Single and Multichannel Speech Source Separation using Non-Negative Matrix Factorisation Incorporating Spectral Masks

Yuxiao Feng

University of Wollongong

Follow this and additional works at: https://ro.uow.edu.au/theses1

Recommended Citation

School of Electrical, Computer and Telecommunications Engineering

Single and Multichannel Speech Source Separation using Non-Negative Matrix Factorisation Incorporating Spectral Masks

Yuxiao Feng

"This thesis is presented as part of the requirements for the Award of the Degree of Master of Philosophy from the University of Wollongong"

March 2017
ABSTRACT

The problem of separating mixtures of speech signals has always been a heated topic in speech processing. Multiple speech separation approaches have been proposed and a successful separation system benefits numerous applications, such as hands-free communication systems. However, separation performance of existing techniques is still unsatisfactory in terms of both speech quality and speech intelligibility. Recently, data driven approaches to solving speech signal processing problems, where information learnt from example databases of speech recordings is used to derive new signal processing algorithms has shown significant success. Consequently, this thesis investigates one of the data-driven models for speech separation, namely non-negative matrix factorization (NMF) and relevant methods, with the expectation of achieving increased speech quality and speech intelligibility of separated speech sources compared to existing approaches. Specifically, Chapter 3 proposes an NMF approach modified with spectral magnitude masks typically derived for single-channel speech separation. Chapter 4 then proposes an enhanced NMF approach that utilises estimated direction-of-arrival information to realize multi-channel speech separation. Compared with corresponding baseline methods, the proposed approaches demonstrate improvements in speech quality and intelligibility metrics, which verifies the success of the proposed approaches in this thesis.
THESIS DECLARATION

I, Yuxiao Feng, hereby declare that all material in this thesis, submitted in fulfilment of the requirements of the award of Master of Philosophy, in the School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Australia, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Yuxiao Feng
March 2017
ACKNOWLEDGEMENTS

First and foremost, I would like to sincerely thank my supervisor, Assoc. Prof. Christian Ritz. He guided me to manage this thesis step by step, with his professional knowledge and inspiring suggestions. Besides, throughout the thesis, his explicit plan extremely helped me to schedule the progress, founding the most crucial part to all my achievements today. Moreover, I am greatly grateful for him being patient and responsible, which has always been my solid support during the hard time. He is not only a supervisor but also a mentor, who sets the best example for the rest of my life. Besides, I also would like to thank Assoc. Prof. Lei Wang, who gives me useful and experienced advises with his profound understanding on machine learning during my beginning time. This brings many opportunities for me to start and continue the research on the thesis.

In addition, I would like to thank my parents, who have been always backing me up whatever the hardship is. Through all the years, they underwent a hard time, but still support me at all times. I sincerely appreciate what they have done to help me fulfilling the thesis.

At last, I would like to thank all my lab colleagues, Jacob, Shahab, Jie, Yi, Xiguang and many of my friends. Your heartfelt suggestions help me, bringing me many new ideas and joys.

I sincerely hope you will always stay happy and healthy.
TABLE OF CONTENTS

ABSTRACT .................................................................................................................. i
THESIS DECLARATION ......................................................................................... ii
ACKNOWLEDGEMENTS ....................................................................................... iii
TABLE OF CONTENTS ......................................................................................... iv
LIST OF FIGURES ................................................................................................. vi
LIST OF TABLES .................................................................................................... vii
1 INTRODUCTION ................................................................................................. 1
  1.1 Thesis overview ............................................................................................... 1
  1.2 Background on the speech separation problem ............................................. 2
     1.2.1 Problem description .............................................................................. 2
     1.2.2 Possible solutions ................................................................................. 4
  1.3 Contributions of this thesis ........................................................................... 6
  1.4 Publications ................................................................................................... 7
  1.5 Outline of this thesis ...................................................................................... 7
  1.6 Chapter conclusion ......................................................................................... 7
2 LITERATURE REVIEW ......................................................................................... 9
  2.1 Background for audio signal processing ....................................................... 9
  2.2 Typical microphone arrays .......................................................................... 12
     2.2.1 Non-directional microphone arrays ...................................................... 14
     2.2.2 Directional microphone array .............................................................. 15
  2.3 Signal transform ............................................................................................ 17
     2.3.1 Fourier transform and relevant alternatives ........................................ 17
     2.3.2 W-disjoint orthogonality in speech separation .................................... 19
     2.3.3 Alternative transform methods ........................................................... 21
  2.4 Approaches to speech separation ................................................................. 21
     2.4.1 Spatial-assistant approaches ................................................................. 22
     2.4.2 CASA associated methods ................................................................. 27
     2.4.3 Statistics approaches ........................................................................... 28
     2.4.4 Matrix-decomposition approaches ....................................................... 29
  2.5 Baseline methods of thesis ........................................................................... 34
     2.5.1 Sparsity & discriminative constraints for single-channel NMF .......... 34
5.2 Recommendations for future work ................................................................. 89
REFERENCES ........................................................................................................ 93

LIST OF FIGURES

Figure 1.1 Soundfield in a conference room ......................................................... 3
Figure 2.1 Uniform linear array ........................................................................... 14
Figure 2.2 B-format microphone array ................................................................. 16
Figure 3.1 Training stage of the proposed speech separation method ................. 48
Figure 3.2 Input SIR vs. output SIR ................................................................... 56
Figure 3.3 Input SIR vs. Output SAR ................................................................. 56
Figure 3.4 Input SIR vs. Output SDR ................................................................. 57
Figure 3.5 Input SIR vs. Output PESQ ............................................................... 57
Figure 3.6 Input SIR vs. Output STOI ............................................................... 58
Figure 4.1 DMNMF separation system ............................................................... 64
Figure 4.2 Localization of 3 active sources with B-format array ......................... 74
Figure 4.3 SDR performance with 2 active sources ........................................... 77
Figure 4.4 PESQ performance with 2 active sources ......................................... 77
Figure 4.5 STOI performance with 2 active sources .......................................... 78
Figure 4.6 SDR performance with 4 active sources ........................................... 79
Figure 4.7 PESQ performance with 4 active sources ........................................ 79
Figure 4.8 STOI performance with 4 active sources ........................................ 80
Figure 4.9 Output SDR performance versus different input SDR ...................... 82
Figure 4.10 PESQ performance versus different input SDR ................................. 82
Figure 4.11 STOI performance versus different input SDR ................................. 83
Figure 4.12 SDR performance against different errors in DOA initialization ....... 85
Figure 4.13 PESQ performance against different errors in DOA initialization ....... 85
Figure 4.14 STOI performance against different errors in DOA initialization ....... 86
LIST OF TABLES

Table 2.1  WDO of activities in a frequency band from two speakers..................20
Table 3.1  Algorithmic of the proposed training and separation approach...........53
Table 3.2  Input SIR vs. MEAR..........................................................................57
Table 4.1  Source position (2 src)......................................................................74
Table 4.2  Source position (4 src)......................................................................76
1 INTRODUCTION

1.1 Thesis overview

In modern daily life, speech communication applications, such as mobile phones, voice assistants, tele-meetings and automatic speech recognition play a prominent role and require speech signal processing to serve human more naturally and efficiently. In general, to realize these applications, a complete processing system needs to be established, basically guaranteeing three points: first, the system should be able to receive and enhance the useful part of a speech signal; second, the system is supposed to understand the human’s meaning in the received speech signal; third, the system could think and judge like a human-being, and thus give back proper reactions to satisfy human’s demand. Based on this design, the system could be roughly divided into three parts: enhancement, recognition and feedback [98].

Being the most fundamental part, enhancement plays a crucial role for the realization of the whole system. Theoretically, after a proper enhancement procedure, useful information carried by the received signal is supposed to be clean enough for the remaining processing stages. Ideally, factors, including undesired noise, interferences and reverberations ought to be removed, maintaining the target speech of good quality. Practically, the processing expense should be also taken into account, depending on different application purposes. For instance, mobile-phone applications may focus more on the ability of processing time and less computation-demand; while translation-associated applications may emphasize more on the accuracy and being understandable. Combining these purposes, a generally applicable speech enhancement system becomes a hard problem.

Thus, this thesis targets the speech separation problem. Under this basic assumption, the original signal will be treated as a mixture, where both the desired part and undesired part of the speech signal participate. The recorded mixture signal will be enhanced based on the proposed methods, keeping the desired part and removing the undesired part. As the desired part is the target speech source, the undesired part is mainly composed of environmental noise, reverberation and
interfering speech sources [97]. The enhancement methods aim to increase the quality for the target applications are hands-free communication systems.

1.2 Background on the speech separation problem

1.2.1 Problem description

Generally, the basic model for audio propagation in a 2-dimensional scenario can be described as follows:

\[ y(t) = \sum_{i} h_{r_i,\theta_i}(l) \otimes x_{r_i,\theta_i}(t - l) + d_{noi}(t) \]  

(1.1)

(1.1) indicates the general situation for a sound propagation in time domain, where the index \( t \) represents the continuous time axis. The \( y \) stands for the received signal, while \( h_{r_i,\theta_i}, x_{r_i,\theta_i} \) and \( d_{noi} \) represent acoustic transfer function (ATF), original audio source and additive noise respectively. The ATF \( h_{r_i,\theta_i} \) may cause multi-path effects on the propagation, bringing decay of the source signals and reverberations. The sub-index \( i \) is the index number of possible sources, with \( r_i \) and \( \theta_i \) as the radius and angle attributed to \( i \)-th sound source in 2-dimensional space [98]. The number \( I \) in (1.1) is defined as the number of total active sources. Operator \( \otimes \) is the convolution operation between the ATF filter \( h_{r_i,\theta_i} \) with the speech source \( x_{r_i,\theta_i} \). Figure 1.1 gives an intuitive view on the audio signal propagation inside a conference room. In this case, the whole rectangular stands for the boundary of the conference room. A target source is labelled as \( Sou \), with an interference source and a receiver labelled as \( Int \) and \( Rec \), respectively. Regarding to the target source, there is one directive-propagation path labelled as \( S_{dir} \), with \( S_{rev1} \) and \( S_{rev2} \) as the propagation path of early echo and reverberation, respectively [99]. The difference between early echo and reverberation can be found in [99], which is distinguished with the number of reflections during propagation. Under this assumption, there are three ATFs of the sound propagation between the target source and the receiver, corresponding to three different propagation paths. Likewise, the interference only has one ATF of directive-propagation as \( I_{dir} \) and one ATF of early echo as \( I_{rev} \). Besides, the
additive noise is labelled as $d_{noi}$, where the propagation is indicated by the dense-dash. Despite the hardware noise of receiver itself, most additive noises are diffuse, which do not have a directional-propagation path, thus being demonstrated with a dense-dash in Figure 1.1.

As for the speech separation problem, the target is to recover the desired speech source $x_{r_{i_0} \theta_{i_0}}$ from the signal $y$ arrived at the receiver, where number $i_0$ is the specific number of the speech source. Since $y$ is the mixture between target source, interference source, the reverberations and the noise, the processing should retain the $x_{r_{i_0} \theta_{i_0}}$, and remove the remaining part of the received signal. It is worth noting that since the model of (1.1) includes multiple speech sources, the target source might be a specific speech source, or each speech sources. This is different from most single target source enhancement problems, which are referred to as the traditional speech denoising problems [97] [98].
1.2.2 Possible solutions

Based on the (1.1), it is clear that the separation problem can be resolved with three factors, namely, the ATF, information on the speech sources or information on the undesired sound. This Chapter will only give a general explanation on the corresponding methodologies relevant to the thesis, and specific techniques from all three areas which have strong relation to implement our methods. More technical details will be elaborated in Chapter 2.

As for the ATF, there are considerable methods for estimating the ATF. The key of estimating ATFs is to discover the spatial audio parameters, such as direction-of-arrival (DOA) and sound source localization, which could be exploited to derive ATFs and set up corresponding spatial enhancement solutions, including beamforming and spatial clustering [97]. The ATFs are mainly decided by the characteristics of the environment within which the sound field propagates. Different soundfield type, sound source locations, environmental factors all affect the final outcome. Meanwhile, the environment may be a free-field, where there is no reflection of the sound, such as an anechoic room; or the environment may be with sound reflections, where there exist both echo and reverberations after the sound being omitted by the source [97] [98]. In most situations, these reflections increase the hardships of estimating the ATFs. After the ATF estimation, the separation can be realized by inversing the propagation process or suppressing the undesired speeches [97] [98] within the received signals, thus enhancing the desired speeches. This thesis assumes a far field assumption.

Among all the spatial information, DOA is a particularly important parameter. Basically, a DOA estimation is to estimate every angular position, namely, the $\theta_i$ in (1.1) corresponding to its source index $i$. To have a DOA estimation, usually the system requires multiple microphones, in order to have several received signals and calculate the DOA information from them [96] [97] [98]. Compared with specific localization information (i.e. the exact spatial position of the original source), DOA information is often easier to compute and express.

Meanwhile, estimating information about the target speech is also beneficial to perform separation. This can involve estimating statistics, spectral parameters or other features that are automatically recognised using e.g. machine learning
techniques [12] [96] [97] [98]. Different kinds of data, side information and features have been exploited, focusing on developing a better separation.

In recent years, the popularity of non-negativity computation, such as the NMF model, has been growing significantly, especially in speech processing research [11] [12] [100]. As for speech, the spectrogram has always been chosen as one of the main input terms, since phase of speech is hard to be analysed during separation tasks and less important than the magnitude for the possible following operations, such as recognition or translation [98]. This illustrates the superiority of the NMF model due to the non-negativity of spectrogram. Moreover, under some well-known assumptions of speech separation problems, the additivity of sources is quite similar to the additivity of active features in the NMF model. This makes us believe that choosing the NMF model as the main framework in this thesis is reasonable.

Besides, the information on the undesired sources also has usability, which is mostly linked to noise estimation techniques. The central idea is to use the known information of the undesired noise, estimating the power of the undesired noise. Afterwards, since the mixture is the summation between the target speech and undesired ones, a subtraction operation can recover the desired speech. Sometimes, since the subtraction is not perfect, it may be followed by some post-processing [12]. Besides, it is also useful for combining both the information of undesired sources and the desired ones to compute specific factors, such as calculating the signal-to-noise ratio (SNR) or the relevant ones with the power information of each source received signal. These ratios could be essential for enabling some separation systems, which will also be explained in Chapter 2 [9] [12].

More precisely, spectral mask techniques, which exploit information of each time-frequency (TF) point in a mixture spectrogram against some configurations to distribute the TF point to a certain source or not, have demonstrated their effectiveness for speech enhancement for a long time [1] [99]. Due to the similarity between the ideal binary mask with common sparsity constraints [18] [21] in speech separation methods, this thesis proposes that introducing this into classic NMF models can enhance the separation performance. Therefore, the fusion between the NMF model and a binary mask has also been one of the main research points in this thesis.
To conclude, since the data becomes more and more accessible today, NMF techniques have been chosen as the main research target in this thesis. Meanwhile, this could bypass the difficulty of getting some complex knowledge as well as handcraft features and preserving the robustness against different situations at the same time [9] [12] [97]. Specifically, the thesis proposes integrating the information from a binary mask into single channel NMF, and DOA information into multichannel NMF have been looked into details, with outcomes in comparison between these two proposals and corresponding most-recently relevant speech separation methods under well-acknowledged evaluation criterions [42] [72] [101] [102].

1.3 Contributions of this thesis

The major findings of this thesis are briefed as follows:

1. Chapter 3 describes a novel extension of semi-supervised NMF for the single channel speech separation problem. Based on the traditional NMF, the method fuses the mask information of source-wise spectrogram structure learned during a training-stage into a post-processing procedure, where the sparsity factor can be more effective and reasonable to suppress the errors during the separation stage. Compared with the other existing methods relevant with this similar assumption, the method shows an advantage of a 0.5-1 dB reduction of the signal-to-interference ratio (SIR) and maintains the other evaluation criteria at a comparable level. The superiority is more apparent when the input signal is under low SNR situations. Besides, if applications are pursuing a general use under all cases, the method is better because of its more robust performance on all evaluation criteria [85].

2. Chapter 4 describes a novel framework to utilize the advantages of a sound-field microphone array, DOA estimation and the NMF features. By merging these features into the separation processing, the system is aims to utilize spatial information corresponding to NMF features, thereby suppressing the mistakes of using the wrong features corresponding to the desired speech source during the reconstruction phase.
1.4 Publications

The publications are listed as the follows:

- Y. Feng, C. Ritz, "Single-channel speech separation by including spectral structure information within non-negative matrix factorization", *IEEE China Summit and Int. Conf. Signal and Information Processing (ChinaSIP)* 2015, pp. 411-414. [85]

1.5 Outline of this thesis

In the following parts of this thesis, the second Chapter is the literature review, where all potential techniques will be elaborated in terms of the advantages and disadvantages to define the comprehensive scope of an advanced speech separation system; Chapter 3 & Chapter 4 will be focused on the two main innovation contributions of this thesis by now, namely, frameworks to integrate mask information and DOA information into the traditional non-negative matrix factorization (NMF); the last Chapter is the conclusion, which reviews the achievements and explains the potential future work of this thesis, followed by the reference list.

1.6 Chapter conclusion

In Chapter 1, an overview of this thesis was presented, including background knowledge of audio processing, the speech separation problem and possible applications. Approaches exploiting spatial information, data-driven models or other associated models achieved relative success in source separation problems, with several problems still remaining. Among all methods, this thesis chooses NMF associated techniques for the speech separation problem. With the success of simulation tests in Chapter 3 and Chapter 4, the proposed methods demonstrate great
potential in terms of speech separation, which verifies the contribution of this thesis. The next chapter provides a detailed review of techniques used for speech separation.
2 LITERATURE REVIEW

This Chapter reviews the technology of speech enhancement, especially with the ones specifically designed for the speech separation problem. It is worth noting that the detailed results are studied in the following Chapters, where comparisons between the performances from different but strongly-related methods can be shown under the same evaluation criteria corresponding to the experiments. Therefore, this Chapter mainly serves as the discussion on processing methods from the most basic concepts to the current advanced methods.

To begin with, the most basic definitions of audio, soundwave propagation and soundfields are introduced in Section 2.1. Next, common but representative evaluation criteria are discussed in terms of their targets and scenarios. Following this several practical microphone arrays are presented, including their arrangement for obtaining a signal with different attributes. Section 2.4 demonstrates the well-acknowledged processing step, namely Fourier transform and its offspring along with a crucial speech property, named W-disjoint orthogonality. Based on all the aforementioned concepts, Section 2.5 presents the three most basic designs in terms of implementing a system catering to our problem; among these three designs, the Section 2.6 specifically discusses the possible data-driven approaches, such as sparse coding, NMF, time series models and the neural-networks. Finally, different possible extensions of NMF are discussed, which are the methods used for the comparison baselines of Chapters 3 and Chapter 4.

2.1 Background for audio signal processing

The general case of soundwave propagation is described as the (1.1) in Chapter 1, where a soundwave is a transmission of the pressure released by the sound sources at the physical level. Similarly, as a kind of oscillations, soundwave propagation requires a media, which is air for the target speech applications investigated in this thesis. Under this assumption, the oscillation from each sound source results in the transmission of air particles, heading to all directions. After certain reflections and decay during the transmission, the sound wave will be finally received by the
receivers. The relation between sound speed and medium conditions can be described by:

\[ c = \sqrt{\frac{\varepsilon}{\rho}} \]  

(2.1)

Here, \( c \) is the sound speed in metres per second, with \( \varepsilon \) and \( \rho \) as the Young’s modulus and density of air. Usually, in the environment with a common level of humidity and atmospheric pressure, the speed can be approximated as:

\[ c = 331.4 + 0.6T \]  

(2.2)

where \( T \) is the air temperature in degrees Celsius [96]. Then, the wavelength \( \lambda \) of soundwave at certain frequency \( f \) can be calculated as:

\[ \lambda = \frac{c}{f} \]  

(2.3)

Moreover, a very crucial factor, namely the intensity \( I \) of a soundwave, can be determined by the follows:

\[ I = \frac{E}{4\pi d^2} \]  

(2.4)

In this equation, \( E \) is the power of sound in watt, with \( d \) as the distance of sound propagation in meter. More precisely, as for the human auditory system, the threshold is at \( I_0 = 10^{-12} Watts/m^2 \). Thus, the relative intensity in decibel (dB) with the threshold of human auditory system as the reference intensity will be [98]:

\[ I_{dB} = 10 \log_{10} \left( \frac{I}{I_0} \right) \]  

(2.5)

Obviously, from the (2.4), it can be found that the intensity of a soundwave will decay along with the distance travelled, namely the attenuation.

