Advanced Deep Neural Networks for Speech Separation and Enhancement

by

Yang Xian

A doctoral thesis submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy (PhD), from Newcastle University.

June 2021

Intelligent Sensing and Communications Research Group (ISC),
School of Engineering,
Newcastle University,
Newcastle upon Tyne, UK, NE1 7RU.

CERTIFICATE OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this thesis, that the original work is my own except as specified in acknowledgements or in footnotes, and that neither the thesis nor the original work contained therein has been submitted to this or any other institution for a degree.

..... (Signed)

..... (candidate)

I dedicate this thesis to my loving family.

Abstract

Monaural speech separation and enhancement aim to remove noise interference from the noisy speech mixture recorded by a single microphone, which causes a lack of spatial information. Deep neural network (DNN) dominates speech separation and enhancement. However, there are still challenges in DNN-based methods, including choosing proper training targets and network structures, refining generalization ability and model capacity for unseen speakers and noises, and mitigating the reverberations in room environments. This thesis focuses on improving separation and enhancement performance in the real-world environment.

The first contribution in this thesis is to address monaural speech separation and enhancement within reverberant room environment by designing new training targets and advanced network structures. The second contribution to this thesis is on improving the enhancement performance by proposing a multi-scale feature recalibration convolutional bidirectional gate recurrent unit (GRU) network (MCGN). The third contribution is to improve the model capacity of the network and retain the robustness in the enhancement performance. A convolutional fusion network (CFN) is proposed, which exploits the group convolutional fusion unit (GCFU).

The proposed speech enhancement methods are evaluated with various challenging datasets. The proposed methods are assessed with the state-of-the-art techniques and performance measures to confirm that this thesis contributes novel solutions.

Contents

Statement of Originality

The contributions of this thesis are mainly on the improvement of the speech separation and enhancement with DNNs. The novelty of the contributions is supported by three journal publications and six other publications in the leading conferences in signal processing.

In the first contribution, the direct path ratio mask is proposed to address the speech separation and enhancement for the reverberant room environment. DRM is calculated by using the geometric information of the target speaker and microphone. Moreover, another method utilizes the parallel LSTMs to capture interdependency among the past and current frames. Two LSTMs are employed to estimate the dereverberant mask and ideal ratio mask , respectively. Publications related to this contribution have been accepted and submitted to:

- **Y. Xian**, Y. Sun, J. A. Chambers and S. M. Naqvi,"Geometric information based monaural speech separation using deep neural network", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

- **Y. Xian**, Y. Sun, W. Wang, and S. M. Naqvi, "Monaural speech enhancement based on two stage long short-term memory networks", *IEEE International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2019.

In the second contribution, a multi-scale feature recalibration convolutional encoder-decoder is proposed for single channel speech enhancement. The multi-scale convolutional layers utilize the kernel with varied sizes to capture features in different scales. Moreover, BGRU layers extract the interdependency among past, current, and future temporal frames. Also, the fully connected and bottleneck layers are introduced to improve the

parameter efficiency of the proposed method. Publications related to this contribution have been accepted and submitted to:

- **Y. Xian**, Y. Sun, W. Wang and S. M. Naqvi, "Multi-scale residual Convolutional Encoder decoder with bidirectional long short-term memory for single-channel speech enhancement", *IEEE European Signal Processing Conference (EUSIPCO)*, 2020.

- **Y. Xian**, Y. Sun, W. Wang, and S. M. Naqvi, "A Multi-Scale Feature Recalibration Network for End-to-End Single Channel Speech Enhancement", *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 143-155, 2021

In the third contribution, a convolutional fusion network is proposed for single channel speech enhancement. The outputs of standard convolution and depthwise separable convolution are concatenated to form the fusion output. Also, the channel shuffle is applied to improve the channel interdependency. Moreover, intra skip connections inside encoder/decoder are employed to enhance the feature reuse. Publication related to this contribution have been accepted and submitted to:

- **Y. Xian**, Y. Sun, W. Wang, and S. M. Naqvi, "Convolutional Fusion Network for Speech Enhancement", Elesiver Journal of Neural Networks, 2021

Other related publications:

- **Y. Xian**, Y. Sun, W. Wang and S. M. Naqvi, "Two stage audio-visual speech separation using multimodal convolutional neural networks", *IEEE Sensor Signal Processing for Defence (SSPD) Conference*, 2019.

- Y. Sun, **Y. Xian**, P. Feng, J. A Chambers and S. M. Naqvi, "Estimation of the number of sources in measured speech mixtures with collapsed

Gibbs sampling", *IEEE Sensor Signal Processing for Defence (SSPD) Conference*, 2017.

- Y. Sun, **Y. Xian**, W. Wang, and S. M. Naqvi, "Monaural source separation in complex domain with long short-term memory neural network", *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 359-369, 2019

- Y. Sun, **Y. Xian**, W. Wang, and S. M. Naqvi, "Single-channel speech enhancement with sequentially trained DNN system", *IEEE International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2019.

Acknowledgements

I am incredibly thankful to my supervisor Dr. Syed Mohsen Naqvi for his patient, constant help, consistent support and guidance during my PhD study period. I have learned much from his knowledge and instruction. His advice gives the direction of my research, and his exceptional experience helps me overcome many difficulties in research. It is my great honor to have been one of his research students.

I am also profoundly thankful to Prof. Wenwu Wang, Prof. Jonathon Chambers and Prof. Satnam Dlay. Their knowledge and enthusiasm help me to improve my ability in research, writing and presentation.

I want to express my greet to friends and families. Their constant support gives me great power to face any challenges in the last four years. Their advice encourages me to focus on my research. I would like to dedicate this thesis to my family.

Yang Xian

June, 2021

List of Acronyms

AMS	Amplitude Modulation Spectrum
BGRU	Bidirectional Gate Recurrent Units
CASA	Computational Auditory Scene Analysis
CED	Convolutional Encoder Decoder
CFN	Convolutional Fusion Network
CNN	Convolutional Neural Network
CRN	Convolutional Recurrent Network
DRM	Direct-path Ratio Mask
DM	Dereverberation Mask
DNN	Deep Neural Network
DRNN	Deep Recurrent Neural Network
EM	Expectation Maximization
FC	Fully Connected
GAN	Generative Adversarial Network
GCFU	Group Convolutional Fusion Unit
GMM	Gaussian Mixture Model

GRU Gated Recurrent Unit

IBM Ideal Binary Mask

ICA Independent Component Analysis

ILD interaural level difference

IPD interaural phase difference

IRM Ideal Ration Mask

IVA Independent Vector Analysis

LSTM Long Short-Term Memory

MCARE Multi-resolution Convolutional Auto-encoder

MCFR Multi-scale Feature Recalibration

MCGN Multi-scale Feature Recalibration Convolutional Bidirectional GRU Network

MESSL Model-based Expectation-maximization Source Separation and Localization

MFCC Mel-Frequency Cepstral Coefficients

MSE Mean-Square Error

MOS Mean Opinion Score

NFL No Free Lunch

NMF Non-negative Matrix Factorization

pdf probability density function

PESQ Perceptual Evaluation of Speech Quality

PLP Perceptual Linear Prediction

RASTA	Relative Spectral Transform
RELU	Rectified Linear Unit
RIR	Room Impulse Response
RNN	Recurrent Neural Network
SDR	Signal-to-Distortion Ratio
SEGAN	Speech Enhancement Generative Adversarial Network
SNR	Signal-to-Noise Ratio
STFT	Short-Time Fourier Transform
STOI	Short-Time Objective Intelligibility
T-F	Time-Frequency

List of Symbols

$\|.\|$	Absolute value
$f(.)$	Activation function
β	Attenuation rate
$const.$	Constant
$*$	Convolution operator
D	Depth-wise filter
det	Determinant operator
d_sm	Distance between speech and microphone
$\odot$	Element-wise Multiplication
E	Expectation
X_M	Input of multi-scale feature recalibration layer
$\mathcal{KL}$	Kullback-Leibler divergence
$kurt$	Kurtosis
$G_\theta(.)$	Mapping relation
max	Maximum value
$\|\|.\|\|$	Norm operation

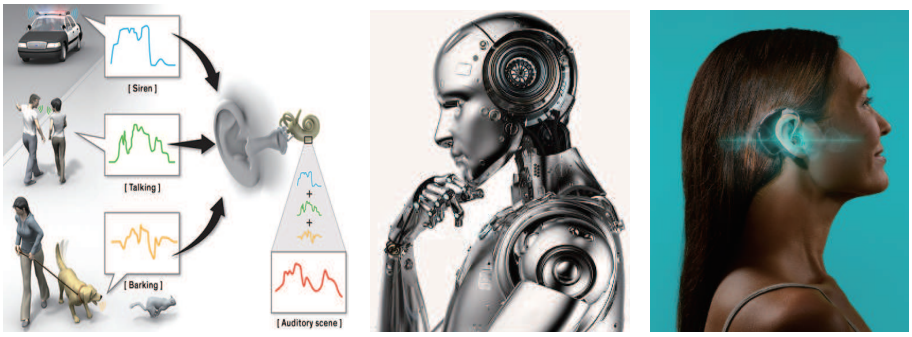
f_s	Sample frequency
S	Spectrogram of the target speech
C_s	Sound velocity
N	Spectrogram of the noise
H	Spectrogram of the room impulse response
Y	Spectrogram of the speech mixture
ν	Standard random Gaussian variable
$\times$	Times operator
β	Tunable parameter to scale the mask
$z_{w,b}$	Output of neuron
D_M	Output of the multi-scale feature recalibration layer
P	Point-wise filter
τ	Propagation time
δ	Unit impulse
C	2-D convolutional output

Chapter 1

INTRODUCTION

1.1 Monaural Speech Separation and Enhancement

"One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others," which is defined as cocktail party problem [1]. It describes there are several speakers are speaking simultaneously, the speech from the particular speaker needs to be separated, which is a common ability for human beings. However, there is a challenge for the machines, because the intelligibility and quality of the speech signal captured in a real acoustic scene are often degraded by noise and interfering sound present in the surrounding environment. Therefore, speech separation and enhancement aim to design machines and algorithms to recover the target speech by removing the background noise and interfering sound from the noisy speech mixture.



(a) (b) (c)

Figure 1.1. Different speech enhancement application contexts, (a) noisy environment [2]; (b) robotic [3]; (c) hearing aids [2]

.

Such a problem can be founded in many real-world applications such as mobile communication, speech recognition, hearing aids and robotics [4–9]. All these applications require the cleared target speech from the noisy speech mixture, which is related to remove the speech or non-speech noises [10, 11]. Therefore, speech separation and enhancement are essential for detection, recognition, which are important front techniques for most speech processing applications. According to the number of microphones (i.e recorded noisy speech mixture), the speech enhancement problem is categorized as multichannel, binaural and monaural (i.e. single channel) [7].

The statistical signal processing methods were first introduced to address speech separation and enhancement for multichannel and binaural cases [12], such as statistical signal processing based methods [13–15], and computational auditory scene analysis (CASA) based methods [16, 17]. The minimum mean-square error (MMSE) based estimator has been introduced for speech enhancement by modeling the speech and noise spectral components as statistically independent Gaussian random variables [18]. The CASA methods are designed for auditory scene analysis by computational means, they aim to imitate the hearing system of human beings, which utilize no more than two microphones to recording the noisy speech mixture. Therefore, it is widely used in hearing aid [17]. For example, In CASA based model-based expectation-maximization source separation and localization (MESSL) algorithm [19], the spatial features are modelled by a Gaussian mixture model, whose parameters are estimated using the expectation maximum algorithm, and then used to derive time-frequency masks for separating the target speech. Statistical approaches mainly focus on statistical modelling of spatial, spectral, or temporal features derived from the sensor signals, while CASA based approaches use computational models of human hearing to separate target speech from sound mixtures [20].

In real life situations, since the availability of limited number of micro-

phones and the distance between them are a major constraint, the multichannel and binaural cases always restrict enhancement performance. There is an extreme case that the noisy mixtures are recorded by a single microphone. Moreover, the target speech, noise interference and transmission paths are unavailable, only the noisy mixture is available. Therefore, the monaural i.e. single channel speech enhancement show more significant potential over the multichannel and binaural cases in real-world applications.

The DNNs based methods are the dominating recent research areas by the community and offer state-of-the-art performance in monaural speech enhancement [21, 22]. The DNNs based speech enhancement methods can be categorized as mapping-based and masking-based methods [23–26]. The DNNs often take time-frequency (T-F) representations of the noisy speech mixture obtained by a time-frequency analysis tool, such as short-time Fourier transform (STFT), as inputs, and train a neural network model to output the estimate of the target speech directly (i.e. mapping method) or the T-F mask. For the masking methods, the target speech can be separated by multiplying the spectrogram of the noisy speech mixture with the T-F mask which is a matrix of weights representing the occupation probability of the source in the noisy mixture at each T-F point [27]. Both binary or soft masks have been considered in the literature, with the ideal binary mask (IBM), and ideal ratio mask (IRM) proposed to benchmark the performance of the T-F masks based speech separation and enhancement systems [28,29]. Recent work shows that the mapping method show greater advantages over the masking method [30]. Moreover, plenty of works have designed new frameworks to improve the robustness of DNNs. The skip connection between the input layer and the output layer has also been incorporated in the DNN model, leading to the S-DNN [31] method, offering improvements over the DNN. In [32], the separation and acoustic models are jointly trained, incorporating additional hidden layers with fixed weights into a DNN.

Although the DNN has dominated the development of single channel speech enhancement, several disadvantages still limit the enhancement performance of DNN frameworks. Since inappropriate training targets and network structures, the speech components are underestimated or overestimated in reverberant environment. Furthermore, the DNN-based methods mainly utilize the current temporal frames to estimate, which underestimates the interdependency among different structures. Moreover, the standard DNN framework often captures the feature on a fixed scale due to the fixed filter size. Besides, in conventional DNN-based methods, the network is constructed by using fully connected layers. The robustness and model capacity may need further improvement. Therefore, a model with a combined structure would be preferable.

1.2 Aims and Objectives

This thesis aims to overcome and mitigate the aforementioned drawbacks of DNN-based speech separation and enhancement methods and improve prediction accuracy. More specifically, the detailed objectives are listed below.

- Objective 1: Contribute to improve the separation and enhancement performance in reverberant environment by exploiting advanced training targets and network structures based on spatial and temporal information.

In Chapter 3, in the first solution, the direct-path impulse response is estimated by using geometric (i.e. spatial) information of the target speaker and microphone. Then, the reflection and noise is removed by using direct-path ratio mask, which is estimated by using direct-path impulse response. In the second solution, parallel long short-term memory networks (LSTMs) are introduced to capture the interdependency (i.e. temporal information) among the past and current temporal frames. Moreover, they are exploited

to estimate dereverberation mask (DM) and IRM, respectively. The reverberations and noises are removed by jointly using DM and IRM.

- Objective 2: Contribute to improve enhancement performance by capturing the feature in different scales, and using multi-scale feature and interdependency.

In Chapter 4, a novel framework is proposed that consists of multi-scale encoder-decoder with bidirectional gate recurrent units (BGRU). Multi-scale encoder-decoder offers features on different scales, and BGRU layers capture the interdependency among the past, current and future temporal frames.

- Objective 3: Contribute to improve generalization ability, model capacity and enhancement performance by employing convolutional fusion encoder-decoder.

In Chapter 5, the depth-wise separable convolution/deconvolution and standard convolution/deconvolution are exploited to build fusion encoder-decoder, which provides better generalization ability and higher parameter efficiency.

1.3 Thesis Outline

The outline of this thesis is listed as follows:

Chapter 2 offers a relevant literature review of speech enhancement by using relevant of deep learning methods. The advantage and disadvantages of each method are also stated. Furthermore, the challenges, that include generalization ability, model capacity and parameter efficiency, associated with these methods are also discussed. Moreover, two feasible directions are offered to these challenges i.e. network framework and training targets.

Chapter 3 provides two solutions for reverberant speech separation and enhancement. In the first solution, a new DNN training target is proposed

to improve the performance in reverberant and noisy room environments, which incorporates geometric information describing the target speaker and microphone. In the second solution, a two-stage approach using LSTM networks is proposed. In the first stage, the dereverberation mask (DM) is obtained by using a trained LSTM, which aims to dereverberate the noisy speech mixture. In the second stage, the IRM is estimated by the second trained LSTM, which is utilized to separate the desired speech signal from the dereverberated speech mixture. The extensive experiments prove two solutions provide advantages over DNN-based baseline methods.

Chapter 4 proposes a multi-scale feature recalibration convolutional encoder decoder with bidirectional gated recurrent unit (BGRU) architecture for end-to-end single channel speech enhancement. The features in different scales are extracted by using the multi-scale feature recalibration 2-D convolutional layers, which efficiently utilize the local and contextual information in the signal. In addition, a feature recalibration network is designed by using a gating mechanism to control the information flow among the layers, enabling different weights to be applied on scaled data, which help to retain features from the target speech while suppressing features from noise. The fully connected layer (FC) is employed to compress the output of the multi-scale 2-D convolutional layer. The BGRU layers is employed to update the current frame, and to exploit the interdependency among the past, current and future frames and thereby improve predictions. The experimental results confirm the proposed MCGN method outperforms several state-of-the-art methods.

In Chapter 5 provides a new convolutional fusion network (CFN) for monaural speech enhancement by improving model performance, inter-channel dependency, information re-use and parameter efficiency. First, a new group convolutional fusion unit (GCFU) consisting of the standard and depth-wise separable CNN is used to reconstruct the signal. Second, the whole

input sequence (full information) is fed simultaneously to two convolution networks in parallel, and their outputs are re-arranged (shuffled) and then concatenated, in order to exploit the inter-channel dependency within the network. Third, the intra skip connection mechanism is used to connect different layers inside the encoder as well as decoder to further improve the model performance. Extensive experiments are performed to show the improved performance of the proposed method as compared with three recent baseline methods.

Finally, conclusions are drawn and future work is then discussed in Chapter 6.

Chapter 2

BACKGROUND METHODS

2.1 Introduction

In this chapter, the background methods relate to speech separation and enhancement are provided, and the discussion of related methods is offered. These methods include statistical signal processing, CASA and DNN based methods. Then, within the monaural case, brief overviews of DNN-based methods are provided, which include network structures and training targets. Furthermore, the limitations of these methods are described. Then, three performance measures and datasets are described. Finally, the chapter summary is provided.

2.2 Statistical Signal Processing based Methods

Speech separation and enhancement have drawn enormous attentions due to the increasing demand of speech-related applications such as mobile communication and robotics. As mentioned in Chapter 1, the statistical signal processing, such as independent component analysis (ICA) and independent vector analysis (IVA) are proposed to solve the over-determined and determined speech separation and enhancement.

2.2.1 Independent Component Analysis

ICA models the observed data as a linear combination of underlying latent variables [33, 34], which assumes each component is statistically independent with other components. Moreover, this model is instantaneous and the time delay is neglected. For the several independent components, the joint probability density function (PDF) is as followed:

$$p(s_1, s_2, s_3 \cdots) = p_1(s_1)p_2(s_2)p_3(s_3)\cdots \tag{2.2.1}$$

where $p_1(s_1)$, $p_2(s_2)$, $p_3(s_3)$ denote PDF of three independent components. For the noisy speech mixture, there is at least one speech have non-Gaussian distribution. And, the unknown mixing matrix is assumed to be invertible.

Moreover, at least one source must have non-Gaussian distribution and the unknown mixing matrix is assumed to be invertible, in which the number of sources is equal or less than the number of mixtures. According to the central limit theorem, any mixture of components will become more Gaussian than the individual components. Thus, separating the target speech is realized by maximization of non-Gaussianity [35]. Besides, the non-Gaussianity is measured by negentropy [36]. Although the ICA provides the feasible solution for speech separation and enhancement, the permutation and scaling problem limited the enhancement performance [37].

2.2.2 Independent Vector Analysis

The main cause of the permutation problem of ICA is the under-estimated inter-independent of components. Therefore, the IVA is introduced to mitigate the permutation problem using the inter-frequency dependencies in the desired speech signals. The multivariate score function is also applied to describe the source prior [38–40], which is the higher-order frequency dependency. There are two important assumptions of IVA. Firstly, each source

is independent with other sources. Secondly, elements of source maintain dependency with other elements. Therefore, the IVA maintains the inter-source dependency by introduce approximated pdfs of individual source vectors $\prod_{i=1}^{L} q(\widehat{\mathbf{s}}_\mathbf{i})$. Mathematically, The cost function and core function of IVA reserve the inter-frequency dependency [38].

$$\mathcal{J} = KL\left(p(\widehat{\mathbf{s}}_\mathbf{1}\dots\widehat{\mathbf{s}}_\mathbf{L})\|\prod_{i=1}^{L} q(\widehat{\mathbf{s}}_\mathbf{i})\right) = \int p(\widehat{\mathbf{s}}_\mathbf{1}\dots\widehat{\mathbf{s}}_\mathbf{L})\frac{p(\widehat{\mathbf{s}}_\mathbf{1}\dots\widehat{\mathbf{s}}_\mathbf{L})}{\prod_{i=1}^{L} q(\widehat{\mathbf{s}}_\mathbf{i})} d\widehat{\mathbf{s}_1}\dots d\widehat{\mathbf{s}_L}$$

$$= \text{const.} - \sum_{k=1}^{K}\log|\det G^{(k)}| - \sum_{i=1}^{L}\int E\{\log(q(\widehat{\mathbf{s}}_\mathbf{i}))\} \tag{2.2.2}$$

where $\mathbf{s}_i = \left[s_i^{(1)}, s_i^{(2)}, ..., s_i^{(K)}\right]^T$ denotes the ith estimated separated source, K represents the kth frequency bin. KL represents Kullback-Leibler (KL) divergence, it is used to measure the difference between one probability distribution and a reference probability distribution.

And the third term of (2.2.2) represents entropy. Then the gradient descent is applied to minimize the cost function. By using this cost function, the dependency between sources are removed but the interdependency of each source is preserved. Although the IVA offers better performance when compared with the ICA, it can only address the determined problem, which means the number of sensors is at least equal to number of sources. Thus, the MESSL algorithm is applied to solve the under-determined case.

However, the separation performance of these methods cannot be improved even the data amount of the speech signals is increased. Therefore, the machine learning and the deep learning algorithms are introduced.

2.3 CASA based Methods

According to Section 2.2, the statistical signal processing based methods are proposed to address the overdetermined and determined speech sepa-

ration and enhancement, which means the number of sources is no more than the number of microphones. CASA methods are introduced to address determined (binaural) speech separation and enhancement [16, 17, 41]. The most well known CASA method is MESSL algorithm, which mainly focus on time-frequency analysis.

In MESSL method, two microphones located in different positions are used to record the noisy speech mixtures. Two binaural cues i.e. interaural phase difference (IPD) and interaural level differences (ILD) are modelled as the Gaussian distributions, respectively. According to W-disjoint Orthogonality theory, only one source is active at each T-F point [42]. More specifically, the IPD and ILD are exploited to build sources' probabilistic model, which is employed to evaluate the hidden variable at each T-F point. The expectation maximization (EM) algorithm is applied to optimized the expectation of hidden variables until convergence. The estimated mask is then multiplied with the spectrogram of noisy speech mixture to generate the estimated sources.

The above methods offer competitive performance in binaural and multi-channel speech enhancement and separation, which can be further used in multi-channel speech recognition and hearing aid [17]. Nevertheless, in single channel speech enhancement, the spatial information is unavailable, as a result, the above methods cannot be used to estimate the target speech.

2.4 Problem Statement of Monaural Speech Enhancement

In speech enhancement, there is an extreme case i.e. monaural speech enhancement, the only one microphone is used to record the noisy speech mixture. As a result, the transmission path and mixing process are unknown, which means the spatial information is unavailable. Therefore, the above methods designed for over-determined and determined cases are not feasi-

ble for monaural speech enhancement, and many methods are proposed to address the monaural issue. The minimum mean-square error (MMSE) estimator realizes speech enhancement by modelling the speech and noise spectral components as statistically independent Gaussian random variables [5]. Also, the non-negative matrix factorization (NMF) is employed to decompose the magnitude and power spectrum of the noisy speech mixture [43]. Furthermore, the weighted sums of non-negative target speech are utilized to model the noisy speech mixture [44].

Mathematically, in monaural i.e. single channel speech enhancement, the noisy speech mixture can be written as:

$$y(m) = s(m) + n(m) \tag{2.4.1}$$

where $y(m)$ denotes the noisy speech, $s(m)$ and $n(m)$ represent the clean speech signal and noise at discrete time m, respectively. If noise is speech signal, it is speech separation. For environment noises, it is speech enhancement. Using STFT, the noisy speech mixture at time frame t and frequency bin f is represented as:

$$Y(t, f) = S(t, f) + N(t, f) \tag{2.4.2}$$

where $S(t, f)$ and $N(t, f)$ are the STFT of the clean speech signal and noise, respectively.

Recently, DNN techniques attract the researchers' attention in monaural speech enhancement. In DNN-based algorithms, the desired speech signal is obtained from the trained neural network model [45], a supervised learning algorithm. In Section 2.5, the structures and training targets of the DNN-based algorithm are discussed firstly, then the different relevant solutions for each of them are reviewed.

