Deep Neural Networks for Monaural Source Separation

by

Yang Sun

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

July 2019

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

In monaural source separation (MSS) only one recording is available and the spatial information, generally, cannot be extracted. It is also an undetermined inverse problem. Rcently, the development of the deep neural network (DNN) provides the framework to address this problem. How to select the types of neural network models and training targets is the research question. Moreover, in real room environments, the reverberations from floor, walls, ceiling and furnitures in a room are challenging, which distort the received mixture and degrade the separation performance. In many real-world applications, due to the size of hardware, the number of microphones cannot always be multiple. Hence, deep learning based MSS is the focus of this thesis.

The first contribution is on improving the separation performance by enhancing the generalization ability of the deep learning-base MSS methods. According to no free lunch (NFL) theorem, it is impossible to find the neural network model which can estimate the training target perfectly in all cases. From the acquired speech mixture, the information of clean speech signal could be over- or underestimated. Besides, the discriminative criterion objective function can be used to address ambiguous information problem in the training stage of deep learning. Based on this, the adaptive discriminative criterion is proposed and better separation performance is obtained. In addition to this, another alternative method is using the sequentially trained neural network models within different training targets to further estimate the clean speech signal. By using different training targets, the generalization ability of the neural network models is improved, and thereby better separation performance.

The second contribution is addressing MSS problem in reverberant room environments. To achieve this goal, a novel time-frequency (T-F) mask, e.g. dereverberation mask (DM) is proposed to estimate the relationship between the reverberant noisy speech mixture and the dereverberated mixture. Then, a separation mask is exploited to extract the desired clean speech signal from the noisy speech mixture. The DM can be integrated with ideal ratio mask (IRM) to generate ideal enhanced mask (IEM) to address both dereverberation and separation problems. Based on the DM and the IEM, a two-stage approach is proposed with different system structures.

In the final contribution, both phase information of clean speech signal and long short-term memory (LSTM) recurrent neural network (RNN) are introduced. A novel complex signal approximation (SA)-based method is proposed with the complex domain of signals. By utilizing the LSTM RNN as the neural network model, the temporal information is better used, and the desired speech signal can be estimated more accurately. Besides, the phase information of clean speech signal is applied to mitigate the negative influence from noisy phase information.

The proposed MSS algorithms are evaluated with various challenging datasets such as the TIMIT, IEEE corpora and NOISEX database. The algorithms are assessed with state-of-the-art techniques and performance measures to confirm that the proposed MSS algorithms provide novel solutions.

Contents

Statement of Originality

The contributions of this thesis are mainly on the improvement of the monaural source separation with deep neural networks (DNN). The novelty of the contributions is supported by two *IEEE* journal publications and eight other publications in the leading conferences in signal processing.

In the first contribution, the adaptive discriminative criterion is proposed in the cost function to address the ambiguous information problem in the training stage. Besides that, another method employs the sequentially trained DNN system, which firstly use two sequentially trained DNNs with different training targets to solve the speech enhancement problem. The sequentially trained DNN system can better predict the desired speech signal by addressing over- or underestimated information from the speech mixture. Pubilications related to this contribution have been accepted and submitted to:

- **Y. Sun**, L. Zhu, J. A. Chambers and S. M. Naqvi,'Monaural source separation based on adaptive discriminative criterion in neural networks', in *Proc. of IEEE International Conference on Digital Signal Processing (DSP)*, 2017, London, UK.

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

In the second contribution, by considering the difference between clean speech signal and the direct sound, the dereverberation mask (DM) is proposed to address the dereverberation problem in source separation in reverberant environment. The DM is utilized in two ways, the first one is using DM to integrate with ideal ration mask (IRM) to generate the ideal enhanced mask (IEM) and trained in one DNN. Alternatively, the DM and IRM are estimated separately to achieve dereverberation and separation tasks, respectively. The results of this solution are included in:

- **Y. Sun**, W. Wang, J. A. Chambers and S. M. Naqvi, 'Enhanced time-frequency masking by using neural networks for monaural source separation in reverberant room environments', in *Proc. European Signal Processing Conference* (*EUSIPCO*), 2018, Roma, Italy.
- **Y. Sun**, W. Wang, J. A. Chambers and S. M. Naqvi, 'Two-stage monaural source separation in reverberant room environments using deep neural network', *ACM/IEEE Transactions on Audio*, *Speech and Language Processing*, vol. 27, no. 1, pp. 125-139, 2019.

In the last contribution, a novel complex signal approximation (cSA) based method is proposed to use the phase information of the clean speech signal from the speech mixture, which prevents the influence from noisy phase information and improves the separation performance. Then, the long short-term memory (LSTM) recurrent neural network (RNN) is selected as the framework of the neural network model. By introducing LSTM RNN, the temporal information is better utilized and the accuracy of the estimation is refined. The results of this scheme are presented in:

- **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

Other related publications which are not included in this thesis as contribution chapters are:

- **Y. Sun**, W. Rafique, J. A. Chambers and S. M. Naqvi, 'Underdetermined source separation using time-frequency masks and an adaptive combined Gaussian-Student's t probabilistic model', in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, New Orleans, USA.

- **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', in *Proc. of IEEE Sensor Signal Processing for Defence (SSPD) Conference*, 2017, London, UK.

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

- Y. Li, **Y. Sun** and S. M. Naqvi, 'Sequentially trained DNNs based monaural source separation in real room environments', in *Proc. of IEEE Sensor Signal Processing for Defence (SSPD) Conference*, 2019, Brighton, UK.

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

- Y. Li, **Y. Sun** and S. M. Naqvi, 'Monaural source separation based on sequentially trained LSTMs in real room environments', in *Proc. European Signal Processing Conference (EUSIPCO)*, 2019, A Coruna, Spain.

Acknowledgements

I am deeply and sincerely indebted to my supervisor Dr. Syed Mohsen Naqvi for his consistent instruction, constant support and generous advice throughout the past four years. I have benefited tremendously from his rare insight, his exceptional knowledge and his great enthusiasm with patient mentoring to students. It is difficult for me to accomplish this thesis without his tireless instruction and encouragement. It is my great privilege and exclusive honour to have been one of his research students. I wish that I will have more opportunities to work with him in the future.

I am also extremely thankful to Prof. Jonathon Chambers, Prof. Satnam Dlay and Dr. Wenwu Wang, who helped me to enter my research field and kept giving me advice and encouraging me during my PhD study period; due to their kindly help, I have learnt much from their instruction.

I would also like to give my appreciation to my friends and colleagues, for their great help and providing a stable and cooperative environment during the past four years in the United Kingdom.

Lastly, but most importantly, I wish to express my deepest gratitude and love to my families, in particularly my loving parents. I really can not find appropriate words to represent my heartfelt thanks to them for their constant encouragement, prayers, attention and support in innumerable ways throughout my PhD and before. I would like to dedicate this thesis to my entire loving family.

Yang Sun

May, 2019

List of Acronyms

AMS	Amplitude Modulation Spectrogram
ARMA	Auto-Regressive Moving Average
ASR	Automatic Speech Recognition
A-V	Audio-Video
CASA	Computational Auditory Scene Analysis
cIRM	complex Ideal Ration Mask
CNN	Convolutional Neural Network
CPP	Cocktail Party Problem
cSA	complex Signal Approximation
DM	Dereverberation Mask
DNN	Deep Neural Network
DRNN	Deep Recurrent Neural Network
DRR	Direct to Reverberant Ratio
EM	Expectation Maximization
GAN	Generative Adversarial Network
GMM	Gaussian Mixture Model

IBM	Ideal Binary Mask
ICA	Independent Component Analysis
IEM	Ideal Enhanced Mask
ILD	interaural level difference
IPD	interaural phase difference
IRM	Ideal Ration Mask
IVA	Independent Vector Analysis
L-BFGS	Limited-memory Broyden-Fletcher-Goldfarb-Shanno
LC	Local Criterion
LSTM	Long Short-Term Memory
MESSL	Model-based Expectation-maximization Source Separation and Localization
MFCC	Mel-Frequency Cepstral Coefficients
MSE	Mean-Square Error
NFL	No Free Lunch
NMF	Non-negative Matrix Factorization
oSA	original Signal Approximation
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
Rt	Reverberant time
SA	Signal Approximation
SDR	Signal-to-Distortion Ratio
SNR	Signal-to-Noise Ratio
SSN	Speech-Shaped Noise
STFT	Short-Time Fourier Transform
STOI	Short-Time Objective Intelligibility
T-F	Time-Frequency
VB	Variational Bayesian
W-DO	W-Disjoint Orthogonality

List of Symbols

$\|.\|$	Absolute value
$\|\|.\|\|_2^2$	L2 norm operation
$\|\|.\|\|$	Norm operation
$\|\|.\|\|_{max}$	Max norm operation
$\|\|.\|\|_P$	P norm operation
det	Determinant operator
$trace(.)$	Trace operator
$kurt$	Kurtosis
$\mathcal{KL}$	Kullback-Leibler divergence
max	Maximum value
S	Spectrogram of the desired speech signal
M	Time-frequency mask
H	Spectrogram of the room impulse response
Y	Spectrogram of the speech mixture
D	Spectrogram of the direct sound
I	Spectrogram of the undesired interference

E	Expectation
$\mathbf{h}_t^l$	Hidden activation at layer l and time t
$\mathbf{y}_t$	Output of DRNN at time t
$\mathbf{x}_t$	Input of DRNN at time t
$g(\cdot)_l$	Activation function at the l-th layer
$\mathbf{R}^l$	Recurrent weight matrix at the l-th layer
$\mathbf{W}^l$	Current connection at the l-th layer
γ	Penalty parameter
$*$	Convolution operator
$\times$	Times operator
β	Tunable parameter to scale the mask
r	Real components
c	Imaginary components
ν	Standard random Gaussian variable
G	Nonquadratic function
$const.$	Constant

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Monaural Source Separation

"One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others. This is such a common experience that we may take it for granted; we may call it the cocktail party problem. No machine has been constructed to do just this, to filter out one conversation from a number jumbled together" - Colin Cherry [1]. The cocktail party problem (CPP) describes the situation that there are several people talking simultaneously in a room environment, and only one of them is desired. During the past decades, a significant research was focused on the CPP and the target is design a machine which can imitate human auditory ability. This problem is called source separation and it has attracted a remarkable amount of attention due to its potential use in many real-world applications such as automatic speech recognition (ASR), assisted living systems and hearing aids [2]. Such technology helps machine to imitate the hearing system of human being, which is useful in human-machine interaction and speech applications. Some application contexts of source separation are represented in Fig. 1.1.

All these example applications require to separate the desired speech signals from mixture, which implies correctly denosing, dereverberation and enhancement [3]. After the desired speech signals are separated from the acquired speech mixtures, the corresponding speech signals will be easy to

(a) (b) (c)

Figure 1.1. Different source separation application contexts, where (a) is teleconference system; (b) is the voice assistant by Amazon and (c) is a hearing aids equipment.

detect, recognize and understand. Hence, source separation is an essential front-end technique in the most of speech processing applications [4]. According to the number of channels (microphones), the source separation problems are divided into three categories: multichannel, binaural (two microphones) and monaural (one microphone) [5].

In multichannel case, when the situation is overdetermined or determined (number of captured speech mixtures is greater than or equal to the number of speech signals), the statistical signal processing based methods can be used to separate the desired speech signals from the acquired noisy speech mixtures [6]. As the humanity, we can use the high-level knowledge to distinguish the speakers, count the source number and localize the position of the speakers from the mixtures. These high-level knowledge, including, but not limited to, types of languages, the context of sentences and the tones of speakers. However, it is very difficult for machine to achieve the same abilities. In statistical signal processing based methods, the most popular algorithms are independent component analysis (ICA) and independent vector analysis (IVA) [7, 8], which rely on central limit theorem and frequency dependency. However, in binaural case, when the situation is undetermined (number of captured speech mixtures is less than the number of speech signals), the statistical based methods cannot provide simple solutions. Hence, according to the W-disjoint orthogonality (W-DO) [9] and spatial information i.e. interaural level difference (ILD), interaural phase difference (IPD), the time-frequency (T-F) based methods are proposed under the framework of computational auditory scene analysis (CASA) [10, 11].

However, due to the limitation of the hardware size and the minimum distance between microphones (less than 2 mm), the binary and multichannel cases cannot always be achieved, the solution to address source separation in monaural case is essential. For monaural source separation, also known as single-channel source separation, it is an undetermined case and because only one speech mixture is captured, the spatial information cannot be utilized. In [12], some algorithms are proposed to address MSS, but their performance is limited. Moreover, for statistical signal processing methods, the increase of the data cannot refine the separation performance significantly, which limits the improvement of the algorithms. Recently, the deep neural network (DNN) is introduced to solve the MSS problem, because the DNN can be trained to fit the non-linear relationship between the input and training target. The DNN-based MSS is a supervised problem, therefore, the training target of the network is very important. According to the training targets, the methods in DNN-based MSS are categorized as: masking-, mapping- and signal approximation (SA)-based [13, 14]. The T-F mask and the spectrogram of the clean speech signal are the most commonly-used training targets. By using the DNN techniques, the undetermined inverse problem in the MSS is addressed and the corresponding separation performance improved. The case of MSS is represented in Fig. 1.2. Besides, as the development of the deep learning, different neural network models are employed to address the MSS problem.

Although the DNN has mainly been employed in MSS, there are several

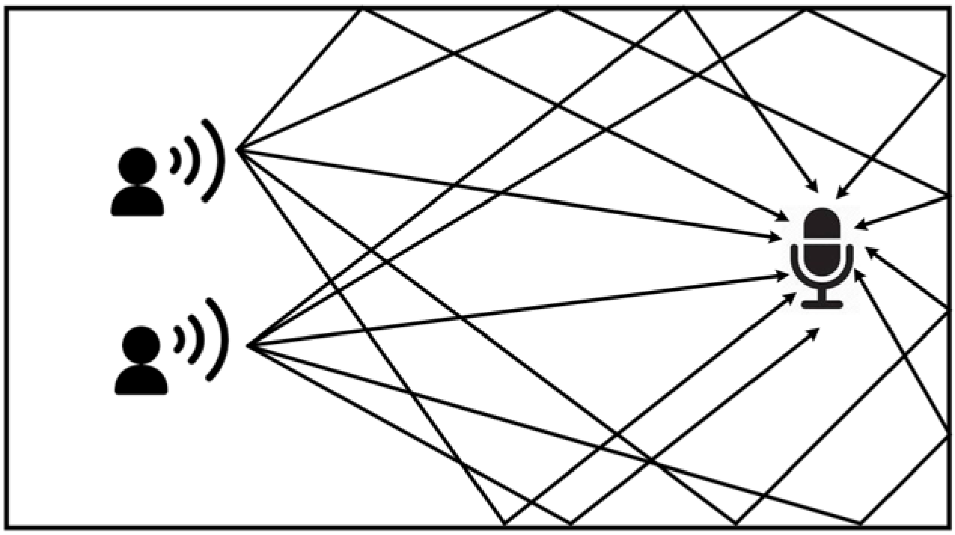

Figure 1.2. Monaural source separation problem with two speakers in the reverberant room environment.

weaknesses within the DNN framework expolited which can be improved. First of all, in the traditional DNN-based method, due to the limitation of the dataset, e.g. the amount of the training data and the unseen data samples in the testing stage, some information of the desired speech signal is over- or underestimated by the trained neural network model and the relationship between the input and the training target is not fully fitted. Secondly, when the MSS is required in the real room environment, the traditional methods can only separate direct path sound of desired speech signal from the mixture instead of desired speech signal. Thirdly, in conventional DNN-based methods, the noisy phase information is used to recover the desired speech singal from the mixture, which impacts the performance negatively [2]. Therefore, the phase information of clean speech signal and temporal information need to be better utilized.

1.2 Aims and Objectives

The overall aims of this thesis are to overcome the aforementioned weaknesses of the DNN-based MSS method and improve its separation performance. The particular objectives are:

- Objective 1: To improve the separation performance by exploiting the information of estimations related to the clean speech signal and refine the generalization ability.

In Chapter 3, in the first solution, the deep recurrent neural network (DRNN) is utilized as the framework and the adaptive calculated discriminative term is introduced in the cost function to solve the ambiguous information problem. Besides, in the second method, the T-F mask and spectrogram of the clean speech signal are used as training targets to train the DNNs sequentially. After the first DNN is trained, the second DNN is used to correct the over- or underestimated information in the estimation of the first DNN.

- Objective 2: Separate clean speech signals from the mixtures in different real room environments and improve the separation performance.

In Chapter 4, a novel dreverberation mask (DM) is proposed, by using the DM, the speech mixture in real room environment is dereverberated. Then, different from the conventional methods, the desired clean speech signal can be separated.

- Objective 3: Contribute to separation performance by utilizing the phase information of clean speech signal and temporal information.

In Chapter 5, the long short-term memory (LSTM) recurrent neural network (RNN) is selected as neural network model. And the phase information of clean speech signal is utilized by the proposed complex signal approximation (cSA)-based method.

1.3 Thesis Outline

The outline of this thesis is listed as follows:

Chapter 2 includes a relevant literature review of MSS by using deep learning, where different methods for source separation are given. Moreover, different challenges associated with MSS work are described, including reverberant environment and noisy phase information. Different MSS algorithms are discussed to address these challenges. Then, MSS methods are categorized according to the training targets and neural network model, where their advantages and disadvantages are given.

Chapter 3 satisfies the first objective of this thesis. In the first method, the DRNN is used as the framework. And the adaptively calculated discriminative term is introduced in the cost function. Hence, the ambiguous information is better estimated. In the second method, the sequentially trained DNNs are firstly employed to replace the single DNN model in the MSS algorithm. In order to correct the over- or underestimated information in estimation of the single DNN model, the second DNN is introduced and both DNNs have different training targets. The second DNN helps to further estimate the separated speech signal. Hence, in both methods, the generalization ability of DNNs is enhanced and the performance is improved.

Chapter 4 proposes a novel dereverberation mask (DM) to achieve the dereverberation before separating the desired speech signal from the mixture in real room environments. Then, the ideal enhanced mask (IEM) is integrated from ideal ratio mask (IRM) and DM. Different two-stage methods are proposed with different system structures (single DNN or two DNNs). By using DM, the estimation of the clean speech signal is obtained, which is different from the conventional methods, where the direct sound of clean speech signal is estimated.

In Chapter 5, the separation performance is improved by utilizing the phase information of the clean speech signal. And the LSTM RNN is applied as framework, the corresponding neural network models can better fit the non-linear relationship and the estimation from the trained model will be more accurate. To utilize the phase information, the cSA-based method is proposed. Evaluation comparisons confirm the improvement of the proposed MSS method over other state-of-the-art techniques.

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

Chapter 2

RELEVANT LITERATURE REVIEW

2.1 Chapter Introduction

In this chapter, the literature review related to source separations and the commonly-used methods in MSS are discussed. These methods are based on statistical signal processing, computational auditory scene analysis (CASA) and the neural networks. Then, within the monaural case, the limitations of these methods are given and the corresponsing solutions are discussed. Finally, some performance mesures are described, which are appiled to evaluate the performance of the MSS algorithms.

2.2 Statistical Signal Processing

Recently, source separation has become very popular due to the development of the speech processed-based applications and human computer interaction [2]. As mentioned in Chapter 1, when the situations are overdetermined and determined cases, the statistical signal processing based methods such as independent component analysis (ICA) and independent vector analysis (IVA) can be used to separate the desired speech signals.

2.2.1 Independent component analysis

ICA is a statistical model where the observed data is expressed as a linear combination of underlying latent variables [15]. In ICA technique, the time delay is neglected and the basic model is instantaneous. The independent components are assumed to be statistically independent of each other. And two variables are independent, if and only if the joint probability density function (PDF) is factorizable in the following way:

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

Moreover, at least one source must have non Gaussian distribution and the unknown mixing matrix is assumed to be invertible, which the number of sources is equal or less than the number of mixtures.

In [7], ICA is used to separate the speech signals from mixtures and the central limit theorem indicates that any mixture of components will become more Gaussian than the individual components. Hence, there are several kinds of methods in ICA and the most popular one is by maximization of non-Gaussianity. To measure the non-Gaussianity, the kurtosis and the negentropy are utilized in ICA. Besides, these two measures can be used to minimize the mutual information with the Kullback-Leibler Divergence [16].

According to [17], the kurtosis (*kurt*) is defined by the fourth-order cumulants of a random variable:

$$kurt(y) = \mathrm{E}\{y^4\} - 3(\mathrm{E}\{y^2\})^2 \tag{2.2.2}$$

where E is the expectation operation. But when there exist outliers, the value of kurtosis becomes large. Hence, the kurtosis is not a robust measure of the non-Gaussainity. A measure that is zero for Gaussian variables and always non-negative can be obtained from differential entropy, and called negentropy [7]. The classical method of approximating negentropy is based on the higher-order cumulants for zero mean random variable y is:

$$J(y) \approx \frac{1}{12}\mathrm{E}\{y^3\}^2 + \frac{1}{48}kurt(y)^2 \qquad (2.2.3)$$

The nonquadratic function G is defined as $G(x) = x^4$. Using nonquadratic function G, the negentropy can be calculated as:

$$J(y) \propto \mathrm{E}\{\mathrm{G}(y)\} - \mathrm{E}\{\mathrm{G}(\nu)\}^2 \qquad (2.2.4)$$

where ν is a standard random Gaussian variable [7]. When the nonquadratic function grows slowly, the obtained estimators are robust [18].

Although the ICA algorithm can be applied to address source separation problem by transfering the mixture into frequency domain, the permutation and scaling problem have negative influence on the separation performance [19]. The permutation ambiguity is caused by different separation order for each frequency bin, which is the main problem in ICA algorithm. Fig. 2.1 shows the ICA algorithm with scaling and permutation problems [20]. Compared with the clean speech signals and the separated signals, it can be observed that the amplitude and the order of the separated signals have changed. In ICA algorithm, the scaling problem can be addressed by using the rescaling algorithm, but it is difficult to solve the permutation problem in ICA algorithm.

2.2.2 Independent vector analysis

From Section 2.2.1, it can be known that the limitations of ICA algorithm will decrease the performance. Hence, the IVA method is introduced to eliminate the permutation problem by using a dependency model which captures inter-frequency dependencies in the desired speech signals. And the multivariate score function is applied to describe the source prior [8], which is the

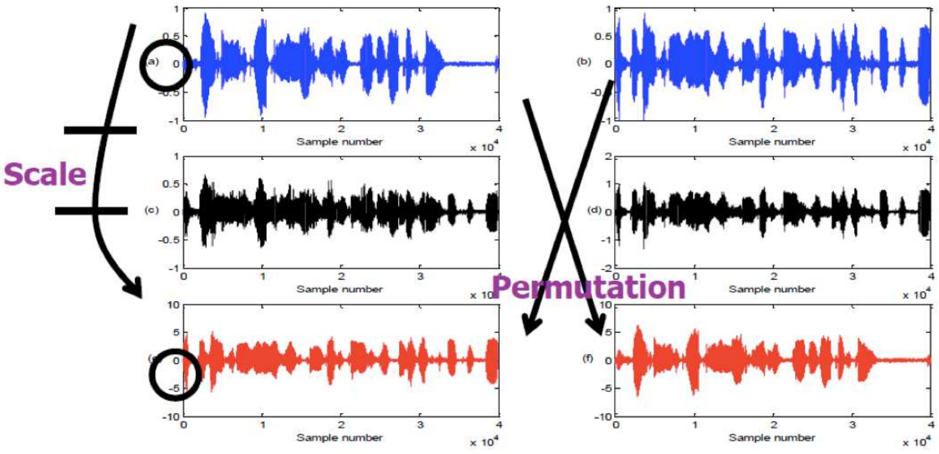

Figure 2.1. Permutation problem and scaling ambiguity with ICA algorithm. The blue signals are the original signals, which are used to generate mixtures, the black signals are mixtures and the red signals are separated signals.

higher order frequency dependency.

Since the multivariate sources need to be separated from the multivariate observations, a cost function is defined for the multivariate random variables. In IVA algorithm, the Kullback-Leibler divergence is defined to measure the independence [19]. One function is the joint pdf $p(\widehat{\mathbf{s}}_\mathbf{1} \ldots \widehat{\mathbf{s}}_\mathbf{L})$ and another is the product of approximated pdfs of individual source vectors, defined as $\prod_{i=1}^{L} q(\widehat{\mathbf{s}}_\mathbf{i})$.

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

$$= const. - \sum_{k=1}^{K} log|detG^{(k)}| - \sum_{i=1}^{L} \int E\{log(q(\widehat{\mathbf{s}}_\mathbf{i}))\} \qquad (2.2.5)$$

where the $\int p(\mathbf{x_1} \ldots \mathbf{x_M}) logp(\mathbf{x_1} \ldots \mathbf{x_M}) d\mathbf{x}_1 \ldots d\mathbf{x}_M$ is the entropy of the observation signals. The dependency between sources are removed but the inherent frequency dependency is preserved. Therefore, by using the IVA algorithm, in the separated speech signal, the permutation problem is eliminated.

Due to the nature of speech signals, the useful samples can be of high amplitude and the longer tails can model the spectrum of speech signals better than the Gaussian distribution. In Student's t-distribution, the tails of the distribution is controlled by the degree of freedom. If the value of degree of freedom is small, the distribution will have heavier tails. If the value of the degree of freedom is increased to infinity, the Student's t-distribution is same as the Gaussian distribution. The shape of the Student's t-distribution with different value of degree of freedom is shown in Fig. 2.2. Therefore, in some improved IVA methods, the Student's t-distribution is exploited to replace the Gaussian distribution to model the spectrum of speech signals perfectly [11, 21, 22].

Generally, both ICA and IVA algorithms are proposed to solve the source separation with overdetermined and determined cases. Moreover, in the audio-video (A-V) methods, the ICA and IVA algorithms are applied with additional video information [23, 24]. After the video information is introduced and combined with the audio information, the ICA and IVA algorithms are applicable to address the underdetermined case. Meanwhile, rather than collecting the video information, using pre-process algorithm with the speech mixture is also beneficial. For example, the variational Bayesian (VB) methods are investigated in source separation problem. In VB algorithm, the distributions of speech signals are estimated and the number of sources in the speech mixture is obtained by counting the number of distributions [25–28]. 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.

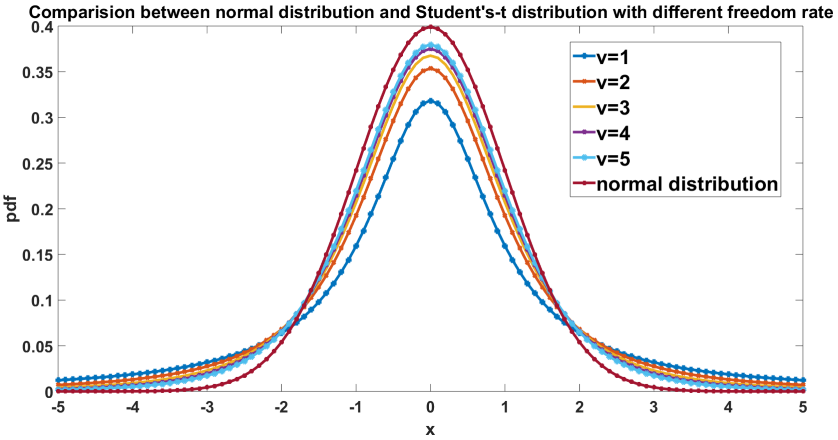

Figure 2.2. Comparison of univariate Gaussian distribution and univariate Student's t-distribution with different degree of freedom rates.

