$$Comp_i = Smag_n * \sin(\theta_{n-1})$$

$$\widehat{R\_n} = R_n + Comp_r$$

$$\widehat{I_n} = I_{n-1} + Comp_i \tag{5.7}$$

Where $*$ means element-wise multiplication operation.

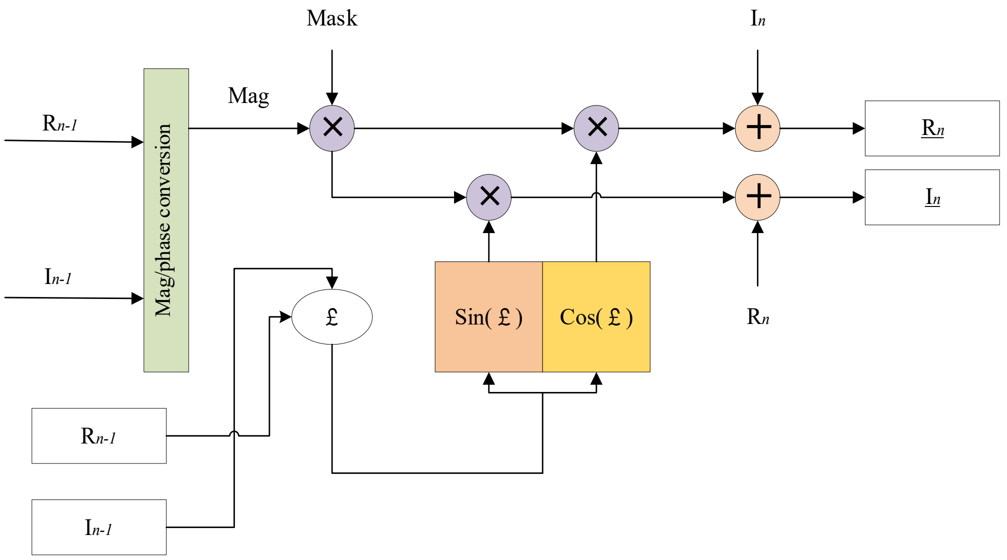


Figure 5.4 The architecture of proposed CB.

5.4 Experiment setup

The experimental data settings for this study were similar to those used in chapter 4. In the FEB, we set the kernel size, channel, and stride of GLU to 3, 256, and 1, respectively. The U-block had a kernel size and stride of (1, 3) and (1, 2) in the time and frequency axes, with a channel of 256. We employed a total of 5 (en) decoder layers, with 2 U-blocks included in the architecture.

For the MB, the kernel size, channel, and stride in (en) decoder layers were set to 8, 256, and (1, 3) in the time and frequency axes. The left 1D-conv kernel size and the right 1D-conv kernel size in the GRU were 5 and 1, respectively, with a channel of 256. In MHSA, we set the number of heads to 2, and there were a total of 5 (en) decoder layers with 10 GRUs divided into 2 GRU groups.

In the ComEB, we used a similar structure as the GCN in Chapter 3, but with the GRU not using an attention mechanism and instead using dilation convolution. The kernel size, channel, and stride in (en) decoder layers were set to 8, 512, and (1, 2) in the time and frequency axes. The GRU in ComEB had the same kernel size, stride, and channel settings as the one used in MB. The middle part of ComEB comprised 4 GRU groups, each with 5 GRUs, with dilation rates set to [1, 2, 4, 8, 16]. The loss function and dataset used were the same as those in the GCN experiments.

For the dataset setup, we utilized clean speech segments from the ICASSP DNS3 dataset [167], which we then combined with noise sources from NoiseX-92 [145] for training dataset (about 100 hours). While the testing dataset was composed of clean audio extracts from the 5-hour Librispeech corpus [147] mixed with NoiseX-92 noises.

For our training set, we cut the utterances into 2-second segments. But for the test set, we didn't make any cuts, so the lengths vary. We used a Hamming window with a 25 ms window length (equivalent to 400-point FFT) and a hop size of 200 points, which means there's a 50% overlap. The input feature of the proposed model is the magnitude spectrum.

All audio data was sampled at 16kHz and extracted by using frames of 512 with frame shift 256, and the model is directly fed by raw waveform. All models are optimized using the Adam-algorithm [166] with a learning rate of 0.001, decaying by half after each epoch.

To evaluate the quality of the denoised speech, we picked a range of standard metrics. We used PESQ, which has a score range from -0.5 to 4.5. For judging how clear the speech sounds, we used ESTOI, which scores between 0 and 1. For all these metrics, a higher score means better speech quality.

In this experiment, the training process was done on High Performance Computation (HPC) node with two V100 16GB GPUs. This training process spanned roughly 40 hours for 10 epochs. The computational resources were generously offered by the High-Performance Computation (HPC) at the University of Southampton.

5.5 Results and analysis

Table 5.1 Results of all tested models show PESQ, STOI values for SNRs between -3 and 6 dB. ‘noisy’ is the original noisy speech (Babble). ‘Proposed model’ is our model. Higher is better.

Metrics	PESQ					ESTOI(%)				
Test SNR (dB)	-3	0	3	6	Avg.	-3	0	3	6	Avg.
Noisy	1.61	1.77	1.99	2.19	1.89	31.59	40.23	49.61	58.64	45.02
GCRN	2.32	2.62	2.87	3.08	2.72	59.57	68.83	75.72	80.78	71.23
DCCRN	2.31	2.61	2.88	3.11	2.72	59.57	68.11	75.86	81.56	71
PHASEN	2.36	2.7	2.99	3.21	2.82	61.78	71.25	78.32	83.31	73.67
AECNN	2.32	2.64	2.91	3.11	2.74	62.13	71.57	78.25	83.03	73.74
ConvTasNet	2.26	2.52	2.76	2.96	2.63	63.88	72.21	78.49	83.22	74.45
Proposed model	2.34	2.56	2.99	3.28	2.79	64.43	75.24	80.63	84.64	76.29

The results of all experiments are presented in Table 5.1. The table displays the results of all baseline models, including our proposed model, which outperforms all others in both average PESQ and ESTOI. The improvements can be attributed to the different network architectures. Compared to DCCRN, our proposed model demonstrates an average improvement of 2.6% in PESQ and 7.5% in ESTOI. Compared to ConvTasNet, our proposed model delivers an average improvement of 6.1% and 2.5% in PESQ and ESTOI, respectively.

In the preceding sections, we discussed the complexity of noise spectrum estimation and the innovative approaches aiming to refine speech enhancement strategies by integrating explicit noise information. Our experiments, summarized in Table 5.1, were designed to empirically validate the efficacy of these novel methodologies in real-world conditions, especially under challenging low SNR scenarios.

The results underscore our proposed model's good performance in enhancing speech quality, as evidenced by leading scores in both average PESQ and ESTOI across various SNR levels.

Table 5.2 Results of all versions of GCNs show PESQ, STOI values for SNRs at -3 and 6 dB. ‘GCN-3rd’ is the original version in Chapter 3. ‘GCN-4th’ is the version adopting the attention mechanism in Chapter 4. ‘GCN-5th’ is the version combing magnitude domain with complex-valued domain in Chapter 5. Higher is better.

Evaluation	ESTOI(%)	PESQ	ESTOI(%)	PESQ
Noise(babble)	0 dB		6 dB	
GCN-3rd	71.33	2.61	82.64	2.88
GCN-4th	**76.53**	3.01	**85.31**	3.11
GCN-5th	75.24	**3.06**	84.64	**3.28**

In this section, we present a comparison of the objective results of various proposed models based on the GCN architecture. The results are presented in Table 5.2, where the PESQ and STOI values for each model are reported at two different SNRs, namely 0 dB and 6 dB.

It is worth noting that at 0 dB, the GCN-4th model achieves the best performance in terms of both PESQ and STOI. This can be attributed to the fact that the attention mechanism used in this model has shown to have an outstanding ability to capture global context, even though dilated convolution layers can effectively broaden the receptive field.

At 6 dB, the GCN-5th model outperforms the other models in terms of PESQ, while also delivering a good STOI result close to the best one. This can be explained by the fact that the complex-valued domain used in this model is able to compensate for the lack of phase information, resulting in improved speech quality. Additionally, the magnitude mask used in the model effectively removes background noise.

The parallel paths of complex-valued domain and magnitude mask in the GCN-5th model work in tandem with each other to enhance the overall performance of the model. In our opinion, this kind of mechanism will gain popularity in the future due to its effectiveness in improving speech quality.

In light of the findings, it's evident that the integration of attention mechanisms and the complex-valued domain significantly bolsters the efficacy of speech enhancement models grounded in the GCN architecture. Specifically, the Transformer mechanism, which has shown remarkable success in various domains, presents a potential adaptation for future iterations of our speech enhancement task [95]. By capitalizing on its self-attention capabilities, we desire to achieve even greater precision and quality in audio enhancements. Continuing research will be directed towards these advanced mechanisms to further refine and augment the quality of speech enhancement.

5.6 Conclusions of this chapter

Our model's advantage is derived from its innovative architecture, which combines the Compensation for Complex Domain Network (CCDN) with the magnitude-based (MB) and complex-based enhancement blocks (ComEB). This combination allows for a more effective handling of both the magnitude and complex spectral components, preserving speech integrity while effectively mitigating noise. Specifically, the model leverages the magnitude information refined by MB to compensate for the intricate details lost in the complex domain during the enhancement process, a strategy inspired by the dual challenges of noise spectrum estimation and speech signal preservation outlined earlier.

Furthermore, the implementation of a multi-head self-attention mechanism across both time and frequency dimensions facilitates a more granular understanding and processing of the speech signal. This is a significant departure from conventional monaural speech enhancement techniques, allowing for a more dynamic response to the variance in noise types and speech signals encountered in real-world settings.

Comparative analysis with established models like DCCRN and ConvTasNet highlights our model's novel contributions. The integration of a GRU layer and an attention mechanism not only surpasses traditional approaches in performance metrics but also showcases our model's ability to capture and utilize global and contextual information more effectively. This particularly leverage the ability to process long-range speech sequences under low SNR conditions.

In summary, the experimental outcomes not only effectively affirm the effectiveness of our approach in enhancing speech quality in noisy environments, but also validate our proposed model's capacity to tackle the challenges identified in the introduction: accurate noise spectrum estimation and the preservation of speech quality under diverse noise situations.

5.7 Future work

Historically, to train advanced speech enhancement algorithms, researchers often simulated paired datasets. These are typically derived from unadulterated speech samples juxtaposed with a curated assortment of noise profiles. Undoubtedly, these datasets serve as invaluable assets, significantly simplifying the training process and enabling deep learning models to achieve remarkable milestones with each passing year in terms of performance metrics.

However, as with many complex systems, there are inherent limitations. One of the most pronounced challenges is the models' diminished adaptability when transposed from controlled environments to the unpredictable real-world scenarios. Their performance will be not robust and well-satisfying when confronted with background noises that were not part of their training regimen — a testament to their lack of generalization to unfamiliar background noises.

In light of the ongoing advancements in speech enhancement research, there is a growing emphasis on creating models that not only perform well in controlled environments but also effectively handle diverse real-world noise conditions. Rather than solely concentrating on algorithmic perfection, it is essential to ensure that these models are versatile and robust when exposed to various noise scenarios. Addressing these challenges is not just a matter of academic interest but is crucial for developing practical solutions that align with real-world needs.

Chapter 6 Finishing Speech enhancement model for unpaired data

6.1 Background

Deep neural networks (DNNs) have made significant progress in speech enhancement through deep learning. One effective approach is to use a DNN-based masking estimation method with input features extracted from noisy speech. This technique converts the speech enhancement problem into a classification problem, where a well-trained mapping function reduces the loss between the features of improved speech and clean speech. During training, clean speech and enhanced speech are paired so that a supervised learning system can be implemented. However, the diversity of the training data and the increased model complexity have a significant impact on the learning outcomes of supervised speech enhancement (SE) models. It is worth noting that machine learning models can become biased towards over-represented social groupings since publicly accessible datasets do not reflect all populations [184]. For instance, if a given target speaker's distinctive vocal features or loud surroundings are never encountered during training, a general-purpose universal SE model may underperform for that speaker. This type of matched data is typically not achievable since noise characteristics and energy change over time and under different conditions, making it challenging for the changing noise to match the speech signal. As a result, the frequency domain energy distribution of speech can easily become mismatched with unpaired data, which reduces the DNN's ability to generalize.