To connect with Section 2.2, here the relation between air particles and sound pressure is presented. Based on the intensity from (2.4), this relation can be represented as:

\[ P = \frac{I}{c_{par}} \]  

(2.6)

where the \( P \) is the sound pressure, with the \( c_{par} \) as the particle velocity, typically as the air particles in the common environment. To measure the particle velocity, the following equation can be used:

\[ c_{par} = \frac{j}{KZ \nabla P} \]  

(2.7)
In this equation, \( j \) is square root of \(-1\), which indicates that the driving force is at a \( \pi/2 \) radians in front of the particle velocity. \( K \) is the wavenumber of the soundwave, which can be computed by the division between the soundwave’s angular frequency and its velocity:

\[
K = \frac{\omega}{c}
\]  

(2.8)

\( Z \) is the impedance, equal to air density \( \rho \) times sound velocity \( c \):

\[
Z = \rho * c
\]  

(2.9)

Moreover, \( \nabla P \) is the gradient of sound pressure. These will form the fundamental part of some specific microphone arrays, such as the acoustic vector sensor (AVS) [103].

All of the aforementioned factors are involved with the soundwave propagation. Besides, there could be several other effects. For a free field, the soundwave will be always considered as travelling in direct lines, heading to infinity; however, as for the practical scenario, there are always some reflections, scattering, diffractions and absorptions due to the obstacles standing on the way of sound traveling [96] [97] [98].

Due to the length limit of this report, the discussion cannot be expanded upon for most of these factors other than reflection. Reflection is a typical issue, especially in a reverberant environment where both reverberation and echo could exist. As described by (1.1) in Chapter 1, the convolution operation and time delay index \( l \), correspond to the effect on sound propagation under a reverberant environment.

Among different parts of reflection, one possible phenomenon is about early reflections: after the omission from the sound source, a soundwave arrives at boundaries such as surfaces of walls, is then reflected, and directly propagates to the receiver. If the delay of the early reflection is more than 30 milliseconds, then it is called an echo. In addition, another possible effect might be the reverberation, which is how the soundwave gets reflected multiple times because of obstacle surfaces and finally arrives at receivers [96].

Both of these could trigger serious problems when a separation or denoising system is implemented at the end of receivers, and affect the speech quality or intelligibility [42] [72] [101] [98].

More precisely, the reflections are affected by several physical factors, such as the material and area of the reflection wall, sound sources’ and receivers’ locations,
propagation paths and conditions in terms of humidity, temperature, density, etc. All these factors contribute to the changes of sound absorptions and reflections, thereby altering the situation of final received signals [96]. Especially with reverberation, reverberation time \( R_{T_{60}} \) is defined to measure the level of reverberation [96] [104]:

\[
R_{T_{60}} = \frac{-0.161V}{S \cdot \ln(1 - \alpha)}
\]  

(2.10)

The physical significance of \( R_{T_{60}} \) is that the sound level drops 60dB compared with the direct arrival sound wave. In (2.10), \( V \) stands for the volume of the soundfield, usually as a room; \( S \) represents the surface area; \( \alpha \) is the absorption coefficient with a range from 0 (no absorption effect) to 1 (complete absorption effect).

In fact, the difference between the reverberation and early reflection is quite large: for both direct sound and early reflected sound, the propagation follows (2.4), where the intensity of the soundwave at the receiver remains as an inverse square relation with the soundwave traveling distance (the early reflection is also affected by the absorption effect from the only reflection during its propagation). However, due to being reflected multiple times, the reverberation will come from almost every direction to the receiver, wherever the receiver is, thus mixing into together and keeping basically the same level of reverberation intensity at every point inside the room. Due to this, researchers usually refer to the reverberant part as the ‘reverberant field’.

2.2 Typical microphone arrays

In order to realize the processing system in real life, choosing a proper hardware as the receivers is a necessary step. Generally, researchers use single or multiple sensors to capture the energy of soundwaves and convert them into electrical signals, thereby enabling analysing and processing the audio signals. These devices are often referred to as a microphone or microphone array [96].

Microphones can be categorised into two types, namely temperature-sensors microphone and pressure-sensors microphone [96]. As for the former, it can be also referred as a particle-velocity-sensor microphone. In the previous section, concepts of particle velocity and sound pressure are discussed in (2.7) and equation (2.6), respectively. Apparently, a temperature-sensors microphone is designed to measure
the particle velocity, thus converting it into sound energy information; on the contrary, the pressure-sensors microphone is used to measure the pressure directly from soundwave.

Besides, microphones can also be grouped into different types based on their directivity. To decide the directivity, researches define the polar pattern as the gains of the microphone receiving a signal corresponding to all directions. If the polar pattern is a circle, which means the levels of receiving gain are the same at every point for sources at the same distance to the microphone, then this microphone is called an omni-directional microphone. Likewise, there are other polar patterns, such as sub-cardioid, cardioid, or hyper-cardioid [96].

Although a single microphone has a large advantage on the size scale, a microphone array can grant more benefits, especially with the spatial information and associated applications. By placing several microphones together and potentially knowing the geometry information between them (distance, angles, facing directions, etc.), this combination can be easily utilized to derive spatial information, including DOA, sound source location and room information [92] [93] [96] [103], thereby enabling further usage of them, such as beamforming or tracking movement of objectives [92] [96]. Meanwhile, with development of today’s manufacturing technology, the size of the microphone array can be shrunk into a relatively small level. As a consequence, this discussion is thus expanded more on microphone arrays, which is also highly involved with part of the thesis contributions.

Based on different types of information that microphone is able to derive from capturing the soundwave’s energy, microphone arrays could be roughly grouped into a non-directional microphone array and directional microphone array. Compared with non-directional microphone arrays, directional microphone arrays can derive directional signals from received soundwaves after simple mathematical calculations, such as an AVS or soundfield microphone array (B-format microphone array) [93] [96] [103]. Moreover, a directional microphone is usually very compact, especially practical when the application scenario only allows small-size devices, such as a hearing-aid system or microphone array of a cell phone. The following two sections elaborate several examples of both types of microphone arrays.
2.2.1 Non-directional microphone arrays

Basically, a ‘non-directional microphone array’ is a combination of multiple microphones (mostly omni-directional microphones), arranged into some typical shapes, such as a direct line, a circle or a sphere. Consequently, these microphone arrays are named as a uniform linear array (ULA), circular microphone array or spherical microphone array [96].

Since the shape of a microphone array is fixed, the geometric relation between each microphone is determined. With basic knowledge of soundwave propagation, these geometric relations can be used as the input to calculate spatial factors [96].

Take the ULA in Figure 2.1 as an example. To simplify the mathematical representation, the source is assumed in the far field, which means the soundwave arriving at each microphone remains a parallel relation. Here, the $\theta$ is the incident angle of arrival soundwaves, and $d$ is the distance between two nearest microphones. It is very clear that the wave arriving at microphone 1 ($M_1$) will travel $d \cos(\theta)$ longer than compared with the wave arriving at microphone 2 ($M_2$).

If the information on one of two factors, namely $\theta$ or $d \cos(\theta)$ is known, the other one factor can be worked out with the known $d$. This will be useful to decide either DOA of sound or the delay between two adjacent soundwaves, thus bringing
possibility to solve further problems such as sound source localization or speech enhancement [96].

Similarly, circular microphone array and spherical microphone array also have their corresponding methods to get this information [92] [96].

To compare these three, from ULA to circular microphone array to spherical microphone array, the accuracy of estimating spatial information is increasing, since there are usually more microphones, thus enabling estimation more accurate recording of the sound field [92]. This is especially true when the position of sound sources changes from positions at the same level to positions with different heights. In three-dimension scenarios, spherical microphone arrays will have an apparent superiority to distinguish the targets with the same coordinates on the horizontal plane but different vertical coordinates, because it can overcome the errors about mistaking the source location to the symmetric location on the other side of horizontal plane [96]. However, these advantages come at the expense of microphone array size growing larger, which might not be proper for some applications with limited space, such as a mobile phone.

2.2.2 Directional microphone array

In this thesis, the term ‘directional microphone array’ is defined to simplify description of the microphone arrays which can derive directional signal with simple mathematical computation [93] [96] [103]. Directional signals benefit the following parts in the whole system, especially as the DOA estimation which will be discussed with in more detail in the later sections [93] [96]. In addition, this type of microphone array is highly related to part of this thesis contribution, where a fusion between data-driven model and DOA characteristics is implemented. Thus, the microphone arrays with ability to derive directional signals easily are thus chosen as the topic of this section.

Among common microphone arrays, AVS and soundfield microphone (B-format microphone) are two typical ‘directional microphone arrays’. An AVS is a set of three directional microphones plus an omni-directional microphone. Ideally, the three directional microphones are used to receive only signals from $x$, $y$ and $z$ direction of a Cartesian coordinates, while the omni-directional microphone are used to measure
signals from all directions [96]. It is also worth noting that these directional microphones actually measure the gradient of sound pressure defined in (2.6) and (2.7) rather than sound pressure itself [96] [103]. A more detailed description of the AVS can be found in [103], but the most important point is signals heading to $x$, $y$ or $z$ directions could be simply derived, which is a very potential advantage for further investigations.

Compared with the AVS, a B-format microphone array is with some similarity to the ‘non-directional microphone array’. Basically, it is composed by four cardioid microphones and the geometry of these microphones grants the array to compute directional signals [93].

Figure 2.2 presents an intuitive view of a B-format microphone array. As the Figure shows, four microphones are at the corner of a tetrahedron. They are named as front left (LF), front right (RF), back left (LB) and back right (RB), respectively. For the two microphones at the front side in the Figure 2.2, namely the LF and RF in the figure, they are symmetric from vertical axis. Likewise, the LB and RB have the same relation.

These four microphones realize the B-format output, which is the combination between four channels and can represent directional and omni-directional signals. Specifically, if $y_{xc}, y_{yc}, y_{zc}, y_{oc}$ are referred as the $x$, $y$, $z$, and omni-directional
signals, then with a soundfield microphone, the following equations can be derived [93]:

\[
\begin{align*}
    y_{xc} &= LF - RB + RF - LB \\
    y_{yc} &= LF - RB - RF + LB \\
    y_{zc} &= LF + RB - RF - LB \\
    y_{oc} &= LF + RB + RF + LB
\end{align*}
\]

(2.11)

To conclude this section, ‘directional microphone arrays’ are extremely superior when they are compared with ‘non-directional microphone arrays’ in terms of obtaining directional signals. Additionally, for practical implementation, both ‘directional microphone arrays’ have only four elements setting in an extremely small structure, but still enable three dimensional recordings. Compared with ‘non-directional microphone arrays’, they are open to more applications such as ones with tiny space to install the recording system. However, due to this small size and distance between each microphone, both ‘directional microphone arrays’ suffer under certain cases [96].

2.3 Signal transform

Following an audio signal being received, transforming the signal is usually performed. Basically, signal transform benefits the following signal analysis, lightening hardships of processing and analysing. The following contents cover the common transform methods in audio, especially speech signals, with discussion of some crucial speech attributes after the transform.

2.3.1 Fourier transform and relevant alternatives

The Fourier transform is one typical transform that most transform methods are based on. In short, French mathematician Joseph Fourier initializes the proposal that all signals can be decomposed as a combination between several sinusoidal components with different frequencies [98]. As the consequence, Fourier transform converts the originally time-domain signal into frequency-domain representations. The following two equations are the computation for continuous Fourier transform (CFT) and inverse Fourier transform [98]:
Analysis: \[ X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft} \, dt \]  
(2.12)

Synthesis: \[ x(t) = \int_{-\infty}^{\infty} X(f)e^{j2\pi ft} \, df \]

Here \( x(t) \) is the continuous signal in the time domain, corresponding to its frequency domain representation \( X(f) \).

More specifically, as for the purpose of this thesis, all signals are assumed to be finite and discretely sampled, which correspond to the most common cases for today’s use of digital devices and audio signal processing. Hence, the discrete Fourier transform (DFT) follows [98]:

Analysis: \[ X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N} \]  
(2.13)

Synthesis: \[ x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k)e^{j2\pi kn/N} \]

where index \( n \) is the discrete signal sampling indexes and \( k \) is the index of frequency point index.

Speech is a typical audio signal with short-time stability for frequency-domain representation. Usually, this short duration is assumed to be at least 20 milliseconds [98]. However, if a DFT analysis is directly performed on a speech signal, the analysis duration is assumed to cover the whole duration of speech, thereby diminishing the analysis resolution. Thus, in order to enhance the analysis performance, researchers prefer to choose short-time Fourier transform (STFT) in discrete-time form with its inverse as follows:

Transform: \[ X(m_w, \omega) = \sum_{n=0}^{N-1} x(n)\text{win}(n - m_w) \, e^{-j\omega n} \]  
(2.14)

Inverse: \[ x(n) = \frac{1}{2\pi \text{win}(n - m_w)} \sum_{m_w=0}^{N-1} X(m_w, \omega) \, e^{j\omega n} \]

Here, win\((n)\) is a window function to pack the signal into frames; \( \omega \) is the angular frequency \( (\omega = 2\pi f) \). Index \( m \) is the index of the frame number, which is involved with the time delay of the mth window. According to (2.14), it is obvious that the signal representation is in the time-frequency domain, which means different frequency-points \( \omega \) aligned corresponding to time-line. Therefore, the representation
will concentrate more on a specific time duration defined by $m_w$-th window and present the signal inside this duration into frequency points [98].

As for the choice of window function, there are plenty of choices, such as rectangular window, triangle window, hamming window or Blackman window [98]. However, in order to avoid leak-out of frequency components, rectangular window are not often chosen. For the other ones, researchers also overlap the adjacent windows at different ratios, expecting the leak-out effect will be reduced as much as possible. Meanwhile, the length of the window, namely the time-duration size is mostly determined by specific purpose corresponding to applications [98]. For speech denoising and separation, this is usually set from 20 milliseconds to 32 milliseconds, which is the common duration with short-time stability. On the contrary, speech recognition processing might choose a bit longer duration (46 milliseconds); while the other type audio signals, such as music signal, can extend this duration to around 0.5 seconds since they are more sinusoid-like during the whole music duration [98].

2.3.2 W-disjoint orthogonality in speech separation

With knowledge of STFT, it is possible to represent the audio signal in the time-frequency (TF) domain as a set of TF points $X(m_w, \omega)$. For speech separation problem, the purpose is to separate the whole set of TF points into different sources, and guarantee the errors of separation on target speech source are as low as possible.

To realize the separation, there is a well-known concept, namely ‘W-disjoint orthogonality’. Briefly, W-disjoint orthogonality (WDO) is a property that different sources’ TF representations do not overlap with each other. This term is initially proposed by O. Yilmaz et al [105] with the following mathematical expression:

$$X_i(m_w, \omega)X_j(m_w, \omega) = 0, \forall m_w, \omega, i \neq j$$ (2.15)

where $i$ and $j$ are indexes of two sources. The equation implies that in the received mixture signal, any TF point will only belong to one source [64] [105].

A more intuitive view can be found in the Table 2.1, where the contents demonstrate two speakers’ activities in TF domain. To simplify the explanation, only the activities in one frequency band along time frames are analysed, which represents
Table 2.1 WDO of activities in a frequency band from two speakers

| Speaker 1 | | | | | | | |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
|           | [       ] | [       ] | [       ] | [       ] | [       ] | [       ] |
| Speaker 2 | | | | | | | |
|           | [       ] | [       ] | [       ] | [       ] | [       ] | [       ] |

Time axis

The same happens on the other frequency bands according to the WDO theory.

As the table showing, one block stands for one frame in one specific frequency band (one TF point) from both speakers, where shaded blocks represent the ones with corresponding speakers’ activities and the others represent the ones without activities. The horizontal axis stands for timeline. Obviously, two speakers’ activities in most TF points do not overlap with each other.

However, this condition cannot be satisfied when the purpose is to separate speech mixtures. Considering the similarity between different persons’ speech and finite resolution of analysis, plus the effect of sound reflection, the WDO condition will be seriously violated. For example, the third and the ninth blocks in Table 2.1 are overlapping, which means the two speakers’ activities in this two TF points do interfere each other. Hence, researchers proposed different methods to measure WDO level in practical scenarios [64]. Meanwhile, the others also studied the relation between WDO level and final evaluation level in terms of separation performance [64].

Though, WDO or nearly WDO still remains as one of crucial fundamental assumptions for major part of speech separation. In the following part of this chapter, more details will be reviewed corresponding to each specific speech separation approach.
2.3.3 Alternative transform methods

Considering that Fourier transform is based on linear-located basis \( (e^{-j2\pi kn/N}) \), there has always been a debate whether a better transform can replace pure Fourier transform, with the expectation for capturing more useful information from signals. Basically, it is reasonable to believe that to study other Fourier-family members and non-linearly located frequency basis is beneficial.

The major motivation is because of the fascinating performance of human auditory system when a human being faces various separation scenarios. Different from the Fourier transform, human auditory system has a pre-processing step, where signals firstly pass through bunch of logarithmically-located bandpass filters, namely the cochlea structure. Meanwhile, the human auditory system is able to mask part of the audio components according to the auditory masking \([1][100]\). These are the important features which are not taken account in classic Fourier transform processing.

Hence, for speech processing, researchers proposed many alternative processing methods to improve the final performance, including discrete cosine transform (DCT) \([39]\), constant-Q transform (CQT) \([98]\), sinusoid models \([58]\), linear predictive coding (LPC) processing \([98]\), gammatone filter processing \([41][45]\) and wavelet processing \([2][13][28]\).The remaining methods in this paragraph, these sorts of models are referred as sub-band models, implying logarithmically-located-bandpass filtering at the beginning of processing.

Besides, several different transforms have been employed as well, with the purpose for sidestepping the disadvantages of STFT, such as empirical model decomposition (EMD) \([58][61][62][63]\), cepstral spectrum \([26]\) or modulation-domain techniques \([43][55]\).

2.4 Approaches to speech separation

Different separation system approaches can be applied to the transformed signals. Based on the different types of information they use, these approaches can be divided into four groups, namely: spatial-assistant approaches, computer-auditory-scene-analysis (CASA), statistics approaches and matrix-decomposition approaches.
2.4.1 Spatial-assistant approaches

Since the objects of sound sources cannot overlap with each other in the same spatial position, soundwaves from different sources always travel along different paths to the receiver. Spatial-assistant approaches mainly focus on the discovery of different spatial information corresponding to different sound sources, thereby complying separation.

2.4.1.1 DOA estimation

For the purpose of separation, the DOA can be used as the essential information for beamforming methods [92] [96] [97] [98] to accomplish separation, or fed into source localization methods [38] [92] [106] [107]. At the same time, the DOA is also a crucial parameter in the proposed system of Chapter 4. Thus, this section describes DOA estimation. For the simplification of mathematic terms, discussions about DOA in this thesis only consider far-field as the basic assumption, where all waves arrive in parallel to the microphones in one receiver array [92] [93] [96]. As for the near-field cases, all methods can be extended by certain microphone array processing techniques [96].