2.3 Computational Auditory Scene Analysis

According to Section 2.3, the statistical signal processing based methods are valid only for determined (number of sources is equal to the number of sensors) and overdetermined (number of sources is less than the number of sensors) cases. For undetermined case, the time-frequency (T-F) based methods are proposed under the framework of computational auditory scene analysis (CASA) e.g. the model-based expectation-maximization source separation and localization (MESSL) algorithm [29, 30].

In the MESSL with binaural cases, which is similar to human auditory system, two speech mixtures are captured by two microphones with different locations. Therefore, the binaural cues i.e. interaural level difference (ILD) and interaural phase difference (IPD) are modelled as the mixture of Gaussians, which can be utilized to infer the T-F mask [10]. The main assumption for the applicability of this method is that only one source is active at each T-F point [9]. In details, the MESSL algorithm exploits the IPD and ILD as clues to build the probabilistic model of sources to evaluate the hidden variable at every spectrogram point. Then, the masks are generated to separate the sources from the mixture according to the probability of activate points. The expectation maximization (EM) algorithm computes the expectation of hidden variable in E step, which is 1 for the active point in the spectrogram. In the M step, the value of E step is exploited to calculate the maximum-likelihood parameters. After several iterations, the EM algorithm arrives the convergence point and the the T-F mask will be applied to separate the original sources from the mixture.

In the MESSL algorithm, the parameters are assumed as the mixture of Gaussian distributions [31]. However, in reverberant room environments, the observed data may have heavy tails and therefore assumption of Gaussian mixture model is limited. The Student's t-distribution has been exploited to replace the Gaussian distribution to model the mixture and obtain more information from outliers to solve this limitation in MESSL [32]. Moreover, since the shape of the distribution of the mixture is not fixed, a combined probabilistic model is proposed with Student's t-distribution and Gaussian distribution. A weighting parameter is utilized to modify the contribution of each distribution in the combined model to ensure most of the information in the mixture is captured. In order to enhance the robustness in the separation performance, an adaptive process is introduced to determine the value of the weight according to the mixture energy and the mixture distribution is jointly modelled with the Gaussian distribution mixture model and Student's t-distribution model [11]. The tails of the distribution are better modelled by the Student's t-distribution model whereas the lower amplitude information by the Gaussian distribution mixture model. Meanwhile, in [10, 11, 32], the EM algorithm is applied to solve the clustering problem of the interaural cues. In addition, the variational inference is introduced to replace the EM algorithm to overcome the difficulties associated with the likelihood optimization in [33]. In binaural and multi-channel cases, all of these unsupervised learning methods achieve promising performance in source separation problem. However, in monaural case, the above mentioned methods cannot be used to separate the desired speech signal, due to the lacking of spatial information [34].

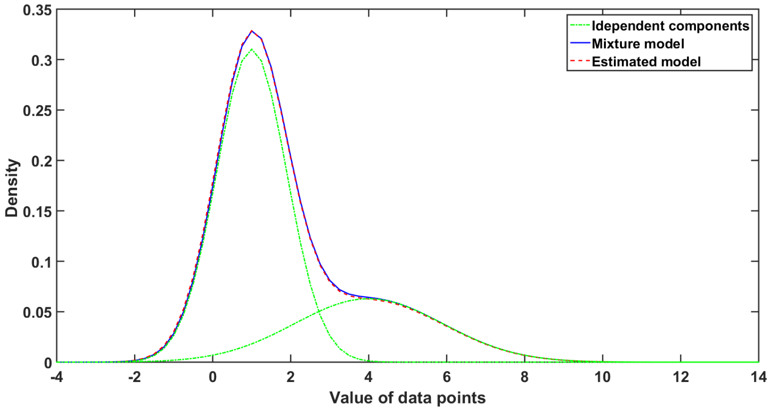

Figure 2.3. The distribution of the Gaussian mixture model (GMM), the red distribution is the estimated distribution of GMM, and the blue distribution is the true GMM and the greens are the idependent components of the GMM.

2.4 Monaural Source Separation

In monaural source separation (MSS), only one speech mixture is available, it leads the spatial information cannot be utilized in this case, and the desired speech signal needs to be separated from it. Many approaches have been developed to address the MSS problem. For example, in signal processing-based methods, in [35], an ideal Wiener filter is estimated and the target signal is reconstructed in the minimum mean squared error (MMSE) sense. While in model-based methods, the non-negative matrix factorization (NMF) [36] is exploited to separate signals from a single channel mixture [37], where non-negative matrix of magnitude or power spectrogram is decomposed. Then, in [12], the noisy observations is modelled based on weighted sums of non-negative sources to separate the desired speech signal from the noisy mixture. In [38], a variational Bayesian inference procedure is developed to learn variational parameters and model parameters with NMF, which is called Bayesian NMF algorithm. All the above mentioned methods are unsupervised learning algorithms, no labelled dataset is needed when drawing interferences from the input data [39, 40].

Recently, the developments of deep neural network (DNN) techniques attract the attention from the researcher in MSS. In DNN-based MSS algorithms, the desired speech signal is obtained from the trained neural network model [41], which is a supervised learning algorithm. In Section 2.6, the challenges of DNN-based MSS algorithm are discussed firstly, then the different relevant solutions for each of them are reviewed.

2.5 Deep Neural Network

Recently, the deep neural network (DNN) has been applied to address many practical speech signal processing problems, e.g. speech recognition, speech enhancement and source separation. A DNN model is constructed by three different types of layers: input layer, hidden layer and output layer, and these layers contain numerous neurons.

The DNN model uses weights and bias to fit the relationship between the training target and the output. By minimizing the loss function with the training target and the output, the DNN model is trained. Generally. the gradient decent algorithm is applied to change the weights and bias of the model to find the best fitted model.

Meanwhile, according to the structure of the DNN model, different from the vanilla DNN model, the recurrent neural network (RNN), convolutional neural network (CNN) et. al are proposed in order to achieve the best performance for specific tasks.

2.6 Challenges Associated with Monaural Source Separation

Although introducing deep learning techniques into MSS leads to the improvement in separation performance, it still has some limitations due to the various situations of environments and the high randomness of speech signals. For example, the energy distribution cannot be totally same even the speech signal is obtained from the same person with same length and context, which means there is no same speech signals exist. The separation performance in DNN-based MSS algorithm depends on many factors, and all these factors will have influence on the separation performance. Therefore, in Section 2.6, the challenges of the DNN-based MSS algorithms are described as follow:

- Training target: Choosing a proper training target of the neural network is very important, it helps the neural network model to be trained accurately and the estimation from the trained model is better used to reconstruct the desired speech signal.

- Neural network architecture: There are many different types of neural network models, each of them has its own advantages and disadvantages in solving MSS problem. Using the appropriate architecture and configurations of the model can lead to performance improvement.

- Generalization ability: Because the speech signal is highly random, it is impossible to find two speech signals which are totally same. In addition, the types of noise interferences in the real world are countless. Therefore, the trained neural network model needs to have a strong generalization ability to retain the performance with unseen speaker and noise interference cases [42].

- Room environment: In real room environment, due to the reflections from floor, walls, ceiling and furnitures, the acquired speech mixture is difficult to separate, the dereverberation is essential before separation, otherwise, the separation performance will be impacted [43].

2.7 Relevant Solutions

As mentioned in Section 2.6, the challenges exist in the MSS algorithm limit the separation performance. In the following subsections, the main challenges for DNN-based MSS work related to this thesis are studied, and their different relevant solutions are discussed.

2.7.1 Training target

Compared with the traditional methods, the DNN-based MSS algorithm is a supervised learning problem, it contains training and testing stages. Therefore, the target in the training stage is crucial.

Many developments have been proposed recently to address the challenges of training targets. The straightforward idea is using the neural network model to find the relationship between noisy speech mixture and clean speech signal [44]. For instance, the mapping-based method is firstly proposed, where the spectrogram of the clean speech signal is used as training target. In [45], the spectrogram of the desired speech signal is estimated directly from the noisy mixture. Then, some pre-and post-processing operations are added into mapping-based method to further improve the performance, e.g. noise estimation and global variance equalization [46].

Then, the T-F mask is given as the training target and the estimated desired speech signal is obtained by using the predicted T-F mask. In [47], the DNN is exploited to generate an ideal binary mask (IBM) to separate the speech mixture as described in (2.7.1).

$$IBM(t,f) = \begin{cases} 1, & if \ \mathrm{SNR}(t,f) > LC \\ 0, & otherwise \end{cases} \tag{2.7.1}$$

where local criteria (LC) is used to determine the value of the T-F unit in IBM. But the IBM is a binary mask, and the associated hard decision causes loss in the separation performance [47]. Then, the soft mask is proposed, also known as the ideal ratio mask (IRM), for which the T-F unit is assigned as the ratio of desired source energy to mixture energy [14] and the IRM-based method outperforms the IBM-based method. However, the above mentioned methods do not utilize the phase information of the desired signal when synthesizing the clean signal. In [48], the phase information is considered to be unimportant, but further research has shown that the phase information is beneficial to predict an accurate mask and the estimated source [49]. Consequently, in [50, 51], complex IRM (cIRM) is employed and both the magnitude and phase spectra are used to estimate the desired speech signal. An IRM plot is shown in Fig. 2.4, the spectrogram is obtained with a 1024-point short time Fourier transform (STFT) with 50 % overlap, zero padding and Hamming window is explored.

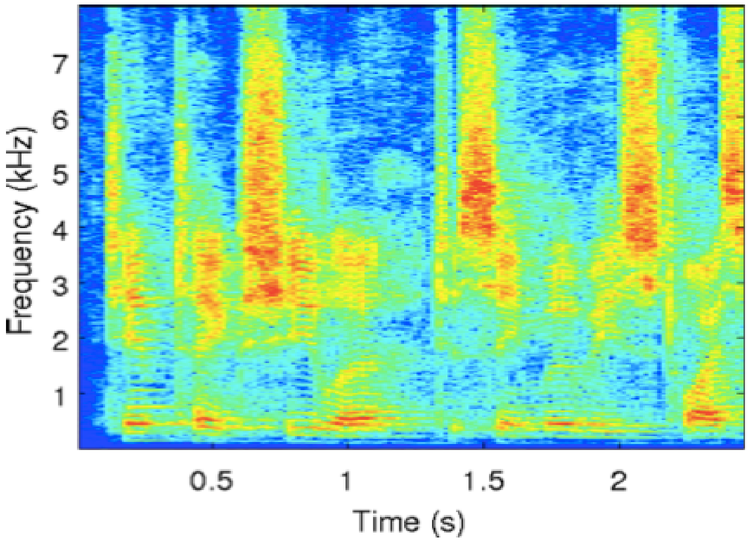

Figure 2.4. The plot of the ideal ratio mask.

Different from mapping- and masking-based methods, the signal approximation (SA)-based method using the spectrogram of the clean signal as training target but the estimation of the trained neural network model is a T-F mask [13], which becomes a popular method in recent research.

2.7.2 Architecture

Besides the challenge of training targets in neural network model, the architecture of neural network model is also a potential problem for MSS. Many researchers have explored various architectures in MSS-based methods to address this challenge.

Since DNN is used to solve the undetermined inverse problem in MSS, the vanilla DNN (DNN model without temporal connection) is firstly applied. The IBM is used as training target with vanilla DNN in [52], then the IRM is utilized with same neural network model. Moreover, dual outputs DNN is used in [53] to further improve the separation performance with vanilla DNN. However, in training stage, the temporal information is very important, which cannot be fully used in traditional vanilla DNN. The context window is proposed in order to utilize the temporal information, which combines number of frames as the input to estimate the single frame in the training target. Although the vanilla DNN has some limitations, it is still the most popular neural network structure due to its low computational cost and good performance [54].

Compared with vanilla DNN, recurrent neural network (RNN) has better ability to utilize the temporal information. In RNN each hidden state is determined by the current state and the previous state. For example, in [55], the RNN is introduced as the trained model to separate speech mixture with SA-based method. Then, the deep recurrent neural network (DRNN) is proposed, for which only the selected layers in the networks have the temporal connection [56]. The MSS method with DRNN is used in [13], which is trained with discriminative training criterion. Based on method in [13], the parameter in the discriminative term is calculated adaptively to penalize the objective function, which is proposed in [41]. Fig. 2.5 shows the basic structure of a recurrent unit, where x_t, s_t and o_t are the input, the hidden state and the output at time step t, respectively. U, W and V are the weights.

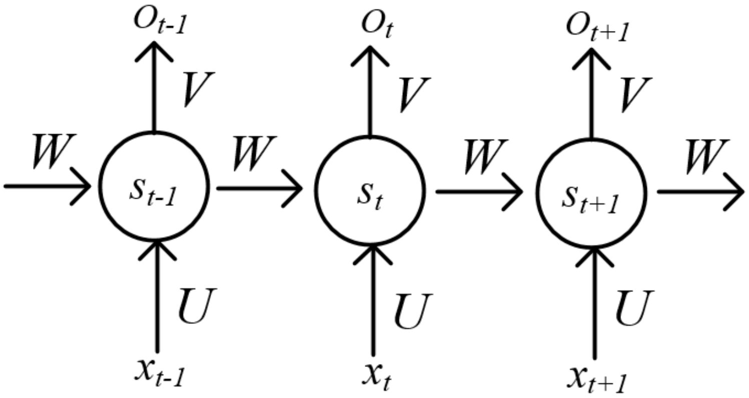

Figure 2.5. A recurrent neural network and the unfolding in time of the computation involved in its forward computation.

By using the long short-term memory (LSTM) block instead of the regular network units, the LSTM RNN is utilized in the monaural source separation in [57], and the evaluations confirmed the improvement of the separation performance. Then, the mapping- and masking-based LSTM RNN methods are compared with different SNR levels and background noise in [58]. More details about the LSTM RNN-base MSS methods are given in Chapter 5.

Moreover, some other deep learning techniques have been explored and attempted to solve the MSS problem. For example, generative adversarial network (GAN) is used to generate the T-F mask to separate desire speech signal in [59]. And the reinforcement learning is applied to address MSS by self-optimized [60, 61]. However, in these novel techniques, separation performance is not consistent, and the further research is needed.

2.7.3 Generalization ability

The speech signal is highly random, therefore, the MSS algorithm needs to have a robust performance when facing different speech mixtures. In term of the DNN-based MSS algorithm, the trained neural network model is required to have the ability to performance well on the unseen speech mixtures rather than the data that it has been trained with, it is called generalization ability. Moreover, the generalization ability is strongly related to the overfitting problem, if the trained model is overfitting, the generalization ability will be not good enough [62].

In [57], the LSTM RNN is selected as the neural network model and its generalization ability is improved by the unique property of the LSTM RNN. Then, some two-stage methods are proposed, which introduce a separation system. In the system, the ensemble learning is employed, some neural network models are trained, which give better generalization ability. For example, in [63], two vanilla DNNs are trained and the training data is divided into two parts in the training stage.

2.7.4 Performance measures

To evaluate the performance of the separated speech signal, some measures are proposed. In [64], the perceptual evaluation of speech quality (PESQ) is introduced to evaluate the quality of the speech signal. By comparing the clean speech signal with the separated speech signal from the noisy speech mixture, the PSEQ is measured. The value range of the PESQ score is given ranging from -0.5 to 4.5, the higher value of the score means better quality of the separated speech signal.

Then, the short-time objective intelligibility (STOI) is proposed in [65], which is a function of a T-F dependent intermediate intelligibility measure. The STOI compares the temporal envelopes of the clean speech signal and the separated speech in short-time regions by means of a correlation coefficient. The value range of the STOI is between 0 and 1, similar to PESQ, the higher value of STOI indicates better speech quality. The STOI can be calculated as:

$$STOI = \frac{1}{JM}\sum d_{j,m} \tag{2.7.2}$$

where $d_{j,m}$ is the sample correlation coefficient between estimated speech signal and desired speech signal, j is the jth frequency band, m is the time frame, J is the total number of frequency bands and M is the total number of time frames.

Based on signal to noise ratio, the frequency-weighted segmental SNR (SNR$_{fw}$) is proposed in [66], the averaging frame level SNR is estimated and a lower threshold is often set to provide a bound on frame based SNR. The SNR$_{fw}$ can be calculated as:

$$SNR_{fw} = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=0}^{J-1} W(j,m)log_{10}\frac{X(j,m)^2}{X(j,m)-\hat{X}(j,m)}}{\sum_{j=0}^{J-1} W(j,m)} \tag{2.7.3}$$

where $W(j,m)$ is the weight assigned to the jth frequency band and mth time frame, K is the number of bands, $X(j,m)$ is the critical band magnitude of the clean speech signal and the $\hat{X}(j,m)$ is the corresponding spectral magnitude of the processed signal [66].

Another measure is source to distortion ratio (SDR), which is proposed according to the SNR. In SDR, firstly, the separated speech signal is decomposed into four parts, the noise, interference, the artifacts error terms and the target speech signal. By comparing the target speech signal and these error terms, the SDR is calculated which is described as:

$$SDR = 10log_{10}\frac{desired\ speech\ signal + noise\ interference + artifacts}{noise\ interference + artifacts} \tag{2.7.4}$$

All these metrics are the most common-used performance measures in source separation.

2.8 Chapter Summary

In this chapter, a literature review of source separation algorithms was presented and discussed. Firstly, the main algorithms of source separation with over- and determined cases were described. Then, the unsupervised learning solutions in MSS were discussed. Secondly, the main challenges of the DNN-based MSS algorithms were given. In the third part of this chapter, the main existing solutions related to those challenges were viewed. From the different algorithms in the literature reviewed in this chapter, the requirements of the source separation algorithms can be summarised as to

- strong generalization ability to the unseen speech mixture and retain good separation performance;
- cope with the reverberant speech mixture, which is recoded in the real room environment;
- give solutions to use the phase information of clean speech signal in separating desired speech signal from the mixture.

Therefore, the focus of this thesis is to achieve these abilities and the above requirements.

In the next chapter, the DRNN with adaptively calculated discriminative term method and the sequentially trained DNN system will be proposed to improve the generalization ability and separation performance.

Chapter 3

SINGLE-CHANNEL SPEECH ENHANCEMENT AND SEPARATION

3.1 Chapter Introduction

Single-channel speech enhancement and separation have been studied for many years due to its importance in a number of real-world applications such as automatic speech recognition (ASR), assisted living systems and hearing aids [2,67–71]. The aim of enhancement and separation is increasing the intelligibility and the quality of the desired speech signal from a noisy speech mixture [46].

Masking and mapping are two commonly-used methods in speech enhancement and separation. In masking-based deep neural network (DNN) methods, the time-frequency (T-F) masks are used to enhance the quality of the desired speech signal from the noisy speech mixture [72]. By calculating the ratio of energy between clean speech signal and noisy speech mixture, the T-F mask is generated and utilized to obtain the desired speech signal [47, 50]. In masking-based method, there are two main categories according to the decision strategy, e.g. ideal binary mask (IBM) and ideal ratio mask (IRM) [73]. The IBM is a binary mask, and the associated hard decision introduces spectral artefacts [41]. On the contrary, the soft decision is applied in the IRM, the T-F unit is assigned as the ratio of desired source energy to the energy of noisy speech mixture [14]. In general, because the hard decision is employed, some noisy information is added, therefore, the IRM-based method is shown to outperform the IBM-based method. The mapping-based DNN is another promising method to address speech enhancement and separation problem, where the DNN is trained to generate the clean spectrum of the desired speech signal by using the spectrum of the noisy speech mixture [74]. Compared with the masking-based method, the mapping-based method offers competitive performance and does not need further operation to recover the spectrum of the desired speech signal. These two methods will be described in details in Section 3.2.

However, there are two limitations associated with the existing methods: (1) The DNN-based methods need better generalization ability when facing unseen interferences. (2) It is well known that the trained DNN may not perfectly reflect the relationship between input and training target, and some information of the clean speech signal may be underestimated or overestimated when it is recovered from the noisy speech mixture [75].

In this chapter, two different methods are proposed to solve the above mentioned limitations. In the first method, the discriminative term is added in the cost function to address ambigous information problem. Besides, the parameter in the discriminative term is calculated adaptively to penalize the cost function and the deep recurrent neural network (DRNN) is used as the framework. Hence, the influence from the ambiguous information will be eliminated and the temporal information is better used. The second method is adding another DNN to build a system, these two DNNs are trained with different training targets sequentially. Therefore, the overestimated and underestimated information in the estimation is corrected from the second DNN, which helps improve the generalization ability and performance.

The performance of the proposed adaptive penalty term cost function with DRNN and sequentially trained DNN system are evaluated with TIMIT and NOISEX datasets. These proposed methods are compared with the traditional masking- and mapping-based methods. The performance and comparison results are shown at the end of this chapter, which confirm the improved performance from both proposed methods.

3.2 Adaptive Discriminative Criterion with DRNN

In the MSS problem, which is solved via neural networks, the separation performance can be improved by utilizing the temporal information of the speech signals in the training stage of networks. Commonly, the temporal information is exploited in two ways: concatenating neighbouring features and using recurrent neural unit [76]. In the concatenating features method, a larger window size can utilize more temporal information with the trade off being computational and memory resources. Therefore, an appropriate window size is required. The RNNs have a recurrent architecture, which is a powerful model for temporal information. The deep recurrent neural network (DRNN) combines the multiple levels of representation that have proved so effective in DNNs with the flexible use of long range context that empowers RNNs [56]. Besides, in the training stage, when the features are similar, the neural network will be conservative. Because of the similarity, a feature can be attributed to more than one target in some cases. To maintain the efficiency of the training stage, the neural network will attribute the feature to both targets, it is called the conservative strategy [77]. However, if the ambiguous features are attributed repeatedly, the separation performance is decreased. In this section, the DRNN and adaptive discriminative criterion are proposed to solve temporal and ambiguous information problem.

3.2.1 Mapping-based DNN

In mapping-based DNN approach, the training target is the spectrum of the clean speech signal and the neural network model is trained to estimate the clean spectrum of desired speech signal.

The cost function of mapping-based DNN approach is expressed as:

$$Loss_{\text{mapping}} = \sum_{t}\sum_{f}(|\hat{S}(t,f)| - |S(t,f)|)^2 \tag{3.2.1}$$

where $|\hat{S}(t,f)|$ is the estimated spectrogram of the desired speech signal and $|S(t,f)|$ is the training target, which is the spectrogram of the clean speech signal.

In [45], the DNN model is trained to learn the relationship between the spectrum of the noisy mixture and the clean spectrum of the target signal to address the enhancement and separation problems. Different from the masking-based method, the prediction of the trained neural network model is the clean speech signal. While in masking-based method, the prediction is the T-F mask, it needs to be operated with the mixture to obtain the estimation of the clean speech signal. But the large value range of T-F points in the spectrum of the desired speech signal leads to a training problem, where the neural network model is difficult to be trained properly [78].

3.2.2 Masking-based DNN

In masking-based DNN approach, the ideal T-F mask is applied as the training target of the neural network models, which happens in training stage. Then, in the testing stage, the T-F mask is predicted from the trained model. The predicted T-F mask is applied to the mixture to reconstruct the desired speech signal and the predicted T-F mask can be categorized as a binary or soft mask.

In the binary mask, each T-F unit of the mask was assigned as 1 or 0 according to the criterion for the active source [52, 79]. However, due to the hard decisions from the IBM, the separated speech signal of the IBM-based method is distorted. In the soft mask, also known as ideal ratio mask (IRM), the T-F unit was assigned as the ratio of target source energy to mixture energy [74]. Compared with the IBM, the desired speech signal separated by IRM often has better quality, e.g. with less musical noise artefacts.

The cost function of masking-based DNN approach is expressed as:

$$Loss_{\text{masking}} = \sum_t \sum_f (|\hat{M}(t,f)| - |M(t,f)|)^2 \tag{3.2.2}$$

where $|\hat{M}(t,f)|$ is the estimated T-F mask and $|M(t,f)|$ is the training target, which is the ideal T-F mask. Because the T-F mask is a ratio mask, compared with the mapping-based, it can better extract the sparse information from the speech mixture.

3.2.3 Deep recurrent neural network

In order to better utilized the temporal information and save computational cost, the DRNN is introduced as the framework. According to [80], two DRNN architectures are defined: 1) an L hidden layer DRNN with temporal connection only at the l-th layer (DRNN-l) and 2) a full RNN. Assume $\mathbf{h}_t^l$ is the hidden activation at layer l and time t :

$$\mathbf{h}_t^l = f_{\text{h}}(\mathbf{x}_t, \mathbf{h}_{t-1}^l)$$

$$= g_l(\mathbf{R}^l \mathbf{h}_{t-1}^l + \mathbf{W}^l g_{l-1}(\mathbf{W}^{l-1}(\cdots g_1(\mathbf{W}^1 \mathbf{x}_{\boldsymbol{t}})))) \tag{3.2.3}$$

The output $\mathbf{y}_t$ is expressed as:

$$\mathbf{y}_t = f_o(\mathbf{h}_t^l)$$

$$= \mathbf{W}^L g_{L-1}(\mathbf{W}^{L-1}(\cdots g_l(\mathbf{W}^l \mathbf{h}_t^l))) \tag{3.2.4}$$

where f_h and f_o are the state transition and output function, respectively. The input at time t is $\mathbf{x}_t$, $g(\cdot)_l$ represents the activation function at the l-th layer, $\mathbf{R}^l$ is the recurrent weight matrix and $\mathbf{W}^l$ is the current connection at the l-th layer. In the layers without temporal connection, the previous weight matrices are the zero matrices.

The full connection DRNN has the same architecture as the vanilla RNN [81], the hidden state of the l-th layer at time t is:

$$\mathbf{h}_t^l = f_{\mathrm{h}}(\mathbf{h}_t^{l-1}, \mathbf{h}_{t-1}^l) = g_l(\mathbf{R}^l \mathbf{h}_{t-1}^l + \mathbf{W}^l \mathbf{h}_t^{l-1}) \tag{3.2.5}$$

In the first layer, where $l = 1$, the activation $\mathbf{h}_t^1$ is calculated by $\mathbf{h}_t^0 = \mathbf{x}_t$. In the DRNN, the activation function is selected as a rectified linear unit (ReLU) to avoid gradient vanishing and reduce the computational cost. The ReLU function is expressed as:

$$g(\mathbf{x}) = max(\mathbf{0}, \mathbf{x}) \tag{3.2.6}$$

After the DRNN is selected as the framework, the adaptive discriminative term is added into the cost function to address ambiguous information issue, which is introduced in the next subsection.

3.2.4 Adaptive discriminative term

Assume two sources are used to generate speech mixture, which is represented by $\mathbf{s}_1$ and $\mathbf{s}_2$. By optimizing the parameters of the neural network, the mapping relationship between the input, $\mathbf{x}_t$, and the estimations, $\hat{\mathbf{s}}_{1t}$ and $\hat{\mathbf{2}}_{2t}$, can be obtained. The sum of the squared errors is selected as the objective function as:

$$J = \frac{1}{2}\sum_{t=1}^{T}(||\hat{\mathbf{s}}_{1t} - \mathbf{s}_{1t}||_2^2 + ||\hat{\mathbf{s}}_{2t} - \mathbf{s}_{2t}||_2^2) \tag{3.2.7}$$

where $\hat{\mathbf{s}}_{1t}$ and $\hat{\mathbf{s}}_{2t}$ are the predictions of the spectra and $\mathbf{s}_{1t}$ and $\mathbf{s}_{2t}$ represent the target spectra, $|| \cdot ||_2^2$ is the l_2 norm operation, and (3.2.7) needs to be minimized to optimize the parameters in the neural network.