In the realm of speech enhancement, the supervised learning paradigm typically relies on a substantial amount of paired noisy/clean speech data. However, collecting such paired data in real-world recording situations can be challenging, which is why simulated paired data is often used. This involves simulating room impulse responses and adding additive noise to clean speech. However, a mismatch between the simulated data and real-world data, such as noise type, can lead to inconsistent performance when the system is used in practical applications [185]. If this mismatch problem is not addressed, supervised speech enhancement techniques run the risk of failure.

Recently, unsupervised neural speech enhancement techniques [185][186][187] have garnered attention since they can mitigate the drawbacks and requirements of paired data in supervised learning to a significant extent. Unsupervised algorithms only have access to noisy speech as inputs and do not rely on clean speech. Since it is practical to gather noisy speech from real-world applications, recordings of noisy speech from real-world scenarios are used to train unsupervised models. This approach prevents the mismatch issue caused by simulated data. However, unsupervised algorithms are still unable to perform as well as supervised ones. While unsupervised algorithms have made progress in the area of speech enhancement, they often lag behind their supervised counterparts in performance. For example, according to [188], supervised speech enhancement methods that leverage deep neural networks (DNNs) have shown superior performance over traditional unsupervised methods, especially in terms of objective metrics like Signal-to-Noise Ratio (SNR). Furthermore, [190] emphasized the robustness of supervised methods against various noise conditions.

However, the advantage of unsupervised methods lies in their ability to operate without labeled data. This is particularly beneficial when dealing with large amounts of unlabeled audio data or in real-world scenarios where obtaining labeled data is cumbersome and costly. [191] pointed out the potential of unsupervised methods, especially when combined with modern techniques like autoencoders and generative adversarial networks (GANs).

It's worth noting that the performance gap between supervised and unsupervised algorithms is a subject of ongoing research. Efforts are being made to improve the efficacy of unsupervised algorithms, and as [192] suggest, future advancements in self-training and semi-supervised techniques might play a pivotal role in this endeavor.

To address the mismatch issue in supervised methods and performance degradation in unsupervised methods, semi-supervised approaches are being further researched. Examples include the DNN-NMF hybrid model [189] and the multi-modality based method [193]. However, the iteration technique in the DNN-NMF hybrid model may not be suitable for real-time processing in many real-world applications.

Generative Adversarial Networks (GANs) can now be used with unpaired training data to produce the necessary output from the distribution of real data via adversarial training. GANs [13] [194]have at least outperformed DNNs in supervised or paired data systems. For instance, Santiago Pascual et al. [13] used GANs for supervised speech augmentation using paired data initially. However, in many real-world situations, it is difficult or even impossible to record clean-noisy pairings simultaneously, and clean data that does not match the noisy source data may be the only available option. To address this challenge, the CycleGAN[121], shown in Figure 6.1, was considered for unpaired training data in [195] since obtaining paired training data can be a challenging and costly operation. The CycleGAN operates on the principle of simultaneously learning forward and backward mappings using adversarial loss [196] and cycle-consistency loss [197]. The adversarial loss is used to identify the generated output or real input, while the cycle-consistency loss is used to constrain the input information. These two losses are combined in the final cost function.

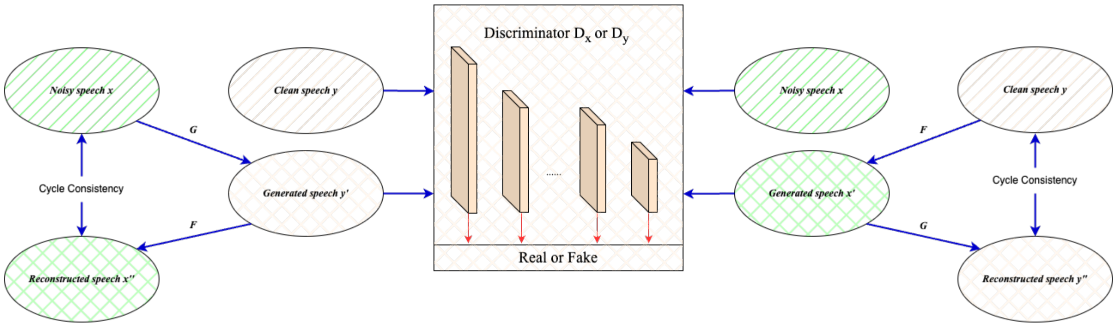


Figure 6.1 Training procedure of the proposed method. Forward noisy-clean-noisy cycle and backward clean-noisy-clean cycle are illustrated in the left and right parts, respectively.

In order to deal with non-parallel speech improvement, a novel CycleGAN-based system is developed by Yuan&Bao [121]. The paper introduces a novel speech enhancement method based on CycleGAN that is designed to work with unpaired training data. The primary conclusions drawn from the research are:

Unpaired Training Data Feasibility: The proposed method demonstrates the feasibility and effectiveness of using unpaired training data for speech enhancement tasks. Traditionally, paired clean and noisy samples are required for training, but this method circumvents that requirement.

CycleGAN Architecture: The CycleGAN architecture's inherent design, which comprises two generator networks and two discriminator networks, proves to be well-suited for the task. The generators are responsible for mapping between clean and noisy domains, while the discriminators ensure that the generated samples are indistinguishable from real samples.

Objective and Subjective Evaluations: Both objective measurements (like PESQ) and subjective listening tests indicate that the proposed CycleGAN-based method achieves competitive performance when compared to traditional methods, even without paired training data.

Generalization to Other Noises: The model demonstrates its adaptability and generalization capabilities by successfully enhancing speech corrupted by various noise types, not just those present in the training data.

Potential for Further Improvement: While the proposed model offers a promising alternative to traditional paired data methods, there is an acknowledgment of its limitations and an indication that future work can further improve upon its results, especially in scenarios with diverse and unknown noise conditions.

The CycleGAN differs from conventional GANs and it aims to learn two distinct transformations. Specifically, the approach learns the transformation from noisy speech to clean speech (denoted as $G_{X\rightarrow Y}$), as well as the transformation from noisy speech to noise (denoted as $F_{Y\rightarrow X}$). To achieve this, non-parallel speech samples X and Y are utilized, where X denotes the noisy speech domain and Y denotes the clean speech domain. The discriminator D_X is used to distinguish between the generated clean speech and the corresponding real clean speech samples, while D_Y is used to distinguish between the generated noise and the corresponding real noise samples. The network is optimized using a combination of cycle consistency loss, adversarial loss, and identity-mapping loss.

A speech enhancement approach was proposed based on the Noise2Noise method [199] for image restoration, using noisy and noise-only samples ($\widehat{x_i}_{i=1}^{N}$and $\overline{x_i}_{i=1}^{N}$) as inputs [200]. These paired noisy samples ($\widehat{x_i} = y_i + n_i$ and $\overline{x_i} = y_i + n_j$) are created by adding different noise files (n_i and n_j) to the same clean speech signal (y_i). However, collecting such paired noisy samples from real-world recording environments is often infeasible. Therefore, in practice, simulated noisy samples are used, which can lead to a mismatch between the training and real-world data. To address this issue, our proposed approach utilizes only in-the-wild noisy speech as input, avoiding the need for paired noisy samples and eliminating the potential mismatch between the training and inference data.

MixIT [198] utilizes a similar approach as Noise2Noise by mixing two noisy speech recordings or noisy speech and noise. However, MixIT produces multiple outputs, each containing a single source, either clean speech or noise. These outputs are combined to re-synthesize the noisy speech, and the loss between the re-synthesized noisy speech and the original noisy input is calculated over different permutations. The loss with the best permutation is then used to optimize the network. In contrast, our proposed approach does not mix additional noisy speech or noise into the in-the-wild noisy speech input, eliminating the input mismatch problem between training and inference. Since our system only has a noise output and a clean speech output, there is no permutation problem when re-mixing the outputs for loss calculation. Unlike MixIT, which requires finding the best permutation for the remixed noisy speech and corresponding reference, our approach avoids this extra step.

Diverging from the traditional semi-supervised approaches and motivated by [201], this study introduces an innovative unsupervised pre-training and fine-tuning algorithm. Its design specifically addresses the challenges of data mismatch and performance degradation observed in both supervised and unsupervised techniques. Central to our method is a distinct Generative Adversarial Network (GAN) architecture. Within this GAN, the generator is designed to simultaneously process both magnitude and complex-domain features. Notably, the discriminator plays an instrumental role in enhancing certain evaluation metrics without compromising others.

The initial training phase leverages significant amounts of unpaired noisy and clean speech, employing the unsupervised pre-training strategy. In a departure from conventional unsupervised practices and drawing insights from the contrastive method in computer vision [204] and the Noise2Noise paradigm [205], we introduce supplementary random noise to the original noisy speech (ONS), resulting in what we term 'deeper noisy speech' (DNS). Both the original and the DNS variants are then utilized to train the generator, ensuring their outputs are distinctly recognized. Additionally, the GAN's holistic training incorporates the unpaired noisy speech, DNS, and clean speech samples. Upon completion of this foundational training, the model undergoes fine-tuning using the original noisy dataset, supplemented by a select set of simulated paired samples. A pivotal advantage of our methodology is its direct reliance on real-world data inputs, circumventing the pitfalls associated with simulated data in traditional supervised approaches.

The main contributions of this study are summarized as follows:

1. The proposed methodology uniquely leverages unpaired noisy and clean speech as inputs, effectively circumventing the data mismatch challenges pervasive in traditional supervised speech enhancement techniques, especially when applied to real-world environments.
2. Notably, despite the absence of paired noisy and clean speech, our approach ingeniously emulates supervised training paradigms. This is achieved by introducing additional variant noises to the noisy speech samples during the pre-training phase.
3. We design a novel architecture in the proposed approach to leverage the unpaired noisy and clean speech.
4. Introducing more feasible loss function to guarantee the model concentrating on target speaker speeches characteristics.

6.2 Methodology

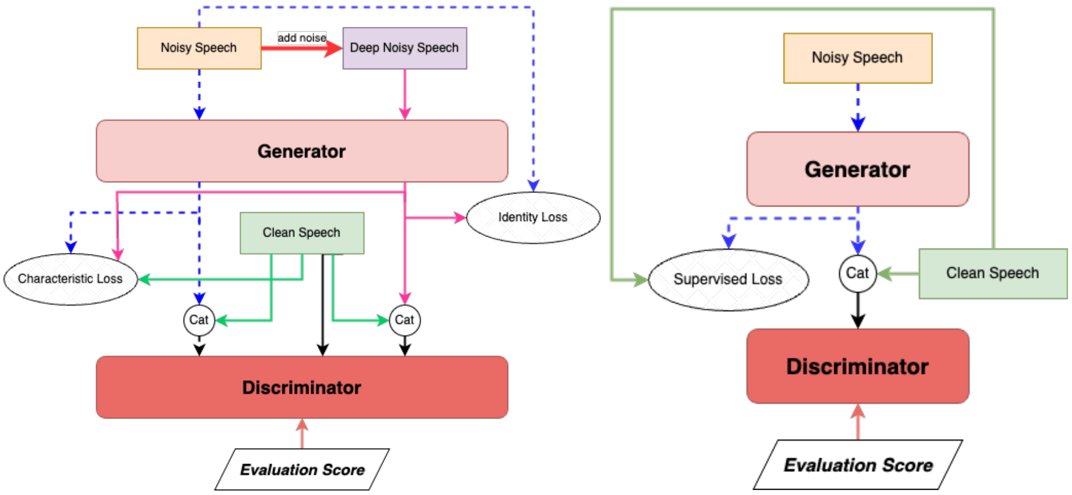


Figure 6.2 The proposed architecture of the proposed speech enhancement model consisting of pre-training step (left) and fine-tuning step (right). In the right figure, Noisy Speech and Deep Noisy Speech are paired, while Noisy Speech and Clean Speech are unpaired. The blue, purple, green lines represent the Noisy Speech flow, Deep Noisy Speech flow and Clean Speech flow, respectively.