2.5 Deep Neural Network based Methods

2.5.1 Network Structure

Meanwhile, the DNNs show great potential for signal processing problems, e.g. speech recognition, speech separation and enhancement [46]. The DNNs imitate human beings, learn information from the training data structure, and make predictions based on the testing data and learned information. Thus, the DNNs are introduced to address the speech enhancement problem, and they play a role as black-box, learning the masking and mapping relation between the target speech and noisy speech mixture [25, 47]. There are two essential aspects of DNN-based methods, and they are network structures and training targets. In this section, we will provide a brief review of them.

A DNN model is constructed by three kinds of layers: the input layer, hidden layer, and output layer, and these layers contain numerous neurons. The core of DNNs is the neuron that is a fundamental computational unit, and the followed diagram can represent the single neuron.

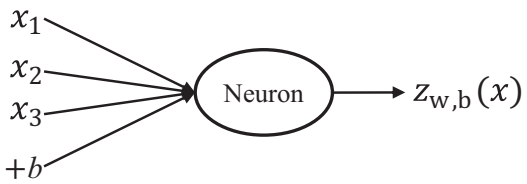


Figure 2.1. Typical structure of single neuron

The "plus one " is called intercept term. x_1, x_2, x_3 are inputs, $z_{w,b}(x)$ is the output. Mathematically, this process can be written as:

$$z_{w,b}(x) = f(\mathbf{w}x) = f\left(\sum_{i=1}^{3} w_i x_i + b\right) \tag{2.5.1}$$

where $f(\cdot)$ is called the activation function, w_i denotes the weight of ith input.

For each neuron, there are plenty of activation functions, and they have

different characteristics. The widely used activation functions are summarized as below.

- Sigmoid: $f(a) = \dfrac{1}{1+e^{-a}}$
- Hyperbolic tangent: $f(a) = tanh(a) = \dfrac{1-e^{-2a}}{1+e^{-2a}}$
- Softmax: $f(a) = \dfrac{e^{a_i}}{\sum_j e^{a_j}}$. And $\sum_i f_i(a) = 1$ and $f_i(a) > 0$
- Softplus: $f(a) = \zeta(a) = log(1+e^a)$
- Absolute value rectification: $f(a) = |a|$ [48]
- Hard tanh: $f(a) = max(-1, min(1, a))$. [49].
- Rectified Linear Unit(ReLU): $f(a) = max(0, a)$
- Leaky ReLU: $f(a) = max(a, ka)$, and $0 < k < 1$.

Combine together many neurons, the neural networks can be built as shown below.

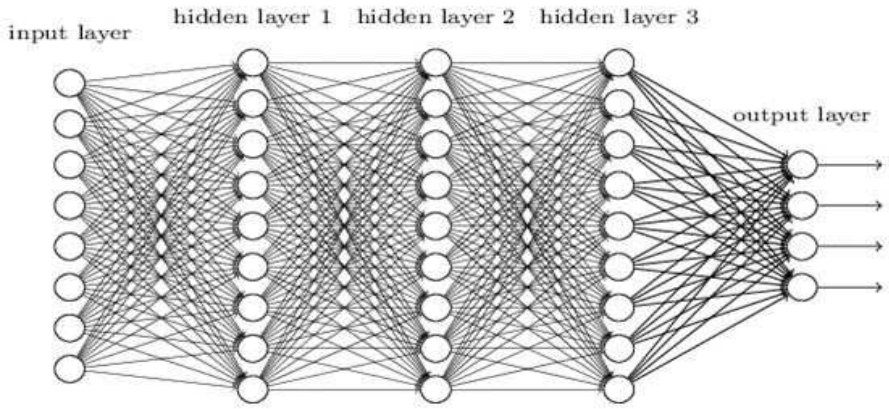


Figure 2.2. Example of DNNs

The DNNs are capable to fit the relationship between the training target and the output by adjusting the inter parameters i.e. weights and bias. The loss function is employed to measure the difference between the training target and the output of DNNs, which is minimized by using gradient decent algorithm. Finally, the optimized model is selected.

2.5.2 DNN-based Mapping methods

Unlike the traditional methods, the DNNs are exploited to learn the training targets include mapping and masking relations [25,50]. Mathematically, the neural network model is trained to find the mapping relation G_θ between the magnitude spectrum of the clean speech signal $|S(t,f)|$ and the noisy speech mixture $|Y(t,f)|$. The mapping function is estimated by optimizing the loss function as:

$$Loss_{mapping} = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}[G_\theta(|Y(t,f)|) - |S(t,f)|]^2$$

$$= \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}(|\hat{S}(t,f)| - |S(t,f)|)^2 \tag{2.5.2}$$

where $|\hat{S}(t,f)|$ is the magnitude spectrum of the estimated target speech, which is combined with phase information of the noisy mixture to recover the target speech.

2.5.3 DNN-based Masking methods

For the DNN-based masking methods, similarly, the neural network model is trained to find the masking relation between the representation of target speech and noisy mixture.

$$Loss_{masking} = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}(\hat{M}(t,f) - M(t,f))^2$$

where the $\hat{M}(t,f)$ represents the estimated mask. More specifically, the mainly used masks are IBM [51] and IRM. The masks are multiplied with the spectrum of noisy speech mixture to generate the enhanced target speech.

The IBM can be represented as:

$$IBM(t,f) = \begin{cases} 1, & if \ \mathrm{SNR(t,f)} > LC \\ 0, & otherwise \end{cases} \tag{2.5.3}$$

where LC denotes the local criteria, and it is employed to determine the value of each T-F unit in IBM. If the value of IBM equal to one, it means this T-F point belongs to the target speech. Otherwise, this T-F point belongs to the noise. Therefore, IBM is associated hard decision, which causes information loss in speech enhancement [51]. To mitigate the information loss, the mask i.e. IRM associated soft decision is proposed [26]. The representation of IRM is shown as:

$$IRM(t,f) = \frac{\sqrt{S^2(t,f)}}{\sqrt{S^2(t,f) + N^2(t,f)}} \tag{2.5.4}$$

where $S^2(t,f)$ represents the energy of target speech, and $N^2(t,f)$ denotes the energy of noise. The IRM is the ratio between the energy of target speech and energy of the noisy speech mixture. As shown in (2.5.4), each T-F point of noisy mixture can be decided how much information from target speech by using the IRM, that ranges from 0 to 1. The experiments results prove the IRM outperforms the IBM.

2.5.4 Advanced Network Architecture

Apart from selecting a proper training target, the neural network architecture (structure) is also essential for speech enhancement. Plenty of researchers have proposed advanced architectures that offer varied advantages in signal processing. Meanwhile, many advanced architectures are introduced to address speech separation and enhancement problems.

Conventional DNNs often consider the local temporal frames, the temporal information is not well utilized, which is vital to capture interdependency

among different frames. As a result, the conventional DNNs are less effective in generalization to mismatch conditions such as speaker independent and noise independent cases. The context window is proposed to utilize the temporal information, which feeds several temporal frames to the DNNs and estimates the single frame. However, the larger size would increase the computational cost. Furthermore, the recurrent neural network (RNN) has been introduced to address speech enhancement problems to better utilize temporal information. In RNN, each neuron is connected with the neurons of last and same layers, which employs the past hidden state to update the current hidden state. Thus, the interdependency between the past and current temporal frames are extracted. In [52], the DRNN is employed to estimate target speech, and the discriminative term is used to optimize the objective function.

Although the DRNN can extract the interdependency among adjacent temporal frames, the temporal information with long-term interval is neglected. Therefore, the LSTM RNN is proposed to capture interdependency among long term interval [53, 54]. The LSTM exploits the cell memory to keep and memory the temporal information even with long term interval. Also, it uses the input, output and forget gates to control how much past information is used to update the current temporal frame. The evaluation proves the past information can improve the enhancement performance and LSTM outperforms the DNN. The details of LSTM based speech enhancement are discussed in Chapter 3.

Recently, many other network structures are introduced for speech enhancement. For example, inspired by the success of computer version [55,56], the convolution neural network is used to learn the masking or mapping relation in speech enhancement [57, 58]. Moreover, the generative adversarial network (GAN) is employed to estimate the target speech [59]. The GAN includes generator (G) and discriminator (D). The G aims to learn a map-

ping that can imitate the real data distribution. The D is a binary classifier, which is employed to classify the G's output is real or fake. By using adversarial training, the G is optimizing its parameter to fool the D and generate the final output. Although the above network structures provide completive enhancement performance, further improvement is desired.

2.5.5 Generalization Ability

The generalization ability of speech enhancement means the model can well estimate the target speech with unseen speakers and noises. The speech signal is highly random, related to the speaker's accent, gender, age, etc. In addition, environmental noises are countless. Therefore, it is impossible to train the neural network model with all kinds of noisy speech mixture. Meanwhile, since the computational resource limitation, the neural network model is trained with limited data. Consequently, a model with strong generalization can retain enhancement perform with unseen speakers and noises, which is positively related to the overfitting [60].

2.6 Research Challenges Associated with Monaural Speech Enhancement

Although DNNs show advantages over the conventional speech enhancement methods, they still have limitations and their performances need to be improved since the speech and noise signals are high randomnesses. Therefore, we summarize the challenges of DNN-based methods as below:

- Training target: selecting robustness training targets to refine the masking or mapping relationship between target speech and noisy speech mixture, which can help the network to improve prediction accuracy.
- Neural network framework: selecting proper network architectures and

hyperparameters, which utilize advantages of different network frameworks to improve parameter efficiency and enhancement performance.

- Generalization ability and model capacity: the network models need to offer strong generalization ability and model capacity to address the speech enhancement problem with varied unseen speakers and noises.

- Room environment: the recorded noisy speech mixtures also contain components of reverberations in room environment, which is generated by the reflections of wall, window, furnitures, etc. The reverberations increase the difficulty of speech enhancement. Consequently, DNNs based methods need to overcome the reverberation environment.

2.7 Performance Measures and Datasets

2.7.1 Performance Measures

Three measures are introduced to evaluate the experimental results of the enhanced speech signal. They are perceptual evaluation of speech quality (PESQ) [61], short-time objective intelligibility (STOI) [62] and signal to distortion ratio improvement (ΔSDR) [63].

For PESQ, the estimated speech and target speech are level aligned to a standard listening level. Then an input filter is used to model them to the standard telephone handset. The filtered speech signals are processed by an auditory transform, which is employed to estimate the distortion parameters from the transformed signals. Two distortions are aggregated in time frequency, and mapped to subjective mean opinion score (MOS). The PESQ ranges from -0.5 to 4.5, and the higher value indicates better enhancement performance.

The STOI is introduced to evaluate the intelligibility of speech quality by calculating the correlation coefficient between the temporal envelope of

the clean target speech and enhanced target speech within the short-time region. STOI is ranged from 0 to 1. The higher value of STOI means better intelligibility quality. More specifically,

$$STOI = \frac{1}{TF}\sum d_{t,f} \tag{2.7.1}$$

where $d_{t,f}$ is the sample correlation coefficient between estimated speech signal and desired speech signal, f is the fth frequency band, t is the time frame, F is the total number of frequency bands and T is the total number of time frames.

To further evaluate the performance, the SDR improvement is introduced to measure the distortion ratio. The enhanced speech is decomposed into four parts: target signal, error terms for interference, noise, artifacts. The SDR is calculated as:

$$SDR = 10log_{10}\frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2} \tag{2.7.2}$$

Based on the SDR of unprocessed noisy mixture and enhanced target speech, we can estimate the SDR improvement as:

$$\Delta SDR = SDR_{enhanced} - SDR_{mixture} \tag{2.7.3}$$

2.7.2 Datasets

In this thesis, different databases are exploited to generate the training and testing noisy mixtures. The speech signals are selected from TIMIT [64], IEEE [65], VCTK [66] databases. And noise signals are selected from NOISEX-92 [67], NON-speech Sound [68] and DEMAND [69] databases. The clean speech signals are mixed with noise signals to generate the training and testing datasets.

The TIMIT database includes 6300 utterances spoken by 630 female and male speakers. IEEE database contains 720 recording utterances from one male speaker. The VCTK database includes about 40000 utterances from 110 speakers, and each speaker read about 400 sentences. For the noises, the NOISEX-92 database includes 15 noises, such as human conversation and machine noises. The Non-speech Sound database offers 100 environment noises, and the DEMAND database provides 15 recorded noises. In total, over 40 noises are used to train the networks, and more than 15 noises are used test the networks.

2.8 Summary

This chapter provided a literature review of speech separation and enhancement methods. Firstly, we discussed the existing methods for over-determined and under-determined speech enhancement problems, and their advantages and disadvantages were provided. Then, the fundamental network structure of DNN-based speech enhancement methods was discussed. Furthermore, two commonly used training targets of DNN-based methods were stated. Then, the main challenges of the DNN-based methods were discussed. In summary, the requirements of the advanced speech enhancement methods were listed as:

- best possible estimation the target speech within noisy reverberant room environment.
- generalization ability for mismatch conditions that include unseen-speakers, unseen-noises and unseen signal-to-noise ratios.
- improvement in the model capacity and parameter efficiency of the speech enhancement system

In the next chapter, the DNN-based method with a direct-path ratio

mask is proposed to improve performance enhancement with the noisy reverberant room environment. Moreover, we also introduce the LSTM method to improve the generalization ability and enhancement performance.

Chapter 3

SPATIAL AND TEMPORAL INFORMATION BASED SPEECH SEPARATION AND ENHANCEMENT

3.1 Introduction

Recently, DNN dominates the development of speech separation and enhancement. The DNN-based methods can be categorized as mapping and masking methods. The time-frequency features of noisy mixtures are fed into DNN, and it is employed to learning mapping or masking relations between the target speech and noisy mixture. For DNN-based masking methods, two important masks are proposed. The IBM judges the belonging of each time-frequency point, this procession is a hard decision [28]. In addition, the soft decision is employed in the IRM, each T-F points is assigned by the ratio between the energy of target speech and noisy mixture [26], which shows advantages over the IBM.

However, the existing methods still have limitations for addressing the speech enhancement in the reverberated room environment. (1) The DNN-based masking methods offer limited generalization to the dereverberation

problem, and new training targets that can better reflect the clean speech and noise are still needed. (2) The vanilla DNN utilizes a window to capture temporal dynamics, which is insufficient for speaker characterization and speech separation [54]. The enhancement performance of these state-of-the-art methods needs to be improved within reverberant room environments for speaker independent case.

In this chapter, two different methods are proposed to solve the limitations above. First, Based on spatial information, the direct-path ratio mask (DRM) is proposed to realize the dereverberation and denoising simultaneously. The geometric i.e. spatial information is used to describe the target speaker and microphone to calculate the direct-path impulse response, which is used to estimate the direct-path speech. The DRM is proposed using direct-path speech, which improves performance in noisy and reverberant room environments. Second, the parallel LSTMs are introduced to capture the interdependency i.e. temporal information between the past and current temporal frames even among long-term interval, which firstly achieve dereverberation, then realize denoising. The long-term speech context is captured by the LSTM, which improves the robustness of the system. Two parallel LSTMs are used to estimate two different training targets. One of the LSTMs is used to estimate DM, and another LSTM is applied to estimate IRM. Then, both DM and IRM are integrated for speech enhancement.

The remainder of this chapter is organized as follows. Section 3.2 states the algorithm of DRM. Section 3.3 states the algorithm of parallel LSTMs method. Section 3.4 provides the experimental setup and evaluation of DRM. Section 3.5 offers the experimental setup and evaluation of parallel LSTMs. Section 3.6 draws conclusions.

This chapter focuses on the first objective of this thesis, which relate to new training targets and advanced structures based on spatial and temporal information for published two conference papers [70, 71].

3.2 Direct-path Ratio Mask

3.2.1 Mixture Model and Direct-path Impulse Response

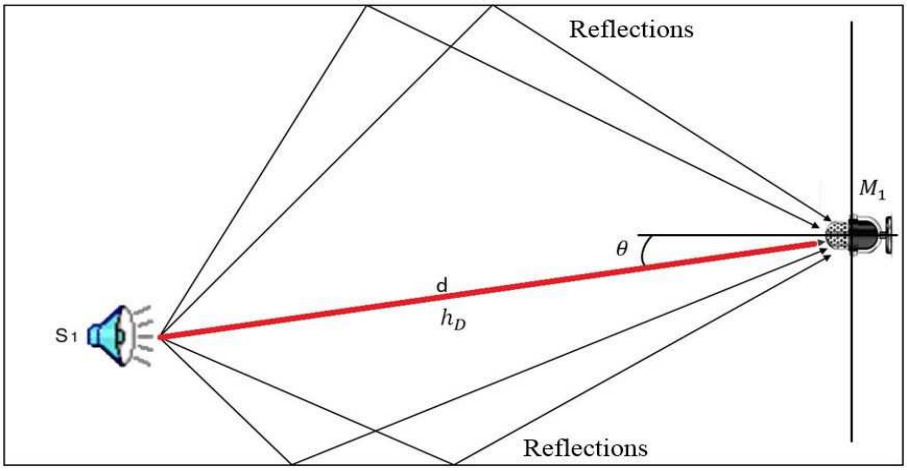


Figure 3.1. Monaural speech separation setup within a reverberant room environment, the distance and angle between the target speaker and sensor are shown.

The reverberant speech can be modelled as the convolution result of the speech source and impulse response as:

$$s_r(m) = s(m) * h_s(m) \tag{3.2.1}$$

where $*$ represents the convolution operator, $s_r(m)$ denotes the reverberant speech at discrete time m, $s(m)$ represents the speech source and $h_s(m)$ is the impulse response. The impulse response can be divided into the direct-path and reflections as:

$$h_s(m) = h_d(m) + h_a(m) \tag{3.2.2}$$

where $h_d(m)$ is the impulse response of the direct-path and $h_a(m)$ denotes the impulse response of reflections.

The geometric information provides the distance and bearing between the speech source and the microphone, which helps to estimate direct-path impulse response. The direct-path impulse response, as shown in Fig. 3.1,

is calculated as:

$$h_d(m) = \beta\delta(m-\tau) = \frac{\kappa}{d_{sm}^2}\cos\left(\frac{\theta}{\mathrm{r}}\right)\delta\left(\mathrm{m} - \frac{\mathrm{f_s}}{\mathrm{C_s}}\mathrm{d_{sm}}\right) \tag{3.2.3}$$

where β denotes the attenuation rate, δ represents the unit impulse, κ represents the attenuation per unit length in air, and d_{sm} is the distance between the speech source and microphone. The parameter θ represents the angle between the speech source and microphone, and r is the directionality coefficient. Besides, τ is the propagation time, f_s is the sample frequency, and C_s denotes the sound velocity in air.

Based on the distributive property of convolution, the reverberant speech can be represented as [72]:

$$\begin{aligned} s_r(m) &= s(m) * h_d(m) + s(m) * h_a(m) \\ &= s_d(m) + s_a(m) \end{aligned} \tag{3.2.4}$$

where $s_d(m)$ is the direct-path speech and $s_a(m)$ includes only reverberations. To simulate the real room environment, the mixture of reverberant speech with additional noises is provided as:

$$y_{ad}(m) = s_d(m) + s_a(m) + n(m) \tag{3.2.5}$$

where $n(m)$ denotes the noise at time m. Using the Fourier transform, (3.2.5) can be represented as:

$$Y_{ad}(t,f) = S_D(t,f) + S_A(t,f) + N(t,f) \tag{3.2.6}$$

where $Y_{ad}(t,f)$ denotes the mixture of reverberant speech and additional noise in time frame t and frequency bin f.

The DRM can be calculated as:

$$DRM(t,f) = \left(\frac{S_D^2(t,f)}{S_D^2(t,f) + N^2(t,f)} \right)^{\eta} \tag{3.2.7}$$

where $S_D^2(t,f)$ denotes the energy of the direct-path speech at time t and frequency frame f, and $N^2(t,f)$ is the energy of noise. And η is the tunable parameter to scale the mask. The proposed DRM is used as a training target, which requires less accuracy in the separation of noisy reverberant speech mixture, because the DRM mitigates reflections and noise. The direct-path impulse response based speech is estimated as:

$$\hat{S}_D(t,f) = Y_{ad}(t,f) DRM(t,f) \tag{3.2.8}$$

3.2.2 Speech Reconstruction

Since the DRM can only separate the direct-path signal from the noisy reverberant mixture, the speech reconstruction module is used to separate the desired speech source. At the testing stage of the speech reconstruction module, there are two inputs: (1) the estimated direct-path speech $\hat{S}_D(t,f)$ based on the DRM, (2) direct-path impulse response $H_D(t,f)$ based on geometric information. The frequency domain separated speech source is calculated as:

$$\hat{S}(t,f) = \left[\left(\hat{S}_D(t,f) \right) \left(H_D(t,f) \right)^{-1} \right] \tag{3.2.9}$$

Then, the time domain target speech can be obtained by using the inverse fast Fourier transform (IFFT) operation.

3.2.3 System Architecture

The system architecture is shown in Fig. 3.2. The geometric information of the target speaker and microphone for monaural speech separation can

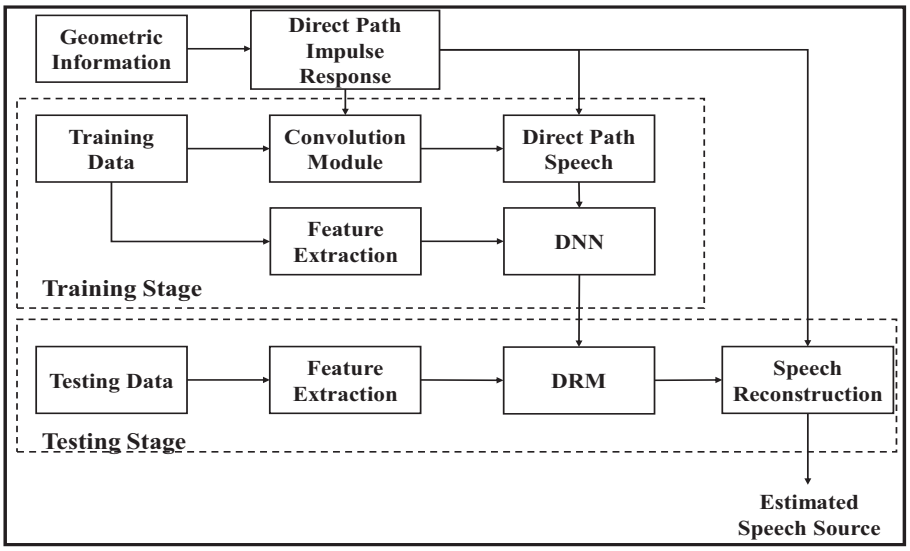


Figure 3.2. The block digram of the propose reverberant and noisy speech separation system.

be obtained from our multiple human tracking systems [73, 74], which are successfully used in multimodal binaural and overdetermined speech separation [6, 20]. At the training stage, the geometric information is applied to generate the proposed DRM and at the testing stage, the trained DNN with geometric information is used to estimate the final desired speech signal.