In the DRNN, the input is a concatenation of features; when the features are similar, the neural network will be conservative in the training stage. Because of the similarity, a feature can be attributed to source 1 or source 2 in some cases. To maintain the efficiency of the training stage, the neural network will attribute the feature to both source 1 and source 2, which is called the conservative strategy. However, if the ambiguous features are attributed repeatedly, the separation performance is decreased due to this strategy.

In [13], a discriminative network training criterion was proposed. The new discriminative objective function is defined as:

$$J_{DIS} = \frac{1}{2}\sum_{t=1}^{T}(||\mathbf{s}_{1t} - \hat{\mathbf{s}}_{1t}||^2 + ||\mathbf{s}_{2t} - \hat{\mathbf{s}}_{2t}||^2 -$$

$$\gamma||\mathbf{s}_{1t} - \hat{\mathbf{s}}_{2t}||^2 - \gamma||\mathbf{s}_{2t} - \hat{\mathbf{s}}_{1t}||^2) \tag{3.2.8}$$

where γ can be treated as the penalty parameter. In the ideal case, $\hat{\mathbf{s}}_{1t}$ and $\hat{\mathbf{s}}_{2t}$ are only estimated by the corresponding target features. However, because of the indeterminacy and conservative strategy, this case cannot happen. Therefore, how to minimize the negative influence from these ambiguous features is important. The $||\mathbf{s}_{1t} - \hat{\mathbf{s}}_{2t}||^2$ and $||\mathbf{s}_{2t} - \hat{\mathbf{s}}_{1t}||^2$ terms are used to represent the squared errors, which are caused by attributing the estimated features, $\hat{\mathbf{s}}_{1t}$ and $\hat{\mathbf{s}}_{2t}$, incorrectly. To be noted, in the training stage of the proposed method, the levels of amplitude of the speech signals and interferences are same.

According to work in [13], γ is selected in the range of 0.01$\sim$0.1, empirically. Whereas the speech signals are random with high indeterminacy. If the value of γ is irrelevant to inputs, when the inputs for training stage are changed, the performance and the trained network may not be amenable. Therefore, the penalty parameter is calculated adaptively, which is applied to penalize the cost function to train the neural networks.

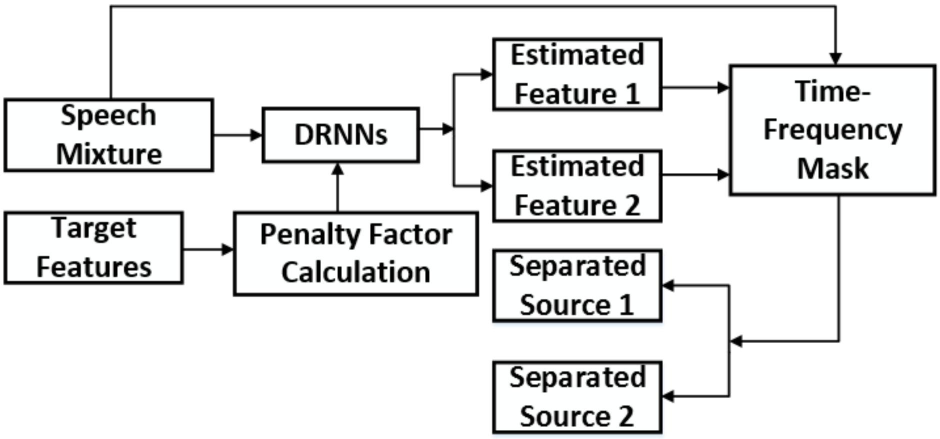

Figure 3.1. The training stage of the proposed DRNN-based MSS system with adaptive discriminative term.

Fig. 3.1 is the flow diagram of the proposed DRNN-based adaptive penalty method in training stage. Before training the neural network, a penalty factor calculation module is added to compute the parameter in the discriminative term to penalize the objective function. Then, in the training stage, the parameters of the DRNN are optimized with the penalty factor and discriminative criterion.

In this proposed method, the value of γ in (3.2.8) is changed with the input features. To be specific, if the input features are almost the same, it indicates that features are more likely to be attributed to both source 1 and source 2. Therefore, the penalty term needs to be significant and the γ

requires a greater value. In contrast, when the targets have huge differences, the conservative strategy and penalty factor are trivial in this situation and γ should be close to zero. According to the analysis above, the value of the penalty factor is inversely proportional to the discrepancy between target features.

Generally, norms of matrix are used to measure the discrepancy and three types of norms are explored.

Assume the spectra of source 1 and source 2 are, respectively, $\mathbf{A} \in \mathbb{R}^{F \times T}$ and $\mathbf{B} \in \mathbb{R}^{F \times T}$. The discrepancy between the features is defined as:

$$\mathbf{D} = \mathbf{A} - \mathbf{B} \tag{3.2.9}$$

The penalty factor is calculated as:

$$\gamma = \frac{1}{\|\mathbf{D}\|_{norm}} \tag{3.2.10}$$

Because the discrepancy between two features needs to be measured, firstly, the max norm is utilized, which is defined as:

$$\|\mathbf{D}\|_{max} = \max|d_{t,f}| \qquad \forall\ t, f \tag{3.2.11}$$

where $d_{t,f}$ is the element in the matrix $\mathbf{D}$, t and f represent the frame and frequency index: $t = 1, \ldots, T$ and $f = 1, \ldots, F$.

However, the max norm only finds the maximum value of the matrix, it cannot fully measure the total discrepancy. Hence, the P-norm will be discussed below [82].

The P-norm of matrix $\mathbf{D}$ is defined as:

$$\|\mathbf{D}\|_P = (\sum_{t=1}^{T}\sum_{f=1}^{F}|d_{t,f}|^P)^{\frac{1}{P}} \tag{3.2.12}$$

where P is the positive integer.

Two cases in the P-norm are discussed, where the value of P is selected as 1 or 2.

For $P = 1$:

$$\|\mathbf{D}\|_1 = \sum_{t=1}^{T}\sum_{f=1}^{F}|d_{t,f}| \tag{3.2.13}$$

For $P = 2$:

$$\|\mathbf{D}\|_2 = (\sum_{t=1}^{T}\sum_{f=1}^{F}|d_{t,f}|^2)^{\frac{1}{2}} = \sqrt{trace(\mathbf{D}\cdot\mathbf{D}^*)} \tag{3.2.14}$$

where $\mathbf{D}^*$ denotes the conjugate transpose of $\mathbf{D}$ and *trace* is the trace operation of the matrix. It is well known as the Frobenius norm.

Theoretically, from the definition of the 2-norm, it can be known that it shrinks the difference between inputs. Therefore, the algorithm based on the 1-norm should have a better separation performance.

Moreover, for any two matrix norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$, they have the relationship for some positive constants δ and θ and all matrices $\mathbf{D}$ in $\mathbb{R}^{F\times T}$. It is defined as:

$$\delta\|\mathbf{D}\|_\alpha\leqslant\|\mathbf{D}\|_\beta\leqslant\theta\|\mathbf{D}\|_\alpha \tag{3.2.15}$$

The above equation indicates that all norms on $\mathbb{R}^{F\times T}$ are equivalent [83]. However, in a specific algorithm, the 1-norm and the 2-norm will show different performance.

Finally, the type of norm in (3.2.10) is selected as the 1-norm and the penalty factor is calculated as:

$$\gamma = \frac{1}{\|\mathbf{D}\|_1} = \frac{1}{\|\mathbf{A}-\mathbf{B}\|_1} \tag{3.2.16}$$

Therefore, the γ can be calculated adaptively with the changes of target features.

The evaluations with the propose DRNN will be shown in Section 3.4 and in the next section, the sequentially trained DNN system is described, which is another promising method to address speech enhancement problem.

3.3 Sequentially Trained DNN System for Speech Enhancement

In this section, two DNNs are employed to build a enhancement system. In the system, different from the single DNN method, one more DNN is trained to further eliminate the overestimate or underestimate problems in the enhanced speech signal.

Based on the mapping- and masking-based methods, a DNN-based system with two sequentially trained DNNs is proposed to further improve the enhancement performance. According to no free lunch theorems (NFL) [84], it is impossible to find the neural network model which can estimate the training target perfectly in all cases. Hence, the desired speech signal is divided into two components, the estimated component and the enhanced component, $|\hat{S}_1(t, f)|$ is used to represent the estimated component which can be obtained from the trained mapping-based DNN and $|S_2(t, f)|$ is used to represent the enhanced component of the magnitude information.

Hence, the desired speech signal can be rewritten as:

$$|S(t,f)| = |\hat{S}_1(t,f)| + |S_2(t,f)| \tag{3.3.1}$$

In the proposed system, a T-F mask is generated from the second trained DNN to obtain the magnitude information to enhance the estimated speech signal. Different from the first DNN (i.e. the mapping-based DNN), the training target of the second DNN is a T-F mask. According to (3.3.1), the training target of the second DNN is expressed as:

$$M(t,f) = \frac{\Big(|S(t,f)| - |\hat{S}_1(t,f)|\Big)}{|Y(t,f)|} \tag{3.3.2}$$

The value of $M(t, f)$ can be negative and according to [50], the value range of mask is not limited within [0,1]. Since $|\hat{S}_1(t,f)|$ is not very far from $|S(t,f)|$, the value range of $M(t,f)$ is not very large and the compression module is not required in our proposed method. If the value in the mask is positive, it shows that $|\hat{S}_1(t,f)|$ is underestimated. If the value in the mask is negative, it indicates that $|\hat{S}_1(t,f)|$ is overestimated. The cost function of this DNN is expressed as:

$$Loss_{\text{proposed}} = \sum_t \sum_f (\hat{M}(t,f) - M(t,f))^2 \tag{3.3.3}$$

where $\hat{M}(t,f)$ is the estimation of the proposed T-F mask.

Therefore, after both DNNs are trained in the proposed system, the enhanced target speech signal is the combination of the directly estimated spectrogram and the speech information obtained from the noisy speech mixture with the proposed T-F mask.

$$|\hat{S}(t,f)|_{final} = |\hat{S}_1(t,f)| + \hat{M}(t,f) \times |Y(t,f)| \tag{3.3.4}$$

where $|\hat{S}(t,f)|_{final}$ is the feature of the enhanced speech signal, which can be used to recover the desired speech signal and $|\hat{S}(t,f)|_{final} \geq 0$.

Fig. 3.2 shows the block diagram of the training and testing stages of the proposed system with two DNNs. The feature we used in the proposed system is the log spectrogram, and the two DNNs in the proposed method are trained sequentially. In the training stage, firstly, the log spectrogram of noisy speech mixture is used as input of the mapping-based DNN (DNN 1). The corresponding training target is the log spectrogram of clean speech signal. After the mapping-based DNN is trained, the feature of noisy speech mixture is given as input to DNN 1 to obtain the estimated log spectrogram of the desired speech signal. It is used to calculate the ideal T-F mask with features of noisy speech mixture and clean speech signal using (3.3.2).

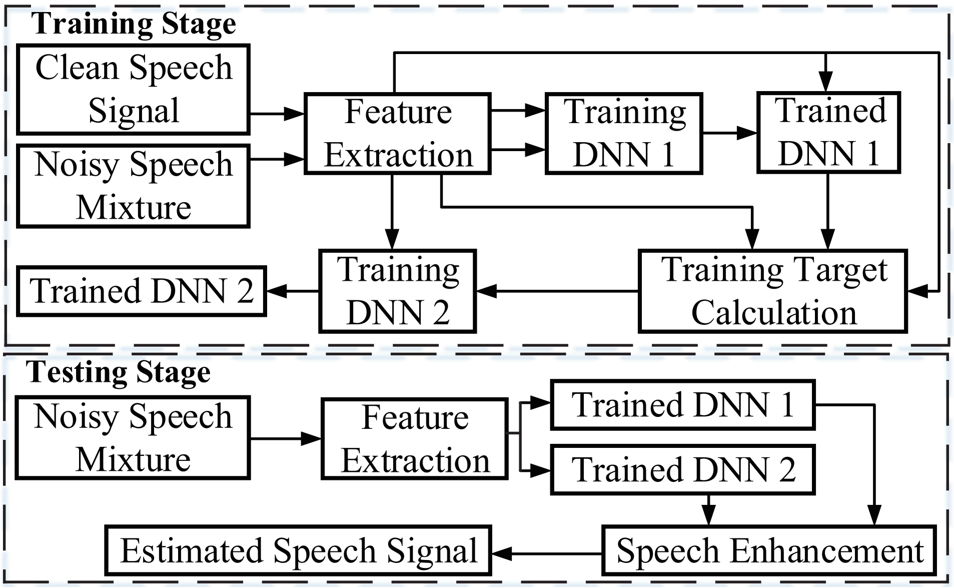

Figure 3.2. The block diagram of the training and testing stages in the proposed two-DNN system. In the training stage, the DNN 1 is mapping-based and DNN 2 is masking-based, and both DNNs are trained with the same input. In the testing stage, both trained DNNs have same input and the final estimated speech signal is obtained from speech enhancement module.

Finally, the ideal T-F mask is applied as the training target of the masking-based DNN (DNN 2) with feature of noisy speech mixture as input. In the training stage, DNNs 1 & 2 are trained sequentially, but in the testing stage, both trained DNNs output estimation in parallel.

In the testing stage, the feature of the noisy speech mixture is given as input to the proposed system, the main part of the desired speech signal is obtained by the trained DNN 1 e.g. $|\hat{S}_1(t,f)|$, and the information that cannot be estimated in DNN 1 is obtained from the trained DNN 2 by using the estimated T-F mask. The speech enhancement module is used to output the enhanced speech signal from the noisy speech mixture by (3.3.4), which yields the enhanced desired speech signal.

Fig. 3.3 is an example of the process and it can be observed that by using the proposed method, the spectrogram of the estimated speech signal

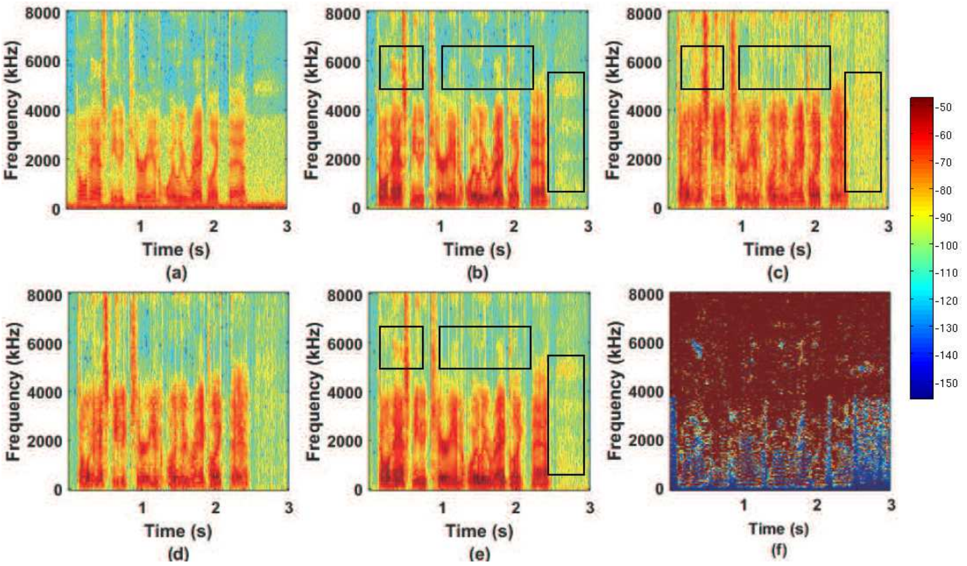

Figure 3.3. (a) The mixture; (b) Clean speech signal; (c) The estimated speech signal with DNN 1; (d) The estimated speech signal with DNN 2; (e) The estimated speech signal of the proposed method and (f) The estimated T-F mask from DNN 2.

with the proposed method is more similar to that of the clean speech signal (comparing (c) and (e) with (b)). Because the DNN 2 is introduced, the information of clean speech signal can be accurately obtained in the desired speech signal. In the unseen noise interferences case, the proposed system can utilize DNN 2 to further improve the quality of the estimated speech signal from noisy speech mixture. In Fig. 3.3 (d), the estimated speech signal with DNN 2 is shown, it represents the enhanced speech information which is used to improve the quality of the estimated speech signal. The values in some estimated T-F masks from DNN2 are checked, the negative values in the T-F mask are very sparse and the ratio of number of negative values to the number of values in T-F mask is 1.07%. Hence, it is difficult to observe these negative values in Fig. 3.3 (f).

When the noise interferences are unseen, it means that the noise interferences in the testing data are totally different from those in the training data. The trained DNN model in the mapping-based method cannot fit the relationship between the noisy speech mixture and clean speech signal accurately. Hence, some information of the clean speech signal may be lost. By using the proposed method, this problem is mitigated, leading to a better generalization ability for unseen mixtures as confirmed by evaluations with different performance measures in Section 3.4.

3.4 Simulations

In this section, the configurations and evaluations with DRNN-based method are shown in subsections 3.4.1 and 3.4.2, respectively. The configurations and the evaluations with sequentially trained DNN system are given in subsections 3.4.3 and 3.4.4, respectively.

3.4.1 Configurations with DRNN-based method

The separation performance is evaluated based on the famous TIMIT database, which contains broadband recordings of 630 speakers [85]. In these experiments, speech signals are selected from the TIMIT corpus randomly to constitute the training, validation and testing sets. The number of mixtures in training, validation and testing set is 972, 216 and 108, respectively. The mixtures in these experiments are generated with different speech sources having different genders. To extract the proper spectral representation to train the networks, a 1024-point short time Fourier transform (STFT) with 50 % overlap, zero padding and Hamming window is explored. The initialization method in [86] is utilized to reduce the training difficulty of deep networks.

The circular shift in the time domain is explored to increase the variety of training set [72]. The spectra and log power spectra are utilized as the types of input features, which are calculated by using the HTK toolkit [87].

The basic DNN, the DRNN with first layer connection, the DRNN with second layer connection and full connected DRNN are the four different architectures of neural networks. All of experiments are based on these architectures to identify generalization ability of the proposed method.

In these networks, the number of hidden layers is two and the number of hidden units on each layer is 1000. The SDR is utilized to measure the separation performance of the proposed method [88]. The limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method is an optimization algorithm in the family of quasi-Newton methods, which is used to train the models [89]. In the experiments, the values of γ are selected as 0, 1 and 0.05 (in the range of 0.01 and 0.1) for comparison. The size of context window in these networks is 1, the concatenation contains three frames, one central frame and two window frames. According to the analysis in subsection 3.2.4, the 1-norm is applied to calculate the discrepancy in the target features.

3.4.2 Results with DRNN-based method

By using the configurations mentioned in subsection 3.4.1, the corresponding neural network models are trained. After these different neural networks are trained, the mixture is separated by using different mask functions and values of γ.

Table 3.1. Separation performance comparison in terms of SDR (dB) with different values of γ and neural network architectures via binary mask and the input features are spectra.

Penalty factor γ	DNN	DRNN-1	DRNN-2	RNN
$\gamma = 1$	5.49	5.61	6.60	6.56
$\gamma = 0$	5.25	5.38	6.57	5.91
$\gamma = 0.05$	5.50	5.58	6.52	6.72
Adaptive γ	5.81	5.96	6.66	6.84

The experimental results are compared in terms of different aspects. Firstly, it can be seen from Tables 3.1 & 3.2 and Tables 3.3 & 3.4 that

Table 3.2. Separation performance comparison in terms of SDR (dB) with different values of γ and neural network architectures via binary mask and the input features are log power spectra.

Penalty factor γ	DNN	DRNN-1	DRNN-2	RNN
$\gamma = 1$	5.56	5.89	6.12	6.62
$\gamma = 0$	5.13	6.16	5.88	7.01
$\gamma = 0.05$	6.28	6.56	6.27	6.94
Adaptive γ	6.79	6.89	6.87	7.11

Table 3.3. Separation performance comparison in terms of SDR (dB) with different values of γ and neural network architectures via soft mask and the input features are spectra.

Penalty factor γ	DNN	DRNN-1	DRNN-2	RNN
$\gamma = 1$	6.08	6.17	7.00	7.07
$\gamma = 0$	5.72	6.20	7.12	6.60
$\gamma = 0.05$	6.14	6.25	7.24	7.52
Adaptive γ	6.30	6.70	7.48	7.56

the separation performance is impacted by the types of features in different architectures of networks. Generally, in DNN and DRNN-1, using the log power spectra as the input features has better performance. In contrast, the spectra can yield a higher SDR in DRNN-2 and full RNN. Then, according to the Tables 3.1 & 3.3 and Tables 3.2 & 3.4, the soft mask based models outperform binary mask based models greatly. It is evident that the soft mask can have around 10% more improvements in SDR.

Finally, the performance between different architectures is compared. The results in all Tables confirm the separation performance and robustness of the proposed method are improved in all architectures of DRNNs. Besides, comparing the separation performance of DNN and DRNNs, introducing the connected layer in networks can provide improvement. In DRNNs, almost all of the full RNN maintains the highest SDR, but demands high computational power and larger memory. In these architectures with connection in hidden layers, DRNN-1, DRNN-2 and full RNN, increasing the complexities of DRNNs gains the SDR. Although the performance is affected differently

Table 3.4. Separation performance comparison in terms of SDR (dB) with different values of γ and neural network architectures via soft mask and the input features are log power spectra.

Penalty factor γ	DNN	DRNN-1	DRNN-2	RNN
$\gamma = 1$	6.01	6.13	6.75	7.21
$\gamma = 0$	6.13	6.51	6.31	6.77
$\gamma = 0.05$	6.82	7.23	7.26	7.40
Adaptive γ	7.07	7.52	7.33	7.74

for DNN and DRNNs, the proposed approach outperforms the DRNN-based method in [13].

In these experiments, the proposed method is compared with different architectures and values of penalty factors. According to Tables 3.1-3.4, the results of the proposed method surpass the experimental results, which are produced by the irrelevant parameter method. The soft masking function can assist to achieve a better separation performance. Generally, the full RNN is the better choice than DNN, DRNN-1 and DRNN-2, but the requirement of computational resource will be higher, when the complexity of the network is increased.

3.4.3 Configurations with sequentially trained DNN system

In this subsection, the evaluations with sequentially trained DNN system are given. For experiments, the clean speech signals are randomly selected from the TIMIT corpus [85], which has 6,300 utterances, 10 utterances spoken by each of 630 speakers. The noise interferences are selected from the non-speech noise database [90] and the NOISEX database [91]. In the experiments, we select 4,680 utterances from the TIMIT corpus to generate the training dataset. 200 utterances are used to generate the testing dataset. These clean speech utterances are mixed with noise at the different signal-to-noise ratio (SNR) levels. In the training dataset, 100 different types of non-speech noise interferences are used to mix with clean utterances to generate the noisy speech mixtures. In the testing data, 10 different unseen noise interferences are used to generate the noisy speech mixtures of the testing data. In our evaluations, we generate testing data at three different SNR levels (-5, 0 and 5 dB) using different types of unseen noise interferences.

In the experiments, both DNN 1 and DNN 2 have three hidden layers and each hidden layer has 2048 units. The activation function for each hidden unit is selected as the rectified linear unit (ReLU) to avoid the gradient vanishing problem and the output layer has linear units. Both DNNs are trained with 10,000 epochs and the learning rate is 0.0001. The performance metrics used are the perceptual evaluation of speech quality (PESQ) [64], the short-time objective intelligibility (STOI) [65] and the signal to distortion ratio (SDR) [88]. The values of PESQ and STOI are in the range of [-0.5, 4.5] and [0, 1], respectively. These values indicate the human speech intelligibility scores. The SDR is exploited to evaluate the overall enhancement performance. The higher values of these metrics means that the desired speech signal is better reconstructed.

Table 3.5. Enhancement performance comparison in terms of PESQ and STOI with different SNR levels. The number of types for unseen noise interferences in the experiments is **3**. Each result is the average value of 200 experiments. **BOLD** indicates the best result.

Measures	PESQ			STOI		
SNR Levels	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
Mapping-based [45]	1.61	1.74	1.92	67.53 %	72.04 %	77.27 %
Masking-based [14]	1.84	2.01	2.12	73.22 %	74.51 %	79.33 %
Proposed	**1.94**	**2.13**	**2.28**	**75.12** %	**78.51** %	**81.65** %

Table 3.6. Enhancement performance comparison in terms of PESQ and STOI with different SNR levels. The number of types for unseen noise interferences in the experiments is **5**. Each result is the average value of 200 experiments. **BOLD** indicates the best result.

Measures	PESQ			STOI		
SNR Levels	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
Mapping-based [45]	1.57	1.67	1.89	66.26 %	71.34 %	76.11 %
Masking-based [14]	1.73	1.88	2.07	72.51 %	73.38 %	78.06 %
Proposed	**1.88**	**1.99**	**2.20**	**74.11** %	**76.39** %	**80.81** %

Table 3.7. Enhancement performance comparison in terms of PESQ and STOI with different SNR levels. The number of types for unseen noise interferences in the experiments is **10**. Each result is the average value of 200 experiments. **BOLD** indicates the best result.

Measures	PESQ			STOI		
SNR Levels	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
Mapping-based [45]	1.25	1.43	1.64	59.71 %	66.43 %	71.83 %
Masking-based [14]	1.55	1.82	2.01	69.87 %	73.05 %	76.73 %
Proposed	**1.75**	**1.92**	**2.01**	**70.80** %	**75.59** %	**77.51** %

3.4.4 Results with sequentially trained DNN system

According to the configurations and performance measures in subsection 3.4.4, the simulation results of these methods are shown.

Firstly, three different types of unseen noise interferences are used to generate the testing data, then the type of unseen noise interferences is increased to 5 and 10, respectively. Tables 3.5 - 3.7 show the enhancement performance of the proposed system and the comparison group with different number of types in noise interferences and SNR levels in terms of the PESQ and STOI. From these tables, it is clear that the proposed method achieves the best performance in all scenarios and SNR levels. For example, in Tables 3.5, when the number of types in noise interferences is 3 with 0 SNR levels, the proposed system achieves 2.13 in PESQ while the mapping-based method achieves 1.74 and the PESQ value of masking-based method is 2.01.

It can be seen that when more types of unseen noise interferences exist in the testing data, the enhancement performance of both methods is decreased. For instance, comparing the STOI performance in Table 3.5 with Table 3.7. When the 3 unseen noise interferences are used to generate the speech mixture with 5 dB SNR level, the value of STOI by using the proposed system is 81.65 %. If 10 unseen noise interferences are used to generate the speech mixture with same SNR level, the performance of the STOI of the proposed system is 77.51 %, which has 4.14 % decrease. Moreover, when the SNR level is increased, the enhancement performance is improved. It can be observed from Table 3.5, when the value of the SNR level becomes larger, better performance in terms of the PESQ and STOI can be obtained. To further confirm that the proposed method outperforms the mapping- and masking-based methods, the SDR improvements are evaluated.