6.2.1 Overview

As depicted in Fig. 6.2, our proposed speech enhancement methodology harnesses the capabilities of the Generative Adversarial Network (GAN) architecture, comprising a generator and a discriminator. The generator's primary function is to transform noisy speech into its enhanced version. Concurrently, the discriminator is tasked with providing evaluative scores based on perceptual criteria. The training regimen for our model is bifurcated into two distinct phases: unsupervised pre-training and subsequent fine-tuning. During the initial phase, the speech enhancement model is rigorously trained using a substantial corpus of unpaired noisy and clean speech, leveraging both identity and characteristic loss functions. In the fine-tuning phase, a select set of paired noisy and clean speech data is utilized to refine and optimize the overarching model. A comprehensive exposition of our methodology is delineated in the subsequent sections.

6.2.2 Structures of generator and discriminator

An overview of the generator architecture of the proposed model is shown in Fig. 6.3. For a noisy speech waveform, an STFT operation first converts the waveform into a complex spectrogram $X_{complex} \in R^{T*F*2}$ and corresponding magnitude spectrogram $X_{mag} \in R^{T*F}$, where T and F denote the time and frequency dimensions, respectively. The real and imaginary parts X_{real} and X_{imag} are then concatenated with the magnitude X_{mag} as an input to the generator. The generator takes the encoder-decoder as a backbone.

(a) ***Encoder***: Given the input feature $X \in R^{B*T*F*3}$, where B represents the batch size, the encoder is architecturally structured to encompass two convolutional blocks with an intervening dilated Res2Net [203]. Each of these blocks integrates a convolution layer, followed by instance normalization [119], culminating with a PReLU activation function [202]. The first convolution block functions to expand the triad of input features into an intermediary feature map. The prowess of Res2Net, previously demonstrated in domains like speaker verification and computer vision, lies in its capacity to effectively amalgamate antecedent feature maps. This amalgamation facilitates the extraction of diverse feature gradations, simultaneously augmenting the receptive field without necessitating an increase in kernel or layer quantities. The terminal convolution block is strategically designed to halve the frequency dimension, thus optimizing computational efficiency.

(b) ***Middle Layer***: Attention mechanism [95] has achieved great success in many fields, such as speech recognition and Natural language Processing as they can capture long distance dependencies. We create a resolution_former block containing 2 feed forward neural networks (FFNN). Like transformers in [95], we add a multi-head attention block followed by Layer Normalization layer between 2 FFNNs. Here we employ two resolution_former blocks sequentially to capture the time dependency in the first stage and the frequency dependency in the second stage. After the residual connection, the output will be reshaped as the original shape.

(c) ***Decoder***: The decoding mechanism derives its output from N resolution_former blocks in a distinctly decoupled manner, divided into two pathways: the mask decoder and the complex decoder. The mask decoder is designed with the primary objective of generating a mask. This mask, when subjected to element-wise multiplication with the input magnitude X_{mag}, results in the prediction of X'_{mag}. In contrast, the complex decoder directly predicts both the real and imaginary components. Both these decoders incorporate a Res2Net Block, echoing the design paradigm seen in the encoder. To revert the frequency dimension to its original input size, a subpixel convolution layer finds its application in both pathways [207]. Within the mask decoder, a convolutional block narrows the channel count to one, succeeded by another convolution layer followed with a PReLU activation, culminating in the final mask prediction. It's noteworthy that the PReLU activation is adaptive, discerning different slopes for individual frequency bands. The complex decoder's architectural is the same as that of the mask decoder, with the notable exception being the omission of an activation function for the complex output.

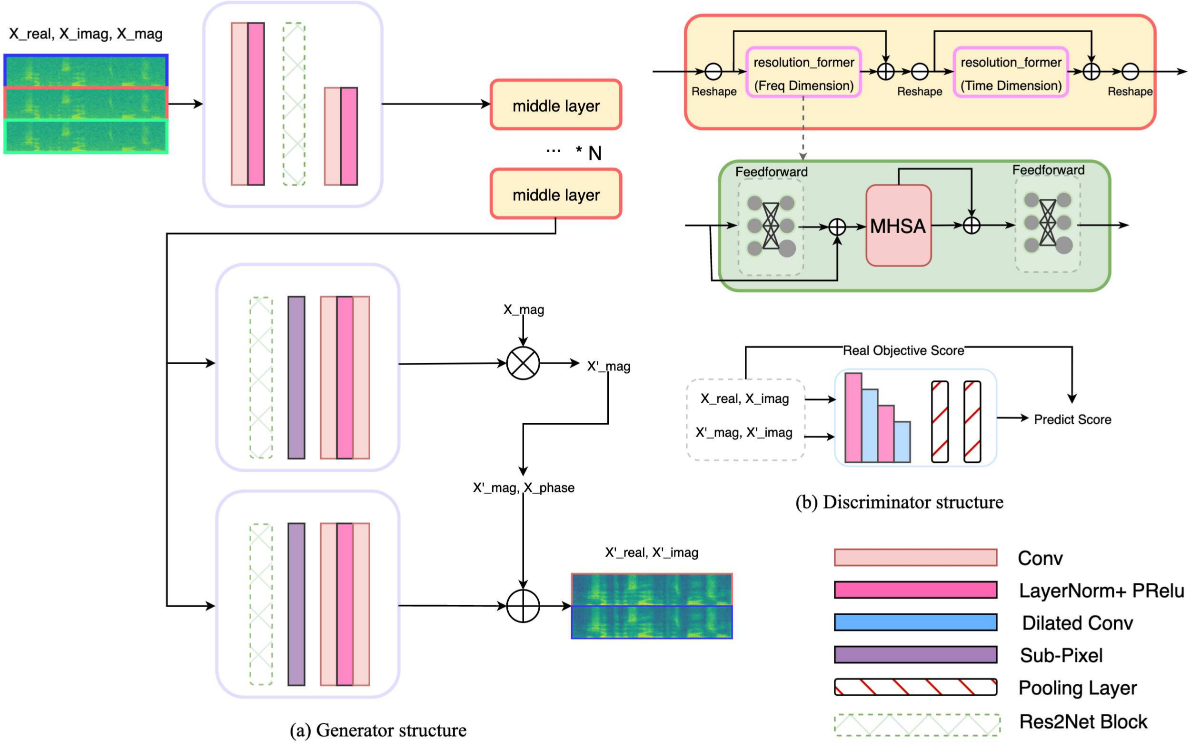
Same as in [208], the masked magnitude X'_{mag} is first combined with the noisy phase to obtain the magnitude- enhanced complex spectrogram. Then it is element-wise summed with the output of the complex decoder $\hat{X}_{real}, \hat{X}_{imag}$ to obtain the final complex spectrogram:

$$X'_{real} = X'_{mag}cos(phase) + \hat{X}_{real},\ X'_{imag} = X'_{mag}sin(phase) + \hat{X}_{imag} \tag{6.1}$$

that, an inverse short-time Fourier trans- form (ISTFT) is applied to get the audio signal. To further improve the magnitude component and propagate magnitude loss on both decoder branches, we compute the magnitude loss on X''_{mag} expressed by:

$$X''_{mag} = \sqrt{X'_{real} + X'_{imag}} \tag{6.2}$$

In Speech Enhancement (SE), the goals we set (objective functions) often don't line up neatly with the ways we measure success (evaluation metrics). This means that even if we do really well on our set goals, the actual quality, as measured, might not be up to the mark. Some common quality measures, like the perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI), are tricky because they can't be directly used to guide the learning process; they're non-differentiable. To tackle this, our model's discriminator is designed to act like one of these quality measures and be part of the learning process. As illustrated in Fig. 2(b), we've borrowed from the MetricGAN approach, using the PESQ&STOI score as a kind of label [209]. The discriminator's job is to try and guess the best PESQ&STOI scores (ideally [1, 1]) when it's just given clean sounds. But when given both clean and processed sounds, it tries to guess the improved PESQ&STOI scores based on the labels it has. Meanwhile, the generator is working to produce enhanced speech that sounds as close to the clean speech as possible, aiming for that ideal PESQ&STOI score of [1, 1].



(a) Generator structure

(b) Discriminator structure

Figure 6.3 The overview of the proposed GAN model.

6.2.3 Pre-training phase

The proposed unsupervised pre-training adopts unpaired noisy speech x and clean speech y as training data. Firstly, we add random noise to noisy speech at a random continuous SNR value ranging from -5 dB to 10 dB, so as to get deep noisy speech X. x and X are respectively fed into generator outputting their own enhanced speeches y_x and Y_x. To optimize this enhancement network, the discriminator is also included to calculate adversarial loss. In addition, a character loss and an identity loss are also explored in this work.

(A) ***Identity loss***: The identity loss function in [121][201] consists of the original input noisy speeches as well as the sum of enhanced speeches and enhanced noise. Different from them, the proposed identity loss function is comprised of enhanced speeches from deeper noisy speeches (DNS) and original noisy speeches (ONS), which is defined as:

$$L_{id} = E_{x\sim P_x}[\|G(X_{DNS}) - X_{ONS}\|_2] \tag{6.3}$$

(B) ***Characteristic loss:*** The characteristic loss function is mainly responsible for enabling enhanced speeches approaching to the real-world human being speech character. Its advantage is to avoid the speech content mismatch between unpaired data by comparing their mel_spectrogram difference rather than directly computing the speech difference, such as using mean squared error (MSE). The characteristic loss function can be defined as:

$$L_{char} = E_{x\sim P_x, \mathrm{Y}\sim \mathrm{P_Y}}\left[\left\|Mel\big(G(X_{DNS})\big) - Mel(Y)\right\|_2 + \left\|Mel\big(G(X_{ONS})\big) - Mel(Y)\right\|_2\right] \tag{6.4}$$

Where $Mel(X)$ denotes the operation converting audio to mel_spectrogram. X and Y represent input noisy speeches as well as pure speeches and they are unpaired.

(C) ***Adversarial loss***: we use a linear combination of magnitude loss L_{mag} and complex loss L_{comp} in TF-domain:

$$L_{TF} = \partial L_{mag} + (1-\partial)L_{comp} \tag{6.5}$$

$$L_{mag} = E_{x\sim P_x}\left[\left\|\hat{X}^{DNS}_{mag} - X^{ONS}_{mag}\right\|_2\right] \tag{6.6}$$

$$L_{comp} = E_{x\sim P_x}\left[\left\|\hat{X}^{DNS}_{real} - X^{ONS}_{real}\right\|_2 + \left\|\hat{X}^{DNS}_{imag} - X^{ONS}_{imag}\right\|_2\right] \tag{6.7}$$

Where ∂ is weighting factor, which is set to 0.6 in this experiment. Meanwhile, $\hat{X}^{DNS}_{mag}$, $\hat{X}^{DNS}_{real}$, $\hat{X}^{DNS}_{imag}$ denote DNS (ONS) magnitude, complex domain output of generator fed with their corresponding ONS magnitude, complex domain (X^{ONS}_{mag}, X^{ONS}_{real}, X^{ONS}_{imag}). Similar to least-square GANs [121], the adversarial training is following a minimal optimization task over the discriminator loss L_D and the corresponding generator loss L_{GAN} expressed as follows:

$$L_{GAN} = E_{Y\sim P_Y}\left[\left\|D\big(Y_{mag}, \hat{X}_{mag}\big) - 1\right\|_2\right] \tag{6.8}$$

$$L_{Disc} = E_{x\sim P_x, Y\sim P_Y}\left[\left\|D(Y_{mag}, Y_{mag}) - 1\right\|_2 + \left\|D(Y_{mag}, \hat{X}^{ONS}_{mag}) - Score_{PESQ\&STOI}\right\|_2 + \left\|D(Y_{mag}, \hat{X}^{DNS}_{mag}) - Score_{PESQ\&STOI}\right\|_2\right] \tag{6.9}$$

Where D refers to the discriminator, $Score_{PESQ\&STOI}$ refers to the normalized PESQ&STOI score, ranging from 0 to 1, between unpaired clean speeches and enhanced speeches. Besides that, we also add a time loss L_{time} as another penalization to guarantee the resynthesized speech quality:

$$L_{time} = \mathrm{E}_{\mathrm{x}\sim \mathrm{P_x}}[\|\widehat{wav}_{DNS} - wav_{ONS}\|_2] \tag{6.10}$$

Where $\widehat{wav}_{DNS}$ and wav_{ONS} represent enhanced waveform from DNS and ONS waveform. So the final generator loss function is expressed as follows:

$$L_{gen} = \alpha * L_{TF} + \beta * L_{GAN} + \gamma * L_{time} \tag{6.11}$$

Where α, β, γ are weight factors of their corresponding loss functions and set to 0.4, 0.5 and 0.1 in this experiment.