3.3 Parallel Long-short Term Memory

3.3.1 The Proposed Method

The reverberant speech mixture can be modelled as:

$$y_r(m) = s(m) * h_s(m) + n(m) * h_n(m) \tag{3.3.1}$$

Where $y_r(m)$ denotes the reverberant speech mixture at discrete time m, $*$ denotes the convolution operator, $s(m)$ and $n(m)$ represent the speech source signal and noise signal at time m, respectively. And $h_s(m)$ and $h_n(m)$ are impulse responses of speech signal and noise signal, respectively. Besides,

the noise can be background noise or speech interference signal. The spectra of reverberant speech mixture is obtained by using Fast Fourier Transform (FFT), and can be written as:

$$Y_r(t,f) = S(t,f)H_s(t,f) + N(t,f)H_n(t,f) \tag{3.3.2}$$

where $H_s(t,f)$ is the impulse response of clean speech signal, and $H_n(t,f)$ denotes the impulse response of noise signal both in frequency domain. $N(t,f)$ and $S(t,f)$ are the spectra of noise and clean speech signal, respectively. The dereverberanted speech mixture can be represented as:

$$Y(t,f) = S(t,f) + N(t,f) \tag{3.3.3}$$

According to (3.3.2) and (3.3.3), the reverberant speech mixture can be rewritten as:

$$Y_r(t,f) = Y(t,f)\left(\frac{H_s(t,f)}{1+\dfrac{N(t,f)}{S(t,f)}} + \frac{H_n(t,f)}{1+\dfrac{S(t,f)}{N(t,f)}}\right) \tag{3.3.4}$$

3.3.2 Training Targets

According to (3.3.4), the DM is expressed as [75]:

$$DM(t,f) = \left(\frac{H_s(t,f)}{1+\dfrac{N(t,f)}{S(t,f)}} + \frac{H_n(t,f)}{1+\dfrac{S(t,f)}{N(t,f)}}\right)^{-1} \tag{3.3.5}$$

By using the $DM(t,f)$, the reflections in the reverberant mixture are removed, the estimated dereverberanted speech mixture can be generated as:

$$\hat{Y}(t,f) = Y_r(t,f)DM(t,f) \tag{3.3.6}$$

The IRM is calculated as [26]:

$$IRM(t,f) = \left(\frac{S^2(t,f)}{S^2(t,f) + N^2(t,f)} \right)^{\eta} \tag{3.3.7}$$

where $S^2(t,f)$ is clean speech signal energy, and $N^2(t,f)$ is the noise energy. And η is the tunable parameter to scale the mask, and it is fixed to 0.5. According to (3.3.6) and (3.3.7), the estimated desired speech signal can be separated as:

$$\begin{aligned} \hat{S}(t,f) &= \hat{Y}(t,f) IRM(t,f) \\ &= Y_r(t,f) DM(t,f) IRM(t,f) \end{aligned} \tag{3.3.8}$$

Since the DM ranges from 0 to $+\infty$, which is not consistent with the IRM, the compression is applied to constraint the value of DM to $(0, V]$ [75]. The compressed DM is written as:

$$DM_c(t,f) = V \frac{1 - e^{C \cdot DM(t,f)}}{1 + e^{C \cdot DM(t,f)}} \tag{3.3.9}$$

Where C is the steepness constraint, and V is the scaling parameter. Empirically, the values of C and V are 1 and 10 respectively. At the test stage, the DM is decompressed to its original value:

$$\hat{DM}(t,f) = -\frac{1}{C} \log \left(\frac{V - DM_c(t,f)}{V + DM_c(t,f)} \right) \tag{3.3.10}$$

3.3.3 LSTM

The LSTM utilizes forget, input, output and memory gates to control how much past information is employed to update current output. A LSTM block is shown in Fig. 3.3. The LSTM is defined as:

$$f_t = \sigma \left(W_f x_t + U_f h_{t-1} + b_f \right) \tag{3.3.11}$$

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{3.3.12}$$

$$\bar{c}_t = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \tag{3.3.13}$$

$$c_t = f_t c_{t-1} + i_t \bar{c}_t \tag{3.3.14}$$

$$o_t = \sigma\left(W_o x_t + U_i h_{t-1} + b_o\right) \tag{3.3.15}$$

$$h_t = o_t \cdot \tanh\left(c_t\right) \tag{3.3.16}$$

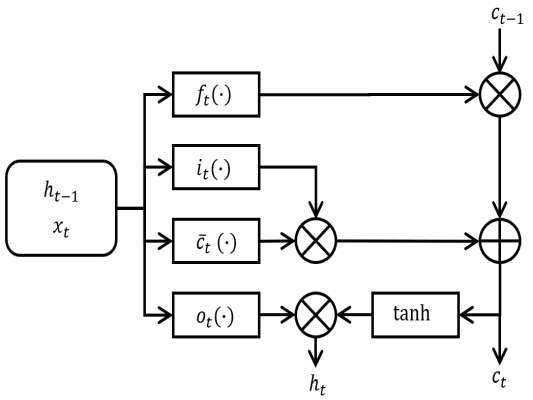


Figure 3.3. The block digram of LSTM network

The f_t and i_t denote the forget gate and input gate at LSTM, $\bar{c}_t$ is the block input, o_t represent the forget gate. There are three inputs h_{t-1}, x_t, c_{t-1} and two outputs h_t, c_t. The W, U denote weights, b's represents biases. σ and $tanh$ represent the sigmod function and hyperbolic tangent function.

More specifically, the forget gate is employed to control what information is forget from the cell state c_{t-1}. It is calculated by using sigmoid function based on time frame x_t and hidden state h_{t-1}. Then, the input gate uses sigmoid function to control how much information is used to update cell memory. Besides, tanh function is used to calculate the $\bar{c}_t$ which will be added to the cell memory of last frame (c_{t-1}) to generate the new cell cell memory c_t. After that, a sigmoid function is utilized to control how much cell memory are outputting, which is fed to a tanh function. Finally, the hidden state h_t is estimated by using output gate. Since the LSTM can preserve

the previous information, which can provide the model with the sufficient information, the mask prediction of LSTM exploited not only present information but also the information from previous frames. Therefore, the LSTM structure is exploited to solve monaural speech separation problem.

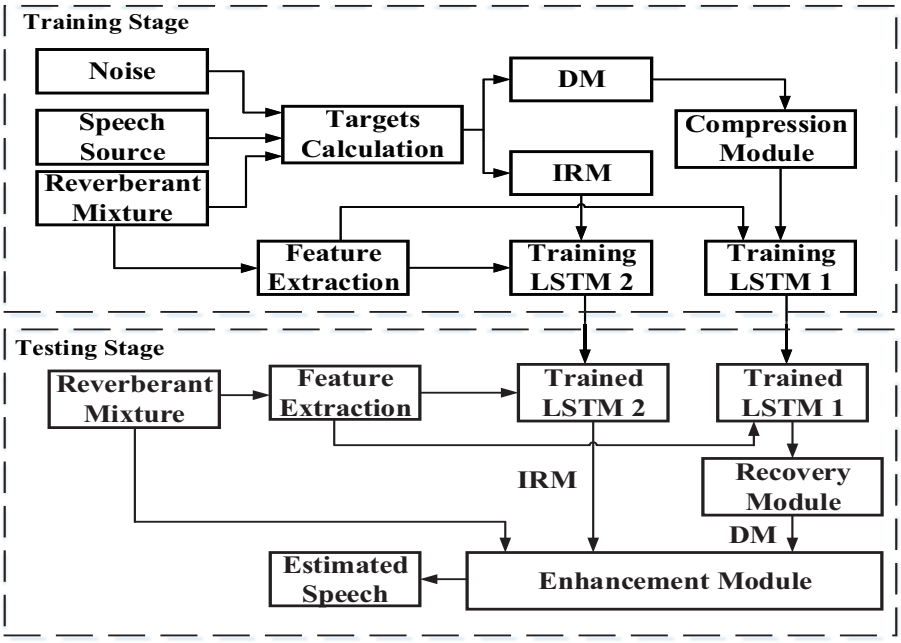


Figure 3.4. The block diagram of the propose two-stage speech enhancement system. Two LSTMs are trained separately. The LSTM1 is used to estimate the DM, and the LSTM2 is exploited to estimate the IRM.

3.3.4 System Architecture

The block diagram of proposed system is shown in Fig. 3.4. At the training stage, the two training targets DM and IRM are calculated by using the speech signal, noise and reverberant noisy speech mixture. The feature combination of training data is extracted from the reverberant mixture. Feature combination and DM are applied to train the LSTM 1. Besides, the LSTM 2 is trained by feature combination and IRM. The relationship between the training targets and feature combination are learnt by two LSTMs.

For the testing stage, the feature combination from testing data is also

extracted, then input to the two trained LSTMs which predict the DM and IRM to be exploited in the enhancement module. Input is the reverberant noisy speech mixture to the enhancement module, the speech source is estimated from the reverberant mixture. Besides, the compression module is applied to map the range of DM. The DM is decompressed to its original value by using the recovery module.

3.4 Simulation for DRM

3.4.1 Datasets

The speech signals are selected from the IEEE corpus which contains 720 utterances [65]. 500 utterances are used to genenrate the training data samples, 100 utterances are applied as development data and 120 utterances are exploited to generate the testing data. factory noise and babble noise are used as background noise, which are selected from the NOISEX datasbase, and both of them are non-stationary [67]. The direct-path impulse responses are obtained by using the geometric information, which is assumed to be avaiable and can be estimated from our previous multimodal human tracking systems [20, 74]. The simulated and real room impulse responses (RIRs) are used to generate the noisy reverberant speech mixtures. The simulated RIRs are generated by the image method [76]. The room dimensions are 9 m $\times$ 5 m $\times$ 3 m, and the target source and microphone are located at 5.5 m $\times$ 2.5 m $\times$ 1.5 m and 4.5 m $\times$ 2.5 m $\times$ 1.5 m, respectively. The RT60 is increased from 0.3 s to 0.9 s with the stepsize of 0.2 s. The database recorded by Surrey University is used for the real RIRs [77], and the RT60s are 0.32 s, 0.47 s and 0.68 s. The SNR levels are set to 3 dB, 0 dB and -3 dB as in [72]. In summary for detailed evaluation of proposed method, there are 21000 training samples, and 4200 testing samples.

The separation performance is evaluated quantitatively by two measures,

they are STOI and PESQ [61, 62].

3.4.2 DNN Settings and Speech Features

The DNN includes four hidden layers, and every hidden layer has 1024 units. The rectified linear unit (ReLU) function is used as the activation function of each unit at hidden layers and the activation function of the output unit is the sigmoid. The maximum number of epochs is 50. The dropout is applied to solve the over-fitting problem, and the rate of dropout is 0.2 [26]. The parameters of the DNN are initialized by random initialization, then they are optimized at every epoch by using adaptive subgradient descent algorithm that has 0.005 learning rate. After 50 epochs, the epoch with minimum cost function value is selected to perform the speech separation task, which is measured by the mean squared error (MSE) cost function.

A complementary set of features is applied [72]. These features are mel-frequency cepstral coefficient (MFCC), spectral transform and perceptual linear prediction (RASTA-PLP) and amplitude modulation spectrum (AMS), and they are spectrum based features [78]. Also, the deltas of RASTA-PLP, AMS and MFCC are appended to the features. The features are normalized to zero mean and unit variance.

3.4.3 Evaluations with Synthetic RIRs

The IRM is used as the benchmark. Table 3.1 and Fig. 3.5 show the PESQ and the STOI values of unprocessed and processed signals with different background noise and RT60s.

In terms of PESQ, both the IRM and the proposed DRM provide considerable improvement over the unprocessed noisy reverberant signal. The proposed DRM outperforms the IRM at all RT60s. And the best PESQ performance is obtained by the DRM at the lowest RT60 (0.3 s). For example, at -3 dB SNR level with factory noise, the proposed method obtains

the PESQ-improvements over the IRM as 0.16, 0.12, 0.16, 0.14 at different RT60s (0.3 s, 0.5 s, 0.7 s, 0.9 s), respectively. Because the higher RT60 increases the complexity in noisy reverberant speech mixture, the PESQ-improvement with the lower RT60 (0.3 s) is better than the higher RT60 (0.9 s). Since the noise has less effect in higher SNR levels speech mixtures, the speech separation performance will be better.

In terms of STOI scores, it is similar with the trend of PESQ. The DRM and the IRM improve the STOI scores, and the average improvement of the DRM over the IRM is approximately 0.021.

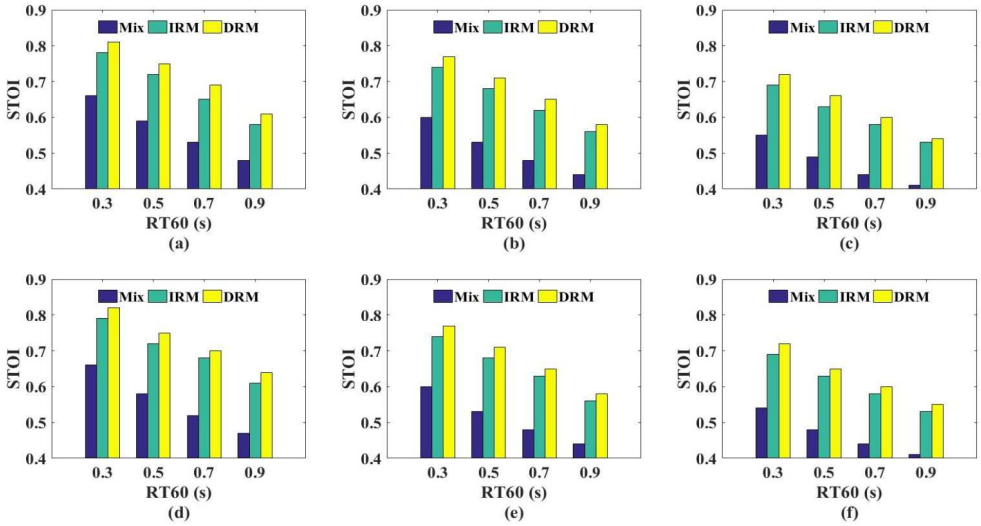


Figure 3.5. Averaged STOI scores of 120 experiments for unprocessed reverberant signals, the IRM [26] and the proposed DRM systems with simulated impulse responses, subfigure:(a) 3 dB factory noise, (b) 0 dB factory noise, (c) -3 dB factory noise, (d) 3 dB babble noise, (e) 0 dB babble noise and (f) -3 dB babble noise.

3.4.4 Evaluations with Real RIRs

Fig. 3.6 and Table 3.2 show the evaluation performance of the proposed approach and the IRM with the real impulse responses. For the STOI performance, the average STOI improvement of the DRM over the IRM is 0.20. When comparing with STOI at different RT60s (0.32 s, 0.47 s), the higher

RT60 (0.47 s) causes worse separation performance, due to higher complexity. Besides, the direct to reverberant ratio (DDR) has positive effect on separation performance. For instance, by using the DRM, when the RT60 is 0.68, the performance is better than the one with lower RT60 (0.47 s), due to the influence of DDR, which strongly justifies another advantage of the geometric information based approach. PESQ performance is consistent with STOI performance.

In summary, the above experimental results confirm the proposed method can separate the target speech from the noisy reverberant mixture in both simulated and real room environments effectively. The proposed method outperforms the state-of-the-art method [26].

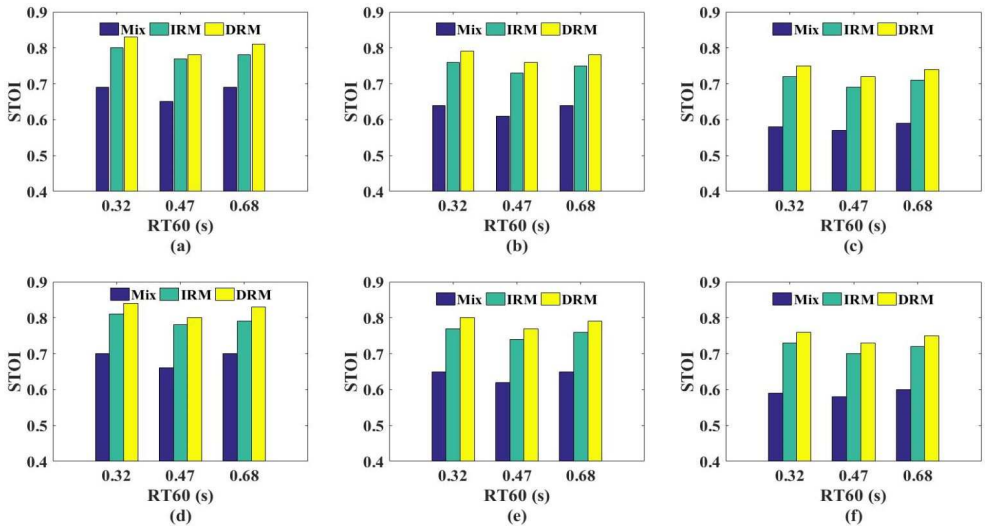


Figure 3.6. Averaged STOI scores of 120 experiments for unprocessed reverberant signals, the IRM [26] and the proposed DRM systems with real impulse responses, subfigure: (a) 3 dB factory noise, (b) 0 dB factory noise, (c) -3 dB factory noise, (d) 3 dB babble noise, (e) 0 dB babble noise and (f) -3 dB babble noise.

Table 3.1. Averaged PESQ scores of 120 experiments for the IRM and the proposed DRM [26] systems at 3 dB, 0 dB and -3 dB SNR levels. The noisy reverberant speech mixtures are obtained by using the IEEE corpus and the factory and the babble background noise under simulated impulse responses. The bold numbers represent the best performance.

SNR Level		3 dB		0 dB		-3 dB	
RT60(s)	Targets	factory	babble	factory	babble	factory	babble
0.3	Unprocessed	0.92	1.06	0.65	0.87	0.48	0.52
	IRM	2.40	2.45	1.95	2.25	1.72	2.03
	DRM	**2.49**	**2.50**	**2.05**	**2.35**	**1.83**	**2.19**
0.5	Unprocessed	0.64	0.83	0.51	0.68	0.45	0.55
	IRM	1.89	2.18	1.69	2.00	1.48	1.83
	DRM	**2.05**	**2.25**	**1.79**	**2.12**	**1.60**	**1.95**
0.7	Unprocessed	0.50	0.64	0.47	0.55	0.44	0.52
	IRM	1.74	1.92	1.55	1.74	1.31	1.62
	DRM	**1.85**	**2.11**	**1.61**	**1.94**	**1.44**	**1.78**
0.9	Unprocessed	0.40	0.60	0.35	0.47	0.31	0.41
	IRM	1.51	1.75	1.32	1.61	1.23	1.46
	DRM	**1.59**	**1.90**	**1.43**	**1.74**	**1.34**	**1.60**

Table 3.2. Averaged PESQ scores of 120 experiments for the IRM [26] and the proposed DRM systems at 3 dB, 0 dB and -3 dB SNR levels. The noisy reverberant speech mixtures are obtained by using the IEEE corpus and the factory and the babble background noise under real recorded impulse responses. The bold numbers represent the best performance.

SNR Level		3 dB		0 dB		-3 dB	
RT60(s)	Targets	factory	babble	factory	babble	factory	babble
0.32	Unprocessed	1.02	1.25	0.74	0.99	0.56	0.78
	IRM	2.31	2.65	2.24	2.51	1.99	2.31
	DRM	**2.42**	**2.70**	**2.37**	**2.57**	**2.11**	**2.39**
0.47	Unprocessed	0.64	0.85	0.49	0.67	0.41	0.57
	IRM	2.17	2.43	1.99	2.31	1.80	2.14
	DRM	**2.28**	**2.53**	**2.11**	**2.40**	**1.89**	**2.21**
0.68	Unprocessed	0.74	0.91	0.69	0.80	0.52	0.61
	IRM	2.21	2.49	2.00	2.24	1.79	2.13
	DRM	**2.33**	**2.51**	**2.11**	**2.42**	**1.92**	**2.22**

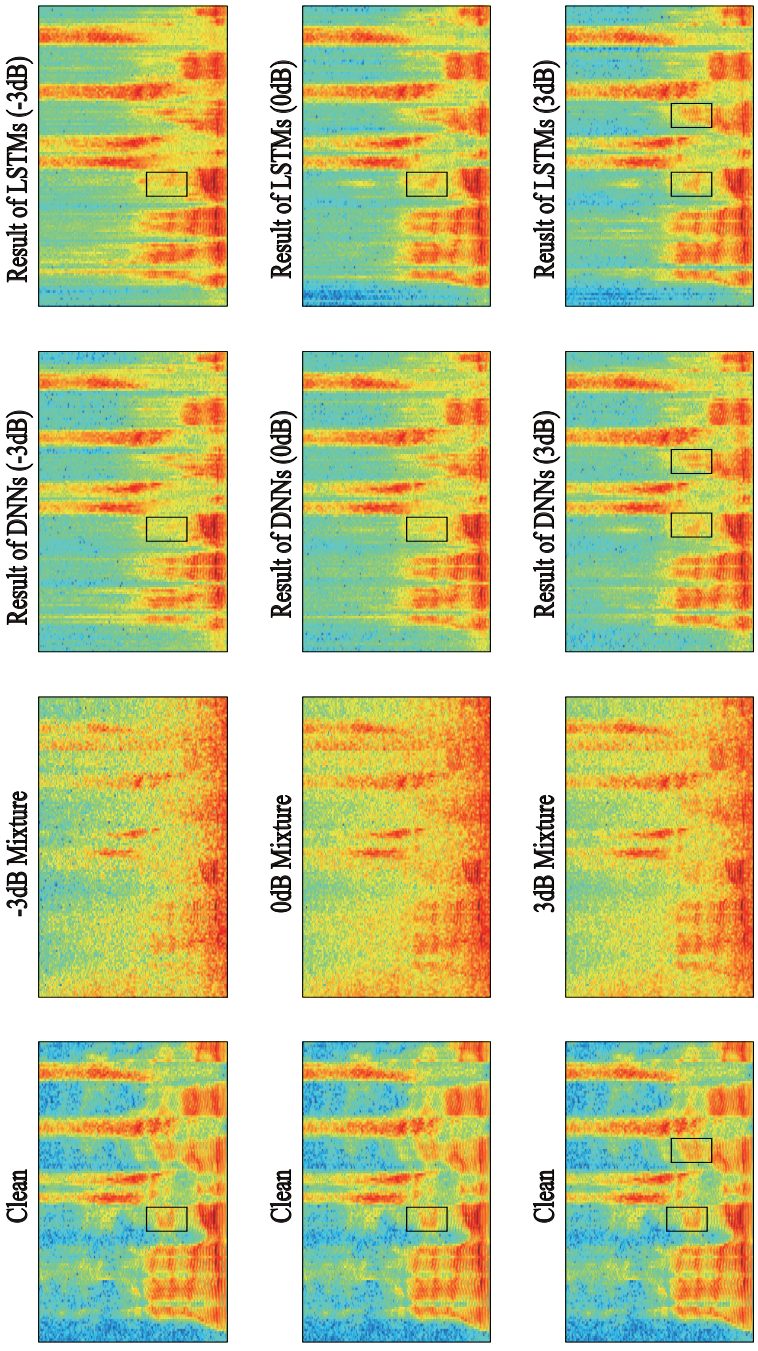


Figure 3.7. Spectrograms of different signals, including the clean speech, noisy mixture, enhanced speeches by DNNs [75] and Proposed LSTMs. The reverberant noisy speech mixture is generated by $factory$ noise at -3dB, 0dB and 3dB SNR levels. The color version is better to understand.

3.5 Simulation for parallel LSTM

A set of spectrograms are shown in Fig. 3.7. It can be observed that both of the DNNs and the proposed LSTMs based methods can be used to recover speech signal. However, the spectrogram of LSTMs based method is more similar to the spectrogram of clean speech.

3.5.1 Datasets

The IEEE [65] and TIMIT corpora [64] are database of speech source. The IEEE contains 720 utterances spoken by a single male speaker, all sentences are downsampled to 16kHz. The TIMT corpora has 630 male and female speakers, everyone spoken 10 utterances, the utterances are recorded as 16-bits, 16kHz speech waveform. To teset the proposed system particularly for speaker-independent case, the training data has 150 male and female speakers from the database, and the 50 unseen speakers are selected from the database in testting set. The factory and babble noise signals are selected from NOISEX database [67]. In general, the noises have duration of four minutes, 16-bits word length and 20kHz sample rate. The factory noise is applied to represent the industrial noise, and the babble noise is the recording of several unseen speakers' voice, both of them are non-stationary. The clean utterances are mixed with noise signals with three signal-to-noise ratio levels (3 dB, 0 dB, -3 dB).

The real room impulse responses (Real RIRs) [77] are convoluted with speech and noise signal to generate the reverberant speech mixture. The RIRs include four types of rooms with different dimensions and RT60s. The detailed parameters are shown in Table 3.3. In total, 12,000 monaural mixtures are generated for training the proposed system, and testing data includes 2880 monaural mixtures.

The separation performance is evaluated quantitatively by SDR improve-

ment [63]. The higher value means better performance.

Table 3.3. The Parameters of Real RIRs for Different Rooms [77]

Room	Size	Dimension (m^3)	RT60(s)
A	Medium	5.7×6.6×2.3	0.32
B	Small	4.7×4.7×2.7	0.47
C	Large	23.5×18.8×4.6	0.68
D	Medium	8.0×8.7×4.3	0.89

3.5.2 LSTM Settings and Speech Features

Both LSTM networks have three hidden layers, each hidden layer has 512 units. To justify the comparison, the DNNs have the same configuration [75]. The number of epoch is 30. The LSTM is trained using stochastic gradient descent (SGD) with momentum. The learning rate is selected as 0.001. The initial momentum is fixed to 0.5 with change for every 5 epoch, and the final momentum is selected as 0.9. The batch size is fixed to 64.

The mel-frequency cepstral coefficient (MFCC), spectral transform and perceptual linear prediction (RASTA-PLP) and amplitude modulation spectrum (AMS) [26] are used to generated the feature combination, which is used to train and test the proposed system.

3.5.3 Evaluations with RIRs

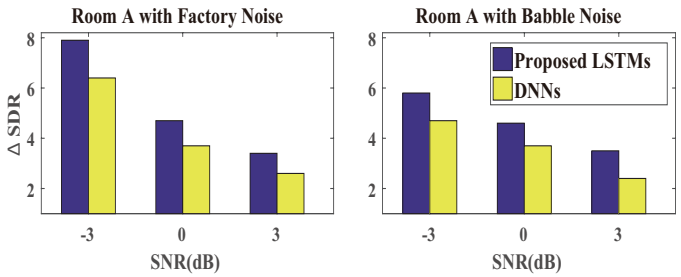


Figure 3.8. Averaged ΔSDR of DNNs method [75] and the proposed LSTMs mentod in Room A with factory and babble noises.

Figs. 3.8, 3.9, 3.10 & 3.11 show the ΔSDR performances of the baseline [75] and the proposed methods with reverberant room environments and two background noises. Since the method in [75] has been confirmed to outperform the IRM- and cIRM-based methods in [72]. Therefore, the method in [75] is used for state-of-the-art comparison.

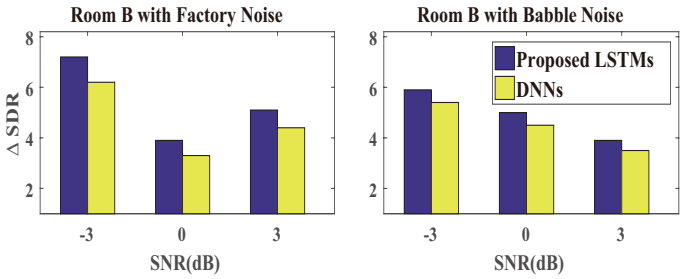


Figure 3.9. Averaged ΔSDR of DNNs method [75] and the proposed LSTMs mentod in Room B with factory and babble noises.