One typical approach for DOA estimation is to compute the time-delay between waves from the same sound source but arriving at different microphones. Based on these time-delay estimations, it is able to calculate DOA with the help from microphone array geometries. Take the situation in Figure 2.1 (Section 2.2.1) as the simplest example. If \( \tau_{ij} \) is defined as the time delay between microphone \( i \) and microphone \( j \), then it is easy to have:

\[
d_{\text{wav}_{12}} = \tau_{12} * c \tag{2.16}
\]

\[
\cos(\theta) = \frac{d_{\text{wav}_{12}}}{d} \tag{2.17}
\]

where \( d_{\text{wav}_{12}} \) is the travelling-distance difference between the waves arriving at microphone 1 and microphone 2; \( c \) is the speed of sound; \( d \) is the distance between two nearest microphones with \( \theta \) as the estimated DOA. Moreover, if there is more than one pair of microphones, or the microphone array’s geometry is different (circular microphone array, spherical microphone array e.g.), it can be adapted corresponding to certain geometry [92] [106] [107] or utilization of redundancy [64] [96].
Due to the fact that the most fundamental factor is time-delay estimation, this type of approach is referred to as time-difference-of-arrival (TDOA) estimation. Researchers mainly propose using cross-correlation between signals from different channels to handle this problem [96] [106]. Obviously, with clear information on microphone array's geometry, the potential estimation error only comes from the estimated time-delay. Unfortunately, the time-delay estimation in practical cases is often a hard problem because of the overlap with other correlated sources and reflections misleading the correlation results [96]. To counter these negative effects, there are many following operations on the correlation matrix, such as generalized-cross-correlation (GCC) methods [96] or multiple-signal-classification (MUSIC) [96]. However, since these operations need the correlation matrix and potential computations on it (eigenvector decomposition for MUSIC e.g.), the computation expense is extremely high when the number of microphones and length of signal increase [96].

Another choice of DOA estimation benefits from the directional microphone arrays. For example, the two-dimension problem could be handled with by a B-format microphone in Figure 2.2 (Section 2.2.2). There is information contained in the $x$ channel (sub-index as $xc$) and $y$ channel (sub-index as $yc$), which can be modelled as the omni-directional channel’s (sub-index as $oc$) signal multiplied by $\cos(\theta)$ and $\sin(\theta)$, respectively [103] [107]:

$$
\begin{bmatrix}
    y_{xc}(t) \\
    y_{yc}(t) \\
    y_{oc}(t)
\end{bmatrix} = 
\begin{bmatrix}
    \cos(\theta) \\
    \sin(\theta) \\
    1
\end{bmatrix} \ast h(l) \otimes x(t - l) + d(t) \quad (2.18)
$$

Here, the assumption only considers one-source case in order to simplify the mathematical description and $h(l), x(t)$ and $d(t)$ are as specified in (1.1). After signal transform (STFT as an example), these relations remain same. Therefore, in TF domain, the DOA can be calculated as follows [103] [107]:

$$
\theta(m, n) = \arctan \left[ \frac{\text{Re}\{P_{pc}(m, n)P_{yc}(m, n)\}}{\text{Re}\{P_{pc}(m, n)P_{xc}(m, n)\}} \right] \quad (2.19)
$$

where $P$ is the pressure in (2.6) recorded by each channel, $\ast$ stands for conjugate. From this section, $m \& n$ will represent the index of frequency and time after STFT, in order to align with the indexes used in following sections. Furthermore, due to the linear relation between pressure and intensity in (2.6), (2.19) could be also changed to the intensity form [107]:

23
\[ \theta(m, t) = \arctan \left[ \frac{\text{Re}\{I_{oc}(m, n)I_{yc}(m, n)\}}{\text{Re}\{I_{oc}(m, n)I_{cc}(m, n)\}} \right] \]  

(2.20)

This is referred to as intensity-direction-of-arrival (IDOA). The major disadvantage of IDOA is similar to TDOA, where the process results in the estimation presenting multiple directions due to the effect from other source signal overlapping on the spectrum with target signal. In contrast, since this method only needs signal transform and some basic calculations, it surely reduces the computation expense [107].

In order to solve the common problem, there are many proposals following the initial DOA estimation. One typical solution is about clustering [93] [107] [108]. The most basic one is to use histogram, which can be explained by the following:

\[ \text{his}_\mu = \sum_{i_\theta} D(i_\theta) \]  

(2.21)

where \( i_\theta \) stands for the number of all DOA estimations. If the DOA estimation is derived by IDOA, then \( i_\theta = \text{number}(\omega) \ast \text{number}(t) \). \( D(i_\theta) \) implies if \( i_\theta \)-th DOA estimation is inside the boundaries of \( \mu \)-th segment as following:

\[ D(i_\theta) = \begin{cases} 1, & \text{if } (\mu - 1)\theta_\text{seg} \leq \theta(i_\theta) < \mu \theta_\text{seg} \\ 0, & \text{else} \end{cases} \]  

(2.22)

In this equation, \( \theta(i_\theta) \) represents the initial \( i_\theta \)-th estimated DOA from any initial DOA estimation method, and \( \theta_\text{seg} \) is the angle-resolution for segmentation [93] [107]. After being relocating into histogram, the \( \mu \)-th segment can be decided as a true source–direction with certain threshold if:

\[ \frac{\text{his}_\mu}{\text{his}_\text{max}} \geq \gamma \]  

(2.23)

where \( \text{his}_\text{max} \) is the maximum of \( \text{his}_\mu \), among all \( \mu \) and \( \gamma \) is the threshold.

However, the performance of this method majorly depends on the resolution and threshold setting. If the level of noise is relatively high, then too small a resolution or too high a threshold could cause underestimation of certain directions and vice versa [107].

Hence, Zheng et al. propose the weighted-histogram method as follows:

\[ \text{his}^\mu_{\text{weighted}} = \sum_{i_\theta} D(i_\theta) \ast \| E(i_\theta) \| \]  

(2.24)

where \( E(i_\theta) \) is the omni-signal corresponding to \( i_\theta \)-th DOA estimation, and \( \| \cdot \| \) is the L1 norm operator [108]. It is believed that the weighted processing will improve the
histogram to have a more apparent peak corresponding to important sources, since important sources often have relatively large energy compared with the remaining ignorable sources with less energy. Low energy sources are more likely to correspond to silence regions, background noise (if SNR is high) or late reflections due to reverberation, while they still contaminate the separation decision if not weighted [108]. However, these methods require at least 100 ms of signal to achieve accurate DOA estimation results and so cannot be regarded as real-time in the context of speech communication applications [108].

Another type of clustering is based on probability estimation. Here the highly-related one for this thesis is discussed, namely the Von-Mises-distribution approach. Von-Mises-distribution is a very well-known probability function for directional data clustering, which is illustrated as follows [93] [107]:

$$f(\theta; \mu_i, \kappa_i) = \frac{e^{\kappa_i \cos(\theta - \mu_i)}}{2\pi I_0(\kappa_i)}, \quad i \in I$$  \hspace{1cm} (2.25)

where variable $\theta$ is still one DOA of the whole initial DOA-estimation set, $I$ represents the number of sources, $\mu_i$ stands for angle of i-th cluster with $\kappa_i$ as the concentration parameter of i-th cluster. $I_0$ is the modified Bessel function of order zero. This function gives a way to decide the level of clustering certain DOA $\theta$ into i-th source’s cluster [107]. For the mixture of all sources, (2.25) can be changed into the following form in practical scenarios [107]:

$$f(\theta; \mu, \kappa) = \sum_{i=1}^{I} a_i e^{\kappa_i \cos(\theta - \mu_i)}$$  \hspace{1cm} (2.26)

where $a_i$ is the weight parameter for each cluster, usually as $1/(I+1)$ since it does not have large effect on the final estimation result [107].

With the von-Mises-function, checking the reliability of DOA estimations becomes possible. However, the parameter $\kappa_i$ is hard to decide if there is merely von-Mises-function’s information in the practical scenario. Thus, Gunel et al. proposes to use the discrete $\kappa_i$, which is decided by:

$$\kappa = \frac{\ln 2}{1 - \cos(\theta_{BW}/2)}$$  \hspace{1cm} (2.27)

where $\theta_{BW}$ is a set of discrete beamwidth (3dB, 5dB … e.g.), and the optimal value is determined by enumerating during the processing [107]. Apparently, compared with the aforementioned methods, the von-Mises-function-based method is more explicit and flexible. This is because the rectified DOA estimation, namely $\mu$ is not affected
by segmentation resolution from the histogram-based methods, and \( \kappa \) benefits the cluster to be flexible. On the contrary, since the parameters require more optimization, the computation expense and robustness do need to be taken into consideration [107].

2.4.1.2 The DUET algorithm

The degenerate unmixing estimation technique (DUET) is another famous type of source separation method that can use spatial information [64]. If the basic assumption about WDO as in Table 2.1 is true, then each TF point should be dominated by only one source. Then, on every TF point, the difference between every channel should be only depended on the attenuation and delay effect corresponding to the microphone array geometry. Thus, a spatial feature could be built up based on this. Without losing generality, one TF point \((\omega_m, t_n)\) is taken from a 2-channel microphone array in an anechoic environment as an example:

\[
\begin{bmatrix}
Y_1(m, n) \\
Y_2(m, n)
\end{bmatrix} = \left[ a_i e^{-j\omega_m \delta_i} \right] X_i(m, n), m \in M, n \in N
\] (2.28)

Here, \( M \) and \( N \) are the number of frequency samplings and the number of frames, respectively. Symbol \( \omega_m \) stands for the frequency of \( m \)-th frequency point. Index \( i \) is the dominant source in this particular TF point \( a_i e^{-j\omega_m \delta_i} \) is named as the mixing parameter, which represents the spatial information caused by signal from one source to different microphones. Because of the assumption about WDO, the signal from other sources contributes nearly nothing to this TF point, thus being able to approximate as 0. Then it is able to calculate the following two parameters for each TF point in the whole TF-point set \( \Omega_{M,N} \):

\[
a_i = \left| \frac{Y_2(m, n)}{Y_1(m, n)} \right|
\] (2.29)

\[
\delta_i = \left( \frac{-1}{\omega_m} \right) \angle \left( \frac{Y_2(m, n)}{Y_1(m, n)} \right), \quad \forall m \in M, n \in N
\] (2.30)

As the reference from S. Rickard [64], \( a_i \) is referred to as the local attenuation parameter and \( \delta_i \) as the local delay parameter. Correspondingly, this computation can operate on every TF point, creating a set of combinations of \( [a_i \quad \delta_i] (i \in I) \). Then the demixing mask can be presented as:
\begin{equation}
M_i(m, n) = 1, \quad if \quad [a(\omega_m, t_n) \quad \delta(\omega_m, t_n))] = [a_i \quad \delta_i]
\end{equation}

This mask is able to be multiplied with the spectrum of the received signals to select certain TF points together as the spectrum corresponding to one source [64].

Compared with all previous separation schemes based on spatial DOA estimation, DUET demands less computation power, and builds a direct way to obtain separation results. However, the basic assumption is based on WDO, which is not true for many real scenarios [64]. Hence, the performance of the algorithm will be relatively limited. Certainly, several algorithms are proposed as the enhancement to DUET [64] [106]. With the use on mixing parameters, one of them is an important part of this thesis contribution’s baseline in the following Section 2.5.2 and Chapter 4, which will be elaborated later in this thesis.

2.4.2 CASA associated methods

In brief, CASA is a type of method that builds computer processing scheme imitating the human auditory system [98]. As aforementioned information, a human auditory system has advantages of non-linearly located filters of the cochlear and the auditory masking effect. On the other hand, there are also many studies into the masking effect, which leads to the various computer-mask methods in CASA [1] [5] [8] [93] [99]. Since one of them, namely ideal binary mask (IBM) plays a crucial role in part of this thesis contribution, the mask methods and associations will be analysed in this section as the main discussion of CASA-relevant techniques [1] [5] [8].

To begin with, the concept of the most basic mask, namely the IBM is presented as an example. Based on the SNR in Section 2.1.2, IBM can be computed as follows:

\begin{equation}
IBM(\omega, t) = \begin{cases} 
1, & if \quad SNR(m, n) > LC \\
0, & else 
\end{cases}
\end{equation}

where \(SNR(\omega, t)\) is the local SNR value of TF point \((m, n)\), and \(LC\) is the threshold value designed based on the known background knowledge of the noisy environment [1] [8]. After the estimation, the IBM will be multiplied by the spectrogram of received signals to realize the separation, which is similar to the use of masks from DUET.

With recent research results, the IBM shows a large benefit for separation performance in terms of intelligibility. However, it is obvious that there are two main contributors for errors in IBM estimation, including the improper local SNR
estimation and improper local SNR-threshold implementation [8]. Various methods are proposed to handle this problem, such as using spatial sparsity [109] to replace SNR or using voice-activity-detection (VAD) to improve SNR estimation [99]. However, they are still not able to solve the root cause of IBM errors [99].

Furthermore, there are also other types of masks, such as ideal ratio mask (IRM) and short time Fourier transform magnitude mask (FFTM), which are shown as the following, respectively [99]:

\[
IRM(m, n) = \left( \frac{X^2(m, n)}{X^2(m, n) + D^2(m, n)} \right)^\beta \\
FFTM(m, n) = \left( \frac{X^2(m, n)}{Y^2(m, n)} \right)Y^2(m, n)
\]  

where \(D^2(m, n)\) is the local additive noise spectral power, and \(\beta\) is the function constant. From (2.33), IRM replaces IBM to multiply with the noisy spectrogram and get the separated speech’s spectrogram. As for FFTM, the physical implementation of separation is presented as (2.34), while in the practical scenario the value \(\frac{X^2(m, n)}{Y^2(m, n)}\) needs to be directly estimated. For IRM, the major problem is that it assumes all noise (including interference) to be additive on the spectrogram, and needs either estimation of the target speech spectrogram \(X^2(m, n)\) or noise spectrogram \(D^2(m, n)\), which is hard to get under practical cases [99]. Similar to IRM, FFTM requires estimation of \(\frac{X^2(m, n)}{Y^2(m, n)}\) which might bring large error into final separation [99].

2.4.3 Statistics approaches

With the development of digital devices, there is a great increase of audio data. Meanwhile, the computation power has also been exponentially increasing, which allows more and more learning-based models to become practical [97]. One way to exploit data for source separation is to extract different sources’ statistics and the relations between them, which is usually referred as statistical models.

One of the most famous statistical models is independent component analysis [70]. Specifically, it assumes the independence between different sources, which can be utilized as the key feature for the estimation on separated source. With the basic assumption, the received mixture should be decomposed into several independent components, which leads to the solution of a similar mixing matrix in (1.1) with
corresponding solution of original sound-set from each source. Although in the original ICA system, the mixing process is supposed to be a linear one, the extension on convolutive mixture can be found in several successful implementations [97]. As for the role of statistics knowledge in ICA, the theory builds a separation system based on the level of independence, which is mathematically defined as a penalty term during decomposition process. The Non-Gaussianality between different source’s signals after decomposition is the most common one used in this case [70]. Despite the unnecesary of data and statistics on the original sources, they still benefit the final estimation in several terms, such as the scale of the final separated signals, the permutation problems and etc. [97].

However, ICA has its drawbacks, which is involved with the basic assumption to assume independence between different sources. Meanwhile, the system requires an equal number between source and microphone-array channels at least in order to derive the meaningful math solution [97]. These limit the performance of ICA in practical implementation.

In spite of ICA, there are other statistical models, which mostly give the sources’ statistics inference and exploit these characteristics. These are highly-related to the following contents in the matrix decomposition approaches, where most probability models are used either on sources or optimization penalties. Due to limits of this thesis’s scope, only the related statistical knowledge is discussed in the following sections.

2.4.4 Matrix-decomposition approaches

As it stated above, data greatly benefits the separation system in many ways. Despite the complex statistical analysis, one promising way is about matrix-decomposition approach. Compared with handcrafted features, learning-based features are more robust, catering to the specific but considerable data beyond human handcrafting ability [97]. On the other hand, learning-based features are based on computer-computation structure, which introduces less potential errors from tuning human-based features to the real computer-computations. Recently, a large amount of learning-based models have been invented [11] [18] [21] [27] [30] [50] [51] [70] [100], being proved to be effective on the speech separation problem as well [9] [10]
Among them, a typical and highly-relevant model is matrix decomposition models, which will be elaborated in the following [46] [57] [70].

The reason for choosing matrix decomposition models is that this kind of model assumes a linear relation between feature and data, which is easy to realize and guarantees the robustness when the amount of data is relatively small [11] [12] [97]. Generally speaking, the matrix decomposition problem can be basically shown as follows:

$$\min f(S; W \ast H) = \langle \text{errors} \rangle, \quad \text{subject to} \langle \text{constraints} \rangle$$

(2.35)

Where $f(S; W, H)$ is the loss function, composed by data matrix $S$ and features $W$ times corresponding weights $H$. The term $\langle \text{errors} \rangle$ implies certain error-measuring term, such as $L_p$ norm or divergences between the data matrix and reconstruction from $W \ast H$. The subject $\langle \text{constraints} \rangle$ is to require the decomposition obeying certain constraint conditions [12] [18] [46] [97] [100].

Due to the similarity of the data from the same cluster, features of this data should be capable to represent this certain cluster under any decomposition, while have enough small reactions when the data of the other clusters shows up. As for speech separation, this could mean features of the target source will only have valid weights when the target source is active in the received mixture signals; on the contrary, weights will be small enough to be ignore when the target source is inactive [14] [17] [20] [46] [47] [57]. Then, the separation can be realized with the reconstruction from the features times their weights corresponding to the target source.

In the following context, the discussions concentrate on two type of matrix decomposition model, namely sparse coding (SC) [18] and NMF [100]. These two are highly involved with the thesis contribution, thereby being treated as important discussion contents. There are other data driven methods that have recently shown promise [2] [12] [28] [37] [52] [97], but they typically require a large amount of training.

### 2.4.4.1 Sparse coding

The definition of sparse data is that the data is overcomplete. In other words, any data point can be decomposed as the multiplication between parts of the feature set with corresponding weights, where the number of active features is largely smaller
than the number of all features. The following equation indicates this relation [18] [19] [21] [73] [69] [111]:

$$\min_H \|H\|_0, \quad \text{s.t. } \|S - WH\|_2 < \epsilon$$ (2.36)

Here, $\|H\|_0$ is the $L_0$ norm of weight matrix $H$, which means the number of active features. Besides, $\|S - WH\|_2$ is the $L_2$ norm, measuring the Euclidean distance between data matrix $S$ and the reconstruction $W \ast H$. From (2.36), it can be seen that the reconstruction must be close enough to the original data matrix, namely the decomposition being accurate enough; at the same time, the number or active features must be minimized to a relatively small level [18].

However, in practical scenarios, (2.36) is not easy to be calculated, thus researchers propose to relax of the whole problem [18]:

$$\min_{W,H} \sum_{n=1}^{N} \|S - WH\|_2, \quad \text{s.t. } \|H\|_0 < T_0$$ (2.37)

Regarding the speech separation problem, data matrix $S$ is usually the spectrogram of received mixture signals in one frame. $T_0$ is the threshold, standing for the largest number of possibly active features. The summation with $n$ means this error should be optimized with data from all frames. While $W$ is the feature set combining features from each source with weight set $H$ corresponding to the weight matrix of every source. In other words, the whole reconstruction is the summation between source-wise reconstructions:

$$WH = [W_1, W_2, \ldots W_i, \ldots W_I] \ast [H_1, H_2, \ldots H_i, \ldots H_I]^T$$ (2.38)

where index $i$ is still the index sources. It can be seen that the weight matrix is actually the parallel source-wise feature matrices, corresponding to their own weight matrix, respectively. When a certain source is the target source and needs to be separated from others, it can be realized as follows [44] [73] [76] [111]:

$$S_{\text{tar}} = W_{\text{tar}}H_{\text{tar}}, \quad \text{tar } \in 1$$ (2.39)

where index $\text{tar}$ is the index of target source. (2.39) gives the reconstruction of the target-source spectrogram. Following this, point-wise multiplying the target spectrogram with the phase matrix from the original mixture will obtain the final reconstruction of the target source [73] [111]. The phase mixture is used because the quality of reconstructed speech mainly depends on the spectrogram shape, where the potential errors from the phase matrix can be ignorable [73] [111].
The major benefit of SC comes from the effect of increasing feature’s representativeness. Since the number of active features is always less than the number of all features, every feature must be very typical in order to ensure the reconstruction errors being small enough. Moreover, the orthogonality between features is increased, thereby reducing the possibility of overlapping during the separation phase [73] [111].