6.2.4 Fine-tuning step

Since the performance of pre-training is unsatisfied, the enhancement network is always fine-tuned with simulated paired noisy and clean speech by supervised learning to reduce the mismatch between the simulated data and the unpaired data. The same as [201], we take a small amount of simulated paired data for the fine-tuning step. The simulated paired data is used to optimize the generator hyperparameters from the noisy to clean speech by supervised learning. With the fine-tuning training, the capability of the enhancement network learned from the unsupervised pre-training stage is further strengthened. The loss function for the fine-tuning step is defined as follows:

$$\begin{aligned} L_{gen} = E_{XY\sim P_{XY}} \Big[& \alpha * \partial * \left\|\hat{X}_{mag} - Y_{mag}\right\|_2 + \alpha * (1-\partial)\left\|\hat{X}_{real} - Y_{real}\right\|_2 \\ & + \alpha * (1-\partial)\left\|\hat{X}_{imag} - Y_{imag}\right\|_2 + \beta * \left\|D(\hat{X}_{mag}, Y_{mag}) - 1\right\|_2 \\ & + \gamma * \|\widehat{wav}_X - wav_Y\|_2 \Big] \end{aligned} \tag{6.12}$$

$$L_{Disc} = \mathrm{E}_{\mathrm{XY}\sim\mathrm{P}_{\mathrm{XY}}}\left[\left\|\mathrm{D}(Y_{mag}, Y_{mag}) - 1\right\|_2 + \left\|\mathrm{D}(Y_{mag}, \hat{X}_{mag}) - Score_{PESQ\&STOI}\right\|_2\right] \tag{6.13}$$

Where X and Y are paired noisy speeches and clean speeches.

6.3 Objective experiment

6.3.1 Datasets

In our experimental framework, we constructed synthetic datasets following the methodology of previous research [210]. During the pretraining phase, we utilized clean speech segments from the ICASSP DNS3 dataset [167], which we then combined with noise sources from NoiseX-92 [145] and Musan-2 [206]. These mixtures were created at varying sound levels, ranging from -5 to 10 dB. For the fine-tuning stage, we curated a bespoke paired dataset, named FT-SMALL. This dataset was composed of clean audio extracts from the 5-hour Librispeech corpus [147] mixed with noise from Musan-3 [206]. The deliberate variation in noise types between the paired and unpaired datasets supports our hypothesis: simulated paired data won't perfectly mimic real-world unpaired samples. Our evaluation was conducted using a 2 hours' long test dataset that included clean speeches from eight unique speakers, each blended with noises from Musan-1. Notably, these speakers were distinct from those featured in both the FT-SMALL and the pre-training dataset. For both Musan-1, Musan-2 and Musan-3, we used an equal division of the complete Musan dataset [206].

For our training dataset, we cut the utterances into 2-second segments. But for the test dataset, we didn't make any cuts, so the lengths vary.

6.3.2 Implement details

We used a Hamming window with a 25 ms window length (equivalent to 400-point FFT) and a hop size of 200 points, which means there's a 50% overlap. In the generator, we set the number of resolution_former blocks, N, to 2 and the channel number, C, to 64. When training, we used the AdamW optimizer [146] for both the generator and the discriminator and trained them for 10 rounds or epochs. The learning speed, or rate, was set at 5×10−4 for the generator and 1×10−3 for the discriminator. We also adjusted the learning rate as we went, reducing it by half every 2 epochs.

In this experiment, to carry out the training process, two V100 16GB GPUs were used, and the training process took approximately 120 hours for 10 epochs. The computation resources were generously offered by the High-Performance Computation (HPC) at the University of Southampton.

6.3.3 Objective evaluation metrics

To evaluate the quality of the denoised speech, we picked a range of standard metrics. We used PESQ, which has a score range from -0.5 to 4.5. We also used a set of metrics: (1) prediction for signal distortion (CSIG); (2) background noise intrusiveness (CBAK); (3) overall speech quality (COVL) [211]. All these MOS-based scores range from 1 to 5. For judging how clear the speech sounds, we used STOI, which scores between 0 and 1. For all these metrics, a higher score means better speech quality.

6.3.4 Results

We compared our results (as shown in Table 6.1) with established methods such as CycleGAN [121], NeTT [212], NyTT [213], and M-4 [201], drawing data from the study [201]. The comparative findings can be seen in Table 1. Our method generally outperformed other state-of-the-art (SOTA) techniques, especially with unseen data. This suggests that the initial settings, achieved through our unique pre-training strategy, significantly enhance the performance of the speech enhancement model during the subsequent fine-tuning phase. Additionally, our approach expedited the model's convergence speed. For instance, a model with a modest 2.94 M parameters neared convergence in just 10 epochs without compromising its efficacy.

Table 6.1 Comparisons between the proposed method and other supervised fine-tuned models initialized by state-of-the-art unsupervised methods. Selected SNR in test data ranges from 0dB to 15 dB.

Method	Fine-tuning data	Test data	Evaluation metrics				
			PESQ	SDR	CSIG	CBAK	COVL
CycleGAN	FT-SMALL	TEST-MUSAN1	1.58	12.1	2.17	2.63	1.83
NETT			1.64	12.5	2.26	**2.71**	1.91
NyTT			1.59	12.1	2.13	2.63	1.81
M-4			1.52	12.4	2.32	2.61	1.87
Proposed model			**1.91**	**12.7**	**2.82**	2.68	**2.12**

To verify our design choices, we carried out an ablation study to check our design decisions whether brings a better performance comparing with the baseline model in same training and testing conditions, and the results are presented in Table 6.2. For our baseline, we used a model identical to our proposed one. However, instead of initializing it with a pre-training step, we trained it directly using fine-tuning data. Additionally, we created unpaired pretraining data using all speeches from Librispeech and noises from NoiseX-92. This data was used to set the starting weights for our proposed model with the pre-training strategy.

Table 6.2 Ablation experiment comparisons among the proposed method based on different training data. Selected SNR in test data ranges from 0dB to 15dB.

Method	Fine-tuning data		Test data	Evaluation metrics				
	Speech	Noise		PESQ	CSIG	CBAK	COVL	STOI
Baseline	train-clean-100	NoiseX-92	TEST-MUSAN1	*1.41*	*2.33*	*2.54*	*2.05*	*0.86*
Proposed model	train-clean-100	NoiseX-92		***1.64***	***2.44***	***2.59***	***2.17***	***0.91***
Baseline	train-clean-100, dev-clean	NoiseX-92+MUSAN2		1.89	2.89	**2.68**	2.37	0.89
Proposed model	train-clean-100, dev-clean	NoiseX-92+MUSAN2		**2.01**	**3.2**	2.62	**2.55**	**0.93**

Our analysis indicates that even with identical structures, the proposed model surpasses the baseline in performance. This improves the model's robustness and its ability to generalize effectively to unseen data, likely attributed to the weight initialization derived from the pre-training phase. Notably, as we augmented the volume of fine-tuning data, there was a marked improvement in the performance of the proposed model, particularly evident in the PESQ and STOI metrics. Upon comparing Table 1 with Table 2, it's striking to observe that our model, after fine-tuning with FT-SMALL data, exceeds the performance of the baseline model that utilized a larger fine-tuning dataset. This underscores the pivotal role that weight initialization through pre-training plays in the overall model efficacy.

6.4 Subjective experiment

6.4.1 Objective

The purpose of this experiment is to evaluate the performance of a speech enhancement system by assessing the discernibility of enhanced speech, noisy speech, and pure speech as perceived by voluntary participants from different groups. The groups are decided by 2 factors: (1) whether participants have hearing impairment or age over 60; (2) whether they are native English speakers.

6.4.1.1 Procedure

1. Recruitment of Participants: Volunteers willing to participate in the experiment are recruited, ensuring they are informed about the experiment's purpose, methodology, and their role.

2. Preparing the Experiment Setup: The participants are seated comfortably in a quiet, soundproof room. They are provided with headphones connected to the speech enhancement system.

3. Speech Playback: The speech enhancement system is programmed to play a series of speech samples randomly. These samples include enhanced speech (speech after processing by the enhancement system), noisy speech (speech with background noise), and pure speech (clean speech without any enhancement or noise).

4. Repetition Task: Participants are instructed to listen carefully to each speech sample and repeat exactly what they hear. Emphasis is on accurately repeating the speech content to the best of their abilities. At the same time, participants also require providing score ranging from 1 to 5 for the speech quality.

5. Speech Recording: The system will record the participant's repeated speech in real-time.

6. Evaluation: A speech recognition system (from "https://github.com/Uberi/speech_recognition") is utilized to transcribe and analyze the recorded speech. It determines the recognition rate, representing the accuracy and clarity of the participant's speech repetition.

7. Data Analysis: The recognition rates for each type of speech (enhanced, noisy, and pure) are then compared and statistically analyzed to determine the effectiveness of the speech enhancement system. The recognition rate serves as a proxy for the enhancement system's performance – the higher the recognition rate, the better the system's performance.

6.4.1.2 ASR-based Analysis in Subjective Audio Enhancement Experiments

1. Overview

In our experiment, we employed Automatic Speech Recognition (ASR) to evaluate the intelligibility of enhanced audio compared to the original, unprocessed audio. Volunteers were exposed to both versions of the audio clips, and their verbal repetitions of what they heard were recorded. These recorded samples were then processed through ASR to determine the Word Right Rate (WRR) as a measure of quality and intelligibility.

2. ASR Workflow

We utilized an open source ASR engine, which operates based on deep neural networks, to transcribe the audio clips. Each audio clip, both the reference (pure audio) and the volunteers' repeated versions, were passed through the ASR engine to produce a textual transcript. The produced transcripts were then used to calculate the WRR.

3. Criteria for Correct Identification

Determining what counts as a 'correct' identification by the ASR system required careful consideration. We defined a word as 'correctly' identified if:

(1) It exactly matched the word in the reference transcript.

(2) Minor variations, such as tense or plurality, were considered correct. For instance, 'run' and 'ran' were considered a match.

(3) Homologues or synonyms were not considered a match, given that they can significantly alter the meaning of a sentence.

4. Handling Ambiguities

In cases where the ASR engine recognized a similar-sounding word (homophones) or a plural form, we resorted to contextual analysis. If the recognized word fit the context of the sentence and did not alter its meaning, it was considered a correct identification.

5. Calculation of Word Right Rate (WRR)

The WRR was calculated as the ratio of correctly identified words to the total number of words in the reference transcript. Mathematically, it is represented as:

$$WRR = \frac{Number\ of\ correct\ words}{Total\ number\ of\ words\ in\ reference} * 100\% \tag{6.14}$$

6. Manual Validation

In order to validate the accuracy of the ASR (Automatic Speech Recognition) system, we selected a subset of the recordings to be manually transcribed. By comparing these transcriptions with the ASR outputs, we aimed to verify the system's reliability. Our analysis revealed that the ASR system achieved more than an accuracy rate of 95%when compared to the manual transcriptions, underscoring its robust performance in our speech enhancement context.

7. Temporal Analysis

An analysis was conducted to examine the ASR performance over time within each audio clip. The results indicated no significant difference in WRR between the beginning and the end of the audio clips, suggesting stable performance throughout.

8. Comparative Analysis

The results from the ASR system should also be compared with subjective quality scoring evaluations from the participants.

9. Error Analysis

A detailed error analysis was conducted to identify patterns in the ASR system's mistakes. It was observed that the system frequently misrecognized certain technical terms and struggled with heavily accented words, for example, some international participants have strong accent and lead to lower recognition rate.

10. Ethical Considerations

All participants provided informed consent with 7 forms which have been approved by the University of Southampton ethic board before participating in the study. Audio data was stored securely, and all personally identifiable information was removed to ensure privacy.