The improvements of SDR ($\triangle$SDR) with different SNR level and noise cases are shown in Fig. 3.4. It is clear to observe that when the value of SNR level is increased, the $\triangle$SDR becomes less. If more types of unseen

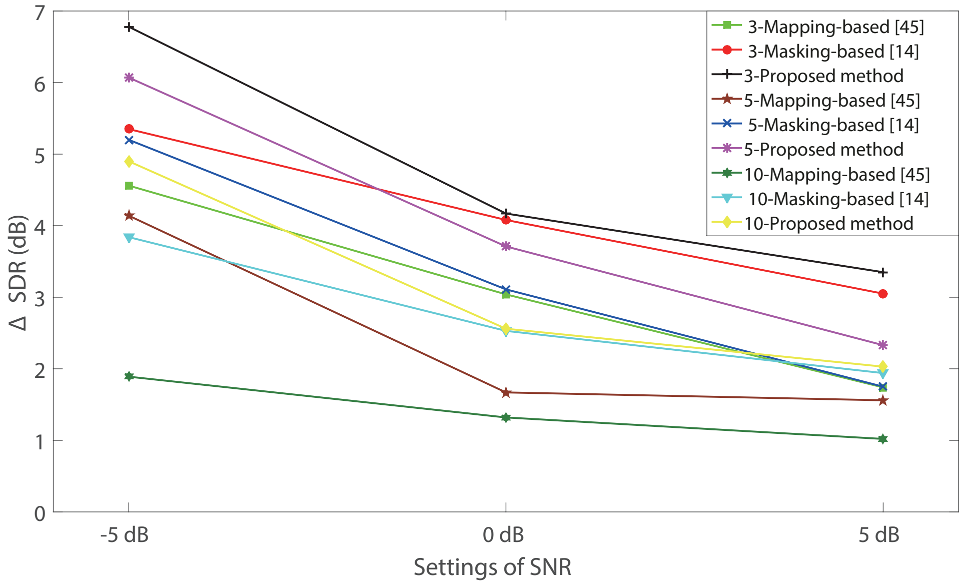

Figure 3.4. The SDR improvement (dB) in terms of different methods with various SNR levels and noise scenarios. Each result is the average value of 200 experiments. The number before each method shows the number of types for unseen noise interferences.

noise interferences are added, the value of related △SDR is decreased. From Tables 3.5 - 3.7 and Fig. 3.4, it can be seen that when the SNR level is increased, although the values of SDR for both methods are increased, the △SDR become less. When more unseen noise interferences are mixed in the testing dataset, the enhancement performance of both methods is decreased. However, comparing the results in Tables 3.5 - 3.7 and Fig. 3.4, when more types of unseen noise interferences are used in the testing dataset, e.g. 3, 5 and 10 different unseen noise interferences, the enhancement performance of the proposed method is better than the mapping- and masking-based methods. This shows that the proposed method has better generalization ability.

3.5 Chapter Summary

In summary, for the DRNN-based method with an adaptive penalty factor. From subsection 3.4.2, it can be confirmed that the adaptive criterion method outperformed the approach with irrelevant penalty factor method [13]. By introducing the penalty factor, the ambigous information problem is solved, and the separation performance is further improved. Because of the indeterminacy of speech signals in the real-world scenarios, this method can be more applicable.

In the sequentially trained DNNs system, when these two DNNs are trained sequentially, the desired speech signal is enhanced with the estimated T-F mask and the underestimated or overestimated information in the estimated speech signal can be further mitigated. The enhancement performance is influenced by the mixing SNR level and types of unseen noise interferences. Besides, more different types of unseen interferences lead to degradation in the enhancement performance. The speech enhancement system with sequentially trained DNNs is proposed to improve the performance. One DNN was mapping-based and the other masking-based. By using the proposed system, the information of a clean speech signal can be extracted accurately from the noisy speech mixture.

The evaluations with unseen noise interferences confirmed that both proposed methods outperformed the existing state-of-the-art methods at enhancement and separation performance and had a better generalization ability.

However, the separation performance of the proposed methods in this chapter is limited when the mixture is captures within real room environments. In the next chapter, a two-stage MSS method is proposed to solve the problem with reverberant environments.

Chapter 4

TWO-STAGE MONAURAL SOURCE SEPARATION WITH MAGNITUDE DOMAIN IN REVERBERANT ROOM ENVIRONMENTS USING DEEP NEURAL NETWORK

4.1 Chapter Introduction

In the previous chapters, the IRM has been introduced to address the monaural source enhancement and separation problem. Chapter 2 provided the fundamental preliminary knowledge about the masking-based method in reverberant environment. In Chapter 3, the deep recurrent neural network (DRNN) with adaptive discriminative cost function and sequentially trained DNNs were employed to obtain a more accurate prediction. However, the above mention methods have limited performance when they are used to address the monaural source separation (MSS) problem in reverberant room environments. In order to improve the separation performance in real room environments, the dereverberation mask (DM) is proposed with a two-stage algorithm. This chapter focuses on the third objective of this thesis, which is the novel masking-based method with MSS accepted by ACM/IEEE Transactions on Audio, Speech, and Language Processing.

When addressing the MSS problem in reverberant environment, the acquired speech mixture contains reflections, which will lead decrease in the separation performance [68]. The DNN-base algorithm has been employed in many scenarios, such as the speech enhancement [46] and source separation [50]. However, when they are applied in reverberant environment, in [46], because the complexity of the speech mixture is increased, the proper trained neural network model is difficult to obtain. While in [51], because the clean speech signal cannot be used to calculate the T-F mask, the direct sound is used to calculate the ideal T-F mask. Therefore, the final estimation in [51] is the direct sound, which is different from the clean speech signal.

In this chapter, the DM is firstly introduced in Section 4.2.1; based upon the DM, the IEM is proposed in Section 4.2.2. By applying DM or IEM, two different two-stage algorithms are proposed, which will be described in Section 4.3. Evaluation and comparisons will be shown at the end of this chapter in Section 4.4, which show the improvement from the proposed two-stage algorithms.

4.2 Time-Frequency Mask

In the real room environment, the sound or signal will be reflected from the the ceiling, walls and floors. Hence, in the captured speech mixture, it will contain a large number of reflections of speech signals and interferences [92]. To measure the reverberant time, RT60 is proposed and it is the time that signal takes for the pressure level to reduce by 60 dB [93]. The room impulse response is the impulse signal used to generate the reverberations in the room environment, it contains three different types signal, the direct-path, the early- and late-reflections. In recent studies, masking-based DNN methods have been introduced to solve the monaural source separation problem in reverberant environments and the corresponding performance is promising.

Assume that $s(m)$, $i(m)$ and $y(m)$ are the desired speech signal, the interference and the acquired mixture at discrete time m, respectively. The terms $h_s(m)$ and $h_i(m)$ are the room impulse responses (RIRs) for reverberant speech and interference, respectively. The convolutive mixture is expressed as:

$$y(m) = s(m) * h_s(m) + i(m) * h_i(m) \tag{4.2.1}$$

where '$*$' indicates the convolution operator. By using the short time Fourier transform (STFT), the mixture is written as:

$$Y(t, f) = S(t, f)H_s(t, f) + I(t, f)H_i(t, f) \tag{4.2.2}$$

where $S(t, f)$, $I(t, f)$ and $Y(t, f)$ are the spectra of speech, interference and mixture, respectively. The qualities $H_s(t, f)$ and $H_i(t, f)$ are the RIRs for speech and interference at time frame t and frequency f, respectively.

By employing the ideal T-F mask $M(t, f)$, the spectrum of the clean speech can be reconstructed as:

$$S(t, f) = Y(t, f)M(t, f) \tag{4.2.3}$$

The IRM and the cIRM are the two targets often chosen in state-of-the-art masking-based DNN methods, which can be used to dereverberate and separate the speech mixture.

If there is no RIR, the IRM for time frame t and frequency f can be expressed as [14]:

$$IRM(t, f) = \left(\frac{|S(t, f)|^2}{|S(t, f)|^2 + |I(t, f)|^2} \right)^{\beta} \tag{4.2.4}$$

where β is a tunable parameter to scale the mask, $|S(t, f)|$ and $|I(t, f)|$ denote the target speech signal and the noise interference magnitude spectra, respectively. Typically, the tunable parameter is selected as 0.5.

When the environment is reverberant, the direct sound at discrete time m is expressed as [51]:

$$d(m) = h_d(m) * s(m) \tag{4.2.5}$$

where $h_d(m)$ is the impulse response for the direct sound. Hence, the IRM for a reverberant environment in the time-frequency domain is expressed as [51]:

$$IRM_{rev}(t, f) = \left(\frac{|D(t, f)|^2}{|Y(t, f)|^2} \right)^{\beta} \tag{4.2.6}$$

where $|D(t, f)|$ and $|Y(t, f)|$ denote the direct sound and noisy reverberant mixture magnitude spectra, respectively.

The cIRM is a complex T-F mask which is obtained by using the real and imaginary components of the STFTs of the desired speech signal and mixture [50].

To calculate the cIRM, the STFTs of the reverberant mixture, direct sound and cIRM are written as:

$$Y(t, f) = Y_r(t, f) + jY_c(t, f) \tag{4.2.7}$$

$$D(t, f) = D_r(t, f) + jD_c(t, f) \tag{4.2.8}$$

$$cIRM(t, f) = cIRM_r(t, f) + jcIRM_c(t, f) \tag{4.2.9}$$

where $j \triangleq \sqrt{-1}$ and the subscripts 'r' and 'c' indicate the real and the imaginary components in the STFTs, respectively.

By using the ideal cIRM, the desired speech signal can be separated from the mixture. The T-F unit of the cIRM is defined as:

$$cIRM(t,f) = \frac{Y_r(t,f)D_r(t,f) + Y_c(t,f)D_c(t,f)}{Y_r^2(t,f) + Y_c^2(t,f)}$$

$$+j\frac{Y_r(t,f)D_c(t,f) - Y_c(t,f)D_r(t,f)}{Y_r^2(t,f) + Y_c^2(t,f)} \tag{4.2.10}$$

The above IRM and cIRM have the same limitation in solving dereverberation and separation problems, the final estimation is the direct sound instead of clean speech signal. To address this problem, the DM is proposed, which is described in Section 4.2.1.

4.2.1 Dereverberation mask

Because the neural network model can be used to find the relationship between the input and the training target [81]. However, estimating the separation mask directly from the reverberant mixture is challenging and the mask obtained is often noisy due to the presence of acoustic reflections. To address this issue, a DM is used to eliminate reverberation, and then the IRM is applied to separate the desired speech signal. According to (4.2.2), the reverberant mixture can be written as:

$$Y(t,f) = [S(t,f) + I(t,f)]\left(\frac{H_s(t,f)}{1+\frac{I(t,f)}{S(t,f)}} + \frac{H_n(t,f)}{1+\frac{S(t,f)}{I(t,f)}}\right) \tag{4.2.11}$$

Therefore, by using $Y(t,f)$ and $[S(t,f)+I(t,f)]$, the relationship between the reverberant and dereverberated mixtures is obtained. The DM is defined as:

$$DM(t,f) = \left(\frac{H_s(t,f)}{1+\frac{I(t,f)}{S(t,f)}} + \frac{H_n(t,f)}{1+\frac{S(t,f)}{I(t,f)}}\right)^{-1} \tag{4.2.12}$$

In the training stage, the spectra of speech, noise and mixture with reverberations are available, therefore, the DM can be learned as:

$$DM(t,f) = \big[S(t,f){+}I(t,f)\big]Y(t,f)^{-1} \tag{4.2.13}$$

From (4.2.13), it is clear that in the training stage, the training target $DM(t,f)$ can be calculated by using $S(t,f)$, $I(t,f)$ and $Y(t,f)$. Therefore, before the target signal is separated from the mixture, the DM is applied to the reverberant mixture to eliminate most of the reflections. In the training stage, the DM is compressed, and its value range is limited to be consistent with that of IRM, and thereby facilitate the fusion with IRM. According to (4.2.13), when there are no RIRs, the elements of the DM will all be ones and the proposed two-stage approach will be reduced to one-stage using only the estimated IRM.

According to (4.2.11) and (4.2.13), it can be known that the DM is a dereverberation operation. Thus, the dereverberated mixture is described as:

$$S(t,f) + I(t,f) = Y(t,f)DM(t,f) \tag{4.2.14}$$

Because the DM can only dereverberate the speech mixture, further processing is required for separating the mixture. Compared with the cIRM, the IRM requires less computational cost and both the DM and the IRM are soft masks which are applied in the T-F domain, while the cIRM is applied in the complex domain. Hence, the IRM is applied to separate the desired signal from the mixture. The desired speech signal is extracted from the dereverberant mixture by using the IRM:

$$S(t,f) = \Big(S(t,f) + I(t,f)\Big)IRM(t,f) \tag{4.2.15}$$

4.2.2 Ideal enhanced mask

From (4.2.14), it can be observed that the DM only achieves the dereverberation, a separation operation is essential to separate the desired speech signal from the dereverberated mixture and the IRM has been selected to undertake the separation operation.

As mentioned in Section 4.1, the DM and IRM are integrated together to generate the IEM. When IEM is used as the training target, only one DNN is trained. The IEM is defined as:

$$IEM(t,f) = DM(t,f)IRM(t,f) \tag{4.2.16}$$

Comparing the proposed IEM with the IRM_{rev}, the proposed single DNN method is essentially different from the one in [51]: the $IRMrev$ is calculated based on the direct sound, which is a delayed and attenuated version of the clean speech signal. Hence, after using the T-F mask, the STFT of the direct sound is obtained [94]. However, in real scenarios, $h_d(m)$ in (4.2.1) is not equal to 1 and as a result, $IRMrev$ is not always effective in mitigating the reverberation effect. While in our proposed IEM, the IRM is calculated by using the clean speech signal and the dereverberant mixture, after using the T-F mask, the STFT of the clean speech signal can be obtained. Therefore, compared with the $IRMrev$, the IEM achieves better separation performance. In addition, the compression module is added to restrict the range of the values within the IEM, which is conducive for training the DNN.

According to (4.2.14) and (4.2.15), it can be seen that the DM is a dereverberation operator and the IRM is the separation operator. Thus, the separated speech signal is obtained as:

$$S(t,f) = Y(t,f)IEM(t,f) \tag{4.2.17}$$

The value range of the proposed DM is (0, $+\infty$), when the DM is integrated with the IRM as the training target, the value range of the DM is not consistent with IRM, and hence the mapping relationship is difficult to find. To address this issue, (4.2.18) is used to compress the DM to restrict its value range in order to make it consistent with the IRM and convert it back to the original value range in the testing stage by using (4.2.19). Empirically, in the training stage, the compressed IEM is written as:

$$IEM_c(t,f) = V\frac{1-e^{-C\cdot IEM(t,f)}}{1+e^{-C\cdot IEM(t,f)}} \tag{4.2.18}$$

where C is the steepness constraint and the value of $IEM_c(t,f)$ is limited in the range $[-V,V]$. Because the magnitude information is used to calculate the IEM, the value of $IEM_c(t,f)$ is restricted in the range $(0,V]$. After the validation tests in the experiments, the values of C and V are chosen as 1 and 10, respectively. These values were found based on the datasets described in Section 4.4.

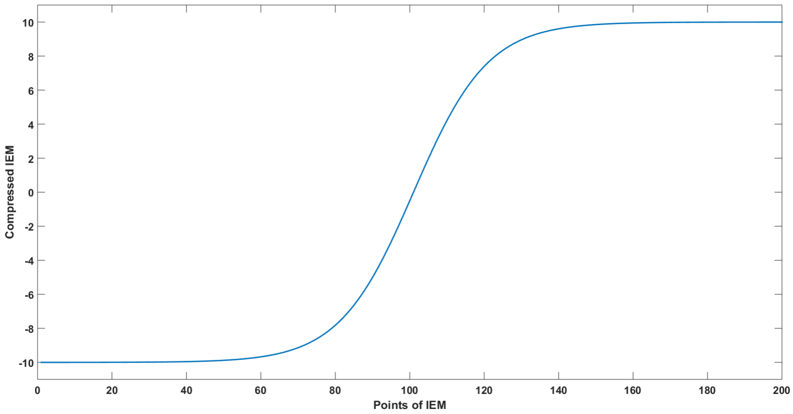

Figure 4.1. Value range of compressed IEM with different value of IEM.

For other datasets, C and V could be choosen in a similar way.

In the testing stage, the estimation of the compressed IEM is recovered and the final predicted IEM is expressed as:

$$I\hat{E}M(t,f) = -\frac{1}{C}\log(\frac{V - O(t,f)}{V + O(t,f)}) \tag{4.2.19}$$

where $O(t,f)$ is the estimation of the compressed IEM.

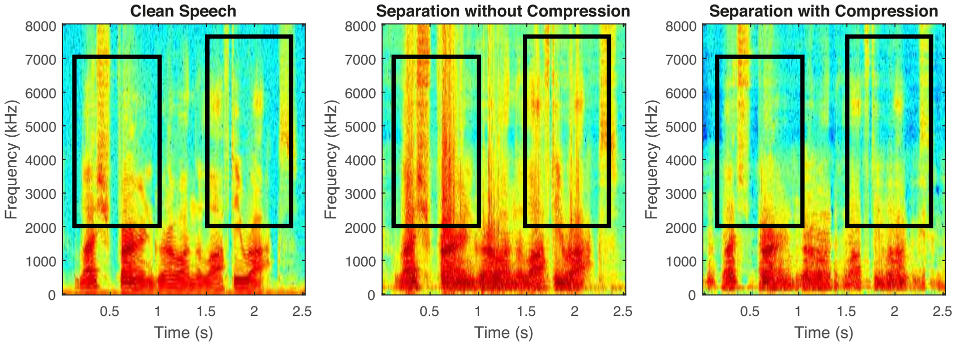

Figure 4.2. Spectrogram plots of the clean speech signal (left), separated speech signal without compression module (middle) and separated speech signal with compression module (right). The reverberant mixture is generated with *factory* noise and $0dB$ SNR level in the *unseen* RIR case for $RT60 = 470ms$. The hyperparameters $C = 1$ and $V = 10$.

As an example, the spectrograms of the clean speech signal, the separated speech signal without compression module and the separated speech signal with compression module are shown in Fig. 4.2. It can be seen that the compression module is important for the DM, which can eliminate noise in the high frequency component of the separated speech signal.

4.3 Two-Stage Algorithm

Based on the DM and the IEM in Section 4.2, two different two-stage algorithms are proposed. The differences between these two algorithms are the training targets and the number of neural network models.

4.3.1 Integrated Training Target

In the proposed two-stage approach with integrated training target, inspired by [51, 95], the feature combination is given to train the DNNs to refine the performance. The amplitude modulation spectrogram (AMS) [96], relative spectral transform and perceptual linear prediction (RASTA-PLP) [97], Mel-frequency cepstral coefficients (MFCC), cochleagram response and their deltas are extracted by a 64-channel gammatone filterbank to obtain the compound feature [98]. MFCC are coefficients that collectively build an MFC, which are derived from a type of cepstral representation of the audio clip. In the MFC, the frequency bands are equally spaced on the mel scale, which imitate the human auditory system's response. The MFCC is an important to represent the feature of speech signal. The feature combination is extracted in the feature extraction module. To update the DNN weights, the backward propagation algorithm is exploited and the mean-square error (MSE) function is used in the cost function.

The cost function of the proposed single DNN-based method is expressed as:

$$J_1 = \frac{1}{2N} \sum_t \sum_f [O(t, f) - IEM_c(t, f)]^2 \tag{4.3.1}$$

where N represents the number of time frames for the inputs, $O(t, f)$ is the estimation of the compressed IEM and $IEM_c(t, f)$ is the compressed IEM at a T-F unit.

Fig. 4.3 is the flow diagram of the proposed single DNN-based method with integrated training target, where (4.2.18) and (4.2.19) are achieved in the compression module and the recovery module, respectively. In the training stage, the DM and the corresponding IRM are calculated by using the target calculation module and integrated as the IEM. The IEM is compressed in the compression module to generate the training target of the single DNN. In the training stage, (4.3.1) is used to update the weights of the DNN. In

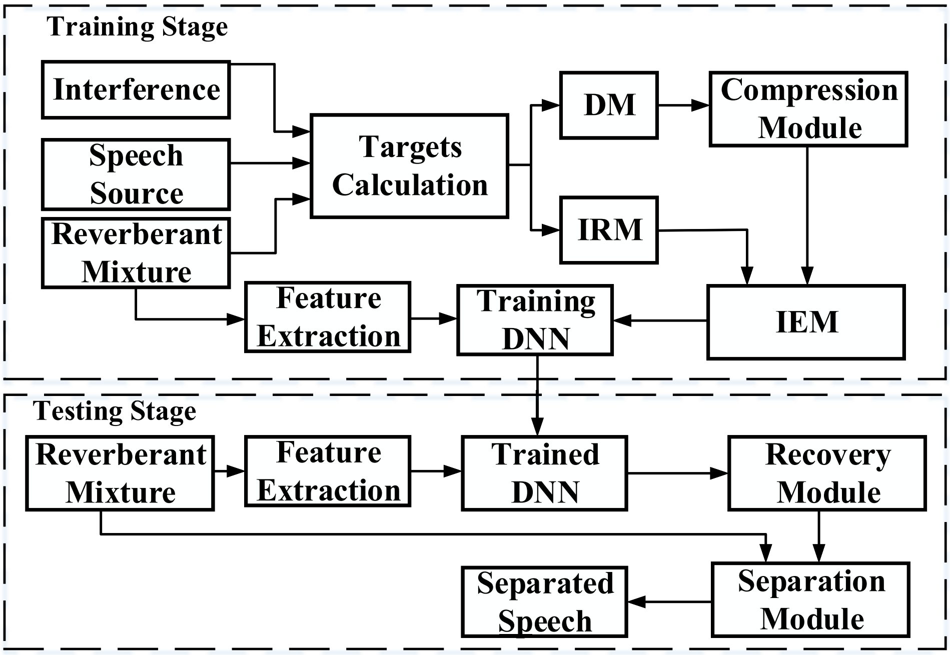

Figure 4.3. The block diagram of the proposed single-DNN based method. One DNN is trained with the integrated training target i.e. IEM. The trained DNN is given by the training stage and in the testing stage, the output of the separation module is the desired speech signal.

the testing stage, once the trained DNN is obtained, the feature combination of the mixture is extracted and input to the trained DNN. The output of the DNN is obtained in the recovery module and used to separate the desired signal. Finally, the desired speech signal is separated from the convolutive mixture with the predicted IEM in the separation module.

It is clear to see the advantages of the proposed single DNN-based method with integrated training target:

(1) Only one DNN is trained, the computational cost and the storage space requirement will be lower than the method based on two training targets with two DNNs.

(2) The dereverberation and separation are achieved by the IEM, in the training stage, the estimation error will be decreased by generating the integrated training target. Compared with the traditional IRM, the IEM can

achieve better separation performance because the DM is used to eliminate the reflection and the IRM is exploited to estimate the source from the dereverberated mixture.

4.3.2 Separate Training Targets

In the proposed second method, two DNNs are trained to model the relationships from the inputs to the DM and the IRM, respectively. In this method, the two T-F masks are predicted, the DM is applied for dereverberation, then the dereverberated mixture is separated by using the IRM. The compression and recovery processes are only applied to the DM, which is similar to the first method.

Assume the predicted dereverberation mask is $D\hat{M}(t,f)$ and the predicted ideal ratio mask is $I\hat{R}M(t,f)$, the separated speech signal is expressed as:

$$\hat{S}(t,f) = Y(t,f)D\hat{M}(t,f)I\hat{R}M(t,f) \tag{4.3.2}$$

Fig. 4.4 is the flow diagram of the proposed two DNN-based method with separate training targets. Because the DM is predicted by the trained DNN, the compression module and the recovery module are essential. In the training stage, the compound features extracted from the reverberant mixture are used as input to DNN2, where IRM is used as the the training target. The same compound features are used as input to DNN1, where DM (modified by the compression module) is used as the training target. In the testing stage, the reverberant mixture is used as input to estimate the DM and IRM, respectively. Since the reverberant mixture is used in the training stage for both DNN1 and DNN2, the trained network is able to generalise to reverberant mixtures in the testing stage.

$$J_2 = \frac{1}{2N}\sum_{t}\sum_{f}[O_1(t,f) - DM_c(t,f)]^2 \tag{4.3.3}$$

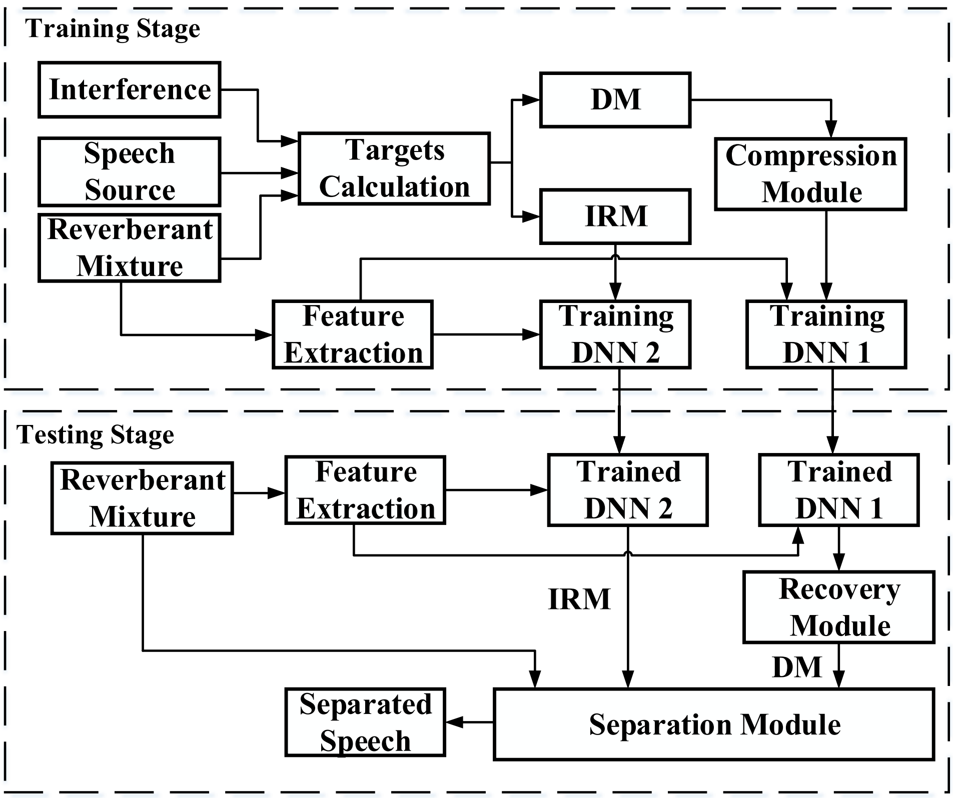

Figure 4.4. The block diagram of the proposed two-DNN based method. Two DNNs are trained with the separate training targets. Two trained DNNs are found by the training stage. In the testing stage, the dereverberated speech mixture is obtained by using the predicted DM in the dereverberation module and the desired speech signal is obtained by using the predicted IRM in the separation module, respectively.

where $O_1(t, f)$ is the output of the DNN1 at a T-F unit and $DM_c(t, f)$ is the compressed DM at a T-F unit by using (4.2.18). Similarly, for DNN2, its cost function is expressed as:

$$J_3 = \frac{1}{2N} \sum_t \sum_f [O_2(t, f) - IRM(t, f)]^2 \qquad (4.3.4)$$

where $O_2(t, f)$ is the output of the DNN2 at a T-F unit and $IRM(t, f)$ is the ideal ratio mask at a T-F unit.