Fig. 3.8 shows both the proposed LSTMs method and baseline DNNs method can provide the consistent ΔSDR in the lowest reverberant environments, which proves they successfully remove the noise component from the noisy speech mixture. Meanwhile, it can be observed that the proposed LSTMs method generates, on average, 1.4dB improvement over DNNs method.

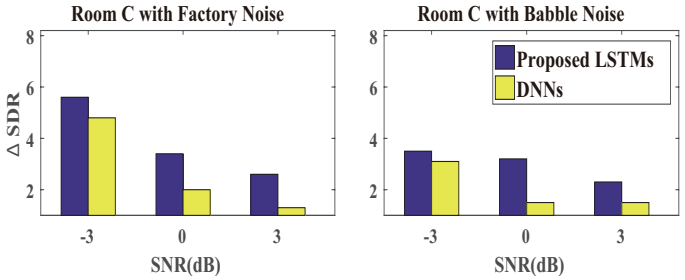


Figure 3.10. Averaged ΔSDR of DNNs method [75] and the proposed LSTMs mentod in Room C with factory and babble noises.

In Room B, the LSTMs method provides, on average, 0.6dB improvement. When compared Room A and Room B, for factory noise, the proposed LSTMs and DNNs obtain the better ΔSDR in Room B. Although Room B has the higher RT60s, which proves the proposed LSTMs can efficiently ad-

dress the dereverberation problem.

In Room C, the proposed LSTMs method obtains, on average, 1dB improvement over the DNNs method. The ΔSDR of Room C is less than other reverberant rooms for both DNNs and LSTMs, because the direct-to-reverberation ratio (DDR) is higher than other rooms [75]. When compared the factory noise with babble noise, the proposed LSTMs method obtains a better ΔSDR with factory noise, because babble noise is a recording of people's conversation, when it is mixed with the speech signal, it increases the complexity in speech enhancement.

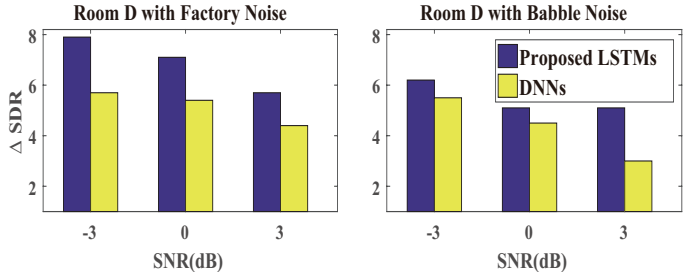


Figure 3.11. Averaged ΔSDR of DNNs method [75] and the proposed two-stage LSTMs mentod in Room D with *factory* and *babble* noise.

In Room D, the proposed LSTMs method generates, on average, 1.5 dB improvement over the DNNs method. Meanwhile, the proposed LSTMs method provides the highest ΔSDR across four reverberant room environments even in the highest RT60s environment. It proves the main advantage of the proposed method over the baseline method is the improved performance at high RT60s. The reduced performance is observed with the increase of the SNR level. The aforementioned results show the DM can remove the speech reverberations in the high RT60s.

In summary, the proposed two-stage LSTMs method obtains, on average, 1.1dB improvement over the two-stage DNNs. The LSTMs can use temporal information to estimate the training targets. Therefore, estimated masks are more accurate, which increase the generalization ability of the system. Moreover, in the high reverberant room environment, the LSTMs provide

significant enhancement performance improvements over the DNNs, which again confirm the temporal information is important for the estimation of the DM.

3.6 Summary

In the this chapter, two methods were proposed to address the speech enhancement in reverberant environment. In first method, the geometric information is utilized to provide the position information of the target speaker and microphone to estimate the direct-path impulse response, which is used to calculate the direct-path speech. Based on the direct-path speech, the DRM was calculated, which is a new training target. The experimental results confirmed the DRM outperforms the state-of-the-art method. In second method, LSTMs were introduced to solve the monaural speech enhancement problem with the speaker-independent case in real reverberant room environments. Two T-F masks were trained separately in the LSTM models to solve the dereverberation and speech enhancement tasks. The proposed method was evaluated with speaker-independent signals and real RIRs to confirm its generalization ability. The experimental results prove the proposed LSTMs method outperforms state-of-the-art DNNs method.

In next chapter, a multi-scale CNN will be provided to capture the features in different scales.

Chapter 4

A MULTI-SCALE FEATURE RECALIBRATION NETWORK FOR END-TO-END MONAURAL SPEECH ENHANCEMENT

4.1 Introduction

Nowadays, a promising direction has been on the exploitation of convolutional neural network (CNN), such as [57], where a convolutional encoder decoder (CED) is introduced to estimate the mapping relation between the noisy mixture and target speech. This is further improved for learning multi-resolution features, with a multi-resolution convolutional auto-encoders (MCARE) model [58], learning with dilated convolution to enlarge the receptive fields of the network in Wavenet, and learning with a gated mechanism to control the information flow among each layer [79]. Furthermore, the gated recurrent network (GRN) method is used with dilated 2-D convolutional layers to enlarge the receptive fields in the time-frequency (T-F) domain [30].

The recurrent and convolutional architectures have been used together to further improve enhancement performance. For example, in the convolutional recurrent network (CRN) [80], the convolutional encoder-decoder is integrated with the LSTM, where the CED is used to capture the local T-F patterns, and the LSTM is used to capture long-term interdependency [80]. The CRN method was shown to perform better than the LSTM.

All the above methods are supervised methods where class labels are required for training the model. In contrast, unsupervised methods have also been proposed for speech enhancement without the requirement of class labels. A well-known method is the speech enhancement generative adversarial network (SEGAN) method [81].

The aforementioned methods are promising and represent current state-of-the-art. However, there are still several limitations. For the CED and CRN methods, a fixed kernel (filter) size is often used. The local information (i.e. feature) in the signal can be extracted by using a kernel of small size, while the contextual feature needs to be extracted with a larger kernel size. A method that can extract both local and contextual information is desired. In the LSTM and CRN models, causal systems are often designed by considering only current and past samples from the signal. However, in terms of [79], the prediction performance of the model can be further improved by considering the future samples. Therefore, in the proposed work, the future information (i.e. a non-causal system) is considered to improve the enhancement performance.

In addition, the implementation of LSTM often involves computational loads for calculating the input, output, forget gates and cell memory [53, 82, 83], sometimes, this can be problematic when the models are deployed on resource-limited devices. It would be desirable to use more efficient RNN models such as BGRU, with performance comparable to BLSTM but less memory requirements. In addition, in the Inception network [55], the fea-

tures of different scales are concatenated directly, and they are assigned with the equal weight. This means that features are considered as equally important, which may be problematic especially when the features are induced by noise. This could be further improved by assigning features with different weights, as shown in proposed work.

In this paper, the MCGN is proposed, with following specific contributions.

First, a multi-scale feature recalibration (MCFR) convolutional encoder-decoder module is introduced, where the kernels with different sizes are exploited in each convolutional layer, to obtain features in different scales. This helps capture the interdependency between the local and contextual information within the signal, and allows the feature in each scale to be assigned with a different weight in order to retain the components from speech while suppressing the components from noise.

Second, the bottleneck convolutional layers are introduced, which uses the 1-D convolutional layer with kernels of size (1,1) to compress the information flow inside the proposed MCGN.

Third, connection layers are used in MCGN, including fully connected (FC) layer and BGRU layers. The FC layer is exploited to reduce the dimension of encoder output. The BGRU layers can capture the interdependencies among the past, current and future temporal frames. Compared with BLSTM, they offer similar performance but require fewer parameters.

Fourth, the multi-scale convolutional output layer is proposed to accelerate the convergence. The output layer enables the enhanced output with access to the different scale convolutional operators, which facilitate network training.

The remainder of this chapter is organized as follows. The a multi-scale feature recalibration (MCFR) convolutional encoder-decoder module with

bidirectional GRU is firstly introduced in Section 4.2. Then, the experiments are performed to make comparison and evaluation between the baselines and the proposed methods in Section 4.3. Section 4.4 stated the summary of this chapter.

This chapter focuses on the second objective of this thesis, which relate to multi-scale feature for monaural speech enhancement accepted by IEEE Journal of selected topics [84] and leading conference [85].

4.2 Algorithm of MCGN Method

4.2.1 Proposed Network Architecture

The details of the proposed MCGN architecture are shown in Fig. 4.1. The MCGN contains four parts, i.e. convolutional encoder, convolutional decoder, connection layers, and multi-scale convolutional output layers. The magnitude spectrum of the noisy mixture is fed to the proposed MCGN, which outputs the estimated magnitude spectrum of the target speech. The convolutional encoder consists of six convolutional layers, except the first convolutional layer and bottleneck convolutional layer, other layers are multi-scale convolutional layers which contain five sub convolutional blocks with varied kernel sizes. Similarly, the convolutional decoder has the symmetric structure with the convolutional encoder. The output of the convolutional encoder is fed to the connection module. After processed by the connection layers, the information flow is fed to the convolutional decoder. In addition, the skip connections are added among the convolutional encoder and decoder. More specifically, the outputs of the convolutional encoder are concatenated with the input of the convolutional decoder, which prevents the information loss and encourage the information reuse. The layer hyper-parameters can be found in Fig. 4.1. Note, the stride size of all layers is (1,2) except the multi-scale output layer which has a fixed stride size of (1,1).

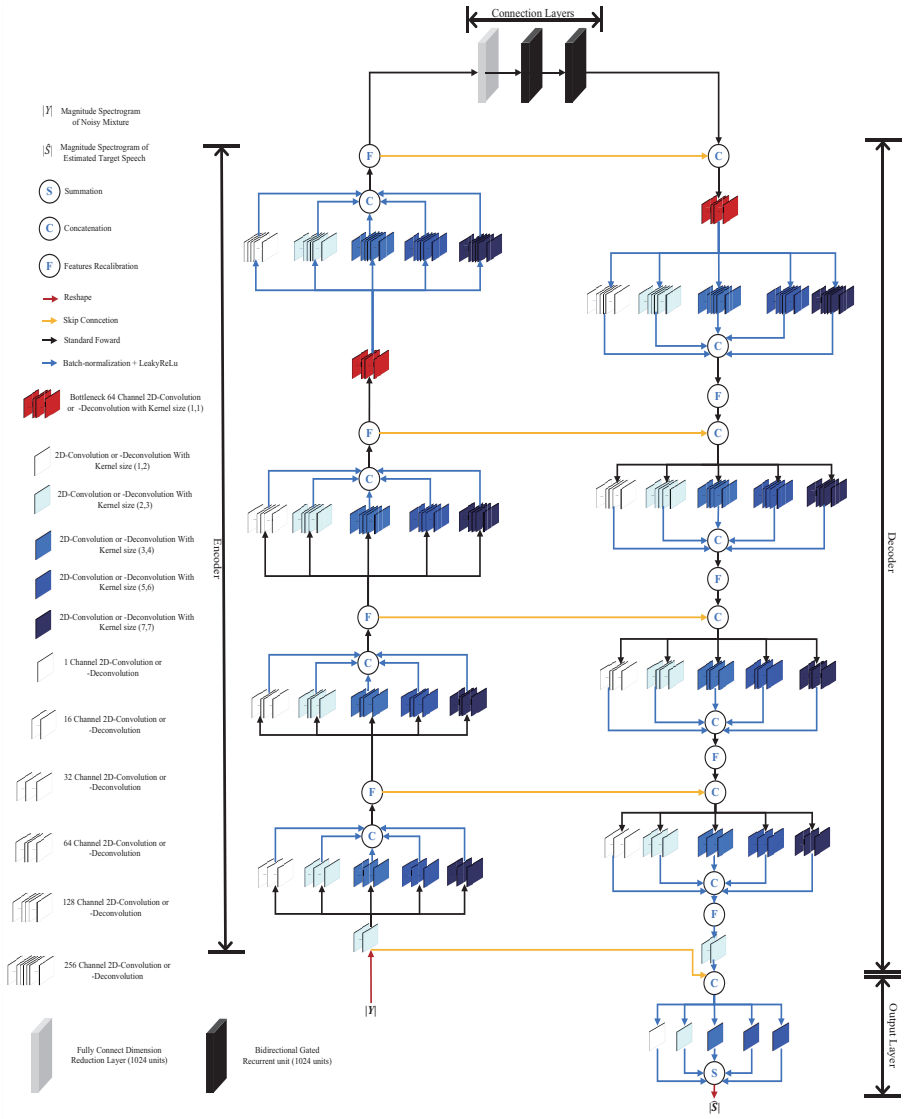


Figure 4.1. The architecture of the proposed MCGN. The components and their functions are shown at the left of the figure. The overlapped 3-D boxes represent the multi-channel 2-D convolutional neural networks. The colored arrows and named circles represent the information flow and operations. The convolutional encoder is on the left of the figure, and the convolutional decoder is on the right of the figure, connection layers are shown in the middle of figure. The figure is color-coded to facilitate understanding.

4.2.2 Multi-Scale Feature Recalibration Convolutional Layer

The receptive field is a region where CNN can affect a particular high-level feature. A small receptive field is feasible to extract local information, and a large receptive field offers contextual information [30]. In conventional CNN, a fixed kernel size is often used, as a result, it compromises between local and contextual information extracted from the signal. To address this limitation, a multi-scale convolutional feature recalibration (MCFR) layer is designed to capture the information on different scales and generate the multi-scaled feature. As shown in Fig. 4.2, MCFR contains several convolutional operators, which use the kernels of different sizes to capture the information with various scales. The convolutional operators with the small kernel sizes can extract the feature from the short duration speech, thus capturing the adjacent T-F points local dependency. The smallest kernel size (1,2) is employed, which allows the feature from two adjacent T-F points to be extracted. The convolutional operators with large kernel sizes offer large receptive fields and can extract features from long-duration speech. These features contain contextual information compared with the feature extracted by kernels with smaller sizes. The batch-normalization is used after each convolutional operator. Different from the standard CNN, which uses the ReLU activation function [86], the proposed MCGN utilizes the activation function LeakyReLU [87]. Then, outputs of each convolutional operator are connected into a single output vector, forming the input of the next stage, as shown in Fig. 4.2. The multi-scale decovolutional layer has a similar structure as the one in MCFR, by replacing the convolutional operators with deconvolutional operators.

After the features at different scales are extracted by using the convolutional operators with varied kernel sizes, a feature recalibration module is introduced to help the network to be selective when using these scaled features, i.e. by assigning different weights to features. It is shown on the

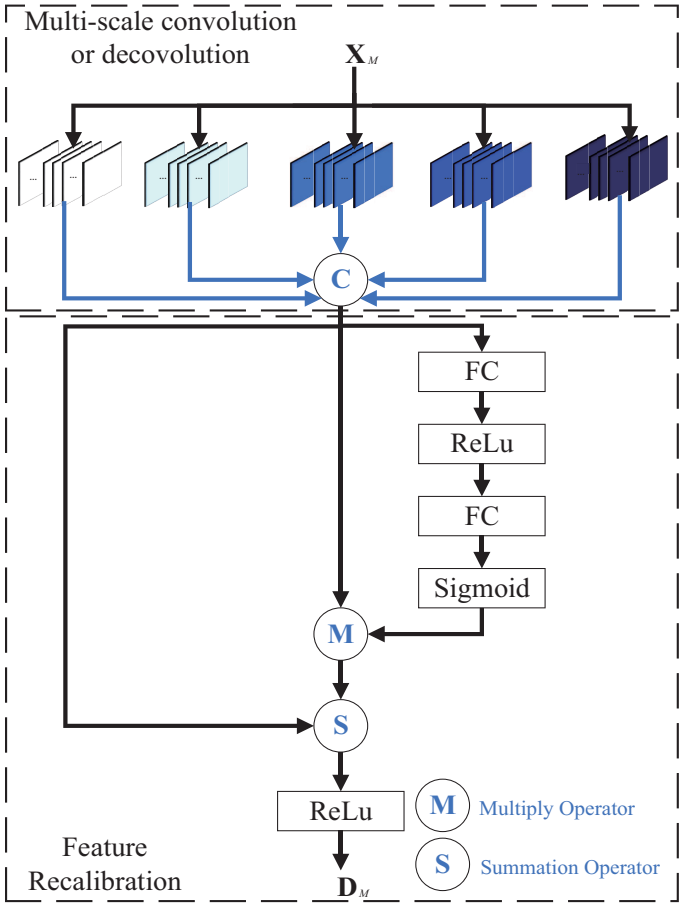


Figure 4.2. Multi-scale feature recalibration network, where $\mathbf{X}_M$, and $\mathbf{D}_M$ represent the input and output of the MCFR module, respectively. The multi-scale convolution or decovolution is shown on top of the figure, kernels of varied sizes are employed to capture the feature in different scales. The bottom of the figure shows the feature recalibration module, features are assigned different weights to retain speech components and suppress the noise components in the noisy mixture.

bottom of Fig. 4.2. The proposed multi-scale convolutional feature recalibration layer is referred as the MCFR layer. In the MCFR layer, there are n sub-convolutional blocks, and each block has the same number of channels but different kernel sizes to capture the features in different scales. The input of the multi-scale layer is $\mathbf{X}_M$, and the output is $\mathbf{K}_M = [\mathbf{k}_1, \mathbf{k}_2, .., \mathbf{k}_n]$, where $\mathbf{k}_n$ is captured by the n-th sub 2-D convolutional block that has different kernel size compared with other 2-D convolutional blocks.

There are several operations for estimating the recalibration coefficients, based on two criteria: the recalibration coefficient could capture the non-linear relation inside the multi-scaled feature, and allocate relatively higher weights to speech components and lower weights to noise components within the feature. The following operations is used to meet these criteria: two FC layers, ReLU and Sigmoid activations. These operations are shown as follows,

$$\mathbf{c}_{1n} = \mathbf{w}_{1n} \odot \mathbf{k}_n + \mathbf{b}_{1n} \tag{4.2.1}$$

$$\mathbf{a}_n = \max[0, \mathbf{c}_{1n}] \tag{4.2.2}$$

$$\mathbf{c}_{2n} = \mathbf{w}_{2n} \odot \mathbf{a}_n + \mathbf{b}_{2n} \tag{4.2.3}$$

$$\mathbf{rs}_n = e^{\mathbf{c}_{2n}} ./ (e^{\mathbf{c}_{2n}} + \mathbf{j}) \tag{4.2.4}$$

where $\mathbf{w}_{1n}$, $\mathbf{w}_{2n}$ denote the weight parameters, $\odot$ denotes element-wise multiplication, $\mathbf{b}_{1n}$, $\mathbf{b}_{2n}$ represent the biases. $\mathbf{c}_{1n}$ and $\mathbf{c}_{2n}$ represent the operations in FC1 and FC2 layers, respectively. $\mathbf{j} = [1, 1, ..., 1]$, and it has the same dimension as $\mathbf{c}_{2n}$. The exponential function e is operated element-wise on $\mathbf{c}_{2n}$, so is the division in the right hand side of equation (4.2.4). The vector $\mathbf{rs}_n$ contains the recalibration coefficient of the n-th scaled feature. Empirically, the ReLU function as (4.2.2) is employed as a non-negative constraint. Inspired by the success of the gating mechanism, Sigmoid is introduced as a gating function to control the information flow, which aims to assign dif-

ferent weights to speech and noise components. The rescaled n-th feature is:

$$\mathbf{p}_n = \mathbf{k}_n \odot \mathbf{rs}_n \tag{4.2.5}$$

Therefore, the rescaled multi-scale feature is $\mathbf{P}_M = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n]$. The deep skip connection (as in residual learning [88]) is introduced inside the MCFR layer. In addition, the residual learning does not introduce any additional parameters. Mathematically, the original relation for the MCFR layer is $\mathbf{D}_M = \mathbf{P}_M$, by using the residual learning and the ReLU function, the relation becomes:

$$\mathbf{D}_M = \max[0, \mathbf{K}_M + \mathbf{P}_M] \tag{4.2.6}$$

Following the extraction of multi-scale features, the proposed MCGN learns the weights and applies them to these features which help retain speech components and suppresses the noise components in the noisy mixture.

4.2.3 Bottlenecks Convolutioal Layers

One of the practical problems in multi-scale convolutional layers that need to be solved is the concatenation of the multi-scale features, which would increase the dimension of the features and cause an increase in computational cost. Therefore, a structure that can retain the information while reducing the complexity (e.g. dimension) is needed. Inspired by the embedding techniques that a low dimensional embedding might contain sufficient information about a relatively large patch [55, 56, 89], the bottleneck convolutional layers are introduced in the proposed MCGN architecture. The bottleneck convolutional layer is a 2-D convolutional layer with (1,1) kernels and 64 channels, followed by the batch-normalization and LeakeyReLU [87]. It is located before the last convolutional encoder layer and the first decoder layer, as shown in Fig. 4.1 (red convolutional blocks). The first bottleneck convolutional layer reduces the dimension from 640-D to 64-D for the last

encoder layer, and the second bottleneck convolutional layer reduces the dimension from 128-D to 64-D for the first decoder layer

4.2.4 Connection Layers

The original convolutional encoder-decoder does not well utilized the long-term temporal information, which, nevertheless, may be valuable in speech enhancement [54, 80]. The CRN method uses the LSTM to capture the long-term interdependency between the past and current temporal frames. However, CRN is designed for the casual problem, which utilizes long-term interdependency between past and current temporal frames. According to [79], the future frames could be used to improve enhancement performance. BGRU layers are introduced to capture the long-term interdependency among the past, current and future temporal frames. In comparison, GRU offers comparable performance to LSTM [83, 90–92], but has an advantage in parameter efficiency. However, the merging of the multi-scaled convolutional sub-blocks would lead to an inevitable increase in its dimension. Therefore, it is necessary to find a way to retain the information and, at the same time, to reduce the dimension and computational cost. To address this, we use a fully connected (FC) layer, as the number of parameters of the fully connected dense layer is smaller than that of the RNN based layer, leading to a reduced dimension in the output of the FC layer, as compared with the output of the encoder.

4.2.5 Multi-Scale Output Layer

The skip connection from the input to the multi-scale output layer is added, as shown at the bottom of Fig. 4.1. As a result, the multi-scale output layer can estimate the magnitude of the target speech from the previous layer's information flow and the input magnitude of the noisy mixture. The multi-scale output layer is a 2-D deconvolutional layer, which contains five

sub-blocks, and the kernel sizes of these sub-layers are different. Unlike the MCFR layer, these varying scaled features are concatenated, the different scaled features are summed together to generate an output matrix with the same size as the input matrix. Thus, the multi-scale output layer utilizes local and contextual information. The stride size of the output layer is set to (1,1). Batch-normalization and linear activation are followed.

4.3 Experimental Evaluations

4.3.1 Datasets

The proposed system is evaluated with three experiments using three different datasets. In the first experiment, we use 1000 clean utterances mixed with 20 noise signals to generate the training set in first experiment. The clean utterances are randomly selected from the TIMIT corpus [64], and noise files are selected from Non-Speech Sounds [68] and NOISEX-92 [67] datasets. Similarly, 100 clean utterances are mixed with 6 noise signals to generate the testing datasets. To better evaluate enhancement performance, the speakers in the training set are different from the speakers in the testing dataset. Meanwhile, the testing noisy interferences are categorized into two types, the seen noises (Babble, Leopard, F16) and the unseen noises (N56, N72, White). Babble, Leopard, F16, N56, N72 are non-stationary noises, and White is stationary noise. N56 and N72 are wind and water sounds, respectively. The noisy mixtures are generated by mixing the clean utterances and noises at -5dB, 0dB and 5dB signal-to-noise ratio (SNR) levels. In total, about 50 hours ($3\times3\times1000\times20\div3600$) noisy mixtures are used to train the networks.

In the second experiment, the proposed method is evaluated on a published dataset [79, 81]. The datasets are generated by using the VCTK corpus [66] and DEMAND Database [69]. The utterances from 28 speakers and

2 speakers are used for training and testing, respectively. Each speaker has spoken around 400 sentences. The training utterances are mixed with 10 types of noise in four SNR levels (0dB, 5dB, 10dB and 15dB). In total, there are 11,572 noisy mixtures for training. Similarly, the testing utterances are mixed with 5 types of noise in four SNR levels (2.5dB, 7.5dB, 12.5dB and 17.5dB). In total, the testing set includes 824 noisy mixtures, where both the speakers and noises are unseen in the training set.

In the third experiment, the proposed MCGN method is evaluated with a larger dataset. For the training set, 2500 clean utterances are randomly selected from the TIMIT [64] and VCTK [66] corpora, mix them with 20 different noise signals selected from the Non-Speech Sounds [68] and NOISEX-92 [67] datasets, to generate 50000 training mixtures for each SNR level (-5dB, 0dB, and 5dB). Similarly, for the testing set, we randomly select 500 clean utterances and mix them with 5 different noise signals, to generate 2500 noisy mixtures for each SNR level. The speakers of the training dataset are different from those in the testing dataset. The Babble, Leopard, F16 are seen noises, while N56 and N72 are unseen noises.