However, to build proper features corresponding to each source is actually very hard [73]. Meanwhile, a reasonable value of threshold $T_0$ is also a difficulty when the level of sparsity varies due to different situations [111]. Although researchers proposed many algorithms to solve the decomposition problem, such as method of optimal directions (MOD) or K-SVD algorithm [17] [110], there are still certain limits when implementing SC into real speech separation [111]. Another point worth noting is that there are possibilities of occurring negative numbers in both features and weights. It is impossible to be interpreted by real audio spectrogram’s components since all elements are with non-negative values and the mixture spectrogram is supposed to be positive summation with each component [21] [82].

2.4.4.2 Non-negative matrix factorization

Likewise, NMF is another type of matrix decomposition method. The key-difference between NMF and SC is that there assumes to be no negative numbers for all present elements during the decomposition [21].

As originally proposed by Lee et al [100], NMF is designed to decompose any original target spectrogram $S$ into two parts, namely, the basis matrix $W$ and the weight matrix $H$ as follows:

$$S \approx W \ast H$$ (2.40)

To realize such a procedure, multiple cost functions and optimization approaches have been proposed [100]. Without losing generality, this section concentrates on a particular but most common cost-function, ‘Kullback-Leibler divergence’ (KL divergence) and ‘multiplicative optimization’ in follows:

KL divergence:

$$D(S||WH) = S \ast log(S/W \ast H) - S + W \ast H$$ (2.41)

Multiplicative optimization:
\[
H_{kn} \leftarrow H_{kn} \ast \frac{\sum_m W_{mk} \left( S_{mn} / (WH)_{mn} \right)}{\sum_m W_{mk}} \tag{2.42}
\]
\[
W_{mk} \leftarrow W_{mk} \ast \frac{\sum_n H_{kn} \left( S_{mn} / (WH)_{mn} \right)}{\sum_n H_{kn}} \tag{2.43}
\]

Here \( k \) represents the number of feature vectors in the basis matrix, which can be designed accountable for different application-considerations \([3] [10] [12] [17] [24] [28] [33] [44] [52] [100]\). In this thesis, operator \( \ast \) and \( / \) stand for element-wise multiplication and division respectively. Subscripts \( m \) and \( n \) represent the index of frequency and time-frame, respectively, in the spectrogram. In terms of representation for all matrix factorization, this thesis use subscripts in order to simplify the representation with complex matrix product as parameters.

To use NMF for speech separation, it follows a similar scheme which is described in Section 2.4.3.1. In other words, once the decomposition is finished based on (2.42) and (2.43), (2.39) can be implemented to get the target source’s spectrogram, thereby followed by multiplying the phase matrix from mixture signal in order to finalise the reconstruction on target source’s signal \([12] [14] [22]\).

Obviously, the major part of NMF’s advantages is based on the non-negativity. Since the features are never negative, each feature vector is highly similar with the real elements in the audio spectrogram, such as the spectrogram of certain instrument’s scale or specific speaker’s phoneme. This enhances the interpretability of NMF’s features \([46] [47]\). Moreover, since all elements are positive, the summation between different elements’ spectrogram is a positive summation, which is more reasonable corresponding to the real life \([98]\). Besides, due to some optimization algorithm, such as the multiplicative optimization, the update of NMF’s elements is in parallel, thereby extremely reducing the computation time \([100]\).

On the contrary, there is no sparsity constraint in classical NMF, which means the level of overlapping between features could be very high. This will cause a serious problem that some information in the target spectrogram could be leaked out due to being mistaken as the part of overlapping features’ activities from other sources \([17] [61] [82]\).

Due to the success of NMF-associated methods on the audio separation problem \([3] [10] [12] [17] [24] [28] [33] [44] [52] [100]\), this thesis chooses NMF as the main approach to solve the speech separation problem. As the following section and the Chapter review show, NMF is a more expandable framework, since it can be
fused with keys from other approaches, such as the sparsity constraints from SC, or neural network as a pre-processing part. Thus, it clarifies this thesis focusing on NMF is a very reasonable decision in order to handle the speech separation problem.

### 2.5 Baseline methods of thesis

As the descriptions in Section 2.4.3.2, NMF is a potential framework to merge the advantages of other approaches and build a comprehensive system for solving the speech separation problem. This section thus reviews some recent NMF-based speech separation frameworks, which also sets up the baseline methods used to compare with this thesis’s contributions.

#### 2.5.1 Sparsity & discriminative constraints for single-channel NMF

This section starts by describing different constraints used in the cost function used in to solve the optimisation problem in single channel NMF. As discussed in Section 2.4.3.1, sparsity constraints benefit the separation performance by forcing features to be more orthogonal and thus less overlapped with each other [17] [61] [82]. Similar to SC, NMF has a similar structure except the non-negativity constraints. Meanwhile, in order to get rid of the computationally-consuming update in SC, it would be better to keep the optimization operations, such as multiplicative optimization the same. This requires relaxation of the original $L_0$ norm sparsity constraints into a continuous and derivable term [21].

Fortunately, researchers proved the L-1 norm is an effective replacement for the L-0 norm as sparsity constraints [18] [110]. The $L_1$ norm is continuous and almost-completely derivable except at the origin point [18]. Corresponding to the error distance between the reconstruction matrix and original data matrix, the origin point implies there is absolutely no error, which is not the common case. Thus, to use $L_1$ norm as sparsity constraints in the practical scenario is reliable.

Then, for sparse NMF (SNMF), it is originally proposed as follows [12] [21] [67] [82] [83]:

$$D(S∥WH) = S \ast log(S/WH) - S + W \ast H + \lambda \ast \|H\|$$

(2.44)
Here the $\lambda$ controls the level of sparseness. Because of the sparseness factor, SNMF usually has less active basis vectors during the decomposition of each spectrogram sample. This actually imposes regularization on the calculation, which also reduces the overlapping between active basis vectors and therefore benefits the separation [21]. In [21], the multiplicative optimization is correspondingly changed into:

$$H_{kn} \leftarrow H_{kn} \cdot \frac{\sum_m W_{mk}(S_{mn}/(WH)_{mn})}{\sum_m W_{mk} + \lambda}$$

$$W_{mk} \leftarrow W_{mk} \cdot \frac{\sum_n H_{kn}(S_{mn}/(WH)_{mn}) + W_{mk} \sum_m W_{mk}}{\sum_n H_{kn} (1 + W_{mk} \sum_m W_{mk} S_{mn}/(WH)_{mn})}$$

Meanwhile, in most NMF approaches, the basis matrix is supposed to be column-wisely normalized, thus maintaining every feature vector with an equal energy level [100]. This purpose can be simply achieved by the following operations:

$$W_{mk} = W_{mk} / \sum_m W_{mk}$$

$$H_{kn} = H_{kn} * (\sum_m W_{mk})$$

However, it is argued the effectiveness of merely sparsity constraints is not enough [79] [80]. This is obvious when different speakers (assumed to be different sources) speak similar utterances at the similar time, which leads to the large overlapped part between the spectrograms of different sources [79].

Discriminative learning has existed for a while, where the most common used is to enhance the classification performance [78] [79] [80] [81]. Generally speaking, discriminative learning is to reduce the overlapping of features from different sources, thereby easing the classification [80].

One typical measurement of overlapping is the redundancy between input vectors. To calculate the level of redundancy between features from different sources, Grais et al. proposes to compute the coherence between different source’s features. Without losing generalizability, here it is assumed a 2-source mixture scenario. Consequently, there will be two feature matrices: $W_1$ & $W_2$ corresponding to source 1 and source 2, respectively. Then the simplified cross-coherence is [79]:

$$coh(W_1, W_2) = \sum_{k_1} \sum_{k_2} W_{1k_1} \cdot W_{2k_2}, \quad k_1, k_2 \in K$$

In this equation, $k_1, k_2$ are the indices of two sources’ feature matrices (basis matrices), respectively. Operator $\cdot$ stands for the dot-product. Due to the non-negativity, the minimum of $coh$ will only be 0 when any basis vector from $W_1$ is
completely orthogonal to any basis from $W_2$ [79]. Next, they argue to change the original separated source-wise divergence into a connected divergence including both sources:

$$D_{mix} = D_1(S_1 || (WH)_1) + \alpha D_2(S_2 || (WH)_2) + \lambda \text{coh}(W_1, W_2)$$

(2.50)

where $S_1$ and $S_2$ are the spectrogram from training sets corresponding to two sources, respectively. The $\alpha$ & $\lambda$ are hyper-parameters, which are set based on experimental optimal values [79]. These two parameters control the level of importance corresponding to target source, the remaining source(s) and the cross-coherence [79].

Then, the optimization can be presented as the follows:

$$W_{1mk} = W_{1mk} \ast \frac{\sum_n H_{1kn} S_{1mn} / ((WH)_{1mN})}{\sum_n H_{1kn} + \lambda \sum_k W_{2mk}}$$

(2.51)

Likewise, the $W_{2mk}$ computation is symmetric on index $i$ (1 or 2) with the computation for $W_{1mk}$. The updating criterion for matrix $H$ stays the same, since the derivation won’t be affected by the changed divergence. After training, these trained basis matrices can be operated similar to (2.38) and (2.39) to realize the separation.

The main purpose is to introduce the cross-coherence parameter into divergence, thus punishing the features with high similarities but from two sources during the training phase. These enhanced features can improve the separation performance, even with the same separation stage for standard single-channel NMF [84].

Another interesting proposal for recent discriminative NMF algorithm is proposed by Wang et al. [80]. In this paper, the authors propose a different divergence function, which is the combination of decomposition errors from target-source-spectrogram-only divergence and mixture-spectrogram divergence plus sparsity constraints [80]. The following are each part of the new divergence:

$$\text{div}_1 = \sum_{m_1n_1} \text{div}(S_{m_1n_1}, (WH)_{11})$$

(2.52)

$$\text{div}_2 = \sum_{m_1n_1m_2n_2} \text{div}(S_{m_1n_1m_2n_2}, (WH)_{12})$$

(2.53)

$$\text{div}_3 = Tr[H]$$

(2.54)

Here (2.52) is about training on clean target speech, while (2.53) is about combining basis matrices from two sources to train on mixtures. (2.54) is the sparsity constraints. Due to the imbalance between the number of training utterances for target-only and number of training utterances for joint mixtures, the authors propose to use the following equation as the complete training divergence:
\[ \text{div} = U \ast \text{div}_1 + \text{div}_2 + \lambda \ast \alpha \ast \text{div}_3 \] 

where \( U \) is the number of utterances for each source (speaker). Besides, \( \alpha = U \ast SP \) where \( SP \) is the number of source in total.

Since these two discriminative NMF methods are extendable to other methods, and these proposals are most recent, they are chosen as the baseline methods to compare with the proposed single channel NMF. Obviously, both of them introduce a hyper-parameter into the optimization function, which is largely dependent on the SNR situation of the input mixture. In their papers, training utterances are all normalized into the same energy level. The final hyper-parameters are likely not robust enough, especially when the real mixture during separation stage is with a frequently-changing SNR [85]. These disadvantages of the baseline approaches are verified in the following sections.

2.5.2 Multi-channel NMF (MNMF) with utilizing spatial information

As the previous discussion in Section 2.4.1, spatial information is often very useful in the speech separation problem. Accurate spatial information benefits the separation in various aspects, bring an explicit feature for clustering and separation [38] [68]. Yet, the spatial information usually comes from processing signals from multiple channels where the differences among these simultaneously-recorded signals give clues about source spatial information [96]. Unfortunately, major matrix-decomposition approaches in the current mainstream of source separation tend to assume the signal is only recorded by one channel [4] [7] [47] [65] [66] [67] [79] [80] [83]. When multiple microphones exist, these models prefer to be performed on each channel, followed by being averaged in terms of initial results. In other words, it cannot use the full strength of multichannel signal, since it does not include the similarity between channel-wise signals into the data model and its optimization. On the other hand, independent-component-analysis-like (ICA) methods have their own inherent disadvantages, where the sources must be assumed as independent component and permutation problems must be solved by certain techniques [70] [96] [97] [98].

Therefore, a reasonable extension of the single channel model into multichannel utilization would bring a potential chance for the improvement in separation
performance. In this thesis, several recently emerging methods, namely multichannel-NMF-associated (MNMF) methods, have been studied [87] [88] [89] [90] [91] [94] [95]. The following contents cover the major part of latest MNMF existing researches.

For the purpose of merging spatial information into classical NMF-type methods, the very first step is to decide what kind of spatial information should be fused. In Section 2.4.1, there are two types of spatial information, including DOA and delay between simultaneous signals from different channels. Since the earliest MNMF model is based on delay, the following discussion will begin with spatial information [87] [90].

A. Ozerov proposed the first MNMF model under the assumption that speech signal is a complex Gaussian mixture model [90] [91] as following:

\[ S_{i,mn} = \sum_{k \in k_i} G_{k,mn} \]  
\[ G_{k,mn} \sim \mathcal{N}_c(0, W_{mk} H_{kn}) \]  

where the sub-scripts \( i, k, m, n \) are still source index, basis index, frequency index and frame index. \( \mathcal{N}_c(\cdot) \) denotes the distribution is complex normal distribution. Obviously, the biggest limit of this model is the assumption may not always be reliable. Since the real part and imaginary part of a speech has a relation based on phase in Fourier transform (same transform as the paper), in practice recordings may not always obey this assumption.

On the contrary, H. Sawada et al. propose another approach to obtain information [87] [89]. In the classical microphone-array signal processing, the cross-correlation matrix is a very common feature to deal with for spatial information [96]. Hence, the authors propose to use the cross-correlation matrix from each channel’s signal as the new input for NMF, which is as the following [87] [89]:

\[ S = \begin{bmatrix} |S_1| & \cdots & |S_1S_c|^{1/2} \text{sign}(S_1S_c^*) \\ \vdots & \ddots & \vdots \\ |S_cS_1|^{1/2} \text{sign}(S_cS_1^*) & \cdots & |S_c| \end{bmatrix} \]  

Here, \( S_c \) is the \( c \)-th channel signal, sub-script \( * \) denotes the complex conjugate. Besides, the operator \( \text{sign}(\cdot) \) stands for the phase:

\[ \text{sign}(S) = \frac{S}{|S|} \]
(2.58) represents one sample from this cross-correlation-spectrogram. For the original one TF-point, it is replaced as a $C \times C$ entry matrix ($C$ is the number of channels) as being shown in (60). Elements on main diagonal are the amplitude of corresponding $c$-th channel received signal, with the elements of the off main diagonal as the multiplication between the row-index-channel signal and column-index-channel signal at the corresponding TF index [87]. Clearly, the delay information between channels is stored in the elements of the main diagonal.

Assuming this new ‘spectrogram’ has a complex Gaussian distribution, then with the original basis matrix $W$ and corresponding weight matrix $H$, the authors brings another $C \times C$ Hermitian positive-semidefinite matrix $O$ as following:

$$P(S|\theta) = \prod_{c_1=1}^{C} \prod_{c_2=1}^{C} \mathcal{N}_c([S]_{c_1c_2} | \sum_{k=1}^{K} [O]_{c_1c_2} W_{mk} H_{kn}, 1)$$

(2.60)

More specifically, it can be expressed as the following calculable form:

$$P(S|\theta) \propto \exp\left(- \left\| S - \sum_{k=1}^{K} O_k W_k H_k \right\|_{Fro}^2 \right)$$

(2.61)

Here the mathematical expression for $O$, $W$, $H$ has been abbreviated. Calculation $\| \cdot \|_{Fro}^2$ denotes the Frobenius norm. The $\theta$ stands for all factors composing the distribution, including $O$, $W$ and $H$. Moreover, in this thesis, $O$ stands for the orientation matrix, which records the relations between signals from different channels. The reliability of this probability distribution modelling the original NMF has been proven in the literature [87] [89].

Yet, this is still not clear how it can be used for separation. In fact, researchers consider every basis vector as one complete speech component, which means one basis vector’s activities only belongs to one source [87]. This is not the absolute truth since the speech from other source might have very similar contents compared with the speech from target source.

However, the central point of this MNMF [87] [89] is to use basis vectors as the samples to compose the mixture, rather than naive TF bins. Generally speaking, this is highly related to the hidden class problem. With the assumption of nearly WDO, one TF bin’s ‘cross-correlation spectrogram’ is supposed to represent the spatial information corresponding to one source. Thus the orientation of one ‘cross-correlation spectrogram’ should be corresponding to only one source, implying that this orientation implies the probability of one frequency point belongs to one certain
source. On the purpose to connect basis-contribution to one certain source and the probability from orientation, the authors propose to introduce a new factor $V$ to represent the portion of one basis’s activity contributing to one source [87] [89]:

$$P(S|W,H,O,V) \propto \prod_{m,n} \exp\left( - \frac{\| S_{mn} - \sum_{k,l} O_{mi} V_{ki} W_{mk} H_{kn} \|_{Fro}^2 }{2} \right)$$  \hspace{1cm} (2.62)

In order to avoid random scale problem, $O$, $V$ also require to be normalized to scale as 1. It is worth noting that here

To summarize the implementation of whole optimization, the steps can be expressed as the following equations [87]:

\begin{align*}
W_{mk} &= W_{mk,*} \left[ 1 + \frac{\sum_n H_{kn} \sum_i V_{ki,*} \text{Tr}[E_{mn} O_{mi}]}{\sum_n H_{kn} \hat{a}_{mn}} \right] \hspace{1cm} (2.63) \\
H_{kn} &= H_{kn,*} \left[ 1 + \frac{\sum_m W_{mk} \sum_i V_{ki,*} \text{Tr}[E_{mn} O_{mi}]}{\sum_m W_{mk} \hat{a}_{mn}} \right] \hspace{1cm} (2.64) \\
V_{ki} &= V_{ki,*} \sum_{m,n} W_{mk} H_{kn} [\hat{a}_{mn} + \text{Tr}[E_{mn} O_{mi}]] \hspace{1cm} (2.65) \\
O_{mi} &= O_{mi,*} \sum_k V_{ki} W_{mk} \sum_n \hat{a}_{mn} H_{kn} + \sum_k V_{ki} W_{mk} \sum_n E_{mn} H_{kn} \hspace{1cm} (2.66)
\end{align*}

Here, the factor $\hat{a}_{mn}$ denotes the estimated TF-bin spectrogram value:

$$\hat{a}_{mn} = \sum_k W_{mk} H_{kn} = \sum_{k,l} V_{ki} W_{mk} H_{kn}$$ \hspace{1cm} (2.67)

and $E_{mn}$ represents the error matrix between original data matrix and the reconstruction matrix:

$$E_{mn} = S_{mn} - \sum_{k,l} O_{mi} V_{ki} W_{mk} H_{kn}$$ \hspace{1cm} (2.68)

Furthermore, to normalize factor $O_{mi}$, the following steps need to be implemented:

$$O_{mi} = \frac{O_{mi}}{\|O_{mi}\|}$$ \hspace{1cm} (2.69)

Likewise, $V_{ki}$ and $W_{mk}$ need a similar normalization as well [87]. It is easy to find that $\sum_i V_{ki} = 1$, which means the summation on probability of one certain basis vector belonging to one certain source is equal to 1, same as the requirement in classical mixture models.

To fulfil the separation, the last step is to use the factors build a soft spectrogram mask:

$$S_{i,m,n} = \frac{V_{ki} W_{mk} H_{kn}}{\sum_i V_{ki} W_{mk} H_{kn}} * S_{mn}$$ \hspace{1cm} (2.70)
Besides, a similar system with Itakura-Saito (IS) divergence or squared Euclidean distance as the loss function can be derived as in [89].

However, the whole system requires the number of active sources as the premise information. Meanwhile, it is clear that the disadvantages of WDO or nearly WDO assumption also exist here. Additionally, the methods need orientation matrix $O$ to be a Hermitian-positive-semidefinite matrix. In practical scenario, authors suggest using eigenvector decomposition (EVD) followed by rectifying negative eigenvalue into extremely-small positive number [89]. Although this is shown acceptable in cross validation experiments, there is no theoretical proof on whether robustness is reliable or possible detrimental effects for performance [89].

To tackle the aforementioned source number problem, the following work of the same authors propose several new methods, including using non-negative tensor factorization (NTF) to replace MNMF and reduce processing for large data [10] [88] [94] or using (2.61) but with certain clustering methods on the orientation matrix in order to realize flexible source-number and separation [89]. Although these proposals achieve promising outcomes, none of them ultimately solve the problem of initialization except performing random initialisation [87] [88] [89] [90] [91] [94]. This is because the orientation matrix is not a real-existing description in real life, frequently changing due to the microphone-array structure changing. However, the initialization of clustering-associated methods usually plays a crucial role in the final performance [89]. Thus, the separation performance is believed to not be very stable.