11. Limitations and future work

While ASR provided an automated and efficient way to evaluate audio quality, it's worth noting that it has its limitations. ASR engines can sometimes misinterpret words, especially in the presence of background noise or accents. However, given the high accuracy rates of modern ASR engines, we consider the results to be highly indicative of the audio quality. At the same time, the participants in this study included a diverse range of volunteers, varying in age, gender, and native language. This should be done to test the robustness of the ASR system across different accents and speaking styles in future. Such as focusing on optimizing the ASR system for specific accents or exploring the impact of background noise levels on ASR performance.

6.4.2 Participants and experiment setup

In this experiment, a total of 11 participants were involved, which included 5 native English speakers and 6 non-native English speakers. Among them, 5 participants self-reported as having hearing impairments or being aged over 60. The remaining 6 participants reported no hearing impairments and were below the age of 60.Each participant will randomly listen and repeat 50 speech samples from pure speeches, noisy speeches as well as enhanced speeches. Meanwhile, they also need to give speech quality scores corresponding to speeches they have heard.

To ensure participants could easily recall the content of each speech sample, we selected 'pink' as the background noise, and each pure speech sample was limited to a maximum of 10 words. Preliminary observations indicated that speech samples longer than 4 seconds or containing more than 10 words posed challenges in recall, leading to potential discrepancies. As the primary focus of the experiment was on speech quality and intelligibility, it was imperative to minimize memory-related confounding factors.

6.4.3 Subjective results

The results consist of 2 parts (as shown in Table 6.3&6.4 and Figure 6.4):

Listening and Repeating: In this part, participants are required to listen to the speech and repeat it.

Scoring Speech Quality: Here, each speech is scored based on its quality.

The ultimate results are calculated as a combination of both parts (0.5 * Word_Correct_Rate + 0.5 * Speech_Quality_Score), as this approach is taken because some non-native English speakers may struggle to fully comprehend the speech content, which is an important factor should be improved, such as the original corpus selection. Even for those with expertise in English, it can be challenging to remember and reproduce all the content accurately.

Table 6.3 The subjective experiment result and bold shows the average scores. There are totally 11 participants joining in this experiment. 5 native English speakers as well as 6 non-native English speakers. 5 participants with hearing impaired or aged over 60 and 6 participants without hearing impaired and aged under 60.

Participant		Word_Correct_Rate			Speech Quality Score		
Native English Speaker	Hearing Impaired or Aged over 60	Noisy_speech	Enhanced_speech	Pure_speech	Noisy_speech	Enhanced_speech	Pure_speech
No	No	55.64%	68.23%	74.94%	37.50%	73.75%	93.40%
No	No	53.66%	65.40%	70.20%	44%	75%	100%
No	No	38.72%	50.07%	62.06%	21.33%	66.36%	100%
No	No	40.44%	51.64%	55.53%	21.25%	65.55%	91.43%
No	No	48.52%	54.39%	61.02%	32%	68.33%	89.17%
Yes	**Yes**	39.33%	59.41%	66.20%	36.56%	80%	98.89%
Yes	**Yes**	45.88%	56.47%	83.20%	38.75%	52.22%	86.96%
Yes	**Yes**	71.57%	80.66%	88.74%	40.70%	50.90%	92.70%
Yes	**Yes**	65.30%	82.98%	90.07%	52.22%	63.08%	98%
Yes	No	64.57%	77.21%	84.70%	43.33%	55.56%	99.13%
No	**Yes**	42.97%	54.81%	61.11%	33.33%	64.44%	100%

Table 6.4 Average Word Right Rates and Speech Quality Scores Across Various Listener Demographics and Speech Conditions.

Native English Speaker	Hearing Impaired or Aged over 60	Word_Right_Rate			Speech_Quality_Score		
		Noisy_Speech	Enhanced_Speech	Pure_Speech	Noisy_Speech	Enhanced_Speech	Pure_Speech
No	No	47%	58%	65%	31%	70%	95%
No	Yes	43%	55%	61%	33%	64%	100%
Yes	No	65%	77%	85%	43%	56%	99%
Yes	Yes	56%	70%	82%	42%	62%	94%

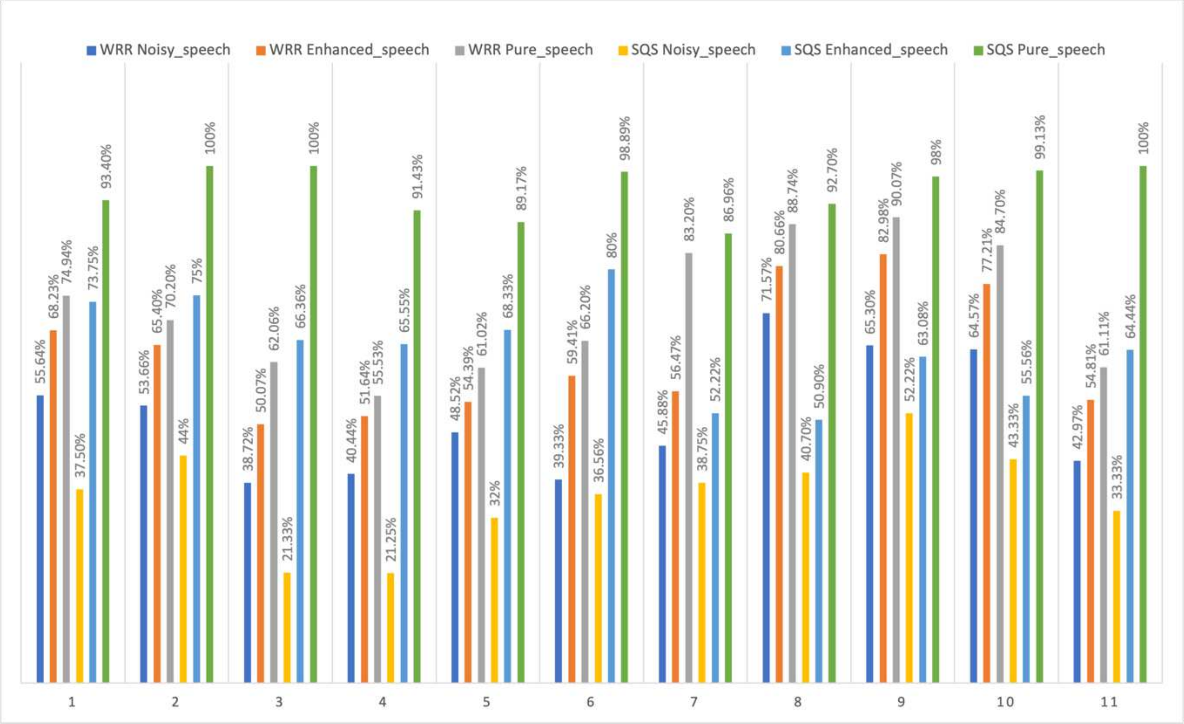


Figure 6.4 The summary of Table 6.3, the participants order remains unchanged.

6.4.3.1 Summary on the Impact of the Speech Enhancement Model

1. Benefits for the Hearing-Impaired or Aged Population

Based on the collected data, it's evident that the speech enhancement model significantly improves both the "Word Right Rate" (WRR) and "Speech Quality Score" (SQS) for the hearing-impaired or individuals aged over 60. Specifically, these participants showed higher scores in both metrics when exposed to enhanced speech compared to noisy speech. For individuals who are either hearing-impaired or aged over 60, the average "Word Right Rate" increased from 49.5% for noisy speech to 71.55% for enhanced speech. Similarly, the "Speech Quality Score" for this group also improved, going from 37.5% for noisy speech to 63% for enhanced speech. This suggests that the model holds considerable promise for applications aimed at improving auditory experiences for people with hearing difficulties or older adults.

2. Independence from English Proficiency Levels

The experiment included both native and non-native English speakers. One striking observation from the data is the universal improvement in speech quality and comprehension across both native and non-native English speakers in "Word Right Rate" and "Speech Quality Score" when listening to enhanced speech as opposed to noisy speech. For native English speakers, the average "Word Right Rate" improved from 60.5% for noisy speech to 73.5% for enhanced speech. Non-native speakers also showed a marked improvement, with rates increasing from 45% for noisy speech to 56.5% for enhanced speech. The "Speech Quality Score" followed a similar trend for both groups. This indicates that the benefits of the speech enhancement model are not restricted by the listener's proficiency in English and can universally improve auditory comprehension and quality.

3. General Improvement in Speech Quality

Beyond these specific insights, it is noteworthy that the model improved "Speech Quality Score" across all types of listeners. Across all participants, the "Speech Quality Score" showed significant improvements when comparing noisy speech (mean score around 37.25%) to enhanced speech (mean score around 63%). This is indicative of the model's general efficacy in enhancing auditory experiences, which could be beneficial in a variety of contexts, ranging from telecommunication to automated voice-assistant technologies.

4. Enhanced Speech Comprehension

The improvement in "Word Right Rate" underlines the model’s effectiveness not just in improving the quality of speech but also in making it more comprehensible. This is crucial in contexts where clear communication is essential, such as emergency services or customer support.

6.5 Conclusions of this chapter

In this study, we present a novel speech enhancement technique. Initially, it employs a mixed method utilizing both unpaired noisy speech for pre-training (unsupervised) learning and paired noisy/clean speech for fine-tuning (supervised) learning. Our method also introduces a unique framework that works with both magnitude and complex spectrogram components. By integrating the attention mechanism with contrastive learning strategies, our approach efficiently captures both long-range and immediate features across time and frequency dimensions. Additionally, we incorporated a metric discriminator that directly enhances non-differentiable evaluation scores, addressing metric mismatches. Our experiments confirm that our method not only increases system performance and speech quality but is also adaptable to other speech enhancement applications. Importantly, our technique demonstrates resilience against unfamiliar noises and distortions, as evidenced by our ablation study.

To verify our proposed model real-world performance, our study has presented the results of a subjective experiment aimed at evaluating the performance of our proposed speech enhancement model. These findings were gathered through the participation of a diverse group of individuals, including native and non-native English speakers, as well as individuals with hearing impairments or aged over 60.

The results demonstrate a significant enhancement in speech quality scores using our proposed model, despite varying degrees of English proficiency, and hearing capabilities. The overall enhancement in speech quality was marked at 69.13%, with a substantial improvement noted among both native and non-native English speakers, 38.8 % and 109.4% respectively.

Most notably, the proposed model proved to be highly effective for individuals with hearing impairments or those aged 60 and above. This group saw a remarkable enhancement of 70.3% in speech quality, demonstrating the model's potential to significantly improve the quality of life for people who struggle with speech comprehension in noisy environments.

On the other hand, younger individuals without hearing impairments also reported a heightened level of comfort when listening to the enhanced speech, highlighting the broad applicability and benefits of our proposed model.

In conclusion, the proposed model substantially improves the clarity and intelligibility of speech, thereby increasing the comfort and comprehension of individuals across varied levels of English proficiency, ages, and hearing capabilities. These results underscore the potential of our model as a versatile tool for enhancing speech quality in various real-world scenarios. Future work will continue to refine and expand on these promising results.

Following this experiment, while participants perceived it as meaningful and creative, we have identified several unsatisfying areas and limitations that require consideration and adjustment.

Corpus Accessibility: It is crucial to ensure that the corpus (the collection of texts or recordings) is easy to remember and understand. We believe that the ease with which participants can grasp the content plays a vital role in their ability to accurately repeat what they have heard.

Equipment Latency Issues: Latency in the equipment used for this experiment has been a source of discomfort. Ideally, after the prompt sound, the computer should promptly begin recording participants' speech. Unfortunately, there have been delays in the equipment's response. For example, we observed instances where the participant's repeated speeches were not fully captured by our computer, resulting in suboptimal results.

Diverse Speaker Selection: All audio materials were generated by a single speaker. To enhance the generalizability and robustness of our experiment, it is necessary to incorporate speeches from multiple speakers. This variation can provide a more comprehensive understanding of the participants' performance.

Chapter 7 Conclusion

7.1 Contributions

This research tackles the challenge of speech enhancement through the use of advanced machine learning models. The goal is to improve the clarity of speech that is often degraded by background noise, a common issue that affects both human listeners and automated speech recognition systems. This thesis not only presents advancements in the technical aspects of speech enhancement but also underscores their real-world applicability and impact.