The signal to distortion ratio improvement (ΔSDR) [63], perceptual evaluation of speech quality (PESQ) [61] and short-time objective intelligibility (STOI) [62] are used to measure the performance. The higher values of the measurements indicate better enhancement performance.

4.3.2 Baselines and Parameters

The proposed MCGN is compared with seven baseline methods, including the standard DNN method from [25], the DNN method with skip connection S-DNN from [31], the LSTM model used in [54], the BLSTM model used in [30], the CNN based methods, the MRCAE method from [58], and the GRN method in [30]. The parameters of the CRN model are set by following [80]. LSTM and BLSTM have four hidden layers, where each hidden

layer contains 1024 units with a dropout rate of 0.2, and the output layer is a dense layer. The MRCAE is a five-layered 1-D convolutional encoder decoder. The encoder consists of two multi-resolution 1-D convolutional layers, and the decoder mirrors the encoder. A deconvolutional layer is used as the output layer of MRCAE. The CRN consists of the 2-D convolutional encoder, two-layered LSTM and 2-D convolutional decoder, which are connected by standard feed-forward connections and skip connections. The GRN is a 62-layered fully connected dilated convolutional neural network with the residual. The aforementioned baseline methods and proposed MCGN method take the STFT magnitude spectrum of the noisy speech mixture as the input features, and output the corresponding magnitude spectrum of the estimated target speech. The estimated magnitude spectrum is combined with the noisy phase to re-synthesize the estimated target speech waveform. Furthermore, the proposed MCGN model trained on the published dataset [79, 81] is compared with the SEGAN and Wavenet. The SEGAN employs generator and discriminator to learn and judge the input data distribution, which uses the adversarial training [81]. The Wavenet is a 30-layered fully connected convolutional neural network [79].

The input and output layers for all methods contain 257 units. The baseline methods and proposed MCGN method are trained with the Adam optimization algorithm [93]. The initial learning rate is set to 0.0001. The mean square error (MSE) is employed as the objective function for the baseline and the proposed MCGN methods. The dropout rate is fixed to 0.2. The sample rate of noisy speech mixtures is 16kHz, and the window length is 512. The time resolution is 32ms, and frequency resolution is 32.15Hz. The next two sections (i.e. Sections 4.3.3 and 4.3.4) report the results based on the first dataset, while Sections 4.3.5 and 4.3.6 present results for the second and third dataset, respectively.

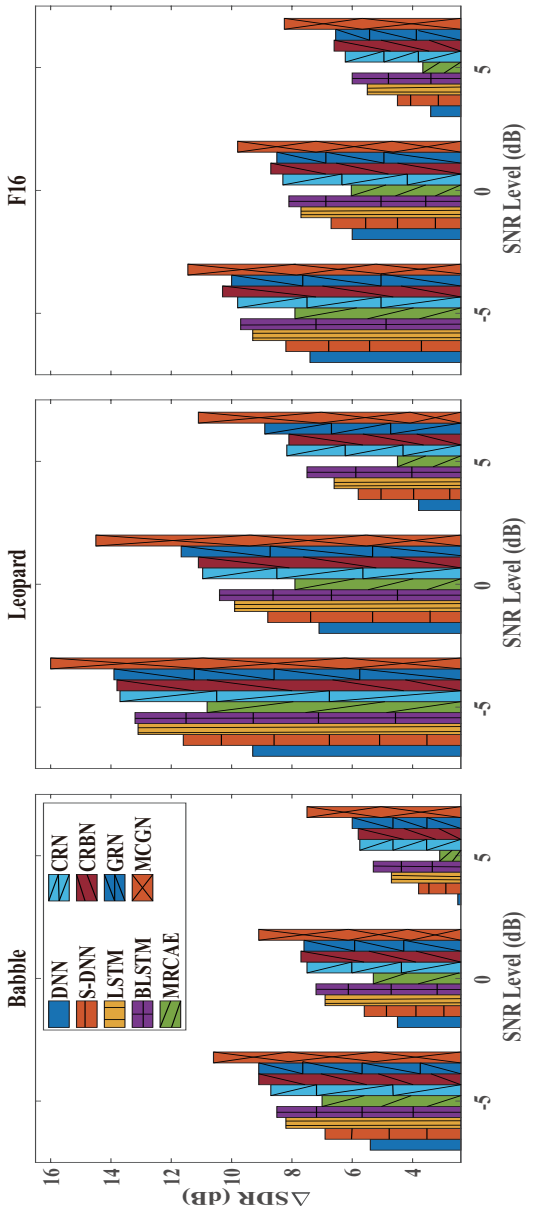


Figure 4.3. Speech enhancement performance comparison in terms of ΔSDR for three types of noise with different methods and SNR levels. Each result is the average value of 100 experiments.

Table 4.1. Speech enhancement performance comparisons in terms of STOI over three different types of noise signals with different baseline methods and SNR levels. Each result is the average value of 100 experiments. *Italic* text refers to the proposed methods. **Bold** number indicates the best performance.

Measure	STOI(%)											
Noises	Babble				Leopard				F16			
SNR	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	53.86	63.05	71.66	62.86	71.68	75.57	78.92	75.39	54.39	64.07	73.37	63.94
DNN [25]	66.36	72.91	79.26	72.83	77.11	80.55	83.26	80.30	69.17	76.24	81.35	78.79
S-DNN [31]	66.80	73.66	79.63	73.36	78.34	81.54	83.98	81.29	69.77	76.46	81.86	75.97
LSTM [54]	68.78	75.81	81.54	75.38	80.76	83.32	85.40	83.16	72.13	77.86	83.39	77.79
BLSTM [30]	69.30	76.63	82.04	75.99	81.10	83.85	85.58	83.59	72.19	78.15	83.61	77.98
MRCAE [58]	65.92	72.83	78.85	72.53	77.50	80.56	83.09	81.25	69.10	75.51	80.88	75.16
CRN [80]	70.10	76.95	81.88	76.31	81.20	84.02	85.80	83.67	72.65	78.98	83.90	78.51
CRBN	70.30	77.08	81.96	76.45	81.20	84.20	85.90	83.77	73.49	79.12	84.14	78.92
GRN [30]	71.60	77.08	82.21	76.94	82.20	84.15	86.08	84.14	72.94	79.27	83.56	78.59
MCGN	**75.02**	**80.52**	**84.43**	**79.99**	**84.10**	**85.78**	**87.31**	**86.55**	**77.49**	**81.56**	**85.60**	**81.55**

Table 4.2. Speech enhancement performance comparisons in terms of PESQ over three different types of noise with different baseline methods and SNR levels. Each result is the average value of 100 experiments. *Italic* text refers to the proposed method. **Bold** number indicates the best performance.

Measure	PESQ										
Noises	Babble				Leopard			F16			
SNR	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	1.28	1.52	1.81	1.53	1.75	1.99	2.22	1.97	1.31	1.54	1.83	1.56
DNN [25]	1.58	1.90	2.20	1.89	2.03	2.31	2.50	2.28	1.73	2.08	2.31	2.04
S-DNN [31]	1.69	2.00	2.28	1.99	2.25	2.45	2.67	2.46	1.82	2.11	2.35	2.09
LSTM [54]	1.82	2.15	2.44	2.14	2.41	2.61	2.80	2.61	1.97	2.28	2.52	2.25
BLSTM [30]	1.84	2.19	2.47	2.16	2.44	2.67	2.84	2.65	2.02	2.30	2.54	2.29
MRCAE [58]	1.72	2.04	2.31	2.02	2.25	2.47	2.69	2.47	1.85	2.15	2.39	2.13
CRN [80]	1.91	2.22	2.49	2.21	2.50	2.70	2.90	2.69	2.02	2.30	2.54	2.29
CRBN	1.93	2.23	2.50	2.22	2.51	2.72	2.90	2.71	2.09	2.38	2.58	2.35
GRN [30]	1.94	2.24	2.49	2.22	2.53	2.73	2.92	2.73	2.05	2.34	2.58	2.32
MCGN	**2.16**	**2.43**	**2.65**	**2.41**	**2.70**	**2.88**	**3.04**	**2.87**	**2.23**	**2.47**	**2.70**	**2.47**

Table 4.3. The p-value of the t-test at 5% significance level, between the proposed method and the baseline methods. H_0 denotes the null hypothesis, and (+) indicates that the difference among the pair is statistically significant at the 95% confidence level.

Measures	STOI		PESQ	
	p-value	H_0	p-value	H_0
Noisy	1.49E-05	(+)	4.14E-12	(+)
DNN [25]	5.22E-06	(+)	1.76E-07	(+)
S-DNN [31]	1.08E-05	(+)	3.20E-09	(+)
LSTM [54]	8.16E-05	(+)	3.49E-07	(+)
BLSTM [30]	1.93E-04	(+)	1.46E-06	(+)
MRCAE [58]	3.19E-06	(+)	1.87E-06	(+)
CRN [80]	1.44E-04	(+)	2.74E-07	(+)
CRBN	7.71E-05	(+)	1.13E-05	(+)
GRN [30]	1.01E-04	(+)	1.08E-06	(+)

4.3.3 Unseen Speakers with Seen Noise

Fig. 4.3 and Tables 4.1 & 4.2 provide experimental results in terms of ΔSDR, STOI and PESQ for the baseline and the proposed methods with real-world noises. The speakers used in testing are unseen in the training data. The noises used in testing include Babble, Leopard, and F16.

The DNN generates, on average, ΔSDR = 5.49dB, STOI = 76.26% and PESQ = 2.07, which offers the worst enhancement performance across all the compared methods. These results show that the generalization ability of DNN remains insufficient. The S-DNN slightly outperforms the DNN, because S-DNN explicates the skip connection to build the residual mapping relation, which mitigates performance degradation. The MRCAE method uses the multi-resolution 1-D convolutional encoder decoder and offers a small improvement over the DNN in terms of ΔSDR, and PESQ.

The LSTM generates, on average, ΔSDR = 8.03dB, STOI = 78.77% and PESQ = 2.33, which shows better generalization ability over the DNN, S-DNN and MRCAE. Different from the DNN, S-DNN and MRCAE method, the LSTM exploits the memory cell to keep the hidden states from the past

temporal frame. Incorporating the past and current temporal frames, the inter-dependencies between them are captured by the LSTM. The BLSTM outperforms the LSTM, due to the use of forward-LSTM and backward-LSTM in every BLSTM layer. The forward-LSTM is the same as the standard LSTM, which is used to capture the interdependency between the past and current temporal frames. However, the backward-LSTM is fed by reverse input sequence, and thus the interdependency between current and future temporal frames are also utilized, to achieve further improvement over the LSTM.

The CRN obtains, on average, ΔSDR = 8.81dB, STOI = 79.49% and PESQ = 2.39, which provides higher improvements over the DNN, S-DNN and LSTM methods. Since the local spatial patterns of the input magnitude spectrum are captured by CRN, it is capable of leveraging the T-F structure of the magnitude spectrum. Moreover, the LSTM layers inside the CRN exploit the temporal dependency by using past and current temporal frames. In addition, experiments are performed for the non-casual version of CRN, namely CRBN, where the LSTM layers are replaced by the BLSTM layers. The experimental results show that the CRBN offers slight improvements over the CRN method, which confirms that the interdependency between the current and future frames is helpful for improving predictions by the model. The GRN outperforms the CRN by using the dilated convolutional layers.

The proposed MCGN gets the highest improvements over the baseline methods, and it achieves, on average, ΔSDR = 10.88dB, STOI = 82.42% and PESQ = 2.58, which are almost 1.7dB, 2.53% and 0.16 higher than those achieved by the CRN method. The MCGN using the MC to encode the input magnitude spectrum in different scales. The local interdependency is captured by the convolutional sub-layers with small kernel sizes. The convolutional sub-layers with large kernel sizes are used to find the inter-

dependency between the remote frames. By using the small and large size kernels, the receptive field of MCGN is enlarged, and the different scaled features are assigned with different weights. Furthermore, the BGRU layers are introduced to connect the multi-scale encoder and multi-scale decoder, which are capable of exploiting the interdependency of the past, current and future temporal frames. Besides, the raw data is fed to the output layer of the MCGN to learn the residual mapping relation.

The t-test [94, 95] between the proposed MCGN method and baseline methods are also performed, noisy mixtures for the unseen speakers with seen noises cases. The t-test results are shown in Table 4.3. It can be seen that the p-values are all smaller than 0.05 and all the null hypothesis is (+), which shows that the proposed MCGN method yields a statistically significant improvement over the baseline methods.

Table 4.4. The p-value of the t-test at 5% Significance Level, comparison of proposed method with the baseline methods. H_0 denotes the null hypothesis, and (+) indicates the improvement of two pairs is statistically significant at the 95% confidence level.

Measures	STOI		PESQ	
	p-value	H_0	p-value	H_0
Noisy	1.49E-04	(+)	2.06E-05	(+)
DNN [25]	2.97E-04	(+)	1.73E-05	(+)
S-DNN [31]	6.75E-04	(+)	3.80E-04	(+)
LSTM [54]	4.18E-04	(+)	3.05E-04	(+)
BLSTM [30]	2.24E-04	(+)	4.17E-05	(+)
MRCAE [58]	9.94E-04	(+)	2.42E-04	(+)
CRN [80]	1.89E-04	(+)	2.62E-05	(+)
CRBN	1.06E-04	(+)	1.18E-05	(+)
GRN [30]	1.94E-04	(+)	2.82E-05	(+)

Table 4.5. Speech enhancement performance comparisons in terms of STOI over three types of noises with different state-of-the-art methods and SNR levels. Each result is the average value of 100 experiments. *Italic* text refers to the proposed methods. **Bold** number indicates the best performance.

Measure	STOI(%)										
Noises	N56				N72			White			
SNR	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	57.07	68.26	78.37	67.90	70.87	76.77	81.63	76.42	53.40	62.57	72.32	62.76
DNN [25]	72.59	78.74	84.07	78.46	75.34	81.43	85.00	80.59	62.73	70.42	77.32	70.16
S-DNN [31]	72.79	78.74	84.31	78.85	76.87	82.18	85.67	81.57	63.00	70.63	77.87	70.50
LSTM [54]	76.99	79.45	86.47	80.97	76.83	82.95	86.80	82.19	67.71	75.91	81.53	75.05
BLSTM [30]	77.64	82.49	86.87	82.33	78.05	83.50	87.12	82.89	72.93	76.43	83.91	77.76
MRCAE [58]	72.74	78.95	83.78	78.49	75.47	80.57	84.72	80.25	65.12	71.36	76.31	70.93
CRN [80]	77.88	83.37	87.09	82.78	78.55	84.12	87.24	83.30	72.90	78.93	84.34	78.72
CRBN	78.95	83.85	87.30	83.37	79.26	84.37	87.64	83.76	76.24	81.30	85.16	80.92
GRN [30]	78.19	83.60	87.33	83.04	78.96	84.27	87.60	83.61	76.31	80.80	84.59	80.57
MCGN	**82.83**	**86.85**	**89.82**	**86.50**	**81.07**	**85.52**	**88.36**	**84.98**	**78.26**	**83.81**	**87.75**	**83.27**

Table 4.6. Speech enhancement performance comparisons in terms of PESQ over three types of noises with different state-of-the-art methods and SNR levels. Each result is the average value of 100 experiments. *Italic* text refers to the proposed method. **Bold** number indicates the best performance.

Measure	PESQ											
Noises	N56				N72				White			
SNR	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	1.14	1.31	1.57	1.34	1.58	1.83	2.06	1.83	1.04	1.21	1.47	1.24
DNN [25]	1.69	1.98	2.16	1.95	1.77	2.05	2.23	2.02	1.32	1.61	1.87	1.60
S-DNN [31]	1.74	2.05	2.20	1.98	1.86	2.11	2.32	2.10	1.34	1.69	1.94	1.66
LSTM [54]	1.93	2.17	2.36	2.16	1.87	2.14	2.40	2.14	1.59	1.96	2.26	1.94
BLSTM [30]	1.97	2.20	2.39	2.18	1.92	2.20	2.45	2.19	1.81	2.19	2.45	2.15
MRCAE [58]	1.83	2.06	2.22	2.04	1.93	2.16	2.35	2.15	1.60	1.85	2.09	1.85
CRN [80]	1.92	2.22	2.49	2.21	1.98	2.20	2.41	2.20	1.90	2.21	2.48	2.20
CRBN	2.05	2.27	2.44	2.25	1.99	2.25	2.47	2.24	2.04	2.28	2.54	2.29
GRN [30]	2.01	2.24	2.43	2.23	2.01	2.25	2.53	2.26	2.14	2.35	2.56	2.35
MCGN	**2.22**	**2.40**	**2.58**	**2.40**	**2.14**	**2.40**	**2.63**	**2.39**	**2.24**	**2.55**	**2.84**	**2.54**

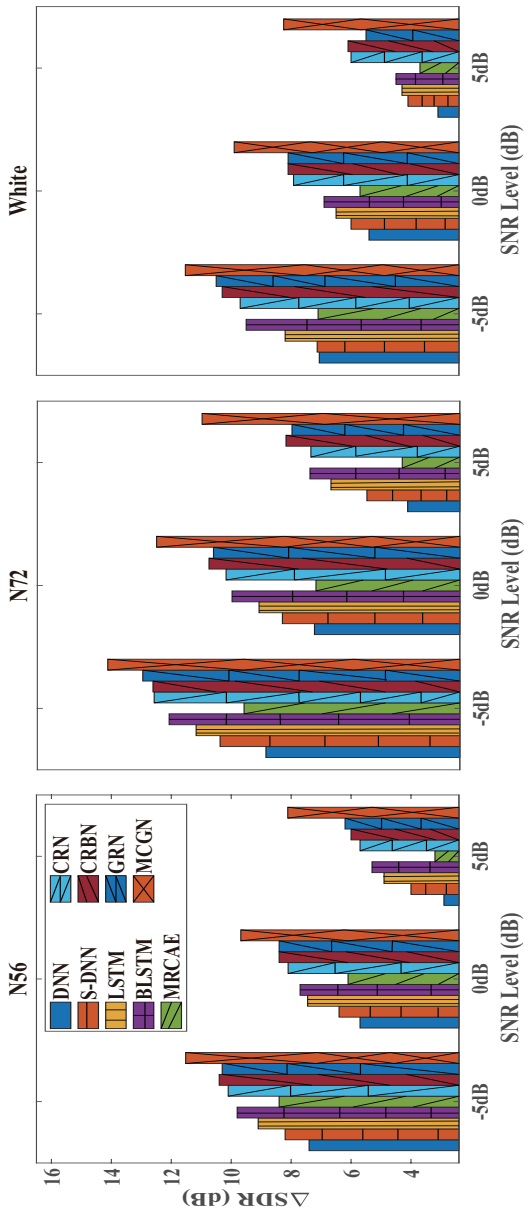


Figure 4.4. Speech enhancement performance comparison in terms of ΔSDR for three types of noise with different methods and SNR levels. Each result is the average value of 100 experiments.

4.3.4 Unseen Speaker with Unseen Noises

Fig. 4.4 and Tables 4.5 & 4.6 provide experimental results in terms of ΔSDR, STOI and PESQ for the baseline and the proposed methods with real-world noise. The testing speakers are unseen in training data. The testing noises

are N56, N72 and White noises, which are also unseen in the training data.

The DNN method offers slight improvement over the original noisy mixture. The MRACE outperforms the DNN method in terms of ΔSDR and PESQ, but its STOI performance is worse than that of DNN and S-DNN. These results show that the shallow structure and small channel numbers can limit the performance of MRCAE. Besides, the large size filters increase computational cost. The skip connection in S-DNN boosts enhancement performance compared to the DNN method. The LSTM obtains further improvement by incorporating the past and current temporal information. The utilization of past, current and future temporal information in BLSTM shows advantages over the LSTM and DNN based method. The CRN method incorporates the CNN encoder-decoder with the LSTM, the convolutional encoder-encoder takes advantage of the convolutional layer and batch normalization to provide a high-level representation of the input feature, which accelerates the training and improves the enhancement performance. Due to the incorporation of the BLSTM layers, the CRBN offers advantages over the CRN method in terms of ΔSDR, STOI and PESQ. The GRN method uses gated linear units to control the information flow, and dilated convolutional layers to expand the receptive fields. These strategies enable the GRN method to outperform the aforementioned methods.

The proposed MCGN method offers improvements over all the baseline methods in terms of ΔSDR, STOI and PESQ. The t-test results in Table 4.4 also show that the improvement of the proposed MCGN methods is statistically significant.

4.3.5 Experiments on Published Dataset

The proposed MCGN method is evaluated on the second dataset i.e. the published dataset generated by the VCTK corpus. Fig. 4.5 shows experimental results. Note that the model size (i.e. the number of parameters)

of SEGAN, Wavenet and the proposed MCGN is 193M, 34.3M, 77.5M respectively. The no-casual, dilated convolutions controlled by the Sigmoid gate in every layer help to enlarge the receptive fields of every kernel, and thus to utilize the interdependency among input features. The future samples help the Wavenet to perform better. The proposed MCGN method produces substantially better enhancement performance, since the MCFR model provides weighted multi-scale feature in every layer, and captures the interdependency among different frames including future frames.

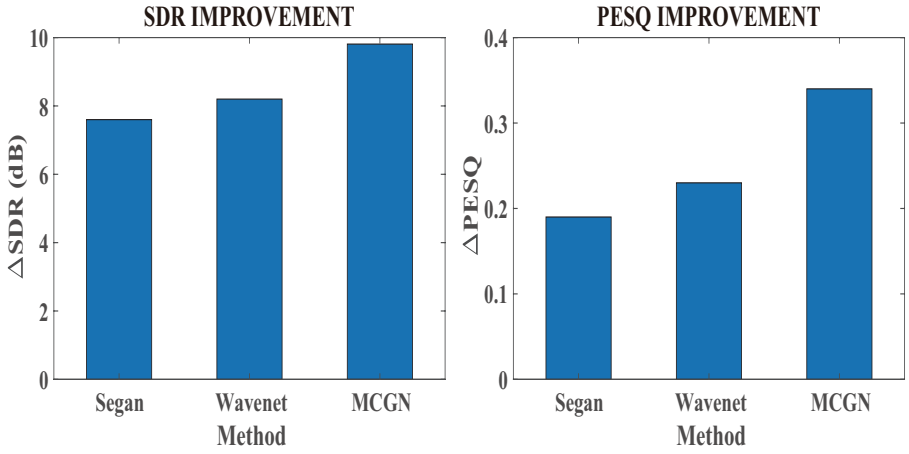


Figure 4.5. Speech enhancement comparison in terms of ΔSDR and ΔPESQ for Segan [81], Wavenet [79] and the proposed MCGN. The enhancement results are the average value of 824 noisy mixtures.

4.3.6 Additional Experiments

Figs. 4.6 & 4.7 and Tables 4.7 & 4.8 provide experimental results in terms of ΔSDR, STOI and PESQ for the proposed MCGN and four baseline methods (i.e. LSTM, BLSTM, CRN and GRN) with seen and unseen noises, for the larger dataset (i.e. 50000 training signals and 2500 testing signals for each SNR level, described in Section 4.3.1).

It can be observed that the proposed MCGN method performs better than all the baseline methods, and shows similar trends as for the smaller dataset tested earlier. All the methods provide some improvements over

the noisy mixtures, which indicate that they are effective for speech enhancement with seen and unseen noises. The BLSTM provides more improvements than LSTM, since it uses additional information from the future frames, in contrast to the information from only current and previous frames used in LSTM. The CRN uses the CED to capture local T-F patterns from input noisy mixtures, also uses the LSTM layers to relate the past frames with current frames, thus offering higher improvements than the LSTM and BLSTM. The GRN shows advantage over LSTM, BLTM and CRN, due to the employment of the dilated 2-D convolutional layers for expanding the receptive fields in the T-F domain, and the gated convolution to control the information flow between layers.

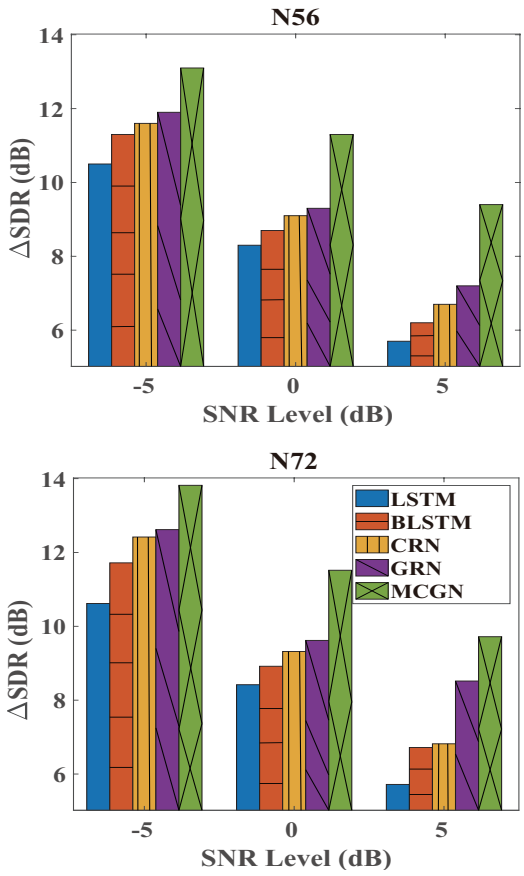


Figure 4.6. Speech enhancement performance comparison in terms of ΔSDR for two unseen noises with different methods and SNR levels. Each result is the averaged value of 500 experiments.