In the testing stage, after the trained DNNs are obtained, the feature combination of the mixture is extracted and input to the trained DNNs. The output of the trained DNN1 is the predicted compressed DM and the output of the trained DNN2 is the predicted IRM. Then, the output of the DNN1 is obtained in the recovery module and used to eliminate the reflections. The mixture without reverberation is given by using the dereverberation module and the desired speech source is obtained from the separation module. Finally, the desired speech signal is separated from the convolutive mixture with the predicted DM and the predicted IRM.

As an example, some spectrogram plots are shown in Fig. 4.5 for the outputs from the different stages of the proposed method. It can be observed that by using the proposed DM, the reflections in the speech mixture can be eliminated. When the compression module is added (comparing (e) and (f) with (b)), the spectrogram of the separated signal with compression module is more similar to that of the clean speech signal. By adding the compression module, the noise in the high frequency component can be better removed.

In the proposed two-stage approach, before speech separation, the room reflections are better eliminated, therefore, the separation performance is improved. Therefore, in both single DNN and two DNNs methods, all factors including the training and testing datasets, the network architectures, hyperparameters and the input feature combination to train the DNNs are the same. It appears that only the training targets and the number of trained DNNs are different between these two proposed methods. Besides, because both the DM and the IRM are estimated, these two masks are more accurate, the performance is further improved with the trade-off of the computational cost.

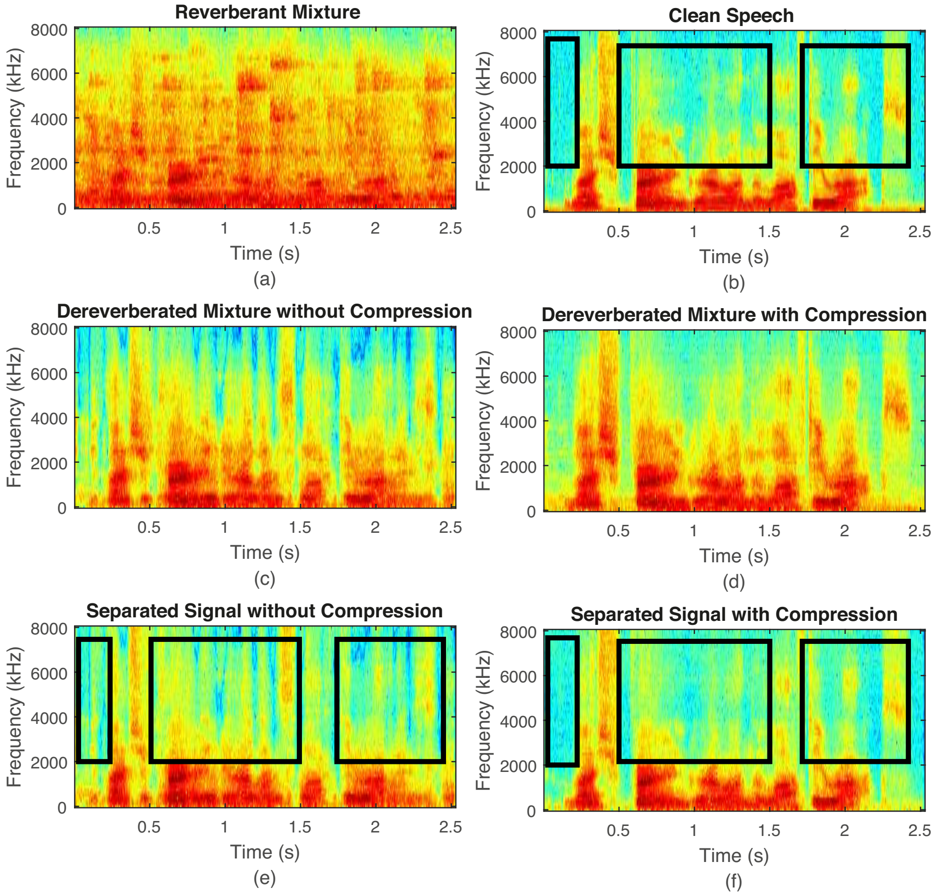

Figure 4.5. Spectrograms of different signals: (a) reverberant mixture; (b) clean speech signal; (c) dereverberated mixture without compression; (d) dereverberated mixture with compression; (e) separated speech signal without compression and (f) separated speech signal with compression. The reverberant mixture is generated with *factory* noise and $0dB$ SNR level in the *unseen* RIR case for $RT60 = 470ms$. The hyperparameters $C = 1$ and $V = 10$.

4.4 Simulations

The simulations with the proposed two-stage algorithms are shown in this section, the proposed method is evaluated with the seen RIRs and the unseen RIRs under these two different interferences. Because in the first DNN-based method with integrated training target, only one DNN is trained, we use single DNN to represent this method. Similarly, two DNNs represents the second DNN-based method with separate training targets.

4.4.1 Dataset selection and configurations

The speech sources are selected randomly from the IEEE [99] and the TIMIT corpora [85]. The IEEE corpus has 720 clean utterances spoken by a single male speaker and the TIMIT database has 6300 utterances, 10 utterances spoken by each of 630 speakers. Therefore, using both the IEEE and the TIMIT corpora can demonstrate that the proposed method is not speaker-dependent. The interferences are categorized into two aspects, the noise interference and the speech interference.

In the experiments, 1000, 100 and 120 utterances are randomly selected from the IEEE and the TIMIT corpora to generate the training, development and testing datasets. These clean utterances are used to mix with interference at three different signal-to-noise ratio (SNR) levels (-3 dB, 0 dB and 3 dB). In the evaluations with seen RIRs, the numbers of mixtures in training, development and testing data are 72,000, 7,200 and 8,640, respectively. In the evaluation with the unseen RIRs, the numbers of mixtures in training, development and testing data are 192,000, 19,200 and 9,600, respectively.

For noise interference, the noise signals are selected from the NOISEX database [91], in these noise signals, a speech-shaped noise (SSN) is generated as the stationary noise [100] and all others are the non-stationary noise, namely factory, babble and cafe. The factory noise is a recording of industrial activities and the babble noise is generated by different number of the unseen speakers in an acoustic environment. The cafe noise is more like a combination of babble and factory noise, it contains the speakers and background noise. The SSN is generated based on the clean speech corpus.

In the evaluation studies, in both training and testing stages, the target speech signals are randomly selected from the TIMIT dataset. Then, interfering speech signals are randomly selected from the remaining signals in the dataset to ensure the speakers of the target speech and the interfering speech signals are different. At the testing stage, the desired speech signals are unseen in the training stage, but the interfering speech signals are seen in the training stage. Therefore, the trained neural network is able to differentiate the target and undesirable speech signals.

To generate the speech mixture, the speech utterances and interferences are convolved with the real RIRs [92] which are recorded in four types of room environments i.e. different RT60s. The position of the desired speech signal is fixed and the azimuth of the interfering source is selected from 0° to 75° with 15° increment. Hence, each room has six different RIRs. In the evaluation with the seen RIRs, we use the RIRs from the same room to generate the training and testing datasets. In the evaluation with the unseen RIRs, for each room, four RIRs are randomly selected and used to generate the training data. The testing data are obtained by using the remaining two RIRs. Therefore, in the testing data, the RIRs are unseen and from different room environments. However, direct signals need to be generated for the baseline systems to enable comparisons with our proposed system. Firstly, the impulse response of the direct path is cropped from the whole impulse response. Then, the direct sounds are generated by using the impulse response of the direct path and clean speech signals in order to train the DNN models in [51]. Table 4.1 illustrates the parameters in the real RIRs: [92].

Table 4.1. The parameters for real RIRs in different rooms

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

The proposed method is compared with two state-of-the-art T-F masks: the IRM [14] and the cIRM [51]. Using different types of interferences, SNR levels and the RIRs in simulations show the performance of the proposed method is consistent. Moreover, when the training target is applied in the complex domain (cIRM), the corresponding DNN outputs the estimates of real and imaginary components of the predicted cIRM. The DNN needs to be Y-shaped, which has dual outputs with one input. The performance evaluation measures are the frequency-weighted segmental SNR (SNR_{fw}) [64], the source to distortion ratio (SDR) [88] and the short-time objective intelligibility (STOI) [65]. The SNR_{fw} computes a weighted signal-to-noise ratio aggregated across each time frame and critical band, it is highly correlated to human speech intelligibility scores [51]. The SDR is exploited to evaluate the overall separation performance. The values of the STOI are in the range of [0, 1], which indicate the human speech intelligibility scores. The higher values of these metrics means that the desired speech signal is better reconstructed. In terms of the STOI, the t-test is also provided to show the significant difference. T-test is the most commonly used when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. The score of the t-test shows the difference between two groups and the difference within the groups. Therefore, the t-test is always used to confirm the significant discrepancy between two sets of results. If the value of the t-test is smaller than 0.05, it indicates significant difference exists between two result sets. Besides, the IRM_{rev} and $cIRM$ in [51] are trained with direct sound, however, in real applications, the direct sound is difficult to obtain and the clean speech signal is used as reference in all performance measures.

In the proposed two-stage approach, the DNNs in the integrated training target and the separate training targets methods have the same architecture. All of the DNNs have three hidden layers and each hidden layer has 1024 units. The activation function for each hidden unit is selected as the rectified linear unit (ReLU) to avoid the gradient vanishing problem and the output layer has linear units [51]. The DNNs are trained by using the AdaGrad algorithm [101] with a momentum term for 100 epochs. The learning rate is linearly decreased from 1 to 0.01, while the momentum is fixed as 0.9 in the first ten epochs and changed to 0.5 till the end. Auto-regressive moving average (ARMA) filtering is a combination of auto-regressive and moving average filters. The output of ARMA filter is a linear combination of both the weighted input and weighted output samples. Besides, the processes of ARMA filter can be considered as a digital IIR filter, with both poles and zeros, which is applied to reduce the interference from the background noise, as in [102].

4.4.2 Evaluations with the noise interferences

In this subsection, the noise is selected as the interference, and both seen RIRs and unseen RIRs are used to generate the testing mixtures to further evaluate the generalization ability of the proposed methods.

In these experiments, the proposed methods are evaluated with the seen RIRs in four rooms. The SNR_{fw} and the SDR performance of the proposed methods and the comparison groups are given in Figs. 4.6 & 4.7, respectively. The STOI performance is shown in Tables 4.2 - 4.5, .

From Figs. 4.6 & 4.7, it is clear that when the type of noise interference varies, the performance of the IRM and the cIRM-based methods is not consistent and robust. In the noise interference case, compared with the proposed two-stage approach with single DNN, the proposed two-stage approach with two DNNs produces better results for source separation from the convolutive mixture. In the high SNR level and low RT60, the proposed two-stage approach achieves high separation performance. Compared with the IRM- and the cIRM-based DNN methods, both our proposed methods provide improved performance in terms of the SNR_{fw} and SDR consistently.

To further analyze the proposed two-stage approach, the STOI perfor-

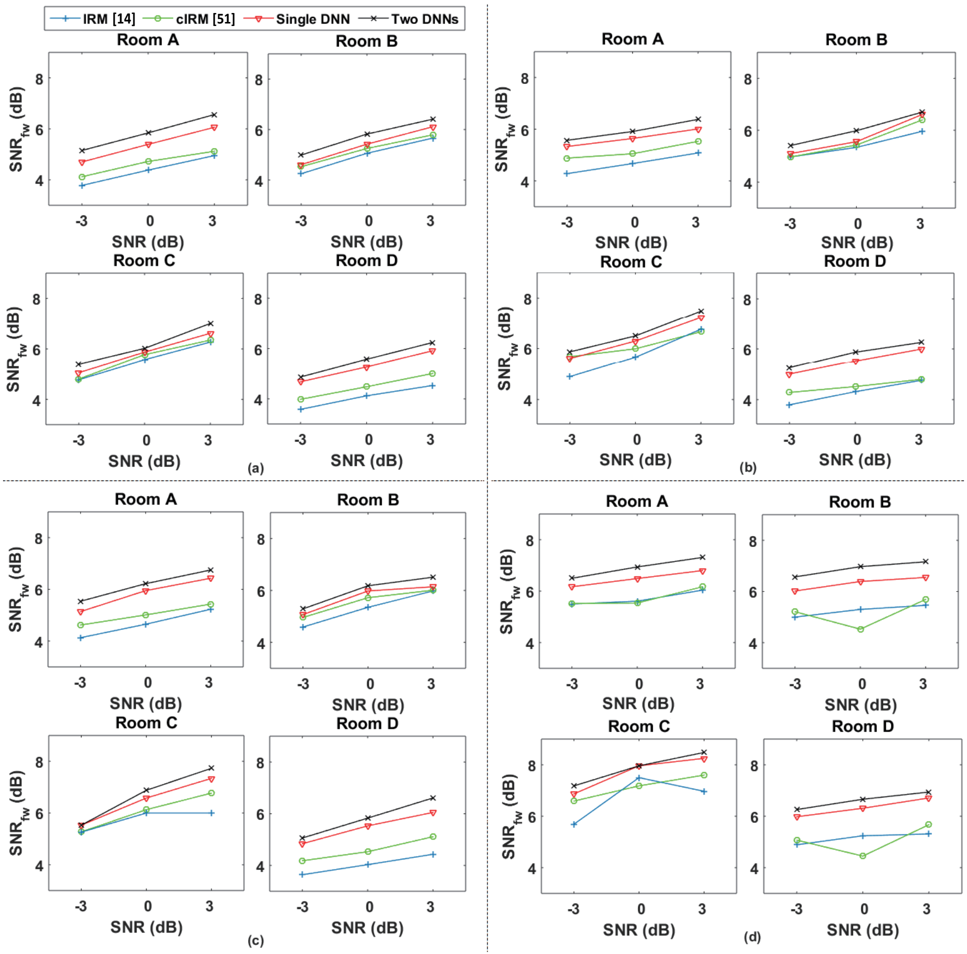

Figure 4.6. The SNR_{fw} (dB) in terms of different methods with various rooms. The X-axis is the SNR level, the Y-axis is the SNR_{fw} (dB), each result is the average value of 120 experiments. The noise types in the subfigures (a), (b), (c) and (d) are factory, babble, cafe and SSN, respectively.

mance is evaluated. The STOI performance of different methods using the IEEE and the TIMIT corpora with different noise and room environments are shown in Tables 4.2 - 4.5. It can be further confirmed that the proposed two-stage approach outperforms the state-of-the-art masking-based methods in different noise interference and reverberant environments from Tables 4.2 - 4.5. With the increase of the RT60, the proposed methods give more STOI improvements. In some cases, the cIRM-based method gives the same STOI

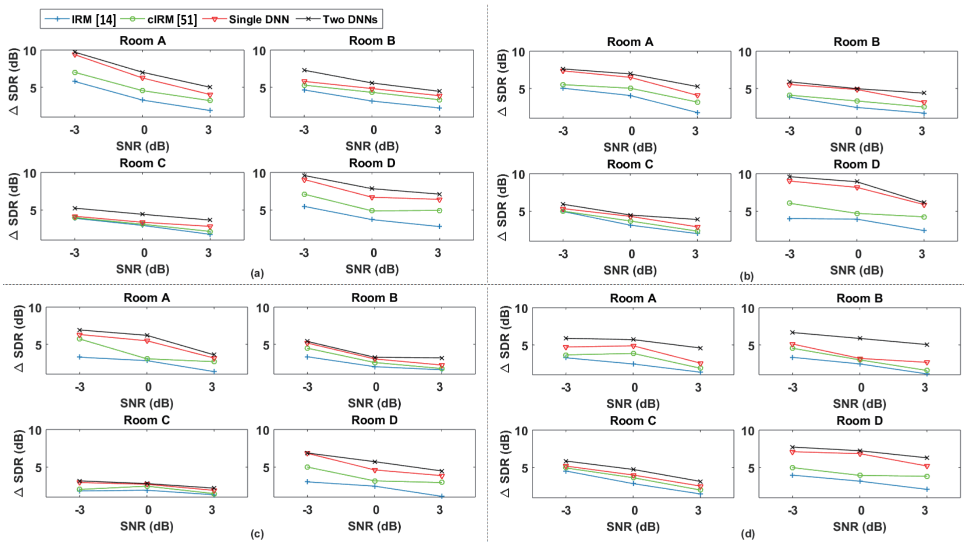

Figure 4.7. The SDR improvement (dB) in terms of different methods with various rooms. The X-axis is the SNR level, the Y-axis is the ΔSDR (dB), the improvements of the SDR. Each result is the average value of 120 experiments. The noise types in the subfigures (a), (b), (c) and (d) are factory, babble, cafe and SSN, respectively.

performance as or does slightly better than the proposed methods, e.g. SSN is used as interference with 0 SNR level in Room C. In terms of the average result, however, the proposed two-stage approach achieves the highest value. The trend of the STOI is the same as that of the SNR$_{fw}$ and the SDR.

Table 4.2. Separation performance comparison in terms of STOI with different training targets, SNR levels and RT60s. The noise in the experiments is ***factory*** noise. Each result is the average value of 120 experiments. **BOLD** indicates the best result.

Factory	Room A (0.32 s)			Room B (0.47 s)			Room C (0.68 s)			Room D (0.89 s)		
Noise	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Mixture	0.54	0.59	0.64	0.52	0.56	0.61	0.54	0.60	0.64	0.46	0.49	0.51
IRM [14]	0.66	0.71	0.76	0.64	0.69	0.73	0.67	0.71	0.77	0.60	0.63	0.66
cIRM [51]	0.66	0.72	**0.77**	0.65	0.69	0.74	0.67	0.73	0.77	0.61	0.64	0.68
Single DNN	**0.68**	0.72	**0.77**	**0.66**	0.72	0.76	0.67	**0.74**	**0.78**	**0.63**	**0.69**	0.73
Two DNNs	**0.68**	**0.73**	**0.78**	**0.66**	**0.73**	**0.77**	**0.68**	**0.74**	**0.78**	**0.63**	**0.69**	**0.74**

Table 4.3. Separation performance comparison in terms of STOI with different training targets, SNR levels and RT60s. The noise in the experiments is ***babble*** noise. Each result is the average value of 120 experiments. **BOLD** indicates the best result.

Babble	Room A (0.32 s)			Room B (0.47 s)			Room C (0.68 s)			Room D (0.89 s)		
Noise	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Mixture	0.54	0.59	0.65	0.53	0.58	0.62	0.55	0.61	0.66	0.47	0.49	0.51
IRM [14]	0.69	0.73	0.77	0.68	0.70	0.73	0.71	0.74	0.78	0.63	0.65	0.66
cIRM [51]	0.70	0.73	0.77	0.67	0.72	0.74	0.71	0.74	0.76	0.65	0.66	0.72
Single DNN	0.70	**0.75**	0.77	0.68	**0.74**	0.74	**0.73**	**0.76**	**0.79**	**0.67**	0.70	0.74
Two DNNs	**0.71**	**0.75**	**0.79**	**0.69**	**0.74**	**0.77**	**0.73**	**0.76**	**0.79**	**0.67**	**0.71**	**0.75**

Table 4.4. Separation performance comparison in terms of STOI with different training targets, SNR levels and RT60s. The noise in the experiments is ***cafe*** noise. Each result is the average value of 120 experiments. **BOLD** indicates the best result.

Cafe	Room A (0.32 *s*)			Room B (0.47 *s*)			Room C (0.68 *s*)			Room D (0.89 *s*)		
Noise	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Mixture	0.59	0.65	0.69	0.57	0.62	0.67	0.61	0.66	0.72	0.48	0.51	0.57
IRM [14]	0.67	0.73	0.76	0.65	0.70	0.74	0.68	0.74	0.79	0.58	0.62	0.65
cIRM [51]	**0.68**	0.76	0.79	0.66	0.71	0.75	0.68	0.75	0.80	0.58	0.63	0.65
Single DNN	**0.68**	0.76	0.79	**0.67**	**0.75**	**0.78**	**0.69**	**0.76**	**0.81**	0.60	0.70	0.73
Two DNNs	**0.68**	**0.77**	**0.80**	**0.67**	**0.75**	**0.78**	**0.69**	**0.76**	**0.81**	**0.65**	**0.71**	**0.76**

Table 4.5. Separation performance comparison in terms of STOI with different training targets, SNR levels and RT60s. The noise in the experiments is ***SSN*** noise. Each result is the average value of 120 experiments. **BOLD** indicates the best result.

SSN	Room A (0.32 *s*)			Room B (0.47 *s*)			Room C (0.68 *s*)			Room D (0.89 *s*)		
Noise	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Mixture	0.60	0.65	0.70	0.59	0.64	0.68	0.62	0.67	0.73	0.51	0.53	0.56
IRM [14]	0.78	0.80	0.81	0.76	0.78	0.79	0.78	**0.82**	0.84	0.70	0.72	0.73
cIRM [51]	0.72	0.77	0.80	0.76	0.79	0.80	**0.79**	0.81	0.85	0.71	0.74	0.75
Single DNN	0.78	0.81	0.82	0.77	**0.80**	**0.81**	**0.79**	**0.82**	**0.86**	0.74	0.76	0.77
Two DNNs	**0.79**	**0.82**	**0.84**	**0.78**	**0.80**	**0.81**	**0.79**	**0.82**	**0.86**	**0.75**	**0.77**	**0.80**

To show the difference of the STOI performance between the cIRM-based method and the proposed method with two DNNs, the t-test is used. *For example, in Room D, the value of the t-test with cafe noise and SSN noise is 0.01 and 0.02, respectively.* It means in Room D, when the noise type is cafe and SSN, the STOI performance of the proposed method with two DNNs and the cIRM-based are significantly different from each other.

From Figs. 4.6 & 4.7 and Tables 4.2 - 4.5, it is clear that with the same amount of training data and DNN configurations, the separation performance of the current state-of-the-art is not consistent and robust when the SNR levels and noise types are varied. The two-stage approach, we proposed, can yield effective performance. Thanks to the DM applied to the mixture, when the RT60 is increased, the relative STOI improvements becomes more prominant at higher RT60s. Compared the masking-based techniques with the proposed two-stage approach, the experimental results demonstrate that using two DNNs in the proposed two-stage approach can further improve the separation performance.

Then, the proposed two-stage approach is evaluated with unseen RIRs. The SNR_{fw} and the SDR performance of the proposed methods and the compared methods are given in Figs. 4.8 & 4.9, respectively. The STOI performance of different methods using the IEEE and the TIMIT corpora with different noise and the unseen RIRs are shown in Table 4.6. In the experiments with the unseen RIRs, the RIRs used in the testing stage are different from those in the training stage.

Fig. 4.8 shows the SNR_{fw} performance in terms of different methods with the unseen RIRs. It can be observed that compared with the IRM and the cIRM, the proposed methods, both single DNN and two DNNs, yield better performance. When the value of SNR level is increased, the performance of SNR_{fw} is refined. Besides, it is observed from the figure that when two DNNs are trained, the values of the SNR_{fw} become higher.

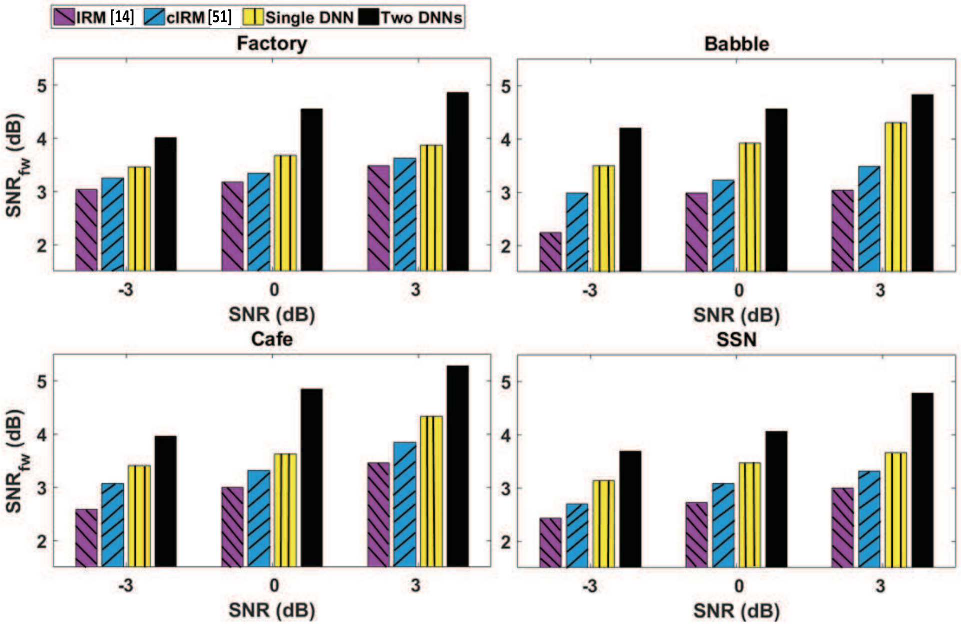

Figure 4.8. The SNR_{fw} (dB) in terms of different methods with the unseen RIRs. The X-axis is the SNR level, the Y-axis is the SNR_{fw} (dB), each result is the average value of 120 experiments. The experimental results with four different types of noise are shown.

For example, according to Fig. 4.8, when the noise type is SSN and the SNR level is 3 dB, the SNR_{fw} value of the IRM-based method is 2.99 dB and the cIRM-based method is 3.32 dB, but the proposed approach with single DNN and two DNNs achieve 3.66 dB and 4.78 dB, respectively.

Fig. 4.9 shows the SDR improvements over all types of noise with the unseen RIRs. It is observed that the proposed two-stage approach further refines the SDR performance (ΔSDR) when compared with the current state-of-the-art methods. In the situation where the RIRs are unseen, with increasing the SNR level, the improvement of the SDR becomes larger and the proposed two-stage approach provides the best performance. It is clear that by training two DNNs in the proposed two-stage approach, the value of the SDR improvement is increased significantly.