As a consequence, J. Nikunen et al. come up with a new approach, where they try to merge DOA information into Sawada’s basic MNMF framework [95]. As the aforementioned sections, DOA can be estimated based on the TDOA approach (2.16) (2.17). Given the $\tau_{c_i|c_j}$ as the time difference between $i$-th microphone and $j$-th microphone, the authors define a $C \times C$ dimensional delay matrix [95]:

$$[O_{m}]_{c_i|c_j} = \exp(j2\pi f_m \tau_{c_i|c_j}(ang_i))$$  \hspace{1cm} (2.71)

where $ang_i$ is the $i$-th angle form a previously-defined direction vectors matrix, sampling on the unit sphere, and $\tau_{c_i|c_j}(ang_i)$ is the time-delay calculated by inverse computation based on (2.16) (2.17). Then, similar to the basic MNMF method, the authors have a same structure for the complete probability modelling as (2.62), except replacing the orientation matrix to the delay matrix, where every element inside it is computed based on the geographic relation between certain pair of
microphones and corresponding delay as (2.17). Concretely, as for the parameter calculations, their methods need following steps:

\[ W_{mk} = W_{mk} \cdot \left[ 1 + \frac{\sum_{l,n} V_{kli} H_{kn} \text{Tr}[E_{mn} O_{ml}]}{\sum_{l,n} V_{kli} H_{kn} a_{mn}} \right] \quad (2.72) \]

\[ H_{kn} = H_{kn} \cdot \left[ 1 + \frac{\sum_{l,m} V_{ki} W_{mk} \text{Tr}[E_{mn} O_{ml}]}{\sum_{l,m} V_{ki} W_{mk} a_{mn}} \right] \quad (2.73) \]

\[ V_{ki} = V_{ki} \cdot \left[ 1 + \frac{\sum_{m,n} W_{mk} H_{kn} \text{Tr}[E_{mn} O_{ml}]}{\sum_{m,n} W_{mk} H_{kn} a_{mn}} \right] \quad (2.74) \]

\[ O_{mi} = O_{mi} \cdot \left[ \sum_{k,n} V_{kli} W_{mk} H_{kn} a_{mn} + \sum_{k,n} V_{kli} W_{mk} H_{kn} E_{mn} \right] \quad (2.75) \]

As for the separation stage, the authors propose to use the DOA kernel clustering, which can be computed as following:

\[ K_{mk} = \sum_{i=1}^{I} O_{mi} V_{kli} \quad (2.76) \]

Obviously, thanks to pre-defined direction vector matrix in (73), this method has a better initialization than the original MNMF. However, the weaknesses are also apparent: since the DOA kernel is based on sampling, the trade-off between sampling resolution and accuracy should be taken into serious consideration. Since there is no guarantee about the environmental factors, such as reverberation and noise, the choice of resolution is hard to decide. With an improper resolution, the methods might cause either information-loss or under-estimated noise.

To conclude, because of the cross-correlation mechanism, both methods expand the original data into \([M*N*C*C]\) dimensional space, which triggers a huge computation price. Meanwhile, both methods require the condition of spatial matrix \(O\) to be a Hermitian-positive-semidefinite matrix, thereby bringing a huge risk in terms of errors in spatial information calculation.

2.5.3 Other extensions of NMF for source separation

Considering that the Fourier transform is based on a linear-located basis \(e^{-j2\pi kn/N}\), there has always been a debate whether a better transform can replace pure Fourier transform, with expectations to capture more useful information from signals. Basically, it is reasonable to believe that to study other Fourier-family members and non-linearly located frequency basis is beneficial.
As it discussed in Section 2.3.3, sub-band models are composed by different sub-band filters to transform a signal into the frequency domain, where there could be more unique representations after the transform. Therefore, a common but inspiring extension is to replace the STFT with certain sub-band model transforms [2] [13] [26] [28] [39] [41] [43] [45] [55] [58] [61] [62] [63]. The study shows an improved performance under multiple evaluation tests. Due to the scope of this thesis, the details will not be elaborated further here but could be investigated in future work.

Moreover, specific decomposition techniques, such as enforcing sparsity [40] [76] [77], side information [23] [24] [35] [55] [74], or basis learning and update [22] [40] [77], deep-learning-associated techniques [3] [5] [6] [28] [29] [31] [32] and certain time-series modelling [12] [14] [15] [16] [25] [36] [37] [38] [44] [48] [50] [51] [52] [53] [54] [56] [69] [73] have demonstrated their success in speech separation, which can be integrated with the basic NMF framework as a probability interface.

### 2.6 Main evaluation metrics

In general, the evaluation of speech separation can be divided into two groups: one focuses on speech quality evaluations and the other focuses on speech intelligibility evaluations [42] [72] [101] [102]. While the former ensures that the level of distortion from separation processing remains at an acceptable level, the latter highlights the importance of speech signals being understandable after the separation processing. This section presents the major evaluation methods that this thesis uses.

Specifically, speech quality evaluation includes objective evaluation and subjective evaluation. For objective evaluation, the Signal-to-Interference Ratio, Signal-to-Distortion Ratio, Signal-to-Noise Ratio and Signal-to-Artefacts Ratio measures are used using the implementations of the ‘BSS eval toolbox’ [72], where the segmental option is used to determine the final values. More precisely, considering $\hat{s}(n)$, namely the estimation of target-source speech in the discrete time domain with the frame number as $n$, after separation processing, it can be commonly described as:

$$\hat{s}(n) = s_{\text{tar}}(n) + s_{\text{noi}}(n) + s_{\text{art}}(n) + s_{\text{int}}(n)$$

(2.77)

where the separation $\hat{s}(n)$ is supposed to be equal as the summation of the ground-truth $s_{\text{tar}}(n)$ corresponding to the target-source speech, background noise-source $s_{\text{noi}}(n)$, artifacts’ part $s_{\text{art}}(n)$ and interference-source part $s_{\text{int}}(n)$. While the last
three parts result from the separation processing, they are supposed to be as small as possible to get a reasonable estimation on the target speech. Correspondingly, the ratio between distortion and source speech is defined as the SDR:

$$ SDR = 10 \log_{10} \frac{\|s_{\text{tar}}\|^2}{\|s_{\text{noi}} + s_{\text{art}} + s_{\text{int}}\|^2} $$ (2.78)

The ratio corresponding to interference against source speech, namely SIR is defined as:

$$ SIR = 10 \log_{10} \frac{\|s_{\text{tar}}\|^2}{\|s_{\text{int}}\|^2} $$ (2.79)

Likewise, the ratio between artefacts with source speech, called as SAR, is defined as:

$$ SAR = 10 \log_{10} \frac{\|s_{\text{tar}}\|^2}{\|s_{\text{art}}\|^2} $$ (2.80)

Similarly, the SNR, Signal-to-Noise Ratio, is defined as:

$$ SNR = 10 \log_{10} \frac{\|s_{\text{tar}}\|^2}{\|s_{\text{noi}}\|^2} $$ (2.81)

Meanwhile, the Perceptual-Evaluation-of-Speech-Quality (PESQ) evaluation is also introduced to compare the separated target signals to the original source signal in terms of subjective quality [101] [102]. Generally speaking, PESQ simulates a subjective assessment of the separation result, which is strongly related to the Mean-Opinion-Score (MOS). With the clean target-speech as the ground-truth, PESQ includes a series of operations, such as time-alignment and auditory transform on the estimated target-speech. The range of PESQ is from 0 to 4.5, with higher scores corresponding to better results in terms of subjective evaluation. Since PESQ system is too complex and beyond the scope, this thesis will not expatiate on its details. It is also worth noting that the PESQ scores in this thesis have been measured and converted into MOS based on the Matlab toolbox in [117]. The other subjective metrics is mainly involved with listening tests with participants. Due to the limit of resources, this thesis will not include this test but to distribute possible ones in the future works.

Despite all the speech quality evaluation, speech intelligibility evaluation is also of importance on separation task. In general, speech intelligibility indicates the level of comprehensibility of the speech after being processed. Although there is no distributed noise in this experiment, the interferences and artefacts could still extremely undermine speech intelligibility. Algorithms which achieve a low
intelligibility score are not very useful for many speech communication applications. In the experiments of this thesis, the Short-Time Objective Intelligibility system (STOI) is used to evaluate the intelligibility of the separated speech sources [42].

2.7 Chapter conclusion

This Chapter describes the details of a common speech separation system, including the system-target signals, hardware devices, signal-transform methods, different separation approaches and related evaluation metrics.

As discussed in the previous sections, the NMF model is chosen as the main basis for the methods proposed by this thesis, due to its recent successes in speech processing and excellent extensibility. The associated contents in Section 2.4 and 2.5 are highly related to the remaining parts of this thesis, where this thesis main contributions include a single-channel NMF model enhanced by IBM features and a multi-channel NMF model collaborated with DOA information are presented, respectively.
This Chapter introduces one of the thesis contributions, which is to integrate structural information of the spectrogram into the classic NMF for the single channel speech separation problem [85]. Section 3.1 presents the theoretical derivation, and is followed by results and analysis with conclusions in Sections 3.2.2, 3.2.3 and 3.2.4, respectively.

### 3.1 Methodology of the MASK-NMF Approach

As discussed in Section 2.5.1, the potential improvement over the single channel NMF separation framework is to bring the different constraints into the original framework, so that the source-wise features, namely basis vectors corresponding to different sources can be more distinguishable. Consequently, when only interference sources are active the possibility of having a feature from the target source being active is reduced.

Generally, the $L_1$ norm sparsity constraints are used to ensure features are distinct from each other. Consequently, source-wise features will be more likely to correspond to one source only rather than multiple sources. This leads to lower errors during separation [21].

However, sparsity-type constraints do not bring actual benefits to speech separation. From the original SC optimization [18] [110], the constraints tend to punish the basis vectors with smaller weightings in the NMF separation model of (2.36) until they diminish to zero, whilst increasing weight-coefficients with large values. For example, certain basis vectors might not be included in the final representation, mainly due to their weight-coefficients being relatively small during the initialization phased compared with other basis vectors. Yet, with respect to speech separation, this penalty on small values of the weights might not be appropriate. Typically, the initial weight matrix is chosen randomly. Considering the above rationale, the optimization may place more emphasis on weights in the matrix with large values,
which may correspond to features from both interfering sources as well as the target source [79].

Although discriminative NMF has reduced the errors resulting from similar features by reducing the similarity between features, this usually comes at the price of more training data and reduced robustness when facing different environmental conditions. In particular, the performance in terms of intelligibility may be even worse than the original mixed signals [79] [80].

On the other hand, spectral mask techniques have a long history of use for speech enhancement applications. The approaches described in Section 2.4.2 usually have a satisfactory outcome, especially for improving intelligibility. Hence, the potential fusion of mask structures with classic NMF is very likely beneficial for the whole system performance. Specifically, the decomposition and subsequent separation might be inaccurate when the same features are used to model more than one source within the NMF model. This may lead to inaccurate spectral masks derived through post-processing of the separated source spectrograms when compared to the ideal binary masks for each source.

In the case of an IBM, it is assumed that one TF bin belongs to either the target source or the other undesired sources. This implies that a sparsity assumption is used in deriving the mask[68] [100]. Thus, if the mask is chosen as an extension to the classic NMF work, it can be operated simultaneously with a sparsity constraint. The approach proposed here is largely inspired by the work of Q. Zhang et al. [81], on discriminative KSVD (DKSVD) for the SC problem. In brief, they pursue a classifier which is based on the coefficients after the optimization of the SC model, and use this to classify the real data set based on the trained model [81].

Typical classification only assumes two possible values as the outcome, namely belonging-to or not-belonging-to the class, which is similar to the IBM where the values of the mask are either 1 or 0. Yet, the difference between a mask structure and a classifier is that the number of entries in the mask structure follows the number of entries in the spectrogram itself, rather than a binary classifier. Fortunately, the IBM entry’s value is either 0 or 1; if the target has a mask value at one TF bin of 1, then the undesired source’ mask will be 0. Consequently, the scale of the whole mask information set is fixed, which prevents potential detrimental scale-variation problems during parameter updates [81].
Therefore, to merge the spectrogram structure information with mask values, the mask matrices of both the target source and undesired source are introduced into the single channel NMF model. Following the similar idea from DKSVD, the system replaces the classifier of DKSVD with the spectrogram mask. For the convenience of parameter calculation, KL divergence remains as the measurement of distance between the true mask with the estimated mask based on NMF’s weight matrices. The training thus aims at obtaining a proper map for rectifying errors in mask estimation. Figure 3.1 gives a direct display on the training stage of the proposed method.

Assume there are two speakers participating in speech mixture. As the figure showing, during training stage, clean-speech spectrogram of speaker 1 (Spectrogram1) and speaker 2 (Spectrogram2) are used to derive IBM information (Mask1, Mask2) and corresponding spectrogram features (W1, W2). Then the speech mixture is simulated by summing two clean speeches into together. Following these pre-processing, the mixture, IBM and spectrogram features are used as the known information for proposed NMF (pNMF), with the expectation to extract mask features (W_M1, W_M2) corresponding to both mixture spectrogram and source-wise IBM information.

After the training-stage described by Figure 3.1, the spectrogram features will be used again during separation stage, where weight information (H1, H2) corresponding to the test mixtures are obtained. The trained mask features are then
supposed to recover correct source-wise binary mask by multiplying corresponding weight information.

Details of algorithms will be elaborated in the following sections.

3.2 NMF post-processing based on spectral masks

Firstly, based on the relative sparsity of different sources in the STFT domain, separation to estimate the $i$-th source $S_{imn}$ can be achieved by a spectrogram mask [8]:

$$S_i(m,n) \approx M_i(m,n) \cdot S_{mix}(m,n)$$  \hspace{1cm} (3.1)

$$M_i(m,n) = \begin{cases} 
1 & P_i(m,n) \geq P_z(m,n) \\
0 & P_i(m,n) < P_z(m,n) 
\end{cases}$$  \hspace{1cm} (3.2)

$$M_z(m,n) = \begin{cases} 
1 & P_i(m,n) < P_z(m,n) \\
0 & P_i(m,n) \geq P_z(m,n) 
\end{cases}$$  \hspace{1cm} (3.3)

Here, $m$ is the label for frequency point, with $n$ as the frame number. The mask value of $(m,n)$ time-frequency (TF) bin $M_i(m,n)$ is based on a Boolean-decision between $i$-th source’s power $P_i(m,n)$ compared with the power summation of all the remaining $z$ sources in the mixture, namely $P_z(m,n)$ in this TF bin. Consequently, the spectrogram part corresponding to the target source, $S_i(m,n)$ is computed by target-source spectral mask $M_i(m,n)$ times mixture spectrogram $S_{mix}(m,n)$. This idea comprises the main part of the proposed approach. The spectrogram mask is referred to as $M_i$ corresponding to the $i$-th target source, with $M_z$ as the mask of all the remaining $z$ sources comprising the interference for the following utilization.

Precisely, the approach first obtains the optimal weight matrix according to the decomposition of the mixture spectrogram, and then maps these weight coefficients to the mask part to generate the correct mask corresponding to each source. More specifically, this problem can be modelled as follows:

$$\text{div}(S_{mix}|(W_{mix}H))$$

$$= S_{mix} \cdot \log(S_{mix}/W_{mix} \cdot H) - S_{mix} + W_{mix}H + \lambda \|H\|$$  \hspace{1cm} (3.4)

$$\text{div}(M_i|W_iH) = M_i \cdot \log(M_i/W_i \cdot H) - M_i + W_iH + \lambda \|H\|$$  \hspace{1cm} (3.5)

where $W_{mix}$ and $H$ are the basis matrix and corresponding weight matrix of mixture, respectively. $\lambda$ is a coefficient to decide the scale of sparsity constraints on weight matrix ($\|H\|$) participating in the complete divergence function of spectrogram part.
In (3.5), index 'I' stands for the union of all sources, including target sources (index as 'i' in (3.2)) and all interferences (index as 'z' in (3.3)). Similarly, regarding the mask part corresponding to a certain source in (3.5), the divergence $\text{div}(M_i || W_i H)$ is introduced to rectify the errors between weight matrix, $H$ and corresponding source's mask-basis matrix, $W_i$. Basically, $W_i$ is trained with the mask values, $M_i$ in (3.2) or (3.3) corresponding to a certain source, which will be elaborated upon in the following Section 3.1.2. Moreover, it is worth noting that the weight matrix, $H$, is the same one in both (3.4) and (3.5), which interactively connects two parts together with respect to both loss functions.

Based on the aforementioned ideas, the training requires a supervised target for the separation of the mixture. Without losing generality, the SNMF is chosen as the basic model. As it can be seen, (3.4) requires the decomposition to be accurate for the mixture spectrogram. Correspondingly, (3.5) requires an accurate estimation of the masks. According to the training data, it is possible to create mixtures between different speech sources with a completely known situation, where the correspondingly correct spectrogram masks can be obtained as the supervised targets. These supervised targets are then included with the original mixture spectrogram as input to the SNMF approach.

However, to implement this approach, the optimization has to be divided into two parts, including all updates of the parameters in (3.4) targeting the accuracy of the spectrogram mixing model and all updates on the parameters in (3.5) targeting accuracy of the spectrogram mask. Theoretically, there is no guarantee to reach the global optimal for the parameters in the whole framework [81]. Besides, since the loss function of the whole system is divided into two parts, the robustness for the updating procedure is not ensured to be optimal [81].

Thus, (3.4) and (3.5) can be merged into a joint form:

$$\text{div}_{\text{joint}} = \text{div}(S_{\text{mix}} || (W_{\text{mix}} H)) + \sum_{i=1}^{I} \text{div}(M_i || W_i H)$$

(3.6)

However, (3.6) is a hard problem in terms of the computational complexity. Thus, the following section elaborates upon a new structure for this joint problem.
3.2.1 Integrating spectral masks into SNMF

The first step is to initialize the basis matrices corresponding to different sources based on the coherence-constrained NMF separation scheme [79]. Then, the data to be modelled is formed as the cascade of the mixture spectrogram and the mask as:

\[ S = [\mu * \mathcal{M}_i; \mu * \mathcal{M}_z; S_{mix}] \]

and where \( \mathcal{M}_i \) and \( \mathcal{M}_z \) are derived from (3.2) and (3.3). The factor, \( \mu \), here is a trade-off factor, which will be discussed later.

Accordingly, the original basis matrix is modified as \( \mathcal{W} = [\mathcal{W}_i; \mathcal{W}_z; \mathcal{W}_{mix}] \), with the same row-dimension arrangement as for \( V \) and \( \alpha \) as the number of basis vectors. \( \mathcal{W}_{mix} \) can be preliminarily obtained by the training stage of the coherence-constrained NMF approach. \( \mathcal{W}_i \) and \( \mathcal{W}_z \) are the supervised correction parts for the post-processing of the mixture weight matrix after the decomposition of \( S_{mix} \) during separation. Mathematically, this can be represented as:

\[
\min ||H_{kn}|| \text{ s.t. } \begin{bmatrix} \mu * \mathcal{M}_i \\ \mu * \mathcal{M}_z \\ S_{mix} \end{bmatrix} = \begin{bmatrix} \mathcal{W}_i \\ \mathcal{W}_z \\ \mathcal{W}_{mix} \end{bmatrix} * H \tag{3.7}
\]

In the original paper describing discriminative K-SVD (DKSVD) [81], this equality is achieved with a K-SVD update. Compared with this, SNMF has its own multiplicative optimization in [21]. Different to [81], optimization of the proposed method is only implemented on \( \mathcal{W}_x, \mathcal{W}_z \) and \( H \) in this re-training step using (2.45) (2.46) to improve the reconstruction of the mixture spectrogram.