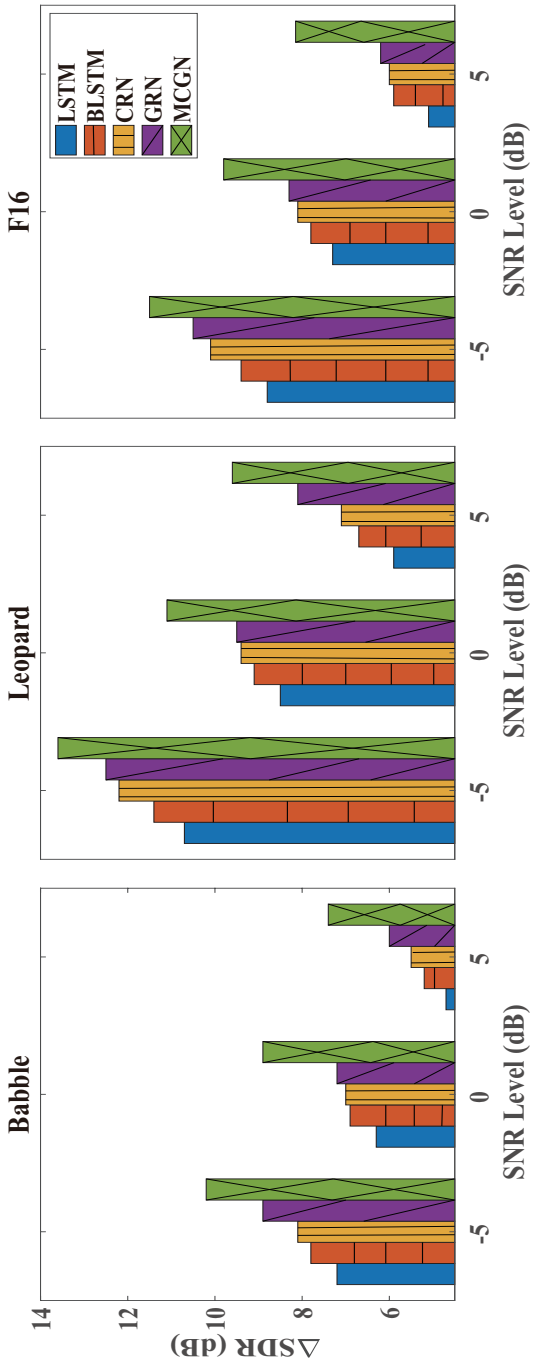


Figure 4.7. Speech enhancement performance comparison in terms of ΔSDR for three types of noise with different methods and SNR levels. Each result is the averaged value of 500 experiments.

Table 4.7. Speech enhancement performance comparisons in terms of STOI and PESQ over two different types of unseen noises with baseline methods and SNR levels. Each result is the averaged value of 500 experiments. *Italic* text refers to the proposed method. **Bold** number indicates the best performance.

Measure	STOI (%)							
Noises	N56				N72			
SNR	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Mixture	57.83	67.19	76.02	67.01	79.19	78.37	83.44	80.33
LSTM [54]	77.29	82.25	85.93	81.82	81.75	85.62	88.36	85.24
BLSTM [30]	77.59	82.54	86.45	82.19	82.59	86.18	89.11	85.96
CRN [80]	77.90	83.62	87.24	82.92	82.61	86.81	89.25	86.22
GRN [30]	78.50	83.88	87.56	83.31	83.00	87.4	89.35	86.58
MCGN	**83.87**	**87.20**	**89.67**	**86.91**	**85.57**	**88.55**	**90.54**	**88.22**
Measure	PESQ							
Noises	N56				N72			
SNR	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Mixture	1.20	1.33	1.56	1.36	1.57	1.84	2.10	1.84
LSTM [54]	2.03	2.22	2.40	2.22	2.15	2.39	2.58	2.37
BLSTM [30]	2.05	2.26	2.44	2.25	2.23	2.42	2.62	2.42
CRN [80]	2.07	2.31	2.47	2.28	2.28	2.46	2.65	2.46
GRN [30]	2.08	2.32	2.45	2.28	2.29	2.46	2.67	2.47
MCGN	**2.27**	**2.51**	**2.64**	**2.47**	**2.42**	**2.61**	**2.82**	**2.62**

Table 4.8. Speech enhancement performance comparisons in terms of STOI and PESQ over three different types of seen noises with different baseline methods and SNR levels. Each result is the averaged value of 500 experiments. *Italic* text refers to the proposed method. **Bold** number indicates the best performance.

Measure	STOI (%)											
Noises	Babble				Leopard				F16			
SNR	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	56.63	66.36	75.28	66.09	73.22	78.06	82.26	77.85	57.4	66.83	75.61	66.61
LSTM [54]	70.30	78.18	83.96	77.48	81.15	84.76	87.45	84.45	74.21	80.91	85.23	80.12
BLSTM [30]	71.24	78.86	84.53	78.21	82.02	85.39	87.99	85.13	74.98	81.28	85.93	80.73
CRN [80]	71.34	79.18	84.67	78.40	82.30	85.77	88.17	85.41	75.06	81.98	86.13	81.06
GRN [30]	73.53	80.06	84.29	79.27	82.77	85.9	88.21	85.63	77.38	82.66	86.11	82.05
MCGN	**78.46**	**83.19**	**86.97**	**82.87**	**85.01**	**87.45**	**89.45**	**87.30**	**79.56**	**84.27**	**87.64**	**83.82**
Measure	PESQ											
Noises	Babble				Leopard				F16			
SNR	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	1.38	1.65	1.96	1.66	1.77	2.09	2.37	2.08	1.31	1.55	1.83	1.56
LSTM [54]	1.88	2.25	2.56	2.23	2.44	2.69	2.92	2.68	2.02	2.37	2.63	2.34
BLSTM [30]	1.91	2.29	2.60	2.26	2.50	2.75	2.98	2.74	2.10	2.41	2.67	2.39
CRN [80]	1.93	2.31	2.61	2.28	2.53	2.79	3.00	2.77	2.14	2.45	2.69	2.43
GRN [30]	2.03	2.33	2.62	2.33	2.53	2.80	3.01	2.78	2.20	2.47	2.70	2.46
MCGN	**2.27**	**2.55**	**2.77**	**2.53**	**2.71**	**2.93**	**3.10**	**2.91**	**2.35**	**2.60**	**2.81**	**2.59**

4.3.7 Kernel Size Analysis

Further experiments are performed to analyse the relation between enhancement performance and kernel sizes. These experiments use kernel size varied from 1×2 to 11×11, thus exploiting different receptive fields in the T-F domain. Table 4.9 provides the experimental results in terms of ΔSDR, STOI, and PESQ. It can be observed that the performance increases with the increase in the kernal size, e.g. from 2×2 to 7×7, but then starts to saturate for the further increase to 11×11. However, the performance difference is relatively small.

A larger kernel size can provide a larger receptive field, which helps the kernel to filter the output from a longer sequence i.e. contextual information, and a smaller kernel size helps capture the local information. However, there is a trade-off between the kernel size and performance, when the kernel size is larger than a certain value, it may cause performance degradation. Using paralleled multi-kernel helps the model to capture the features in different scales, thus exploiting both local and contextual information, as in the proposed method.

To interpret the use of different kernel sizes, an example of the feature map is provided, it is obtained by using kernels of different sizes in the first multi-scale convolutional layer, as shown in Fig. 4.8, using kernels of

Table 4.9. Kernel Size Analysis

Filte Size	ΔSDR	STOI	PESQ
1×2	10.55	72.07	1.71
2×2	10.72	72.21	1.72
2×3	10.76	72.30	1.73
4×5	10.88	72.37	1.73
5×5	11.16	72.97	1.76
7×7	11.18	73.15	1.77
11×11	11.23	73.07	1.75
Multi-Kernel	**11.72**	**76.21**	**1.92**

size 1×2, 3×4, and 7×7. It can be seen, although the kernel at each scale extracted both speech and noise components, as shown in the regions highlighted with blue and black, the feature maps obtained with these kernels characterise different receptive fields, for example, with the large kernel, more heavy smoothing is applied which is effective in mitigating the impact of noise, while the use of a small kernel can retain the fine structure of the spectrum. Therefore, using a bank of kernels, the system has a better chance to capture and distinguish the features from speech and noise, thus further improves the speech enhancement performance.

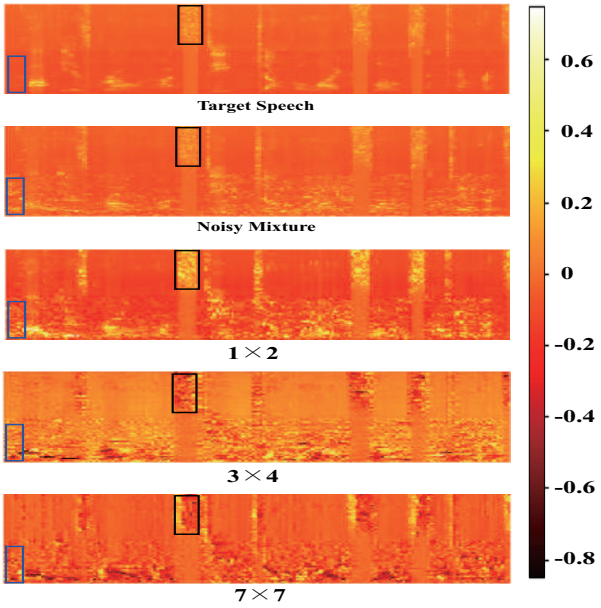


Figure 4.8. The weights obtained by feature recalibration are shown as a hot-map, where the horizontal and vertical axis denote the time and frequency, respectively, and the color-bar shows the values of the weights.

4.3.8 Component Analysis

A series of experiments are conducted to investigate the efficiency of different components in the proposed model. In the component analysis, the ablation experiments are performed, by removing different components to show how it affects the enhancement performance.

Table 4.10 provides the experimental results of using various components in terms of the ΔSDR, STOI, PESQ and the parameters (million). Full means the full MCGN framework. No bottleneck represents removing the bottleneck layers in MCGN. No FC represents removing the fully connected layers in MCGN. No MCFR means using the single kernel in each encoder-decoder layer. No CL represents removing the connection layers that include a dense layer and two BGRU layers. No FR denotes removing feature recalibration, which means that the different scaled features use the same weight and are concatenated directly.

The bottleneck layers employ fewer channels than previous layers to compress the information from previous convolutional layers, and this can reduce the computational cost with slight information loss, as shown in the experimental results. Unlike bottleneck layers in the convolutional encoder and decoder, the FC layer with non-linear activation can produce a compact representation of the encoder output before the BGRU layer is applied. The bottleneck and FC layers help capture global information from the mixture. In addition, the interdependency among the past, current and future frames is captured by the BGRU layers. Therefore, the CL can employ BGRU and FC layers to provide improvements of enhancement performance and parameter efficiency. The results also show that the MCFR module can improve the performance by capturing the features in different scales using paralleled kernels of different size.

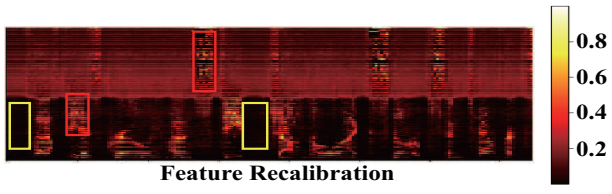


Figure 4.9. The weights obtained by feature recalibration are shown as a hot-map, where the horizontal and vertical axis denote the time and frequency, respectively, and the color-bar shows the values of the weights.

Table 4.10. Component Analysis

Measures	ΔSDR	STOI	PESQ	Parameters(Millions)
Full	10.20	81.40	2.40	77.5
No Bottleneck	10.39	81.60	2.43	133.4
No FC	10.24	81.51	2.40	123.7
No CL	9.27	77.25	2.23	41.9
No MCFR	9.12	77.42	2.20	27.9
No FR	9.61	79.87	2.32	68.8

Fig. 4.9 shows the weights obtained by feature recalibration in the last layer of the multi-scale decovolutional layer. The color-bar shows the weight values, and the deeper color represents a smaller value. Comparing Fig. 4.9 with Fig. 4.11 (A), (B), it can be observed that the weights of high values capture the target speech very well. For example, the areas highlighted with the red blocks represent speech components, while those highlighted with yellow blocks represent components from noise. It can be observed that the feature recalibration tends to assign the features from speech with higher weights, and features from noise with lower weights. Therefore, the feature recalibration helps suppress noise and improve reconstruction of the target speech.

4.3.9 Convergence Lines and Spectrums

Fig. 4.10 demonstrates the testing MSEs of the baseline methods and the propose MCGN and MCGN without multi-scale output (MCGN(NM)) layers over epochs. It can be seen that the MCGN converges faster than the baseline methods and reaches the lowest MSE. After 20 epochs training, the MCGN and MCGN(NM) offer similar MSEs, but the convergence speech of MCGN is faster than MCGN(NM) at 1-5 epochs. It shows that the multi-scale output layer accelerates the convergence speed. Furthermore, the MCGN provides better performance than the state-of-the-art baseline methods.

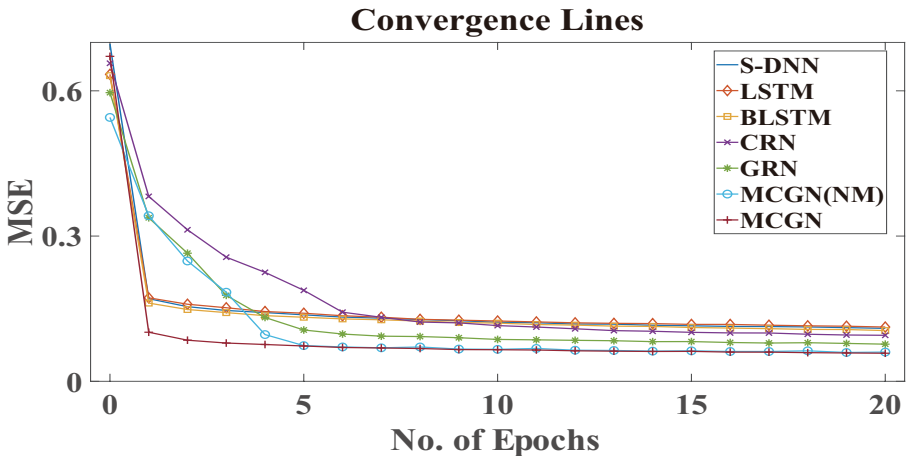


Figure 4.10. Mean squared errors over training epochs for S-DNN, LSTM, BLSTM, CRN, GRN, MCGN and MCGN(NM) on the test set. The MCGN(NM) represents the delete the multi-scale output layers, only use the normal output layer. All models are evaluated with a test set of unseen speakers.

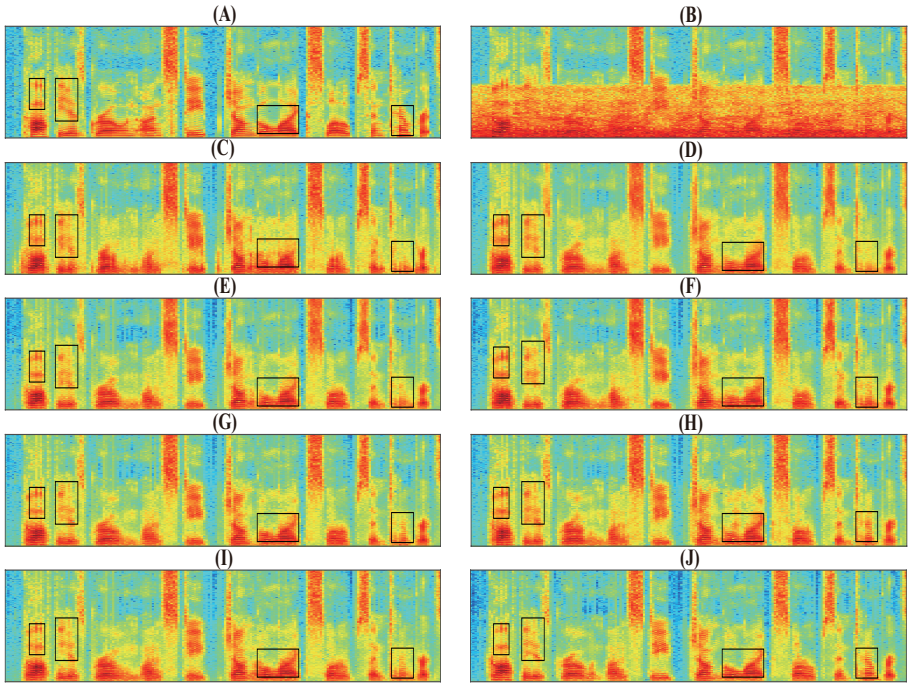


Figure 4.11. Spectrums of different signals: (A) clean speech; (B) noisy speech mixture; (C) enhanced speech by S-DNN [31]; (D) enhanced speech by LSTM [54]; (E) enhanced speech by BLSTM [30]; (F) enhanced speech by the proposed MRCAE [58] (G) enhanced speech by the proposed CRN [80]; (H) enhanced speech by the proposed CRBN; (I) enhanced speech by the proposed GRN [30]. (J) enhanced speech by the proposed MCGN.

A set of spectra are offered in Fig. 4.11. It can be seen that the baseline methods and the proposed MCGN method provide different enhancement performance in terms of reconstruction of target speech. The spectrums of the proposed MCGN method is closer to the specgtrums of the target speech, which again confirmes that the MCGN outperforms the baseline methods.

4.4 Summary

An advanced network structure named MCGN was proposed to address monaural speech enhancement. A series of novel strategies were utilized to refine enhancement performance and improve parameter efficiency. The feature in different scales were captured by using varied sized kernels in MCFR. Larger kernels captured global information. On the contrary, the local information was obtained by smaller kernels. Moreover, the feature recalibration was achieved, which exploit the gating mechanism to assign higher weights to essential features. As a result, the MCFR paid more attention to speech components and suppressed the noise components. In addition, the bottleneck convolutional and deconvolutional layers were introduced to retain the information and reduce parameters. Similarly, the FC layer was introduced to offer compressed information flow. Furthermore, the BGRU layers were utilized to capture interdependency among the past, current and future temporal frames. The unseen speakers and noises were used to examine the enhancement performance of the proposed MCGN method. The experimental results confirmed the improved performance of the proposed method overs the state-of-the-art baseline methods.

In next chapter, a convolutional fusion network will be discussed to improve the model capacity of speech enhancement method.

Chapter 5

CONVOLUTIONAL FUSION NETWORK FOR MONAURAL SPEECH ENHANCEMENT

5.1 Introduction

The aforementioned methods, however, still have limitations. For instance, although a large kernel size used can enlarge the receptive fields of the model in the conventional convolutional encoder-decoder network, it increases the computational cost [96]. The InceptionNet and MRCAE utilize multiple kernels with various sizes to improve the model capacity, it also requires larger computational resource [58], which will decrease the parameter efficiency and limits its applicability in resource-limited applications. The AlexNet uses two group convolutions in parallel at each layer, with each group taking half of the input sequence [97]. As a result, the output from a certain channel will only be partially related to the input channels, this will limit the information flowing between channels [98]. The Shuffle Network (ShuffleNet) introduces the channel shuffling to rearrange and concatenate the outputs of different groups, which entangles the information across the channels [98]. However, for each group of convolution, only part of input sequence is used, which may limit each kernel to only obtaining partial information from the

full input sequence and potentially degrade the model performance. The AECNN model only employs the skip connections between the encoder and decoder, which feds the information flow from the encoder to the decoder, but the information flow within the encoder or decoder is not considered [96].

In this chapter,a new framework is proposed, namely, convolutional fusion network (CFN) to mitigate some of these limitations. More specifically, the following contributions are offered.

First, a convolutional fusion unit is proposed consisting of standard convolution and depth-wise separable convolution with smaller kernel size. The weighted outputs from these two convolutions are concatenated as the output of the convolutional fusion unit. The convolutional fusion units are used to build the encoder, instead of using only standard (vanilla) convolution.

Second, a novel decoder is proposed that includes deconvolution, depth-wise separable convolution and upsampling layers to improve the model capacity.

Third, a full input sequence (full information) based channel shuffling is introduced to exploit the inter-channel dependency. More specifically, the full input sequence is fed to standard convolution and depth-wise separable convolution, and their outputs are re-arranged and concatenated to utilize the inter-channel dependency according to the channel order. As a result, both groups of convolution can exploit the full information from the input sequence.

Lastly, intra skip connection mechanisms are applied inside the encoder and decoder. With intra skip connection, the ability of reusing information flow within the encoder and decoder is refined.

The remainder of this chapter is organized as follows. Section 5.2 presents the proposed CFN method. The experimental settings and results are discussed in Section 5.3. Section 5.4 draws the conclusions.

This chapter focuses on the third objective of this thesis, which relate to

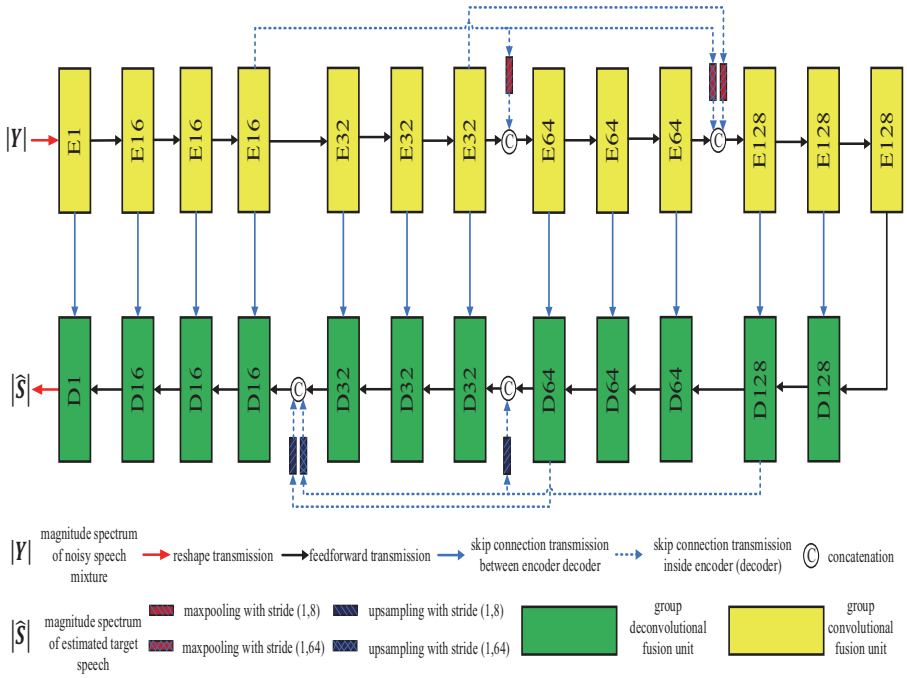


Figure 5.1. The architecture diagram of the proposed CFN. The components and their function are listed in bottom of the figure. For example, $E64$ represent GCFU with 64 output channel in the encoder, and $D64$ represents GDFU in the decoder. The encoder is on the top of the figure, and the decoder is on the bottom of the figure. The kernel sizes of standard convolution and deconvolution are set as (1,3), and their strides sizes are set to (1,2). The kernel size of depth-wise separable convolution is set to (3,3), and strides size is (1,1). The pooling layer with strides size (1,2) is used to reduce the dimension of depth-wise separable convolution output in GCFU, and upsampling layer with size (1,2) is employed to increase its dimension in GDFU. Besides, the last layer of encoder uses stride size (1,1) for convolution and depth-wise separable convolution, and pooling with strides (1,1) for depth-wise separable convolution.

the covolutional fusion structure method for monaural speech enhancement submitted to Elsevier Journal of Neural Network.

5.2 Algorithm of CFN Method

5.2.1 Proposed Network Architecture

The convolutional encoder-decoder structure with multiple skip connections are exploited for monaural speech enhancement. The details of the proposed CFN are shown in Fig. 5.1. The proposed CFN takes the magnitude spectrum of the noisy spectrum as input, and outputs the magnitude spectrum of estimated target speech. The estimated target speech is reconstructed by using the estimated magnitude of the target speech and the phase information of the noisy speech mixture. The encoder has multiple layers of group convolutional fusion units, and each unit consists of a standard convolution and depth-wise separable convolution. The number of output channels of the group convolutional fusion units is increased from 16 to 128 in the encoder. The encoder is applied to reduce the dimension of input sequence by using the strides in group convolutional fusion units. The decoder has a mirror structure with the encoder, and each group deconvolutional fusion unit consists of standard deconvolution and depth-wise separable deconvolution. The decoder is used to recover the dimension of the encoder output and generate the final output. In addition, multiple types of skip connections are applied to improve feature re-use. More specifically, the group deconvolutional fusion unit of the decoder is connected with the output from the corresponding symmetric group deconvolutional fusion unit of the encoder by skip connection. Furthermore, the skip connections are used to connect different group convolutional/deconvolutional fusion units inside the encoder or decoder.