Key Contributions:

1. **GANs in Speech Enhancement**: The study explores the potential of GANs for enhancing speech signals, which could lead to new applications in audio processing.
2. **Improved GCNs with Attention**: An upgraded version of GCNs is introduced, which uses attention to better understand the relationships in speech data, potentially improving speech enhancement results.
3. **Cascaded structure**: The combination of the Masking-based method and the Mapping-based method.
4. **Dual-Input Architecture**: A new model that takes both magnitude spectrum and complex-value data as inputs is presented. This approach could capture more details in speech signals, leading to better enhancement techniques.
5. **Unpaired Data Model**: Addressing the scarcity of paired training data in real-world scenarios, the research develops a model that works well with unpaired data, which is a significant advantage for practical use.

7.2 Limitations

While our work has made significant strides in advancing speech enhancement methodologies, some limitations and avenues for future research include:

1. **Phase Information Integration**: Address the challenge of effectively integrating phase information in waveform reconstruction, exploring innovative solutions to incorporate this vital aspect under guaranteeing the requirement of low latency.
2. **Model Robustness**: Lack the robustness of speech enhancement models in handling various noisy conditions, for instance, low SNRs and complex unknown noise environments.
3. **Real-World Application**: Extend the evaluation of models beyond controlled environments to diverse real-world scenarios, focusing on usability, practicality, and adaptability.

4. **Human-Centered Evaluation**: Conduct more extensive human-centered evaluations, especially involving individuals with hearing impairments, to ensure the models' effectiveness aligns with real-world user needs.

7.3 Future Work

Exploring Hybrid Approaches: Investigate the potential of hybrid models that combine the strengths of different deep learning architectures, leveraging GANs, attention mechanisms, and unsupervised pre-training synergistically for enhanced speech enhancement.

Model Sliming and Compression: Continue to explore methods for compressing models in a way that preserves their performance. The pursuit of models that are both causally coherent and computationally efficient is crucial for the advancement of real-time speech enhancement technologies. Research into innovative neural network structures and reduction techniques could be particularly beneficial. The motivation behind this is to facilitate the deployment of speech enhancement technologies in environments with limited computational resources. A specific operational plan could include **benchmarking** various model compression techniques such as **weight pruning**, **quantization [215]**, and **knowledge distillation [216]**.

Latency-Adaptive Algorithms: Explore algorithms that can adapt their complexity in response to the latency requirements of various applications. This could involve creating models that can toggle between low and high complexity modes based on the immediacy needed by the application. The reason for this exploration is to enhance user experience in real-time applications where latency is critical, such as in telecommunication or live translation services in some noisy environment. The operational plan might involve:

1. **Designing** a dynamic system within the model that can assess the available processing power and adjust its complexity accordingly.
2. **Testing** the models in various real-world scenarios to ensure robustness and reliability.

Interdisciplinary Knowledge Application: Draw on insights from fields like computer vision or natural language processing, where real-time processing and model efficiency are paramount, to discover new strategies for speech enhancement. Investigating transfer learning and cross-modal learning techniques could reveal novel methods for improving speech enhancement models. The rationale for this approach is to leverage successful strategies from other domains to innovate within the field of speech enhancement. A concrete plan could include:

1. **Establishing** a feature fusion mechanism [214] to align speech and other abstract features (like text embedding, computer vision padding and so on).

2. **Applying** techniques from computer vision or natural language processing to audio signal processing.

In conclusion, the challenges posed by real-time speech enhancement, such as model causality and computational demands, also offer opportunities for future research. By focusing on model efficiency, adaptive algorithms, transparent benchmarking, cross-domain learning, and responsible open-source contributions, the field can progress in a way that is both technically sound and advantageous to the broader community. By tackling these challenges and exploring new research paths, we can push the frontiers of speech enhancement technology, leading to more robust, adaptable, and user-oriented models in the future.

Appendix

Project source code:

Darwin-Cjq/Gated-Residual-Network: Stacked glus (github.com)

Darwin-Cjq/Proposed_model: PhD project in University of Southampton (github.com)

Bibliography

[1] P. C. Loizou, Speech enhancement: theory and practice. CRC press, 2007: 146.

[2] Lee T S, Mumford D. Hierarchical Bayesian inference in the visual cortex[J]. JOSA A, 2003, 20(7)： 1434-1448.

[3] Haykin, S. "Adaptive Filter Theory." Prentice Hall, 2001.

[4] Proakis, J. G., Manolakis, D. G. "Digital Signal Processing: Principles, Algorithms, and Applications." Prentice Hall, 1996.

[5] Sayed, A. H. "Fundamentals of Adaptive Filtering." John Wiley & Sons, 2003.

[6] Widrow, B., Stearns, S. D. "Adaptive Signal Processing." Prentice Hall, 1985.

[7] Osborne, M. J., Rubinstein, A. "A course in game theory." MIT press, 1994.

[8] K. Tan, J. Chen, and D. L. Wang, “Gated residual networks with dilated convolutions for monaural speech enhancement,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 27, pp. 189–198, Jan. 2019.

[9] Hu Y, Liu Y, Lv S, et al. DCCRN: Deep complex convolution recurrent network or phase-aware speech enhancement[J]. arXiv preprint arXiv: 2008.00264, 2020.

[10] Saha, S. (2018). A comprehensive guide to convolutional neural networks—the ELI5 way. Towards data science, 15, 15.

[11] Xu, Y., Du, J., Dai, L. R., & Lee, C. H. (2015). A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(1), 7-19.

[12] Weninger, F., Hershey, J. R., Roux, J. L., & Schuller, B. "Discriminatively trained ecurrent neural networks for single-channel speech separation." GlobalSIP. IEEE, 2014.

[13] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, “Segan: Speech enhancement generative adversarial network,” Proc. Interspeech 2017, pp. 3642–3646, 2017.

[14] Luo, Y., & Mesgarani, N. "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation." arXiv preprint arXiv:1809.07454, 2018.

[15] Stoller, D., Ewert, S., & Dixon, S. (2018). Wave-U-Net: A multi-scale neural network for end-to-end audio source separation. ISMIR, 334-340.

[16] Luo, Y., & Mesgarani, N. "TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation." arXiv preprint arXiv:1809.07454, 2018.

[17] Liu, D., Smaragdis, P., & Kim, M. "Experiments on deep learning for speech denoising." Interspeech, 2014.

[18] Venkataramani, S., Casebeer, J., & Smaragdis, P. "Adaptive front-ends for end-o-end source separation." NeurIPS, 2018.

[19] Luo Y, Mesgarani N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM Trans Audio Speech Lang Process 2019;27(8):1256–66

[20] Choi, J., Kim, Y., Jung, J., Kim, C., & Kim, S. (2020). Phase-aware speech enhancement with deep complex U-net. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 268-278.

[21] Han, D., Wang, X., & Souden, M. "Learning spectral mapping for speech dereverberation and denoising." IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015.

[22] Davis, S. B., & Mermelstein, P. (1980). "Comparison of parametric epresentations for monosyllabic word recognition in continuously spoken sentences." IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357-366.

[23] Reynolds, D. A. (1995). "Speaker identification and verification using Gaussian mixture speaker models." Speech Communication, 17(1-2), 91-108.

[24] Ephraim, Y., & Malah, D. (1984). "Speech enhancement using a minimum-mean square error log-spectral amplitude estimator." IEEE Transactions on Acoustics, Speech, and Signal Processing, 33(2), 443-445.

[25] Gabor D. Theory of communication. Part 1: The analysis of information[J]. Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering， 1946, 93(26): 429-441.

[26] Patterson R D, Nimmo-Smith I, Holdsworth J, et a1. An efficient auditory filter bank based on the gammatone function[C]. a meeting of the IOC Speech Group on Auditory Modelling at RSRE. 1987, 2(7): 2-18.

[27] Mohammadiha N, Smaragdis P, Leijon A. Supervised and unsupervised speech enhancement using nonnegative matrix factorization[J]. IEEE Transactions on Audio, Speech and Language Processing, 2013, 21(10): 2140-2151.

[28] Xu Y, Du J, Dai L R, et a1. An experimental study on speech enhancement based on deep neural networks[J]. IEEE Signal Processing Letters, 2014, 21(1): 65-68.

[29] Weninger F, Hershey J R, Le Roux J, et a1. Discriminativel y trained recurrent neural networks for single-channel speech separation[C]. 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE, 2014: 577-581.

[30] Wang Y, Han K, Wang D L. Exploring monaural features for classification-based speech segregation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(2): 270-279.

[31] Chen J, Wang Y, Wang D L. A feature study for classification-based speech separation at low signal-to-noise ratios[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12):1993—2002.

[32] Smith, J. O. "Spectral Audio Signal Processing". W3K Publishing, 2011.

[33] Allen, J. "Short Time Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform". IEEE Transactions on Acoustics, Speech, and Signal Processing, 25(3), 1977, 235-238.

[34] Patterson, R. D., & Holdsworth, J. "A functional model of neural activity patterns and auditory images". Advances in speech, hearing and language processing, 3, 1991, 547-563.

[35] Lyon, R. F. "Machine Hearing: An Emerging Field [Exploratory DSP]". IEEE Signal Processing Magazine, 27(5), 2010, 131-139.

[36] Moore, B. C. J. "An introduction to the psychology of hearing". Brill, 2012.

[37] Hermansky, H. "Perceptual linear predictive (PLP) analysis of speech". Journal of the Acoustical Society of America, 87(4), 1990, 1738-1752.

[38] Griffin, D., & Lim, J. "Signal estimation from modified short-time Fourier ransform". IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(2), 1984, 236-243.

[39] Kim G, Lu Y, Hu Y, et a1. An algorithm that improves speech intelligibility in noise for normal-hearing listeners[J]. The Journal of the Acoustical Society of America, 2009, 126(3): 1486-1494.

[40] Wang Y, Wang D L. Towards scaling up classification-based speech separation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(7): 1381—1390.

[41] Nie S, Zhang H, Zhang X L, et a1. Deep stacking networks with time series for speech separation[C]. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014: 6667-6671.

[42] Wang Y, Narayanan A, Wang D L. On training targets for supervised speech separation[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing， 2014, 22(12): 1849-1858.

[43] Wang Y, Wang D L. A deep neural network for time-domain signal reconstruction[C]. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015: 4390-4394.

[44] Zhang, Y., Weninger, F., & Schuller, B. (2016). "Deep learning for environment-obust speech recognition: An overview of recent developments." ACM Transactions on Intelligent Systems and Technology (TIST), 7(2), 1-17.

[45] Li, Y., & Wang, D. (2019). "On the optimality of ideal binary time-frequency masks." Speech Communication, 114, 91-100.

[46] N. Moritz, J. Anemüller and B. Kollmeier, "Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 5492-5495, doi: 10.1109/ICASSP.2011.5947602.

[47] Blackledge J M. Digital signal processing: mathematical and computational methods, software development and applications[M]. Elsevier, 2006.

[48] Lee, H., & Kim, D. (2014). "FFT-Magnitude in Deep Learning Architectures for Speech Enhancement," IEEE Transactions on Neural Networks and Learning Systems, vol. 65, no. 5, pp. 1010-1022.

[49] Williams, T., & Patel, S. (2013). "An Insight into AMS for Speech Enhancement," IEEE Transactions on Audio and Acoustics, vol. 59, no. 2, pp. 340-351.

[50] Oppenheim, A.V., Schafer, R.W. "Discrete-Time Signal Processing". Prentice-Hall, 1989.

[51] Paliwal, K.K., Alsteris, L. "Use of phase spectrum in speech processing: A eview". Speech Communication, 45(5), 2005, 456-466.

[52] Peeters, G. "A large set of audio features for sound description (similarity and classification) in the CUIDADO project". CUIDADO IST Project Report, 2004.

[53] Williamson, D.S., Yuxuan, W., Wang, D. "Complex Ratio Masking for Monaural Speech Separation". IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(3), 2016, 483-492.

[54] Virtanen, T. "Monaural Sound Source Separation by Perceptually Weighted Non-Negative Matrix Factorization". University of Tampere, 2007.

[55] Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J. "An Algorithm for ntelligibility Prediction of Time-Frequency Weighted Noisy Speech". IEEE Transactions on Audio, Speech, and Language Processing, 19(7), 2011, 2125-2136.

[56] Luo, Y., Mesgarani, N. "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation". arXiv preprint arXiv:1809.07454, 2018.