Compared with [79] [80], the proposed method focuses on creating a classifier during the training. This classifier uses the latent representation (weight matrix, \( H \)) as the input times the trained correction part in basis matrices \( \mathcal{W}_n \), \( \mathcal{W}_n \supset \mathcal{W}_i \cup \mathcal{W}_z \) in (3.7), to classify each time-frequency as belonging to the target or interfering signal based on the spectrogram masks.

3.2.2 Deriving the trade-off factor & re-normalization

The trade-off factor \( \mu \) needs to be designed to achieve a balance between the reconstruction and the correction. A smaller value of \( \mu \) places more emphasis on the reconstruction using \( \mathcal{W}_{mix} \) and \( H \), while a larger value of \( \mu \) places more emphasis on the classification of the corrective spectrogram masks during the optimisation of (2.44). Compared with the original DKSVD [81], our proposal is dedicated to speech source separation. Since speech signals are non-stationary, this thesis proposes to
adaptively vary $\mu$ for each time frame so that the correction part has a similar scale to the reconstruction part. This can be expressed as:

$$\mu_i(n) \approx \sum_m S_i(m,n) \frac{\sum_m M_i(m,n), (i \in I)}{\sum_m M_i(m,n), (i \in I)} \tag{3.8}$$

where $m, i$ represent the frequency axis of the mixture spectrogram and the source index, respectively, and $n$ represents the time-frame. As in [81], basis matrices for the reconstruction and correction parts of (3.7) are re-normalised following re-training such that they correspond to the same scale. For certain purpose, such as emphasizing accuracy of mask estimation or reducing errors in spectrogram decomposition, $\mu_i(n)$ can be adapted to be either larger or smaller, correspondingly.

### 3.2.3 Reconstruction and separation

Finally, in the separation, the proposed model uses only $W_{mix}$ as the basis matrix to decompose the target mixture $S_{mix}$ to a corresponding weight matrix $H$. Although the content of the target mixture and training mixture are not the same in most cases, a well-trained dictionary will mean that the same basis matrix $W_{mix}$ in training and separation provides a similar $H$ distribution. The product between trained $W_n$ and $H$ estimates the spectrogram masks in the target mixture, corresponding to the target and interference signals of the mixture. However, the final result of estimated mask parts after the computation of the separation stage is not perfect, which means the mask values after this processing may not be exactly 1 or 0. Thus, to obtain the separated source signals in the time domain, the dominance probability of the proposed scheme is used to get the spectral masks as also determined in [8]. Then, with these final masks, the spectrogram corresponding to different sources are estimated as [8] and retrieved back to the time domain using the phase of the input mixture signal as the phase of the separated target signal. The key steps of the algorithm are summarized in Table 3.1.
3.3 Experiments & results

This section describes the experimental methodology and results comparing the proposed approach (pNMF) with discriminative NMF (dNMF) [80], coherence-constraint NMF (cNMF) [79], original NMF (bNMF) [100].

3.3.1 Experimental set-up

The experiments focus on instantaneous mixtures of two sources from the ANDOSL speech database [115]. This dataset contains recordings of multiple speakers each
uttering the same 200 utterances and sampled at 16 kHz. A set of 8 speakers are chosen from this database. Then, different combinations of target and interference speakers are randomly chosen to create 5 different target-interference mixture sets. For each set, 20 utterances per speaker are randomly set aside for testing and the remaining 180 utterances per speaker are used for training (a total of 900 mixture signals for training).

As for the evaluation metrics, this section mainly uses the methods in Section 2.6. In regards to this experiment, it is worth noting that background noise is assumed to be zero and so the SNR is not evaluated. Conversely, despite artefact parts, the remaining differences between the estimated speech sources with the ground-truths are attributed to the interferences. Hence, SAR, SIR and SDR are used her for objective evaluation.

Since one purpose of the proposed method in this chapter is to improve the separation mask estimation and thus enhance separation performance, a plain model evaluating the Mask-Estimation Accuracy Ratio (MEAR) is designed as:

$$r_c = \frac{m_{err}}{m_{tot}}$$  \hspace{1cm} (3.9)

where $r_c$ is the accuracy ratio, with $m_{err}$ and $m_{tot}$ as the number of wrongly-estimated masks and total masks, respectively.

Specifically, $m_{err}$ and $m_{tot}$ are linked to the IBM in [99]. Corresponding to this experiment, each algorithm returns an estimation of target speech and an estimation of interference speech. These two estimations are respectively referred to as ‘speech’ and ‘noise’ in the mask-estimation scheme, followed by calculation of the IBM. Correspondingly, the masks based on the calculation above are defined as estimated-masks. On the contrary, a similar process is used but with the ground-truth target speech and interference speech, to obtain true-masks. Then, the number of estimated-masks with different values against true-masks is defined as $m_{err}$. The number of true-masks in total is defined as $m_{tot}$.

The point of this evaluation is to demonstrate the effectiveness of the correct IBM on speech separation, and the potential for improving separation performance. This will be elaborated in Section 3.2.4.

All results are shown with error bars representing 95% confidence intervals. For our training and test databases, a range of input SIR values (which in this section is equivalent to input segmental SNR since there are only two sources and no
background noise in the mixture) is designed, covering the range -15 dB to 15 dB in increments of 5 dB. For each input SIR value, there are 20 random mixtures for one target, and the final objective result comes from the average of 4 target speakers. For our algorithm, the signal is transformed into the TF domain through the STFT with a 0.02 s long Hanning window and incrementing every 0.01 s. As for the value of the trade-off factor we use (3.8) to define the exact balance between discriminative correction and the reconstruction errors. Each source-wise basis $W_n$ has 128 column vectors to gain enough representation ability, with the whole basis $W_m$ being comprised of 256 feature vectors. The sparsity parameter $\lambda$ in our algorithm is set as 0.01, which is set based on results from empirical testing. For the other algorithms (cNMF, dNMF), all configurations are adjusted based on [79] [80].

Additionally, the correct IBM has also been derived based on the ground-truth in order to show the effectiveness of the IBM for source separation problems, which is referred as ‘oNMF’ in the following figures.

### 3.3.2 Speech-quality experiment results & analysis

Table 3.2 indicates the results of SIR, SAR, SDR and PESQ, respectively. As shown, the best results are highlighted in bold within each test criterion. The SIR of the proposed approach is generally higher than the other methods. Compared with the second best performing method, our method gains an improvement of around 0.5-1 dB for low input SIR mixtures. Figure 3.4 indicates our proposed method obtains similar SDR results although not as high as the best performing method. Since SAR and SDR are usually opposite in trend to the SIR, it is suggested that the proposed method is better. This is because it has both a high SIR output and a relatively high SAR and SDR. Likewise, in corresponding PESQ tests, our proposed method mostly achieves the highest results compared to all other methods.

However, it is worth noting that all four separation methods bring huge artefacts, thus worsening the SAR and then SDR. Compared with these, the correct IBM derived from ground truth has no such shortage, which indicates the four separation methods still have disadvantages with respect to a correct and natural separation.

This is also verified by PESQ results, where the PESQ of the simulated mixture before any separation process has also been included in Figure 3.5. Contrary to the
Figure 3.2 Input SIR vs. output SIR

![InSIR vs. OutSIR](image)

Figure 3.3 Input SIR vs. Output SAR

![InSIR vs. OutSAR](image)

A steady improvement in output SIR of Figure 3.2, the four separation methods only present a slight advantage in terms of average PESQ values when the input SIR is around the range of -15 dB to 0 dB. When the input SIR is over 0 dB, the four separation methods result in reduced subjective quality of speech, with around a 0.1 to 0.3 reduction in PESQ scores.
3.3.3 Speech-intelligibility experiment results & analysis

In terms of STOI, the method from C. Taal et al [42] is used to obtain STOI scores, of which the range is from 0 to 1 with higher values as better results. In general, it can be found that all methods have similar scores. However, separation results with
Figure 3.6 Input SIR vs. Output STOI

low-input SIR (-15dB to 0dB) show a small advantage compared with the proposed method. On the other hand, the performance of proposed method for the remaining input SIR cases is similar to the other methods.

This is a similar trend to the speech subjective quality evaluations in Figure 3.5, despite two criteria not being strongly linked to each other. With analysis of the separation results, it is believed that the single-channel NMF features, namely basis vectors in the frequency-domain cannot recover speech elements correctly, especially when there are overlaps between the target speech and interference speech. Although all four algorithms ensure that the mathematical error is small enough after optimization, the separation still results in distortions, which mainly damages both quality and intelligibility.

Likewise, Figure 3.6 also shows the STOI when using the IBM and the unprocessed mixture, which presents an alike relation between the processed and unprocessed speech signals.
### Table 3.2 Input SIR vs. MEAR

<table>
<thead>
<tr>
<th>InSIR</th>
<th>bNMF</th>
<th>dNMF</th>
<th>cNMF</th>
<th>pNMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>-15</td>
<td>0.45 ± 0.020</td>
<td>0.45 ± 0.020</td>
<td>0.44 ± 0.020</td>
<td>0.42 ± 0.020</td>
</tr>
<tr>
<td>-10</td>
<td>0.36 ± 0.023</td>
<td>0.36 ± 0.022</td>
<td>0.36 ± 0.021</td>
<td>0.34 ± 0.022</td>
</tr>
<tr>
<td>-5</td>
<td>0.30 ± 0.020</td>
<td>0.30 ± 0.020</td>
<td>0.30 ± 0.019</td>
<td>0.28 ± 0.019</td>
</tr>
<tr>
<td>0</td>
<td>0.22 ± 0.014</td>
<td>0.22 ± 0.014</td>
<td>0.22 ± 0.013</td>
<td>0.20 ± 0.013</td>
</tr>
<tr>
<td>5</td>
<td>0.18 ± 0.017</td>
<td>0.18 ± 0.016</td>
<td>0.19 ± 0.016</td>
<td>0.17 ± 0.016</td>
</tr>
<tr>
<td>10</td>
<td>0.14 ± 0.018</td>
<td>0.14 ± 0.017</td>
<td>0.15 ± 0.017</td>
<td>0.13 ± 0.016</td>
</tr>
<tr>
<td>15</td>
<td>0.10 ± 0.007</td>
<td>0.10 ± 0.007</td>
<td>0.11 ± 0.006</td>
<td>0.08 ± 0.006</td>
</tr>
</tbody>
</table>

#### 3.3.4 Mask accuracy results (MEAR) & analysis

From the description in Section 3.2.1 and equation (3.21), the range of MEAR values can be found as 0 to 1. Likewise, Table 3.2 shows the output MEAR versus input SIR. Compared with other methods, the proposed approach shows a steady improvement.

The correct IBM (oNMF) has also been derived based on the ground-truth in order to show the effectivity of the IBM for source separation problems. The results of SDR, PESQ and STOI tests are shown in Figure 8.

Moreover, from the Figure 3.4, Figure 3.5, Figure 3.6 and Table 3.2, it is clear that the separation is far better when the applied IBM is derived based on the ground-truth (oNMF). This highlights the potential of integrating a mask-structure to enhance the separation, which is the proposed method’s main contribution for the NMF system. Certainly, with the MEAR of the proposed method and performance-differences between the masks estimated from the proposed method and the correct IBM, the method still requires improvement to obtain a better estimation of the separation masks.
Chapter conclusions

In this Chapter, a new scheme is derived to integrate a spectral separation mask with the coherence constrained NMF separation approach, which utilises information about the source-wise spectrogram masks derived during a training phase. After being evaluated across a database of two-source mixture signals, the performance of the proposed method shows a feasible advantage over baseline methods. Specifically, in terms of output SIR, the proposed method is about 2dB better than the best performing baseline methods while being able to achieve acceptable results for the output SAR, SDR, PESQ and STOI. Compared with baseline methods, the proposed method achieves more consistent results across all evaluation criteria. Meanwhile, the mask estimation results from the proposed method imply the estimation is more accurate than the baseline methods and indicates there is further potential to be explored for this method.

On the other hand, the proposed method still inherits the limitations of single-channel NMF-like methods, where the bases from the training phase are not effective to separate spectrogram-overlap between target speech and interference speech. This results in distortion after separation, undermining the quality and intelligibility of enhanced target speech. Moreover, it is clear that there is still a big gap between the performance of the proposed method and the oracle IBM. As the oracle IBM is effective for this separation task, to improve the proposed method a better mask-estimation is required.

To conclude, this Chapter has researched the advantages and disadvantages of approaches to single-channel NMF source separation with the proposed mask constraint. The next Chapter will thus focus on the multi-channel NMF framework for separating speech sources.
COMBINING DOA WITH MULTICHANNEL NMF FOR SEPARATION MASK ESTIMATION

This Chapter discusses the second part of this thesis contribution, where the single channel NMF separation approach is adapted into the multichannel NMF approach. With the advantages of being able to estimate the DOA, a new framework to combine DOA cues and multichannel NMF (DMNMF) is built, in order to provide more accurate and practical solutions for separation mask estimation.

4.1 Multichannel NMF informed by DOA estimates

Although single channel NMF and associated methods achieve very successful performance on speech separation problems, there are still many limitations of these methods. In Chapters 2 & Chapter 3, different training frameworks and constraints (sparsity, temporal constraints) have been discussed. For the purpose for the best separation performance, these methods require explicit data corresponding to speakers (speech sources), thus extracting useful features by off-line training. Given some typical applications, such as speech enhancement for personal smartphones or hand-free control of personal vehicles, such a speaker-dependent solution is feasible. However, for other cases, it may not be practical to obtain labelled training data covering a wider variety of speakers.

In contrast, multichannel approaches have achieved many successes for speech separation. Compared with single-channel approaches, multichannel approaches are capable of more diverse-information utilization, such as spatial information and joint processing of multiple channels. Specifically, Section 2.4.1 presents approaches using vast spatial information as the main tool for speech separation, based on either suppressing soundwaves from undesired directions or emphasising the parts with similar phase-delays on the mixture speech spectrogram. Given the complexity of speech-mixes in real life and thus the hardship of training source-wise features with perfect separation performance, approaches that utilise spatial information are significantly effective since they sidestep the necessity of training and obtaining perfect features[12].
Likewise, such spatial information can be also converted into matrix-decomposition models. Not only can spatial information be exploited during the training stage, but also the separation can take advantage of known spatial information.

Based on this idea, Section 2.5.2 mainly describes how MNMF is combined with spatial information derived from processing of microphone array recordings. Specifically, mixing-vectors and covariance matrices of each channel signals are obtained as algorithms replace the original spectrogram on single channel NMF with these and utilize their spatial information on multi-channel NMF; while the separation stage has cluster factors which represent the level of each feature belonging to certain sound sources, which formulates the following separation masks. These approaches provide different ways to construct a separation mask rather than just using the spectrogram from a single channel recording, thus effectively improving the separation performance.

One major shortcoming of these MNMF approaches occurs due to the use of an EVD for obtaining the spatial information. For example, (2.71) and (2.75) calculate and update the DOA kernel $O_{mi}$. During the update, EVD can be performed onto it as:

$$O_{mi} = EDE^H$$

(4.1)

where $E$ represents the eigenvectors of $O_{mi}$, with $D$ as corresponding eigenvalues. Since the base of NMF updates requires non-negative values, $O_{mi}$ has to be positive semidefinite. Considering possible bad-scaling situations, all the negative values in $D$ must be replaced by zeros or extremely small positive numbers, which is referred to as $\tilde{D}$. Consequently, the DOA kernel after rectification will be:

$$\tilde{O}_{mi} = E\tilde{D}E^H$$

(4.2)

Although replacing negative eigenvalues with zeros or very small positive numbers can guarantee smooth updates, it cannot ensure the robustness of the algorithms. In the meantime, most MNMF approaches require a lot of computations due to the hard initialization of spatial information.

However, among all common spatial factors, DOA can be easily derived using a range of methods, especially with the IDOA method (Section 2.2.2, [107]) where all computation of the DOA is basically with positive-value numbers (Section 2.4.1.1). Besides, since positions of speech sources are often different, obtaining DOA information actually offers reliable cluster labels, where data such as TF points with
different DOAs could be considered as the data from corresponding sources, respectively. This is indeed equal to a mask decision, which is the same idea as used in the DUET algorithm [64].

Compared with the techniques in Chapter 3, a mask derived from DOA information does not require any discriminative learning, which means the source-wise training is not essential. Moreover, the initialization of DOA could be inherited from any initial DOA estimation, which brings heuristic clues into the following part of the system, thus lessening the computation price for system initialization.

On the other hand, the following part, such as the MNMF framework, is also supposed to enhance DOA-based separation performance. This is because MNMF introduces the features into the estimation, rather than simple TF points. With features, the representation of each source should have less-overlap from the representations of the other sources, thereby improving the accuracy of DOA estimation. Furthermore, the features could benefit separation more with certain training datasets provided corresponding to different sources, which is similar as the scenario of Chapter 3.

The key problem of building a DMNMF framework is to model the combination between DOA and MNMF. Since these are different types of data, a proper joint-possibility could be the best choice for the combination.

As for the probabilistic modelling of DOA estimation, directional statistics using a von Mises distribution [92] [93] could be used. Generally, a von Mises distribution models the data-points on a circle of unit radius, which is commonly expressed as the differential angle between the data-points with the reference point, respectively. Regarding speech separation, these data-points could be referred to as TF points, where the differential angle could be the DOA with a reference chosen as 0 degrees. For the purpose of simplifying the mathematical derivation, this thesis only discusses DOA estimation for the two-dimensional cases. The method could be easily adapted on three-dimensional cases by changing two-dimensional von Mises distribution into three-dimensional one. Given the continuity of the von Mises distribution, the idea of using this joint DOA-MNMF can take this advantage and obtain more accurate DOA estimates.

Meanwhile, authors P. Smaragdis et al. proposed the equivalent model of NMF with regard to probability modelling, named as probabilistic latent component analysis
Figure 4.1 DMNMF separation system

(PLCA) [46]. The model is firstly derived from pure latent component analysis in terms of probability modelling, and then proved to be equivalent to NMF. The final conclusion is that NMF is actually similar to a mixture model, since the optimizations follow a similar rule. Thus, it is reasonable to adapt NMF (MNF) into a mixture model, which is the basic concept in the proposal of H. Sawada et al [87] [89], where MNMF with DOA estimation can compose a joint system (DMNMF) and enhance the final separation performance.

Figure 4.1 gives a glimpse on DMNMF speech separation system.

4.2 Methodology of DMNMF

To begin with, the von Mises distribution is a probability model given by the following probability density function (PDF) (2.25). As for the speech separation problem, the von Mises function could be used as the mixture model described in (2.26).

Different to the usage as a post-processing step in Section 2.4.1.1, researches propose the expectation maximization (EM) algorithm to update the parameters of (2.26). Firstly, the loss function of the EM algorithm could be built as follows [93]:

$$
L(\theta; \mu, \kappa) = \ln \left( \sum_{i=1}^{t} a_i \frac{e^{\kappa_i \cos(\theta - \mu_i)}}{2\pi I_0(\kappa_i)} \right)
$$

(4.3)

For the purpose of speech separation, it is assumed the dataset is the IDOA estimation derived from the spectrogram. Meanwhile, following the similar idea of
the literature [93] [107], the weight coefficients $a_i$ is assumed to be a constant. Then, this log likelihood could be adapted into the following with constants being omitted:

$$L(\theta_{m,n}; \mu, \kappa) = \sum_{m,n}^{M,N} \ln\left( \sum_{i=1}^{l} \frac{e^{i \kappa_i \cos(\theta - \mu_i)}}{2\pi I_0(\kappa_i)} \right) \quad (4.4)$$

Following this is to use the EM algorithm to update the parameters $\mu, \kappa$. This will be elaborated upon more in the following section.