5.2.2 Group Convolutional Fusion Units

The CFN is proposed by employing group convolutional fusion units (GCFU) of two different network architectures, i.e. standard convolution and depth-

wise separable convolution, and the weighted outputs of the two convolutions are concatenated together. The details of the proposed GCFU is shown in Fig. 5.2. For every unit, the input matrix $\mathbf{X} \in \mathbb{R}^{H \times W \times M}$, where H and W represent heigh and width of matrix respectively. M represents channels of matrix.

A standard 2D-convolutional layer can be characterized by an input $\mathbf{X}$, and a bank of filters $\mathbf{F} \in \mathbb{R}^{K \times L \times M \times N}$, K and L represent the width and length of kernel respectively, N represents the number of filters i.e. the number of output channels of the standard 2D-convolutional layer. The operation of the standard 2D-convolutional layer is:

$$\mathbf{C}_{(k,l,n)} = \sum_{i=1}^{K} \sum_{j=1}^{L} \sum_{m=1}^{M} \mathbf{F}_{(i,j,m,n)} \mathbf{X}_{(k+i-1,l+j-1,m)} \tag{5.2.1}$$

The output of the standard 2D-convolutional layer is $\mathbf{C} \in \mathbb{R}^{H \times W \times N}$. Also, the stride sizes are used to control the output size of $\mathbf{C}$. The batch normalization and activation function LeakyReLU [87] are followed to generate the 2D-convolution output.

Different from the standard 2D-convolution, the depth-wise separable convolution has two steps: depth-wise convolution i.e. a spatial convolution performed independently over every input channel, and the point-wise convolution i.e. a standard convolution, which projects every channel's output of the depth-wise convolution to a new channel space. Mathematically, the filter $\mathbf{F}$ are spilt into two filters, the depth-wise filter $\mathbf{D} \in \mathbb{R}^{K \times L \times 1 \times 1}$, and point-wise filter $\mathbf{P} \in \mathbb{R}^{1 \times 1 \times M \times N}$.

$$\begin{aligned} \mathbf{S}_{(k,l,n)} &= \sum_{i=1}^{K} \sum_{j=1}^{L} \sum_{m=1}^{M} \mathbf{F}_{(i,j,m,n)} \mathbf{X}_{(k+i-1,l+j-1,m)} \\ &= \sum_{i=1}^{K} \sum_{j=1}^{L} \sum_{m=1}^{M} \mathbf{D}_{(i,j,m)} \mathbf{P}_{(m,n)} \mathbf{X}_{(k+i-1,l+j-1,m)} \end{aligned}$$

$$= \sum_{m=1}^{M} \mathbf{P}_{(m,n)} \left(\sum_{i=1}^{K} \sum_{j=1}^{L} \mathbf{D}_{(i,j,m)} \mathbf{X}_{(k-i,l-j,m)} \right) \tag{5.2.2}$$

The output of the depth-wise separable convolutional layer is $\mathbf{S} \in \mathbb{R}^{H \times W \times N}$. Similarly, the batch normalization and activation function LeakyReLU [87] are followed after depth-wise separable convolutional layers. In addition, the max pooling operation is used to down-sample the output of the depth-wise separable convolution.

Different from the conventional residual structure that sums two output convolutions [30]. The convolutional fusion is realized by concatenating the weighted outputs of the two convolutions:

$$\mathbf{B} = [\alpha_1 \mathbf{C}, \alpha_2 \mathbf{S}] \tag{5.2.3}$$

where α_1 and α_2 represent the weight parameters of the standard and depth-wise separable convolutions.

5.2.3 Full Information Channel Shuffle

The concept of group convolution is first proposed in AlexNet [97], which employs two parallel convolutions in each layer. Furthermore, the group convolution has been demonstrated its effectiveness in many works [55, 88]. In the proposed method, the outputs of these two parallel convolutions are concatenated to generate the final output.

For the next layer, the output of a certain channel is only related to a small fraction of the input channels, and information flow between channels is limited in conventional group convolution [98]. Therefore, motivated by the idea in [98], a strategy is introduced for shuffling group channels in the proposed CFN model, which allows each channel of the next layer to obtain information flow from two kinds of convolutions. In addition, the input and output of this layer will be related. In the original group channel shuffle

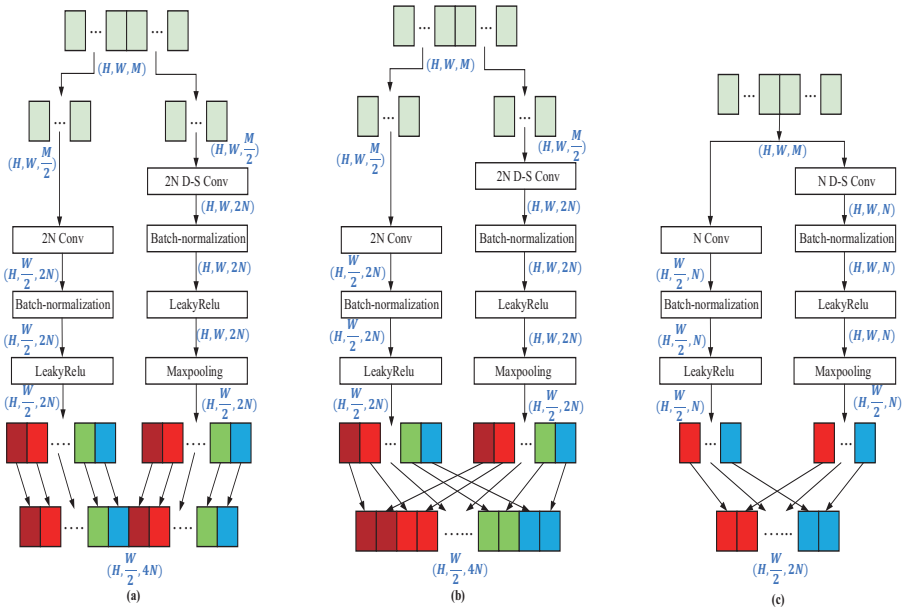


Figure 5.2. (a) Conventional group convolution that consists of the standard convolution and depth-wise separable convolution. $2N$ represents the number of output channels of convolutional layers. The standard convolution and depth-wise separable convolution take half input sequence as input, each of them generates output with $2N$ channels, and the output of convolution and depth-wise separable convolution are concatenated directly. (b) Group convolution with channel shuffle. The input sequence is divided into two parts based channel index, the first part of the input sequence is feed to the standard convolution, and the second part of the input sequence is fed to the depth-wise separable convolution layer. Their outputs are re-arranged the channel and concatenated by using the channel shuffle, and final output with $4N$ channels. (c) Proposed full information channel shuffle. The full input sequence (full information) is fed to the N channel standard convolution and N channel depth-wise separable convolution. Their outputs are re-arranged channel and concatenated by using the channel shuffle, and final output with dimension $2N$.

method [98] as shown in Fig. 5.2.(b), the input sequence is divided into two parts, where the first part is fed to the standard convolution, and the second part is fed to the depth-wise separable convolution. Therefore, the standard convolution and depth-wise separable convolution cannot fully utilize the input sequence, which causes the model capacities of two kinds of convolutions to be not well utilized. To address this problem, in the proposed

GCFU, a new structure is designed to exploit the full information channel shuffle as shown in Fig. 5.2.(c). The full input sequence (full information) is fed to both standard convolution and depth-wise separable convolution in the proposed structure, different from Fig. 5.2.(b). They are employed to generate the feature representations for the full input sequence. As shown in Fig. 5.2.(c), the outputs of standard convolution and depth-wise separable convolution are re-arranged and concatenated according to the channel order.

Both the standard convolution and depth-wise separable convolution have N output channels, and they can be represented as $\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2..., \mathbf{C}_N]$ and $\mathbf{S} = [\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_N]$. The full information channel shuffle of GCFU is:

$$\mathbf{B}_s = [\alpha_1\mathbf{C}_1, \alpha_2\mathbf{S}_1, ..., \alpha_1\mathbf{C}_N, \alpha_2\mathbf{S}_N] \tag{5.2.4}$$

where $\mathbf{B}_s$ represents the channel shuffled group convolution. By using the full information channel shuffle, the output of the standard convolution and depth-wise convolution are fully related, and the next layers can obtain the shuffled information flow. In addition, the channel shuffle enables us to halve the number of convolutions output channels as shown in Fig. 5.2.(c).

5.2.4 Group Deconvolutional Fusion Units

In the convolutional encoder-decoder structure, the decoder is exploited to map the low dimension encoder feature representation to the original dimension of the input sequence that with a higher dimension. The standard decoder uses deconvolutional layers to up-sample the encoder output. However, there is no depth-wise separable deconvolution structure. A group deconvolutional fusion unit (GDFU) is proposed to up-sample the encoder feature map, which can increase the feature dimension. The GDFU architecture is shown in Fig. 5.3. The low dimensional feature representation is fed to

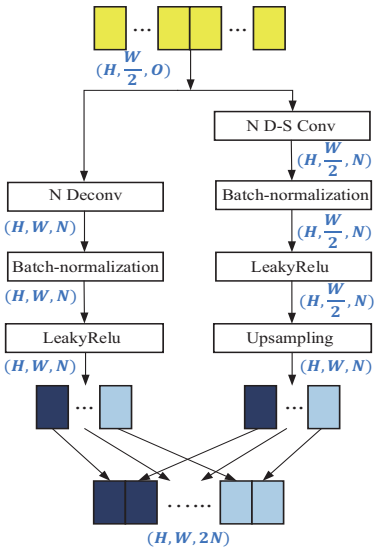


Figure 5.3. Structure diagram of the proposed GDFU with full information channel shuffle. The full input sequence is fed to the standard deconvolution and depth-wise separable convolution. N Deconv represents N channel standard deconvolution, and N D-S Conv denotes the N channel depth-wise separable convolution.

deconvolutional and depth-wise separable convolutional layers respectively, to generate the dense feature representations. The batch normalization and LeakeyRelu [87] steps are followed. Inspired by the work [99], which uses the transferred pooling layers and standard convolutional layers to build the convolutional decoder, the upsampling layer is used to up-sample the feature representation of the depth-wise separable convolutional layers. Finally, full information channel shuffle is utilized to rearrange the outputs of standard deconvolution and depth-wise separable convolution as shown in Fig. 5.3.

5.2.5 Skip Connection inside Encoder or Decoder

In the convolutional encoder-decoder, the input sequence is processed with many layers. Some information may have lost due to the variations in the dimension of feature representation of the signal [96]. The skip connections between encoder and decoder can be introduced to address this issue. For example, in AECNN [96], the encoder layers are connected with the decoder

layers. Unlike this work,the skip connection mechanism is proposed as shown in Fig. 5.1, to connect the layers inside the encoder (decoder), which facilitates the feature re-use inside the encoder or decoder. However, densely connecting all the layers will significantly increase the computational cost. Therefore, block dense connections are proposed, where the encoder layers with the same number of output channels are set as a block, e.g. block-16, block-32, block-64 and block-128. For instance, the output of block-16 is fed to the other blocks (block-32, block-64 and block-128) of the encoder. Since GCFU has stride size $(1, 2)$ in the encoder layers, the output sizes of different layers are varied, as a result, the output of block-16 cannot be directly concatenated with the output of the other blocks. Therefore, a new mechanism is designed to down-sample the features of block-16 using a max-pooling layer with the stride of size (1, 8). Then, the down-sampled output is concatenated with the output block-32, concatenated representation is fed to block-64. Similarly, other skip connections are developed within the encoder as shown in Fig. 5.1. On the contrary, the layers are up-sampled in the decoder to match the size of the skip connections.

5.3 Experimental Evaluations

5.3.1 Data and Setup

To build the training and testing data, the clean utterances from TIMIT [64] and IEEE [65] corpora are mixed with the environmental noises from the NOISEX-92 [67] dataset. 1500 utterances are randomly selected from TIMIT and IEEE corpora as the training utterances. In addition, 100 utterances are selected from TIMIT corpus as the testing utterances, and the speakers of testing utterances are different from the speakers of training utterances, which is speaker-dependent case. The training and testing utterances are mixed with the Babble, Artillery, Airplane, Factory, Tank, and White noises.

The noises' names indicate their recording environments, and they are four minutes long. The training and testing datasets are generated with three signal-to-noise ratio (SNR) levels i.e. -5dB, 0dB and 5dB.

Furthermore, the proposed method is evaluated with speaker- and noise-independent cases. These experiments aim to evaluate the performance of the proposed method under the challenging mismatch conditions. 200 utterances are randomly selected from the TIMIT corpus [64] as the training utterances, and 100 utterances from the TIMIT corpus as the testing utterances, and the speakers of testing utterances are different from the speakers of training utterances. Three types of unseen noise i.e. Water, Wind and Pink noises, are chosen as the testing noises from the Non-Speech Sounds and NOISEX-92 datasets. 108 noises that consist of 98 noises from Non-Speech Sounds dataset and 10 from NOISEX-92 dataset are used as the training noises. The Non-Speech Sounds dataset contains 100 environmental noises. The training noises are mixed with the training speeches in SNR levels of -5dB, 0dB and 5dB. Similarly, the testing speeches are mixed with three unseen environmental noises at SNR levels of -5dB, 0dB and 5dB to generate the testing dataset. In total, 60 hours ($1500\times6\times3\times2.5\div3600+200\times100\times3\times2.5\div3600=60.42$) noisy speech mixtures are used to train the proposed model, and about 2 hours noisy speech mixtures to test the proposed model.

The parameters of the CFN are shown in Fig. 5.1 and its caption. The signals in the training and testing datasets are re-sampled at 16 kHz. The magnitude spectrum of these signals is obtained by using STFT with Hanning window of 512 samples and 50% overlap between the neighboring windows, and then log-compressed. The MAE is used as the cost function for the baselines (discussed in the next sub-section) and the proposed CFN methods, and the Adam optimization algorithm with 0.0001 initial learning rate [93] is employed. The best models are selected. The training and testing

processes are executed by GeForce GTX-1080. For quantitative evaluation, short-time objective intelligibility (STOI) [62] and perceptual evaluation of speech quality (PESQ) [61] are used to measure the enhancement performance. The STOI indicates the intelligibility quality of the estimated target speech which ranges in (0, 1), and the PESQ indicates the perceptual quality of the estimated speech which ranges in (0, 4.5). The higher value of the measurements indicates better enhancement performance.

5.3.2 Baseline Methods

Three state-of-the-art speech enhancement methods are used as the baselines, and they are, respectively, the DNN in [25], GRN in [30], and AECNN in [96]. DNN is a fundamental method in deep learning, and the GRN and AECNN show advantages over RNN. DNN has four hidden layers, and each hidden layer has 1024 units. Also, the dropout with a rate of 0.2 is used in DNN to reduce the over-fitting [100]. The output layer of DNN has the same number of units as the length of the input sequence. The GRN model is a 62-layered deep fully connected convolutional model with residual connections. The stacked convolutional layers use gated convolution with an increased dilated ratio. The dilated convolution offers larger receptive fields, which enables each kernel to filter out information on longer-term of the sequence than standard convolution. Also, the Sigmoid activation function follows the dilated convolution to build a gate mechanism to control the information flow in GRN. Finally, the prediction module takes the information flow from the stacked dilated convolution layers by the feed-forward and skip connections, and generates the magnitude spectrum of the estimated target speech.

The AECNN is a 18-layered convolutional encoder-decoder structure. The convolutional encoder is exploited to reduce the dimension of the input magnitude spectrum by using the convolutional layers with strides sized 2.

The deconvolutional decoder has a minor structure with the convolutional encoder, which is employed to recover the dimension of the output of the convolutional encoder to the original dimension i.e same as the input noisy speech magnitude spectrum. The number of output channel of the convolutional encoder is increased from 64 to 256, but the number of the convolution decoder is reduced from 256 to 64, and the output layer of the AECNN has one channel. The layers of the encoder are connected with layers of the decoder that has the same number of the output channels by skip connections. The MAE between the magnitude spectrum of noisy speech and the estimated target speech is employed in AECNN. To make a fair comparison, magnitude spectrum of 257 units are fed into the baseline methods and the proposed CFN, and they output the magnitude of estimated target speech. The same training and testing datasets are employed for the baselines and the proposed method. The number of parameters for the baseline methods: DNN (5.5 Million), GRN (2.5 Million), AECNN (6.4 Million), and the number of parameters of the proposed CFN is 3.5 Million.

5.3.3 Experimental Results for Seen Noises

Tables 5.1 & 5.2 provide comparisons among the proposed CFN and the baseline methods in terms of STOI and PESQ for speaker-independent case with seen Babble, Artillery, Airplane, Factory, Tank, and White noises. The DNN offers, on average, STOI = 75.39% and PESQ = 2.14, which provides the lowest improvement over the noisy speech mixture across all compared methods. The results show that the DNN under-performs for speaker-independent speech enhancement, where the speakers in the test set are different from those in the training set, as compared with other method including the proposed method.

The GRN provides, on average, STOI = 78.69% and PESQ = 2.27. The GRN offers further improvement over the DNN methods. The GRN utilizes

the dilated convolutional layers to enlarge the receptive fields, which means one kernel (filter) can take information from a longer sequence and generate the output. Therefore, the temporal information from the long-term frames is captured. Also, the convolutional layer with Sigmoid is employed to build the gated mechanism to control the information flow in GRN. In addition, residual learning is employed by using the skip connections among the different layers of GRN. By joint using these strategies, the GRN offers a better enhancement performance in terms of STOI and PESQ, as compared with DNN in the speaker-independent speech enhancement.

The AECNN provides, on average, STOI = 79.75% and PESQ = 2.34, which outperforms GRN and DNN methods, which is consistent with the finding in [96]. The AECNN employs a speech encoder-decoder structure to estimate the magnitude spectrum of target speech. The convolutional encoder takes the magnitude spectrum of the noisy speech mixture as input, which generates a lower dimension output. The convolutional decoder is utilized to recover the dimension of the encoder output. In addition, MAE between the magnitude spectrums of the estimated target speech and the original target speech is utilized as the cost function. The experimental results show the AECNN i.e. convolutional encoder-decoder is an advanced method over DNN and GRN methods.

Table 5.1. Speech enhancement performance comparison in terms of STOI and PESQ for speaker-independent case with Babble, Artillery, Airplane noises. *Italic* text is the proposed method. **Bold** number indicates the best performance.

Measures	STOI(%)											
Noises	Babble				Artillery				Airplane			
SNR / Methods	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	54.62	63.82	72.47	63.64	65.34	73.02	79.66	72.67	54.57	64.13	73.42	64.04
DNN	68.89	72.98	79.69	73.85	74.62	79.29	82.92	78.94	68.18	74.85	80.48	74.50
GRN	69.76	76.89	81.42	76.02	77.80	82.31	85.10	81.74	72.70	78.60	83.10	78.13
AECNN	72.01	77.78	82.51	77.43	79.62	83.68	86.59	83.30	73.87	77.99	84.43	78.76
CFN	**75.67**	**80.33**	**83.85**	**79.95**	**81.81**	**85.88**	**87.43**	**85.04**	**77.55**	**82.15**	**85.56**	**81.77**
Measures	PESQ											
Noises	Babble				Artillery				Airplane			
SNR / Methods	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	1.37	1.63	1.92	1.64	1.66	1.94	2.18	1.93	1.36	1.59	1.87	1.61
DNN	1.77	2.08	2.34	2.06	2.10	2.33	2.54	2.32	1.78	2.12	2.37	2.09
GRN	1.86	2.16	2.42	2.15	2.20	2.47	2.70	2.46	1.93	2.25	2.55	2.24
AECNN	1.92	2.19	2.45	2.19	2.32	2.55	2.75	2.54	2.03	2.32	2.57	2.31
CFN	**2.16**	**2.41**	**2.62**	**2.40**	**2.49**	**2.68**	**2.86**	**2.68**	**2.24**	**2.51**	**2.73**	**2.49**

Table 5.2. Speech enhancement performance comparison in terms of STOI and PESQ for speaker-independent case with Factory, Tank, White noises. *Italic* text is the proposed method. **Bold** number indicates the best performance.

Measures	STOI(%)											
Noises	Factory				Tank				White			
SNR / Methods	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	54.17	63.57	74.07	63.94	72.24	76.17	79.54	75.98	53.26	62.56	72.42	62.75
DNN	62.00	69.76	76.09	69.28	80.34	82.67	83.14	82.05	67.28	74.10	79.90	73.76
GRN	68.06	74.98	80.42	74.49	81.37	83.86	85.42	83.55	72.32	78.36	82.50	77.73
AECNN	69.72	75.77	81.72	75.74	83.57	84.67	86.17	84.48	73.11	79.34	83.00	78.48
CFN	**71.61**	**78.19**	**86.20**	**78.67**	**84.64**	**86.26**	**87.31**	**86.07**	**76.29**	**81.01**	**85.02**	**80.77**
Measures	PESQ											
Noises	Factory				Tank				White			
SNR / Methods	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	1.33	1.61	1.88	1.61	1.76	1.99	2.21	1.99	1.13	1.31	1.57	1.34
DNN	1.67	1.97	2.22	1.95	2.37	2.57	2.60	2.51	1.72	2.05	1.95	1.91
GRN	1.78	2.09	2.35	2.07	2.43	2.60	2.74	2.59	1.81	2.17	2.38	2.12
AECNN	1.80	2.10	2.40	2.10	2.58	2.76	2.83	2.72	1.95	2.25	2.40	2.20
CFN	**1.98**	**2.24**	**2.63**	**2.28**	**2.72**	**2.86**	**2.99**	**2.86**	**2.20**	**2.44**	**2.64**	**2.63**

Table 5.3. The p-value of the t-test at 5% Significance Level, and comparison of proposed method with the baseline methods for speaker-independent case. H_0 denotes the null hypothesis, and (+) indicates the improvement of two pairs is statistically significant at the 95% confidence level.

Measures	STOI		PESQ	
	p-value	H_0	p-value	H_0
Noisy	1.63E-10	(+)	1.29E-14	(+)
DNN	1.75E-10	(+)	5.34E-12	(+)
GRN	4.33E-10	(+)	1.85E-12	(+)
AECNN	1.21E-07	(+)	5.53E-12	(+)

The proposed CFN method offers, on average, STOI = 82.20% and PESQ = 2.52, which provides over 2.45% STOI improvement and 0.17 PESQ improvement over the AECNN, GRN and DNN methods. It shows advantages in processing speaker-independent speech enhancement. In addition, the CFN uses a fewer number of parameters, thus offers a higher parameter efficiency. The reason will be discussed in the next subsection.

To further evaluate whether the improvement in terms of STOI and PESQ is statistically significant, the t-test is performed between proposed CFN with baseline methods and noisy speech mixture at a significant level of 0.05 in Table 5.3. The t-test is performed following statistical analysis in [95]. When p-values smaller than 0.05, it means there is statistical significant difference between values of two group. All p-values are smaller than 0.05, and all H_0 are +, which confirms that the improvements by the proposed CFN over the baselines are statistically significant.

5.3.4 Experimental Results for Unseen Noises

Table 5.4 provides comparisons among the proposed CFN and baseline methods in terms of STOI and PESQ for speaker- and noise-independent cases with unseen Water, Wind, Pink noises.

Similarly, experimental results of speaker- and noise-independent cases

show similar trends of enhancement performances. The DNN offers the worst enhancement performance in terms of STOI and PESQ, which shown the limitation of DNN in processing challenging speech enhancement problems. The GRN offers improvements over the DNN method. Also, the AECNN outperforms DNN and GRN methods, which yields, on average, STOI = 79.07% and PESQ = 2.06.

The proposed CFN method yields the best enhancement performance, on average, STOI = 81.85% and PESQ = 2.25. Several contributions are exploited to boost enhancement performance. The CFN uses standard convolution and depth-wise separable convolution to produce the representation of feature, which reinforce the model capacity of the proposed CFN method. Also, a novel decoder that consists of deconvolution and depth-wise separable convolution is employed to up-sample the encoder output. In addition, full information channel shuffle structure is designed to reduce the parameters and improve the channel related. Also, the two types of skip connections are introduced to improve the feature re-use, especially the intra skip connections are employed in encoder or decoder, to make proceeding layers of the encoder can receive more information from previous layers of the encoder. By combining using the aforementioned contributions, the CFN shows advantages over the DNN, GRN and AECNN for noise- and speaker-independent cases. Also, the results of the t-test in Table 5.5 demonstrates the proposed CFN method yields statistically significant improvements over the baseline methods.

Table 5.4. Speech enhancement performance comparison in terms of STOI and PESQ for speaker- and noise-independent cases with Water, Wind and Pink noises. *Italic* text is the proposed method. **Bold** number indicates the best performance.