The experimental results in terms of the STOI are shown in three dif-

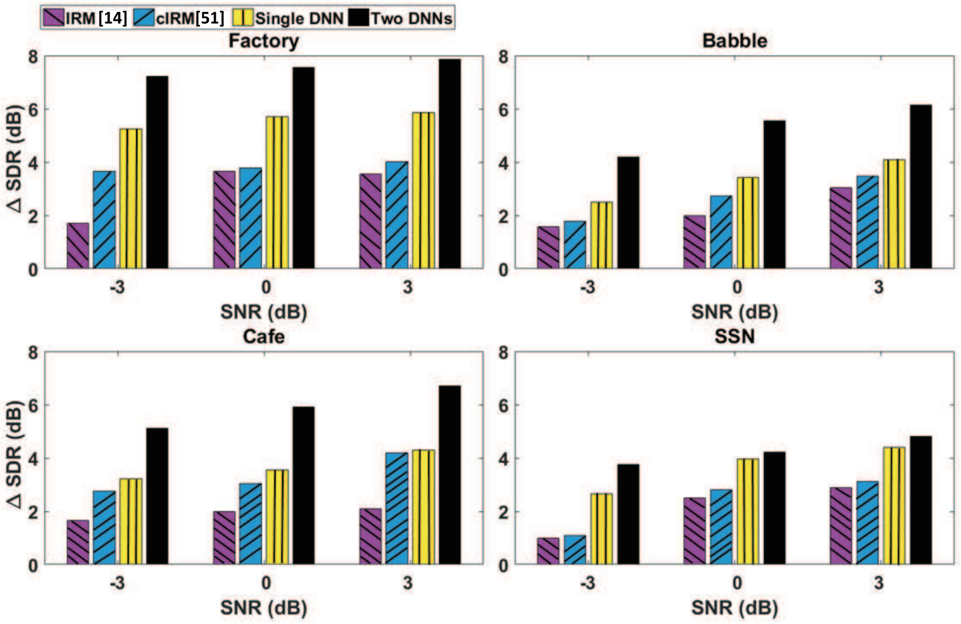

Figure 4.9. The SDR improvement (dB) in terms of different methods with the unseen RIRs. The X-axis is the SNR level, the Y-axis is the SDR improvement (dB), each result is the average value of 120 experiments. The experimental results with four different types of noise are shown.

ferent SNR levels in Table 4.6. As the value of SNR level is increased, the performance of the STOI is improved. From Table 4.6, it is clear that with the same amount of training data and DNN configurations, when the RIRs are unseen, in terms of the STOI, the separation performance of the current state-of-the-art is not consistent and robust when the SNR levels and noise types are varied. *For all types of the noise, the value of the t-test in the STOI results with the unseen RIRs between the cIRM-based method and the proposed method with single DNN and two DNNs is 0.02 and 0.0004, respectively.* It confirms that the proposed two-stage approach outperforms the current state-of-the-art methods in terms of the STOI.

Table 4.6. Separation performance comparison in terms of STOI with the unseen RIRs. Different training targets, SNR levels and RT60s with all types of noise are evaluated. Each result is the average value of 120 experiments. **BOLD** indicates the best result.

Noise Type	Factory			Babble			Cafe			SSN		
SNR Levels	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Mixture	0.46	0.48	0.50	0.47	0.49	0.52	0.49	0.51	0.54	0.50	0.53	0.55
IRM [14]	0.52	0.55	0.56	0.52	0.54	0.55	0.51	0.53	0.57	0.51	0.55	0.59
cIRM [51]	0.57	0.59	0.63	0.54	0.57	0.58	0.52	0.55	0.59	0.53	0.57	0.63
Single DNN	0.62	0.64	0.65	0.58	0.61	0.64	0.57	0.61	0.64	0.57	0.61	0.67
Two DNNs	**0.68**	**0.71**	**0.74**	**0.64**	**0.69**	**0.73**	**0.64**	**0.70**	**0.75**	**0.64**	**0.67**	**0.72**

From Figs. 4.8 & 4.9 and Table 4.6, it can be observed that the proposed two-stage approach can yield effective performance and using two DNNs in the proposed two-stage approach provides the best separation results. Using the noise and unseen RIRs, the proposed methods show better generalization ability. In the testing stage, since the RIR is unseen, compared with the seen RIRs case, the values of the corresponding SNR_{fw}, SDR and STOI are smaller.

4.4.3 Evaluations with the speech interferences

After the evaluations of the proposed two-stage approach with noise interference, the undesired speech signal is exploited as the interference to generate the convolutive mixture. The interfering speech signal is chosen from the above mentioned corpora and both male and female speakers are used. The SNR_{fw} and the SDR performance of the proposed methods and the comparison groups are given in Figs. 4.10 & 4.11, respectively. The STOI performance of different methods are shown in Table 4.7.

For the SNR_{fw}, shown in Fig. 4.10, the proposed two DNN-based method further improves the performance relative to the separated desired speech signal. The largest SNR_{fw} gains in all room environments are achieved by the proposed two DNN-based method. For example, at 3 dB SNR level, from Rooms A to D, the proposed method with two DNNs gives 16.1%, 21.8%, 22.3% and 13.7% more gain, respectively.

Besides, according to Fig. 4.10, it confirms that the higher SNR level helps the two-stage approach to better separate the desired speech signal from the mixture with speech interference. Compared the performance with different SNR levels in terms of the SNR_{fw}, when the SNR levels increases (from -3 dB to 3 dB), the separation performance is improved, which is the same as the situations with noise interferences. For different RT60s, when the RT60 increases, e.g. Room A and Room D, the value of the SNR_{fw} is decreased.

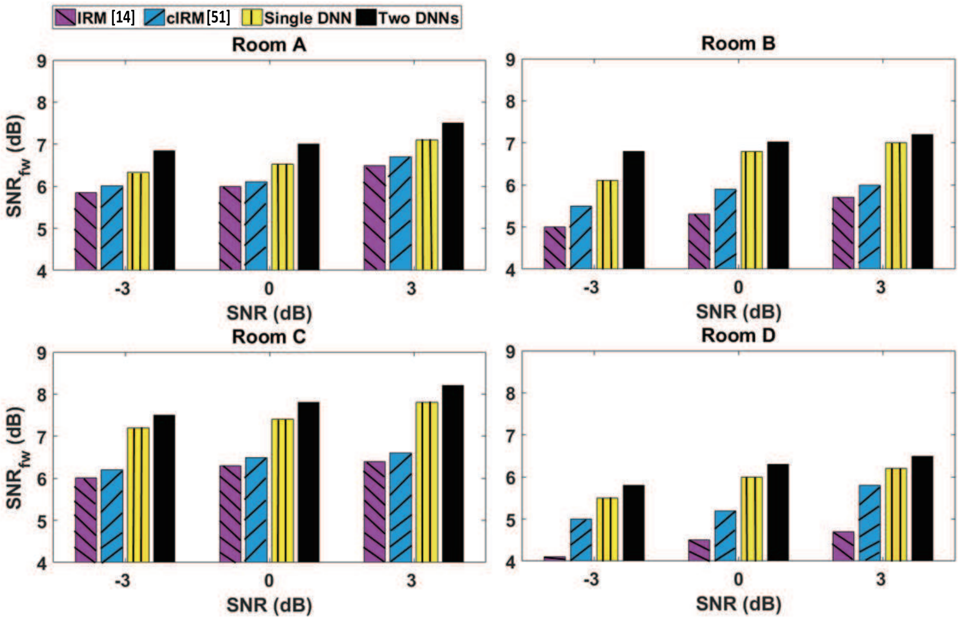

Figure 4.10. The SNR$_{fw}$ (dB) in terms of different methods with various rooms i.e. different RT60s. The X-axis is the SNR level, the Y-axis is the SNR$_{fw}$ (dB), each result is the average value of 120 experiments. The interference is the undesired speech signal, respectively.

Fig. 4.11 displays the SDR improvements over all room environments. It is observed that the proposed two-stage approach significantly improves the SDR performance (ΔSDR), especially in the highly reverberant room environments such as Room C and Room D. With increasing the SNR level, the improvement of the SDR becomes smaller, but the proposed two DNN-based method still provides better results. In Room C, with 0.68 s RT60, compared with the cIRM, the proposed method with single DNN has 1.01 dB, 1.71 dB and 0.49 dB more improvements and the proposed method with two DNNs has 1.81 dB, 3.27 dB and 3.67 dB from -3 dB to 3 dB SNR levels, respectively.

From Table 4.7, it is clear that the two DNN-based method always gives the best performance in the case where the interference is a speech signal.

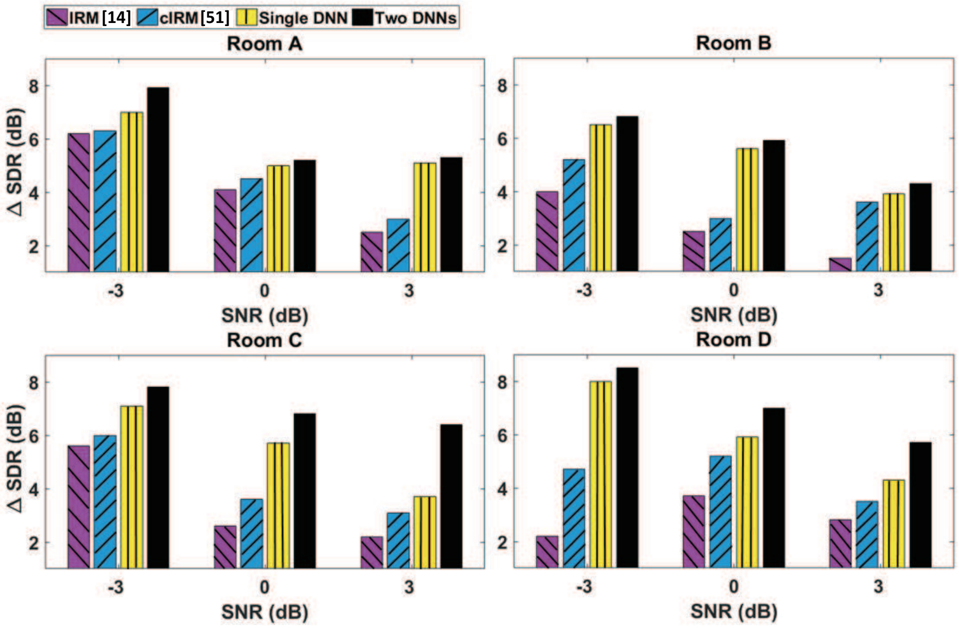

Figure 4.11. The SDR improvement (dB) in terms of different methods with various rooms i.e. different RT60s. The X-axis is the SNR level, the Y-axis is the ΔSDR (dB), the improvements of the SDR. Each result is the average value of 120 experiments. The interference is the undesired speech signal, respectively.

For example, in Room D, the proposed method with two DNNs achieves 13.1%, 8.7% and 12.5% STOI improvements over the proposed method with single DNN (integrated training objective) at -3, 0 and 3 dB SNR levels, respectively. The two DNN-based method provides around 13.9% more STOI improvement in all scenarios. *When the undesired speech signal is the interference, the value of the t-test in the STOI results with the seen RIRs between the cIRM-based method and the proposed method with two DNNs is 0.008.* It proves that the proposed method with two DNNs yields better separation performance in terms of the STOI than the current state-of-the-art methods, e.g. cIRM-based method.

The interfering speech signal is chosen from the IEEE and the TIMIT corpora and both male and female speakers are used. The SNR_{fw} and the SDR performance of the proposed methods and the comparison groups are given in Figs. 4.12 & 4.13, respectively. The STOI performance of different methods using the above mentioned corpora with different undesired speech signal and the unseen RIRs are shown in Table 4.8.

Table 4.7. Separation performance comparison in terms of STOI with different training targets, SNR levels and RT60s. The interference in the experiments is ***the undesired speech signal***. Each result is the average value of 120 experiments. **BOLD** indicates the best result.

Speech	Room A (0.32 s)			Room B (0.47 s)			Room C (0.68 s)			Room D (0.89 s)		
Interference	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Mixture	0.58	0.63	0.67	0.54	0.59	0.63	0.58	0.64	0.66	0.48	0.50	0.51
IRM [14]	0.76	0.78	0.79	0.72	0.73	0.75	0.78	0.79	0.81	0.60	0.61	0.62
cIRM [51]	0.77	0.78	0.80	0.74	0.75	0.76	0.79	0.80	0.81	0.63	0.64	0.64
Single DNN	0.78	0.80	0.82	0.76	0.80	0.81	0.79	0.81	0.83	0.71	0.73	0.75
Two DNNs	**0.80**	**0.82**	**0.84**	**0.79**	**0.81**	**0.82**	**0.81**	**0.82**	**0.84**	**0.74**	**0.75**	**0.78**

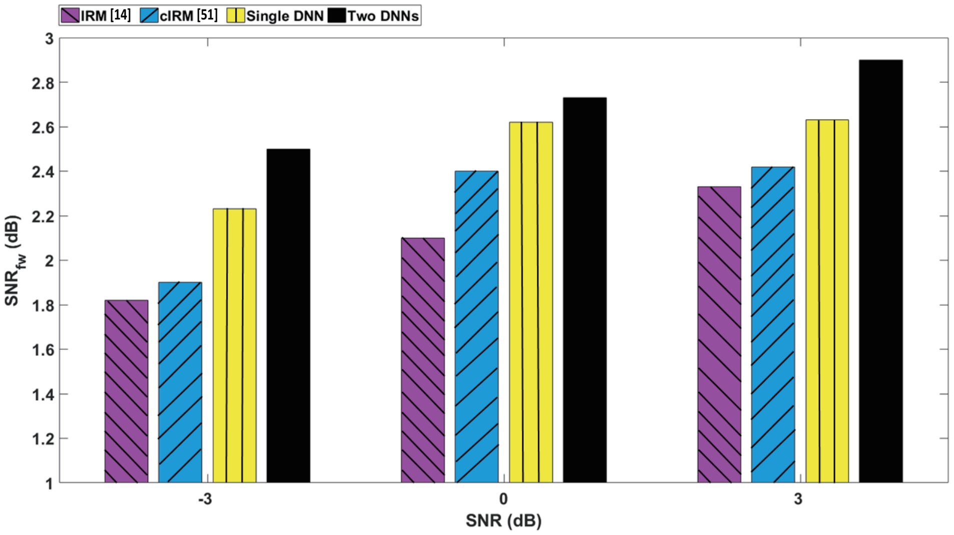

Figure 4.12. The SNR$_{fw}$ (dB) in terms of different methods with the unseen RIRs. The X-axis is the SNR level, the Y-axis is the SNR$_{fw}$ (dB), each result is the average value of 120 experiments. The interference is the undesired speech signal, respectively.

For the SNR$_{fw}$, shown in Fig. 4.12, the proposed two-stage approach provides the largest performance improvements with the unseen RIRs scenarios. The largest SNR$_{fw}$ gains in all SNR levels are achieved by the proposed two-stage approach with separate training targets. According to Figure 11, the proposed two-stage approach with integrate training target can achieve higher value of the SNR$_{fw}$ and by training two DNNs in the proposed method, the separation performance is further improved.

Fig. 4.13 shows the SDR improvements (ΔSDR) over all SNR levels with the unseen RIRs. It is observed that the proposed two-stage approach significantly improves the SDR performance, especially with higher SNR levels. With increasing the SNR level, the improvement of the SDR becomes larger and the proposed two DNN-based method achieves better separation results. For instance, when the SNR level is 3 dB, the value of ΔSDR of the proposed method with separate training objectives is 5.05 dB, while the value of the cIRM-based and the IRM-based method is 3.06 dB and 2.41 dB,

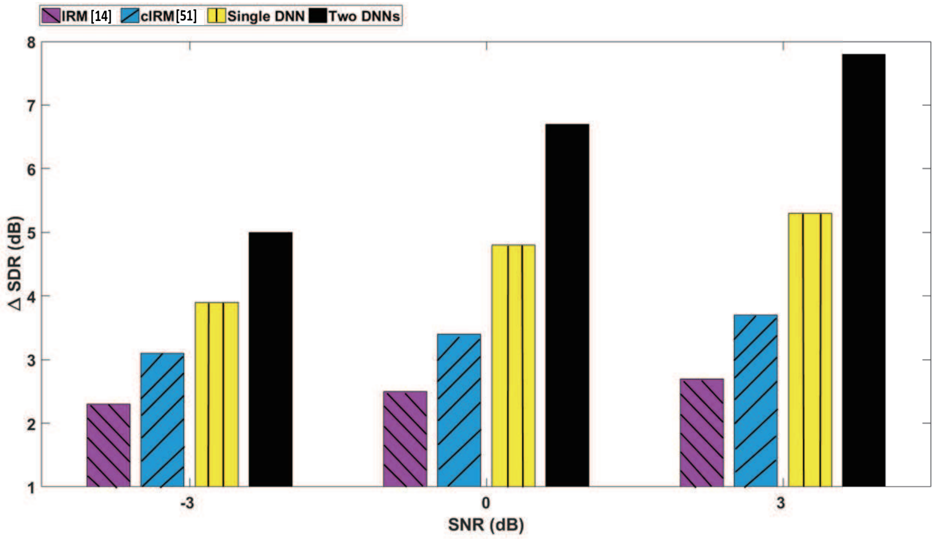

Figure 4.13. The SDR improvement (dB) in terms of different methods with the unseen RIRs. The X-axis is the SNR level, the Y-axis is the ΔSDR (dB), the improvements of the SDR. Each result is the average value of 120 experiments. The interference is the undesired speech signal, respectively.

respectively. It is clear that by training two DNNs in the proposed two-stage approach, the separation performance is increased significantly. In contrast to the evaluations with the seen RIRs, when the RIRs are unseen and the RT60 increases, the value of the SDR improvement increases, which are the same as the situations with noise interferences.

When the interference is the undesired speech signal, Table 4.8, it is clear to observe that in terms of the STOI, the proposed two-stage approach outperforms current state-of-the-art. For example, compared with the cIRM, the proposed method with single DNN has 0.06, 0.08 and 0.07 improvements and the proposed method with two DNNs has 0.11, 0.11 and 0.1 improvements from -3 dB to 3 dB SNR levels, respectively. *When the undesired speech signal is the interference, the value of the t-test in the STOI results between the cIRM-based method and the proposed method with two DNNs is 0.01.* Hence, by using two DNNs in the proposed method, the value of STOI is the highest over all of the SNR levels.

Table 4.8. Separation performance comparison in terms of STOI with different training targets, SNR levels and the unseen RIRs. The interference in the experiments is ***the undesired speech signal***. Each result is the average value of 120 experiments. **BOLD** indicates the best result.

Speech	STOI		
Interference	-3 dB	0 dB	3 dB
Mixture	0.52	0.57	0.59
IRM [14]	0.56	0.59	0.64
cIRM [51]	0.59	0.61	0.66
Single DNN	0.65	0.69	0.73
Two DNNs	**0.70**	**0.72**	**0.76**

Since two system structures of the proposed two-stage approach are exploited in this work, their processing time is different. In order to evaluate their processing time, all of the DNN-based methods are executed ten times and their processing time is averaged. The evaluation results are shown in Table 4.9.

Table 4.9. Averaged processing time of the DNN-based methods with different training targets. The time of training stage and testing stage are shown in seconds.

Training Target	Processing Time (s)	
in DNN-based Method	Training Stage	Testing Stage
IRM [14]	8,398.8	37.4
cIRM [51]	8,655.4	43.1
IEM	8,443.4	39.8
DM & IRM	16,651.9	48.5

The codes of the IRM, cIRM and the proposed methods were written in MATLAB (R2015a version) without any optimization. The experiments were implemented on a desktop with an Intel i5 CPU with 3.5 GHz and 16 GB of memory without parallel processing. In the training and testing stages, no GPU was used.

It is observed from Table 4.9 that in the training stage, the processing time of the proposed method with single training target (integrated objective) is half of the one with two training targets (separate objectives). Because in the second method, two DNNs are trained and these DNNs have the same architectures as the DNN in the first proposed method. While compared with the training stage, in the testing stage, the difference of the processing time with these methods can be ignored. The IRM-based method and the proposed IEM almost have the same processing time. Moreover, because the Y-shaped DNN was used in the cIRM-based method, its processing time is slightly higher than the IRM- and the IEM-based approaches. In the testing stage, all of these methods have a relative lower processing time.

Hence, the proposed two DNN-based method needs longer processing time and the computational cost is almost double than the single training target based method.

4.5 Chapter Summary

In summary, the proposed two-stage approach outperforms state-of-the-art IRM- and the cIRM-based methods, particularly in reverberant room environments. When the RIRs are seen, the noise and undesired speech signal are used as the interferences in the mixture, all the experimental results further confirm that our proposed two-stage approach is effective in separating mixtures at various SNR levels and with different room environments. When the RIRs are unseen, the generalization ability of the proposed method is evaluated, the results confirm that the proposed method can better separate the desired speech signal from mixture than the IRM- and cIRM-based methods. There are two possible reasons that the proposed method has better generalization ability: (1) The compression and recovery modules are conducive for training the DNNs and thus leading to better prediction of the DM from the mixtures. (2) The use of DM can mitigate the adverse effect of acoustic reflections on the estimation of the IRM_{rev} and $cIRM$ for separating target speech from the mixture. As a result, the proposed method has better ability in adapting to unseen RIRs and leading to improved performance in such scenarios.

In addition, using the proposed two DNN-based method, the mixture can be better separated than just utilizing the IEM as integrated training target in the single DNN. From the results, it can be seen that the cIRM had worse performance than IRM in some cases. To estimate the real and imaginary part of the cIRM jointly, the Y-shaped DNN was used. In this architecture, the weights of the hidden layers are shared by the real and imaginary parts of the cIRM and only two sub-output layers are used to distinguish the estimations of real and imaginary components of the cIRM. Hence, compared with the IRM, the cIRM-based DNN is more difficult to train, in order to provide balance for both the real and imaginary part. This can lead to degradation in separation performance.

It is worth noting that although the RT60 of Room C (RT60 = 680 ms) is higher than Room B (RT60 = 470 ms), the separation performance for Room C is better than that for Room B. This is mainly due to the difference in the Direct to Reverberant Ratio (DRR) where the DRR from Room C is higher than that for Room B. And in the proposed method with different training targets, when the DM and the IRM are trained individually, the computational cost is increased almost two times. Therefore, there is a trade-off between the computational cost and the separation performance. If two-DNNs are trained in the proposed two-stage approach, the separation performance is further refined, but more computational cost and storage space are required.

But in the proposed method of this chapter, the noisy phase information is utilized to generate desired speech signal from the mixture. In the next chapter, the phase information of the clean speech is utilized by operating in the complex domain. Moveover, the LSTM RNN is introduced to better use the temporal information and obtain the accurate trained neural network model.

Chapter 5

MONAURAL SOURCE SEPARATION IN COMPLEX DOMAIN WITH LONG SHORT-TERM MEMORY NEURAL NETWORK

5.1 Chapter Introduction

In this chapter, improvements to the monaural source separation are made both in terms of the LSTM RNN and phase information. In the LSTM RNN, because the LSTM unit is introduced in the RNN, which aids the utilizing of the temporal information [103]. Moreover, to further improve the separation performance, the phase information of clean speech signal is estimated via neural network model. In terms of the phase information, the complex IRM (cIRM) is firstly used; however, the cIRM needs further operation to recover clean speech signal which may cause incorrect estimation. To address this issue, the cSA is proposed to directly estimate the real and imaginary components.

This chapter focuses on the third objective of this thesis, which relate to the phase information and LSTM RNN-based method for monaural source separation published in IEEE Journal of Selected Topics in Signal Processing [104].

5.2 Complex Signal Approximation

In order to achieve more robust prediction of clean speech signal, the complex signal approximation is employed to utilize the phase information of the clean speech signal, and the outputs are the real and imaginary components to recovery the desired speech signal.

5.2.1 Signal approximation

In masking-based DNN approach, the ideal time-frequency (T-F) mask is applied as the training target of the neural network models. The T-F mask predicted by the trained model is applied to the mixture to reconstruct the desired speech signal. The cost function of masking-based DNN approach is expressed as:

$$Loss_{\text{masking}} = \sum_t \sum_f (|\hat{M}(t,f)| - |M(t,f)|)^2 \tag{5.2.1}$$

where $|\hat{M}(t,f)|$ is the estimated T-F mask and $|M(t,f)|$ is the training target, which is the ideal T-F mask.

In the mapping-based approach, the training target is the spectrum of the clean speech signal. The cost function of mapping-based DNN approach is expressed as:

$$Loss_{\text{mapping}} = \sum_t \sum_f (|\hat{S}_1(t,f)| - |S(t,f)|)^2 \tag{5.2.2}$$

where $|\hat{S}_1(t,f)|$ is the estimated spectrogram of the desired speech signal and $|S(t,f)|$ is the training target, which is the spectrogram of the clean speech signal.

The SA-based approach combines the mapping- and masking-based approaches. The training target in the oSA-based method is the spectral magnitude of clean speech, which is equivalent to the mapping-based approach. The cost function in the oSA-based method can be written as:

$$J_{oSA} = \sum_{t} \sum_{f} (|Y(t,f)\hat{M}_{oSA}(t,f)| - |S(t,f)|)^2 \tag{5.2.3}$$

where the predicted T-F mask in the oSA-based method is $\hat{M}_{oSA}(t,f)$, which is used to obtain the estimated spectrum $\hat{S}(t,f)$. The T-F mask is predicted in the oSA-based neural network to minimize the discrepancy between the magnitude spectrum of mixture and that of the clean speech signal, which is similar to masking-based approaches. Hence, using the magnitude spectrum of the clean signal as the training target can increase the accuracy of the estimated T-F mask and improve separation performance.

However, the oSA-based method has the same problem as the IRM-based method where the phase information of the target signal is not used when reconstructing the desired signal.

5.2.2 Complex signal approximation

Inspired by the cIRM, the cSA-based method is proposed, which replaces the IRM by cIRM in the training process to estimate both real and imaginary components of the clean speech signal. One could use the magnitude and phase information, instead of the real and imaginary components, as training targets, are exploited. However, our empirical tests show that using the real and imaginary components as training targets offers better separation performance. Hence, in the cSA-based method, the real and imaginary components of the desired clean speech signal are used as training targets. In the cSA-based method, the estimated spectrum of the clean signal is obtained by applying the predicted complex T-F mask, defined as $\hat{M}_{cSA}$. Using the complex mask, the estimated spectrum can be obtained:

$$\hat{S}_r(t,f) + j\hat{S}_c(t,f) = (\hat{M}_{cSA_r}(t,f) + j\hat{M}_{cSA_c}(t,f)) \times (Y_r(t,f) + jY_c(t,f)) \tag{5.2.4}$$

Then, the estimated spectrum is expressed as:

$$\begin{aligned}\hat{S}_r(t,f) + j\hat{S}_c(t,f) = {} & \hat{M}_{cSA_r}(t,f)Y_r(t,f) + j\hat{M}_{cSA_r}(t,f)Y_c(t,f) \\ & + j\hat{M}_{cSA_c}(t,f)Y_r(t,f) - \hat{M}_{cSA_c}(t,f)Y_c(t,f)\end{aligned} \tag{5.2.5}$$

The real component of the estimated clean spectrum in the cSA is expressed as:

$$\hat{S}_r(t,f) = \hat{M}_{cSA_r}(t,f)Y_r(t,f) - \hat{M}_{cSA_c}(t,f)Y_c(t,f) \tag{5.2.6}$$

The imaginary component of the estimated clean spectrum is calculated as:

$$\hat{S}_c(t,f) = \hat{M}_{cSA_r}(t,f)Y_c(t,f) + \hat{M}_{cSA_c}(t,f)Y_r(t,f) \tag{5.2.7}$$

In the proposed cSA-based LSTM RNN method, when the Y-shaped neural network model is used, the shared weights in the hidden layers cannot be fully used for both components, and this may have negative impacts on the estimations, and thus the separation performance. Our empirical tests show that using two networks performs better than stacking the two components in one network. In the cSA-based method, the real and imaginary components are estimated separately and two neural network models are trained with real and imaginary components of the cIRM. The cost functions can be expressed in the complex domain with the real and imaginary components. According to (5.2.6) and (5.2.7), the expanded cost functions of the cSA-based method are:

$$J_1 = \sum_t \sum_f \Big[\Big(\hat{M}_{cSA_r}(t,f)Y_r(t,f)$$

$$-\hat{M}_{cSA_c}(t,f)Y_c(t,f)\Big) - S_r(t,f)\Big]^2 \tag{5.2.8}$$

$$J_2 = \sum_t \sum_f \Big[\Big(\hat{M}_{cSA_r}(t,f)Y_c(t,f)+$$

$$\hat{M}_{cSA_c}(t,f)Y_r(t,f)\Big) - S_c(t,f)\Big]^2 \tag{5.2.9}$$

Hence, by using cSA-based method, the phase information of clean speech signal can be utilized by estimating the imaginary component of the speech signal in complex domain. In DNN, the temporal information can be utilized via context window, however, the ability of using temporal information depends on the length of context window. In practical, the length of context window is difficult to select [105]. The recurrent units is introduced to employ the temporal information.