In the meantime, a soundfield microphone with IDOA algorithm is used to estimate the DOA. From equation (2.11), assume the received TF-representations of signals in the $x$-directional channel, $y$-directional channel and omni-directional channel as $Y_{xc}$, $Y_{yc}$, and $Y_{oc}$, respectively. Then the equation (2.20) could be adjusted as the following form:

$$\theta(m, n) = \arctan \left[ \frac{Re\{Y_{oc}^*(m, n)Y_{yc}(m, n)\}}{Re\{Y_{oc}^*(m, n)Y_{yc}(m, n)\}} \right] \quad (4.5)$$

Meanwhile, similar to the idea of [87] [89], the NMF in this chapter is assumed as the mixture probability as follows (with constants omitted):

$$P(S_{mn}; W_{mk}, H_{kn}) = \prod_{m,n}^{M,N} N_c(0, \sum_k^{K} W_{mk} H_{kn}) \quad (4.6)$$

Then this log likelihood could be expressed as follows:

$$L(S_{mn}; W_{mk}, H_{kn}) = \ln(P(S_{mn}; W_{mk}, H_{kn})) \quad (4.7)$$

Here $S$ is the spectrogram of $Y$, where the relation with the mixture signal phase $\varphi$ is as follows:

$$Y_{mn} = \sqrt{S_{mn}} \varphi_{mn} \quad (4.8)$$

To obtain directional sound, the model (4.6) is applied on both $S_{xc}$, $S_{yc}$ as:

$$\begin{bmatrix} S_{xc} \\ S_{yc} \end{bmatrix} = W \ast \begin{bmatrix} H_x \\ H_y \end{bmatrix} \quad (4.9)$$

Combining equation (4.5) (4.8) (4.9), the DOA can be estimated.

Accordingly, DOA estimation and MNMF t are supposed to benefit each other, thus improving the separation performance of the whole joint model. Precisely, DOA estimation corresponding to each basis implies MNMF whether this basis vector should be chosen as part of a certain source during optimization; while the optimization of the complete joint probability model could also improve results of DOA estimation. Therefore, both equation (4.4) and equation (4.7) should be maximized, in order to have the optimal estimation of DOA and spectrogram parameters. Obviously, separating the whole likelihood into these two parts and
calculating the parameters cannot guarantee the global optimum. Therefore, the following sections describe how to merge these two parts within the EM algorithm to update parameters during optimization.

4.2.1 Combining DOA with MNMF estimators by EM optimization

Apparently, both DOA clues and MNMF clues are represented by probability models, which bring an easy choice to combine them, namely, the EM algorithm. Generally speaking, the EM algorithm could suppress errors from estimation of DOA parameters by checking whether the derived parameters are reasonable for MNMF decomposition, and vice versa. Meanwhile, both DOA estimation and MNMF could be updated under the EM-algorithm framework, which makes the proposed combination more reasonable to realize.

To begin with, the parameter of the whole system is set as: \( \phi = \{ \theta, \mu, \kappa, W, H, Z \} \). Here, \( \theta \) represents the DOA estimation according to IDOA (4.5); \( \mu \) and \( \kappa \) stand for the mean and concentration parameter in von Mises function, respectively; \( W \) is the basis matrix for MNMF part with \( H \) as the corresponding weight matrix. Despite these parameter, a new parameter, the latent variable denoted as \( Z_{ik} \) represents the mixing weights corresponding to the \( k \)-th basis vector in the MNMF basis matrix belonging to the \( i \)-th source. Thus, the complete log-likelihood can be represented with von-Mises distribution in (26) and MNMF as follows:

\[
L(S; \phi) = \sum_{i} \sum_{m,n} \ln(f(\theta; \mu, \kappa)P(S; Z_i, W, H))
\] (4.10)

Here, the likelihood from MNMF is changed with factor \( Z_{ik} \) as follows:

\[
S_{mn} = \sum_{i,k} Z_{ik} W_{mk} H_{kn}
\] (4.11)

Here, the \( S \) actually represents the two channel signals with corresponding \( H \) in the model of (4.9). Correspondingly, the MNMF in (2.61) could be changed as:

\[
P(S; Z, W, H) \propto \exp(- \left\| S - \sum_{i,k} Z_{ik} W_k H_k \right\|_{F,0}^2)
\] (4.12)
4.2.2 Computation of parameter sets

To find the optimal solution, the ultimate purpose is to maximize the joint log-likelihood. Similar to the standard EM algorithm update stage, the following auxiliary function is used for deriving the updating steps:

\[ \mathcal{L}_a(S; \phi) = \sum_i \sum_{m,n} u_{imn} \ln \left( \frac{f(\theta_i; \mu_i, \kappa_i) P(S; Z_i, W, H)}{u_{imn}} \right) \]  

(4.13)

where \( \sum_i u_{imn} = 1 \) is supposed to be the representation based on latent variable \( Z \).

Due to the concavity on \( \mathcal{L}(S; \phi) \), the following relation between original joint-likelihood in equation (4.10) and the auxiliary function in equation (4.13) could be shown as follows:

\[ \mathcal{L}(S; \phi) \geq \mathcal{L}_a(S; \phi), \quad \text{if} \quad \frac{f(\theta; \mu, \kappa) P(S; Z, W, H)}{u_{imn}} = \eta \]  

(4.14)

where \( \eta \) must be a constant. As for the condition, it could be derived as follows:

\[ f(\theta; \mu, \kappa) P(S; Z_i, W, H) = u_{imn} \eta \]  

(4.15)

\[ \sum_i f(\theta; \mu, \kappa) P(S; Z_i, W, H) = \eta \]  

(4.16)

Therefore:

\[ u_{imn} = \frac{f(\theta; \mu, \kappa) P(S; Z, W, H)}{\sum_i f(\theta; \mu, \kappa) P(S; Z_i, W, H)} \]  

(4.17)

In other words, the \( u_{imn} \) could be the post-probability. This is the E-step of the EM algorithm.

Meanwhile, for optimizing other parameters, the auxiliary function is maximized, thus obtaining the optimal parameters at the same time, which is referred as the M-step of the EM algorithm. This can be derived as:

\[ \max(\mathcal{L}_a(S; \phi)) = \max \left( \sum_i \sum_{m,n} u_{imn} \ln \left( \frac{f(\theta; \mu, \kappa) P(S; Z_i, W, H)}{u_{imn}} \right) \right) \]  

(4.18)

which is equal to:

\[ \max(\sum_i \sum_{m,n} u_{imn} \ln(f(\theta; \mu, \kappa)) + \ln(P(S; Z_i, W, H))) - \ln(u_{imn})) \]  

(4.19)

where \( u_{imn} \) is defined as the equation (4.17). At the same time, \( \|W\| \) need to be equal to 1 and \( Z, W, H \) all demand to be non-negative numbers, where the updating will be derived in the following.
The calculation is based on the partial derivative of each parameter. Since the last term, namely \(-u_{imn}\ln(u_{imn})\) remains a constant with variables as the parameter-set, the partial derivative on this part will always be 0. Hence, it is omitted from the following calculation.

For the convenience of derivation, the rest of equation (4.19) is separated into two parts. Specifically, the first part is as follows [93]:

\[
\sum_i \sum_{m,n} u_{imn} \ln(f(\theta; \mu, \kappa))
\]

\[
= \sum_i \sum_{m,n} u_{imn}(-\ln(2\pi) - \ln(I_0(\kappa_i)) + \kappa_i \cos(\theta_{mn} - \mu_i))
\]

(4.20)

where \(I_0\) is modified Bessel function of the first kind [92] [93], \(\mu_i\) and \(\kappa_i\) are the mean DOA estimation and concentration parameter of the \(i\)-th source. The \(\theta_{mn}\) is the DOA estimation of each TF bin. Then, the partial derivatives are:

\[
\mu_i = \arctan\left(\frac{\sum_{mn} u_{imn} \sin(\theta_{mn})}{\sum_{mn} u_{imn} \cos(\theta_{mn})}\right)
\]

(4.21)

\[
\kappa_i = A^{-1} \left(\frac{\sum_{mn} u_{imn} \cos(\theta_{mn} - \mu_i)}{\sum_{mn} u_{imn}}\right)
\]

(4.22)

Here, positions of all sources are supposed to be fixed. The \(\theta_{mn}\) is estimated based on IDOA with (4.5) (4.8) (4.9), representing DOA estimation corresponding to each TF bin. More precisely, the operator \(A^{-1}\) is denoted as a function that can be calculated with Batschelet’s Table [93].

On the other hand, the second part is to maximize the probability term from the MNMF model. However, since the distance kernel, namely the Frobenius norm brings the summation along source number \(i\) before the computation of whole EM framework, it is unable to directly implement the combination.

However, inspired by the similar idea of [87], a matrix relaxation method is introduced here as:

\[
\mathcal{L}^+(S; Z, W, H) = \sum_{i,m,n} \frac{1}{r_{imn}} \left\| G_{imn} - \sum_k Z_{ik} W_{mk} H_{kn} \right\|^2_{Fro}
\]

(4.23)

where the following conditions need to be met:

\[
\sum_i G_{imn} = S_{mn}
\]

(4.24)

\[
\sum_i r_{imn} = 1
\]

(4.25)
According to literature [87] [89] [95], this auxiliary function has the following property:

\[ \mathcal{L}^+(S; Z, W, H) \geq -\mathcal{L}(S; Z, W, H) \]  

(4.26)

Thus, to maximize the original log-likelihood term is equal to minimize the negative log-likelihood in (4.26).

This also leads to the maximization of the second part of the whole EM framework equal to the minimization on following term:

\[
\mathcal{L}_S^+(S; Z, W, H) = \sum_{i,m,n} v_{imn} \ln(\exp(\frac{1}{r_{imn}} || G_{imn} - \sum_k Z_{ik} W_{mk} H_{kn} ||^2_{Fro}))
\]

(4.27)

It can be found that the original Frobenius norm problem is relaxed into the new point-wise Frobenius norm problem corresponding to the auxiliary term \( G_{imn} \), thus enabling to combine the post-probability factor \( v_{imn} \) from the E-step and realizing the M-step.

Based on the conditions in (4.26) (4.27), the following three factors are proposed:

\[
\hat{G}_{imn} = \sum_k Z_{ik} W_{mk} H_{kn}
\]

(4.28)

\[
\hat{G}_{mn} = \sum_{ik} Z_{ik} W_{mk} H_{kn}
\]

(4.29)

\[
r_{imn} = \frac{G_{imn}}{G_{mn}}
\]

(4.30)

Intuitively, the term \( \hat{G}_{imn} \) is denoted as the part of a certain estimated TF bin belonging to the \( i \)-th source, while \( \hat{G}_{mn} \) is the total estimation on this TF bin, with \( r_{imn} \) as the ratio between them.

Thanks to the chain rule for the derivative, the four desired terms \( G_{imn}, Z_{ik}, W_{mk}, H_{kn} \) can be updated as follows:

\[
G_{imn} \leftarrow Z_{ik} W_{mk} H_{kn} + r_{imn}(S_{mn} - \hat{G}_{mn})
\]

(4.31)

\[
Z_{ik} \leftarrow Z_{ik} \frac{\sum_{mn} v_{imn} S_{mn} W_{mk} H_{kn}}{\sum_{mn} v_{imn} \hat{G}_{mn} W_{mk} H_{kn}}
\]

(4.32)

\[
W_{mk} \leftarrow W_{mk} \frac{\sum_{in} v_{imn} S_{mn} Z_{ik} H_{kn}}{\sum_{in} v_{imn} \hat{G}_{mn} Z_{ik} H_{kn}}
\]

(4.33)
$$H_{kn} \leftarrow H_{kn} \frac{\sum_{im} v_{imn} S_{mn} Z_{ik} W_{mk}}{\sum_{im} v_{imn} G_{mn} Z_{ik} W_{mk}} \quad (4.34)$$

Here, the left side is the updated values, compared to the right side with the values from the last duration. Apparently, this form is quite similar to the original multiplicative form in standard NMF, thus making it very easy to implement.

Likewise, in order to prevent the random-scaling problem, $Z_{ik}$ and $W_{mk}$ must be normalized before it is updated into next optimization epoch:

$$W_{mk} = \frac{W_{mk}}{\|W_{mk}\|} \quad (4.35)$$

$$Z_{ik} = \frac{Z_{ik}}{\|Z_{ik}\|} \quad (4.36)$$

$$H_{kn} = H_{kn} \ast \|W_{mk}\| \ast \|Z_{ik}\| \quad (4.37)$$

Through the normalization, these parameters can keep with a scale of 1, which shifts all scaling-factors into $H_{kn}$. The point of this is to fix the energy level on all parameters except weight matrices, which are supposed to be the representation of spectrograms with basis matrices as the base. Besides, for the practical purpose of this proposed method, this normalization also saves computations during parameter update since the only part requiring update in IDOA (4.5) (4.8) (4.9) is $H_{kn}$.

Combining (4.17) (4.21) (4.22) (4.32) (4.33) (4.34) and the IDOA’s derivations (4.5) (4.8) (4.9), the whole EM framework can be established and optimized iteratively until the error is small enough. Usually, in a practical scenario, a certain number of iterations are used to set a stop on the optimization, similar with the standard NMF. The final step is to use a soft-mask to finalise the separation of the mixture spectrogram into individual sources, which could be done by replacing $V_{ik}$ with $Z_{ik}$ in the equation (2.70).

Compared with the aforementioned methods of MNMF in section 2.5.2, the proposed methods do not require any calculations in complex-number domain, thus sidestepping the problem of forcing the negative eigenvalues rectified during EVD. Meanwhile, it creates a very easy initialization. The DOA information is a very general clue, which could be realized by any simple DOA estimation method. Moreover, the DOA could be decided in advance and all computations on non-negative values only, so that the proposed method is supposed to require much less computation power than the other MNMF methods. Therefore, it has strong potential to overcome the problems in the other MNMF methods and surpass the separation performance of these.
On the other hand, the proposed method exploits the features from (M)NMF, which brings the possibility to be expanded with other beneficial factors, such as trained features, sparsity or time-constraints. This is the superiority of proposed method over the standard spatial-assisted methods in Section 2.4.1.

### 4.3 Results and analysis

In order to confirm the proposed DMNMF’s effectiveness for speech separation problems, this sub-section presents a series of associated experiments and corresponding results.

#### 4.3.1 General configurations

Basically, the basis is to simulate different speech-mixing situations with microphone-arrays as the receivers. More precisely, this thesis employs a set of different simulated Room-Impulse-Response (RIR) signals convolved with clean speech signals as the dataset. This sub-section presents the configurations on the general parameters.

#### 4.3.1.1 Baseline methods

With respect to the baseline methods, this section introduces the methods from [87], [89], [90], [93] and [95]. These methods are the major works on multi-channel NMF, which are properly adapted to the experiments. All the general algorithm parameters are initialized as the ones with the best performances in corresponding baseline-papers. For simplifying the representation, the baseline methods [93], [90], [95], [87] and [89] are abbreviated as ‘Cn’, ‘Eu1’, ‘Eu2’, ‘Jp1’ and ‘Jp2’ in the following figures, respectively. Likewise, the DMNMF proposed in this thesis and the unprocessed signal is referred as ‘Prop’ and ‘Input’, respectively.

Despite the general ones, the number of cluster, namely $l$ in [89] [93], is initialized as two times of the number of speech sources, $i$ (2.62), based on the ground truth. The idea is to ensure an appropriate level of stochasticity, effectively inspiring the algorithm to explore a potentially better solution rather than being trapped into local minima; while the configuration could guarantee an acceptable complexity of the
algorithm, saving the computation cost as a relatively low level. Besides, the
program automatically performs a Pairwise-Merge operation after every 10-iteration,
until the number of clusters is reduced to the number of sources from the ground-
truth, in order to ensure the stability of algorithm [89] [93].

Regarding [95] and the proposed one which require the DOA information before
the main optimization, this thesis employs the method from [108] as the initial DOA
estimation, where the number of source from the ground truth is set as the number of
source during this initial phase.

Additionally, the spatial angle resolution for the simulation setup in [93], [93] and
proposed methods is configured as 15 degrees. Moreover, the number of major
optimization iterations is set as 200 for all methods. The number of basis vectors for
all MNMF-associated methods is set as 256, which is empirically resulted in the best
separation performance.

4.3.1.2 RIR & related simulation

As for the generation of RIR signals, software RIR generator [104] is applied to
construct the room with different reverberant conditions. In all the experiments of
this thesis, the room dimensions are set as 8m*6m*4m. Without losing
generalizability, this thesis simplifies the conditions of real life into a 2-dimensional
scenario, where the simulated roof and floor have no reflection. A variety of different
RT60 values are investigated for this room.

Due to the proposed method focusing on the B-format microphone array, the
experiments apply a similar regular-pyramid structure in Figure 2.2 for all methods.
The central point of the simulated room, point (4, 3, 2) is defined as the origin at (0,
0, 0) in all experiments. While its geometrical centre is placed at the origin point, the
coordinates of the microphone-structure four vertexes are at point (0.005, 0.005,
0.005), (0.005, -0.005, -0.005), (-0.005, 0.005, -0.005) and (-0.005, -0.005, 0.005).

With respect to the B-format microphone array used in [93] and proposed DMNMF,
these four vertices stand for the position of microphones on the left-front (LF), the
right-front (RF), the left-back (LB) and the right-back (RB), respectively. All four
microphones are cardiac microphones. In contrast, for the methods in [87] [89] [90],
this thesis still uses a four-channel microphone with the same geometric structure
and positions, except all cardiac microphones replaced with omni-directional
microphones. Since there are no special conditions on microphone structure in [87] [89] [90], it is fair to support the microphone configuration in this thesis is a proper one in terms of implementation. In addition, all related parameter sets are adapted according to this structure, such as the DOA kernel in [95]. Besides, imitated simulated omni-directional microphone (virtual microphone) is employed at the origin as a reference microphone, so that the mixture after propagation from speech sources could be measured in terms of SDR or related ratios.

Meanwhile, all speech sources are placed on the same level of the microphone array, with 1.5 meter distance between any one of them and the geometrical centre of the microphone array. With the positive side of $x$ axis as 0 degree, one of speech sources is fixed at (1.5, 0, 0), which represents the position of primary speech source. This primary source is defined so that the other sources can be relatively defined as interference source. Consequently, all the evaluations are also targeted at this primary source.

To give an intuitive description, the Figure 4.2 presents one scenario of the experiments, where three sources are active. The receiver is located at origin point, which is labelled as $Rec$. A circle with $r$ as radius is drawn by dash line, where all sources are located on it with different angular position. The preliminary source, namely the target source is presented as a star on $x$ axis; while the angular difference between source with interference $Int_1$ and interference $Int_2$ are $\theta_{i_1}$ and $\theta_{i_2}$, respectively. These two values are angular coordinates of two interferences, correspondingly.
4.3.1.3 Dataset, mixture-generation & evaluation metrics

The experimental test dataset is simulated by convolving certain RIR signals with clean speech files. As for the clean speech files, the well-known set, namely TIMIT [116] is used.

Specifically, TIMIT dataset is composed by 8 directories, where each of them contains around 30 to 80 different speakers, including males and females. Besides, 10 speech utterances are recorded by each speaker, with first utterance having repeated contents. Given the common setup for speech separation experiments [87] [89] [90] [93] [95], the thesis only chooses the last 9 utterances of every speaker as test speech files. Each utterance is about 3 to 5 seconds. From these sentences, this thesis randomly picks 30 speakers, building up a mini-set for the evaluation. Thus, there are 270 utterances in total for the input clean speech signals.

Additionally, all speech files in TIMIT are recorded at 16000 Hz, which is perfectly matched with the separation processes and evaluation processes.

According to the experiment design, each experimental sub-set is generally constituted by 30 different mixtures. More precisely, 5 speakers are picked from a total 30 speakers to act as the target source, while the remaining speech sources are attributed to as interference speakers. To generate one certain mixture, one utterance
is stochastically chosen from all utterances of 5 speakers, with a similar procedure used to determine the interference-speakers’ utterances. Care is taken to ensure mixtures of the same sentences are not repeated.

Afterwards, each utterance is convolved with the corresponding RIR signal targeting 4 microphones in the microphone array, which is generated based on the RIR generator. Likewise, all speech utterances are convolved with the RIR signal targeting at the virtual microphone, for the evaluation purpose.

In terms of evaluation metrics, this Chapter employs the SDR, PESQ and STOI as the objective quality evaluation, subjective quality evaluation and intelligibility evaluation, respectively. Due to the limitation of this thesis’s scope, SDR is the only objective evaluation, which presents a comprehensive view on all objective evaluation metrics in Section 2.6.

Results for each sub-experiment (e.g. 2 active sources) are averaged across all combinations of source speakers and DOAs.