Measures	STOI(%)											
Noises	Water				Wind				Pink			
SNR / Methods	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	56.10	67.18	77.22	66.83	71.50	77.37	82.35	77.07	53.75	63.17	72.67	63.20
DNN	72.36	78.08	82.95	77.80	74.75	80.68	84.70	80.04	60.91	66.64	74.97	67.51
GRN	74.85	80.00	86.04	80.29	76.55	82.24	87.13	82.01	63.79	70.97	76.37	70.37
AECNN	76.76	81.78	87.09	81.88	78.72	84.55	87.68	83.65	64.97	71.45	78.65	71.69
CFN	**79.58**	**84.12**	**88.22**	**83.98**	**82.57**	**86.32**	**88.90**	**85.93**	**69.33**	**75.60**	**82.02**	**75.65**
Measures	PESQ											
Noises	Water				Wind				Pink			
SNR / Methods	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	1.22	1.36	1.60	1.40	1.54	1.75	1.95	1.75	1.38	1.44	1.75	1.52
DNN	1.77	1,97	2.15	1.96	1.76	1.99	2.18	1.98	1.56	1.79	2.06	1.80
GRN	1.79	2.00	2.28	2.02	1.82	2.02	2.24	2.02	1.66	1.90	2.18	1.91
AECNN	1.86	2.05	2.32	2.07	1.89	2.09	2.27	2.08	1.79	2.02	2.26	2.02
CFN	**2.03**	**2.22**	**2.44**	**2.23**	**2.04**	**2.27**	**2.46**	**2.26**	**1.92**	**2.20**	**2.45**	**2.19**

Table 5.5. The p-value of the t-test at 5% Significance Level, and comparison of proposed method with the baseline methods for speaker- and noise-independent cases. H_0 denotes the null hypothesis, and (+) indicates the improvement of two pairs is statistically significant at the 95% confidence level.

Measures	STOI		PESQ	
	p-value	H_0	p-value	H_0
Noisy	7.03E-5	(+)	1.04E-06	(+)
DNN	1.21E-06	(+)	2.68E-07	(+)
GRN	2.24E-05	(+)	8.63E-08	(+)
AECNN	1.45E-04	(+)	5.55E-8	(+)

5.3.5 Ablation Analysis and Spectrums

The ablation analysis is realized by removing certain components in proposed network to show the contribution of the component for overall system. More specifically, the ablation analysis is performed in Table 5.6 to show the contribution of every component in the proposed CFN. Full denotes results of the proposed CFN method. No SC denotes deleting the standard convolution, No D-SC represents ablating the depth-wise separable convolution, No ISCED is deleting the intra skip connections of encoder(decoder).

Table 5.6. Ablation analysis in terms of STOI, PESQ and number of parameters.

Measures	STOI	PESQ	No. of Parameters
Full	70.18	1.73	3.5 Million
No SC	65.89	1.57	1.2 Million
No D-SC	66.12	1.55	0.6 Million
No ISCED	69.31	1.70	3.1 Million

The standard convolution yields the most improvements in terms of STOI and PESQ, which proves the standard convolution has a better model capacity over the depth-wise separable convolution in the proposed CFN. Meanwhile, the depth-wise separable convolution has similar importance when compared with standard convolution. However, there are around 4% STOI

and 0.2 PESQ performance decrease when using the standard convolution or depth-wise separable convolution. These results confirm the standard convolution and depth-wise Separable convolution are limited in processing mismatch speech enhancement, but the proposed CFN is capable to provide a better model capacity for speech enhancement. Besides, the intra skip connections of encoder or decoder also have the contribution to enhancement performance, it demonstrated the layers of CFN may not well reconstruct the input sequence, and the intra skip connections fed more information from previous layers which promote the feature re-use in the proposed CFN model.

Fig. 5.4 shows the spectrums of target speech, noisy mixture and enhanced speech of different methods. DNN, GRN, AECNN and the proposed CFN remove most of the noise from the noisy mixture. Meanwhile, enhanced speeches by DNN, GRN and AECNN remain some noise in the low frequency region. The enhanced spectrum of CFN is the closest to that of target speech, which confirms it can remove the noise from the noisy speech mixture successfully and provides the best performance over the DNN, GRN and AECNN methods.

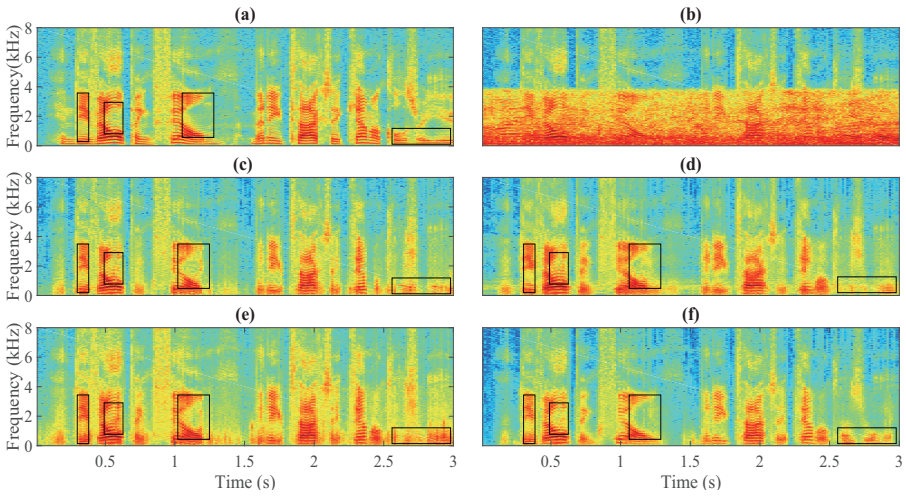


Figure 5.4. Spectrums of different signals: (a) target speech, (b) noisy speech mixture, (c) enhanced speech by DNN, (d) enhanced speech by GRN, (e) enhanced speech by AECNN, (f) enhanced speech by CFN.

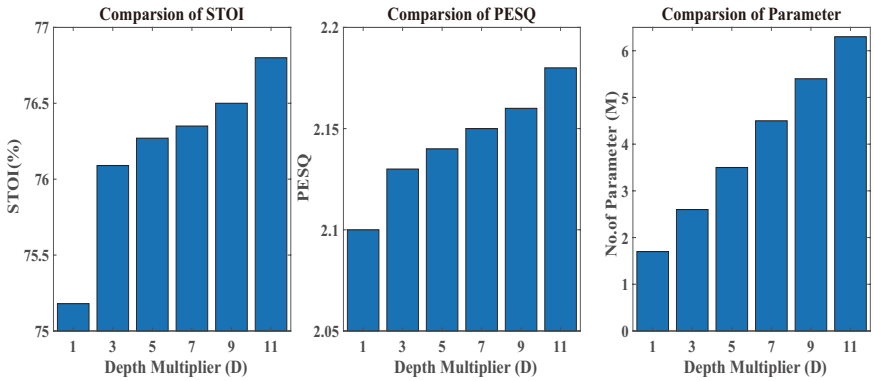


Figure 5.5. The STOI performance, PESQ performance and the number of parameters with different depth multipliers. The depth multiplier D with increment 2 from 1 to 11.

5.3.6 Depth Multiplier of Depth-wise Separable Convolution

The experiments is offered to analyze the effects of depth multiplier in depth-wise separable convolution, the experimental results are provided in Fig. 5.5.

The depth multiplier represents the number of depth-wise convolution output channels for each input channel. The target of this series of experiments aims to find the balance between speech enhancement performance and depth multiplier i.e the parameter efficiency. The experimental results are shown in Fig. 5.5. With the increase D value, speech enhancement performance in terms of STOI and PESQ is improved. However, a larger number of parameters is needed, which means it will need more computational resource. D=1 offers, on average, STOI = 75.18% and PESQ = 2.10, which provides the lowest enhancement performance but requires the fewest number of parameters around 1.7 Million. However, when D=11, it provides on average, STOI=76.83% and PESQ = 2.17, which offers the highest enhancement performance but needs more parameters around 6.3 Million. When D increases to 3, significant improvements are observed in terms of STOI and PESQ. If D larger than 5, the improvements of STOI and PESQ become stable. For example, the D=5 offers on average STOI=76.27%, and D=7 of-

fers STOI= 76.35%. In summary, taking the number of parameters, memory size and enhancement performance, D=5 in the proposed CFN model.

In addition, when comparing the proposed CFN methods with the aforementioned DNN and CNN methods, the proposed CFN obtains better performance and has lesser parameters, which proves CFN has better parameter efficiency. More specifically, the number of parameters for different methods are: DNN-DRM (5.5M), LSTMs (11.8M), CFN (3.5M). The DRM and LSTMs are designed for speech separation in reverberation environments. CFN is designed for speech enhancement, which uses several strategies to further improve parameter efficiency.

5.4 Summary

A novel network model was offered as CFN to address the monaural speech enhancement problem. The CFN considered the speech enhancement problem as a sequence-to-sequence problem. Therefore, the CFN was exploited to finding the mapping relation between the spectrums of the noisy speech mixture and clean target speech. The model capacity, inter-channel dependency, parameter efficiency was improved by using the CFN model. More specifically, the standard convolution and depth-wise separable convolutions were used to build convolutional fusion units. Then, the group channel shuffle was introduced to reinforce the interdependency among different channels. Furthermore, the skip connections were utilized in both the encoder and decoder to promote feature re-use. The dataset with unseen speakers and noises was exploited to test the proposed CFN model. The experimental results confirmed the proposed CNF model shows advantages over state-of-the-art methods.

Chapter 6

CONCLUSIONS AND FUTURE WORK

In this chapter, the contributions of this thesis are summarized in Section 6.1, and the suggestions for future work are given in Section 6.2.

6.1 Conclusions

This thesis contributed neural network-based methods to address the monaural speech enhancement and separation problems. These advanced methods are designed to improve the enhancement performance, generalization ability and model capacity with the real environment. The proposed methods were evaluated over benchmark datasets generated by TIMIT, IEEE, VCTK, NOISEX-92, DEMAND database. Besides, they are compared with state-of-the-art methods.

In this thesis, four methods were proposed to achieve these targets, the advanced training targets, system structure and neural network model architectures were introduced in these algorithms. The contributions satisfy the three objectives. The first contribution offered two methods, which introduced new training targets and network structure to exploit spatial and temporal information to address the speech separation and enhancement in reverberant environment. The second contribution offered MCGN method to capture the weighted multi-scale features, and interdependency among

different frames for speech enhancement. The third contribution offered CFN method to employ depthwise separable convolution and channel shuffle, which improved the generalization and model capacity.

In Chapter 3, in the first proposed method, the geometric information of target speech and microphone was used to estimate the direct path impulse response. Furthermore, DRM was obtained by using the energy of direct path target speech and noisy mixtures. The direct path target speech was obtained by multiplying DRM with the noisy mixture, which could denoise from noisy mixture. Finally, the estimated target speech was recovered from the direct path speech, which realized dereverberation. The experimental results confirmed the DRM outperforms the IRM in terms of speech enhancement. In the second method, the paralleled LSTMs were exploited to estimate the DRM and IRM, respectively. More specifically, the LSTMs could keep and memory the information of past time frames even with long-term intervals, as a result, the interdependency among the past and current frames was extracted. The first LSTM was used to estimated DM, which could remove the reverberation. Simultaneously, the IRM was estimated by the second LSTM to remove the noisy component from the mixture. The experimental results confirmed the proposed paralleled LSTM offered performance improvement over the baseline method. Meanwhile, they proved the LSTM capture past information to improve enhancement performance. The experimental results proved the DRM obtained over 2.5% STOI and 0.1 PESQ improvements over IRM. In additional, the parallel LSTMs offered more than 1dB SDR improvements over DNNs.

In Chapter 4, a new framework was presented for monaural speech enhancement. The proposed MCGN introduced several novel strategies to improve speech enhancement performance and computational efficiency. Firstly, the MCFR structure was introduced to extract the features in different scales, capturing both the local and contextual information from the

speech mixtures. In addition, the feature recalibration network was implemented using the gated mechanism to control the information flow, and to assign different weights to multi-scale feature. Also, the skip connections between the convolutional encoder-decoder were exploited to alleviate performance degradation. Secondly, bottleneck convolutional and deconvolutional layers were introduced to reduce information flow dimension in encoder and decoder, but to retain the information. Thirdly, the efficiency connection module was introduced. The fully connected layer was used to reduce the dimension of the output of the convolutional encoder. The BGRU was exploited to capture the interdependency among the past, current and future temporal frames, which provides comparable performance with fewer parameters than BLSTM. Finally, we introduced the multi-scale convolutional output layer, then summed the multi-scale outputs to accelerate the convergence speed. A variety of noises were used to examine the enhancement performance of the system. The unseen speakers with the seen and unseen noises were exploited to evaluate the efficacy of the proposed method. The experimental results confirmed the improved performance of the proposed MCGN method provided over 1.3dB SDR, 2% STOI, and 0.14 PESQ improvements over the state-of-the-art baseline methods.

In Chapter 5, a novel convolutional model was proposed, named convolutional fusion network (CFN), to address the monaural speech enhancement problem. Speech enhancement was considered as a sequence-to-sequence problem by the CFN, where the magnitude spectrum of the noisy speech mixture is taken as the input, for estimating the magnitude spectrum of the target speech. The proposed CFN model improves the model capacity, inter-channel dependency, parameter efficiency and feature re-use. With the proposed group convolutional fusion units, the standard convolution and depth-wise separable convolution were used to reinforce the model capacity of CFN. Then, the novel decoder allowed the CFN to take advantage of two

different convolutions. The experimental results confirmed that the group convolutional model had better model capacity than standard convolution. The group channel shuffle structure halved the number of output channels, thereby increasing parameter efficiency and exploiting inter-channel dependent information. In addition, utilizing skip connections inside the encoder and decoder can promote feature re-use and improve the performance. The experimental results showed the CFN offered more than 2.1% STOI, and 0.15 PESQ improvement over the baseline method.

Although this thesis offered feasible solutions to the monaural speech enhancement problems, further improvements could be obtained by introducing further advanced strategies in following aspects. More specifically, the proposed methods mainly focused on time-frequency domain speech enhancement. The phase information of noisy speech mixture was used to reconstruct target speech. Moreover, the proposed methods mainly used DNN, LSTM and CNN structures, more advanced network structures can be introduced. Furthermore, the neural network only utilized the audio-based feature, as a result, the video information was underestimated. In addition, the number of parameters of proposed methods can be further reduced, which enables these methods to be further applied in low-computational and resource devices. As a result, the influence of noise will be reduced in speech communication, audio recording for mobile phone an laptops.

6.2 Suggestions for Future Work

To further improve this study, some potential contribution points could be further researched.

Firstly, the phase information can be used to improve enhancement performance. The neural network can take the phase information of noisy mixture as input, and estimate the phase of target speech. Although such ap-

proaches have been investigated, these techniques can be combined with the advanced network structure to obtain accurate estimation.

Secondly, unsupervised learning can be further introduced. Unlike the conventional DNN based methods that learn mapping or masking (i.e. labels) relation between input and output, unsupervised learning without of class label is proposed. The advanced network can be combined with unsupervised learning to improve the generalization ability of the method.

Thirdly, video information can be employed in speech enhancement. The mouth movement and facial expression could provide the length, energy level even the contextual information of the target speaker. This information can generate fusion feature with audio data, and the advanced network structure can learn mapping or masking relation from these fusion features.

Finally, the parameter efficiency structure can be exploited in the enhancement framework, which offers tremendous potential for low-power devices such as mobile phones and laptops. Also, these structures should be easy to utilize in other network structures.

References

[1] C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, 1953.

[2] D. Wang, "Deep learning reinvents the hearing aid," *IEEE Spectrum*, vol. 54, no. 3, pp. 32–37, 2017.

[3] Lance Eliot, "Self-Driving Cars And Asimov's Three Laws About Robots." `https://www.forbes.com/sites/lanceeliot/2021/01/05/self-driving-cars-and-asimovs-three-laws-about-robots/?sh=3d35ccc54768`. online: accessed April 2021.

[4] D. L. Wang, "Deep learning reinvents the hearing aid," *IEEE Spectrum*, vol. March Issue, pp. 32–37, 2017.

[5] P. C. Loizou, *Speech Enhancement : Theory and Practice*. CRC Press, 2013.

[6] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 895–910, 2010.

[7] Y. Sun, Y. Xian, W. W. Wang, and S. M. Naqvi, "Monaural source separation in complex domain with long short-term memory neural network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 359–369, 2019.

[8] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 125–134, 2014.

[9] D. L. 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.

[10] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[11] W. Wang and a. J. A. C. S. Sanei, "Penalty function-based joint diagonalization approach for convolutive blind separation of nonstationary sources," *IEEE Transactions on Signal Processing*, vol. 53, no. 5, pp. 1654–1669, 2005.

[12] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proc. of IEEE*, vol. 86, no. 10, pp. 2009—-2025, 1998.

[13] Y. Liang, G. Chen, S. M. Naqvi, and J. A. Chambers, "Independent vector analysis with multivariate Student's t-distribution source prior for speech separation," *Electronics Letters*, vol. 49, no. 16, pp. 1035–1036, 2013.

[14] Y. F. Liang, J. Harris, S. M. Naqvi, G. J. Chen, and J. A. Chambers, "Independent vector analysis with a generalized multivariate gaussian source prior for frequency domain blind source separation," *Signal Processing*, vol. 105, pp. 175–184, 2014.

[15] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[16] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, no. 4, pp. 297–336, 1994.

[17] D. L. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley, 2006.

[18] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[19] M. I. Mandel, R. J. Weiss, and D. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[20] M. S. Salman, S. M. Naqvi, A. Rehman, W. W. Wang, and J. A. Chambers, "Video-aided model-based source separation in real reverberant rooms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1900–1912, 2013.

[21] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Network*, vol. 61, pp. 85–117, 2015.

[22] X. L. Zhang and D. L. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 967 – 977, 2016.

[23] Y. X. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[24] K. Han, Y. X. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 189–198, 2015.

[25] 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.

[26] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849—-1858, 2014.

[27] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002.

[28] Y. Jiang, D. L. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2112–2121, 2014.

[29] D. L. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplication*, vol. 12, no. 4, pp. 332–351, 2008.

[30] K. Tan and D. L. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 189–198, 2019.

[31] M. Tu and X. X. Zhang, "Speech enhancement based on deep neural networks with skip connections," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[32] A. Narayanan and D. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 92–101, 2015.

[33] P. Comon, R. Mukai, S. Araki, and S. Makino, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[34] W. Taylor, M. L. Seltzer, and A. Acero, "Maximum a posteriori ica: Applying prior knowledge to the separation of acoustic sources," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

[35] A. Hyvarinen and E. Oja, *Independent Component Analysis*. Wiley, 2001.

[36] S. Tu and H. Chen, "Blind source separation of underwater acoustic signal by use of negentropy-based fast ica algorithm," in *IEEE International Conference on Computational Intelligence and Communication Technology*, 2015.

[37] J. Harris, B. Rivet, S. M. Naqvi, J. A. Chambers, and C.Jutten, "Real-time independent vector analysis with student's t source prior for convolutive speech mixtures," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[38] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007.

[39] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Transactions on Circuits and System*, vol. 57, no. 7, pp. 1431–1438, 2010.

[40] J. Hao, I. Lee, and T. Sejnowski, "Independent vector analysis for source separation using a mixture of gaussians prior," *Neural computation*, vol. 22, no. 6, pp. 1646–1673, 2010.

[41] A. S. Bregman, *Auditory scene analysis*. MIT Press, 1990.

[42] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[43] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[44] J. Traa, P. Smaragdis, N. D. Stein, and D. Wingate, "Directional nmf for joint source localization and separation," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2015.

[45] Y. Sun, L. Zhu, J. A. Chambers, and S. M. Naqvi, "Monaural source separation based on adaptive discriminative criterion in neural networks," *IEEE International Conference on Digital Signal Processing (DSP)*, 2017.

[46] E. M. Grais, G. Roma, A. Simpson, and M. D. Plumbley, "Two-stage single-channel audio source separation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 9, pp. 1773–1783, 2017.

[47] Z. Z. Jin and D. Wang, "Learning to maximize signal-to-noise ratio for reverberant speech segregation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

[48] M. Henaff, K. Jarrett, K. Kavukcuoglu, and Y. Lecun, "Unsupervised learning of sparse features for scalable audio classification," *International Society of Music Information Retrieval (ISMIR)*, pp. 441–446, 2011.

[49] P. H. O. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," *International Conference on Machine Learning (ICML)*, 2014.

[50] Z. Z. Jin and D. L. Wang, "A supervised learning approach to monau-

ral segregation of reverberant speech," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[51] Z. Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberation speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 625–638, 2009.

[52] P.-S. Huang, M. Kim, M.-H. Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[53] F. Weninger, J.-L. Durrieu, F. Eyben, G. Richard, and B. Schuller, "Combining monaural source separation with long short-term memory for increased robustness in vocalist gender recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[54] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705—-4714, 2017.

[55] C. Szegedy, W. Liu, Y. Q. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2015.

[56] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *Association for the Advancement of Artificial Intelligence (AAAI) conference*, 2016.

[57] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1993–1997, 2017.

[58] E. M. Grais, D. Ward, and M. D. Plumbley, "Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders," *European Signal Processing Conference (EUSIPCO)*, 2018.

[59] Z.-C. Fan, Y.-L. Lai, and J.-S. R. Jang, "SVSGAN: Singing voice separation via generative adversarial network," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[60] M. Zhang and Z. Zhou, "A review on multi-label learning algorithms," *IEEE/ACM Transactions on knowledge and data engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

[61] 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.

[62] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time frequency weighted noisy speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125—-2136, 2011.

[63] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transanctions on Audio Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[64] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM.* Nat. Inst. Standards Technology, 1993.

[65] IEEE Audio and Electroacoustics Group, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio Electroacoust*, vol. 17, no. 3, pp. 225—-246, 1969.

[66] C. Veaux, J. Yamagishi, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[67] A. Varga and H. Steeneken, "Assessment for automatic speech recognition NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247—-251, 1993.

[68] G. N. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation.," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 2067–2079, 2010.

[69] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database (demand): A database of multichannel environmental noise recordings," *Journal of the Acoustical Society of America*, vol. 135, no. 5, p. 3591, 2013.

[70] Y. Xian, Y. Sun, J. A. Chambers, and S. M. Naqvi, "Geometric information based monaural speech separation using deep neural network," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[71] Y. Xian, Y. Sun, W. Wang, and S. M. Naqvi, "Monaural speech enhancement based on two stage long short-term memory networks," *IEEE International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2019.

[72] D. S. Williamson, Y. Wang, and D. L. 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.

[73] P. Feng, W. Wang, S. Dlay, S. M. Naqvi, and J. A. Chambers, "Social force model-based MCMC-OCSVM particle phd filter for multiple human

tracking," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 725–739, 2017.

[74] A. Rhemen, S. M. Naqvi, L. Mihaylova, and J. A. Chambers, "Multi-target tracking and occlusion handling with learned variational Bayesian clusters and a social force model," *IEEE Transactions on Signal Processing*, vol. 64, no. 5, pp. 1320–1335, 2016.

[75] Y. Sun, W. W. Wang, J. A. Chambers, and S. M. Naqvi, "Two-stage monaural source separation in reverberant room environments using deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, no. 1, pp. 125–139, 2019.

[76] J. B. Allen and D. A. Berkley, "Image method for efficiently simulation small-room acoustics," *Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[77] B. Shinn-Cunningham, N. Kopco, and T. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *Journal of the Acoustical Society of America*, vol. 117, pp. 3100–3115, 2005.

[78] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1983.

[79] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[80] K. Tan and D. L. Wang, "A convolutional recurrent neural network for real-time speech enhancement," *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018.

[81] P. Santiago, B. Antonio, and S. Joan, "Segan: Speech enhancement generative adversarial network," *Annual Conference of the International Speech Communication Association (Interspeech)*, 2017.

[82] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2015.

[83] J. Y. Chung, C. Gulcehre, K. H. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *Conference on Neural Information Processing Systems (NIPS)*, 2014.

[84] Y. Xian, Y. Sun, W. Wang, and S. M. Naqvi, "A multi-scale feature recalibration network for end-to-end single channel speech enhancement," *IEEE Journal of Selected Topics in Signal Processing*, 2021.

[85] Y. Xian, Y. Sun, W. Wang, and S. M. Naqvi, "Multi-scale residual convolutional encoder decoder with bidirectional long short-term memory for single-channel speech enhancement," *IEEE European Signal Processing Conference (EUSIPCO)*, 2020.

[86] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *International Conference on Machine Learning (ICML)*, 2010.

[87] A. L. Mass, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," *International Conference on Machine Learning (ICML)*, 2013.

[88] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR))*, 2016.

[89] C. Szegedy, V. Vanhoucke, S. Ioffe, and Z. W. J. Shlens, "Rethinking the inception architecture for computer vision," *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[90] K. Y. Cho, B. V. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[91] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," *International Conference on Machine Learning (ICML)*, 2014.

[92] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (gru) neural networks," *Midwest Symposium on Circuits and Systems (MWSCAS)*, 2017.

[93] D. P. Kinqma and J. L. Ba, "Adam: A method for stochastic optimization," *International Conference for Learning Representations (ICLR)*, 2015.

[94] S. K. Wajid, A. Hussain, and K. Z. Huang, "Three-dimensional local energy-based shape histogram (3d-lesh): A novel feature extraction technique," *Expert Systems With Applications*, vol. 112, no. 1, pp. 388–400, 2018.

[95] A. Adeel, M. Gogate, A. Hussain, and W. M. Whitmer, "Lip-reading driven deep learning approach for speech enhancement," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–10, 2018.

[96] A. Pandey and D. L. Wang, "A new framework for cnn-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[97] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, no. 2, pp. 1097–1105, 2012.

[98] X. Y. Zhang, X. Y. Zhou, M. X. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[99] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[100] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overftting," *Journal of Machine Learning Research*, no. 15, pp. 1929–1958, 2014.