5.3 Complex Signal Approximation with LSTM RNN

To utilize the temporal information, the recurrent unit is introduced. And there are to common-used architectures: 1). RNN and 2). LSTM RNN, they are described in the following subsections.

5.3.1 Recurrent neural network

In the monaural source separation problem, which is solved via neural networks, the separation performance can be improved by utilizing the temporal information of the speech signals in the training stage of networks. Commonly, the temporal information is exploited in two ways: concatenating neighbouring features and using RNNs [76]. In the concatenating features method, a larger window size can utilize more temporal information with the trade off being computational and memory resources. Therefore, an appropriate window size is required. The RNNs have a recurrent architecture, which is a powerful model for temporal information.

In the vanilla RNN [81], the hidden state of the l-th layer at time t is:

$$\mathbf{h}_t^l = f_{\mathrm{h}}(\mathbf{h}_t^{l-1}, \mathbf{h}_{t-1}^l) = g_l(\mathbf{R}^l \mathbf{h}_{t-1}^l + \mathbf{W}^l \mathbf{h}_t^{l-1}) \tag{5.3.1}$$

In the first layer, where $l = 1$, the activation $\mathbf{h}_t^1$ is calculated by $\mathbf{h}_t^0 = \mathbf{x}_t$. In the RNN, the activation function is selected as a rectified linear unit (ReLU) to avoid gradient vanishing and reduce the computational cost. The ReLU function is expressed as:

$$g(\mathbf{x}) = max(\mathbf{0}, \mathbf{x}) \tag{5.3.2}$$

In (5.3.1), each hidden state is computed with current state and the previous state, it can be observed that the temporal information is better utilized. However, in RNN, there existing exploding gradients and vanishing gradient problems. To address there issues, the LSTM RNN is proposed [106].

5.3.2 Long short-term memory recurrent neural network

Different from the vanilla DNN, which can only use context window to capture temporal dependencies, the LSTM RNN stores the temporal information in the cell, therefore, the long temporal dependencies can be utilized. In the DNN-based method, the neural network model is trained with backward propagation algorithm [50] but in the LSTM RNN-based method, the backward propagation through time algorithm is exploited [107].

The structure of the LSTM block is shown in Figure 3. Assume the current time is m, $\sigma(\cdot)$ represents the sigmoid function, and $tanh(\cdot)$ represents the hyperbolic tangent function. The x_m, h_m and c_m are defined as the input, hidden state and cell memory at time m, respectively. The weights in the input gate, forget gate, output gate and cell are defined as W_i, W_f, W_o and W_c, the corresponding bias are b_i, b_f, b_o and b_c, respectively. Each gate is constructed by the weights, bias and activation function.

For the input gate, the output i_m is:

$$i_m = \sigma(W_i[x_m, h_{m-1}] + b_i) \tag{5.3.3}$$

For the forget gate, the output f_m is:

$$f_m = \sigma(W_f[x_m, h_{m-1}] + b_f) \tag{5.3.4}$$

For the output gate, the output o_m is:

$$o_m = \sigma(W_o[x_m, h_{m-1}] + b_o) \tag{5.3.5}$$

For the memory cell, the current input's cell state is:

$$\tilde{c}_m = tanh(W_c[x_m, h_{m-1}] + b_c) \tag{5.3.6}$$

and the current cell state c_m is:

$$c_m = f_m \times c_{m-1} + i_m \times \tilde{c}_m \tag{5.3.7}$$

The final hidden state of the LSTM block is expressed as:

$$h_m = o_m \times tanh(c_m) \tag{5.3.8}$$

In the LSTM RNN block, from (5.3.3) to (5.3.8), it can be known that the state is determined by those cells and gates. Hence, by storing the state in the cell, the gradient vanishing problem is solved.

5.3.3 Complex signal approximation with LSTM RNN

After the hidden states are obtained from the LSTM blocks, the output layer is added to generate the output of the LSTM RNN. The activation function

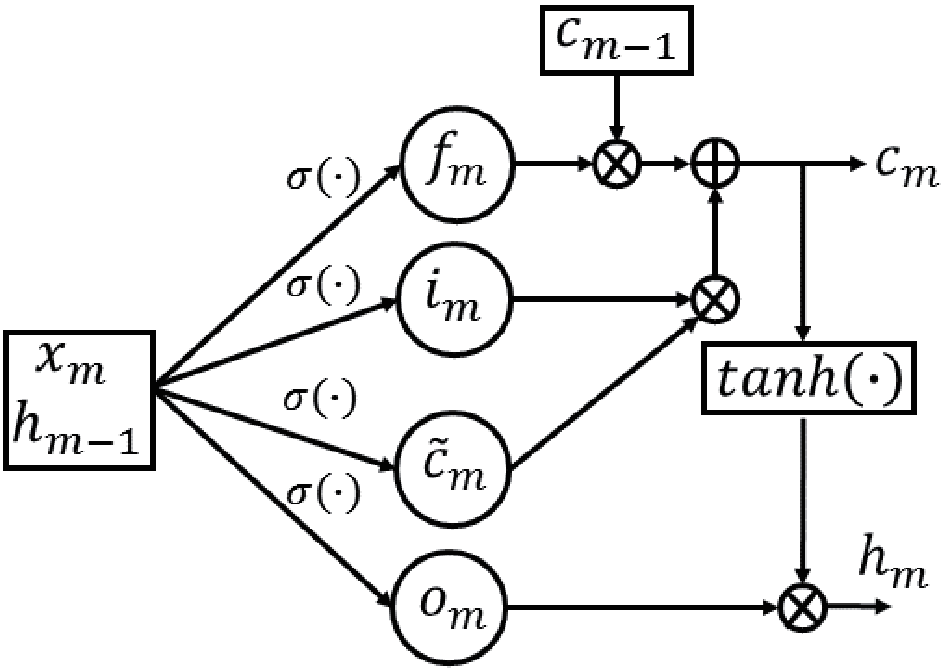

Figure 5.1. Diagram of only one LSTM block, which contains three gates and one memory cell. The block is exploited in the proposed LSTM RNN-based methods.

of the output layer is selected as a linear function. For complex domain monaural source separation, the estimated phase information of clean speech signal is used to recover the desired speech signal. Then, by introducing the LSTM RNN, the temporal information is utilized. Besides, if the training target of the LSTM RNN is the cIRM, the neural network is Y-shape and two sub-output layers are added as shown in Fig. 5.2. In the cSA-based LSTM RNN method, two LSTM RNNs are exploited to predict the real and imaginary components in parallel and both LSTM RNNs have the same configuration.

In the proposed cSA-based LSTM RNN method, inspired by [50, 95] and vanilla DNN methods, the feature combination is given to the input layer to increase the efficiency of the networks and system. The amplitude modulation spectrogram (AMS) [96], relative spectral transform and

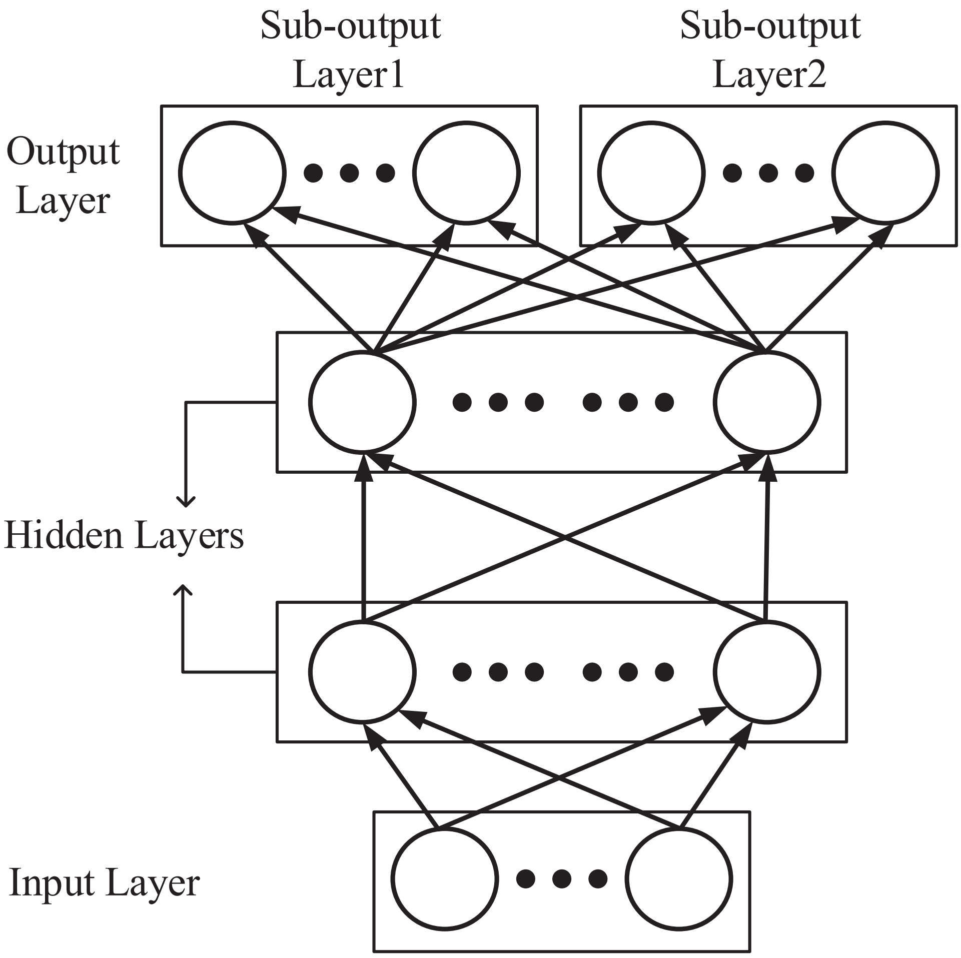

Figure 5.2. The Y-shaped neural network architecture, which has two sub-output layers. The sub-output layer 1 and the sub-output layer 2 yield the real and imaginary components of the estimation, respectively.

perceptual linear prediction (RASTA-PLP) [97], mel-frequency cepstral coefficients (MFCC), cochleagram response and their deltas are extracted by a 64-channel gammatone filterbank to obtain the compound feature [98]. Furthermore, in the oSA- and the cSA-based methods, the spectra of the mixture and the clean signal are given to calculate the spectrograms of the predicted clean signal and the training objective, respectively. The flow diagram of the proposed cSA-based LSTM RNN method is shown in Fig. 5.3.

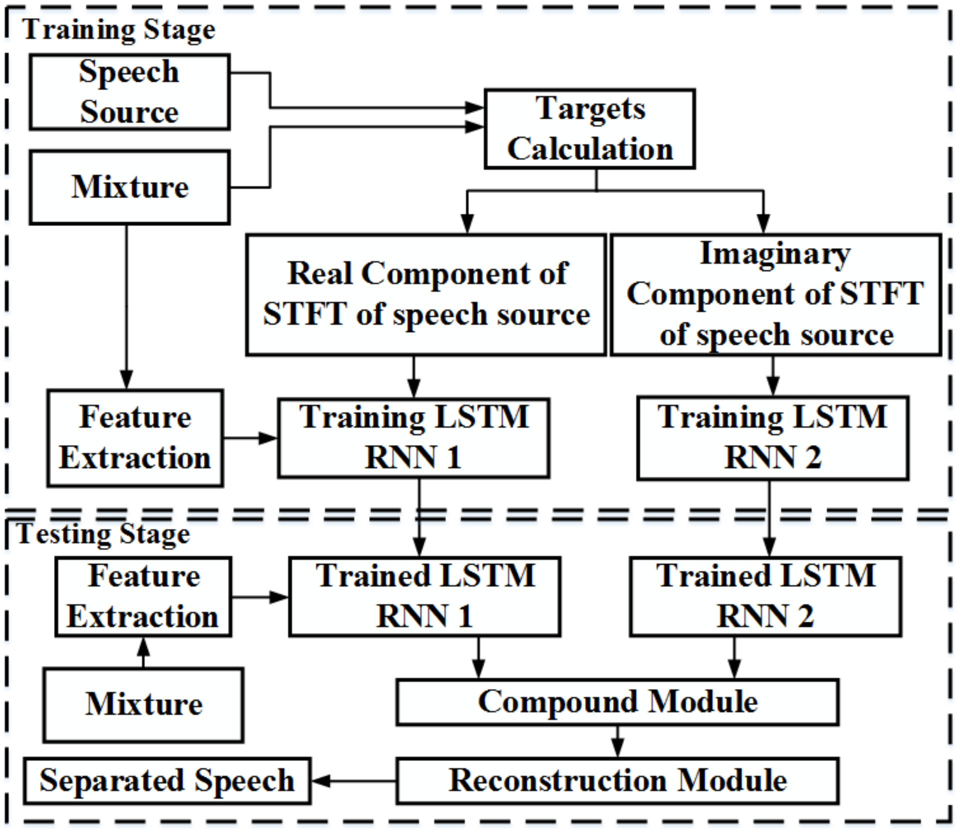

Figure 5.3. The block diagram of the proposed complex signal approximation (cSA)-based LSTM RNN method. Two LSTM RNNs are trained with the separate training targets, e.g. the real and the imaginary components of the STFT of clean speech signal.

In the training stage, by using the targets calculation module, the STFTs of speech source and mixture are obtained. Then, the real and imaginary components of STFT of the speech source are used as the training targets for LSTM RNN 1 and LSTM RNN 2, respectively. The outputs of the LSTM RNN models are obtained by multiplying the estimated T-F mask with the STFT of the mixture. After each iteration, the estimated T-F mask is trained to minimize the discrepancy between the spectrum of the clean speech signal and that of the estimated source signal.

In the testing stage, the trained LSTM RNNs can output the real and imaginary components of the estimated speech signal when the feature combination of the mixture is used as input. Then, the STFT of the separated speech is obtained in the compound module and the separated speech signal is reconstructed in the reconstruction module.

Compared with the oSA-based DNN method, the proposed cSA-based LSTM RNN method has two advantages:

(1) In traditional oSA-based DNN method, the noisy phase information is used to synthesise the desired speech signal. However, in the proposed cSA-based LSTM RNN method, both clean magnitude and phase information are estimated.

(2) The LSTM blocks are introduced with the RNN, the temporal information can be better utilized and the trained LSTM RNN models have better generalization ability.

5.4 Simulations

In this section, the cIRM- and oSA-based method are evaluated with the vanilla DNN and the LSTM RNN to show the advantage of LSTM RNN over the vanilla DNN. Then, the results of the proposed cSA-based LSTM RNN method are shown. Firstly, the interference is selected as the noise, in both seen and unseen scenarios. The unseen interferences mean that the interferences in the training data are not used to generate testing data. Then, the interference is chosen as the undesired speech signal which is unseen in the training stage. Therefore, the generalization ability of these methods can be evaluated.

5.4.1 Dataset selection and configurations

The speech sources are selected randomly from the IEEE and the TIMIT corpora [85, 99]. The IEEE corpus has 720 clean utterances spoken by a single male speaker and the TIMIT database has 6300 utterances, 10 utterances spoken by each of 630 speakers. Therefore, using both the IEEE and the TIMIT corpora can demonstrate that the proposed method is speaker-independent. We randomly select 1000, 100 and 200 clean utterances from the IEEE and the TIMIT corpora to generate the training, development and testing datasets.

The interferences are categorized into two aspects, the noise interference and the undesired speech interference. In the seen noise interference cases, these clean speech utterances are mixed with five different noise types at three different SNR levels (-3 dB, 0 dB and 3 dB). These five noise scenes are named as *factory*, *babble*, *cafe*, *f16* and *tank*. The names of these noise signals indicate their recording situations. The above mentioned noise signals are selected from the NOISEX database [91]. Each noise sequence is four minutes long, which is truncated randomly from the first two minutes to match the lengths of the speech signals to generate the training mixtures. The last two minutes are used to generate the development and testing mixtures. In this case, although the noise interference in the testing dataset is unseen, the noise type is known.

In the unseen noise interference cases, 50 different noise signals are used to generate the training, development and testing datasets and 50 noise signals are only used to generate the testing data. These non speech sounds contain many different types of noise, e.g. animal sounds, tooth brushing sounds and machine noise [90]. Finally, the number of mixtures in training, development and testing data is 12,000, 1200 and 2400, respectively. The training speech duration is around 10 hours and 100 types of different noise signals are used in the unseen cases.

In the evaluation studies where the interference is undesired speech signal, in both training and testing stages, the target speech signals are randomly selected from the TIMIT dataset. Then, interfering speech signals are randomly selected from the remaining signals in the dataset to ensure the speakers of the target speech and the interfering speech signals are different. At the testing stage, the desired speech signals are unseen in the training stage, but the interfering speech signals are seen in the training stage. Therefore, the trained neural network is able to differentiate the target and undesirable speech signals. Similarly, the SNR levels are -3 dB, 0 dB and 3 dB and the number of mixtures in training, development and testing data is 12,000, 1200 and 2400, respectively.

Both the DNNs of the comparison group and the LSTM RNN have three hidden layers and each hidden layer has 512 units. The dimension for the input layer is 1722 $(246\times(3\times2+1))$, the number of neighbouring for each side of central frame is 3 and the input size is 246. In terms of the DNN, according to [50], the activation function for each hidden unit is selected as the rectified linear unit (ReLU) to avoid the gradient vanishing problem and the output layer has linear units [51]. In the LSTM RNN, the activation function for each hidden unit is selected as the sigmoid and the output layer has linear units. When the training target is the cIRM, the corresponding neural network outputs the estimates of real and imaginary components of the predicted cIRM. When the training target is the clean spectrum of the desired speech signal, two LSTM RNNs are trained separately. The DNN and the LSTM RNN are trained by using the RMSprop algorithm [108] with a learning rate of 0.001. The number of epochs is 100 and the batch size is 1024. Auto-regressive moving average (ARMA) filtering is applied to reduce the interference from the background noise, as in [102].

In the experiments, the proposed cIRM- and cSA-based LSTM RNN methods are compared with DNN-based approaches: the cIRM [50] and the oSA estimation [55]. In the oSA-based method, the T-F mask is an IRM, which is estimated by minimizing the discrepancy between the estimated spectrum and the spectrum of the target speech signal. In oSA-based DNN and LSTM RNN methods, the target signal is reconstructed without using the phase information of the clean speech signal, meanwhile, the cIRM- and the cSA-based methods utilize both the amplitude and phase information from the clean signal. The proposed methods are shown in *italics*. The separation performance is evaluated with three measurements. The short-time objective intelligibility (STOI) [65], the perceptual evaluation of speech quality (PESQ) [64] and the SDR [88]. The values of the STOI are in the range of [0, 1] and the PESQ are in the range of [-0.5, 4.5]. The STOI and the PESQ indicate the intelligibility scores and human speech quality scores, respectively. The SDR is exploited to evaluate the overall separation performance. The SDR value of the separated speech signal and the SDR value of the unprocessed speech mixture are used to calculate the improvement of the SDR.

5.4.2 Evaluation with noise interference

In this section, the proposed cSA LSTM RNN-based method is compared with the traditional cIRM DNN-based method in [50] and the oSA DNN-based method in [13]. Then, the cIRM LSTM RNN- and the oSA LSTM RNN-based methods are used as comparion groups.

Seen noise interferences

The separation results based on the STOI are shown in Tables 5.1, 5.2 and 5.3. The results based on PESQ are shown in Tables 5.4, 5.5 and 5.6. Each experimental result in Tables 5.1 - 5.6 is the average value over 200 testing mixtures. In total, 43,200 tests are performed. The baseline is calculated by using the unprocessed mixture and the clean speech signal.

It can be observed in Tables 5.1 - 5.6 that the performance of LSTM RNN-based methods is better than the DNN-based methods. For example, according to Table 5.1, when the noise type is factory and the SNR level is -3 dB, the STOI value of the cIRM DNN-based method is 68.31 % and the value of cIRM LSTM RNN-based method is 70.59 %. This is because the memory component in the LSTM RNN can better exploit the temporal information. In addition, the phase information is also beneficial and cSA LSTM RNN-based method outperforms all other methods. For instance, according to Table 5.5, when the noise type is cafe and the SNR level is 0 dB, the PESQ value of the oSA LSTM RNN-based method is 2.32 and the value of cSA LSTM RNN-based method is 2.49. Besides, both values of the STOI and PESQ are increased when the SNR level changes from -3 dB to 3 dB.

Table 5.1. Separation performance comparison in terms of STOI with different training targets, noises and neural network architectures, the SNR of these mixtures is -3 dB. Each result is the average value of 200 experiments. *Italic* shows the proposed methods. **BOLD** indicates the best result.

STOI	Unprocessed	cIRM-DNN [50]	*cIRM-LSTM*	oSA-DNN [13]	*oSA-LSTM*	*cSA-LSTM*
Factory	60.35%	68.31%	70.59%	70.21%	72.42%	**73.57%**
Babble	57.04%	69.22%	70.00%	68.33%	74.12%	**76.70%**
Cafe	58.07%	65.45%	68.62%	66.11%	69.03%	**75.44%**
F16	62.54%	71.11%	72.58%	72.02%	74.17%	**75.20%**
Tank	70.93%	75.48%	79.04%	76.11%	85.35%	**86.77%**
Averaged	61.79%	69.91%	72.17%	70.56%	75.01%	**77.54%**

Table 5.2. Separation performance comparison in terms of STOI with different training targets, noises and neural network architectures, the SNR of these mixtures is 0 dB. Each result is the average value of 200 experiments. *Italic* shows the proposed methods. **BOLD** indicates the best result.

STOI	Unprocessed	cIRM-DNN [50]	*cIRM-LSTM*	oSA-DNN [13]	*oSA-LSTM*	*cSA-LSTM*
Factory	67.42%	74.20%	77.92%	76.33%	78.92%	**79.59%**
Babble	64.22%	73.87%	76.81%	72.91%	78.99%	**79.47%**
Cafe	63.21%	70.36%	75.38%	71.38%	75.44%	**77.61%**
F16	65.31%	74.20%	77.26%	74.87%	79.77%	**80.13%**
Tank	75.34%	80.92%	83.75%	81.25%	87.51%	**88.03%**
Averaged	67.10%	74.74%	78.22%	75.35%	80.12%	**80.96%**

Table 5.3. Separation performance comparison in terms of STOI with different training targets, noises and neural network architectures, the SNR of these mixtures is 3 dB. Each result is the average value of 200 experiments. *Italic* shows the proposed methods. **BOLD** indicates the best result.

STOI	Unprocessed	cIRM-DNN [50]	*cIRM-LSTM*	oSA-DNN [13]	*oSA-LSTM*	*cSA-LSTM*
Factory	70.36%	81.39%	83.94%	81.95%	84.89%	**85.99**%
Babble	71.22%	80.01%	82.99%	80.03%	85.28%	**86.03**%
Cafe	70.47%	79.20%	81.14%	79.30%	80.97%	**82.06**%
F16	72.45%	81.34%	82.62%	81.66%	84.02%	**84.71**%
Tank	79.66%	84.37%	87.77%	84.66%	89.20%	**89.26**%
Averaged	72.83%	81.26%	83.69%	81.52%	84.87%	**85.61**%

Table 5.4. Separation performance comparison in terms of PESQ with different training targets, noises and neural network architectures, the SNR of these mixtures is -3 dB. Each result is the average value of 200 experiments. *Italic* shows the proposed methods. **BOLD** indicates the best result.

PESQ	Unprocessed	cIRM-DNN [50]	*cIRM-LSTM*	oSA-DNN [13]	*oSA-LSTM*	*cSA-LSTM*
Factory	1.63	2.07	2.33	2.11	2.30	**2.41**
Babble	1.76	2.05	2.12	2.03	2.22	**2.28**
Cafe	1.75	2.03	2.16	2.10	2.14	**2.38**
F16	1.64	2.13	2.25	2.10	2.27	**2.38**
Tank	1.92	2.29	2.49	2.33	2.72	**2.74**
Averaged	1.74	2.11	2.27	2.13	2.33	**2.44**

Table 5.5. Separation performance comparison in terms of STOI with different training targets, noises and neural network architectures, the SNR of these mixtures is 0 dB. Each result is the average value of 200 experiments. *Italic* shows the proposed methods. **BOLD** indicates the best result.

PESQ	Unprocessed	cIRM-DNN [50]	*cIRM-LSTM*	oSA-DNN [13]	*oSA-LSTM*	*cSA-LSTM*
Factory	1.80	2.34	2.54	2.41	2.50	**2.59**
Babble	1.89	2.19	2.37	2.14	2.49	**2.51**
Cafe	1.95	2.27	2.38	2.29	2.32	**2.49**
F16	1.79	2.30	2.47	2.25	2.49	**2.61**
Tank	2.01	2.58	2.67	2.59	2.88	**2.91**
Averaged	1.88	2.34	2.49	2.37	2.54	**2.62**

Table 5.6. Separation performance comparison in terms of STOI with different training targets, noises and neural network architectures, the SNR of these mixtures is 3 dB. Each result is the average value of 200 experiments. *Italic* shows the proposed methods. **BOLD** indicates the best result.

PESQ	Unprocessed	cIRM-DNN [50]	*cIRM-LSTM*	oSA-DNN [13]	*oSA-LSTM*	*cSA-LSTM*
Factory	1.98	2.61	2.73	2.63	2.71	**2.81**
Babble	1.96	2.40	2.56	2.29	2.69	**2.76**
Cafe	2.01	2.46	2.58	2.48	2.55	**2.62**
F16	1.97	2.42	2.64	2.37	2.67	**2.77**
Tank	2.19	2.69	2.88	2.70	3.12	**3.17**
Averaged	2.02	2.51	2.67	2.49	2.75	**2.82**

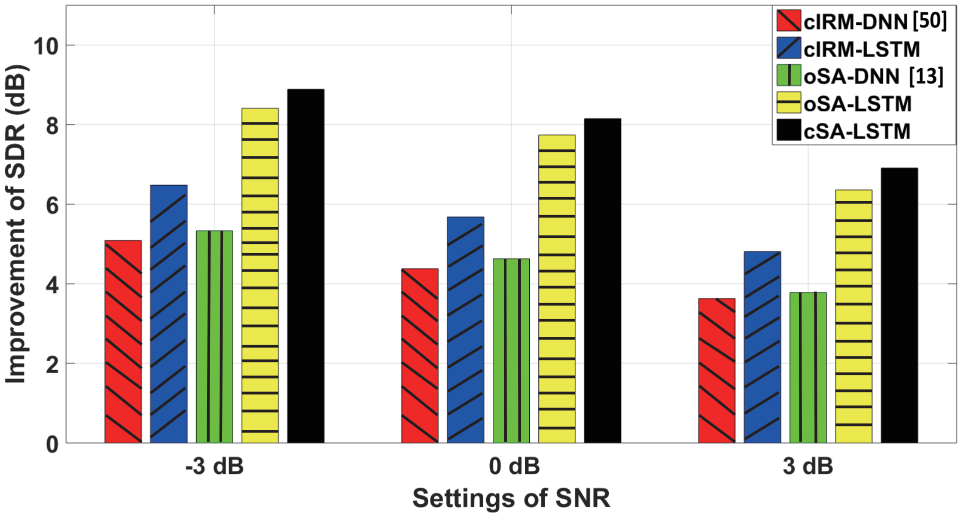

Figure 5.4. Average SDR improvement (dB) for different training targets and neural network models with five types of seen noise. Each result is the average value of 200 experiments.