4.3.2 Multi-source speech separation

The main aim of DMNMF is to separate mixtures of multi-source speech under reverberant environments. Compared with the experiment assumption of Chapter 3, this condition is more practical. Therefore, this sub-section presents the results of multi-source speech separation.

Given the different mixing conditions in real life, this experiment covers the separation with mixture speech constituted of 2 speech sources and 4 speech sources, respectively. The angular coordinates of speech sources (in degree) corresponding to different sets in 2D polar coordinate system are set as Table 4.1 and Table 4.2, where ‘src’ is the abbreviation of source, with ‘tar’ as target and ‘int’ as interference.

The design is to exploit the robustness against the variable number of active speakers, which might bring critical effects on the separation performance. Likewise, the performance under such a diverse range of angular coordinates also projects each method’s ability to counter the changes on spatial relation between sources and all microphones, and removes the possible bias from RIR generator in regards of unbalance reverberation-distributions. Different angular increment on each source position serves a similar purpose.
Moreover, for each combination, there are four different RT60 configurations for the RIR signal generation, namely 0, 130, 250 and 350 ms. Specifically, the first one represents an anechoic environment and the last one indicates a largely reverberant environment [95]. The middle two are designed as [93], in order to enable the comparison on performance of all methods in this experiment with other general methods.

Eventually, one experimental sub-set with a certain combination of angular coordinates, source numbers and RT60 is built up by 30 different mixtures. Following are the evaluations of the separation performance.

The first one is the SDR test on 2 active sources. It can be found from the Figure 4.3 that the SDR at the virtual microphone is gradually dropping while the level of reverberation is increasing. Generally, all methods gain around 10 dB improvement after separation, and show a similar deterioration on separation performance following the reverberation increase.

Among all methods, DMNMF basically achieves the best performance in terms of the average score, where the advantage is about 0.5 to 1 dB over the method with second best performance. It is also worth noting that the DMNMF presents only small confidential intervals, which implies a strong robustness against diverse changes on the mixture environment. Yet, with confidence interval, the performance is basically at the same level, when RT60 is around 130 ms and 250 ms.

<table>
<thead>
<tr>
<th>Table 4.1 Source position (2 src)</th>
</tr>
</thead>
<tbody>
<tr>
<td>src1 (tar)</td>
</tr>
<tr>
<td>src2 (int)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4.1 Source position (4 src)</th>
</tr>
</thead>
<tbody>
<tr>
<td>src1 (tar)</td>
</tr>
<tr>
<td>src2 (int1)</td>
</tr>
<tr>
<td>src3 (int2)</td>
</tr>
<tr>
<td>src4 (int3)</td>
</tr>
</tbody>
</table>
Figure 4.3 SDR performance with 2 active sources

Similarly, figure 4.2 presents the PESQ test in the same experiment sub-sets. Through all experiments in this topic, all methods obtain a 0.5 to 1.4 increment over the score of the input signal at the virtual microphone in terms of PESQ score. Despite an alike trend in SDR test, the PESQ scores show a larger oscillation in
every experiment sub-set, especially when the reverberation is at a large level. This is undesired but reasonable, which is quite similar to the separation performance in Section 3.3.3. DMNMF still shows the competitive performance or even best performance of average score, in spite of only small advantages and similar confidence intervals against the second-best method.

Moreover, Figure 4.5 gives the results of STOI tests. Obviously, the trend of STOI scores against different reverberation levels is consistent to the other two evaluation metrics. However, the enhancement of intelligibility is not that strong such as under the former two evaluation metrics, with the average gain only 10 to 15 percentages in terms of word accuracy. This implies the main limitation of all methods tested in this thesis. As to particular applications, such as recognition-associated applications, these methods might be not adequate to fulfil the purpose.

Regardless this inadequacy of intelligibility, DMNMF still achieves the best average value among all methods but only with a basically same in terms of confidential intervals.

In contrast, the performance of all methods under 4 active sources shows a relatively-worse result, which is described further below.
Compared with the similar tests but with 2 active sources, the separation for 4 active sources’ mixture is more difficult, where the SDR score shows a relative decline.

Accordingly, the PESQ results also present a similar problem as the active-source number increases. Specifically, although the PESQ score of the virtual-microphone input signal barely falls, it extremely hampers the separation process. For example,
the PESQ gain of DNMF with 350 ms reverberation under 4-source-reverberation is about only 0.35, while this number is about 0.9 under 2-source mixture on the contrast. Thus, it highlights the shortage of these methods which focus on the utilization of spatial information.

Additionally, Figure 4.8 shows corresponding STOI results. Although the SDR results and PESQ results with 4 active sources are relative worse than ones with only 2 active sources, the difference between STOI results with 4 active sources and with 2 active sources are not apparent, with constantly a disadvantage as 10 percentages of word accuracy. It is reasonable since three evaluation metrics are not strongly related. Yet, details behind this phenomenon are still under investigation.

To conclude, the reverberation highly impedes the separation process, which justifies the necessity of study of speech separation under reverberant environments. Meanwhile, the increase of the active number of sources also extremely hinders the separation process, especially with respect to speech quality. While most speech enhancement ignores the variety of active source numbers, experiments in this subsection verify that approaches might take this side information into account, ensuring a better performance.
4.3.3 2-source separation under anechoic environment

For the purpose of verifying the effectivity of multi-channel microphone processing with spatial information in terms of speech separation, this section presents a similar experiment design as it in Chapter 3.

Specifically, the general configuration is still set as the description in Section 4.3.1, while the environment is an anechoic environment (RT60 as 0 ms). With the expectation of simulating an instantaneous speech mixture to equalize mixing-condition in Chapter 3, experiments in this sub-section also includes a series of input 2-active-source mixtures with different SDR conditions. Given the reference input SDR from signals at the virtual microphone, the input SDR is adjusted as -15 dB to 15 dB, with 5 dB increments to set up one experiment sub-set.

The polar coordinates of target source and interference source are set to 0 degrees and 60 degrees, respectively. This design is one specific configuration to create an instantaneous speech mixture in Chapter 3. With the consistence of separation performance in Section 4.3.2 (acceptably small confidence-interval), it is reasonable to believer adequate experiments (30 mixture utterances as a test group for each input SDR) with only this specific geographic design is valid to make a general conclusion about the performance of this method.

Figures 4.9 to 4.11 reveal the performance in different respect of evaluations. Obviously, all methods with multi-channel microphone processing reveal a dominant advantage compared to the methods in Chapter 3. For example, with 0 dB SDR as the input condition, the method with best performance, namely DMNMF, prevails the best-performance method MASK-NMF in Chapter 3 by over 10 dB, 0.8 and 0.15 in terms of SDR, PESQ and STOI, respectively.

To explain it, the artefact (revealed by SAR) and distortion after separation process contribute the majority of this difference. Compared with the feature representation purely based on spectrogram in single-channel-NMF-associated approaches, methods in this Chapter take a huge advantage of spatial information. Generally, all parameters are supposed to be optimal after separation processing, including spectra corresponding to each speaker and spatial factors revealed by these spectra.

Considering the limitation of resolution created by STFT computation, there are more likely overlaps in spectrogram than spatial parameters, such as DOA or the phase difference in covariance matrices between multi-channel signals. Thus, the-

81
Figure 4.9 Output SDR performance versus different input SDR

OutSDR vs. InSDR

Figure 4.10 PESQ performance versus different input SDR

PESQ vs. InSDR
-separation in multi-channel processing could exploit the spatial information as heuristic hints to enhance the related parameter optimization, ultimately bringing more chances to obtain a more proper separation. However, in single-channel processing, all optimizations are barely decided by spectrogram features, which only focuses on numerical computation rather properly separating mixture.

Certainly, from a rigorous experimental view, the comparison of this sub-section is invalid, since this sub-section utilizes signals from a 4-channel microphone as the input, rather than the signal from single channel microphone in Chapter 3. Nevertheless, the main purpose is to verify the advantages of multi-channel microphone processing in terms of a better speech separation performance. The capability of using spatial information brings an inherent advantage for multi-channel microphone. Despite that matrix-decomposition approaches are mostly designed with single-channel microphone, it is legitimate to explore more on extending the current approaches into multi-channel-microphone schemes.

4.3.4 Robustness against different initializations

Notwithstanding all tests aforementioned, one of the most important metrics is to evaluate the robustness against possible errors inherited from the previous stages of
the separation system. Regarding the approaches of multi-channel microphones in this Chapter, these errors might come from wrong source-number configuration and errors in previous DOA estimation.

More precisely, this experiment is composed of 5 different DOA initializations, including the four sources’ angular coordinates according to ground-truth, DOA estimates based on approach [108] (referred as ‘Zheng’ in the following graph), 15-degree counter-clockwise error for each source, 30-degree counter-clockwise error for each source and 60-degree counter-clockwise error for each source. The first group of experiments aim to demonstrate the upper limit of the separation performance in the current setting; in contrast, the last three groups are used to test robustness of methods. The approach of [93] and DMNMF are used in this test, since they utilise the DOA as input information for initializations.

All experiments are under a simulated anechoic environment (RT60 as 0 ms), so that the results could present differences based on the algorithm robustness against error in DOA only. As for each group, there are 30 mixture utterances to test.

The results are showed in Figures 4.12 to Figure 4.14, where the horizontal axis stands for different error-settings. The advantage of DMNMF against approach [93] is its robustness to the initial errors in the DOA. Specifically, in the idealistic situation where the DOA is initialized as the source angular coordinate from the ground-truth, DMNMF shows 2dB, 0.25 and 0.04 advantages in terms of SDR, PESQ and STOI respectively. As for the remaining approaches, the advantage of DMNMF over approach [93] gradually increases from 1.5 dB (SDR), 0.15 (PESQ) and 0.03 (STOI) to 2.4 dB, 0.3 and 0.08, correspondingly. This significant change in terms of separation performance implies that DMNMF is more robust to possible errors on initializations.

Additionally, it is obvious that there is still a large improvement margin in case that the initialization is according to ground-truth information. This indicates the potential to enhance the whole separation process, such as replacing the initial DOA estimation with a more accurate one, or taking the advantage of certain side-information, such as previously-known location of the speaker.
Figure 4.12 SDR performance against different errors in DOA initialization

Figure 4.13 PESQ performance against different errors in DOA initialization
4.3.5 Analysis of the algorithm complexity

Despite all the results in the aforementioned contents, algorithm complexity is also a crucial factor to evaluate different approaches. Although hardware is developing rapidly, speech separation still pursues a real-time approach as the ultimate target. Moreover, all algorithms above require iterative process for the optimizations of different parameters. Thus, algorithm complexity should be taken into account in terms of algorithms’ separation performance.

Compared with other baseline methods in this thesis [90] [87], [93] [95] and [89] show generally better performance. Assume the number of microphone array is $C$. Then the element-number of algorithm input in [93] [95] [89], namely the cross-correlation matrix is $C*C$. In contrast, DMNMF requires only $x$ and $y$ channel signal and basic element-wise division for IDOA during optimization. Given that major parameters derived from the spectrogram, including post-probability factor $O$ in [90] [96] and $V$ in [94] and DMNMF, basis vectors $W$, weight vectors $H$, cluster vectors $Z$ are updated in a similar way (Section 2.5.2, Section 4.2.2), DNMF reduces the complexity of these three baseline methods by $C*C/2$ times. When channel number is large to deal with more simultaneously active sources or the application focus on

Figure 4.14 STOI performance against different errors in DOA initialization

![STOI vs. InErr](image-url)
real-time or near-real-time processing, the advantage of less computations in DMNMF is reasonable to be verified.

4.4 Chapter conclusion

This Chapter describes a new multi-channel NMF approach, referred as DMNMF. Unlike the baseline methods, including multi-channel NMF or relative models [93] [90] [95] [87] [89], DMNMF is extremely simplified, where only initial DOA information (including source number) and spectrogram in each channel of a B-format microphone are required. Since it is able to obtain directional information with the B-format microphone, the DMNMF model needs no computation of features with complex number to exploit spatial information. This enables DMNMF to achieve a more proper separation, compared with baseline methods.

Experimental results reveal an expected advantage of DMNMF over other methods. Through all experiments, DMNMF always achieves the best separation performance, with a visible privilege over the second-best-performance method. Additionally, DMNMF presents also a convincible robustness against potential initial errors, which also verifies the effectiveness of this proposed method.

The future work might focus on a better DOA initialization and replace the KL-divergence with more different members from the divergence family [12], so that a similar separation scheme could be realized and possibly surpass the current one. Meanwhile, more tests could be set up, such as the robustness test against errors in source number.
5 CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary and conclusions

This thesis is motivated by the flourishing progress of speech signal processing, where the main target is to develop an effective speech separation method under various propagation scenarios. Given the underperformance of this problem in contrast with the increasing application-demands, the thesis takes advantages of the growth of associated data and matrix-decomposition approaches, and focuses on NMF as the main framework for speech source separation. Compared with other matrix-decomposition methods, the most obvious difference with NMF is the non-negativity constraint, which fundamentally resembles the spectrogram mixture of different acoustic components in real life. Consequently, the features of NMF, namely the basis vectors, are related to a specific piece of the spectrogram. Thus, computing activations of NMF features (the weight matrix) relates to actual parts of the speech spectrogram. In Section 2.4 and Section 2.5, the thesis therefore reviews multiple separation approaches based on single-channel NMF and multi-channel NMF, respectively.

However, due to the similarity and overlap between spectrograms of different sources, there are limitations of the basic NMF approach for speech separation. In order to improve separation performance, this thesis proposes two different extensions of NMF, including the enhancement based on integrating the IBM and combining DOA information.

The following briefly summarizes this thesis:

- Chapter 3 presents a novel extension of single-channel NMF, where IBM information is introduced into the loss function of NMF. The idea is to use IBM information generated from the ground-truth of the mixture and adapt the common loss function to a collaborative loss function that includes both the spectrogram and IBM information in Section 3.2.1. Despite the source-wise training of standard single-channel NMF, this collaborative loss function requires an extra joint training with respect to the simulated speech mixture as
described in Section 3.2.2. In terms of improvements, the proposed methods show about a 2 dB advantages in SIR with a competitive performance in regards to other evaluation tests, when input SNR is in the low range. As the input SNR increases, the proposed method still gives an acceptable performance compared with the next best performing approach.

- Chapter 4 discusses the enhancing of the separation performance of MNMF with DOA information. Compared with other baseline methods, the proposed method employs a B-format microphone (Section 2.2.2), which brings efficiency in terms of DOA estimation before and during the processing. Thanks to the B-format microphone’s special design to perform an IDOA in Section 2.4.1.1, the proposed method can calculate the DOA corresponding to each feature of a common MNMF model as in Section 4.2.2. The EM method is used to optimize related parameters, define the intensity of each basis belonging to a certain source and finalize the separation with the help from the intensity. Generally, this proposed DMNMF achieves the best performance in most cases, which is about a 0.5~1.0 dB SDR advantage, 0.1~0.2 PESQ reduction and around 2% word accuracy measured by STOI in terms of the average performance, with a slightly narrower confidence interval to claim better robustness. Meanwhile, DMNMF largely reduces the computation cost, compared with other baseline methods in Section 4.3.1.1. Compared with the methods in Chapter 3, DMNMF exploits spatial information, which improves the separation performance.

5.2 Recommendations for future work

Regardless of all the improvements in separation performance, NMF and related techniques still face multiple problems and limitations. Since all the features are directly or partially derived from spectrogram information, the similarities between features of different sources tend to mislead the separation process.

Therefore, the main work about NMF in separation focuses on either reinforcing the differences between features of different sources during training, or making a decision whether a feature should be active at a particular frequency and instance to
reconstruct one certain source during separation. Besides, to adapt the separation system into a real-time or near-real-time application is also appealing.

As the answers for the first two problems, the thesis provides methods in Chapter 3 and Chapter 4, respectively. However, the improvement is acceptable but still limited. In recent years, various approaches for speech separation problems emerge, which could be integrated within the NMF framework and lead to a new era of increased performance. Accordingly, several promising approaches are discussed as the follows, to eventually conclude this thesis:

- Section 2.3.3 and Section 2.5.3 outlines multiple sub-band transforms, where the traditional STFT is replaced by the transform based on diverse sub-band filters. One recent model, called ScatNet [2] [28], is reported to have the state-of-art performance. In short, the scattering transform of ScatNet combines wavelet filters with different resolutions, thus calculating the signal coefficients under each wavelet filter. Next, the coefficients of each sub-band filter from the first step are set as one group, where all coefficients inside each group are transformed by performing a similar wavelet transform again. The authors refer the results from each hierarchical transform to one layer.

In this fashion, the transform brings different resolutions. Specifically, the higher the band-frequency is, the higher resolution of the transform, enabling an adaptive frequency resolution corresponding to different band-frequencies. Meanwhile, given that the transform of the next layer is always performed over the sub-bands defined by the previous layer, the whole system guarantees the final results with a fine time resolution, which avoids information loss as in the one-layer only wavelet transform [13] [28].

It is supposed that the scattering transform can provide more representative and distinct information for any matrix-decomposition model to use. With respect to NMF, authors in [2] [28] have already tested NMF with signals provide by scattering transform to replace the traditional STFT, which brings a steady improvement over the baselines.

Yet, the NMF model in [2] [28] is for the single-channel NMF. As the performance of associated models in this thesis prevails over the common NMF,
it is reasonable to believe the fusion between models in Chapter 3 and Chapter 4 with the scattering transform can improve separation performance.

- As previously described, a strategy to decide basis-vector activations is beneficial to enhancing the performance. In [37] [53], a modified Markov model, named as infinite factorial hidden Markov model (IFHMM), is defined for this purpose. In general, IFHMM basically consists of two parts, including a hidden Markov model (HMM) to estimate the post-probability of a component being active and an India Buffet Process (IBP) providing one part of the prior probability for HMM. Besides the well-known HMM, an IBP is designed based on a non-parametric Bayesian factor model, where the process computes distribution over infinite binary matrices to decide if a corresponding component is active or not in the IFHMM system. Contrary to parametric systems as in the aforementioned NMF-associated systems, a structure such as IBP requires no pre-knowledge about component number [53]. Instead, the structure can automatically decide the activation of components, sidestepping a hard decision on the number of components. In regards to NMF, this may adapt the basis matrix of each source into a flexible-sized matrix, where the number of basis vectors can be adaptive corresponding to each particular source. Meanwhile, during the separation stage, an NMF-associated system enhanced by IBP can be built without requiring knowledge of the source-number, which is a major problem of all methods in Chapter 4. Several approaches have already merged NMF with IBP or IFHMM, where an overall advantage of separation performance has been reported [36] [48] [52]. For both methods proposed in this thesis, an IBP or IFHMM can be combined, which enables the auto-computation of features’ activations and source-number.

- Recently, one increasingly topic is deep learning, where vast and different artificial neural networks have been proposed and proved to be successful in solving multiple signal processing problems. With respect to audio processing, the recurrent neural network (RNN) demonstrates a promising potential. Specifically, long short term memory (LSTM) networks have achieved great
successes in problems such as machine translation or speech recognition [112] [113].

Typically, an LSTM is mainly composed by three parts: one forgets information inherited from the previous state; one adds new information of the current network input into the current network state; one computes the output information and propagates the current state to the next instance [112] [113].

Similar with IBP or IFHMM, a LSTM could also provide the activation information of certain components in speech separation problems. However, a bare HMM is usually too shallow and is not able to extract comprehensive information. In contrast, stacking several LSTM together and constructing a deep LSTM network (DRNN) can exploit the high level of non-linearity of the whole network, exploring more detailed information. This can benefit the speech separation system to obtain a more accurate estimation of corresponding information after successful training, where approaches of IRM or soft-mask estimation enhanced by DRNN have already been studied and showed various advantages [112] [113].

As for the methods in this thesis, a similar approach to that described for enhancing the NMF system with IBP can be employed with replacing IBP by a DRNN, where the final benefits remain to be investigated.
REFERENCES


[12] P. Smaragdis, C. Févotte, G.J. Mysore, N. Mohammadiha, M. Hoffman, "Static and


[61] B. Gao, W. L. Woo, S. S. Dlay, "Single channel source separation using EMD-


[74] C. Joder, F. Weninger, D. Virette, and B. Schuller, "Integrating Noise Estimation and


[99] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech


2012.