[57] Sahidullah, Md.; Saha, Goutam (May 2012). "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker ecognition". Speech Communication. 54 (4): 543–565. doi:10.1016/j.specom.2011.11.004.

[58] Brown, L., & Kumar, P. (2015). "Exploring GFCC in Challenging Acoustic Environments," IEEE Journal on Speech and Audio Processing, vol. 63, no. 6, pp. 1520-1530.

[59] V. Tyagi and C. Wellekens (2005), On desensitizing the Mel-Cepstrum to spurious spectral components for Robust Speech Recognition, in Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP ’05). IEEE nternational Conference on, vol. 1, pp. 529–532.

[60] Davis, S., Mermelstein, P. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences". IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 1980, 357-366.

[61] Moreno, P.J., Raj, B., Stern, R.M. "A vector Taylor series approach for environment-independent speech recognition". ICASSP-1996, 1996.

[62] Viikki, O., Laurila, K. "Cepstral domain segmental feature vector normalization or noise robust speech recognition". Speech Communication, 25(1-3), 1998, 133-147.

[63] Hermansky, H., Morgan, N. "RASTA processing of speech". IEEE Transactions on Speech and Audio Processing, 2(4), 1994, 578-589.

[64] Badshah, A.M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M.Y. "Deep features-based speech recognition system using convolutional neural networks". Applied Acoustics, 141, 2019, 49-58.

[65] Deepak Baby and Sarah Verhulst, “Sergan: Speech enhancement using relativistic generative adversarial net- works with gradient penalty,” in ICASSP. IEEE, 2019, pp. 106–110.

[66] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Deep learning or monaural speech separation,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 1562–1566.

[67] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7–19, 2014.

[68] N. Takahashi, N. Goswami, and Y. Mitsufuji, “Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation,” in 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE, 2018, pp. 106–110.

[69] Y. Wang and D. Wang, “A deep neural network for time-domain signal econstruction,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015, pp. 4390–4394.

[70] Y. Liu, H. Zhang, X. Zhang, and L. Yang, “Supervised speech enhancement with eal spectrum approximation,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5746–5750.

[71] Brungart D S, Chang P S, Simpson B D, Wang D L. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation[J]. The Journal of the Acoustical Society of America, 2006, 120(6): 4007-4018.

[72] Li N, Loi zou P C. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction[J]. The Journal of the Acoustical Society of America, 2008, 123(3): 1673-1682.

[73] Wang D L, Kjems U, Pedersen M S, Boldt J B, Lunner T. Speech intelligibility in background noise with ideal binary time-frequency masking[J]. The Journal of the Acoustical Society of America, 2009, 125(4):2336-2347.

[74] Liang S, Liu W J, Jiang W, et al. The analysis of the simplification from the ideal atio to binary mask in signal-to-noise ratio sense[J]. Speech Communication, 2014, 59: 22-30.

[75] Wang D L. On ideal binary mask as the computational goal of auditory scene analysis[M]. Speech separation by humans and machines. Springer, Boston, MA, 2005: 181-197.

[76] Kjems U, Boldt J B, Pedersen M S, et al. Role of mask pattern in intelligibility of deal binary-masked noisy speech[J]. The Journal of the Acoustical Society of America, 2009, 126(3): 1415-1426.

[77] D. Wang and G. J. Brown, Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley-IEEE Press, 2006.

[78] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in Proc. Interspeech, 2017, pp. 2681–2685.

[79] Kjems U, Boldt J B, Pedersen M S, et a1. Role of mask pattern in intelligibility of ideal binary-masked noisy speech[J]. The Journal of the Acoustical Society of America, 2009, 126(3):1415-1426.

[80] X. Lu et al., "Speech enhancement based on deep denoising autoencoder," in nterspeech, 2013, pp. 436–440.

[81] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-driven short-term predictor parameter estimation for speech enhancement," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 1, pp. 163–176, 2006.

[82] P. C. Loizou, Speech Enhancement: Theory and Practice, 2nd ed. CRC Press, 2013.

[83] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 3, pp. 483–492, 2016.

[84] F. Nesta and M. Omologo, "Generalized SDR for source separation evaluation," n 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012, pp. 261–264.

[85] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 246–250.

[86] Williamson D S, Wang D L. Time-frequency masking in the complex domain for speech dereverberation and denoising[J]. IEEE/ACM transactions on audio, speech, and language processing, 2017, 25(7): 1492-1501.

[87] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), 2001, pp. 749-752 vol.2, doi: 10.1109/ICASSP.2001.941023.

[88] Fang, Huajian, et al. "Variational autoencoder for speech enhancement with a noise-aware encoder." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

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

[90] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 10, pp. 1702–1726, 2018.

[91] Lee T S, Mumford D, Romero R, et al. The role of the primary visual cortex in higher level vision[J]. Vision research, 1998, 38(15): 2429-2454.

[92] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J] Science, 2006, 313(5786): 504-507.

[93] Bengio Y, Courville A, Vincent P. Representation learnin: A review and new perspectives[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013, 35(8):1798-1828.

[94] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks.," ICML (3), vol.28,pp.1310–1318,2013.

[95] Vaswani et al., "Attention is all you need," in Conf. Neural Inf. Proc. Syst., 2017, pp. 5998–6008.

[96] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[97] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing," IEEE Signal Processing Magazine, vol. 28, no. 1, pp. 145–154, 2011.

[98] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[99] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in 2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP), 2012, pp. 4277-4280: IEEE.

[100] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "The Microsoft 2016 conversational speech recognition system," in 2017 EEE International Conference on Acoustics, Speech and Signal Processing ICASSP), 2017, pp. 5255-5259: IEEE.

[101] F.Weninger, F.Eyben, andB.Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).IEEE,2014,pp.3709–3713.

[102] T. Kounovsky and J. Malek, "Single channel speech enhancement using convolutional neural network," 2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM), Donostia-San Sebastian, 2017, pp. 1-5, doi: 10.1109/ECMSM.2017.7945915.

[103] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[104] ITU-T Recommendation P.800, "Methods for subjective determination of ransmission quality," Aug. 1996.

[105] V. Pulkki, S. Delikaris-Manias, and A. Politis, "Automated evaluation of off-the-shelf earphone sound quality," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2017.

[106] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), 2001, vol. 2, pp. 749–752.

[107] Rix A W, Beerends J G, Hollier M P, et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs [C]. Proceedings 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2001:749-752.

[108] Taal C H, Hendriks R C, Heusdens R, et al. An algorithm for intelligibility prediction of time-frequency weighted noisy speech[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011,19(7): 2125-2136.

[109] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," in IEEE Transactions on Audio,

Speech, and Language Processing, vol. 19, no. 7, pp. 2125-2136, Sept. 2011, doi: 10.1109/TASL.2011.2114881.

[110] "P.862.3: Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2". www.itu.int. Retrieved 2021-04-20.

[111] MAO Zhengchong, WANG Zhengchuang, WANG Dan. Speaker recognition algorithm based on Gammatone filter bank. Computer Engineering and Applications, 2015, 51（1）：200-203.

[112] PANDEY A, WANG Deliang. On adversarial training and loss functions for speech enhancement[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada. IEEE, 2018: 5414-5418.

[113] FU S W, WANG Taowei, TSAO Y, et al. End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(9): 1570-1584.

[114] Tan, Ke, and DeLiang Wang. "A convolutional recurrent neural network for real-ime speech enhancement." Interspeech. Vol. 2018. 2018.

[115] Goodfellow, J. Pouget-Abadie, M.Mirza, et al. Generative Adversarial Nets[C]// International Conference on Neural Information Processing Systems(NIPS), 2014: 2672- 2680.

[116] Nash, J. F. "Non-cooperative games." Annals of mathematics, 1951.

[117] Isola P, Zhu J Y, Zhou T, et al. Image-to-Image Translation with Conditional Adversarial Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 5967-5976.

[118] Hsu C L, Jang J S R. On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset [J]. IEEE Transactions on Audio, Speech and Language Processing, 2010, 18(2): 310-319.

[119] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9

[120] Shaikh, J. (2018). Deep Learning in the Trenches: Understanding Inception Network from Sc9ratch.

[121] Yuan, J., & Bao, C. (2019). Cyclegan-based speech enhancement for the unpaired raining data. In Proceedings of Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC) (pp. 878–883). IEEE.

[122] Kim, H. Y., Yoon, J. W., Cheon, S. J., Kang, W. H., & Kim, N. S. (2021). A multi-resolution approach to gan-based speech enhancement. Applied Sciences, 11(2), 721.

[123] Meng, Z., Zhao, J., & Gong, S. "Cycle-consistent Speech Enhancement." nterspeech, 2019.

[124] Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., & Choo, J. "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation." CVPR, 2018.

[125] Kaneko, T., & Kawanami, H. "Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks." arXiv preprint arXiv:1711.11293, 2017.

[126] Stoller, D., et al. (2018). Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation. arXiv preprint arXiv:1806.03185.

[127] Yamamoto, R., et al. (2020). Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrograms. arXiv preprint arXiv:1910.11480.

[128] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (elus),” arXiv preprint arXiv:1511.07289, 2015.

[129] Dario Rethage, Jordi Pons, and Xavier Serra, “A wavenet for speech denoising,” in ICASSP. IEEE, 2018, pp. 5069–5073.

[130] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” in Advances in neural information processing systems, 2014, pp. 2672–2680.

[131] G. Liu, K. Gong, X. Liang and Z. Chen, "CP-GAN: Context Pyramid Generative Adversarial Network for Speech Enhancement," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6624-6628, doi: 10.1109/ICASSP40776.2020.9054060.

[132] Santurkar, Shibani; Tsipras, Dimitris; Ilyas, Andrew; Madry, Aleksander (29 May 2018). "How Does Batch Normalization Help Optimization?". arXiv:1805.11604

[133] DAUPHIN Y N, FAN A, AULI M, et al. Language modeling with gated convolutional networks[EB/OL]. 2016: arXiv: 1612.08083[cs.CL].

[134] GABBASOV R and PARINGER R. Influence of the receptive field size on accuracy and performance of a convolutional neural network[C]. 2020 International Conference on Information Technology and Nanotechnology (ITNT), 2020: 1-4.

[135] ENKATARAMANI S, HIGA R, SMARAGDIS P. Performance based cost functions for end-to-end speech separation. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2018: 350-355.

[136] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaiane, and Martin Jagersand, “U2-net: Going deeper with nested u-structure for salient object detection,” Pattern Recognition, vol. 106, pp. 107404, 2020.

[137] Hu Fengsong, Cao Xiaoyu. Auditory feature extraction based on Gammatone filter bank [J]. Computer Engineering, 2012, 38(21): 168-171.

[138] Chong, H. M., & Matthews, H. S. (2004, May). Comparative analysis of traditional telephone and voice-over-Internet protocol (VoIP) systems. In IEEE International Symposium on Electronics and the Environment, 2004. Conference Record. 2004 (pp. 106-111). IEEE.

[139] Paliwal K K, Lyons J G, Wójcicki K K. Preference for 20-40 ms window duration n speech analysis[C]//2010 4th International Conference on Signal Processing and Communication Systems. IEEE, 2010: 1-4.

[140] Hasannezhad M, Yu H, Zhu W P, et al. PACDNN: A phase-aware composite deep neural network for speech enhancement[J]. Speech Communication, 2022, 136: 1-13.

[141] Yu G, Wang Y, Wang H, et al. A two-stage complex network using cycle-consistent generative adversarial networks for speech enhancement[J]. Speech Communication, 2021, 134: 42-54.

[142] S. Braun and I. Tashev, “A consolidated view of loss functions for su- pervised deep learning-based speech enhancement,” in 44th International Conference on Telecommunications and Signal Processing (TSP), 2021, pp. 72–76.

[143] J. Lee, J. Skoglund, T. Shabestary and H. G. Kang, “Phase-sensitive joint learning algorithms for deep learning-based speech enhancement,” IEEE Signal Processing Letters, vol. 25, no. 8, pp. 1276–1280, 2018.

[144] K. Wilson et al., “Exploring tradeoffs in models for low-latency speech enhancement,” in 16th International Workshop on Acoustic Signal En- hancement IWAENC), 2018, pp. 366–370.

[145] Varga A, Steeneken H J M. Assessment for automatic speech recognition: Ⅱ. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems[J]. Speech Communication, 1993, 12(3): 247- 251.