The experiments with SDR aim to evaluate how the variations of the training targets, types of neural network models and SNR levels affect the SDR. It is shown in Fig. 5.4 that the proposed cSA-based LSTM RNN method achieves the largest SDR improvement in all scenarios. When the vanilla DNN is trained, the cIRM- and oSA-based methods offer almost the same SDR improvement. While comparing the cIRM- and oSA-based methods with DNN and LSTM RNN, the performance of the LSTM RNN is again better than the DNN. By using the proposed LSTM RNN, the oSA-based method can gain 3.08, 3.11 and 2.58 dB more SDR improvements at -3, 0, and 3 dB SNR levels, respectively. In addition, the phase information of clean speech signal in complex domain provides further SDR improvement, e.g. by comparing with the oSA- and the cSA-based LSTM RNN methods.

Unseen noise interferences

In the real-world environments where the situations varies, it is important to provide the generalization ability of the proposed methods. Therefore, the evaluation results based on the STOI and PESQ are shown in Table 5.7 for unseen noise cases. Each result in Table 5.7 is the average value of 200 testing mixtures, the baseline is calculated by using the unprocessed mixture and the clean speech signal.

Table 5.7. Separation performance comparison in terms of STOI and PESQ with different methods and the unseen noises, the SNR levels of these mixtures are -3, 0, and 3 dB. Each result is the average value of 200 experiments. *Italic* shows the proposed methods. **BOLD** indicates the best result.

	STOI			PESQ		
SNR level	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Unprocessed	59.50%	66.16%	73.00%	1.61	1.80	2.01
cIRM-DNN [50]	64.33%	70.68%	76.92%	2.07	2.22	2.37
cIRM-LSTM	65.56%	72.78%	79.43%	2.17	2.34	2.53
oSA-DNN [13]	63.17%	69.06%	75.81%	2.09	2.25	2.36
oSA-LSTM	66.30%	75.99%	81.02%	2.24	2.35	2.47
cSA-LSTM	**75.14%**	**78.87%**	**83.52%**	**2.29**	**2.47**	**2.60**

It can be known from Table 5.7 that when the noise interference is unseen, the separation performance is decreased, compared with the seen noise interference case. It is difficult to obtain the accurate estimate in the testing stage with unseen noise interference. For example, when the noise interference is seen, in 0 dB SNR level, the cIRM-based DNN method can gain 7.64% improvement in terms of the STOI. However, if the noise interference is unseen, the improvement decreases to 4.83%.

Besides, in the unseen noise interference case, when the SNR level is increased, the separation performance is improved and the best separation performance is given by the proposed cSA-based LSTM RNN method. For instance, in -3 dB SNR level case, the cSA-based LSTM RNN method achieves 75.14% and 2.29 in STOI and PESQ, respectively. While the oSA-based DNN method only achieves 63.17% and 2.09, respectively.

Hence, if LSTM RNN is selected as the neural network model, the generalization of the related methods is enhanced, which has been confirmed by the experimental results similar to [57].

Then, some experiments are evaluated to show how the variations of the SNR levels affect the SDR performance in terms of the proposed methods with unseen noise interference. Besides, the generalization ability is further evaluated. Fig. 5.5 gives the SDR improvement with different training targets and neural network models.

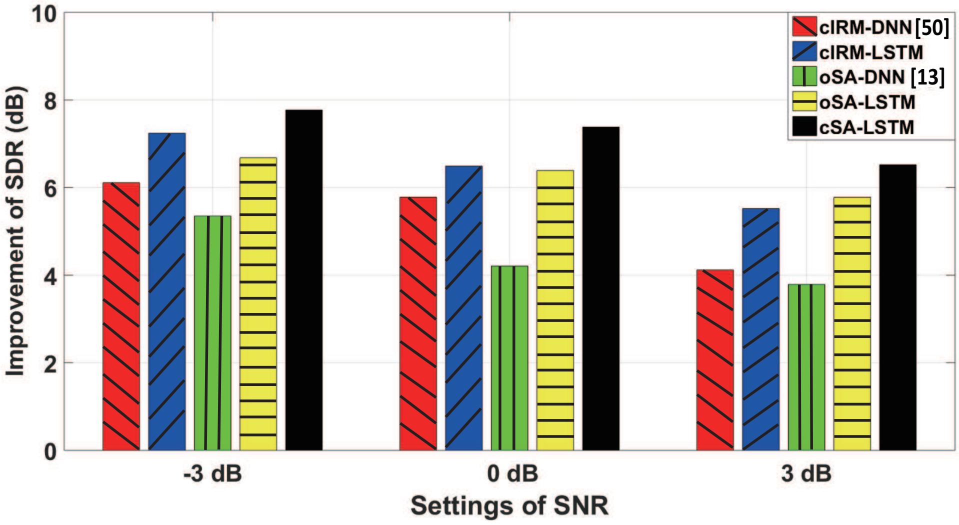

Figure 5.5. Average SDR improvement (dB) for different training targets and neural network models with 100 types of unseen noise. Each result is the average value of 200 experiments.

It can be seen from Fig. 5.5 that in the unseen noise case, compared with the cIRM-based DNN method, the cIRM-based LSTM RNN method gives more SDR improvement from -3 dB to 3 dB SNR levels. Similarly, the oSA-based LSTM RNN method achieves a higher SDR improvement than the oSA-based method by using the vanilla DNN. It is clear to observe that when the SA approach is operated in the complex domain and the LSTM RNNs are trained to predict the corresponding training targets, the separation performance outperforms others. For example, in the scenario, when the SNR level is -3 dB, the separation performance of oSA-based DNN method is 6.68 dB and the cSA-based LSTM RNN method gives 7.77 dB

SDR improvement.

From Tables 5.1 - 5.7 and Figs. 5.4 & 5.5, the best separation performance in noise interference case is given by the proposed cSA-based LSTM RNN method. There are two main reasons: (1) The phase information of clean speech signal is used to recover the desired speech signal; (2) the LSTM RNN exploits the temporal information and the generalization ability is enhanced. Besides, it can be seen from Table 5.7 that by using the proposed cSA-based LSTM method, the best performance in terms of the STOI and PESQ is obtained in all SNR levels, although there are some discrepancies in the level of improvements across these performance metrics. One possible reason is that when the SNR level is low, by using the proposed cSA-based LSTM method, the intelligibility of the separated speech, as assessed by the STOI, is better improved, due to the time-frequency weighting of the speech spectrum. In a high SNR level, less processing is enforced on the separated speech signal. As a result, the level of artefacts introduced by the proposed cSA-based LSTM method is lower, as shown by the PESQ measure.

In summary, in the seen noise interference case, the separation performance is better than the unseen case. When the SNR level is changed from -3 dB to 3 dB, all of the methods achieve better separation performance. Moreover, compared with the vanilla DNN, using the LSTM RNN as the neural network model, the proposed method provides improvement in all performance measures.

5.4.3 Evaluation with speech interference

When the interference is the undesired speech signal, the task is more difficult to address because the speech signals are highly non-stationary. In this subsection, the evaluations with undesired speech interferences are shown in Table 5.6 and Fig. 5.6.

From Table 5.6, it can be observed that when the interference is the undesired speech signal, compared with the noise interference cases, the separation performance decreases in all cases. The proposed cSA-based LSTM RNN method provides the highest values of both STOI and PESQ. Compared with the noise interference, when the interference is speech signal, because the indeterminacy of the speech interference, the related neural network model is more difficult to train, which effects on the overall separation performance.

After introducing the LSTM RNN, the separation performance is improved. For example, when the speech interference is used, in 0 dB SNR level, the oSA-based DNN method can gain 5.34% improvement in terms of the STOI, the oSA-based LSTM RNN method gives 7.51% improvement. In general, the phase information is beneficial and it can be observed that in -3 dB SNR level, the PSEQ value of oSA-based LSTM RNN method is 2.14 and cSA-based LSTM RNN method achieves 2.32.

Table 5.8. Separation performance comparison in terms of STOI and PESQ with different methods and the speech interference, the SNR levels of these mixtures are -3, 0, and 3 dB. Each result is the average value of 200 experiments. *Italic* shows the proposed methods. **BOLD** indicates the best result.

	STOI			PESQ		
SNR level	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Unprocessed	64.84%	69.03%	76.62%	1.63	1.92	2.01
cIRM-DNN [50]	69.27%	73.82%	80.16%	2.02	2.23	2.37
cIRM-LSTM	69.13%	73.11%	80.33%	2.05	2.19	2.39
oSA-DNN [13]	70.84%	74.37%	81.79%	2.02	2.30	2.38
oSA-LSTM	72.84%	76.54%	82.25%	2.14	2.36	2.48
cSA-LSTM	**75.80%**	**79.26%**	**82.59%**	**2.32**	**2.54**	**2.57**

The variations of the SNR levels affect the SDR performance in terms of the proposed methods with speech interference is shown in Figure 6. It can be seen from Fig. 5.6 that in the speech interference case, the cSA-based LSTM RNN method gives the largest SDR improvement over the other methods and SNR levels. It is shown that because the strong ability of using temporal information, the SDR improvement of the LSTM RNN-based method is always larger than the DNN-based methods. For instance, when the SNR level is -3 dB, the SDR improvement of the oSA-based DNN method is 4.11 dB and the improvement of the oSA-based LSTM RNN method is 6.24 dB.

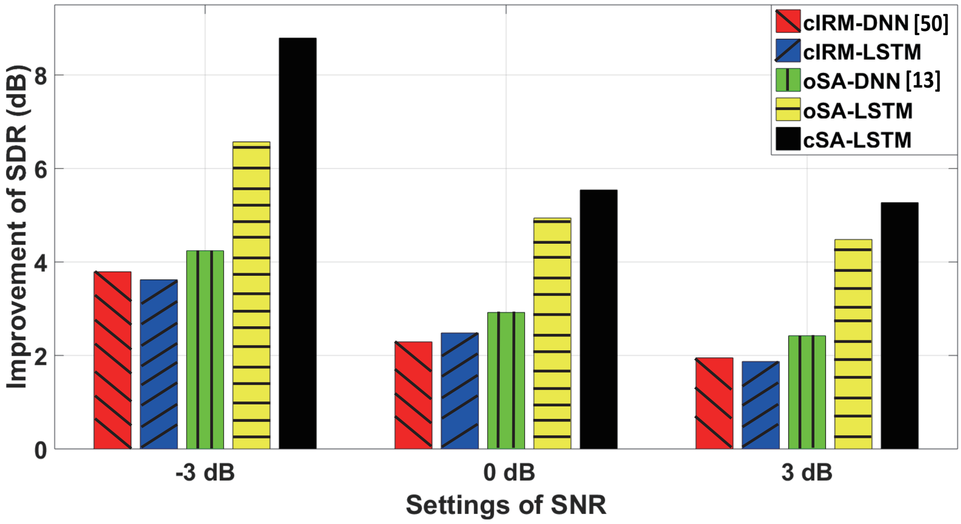

Figure 5.6. Average SDR improvement (dB) for different training targets and neural network models with speech interferences. Each result is the average value of 200 experiments.

However, in cIRM-based methods, due to the indeterminacy of the undesired speech signal, and the corresponding neural network is Y-shape, the T-F mask in the complex domain cannot be accurately estimated sometimes. For example, in Fig. 5.6, when the SNR level is -3 dB, the cIRM-based DNN achieves higher SDR improvement than the cIRM-based LSTM RNN method. To address this issue, in the proposed cSA-based LSTM RNN method, two individual LSTM RNNs are used to estimate the eal and imaginary components separately. It can be observed from Fig. 5.6, when the SNR level is -3 dB, the performance of the proposed cSA-based LSTM RNN method is 8.91 dB, which confirms the efficacy of the proposed method.

In summary, in the speech interference case, the separation performance is less than the noise interference case. When the SNR level varies from -3 dB to 3 dB, all of these methods achieve better separation performance in both noise interference and speech interference cases. From Tables 5.1 - 5.6 and Figs. 5.4 - 5.6, it is confirmed that the LSTM RNN is a better neural network model to utilize the long-term temporal information, which helps the trained model to obtain better separation performance.

5.5 Chapter Summary

In this chapter, contributions for improving separation performance were proposed by using both the phase information of clean speech signal and LSTM RNN. In terms of the phase information used, the complex signal approximation-based method was proposed and the imaginary component of the clean speech signal was estimated. To better utilize the temporal information, the LSTM RNN was used as the neural network model in the proposed cSA-based method. By introducing cIRM, both real and imaginary components can be calculated and estimated in the cSA-based LSTM RNN method. Compared with oSA-based method, if the complex domain training targets were exploited, the phase information can be used in the SA-based approach. Hence, in the cSA-based method, both clean magnitude and phase information were utilized and the separation performance was further improved. In Section 5.4, comparisons were made between the proposed novel cSA LSTM RNN-based method and traditional ones. Moreover, the evaluation confirms the improvement from the proposed method.

However, it should be noted that although the phase information is helpful to improve the separation performance, which can be observed by comparing the results of the oSA-based method with those of the cSA-based method, the major improvement actually comes from the use of the SA-base method, which can be observed by comparing the performance of the oSA-based method with that of the cIRM-based method. The proposed cIRM-based LSTM RNN method not only has the benefits from the SA formulation but also the clean phase information.

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 deep neural network (DNN)-based solutions to monaural source separation (MSS) problem, in particular to handle the challenges of generalization ability, mixture in real room environment and phase information.

In order to achieve these targets, different algorithms were proposed with different training targets, system structure and neural network model architectures. The contributions to improve the separation performance satisfy the three objectives mentioned in the introduction chapter. The first contribution was to use two sequentially trained DNNs with different training targets to build a system to achieve speech separation and the cost function with discriminative term; the second contribution was to provide a two-stage algorithm with dereverberation and separation stages, then two new time-frequency (T-F) masks were proposed as the training targets; and the last contribution was to improve the separation performance by utilizing the phase information of clean speech signal with the proposed complex signal approximation (cSA)-based method and the long short-term memory (LSTM) recurrent neural network (RNN) as the framework. The details of the contributions can be concluded as follows:

In Chapter 3, in the first proposed method, the discriminative term was added in the cost function with deep recurrent neural network (DRNN) to eliminate the influence of ambiguous information and utilize the temporal information. And the discriminative term was calculated adaptively via discrepancy to better model the ambiguous information. Besides, since the DRNN was introduced, it was easier to find the trade-off between separation performance and computational cost. In the second method, the sequentially trained DNNs were used to improve the tracking performance. Firstly, one DNN was trained with the spectrogram of clean speech signal, then based on the estimation from the first trained DNN, a T-F mask was calculated as used as the training targets of the second DNN. Both DNNs were employed to build a separation system, which can correct the over- or underestimated information of clean speech signal. The simulation results confirmed the outcome from the proposed method, for example, with TIMIT dataset and 100 noise interferences in the testing data, when the number of types in noise interferences was 3 with 0 SNR levels, the proposed system achieved 2.13 in PESQ while the mapping-based method achieves 1.74 and the PESQ value of masking-based method was 2.01. In terms of the STOI, when the number of types in noise interferences was 5 with 5 SNR levels, the proposed system gave 80.81% while the mapping-based method achieved 76.11% and the value of masking-based method was 78.06%.

In Chapter 4, following the concept of two-stage algorithm, the dereverberation stage was employed in order to obtain the dereverberated speech mixture before the separation stage; firstly, the dereverberation mask (DM) was proposed, so the ideal enhanced mask (IEM) was obtained by integrating the DM with ideal ratio mask (IRM), which employed the dereverberation and separation together. Then, two different system structures were used, one was single DNN trained with IEM, another was using two DNNs to train the DM and IRM separately. The separation performance of the mixture in real room environment was improved because the clean speech signal was estimated from the system instead of the direct sound of the clean speech signal. By evaluating with the IEEE, TIMIT and NOISEX datasets, the performances of the separation results were improved. For instance, in Room D, the proposed method with two DNNs achieves 13.1%, 8.7% and 12.5% STOI improvements over the proposed method with single DNN (integrated training objective) at -3, 0 and 3 dB SNR levels, respectively. The two DNN-based method provides around 13.9% more STOI improvement in all scenarios.

In Chapter 5, the cSA LSTM RNN-based method was proposed, which can be separated into two main contributions to improve the separation performance. Firstly, a novel cSA-based method was employed to estimate the phase information of the clean speech signal in order to avoid using noisy phase information in synthesis stage; secondly, LSTM RNN was used as the framework of the proposed cSA method. For example, when the speech interference was used, in 0 dB SNR level, the oSA-based DNN method can gain 5.34% improvement in terms of the STOI, the oSA-based LSTM RNN method gave 7.51% improvement. For instance, in -3 dB SNR level case, the cSA-based LSTM RNN method achieved 75.14% and 2.29 in STOI and PESQ, respectively. While the oSA-based DNN method only achieved 63.17% and 2.09, respectively.

6.2 Suggestions for Future Work

In order to further improve this study, there are some potential contribution points which could be further researched.

Firstly, in order to obtain further improvement in separation performance, the reinforcement learning method can be considered, which can help to solve the MSS problem by using unsupervised learning. Although such approaches have been proposed such as [61], this technique can be combined with some existing traditional supervised learning methods to obtain accurate estimation.

Secondly, apart from the whole frequency band estimation, according to the frequency range of the speech signal, the whole frequency range can be divide into two components [109]. One component contains most of the information from speech signal and another only have few of them. Therefore, different neural network architectures and training targets can be used depend on the different frequency sub-bands. According to this, more computational cost is allocated to the frequency sub-bands contains most of useful information.

Thirdly, video information can be considered to be exploited in the separation framework and several such approaches have been proposed such as [67, 110]. The video information gives the spatial information of the speech mixture, therefore, the position of speaker can be localized, which is beneficial for separation performance.

Finally, it is a very popular research topic, the source separation with time-domain. Different from the conventional DNN-based methods, by separating the desired speech signal in the time domain, the frequency decomposition step is removed and reduces the separation problem to estimation of source masks [111]. Meanwhile, the complex-value DNN can be applied to combine the magnitude and phase information of the desired speech signal.

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] Y. Sun, 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, 2018.

[3] 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.

[4] B. Wu, K. Li, F. Ge, Z. Huang, M. Yang, S. M. Siniscalchi, and C.-H. Lee, “An end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1289–1300, 2017.

[5] J.-F. Cardoso, “Blind signal separation: statistical principles,” *Proc. of IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[6] S. I. Amari, A. Cichocki, and H. H. Yang, “A new learning algorithm for blind signal separation,” *Adv. Neural Inf. Process. Syst*, vol. 8, pp. 752–763, 1996.

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

[8] T. Kim, T. Attias, S. Lee, and T. 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.

[9] 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.

[10] M. I. Mandel, R. J. Weiss, and D. P. 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.

[11] Y. Sun, W. Rafique, J. A. Chambers, and S. M. Naqvi, "Underdetermined source separation using time-frequency masks and an adaptive combined gaussian-students t probabilistic model," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[12] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," in *Proc. of IEEE International Conference on Digital Signal Processing (DSP)*, 2011.

[13] 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.

[14] 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.

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

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

[17] R. Wakisaka, H. Saruwatari, K. Shikano, and T. Takatani, "Speech kurtosis estimation from observed noisy signal based on generalized gaussian distribution prior and additivity of cumulants," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

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

[19] 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," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[20] S. M. Naqvi, *Fundamentals of PCA, ICA and IVA.* UDRC Summer School, Surrey University, UK, 2015.

[21] 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.

[22] H. Sundar, C. S. Seelamantula, and T. Sreenivas, "A mixture model approach for format tracking and the robustness of student's t distribution," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 2626–2636, 2012.

[23] 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.

[24] S. Sanei, S. M. Naqvi, and J. A. Chambers, "A geometrically constrained multimodal approach for convolutive blind separation of nonsationary sources," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[25] J. Taghia, N. Mohammadiha, and A. Leijon, "A variational Bayes approach to the underdetermined blind source separation with automatic determination of the number of sources," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[26] J. Taghia and A. Leijon, "Separation of unknown number of sources," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 625–629, 2014.

[27] O. Walter, L. Drude, and R. Haeb-Umbach, "Source counting in speech mixtures by nonparametric Bayesian estimation of an infinite Gaussian mixture model," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[28] 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," in *Proc. of IEEE Sensor Signal Processing for Defence Conference (SSPD)*, 2017.

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

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

[31] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2009.

[32] Z. Y. Zohny, S. M. Naqvi, and J. A. Chambers, "Enhancing MESSL algorithm with robust clustering based on Student's t-distribution," *Electronics Letters*, vol. 50, pp. 552–554, 2014.

[33] Z. Zohny, S. M. Naqvi, and J. A. Chambers, "Variational em for clustering interaural phase cues in messl for blind source separation of speech," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[34] Z. Wang and D. L. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 457–468, 2019.

[35] P. C. Loizou, *Speech Enhancement: Theory and Practice.* Boca Raton, FL, USA: CRC Press, 2007.

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

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

[38] J.-T. Chien and P.-K. Yang, "Bayesian factorization and learning for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 185–195, 2016.

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

[40] O. Dikmen and A. T. Cemgil, "Unsupervised single-channel source sep-

aration using bayesian nmf," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009.

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

[42] H. Guo and Z. Li, "A method of improving generalization ability for neural network based on genetic algorithm," *Proc. of IEEE International Conference on Intelligent Computing and Intelligent Systems*, 2010.

[43] A. Raikar, S. Basu, and R. M. Hegde, "Single channel joint speech dereverberation and denoising using deep priors," in *Proc. of IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2018.

[44] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[45] K. Han, Y. 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. 982–992, 2015.

[46] 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.

[47] 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.

[48] D. L. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[49] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[50] 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.

[51] D. S. Williamson and D. L. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1492–1501, 2017.

[52] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech Sep. Humans Mach*, vol. 60, pp. 63–64, 2005.

[53] J. Du, Y. Tu, L.-R. Dai, and C.-H. Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1424–1437, 2016.

[54] F. Chollet, *Deep learning with python.* Manning, 2017.

[55] F. Weninger, J. R. Hershey, J. L. Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. of IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014.

[56] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[57] 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.

[58] L. Sun, J. Du, L. R. Dai, and C.-H. Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017.

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

[60] L. Busoniu, R. Babuska, B. D. Schutter, and D. Ernst, *Reinforcement learning and dynamic programming using function approximators.* CRC Press, 2010.

[61] Y. Koizumi, K. Niwa, Y. Hioka, C. Kobayashi, and Y. Haneda, "Dnn-based source enhancement self-optimized by reinforcement learning using sound quality measurements," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] J. Hansen and B. Pellom, “An effective quality evaluation protocol for speech enhancement algorithms,” *Proc. of International Conference on Spoken Language Processing (ICSLP)*, 1998.

[67] M. S. Salman, S. M. Naqvi, A. Rehman, 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.

[68] B. Wu, K. Li, M. Yang, and C.-H. Lee, “A reverberation time aware approach to speech dereverberation based on deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 102–111, 2017.

[69] M. Yu, A. Rhuma, S. M. Naqvi, L. Wang, and J. A. Chambers, “A posture recognition-based fall detection system for monitoring an elderly person in a smart home environment,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1274–1286, 2012.

[70] Y. Sun, W. Wang, J. A. Chambers, and S. M. Naqvi, “Enhanced time-frequency masking by using neural networks for monaural source separation in reverberant room environments,” *Proc. of the 26th European Signal Processing Conference (EUSIPCO)*, 2018.

[71] 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, pp. 895–910, 2010.

[72] P.-S. Huang, M. Kim, M.-H. Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[73] 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.

[74] 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.

[75] J. Chen, Y. Wang, and D. L. Wang, "Noise perturbation for supervised speech separation," *Speech Communication*, vol. 78, pp. 1–10, 2016.

[76] Y. Wang, J. Du, L. R. Dai, and C.-H. Lee, "Unsupervised single-channel speech separation via deep neural network for different gender," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016.

[77] D. Yu and D. Li, *Automatic speech recognition a deep learning approach.* Springer, 2014.

[78] H. B. Demuth, M. H. Beale, O. D. Jess, and M. T. Hagan, *Neural network design.* Martin Hagan, 2014.

[79] N. R. S. Srinivasan and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[80] M. Hermans and B. Schrauwen, “Training and analysing deep recurrent neural networks,” in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2013.

[81] A. C. I. Goodfellow and Y. Bengio, *Deep Learning.* MIT Press, 2016.

[82] J. B. Conway, *A course in Functional Analysis.* Springer, 2000.

[83] J. W. Demmel, *Applied Numerical Linear Algebra.* Siam, 1997.

[84] D. Wolpert and W. Macready, “No free lunch theorems for optimization,” *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.

[85] 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.

[86] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proc. Conf. Artificial Intelligence and Statistics*, 2010.

[87] *Hidden Markov Model Toolkit (HTK).* Cambridge University and Microsoft, 2016.

[88] 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.

[89] R. Fletcher, *Practical Methods of Optimization, 2nd Edition.* Wiley, 2000.

[90] G. Hu, “100 nonspeech sounds.” `http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html`, 2014. online: accessed June 2018.

[91] 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.

[92] C. Hummersone, *Binaural Room Impulse Response Measurements.* Surrey University , United Kingdom , 2011.

[93] M. Valente, H. Hosford-Dunn, and R. J. Roeser, *Audiology: Treatment.* Thieme, 2008.

[94] D. H. Griesinger, “The audibility of direct sound as a key to measuring the clarity of speech and music,” *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. 2319–2319, 2011.

[95] Y. Wang, K. Han, and D. L. Wang, “Exploring monaural features for classification-based speech segregation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270–279, 2013.

[96] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, “An algorithm that improves speech intelligibility in noise for normal-hearing listener,” *Journal of the Acoustical Society of America*, vol. 126, pp. 1486–1494, 2009.

[97] H. Hermansky and N. Morgan, “RASTA processing of speech,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[98] M. Delfarah and D. L. Wang, “Features for masking-based monaural speech separation in reverberant conditions,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1085–1094, 2017.

[99] 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.

[100] S.-H. Jin and C. Liu, “English sentence recognition in speech-shaped noise and multi-talker babble for English-, Chinese-, and Korean-native listeners,” *Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 391–397, 2012.

[101] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.

[102] C. Chen and J. A. Blimes, “MVA processing of speech features,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 257–270, 2007.

[103] F. Gers and E. Schmidhuber, “Lstm recurrent networks learn simple context-free and context sensitive languages,” *IEEE Transactions on Neural Networks*, vol. 12, no. 6, pp. 1333–1340, 2001.

[104] 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.

[105] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Interspeech 2015*, 2015.

[106] 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,” in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[107] E. Ceolini and S.-C. Liu, “Impact of low-precision deep regression net-

works on single-channel source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[108] S. Ruder, "An overview of gradient descent optimization algorithms," *in preprint arXiv: 1609.04747*, 2016.

[109] H. Bourlard, S. Dupont, H. Hermansky, and N. Morgan, "Towards subband-based speech recognition," in *Proc. of the 8th European Signal Processing Conference (EUSIPCO)*, 1996.

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

[111] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.