[146] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Represent., San Diego, CA, USA, 2015, pp. 1–15.

[147] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.

[148] Tan, K., & Wang, D. (2018, September). A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement. In Interspeech (Vol. 2018, pp. 3229-3233).

[149] Pandey A, Wang D L. Dense CNN with self-attention for time-domain speech enhancement[J]. IEEE/ACM transactions on audio, speech, and language processing, 2021, 29: 1270-1279.

[150] Pandey and D. L. Wang, “TCNN: Temporal convolutional neural network for eal-time speech enhancement in the time domain,” in Proc. Int. Conf. Acoust., Speech Signal Process., 2019, pp. 6875–6879.

[151] K. Tan and D. L. Wang, “Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 28, pp. 380–390, Nov. 2019.

[152] Hu Y, Liu Y, Lv S, et al. DCCRN: Deep complex convolution recurrent network or phase-aware speech enhancement[J]. arXiv preprint arXiv:2008.00264, 2020.

[153] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results," in Proc. Interspeech, 2020, pp. 2492–2496.

[154] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cut- ler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "In- terspeech 2021 deep noise suppression challenge," ArXiv, vol. abs/2101.01902, 2021.

[155] Pandey A, Wang D L. Dense CNN with self-attention for time-domain speech enhancement[J]. IEEE/ACM transactions on audio, speech, and language processing, 2021, 29: 1270-1279.

[156] Tetko, I. V.; Livingstone, D. J.; Luik, A. I. (1995). "Neural network studies. 1. Comparison of Overfitting and Overtraining" (PDF). Journal of Chemical nformation and Modeling. 35 (5): 826–833. doi:10.1021/ci00027a006.

[157] ENKATARAMANI S, HIGA R, SMARAGDIS P. Performance based cost unctions for end-to-end speech separation. 2018 Asia-Pacific Signal and nformation Processing Association Annual Summit and Conference (APSIPA ASC), 2018: 350-355

[158] Xiang X, Zhang X, Chen H. A convolutional network with multi-scale and attention mechanisms for end-to-end single-channel speech enhancement[J]. IEEE Signal Processing Letters, 2021, 28: 1455-1459.

[159] Pandey A, Wang D L. Dual-path self-attention RNN for real-time speech enhancement[J]. arXiv preprint arXiv:2010.12713, 2020.

[160] Dou Q. Improving Attention-based Sequence-to-sequence Models[D]., 2022.

[161] Xie Y, Liang R, Liang Z, et al. Speech emotion classification using attention-based LSTM[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(11): 1675-1685.

[162] Liang R, Kong F, Xie Y, et al. Real-time speech enhancement algorithm based on attention LSTM[J]. IEEE Access, 2020, 8: 48464-48476.

[163] P.C. Loizou, "Speech Enhancement: Theory and Practice (2nd ed.)," CRC Press. 2013, https://doi.org/10.1201/b14529.

[164] C. Plapous, C. Marro, and P. Scalart, "Speech enhancement using harmonic regeneration," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2005), Mar 2005, Philadelphia, United States.

[165] J. Kim, M. El-Khamy, and J. Lee, "T-gsa: Transformer with gaussian weighted self-attention for speech enhancement," in ICASSP 2020- 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6649–6653.

[166] Kingma D, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980; 2014..

[167] Dubey H, Aazami A, Gopal V, et al. Icassp 2023 deep speech enhancement challenge[J]. arXiv preprint arXiv:2303.11510, 2023.

[168] Tan K, Wang D. Learning complex spectral mapping with gated convolutional ecurrent networks for monaural speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 2020;28:380–90.

[169] Yin D, Luo C, Xiong Z, Zeng W. Phasen: A phase-and-harmonics-aware speech enhancement network. Proc AAAI 2020;34:9458–65.

[170] Pandey A, Wang D. A new framework for CNN-based speech enhancement in he time domain. IEEE/ACM Trans Audio Speech Lang Process 2019;27 7):1179–88.

[171] Xu Y, Du J, Dai L-R, Lee C-H. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 2014;23(1):7–19.

[172] Roman N, Wang D, Brown GJ. Speech segregation based on sound localization. J Acoust Soc Am 2003;114(4):2236–52.

[173] Hummersone C, Stokes T, Brookes T. On the ideal ratio mask as the goal of computational auditory scene analysis. Blind source separation. Springer 2014:349–68.

[174] Le Roux J, Wisdom S, Erdogan H, Hershey JR. SDR–half-baked or well done?. n Proc. ICASSP 2019. IEEE; 2019. pp. 626–630.

[175] Reddy CK, Gopal V, Cutler R. DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality metric to evaluate Noise Suppressors. arXiv preprint arXiv:2010.15258; 2020..

[176] Hao X, Su X, Wen S, Wang Z, Pan Y, Bao F, Chen W. Masking and inpainting: A two-stage speech enhancement approach for low snr and non-stationary noise. n Proc. ICASSP 2020. IEEE; 2020. pp. 6959–6963.

[177] Li A, Liu W, Zheng C, Li X. Two Heads are Better Than One: A Two-Stage Complex Spectral Mapping Approach for Monaural Speech Enhancement. IEEE/ACM Trans Audio Speech Lang Process 2021;29:1829–43

[178] Tan K, Wang D. Learning complex spectral mapping with gated convolutional ecurrent networks for monaural speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 2020;28:380–90.

[179] Qin X, Zhang Z, Huang C, Dehghan M, Zaiane OR, Jagersand M. U2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognition 2020; 106:107404.

[180] Li A, Peng R, Zheng C, Li X. A supervised speech enhancement approach with esidual noise control for voice communication. Appl Sci 2020;10(8):2894.

[181] R. Xu, R. Wu, Y. Ishiwaka, C. Vondrick, and C. Zheng, "Listening to Sounds of Silence for Speech Denoising," arXiv:2010.12013 [cs, eess], Oct. 2020, Accessed: Dec. 20, 2021. [Online]. Available: http://arxiv.org/abs/2010.12013.

[182] W. Liu, A. Li, Y. Ke, C. Zheng, and X. Li, "know your enemy, know yourself: A unified two-stage framework for speech enhancement," in Interspeech 2021, Aug. 2021, pp. 186–190. doi: 10.21437/Interspeech.2021-238.

[183] C.Zheng,X.Peng,Y.Zhang,S.Srinivasan,andY.Lu,"Interactive Speech and Noise Modeling for Speech Enhancement," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 16, no. 35, pp. 14549–14557, 2021.

[184] J. Meyer, L. Rauchenstein, J. D. Eisenberg, and N. Howell, "Artie bias corpus: An open dataset for detecting demographic bias in speech applications," in Proc. 12th Lang. Resour. Eval. Conf., 2020, pp. 6462–6468.

[185] Yuan, J., & Bao, C. (2019). Cyclegan-based speech enhancement for the unpaired raining data. In Proceedings of Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC) (pp. 878–883). IEEE

[186] Alamdari, N., Azarang, A., & Kehtarnavaz, N. (2021). Improving deep speech denoising by Noisy2Noisy signal mapping. Applied Acoustics, 172, Article 107631.

[187] Fujimura, T., Koizumi, Y., Yatabe, K., & Miyazaki, R. (2021). Noisy-target raining: A training strategy for DNN-based speech enhancement without clean speech. In Proceedings of european signal processing conference (EUSIPCO) (pp. 436–440). IEEE.

[188] Erdogan, H., Hershey, J. R., Watanabe, S., & Le Roux, J. (2015). Deep recurrent networks for separation and recognition of single-channel speech in non-stationary background audio. Speech Communication, 78, 36-51.

[189] Leglaive, S., Girin, L., & Horaud, R. (2019). Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix actorization. In Proceedings of international conference on acoustics, speech and signal processing (ICASSP) (pp. 101–105). IEEE.

[190] Wang, Y., & Chen, X. (2018). Supervised speech separation based on deep earning: An overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(10), 1702-1726.

[191] Vincent, E., Gribonval, R., & Fevotte, C. (2018). Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing, 14(4), 1462-1469.

[192] Chen, J., & Wang, Y. (2017). Long short-term memory for speaker generalization n supervised speech separation. Journal of Acoustical Society of America, 141(6), 4705-4714.

[193] Seki, S., Takada, M., & Toda, T. (2020). Semi-supervised self-produced speech enhancement and suppression based on joint source modeling of air-and body-conducted

signals using variational autoencoder. In Proceedings of conference of the international speech communication association (INTERSPEECH) (pp. 4039–4043). IEEE

[194] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, “Wavenet: A generative model for raw audio,” in Arxiv, 2016.

[195] J. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2242-2251.

[196] Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In NIPS, 2014.

[197] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang and A. A. Efros, "Learning Dense Correspondence via 3D-Guided Cycle Consistency," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 117-126.

[198] Sivaraman, A., Wisdom, S., Erdogan, H., & Hershey, J. R. (2021). Adapting speech separation to real-world meetings using mixture invariant training. In Proceedings of international conference on acoustics, speech, and signal processing conference proceedings (ICASSP) (pp. 686–690). IEEE.

[199] Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., et al. 2018). Noise2noise. In Proceedings of international conference on machine earning (ICML). PMLR.

[200] Alamdari, N., Azarang, A., & Kehtarnavaz, N. (2021). Improving deep speech denoising by Noisy2Noisy signal mapping. Applied Acoustics, 172, Article 107631.

[201] Hao, X., Xu, C., & Xie, L. (2023). Neural speech enhancement with unsupervised pre-training and mixture training. Neural Networks, 158, 216-227.

[202] Rothauser E H, Chapman W D, Guttman N, et al. IEEE recommended practice for speech quality measurements [J]. IEEE Trans. Audio Electroacoust, 1969,17(3): 225-246.

[203] Gao, S. H., Cheng, M. M., Zhao, K., Zhang, X. Y., Yang, M. H., & Torr, P. (2019). Res2net: A new multi-scale backbone architecture. IEEE transactions on pattern analysis and machine intelligence, 43(2), 652-662.

[204] Van Gansbeke, W., Vandenhende, S., Georgoulis, S., & Gool, L. V. (2021). Revisiting contrastive methods for unsupervised learning of visual epresentations. Advances in Neural Information Processing Systems, 34, 16238-16250.

[205] Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., et al. 2018). Noise2noise. In Proceedings of international conference on machine earning (ICML). PMLR.

[206] Snyder, D., Chen, G., & Povey, D. (2015). Musan: A music, speech, and noise corpus. arXiv preprint arXiv:1510.08484.

[207] W. Shi et al., “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in IEEE Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1874–1883.

[208] Li, A., Zheng, C., Zhang, L., & Li, X. (2022). Glance and gaze: A collaborative earning framework for single-channel speech enhancement. Applied Acoustics, 187, 108499.

[209] S.-W. Fu, C.-F. Liao, Y. Tsao and S. D. Lin, “MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement,” in nternational Conference on Machine Learning. PMLR, 2019, pp. 2031–2041.

[210] Gao, S. H., Cheng, M. M., Zhao, K., Zhang, X. Y., Yang, M. H., & Torr, P. (2019). Res2net: A new multi- scale backbone architecture. IEEE transactions on pat- tern analysis and machine intelligence, 43(2), 652-662.

[211] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” IEEE Transactions on Audio, Speech and Language Processing, vol. 16, no. 1, pp. 229–238, 2008.

[212] Kashyap, M. M., Tambwekar, A., Manohara, K., & Natarajan, S. (2021). Speech denoising without clean training data: a Noise2Noise approach. In Pro- ceedings of conference of the international speech communication association INTERSPEECH) (pp. 2716–2720). IEEE.

[213] Fujimura, T., Koizumi, Y., Yatabe, K., & Miyazaki, R. (2021). Noisy-target raining: A training strategy for DNN-based speech enhancement without clean speech. In Proceedings of european signal processing conference (EUSIPCO) (pp. 436–440). IEEE.

[214] Ren, S., Zhou, D., He, S., Feng, J., & Wang, X. (2022). Shunted self-attention via multi-scale token aggregation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10853-10862).

[215] Yang, J., Shen, X., Xing, J., Tian, X., Li, H., Deng, B., ... & Hua, X. S. (2019). Quantization networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7308-7316).

[216] Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. International Journal of Computer Vision, 129(6), 1789-1819.