Improving Analytic Algorithms for Speech Separation

by

Yudong He

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy
in Industrial Engineering and Decision Analytics

August 2022, Hong Kong

Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Yudong He

December 9, 2022

Improving Analytic Algorithms for Speech Separation

by

Yudong He

This is to certify that I have examined the above PhD thesis
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by
the thesis examination committee have been made.

Prof. Richard So, Thesis Supervisor

Prof. Jiheng Zhang, Head of Department

Department of Industrial Engineering and Decision Analytics

December 9, 2022

To My Parents

Acknowledgments

I sincerely thank my supervisor, Prof. Richard So, for his gentle and tireless guidance and support during my Ph.D. study. I am deeply inspired and infected by his positive and optimistic attitude towards life and research.

I thank my committee member, Prof. Xiangtong Qi, Prof. Wei You, Prof. Raymond Wong, and Prof. Woon-Seng Gan for their reviewing this thesis and for their valuable suggestions.

I am also grateful to my research group mates and office mates, for making my life colorful. We often go out to eat something delicious, go hiking, make hotpot, and play exercises. Particularly, I thank Jun Hui for his advice on my early research. Thanks, Shutao Chen, we often have some relaxing chats and lunches. I thank Kyle for treating us in his house. I thank Pin Gao and Yue Wei for enlightening me in my study and life. There are so many people to thank, that I won't list them all.

Last but not least, I appreciate my parents' mental support throughout my four-year study in Hong Kong. They give me the greatest freedom to do what I want. They are my powerful and secure backing, my warm harbor. Finally, I thank myself for the effort made in pursuing this Ph.D. Although not doing well in many areas, such as inefficient, poor ability to find research questions, poor writing, etc. I am looking forward to continuous learning, updating, and improving myself every day.

Table of Contents

List of Figures

List of Tables

Improving Analytic Algorithms for Speech Separation

by Yudong He

Department of Industrial Engineering and Decision Analytics

The Hong Kong University of Science and Technology

Abstract

Humans can listen to others in noisy environments, but it is difficult for a humanoid robot to separate the speech of individual persons from a mixture of these sounds. Known as the cocktail party problem, this challenge has intrigued scientists and engineers for more than half a century. Using the sparse characteristic of speech signals, this thesis improves three binaural audio separation algorithms in different scenarios: 1) weak target signal extraction, 2) fast separation, and 3) under-determined reverberant speech separation (i.e., binaural speech separation problem with more than two mixed sources in the presence of echoes).

First, the thesis improves a previously reported audio cancellation kernel to separate weak target signals. Our new version of the cancellation kernel achieves comparable or even better results with 3000 times the speed thanks to our analytical solutions. This solution originated from our realization that the whole extraction process can be performed in the time-frequency domain by the Short-time Fourier transform.

Second, the thesis improves the degenerate unmixing estimation technique (DUET), one of the fastest algorithms in speech separation. As a binary masking technique, DUET cannot completely separate speech signals, resulting in poor performance. We applied post-filtering with multiple linear spatial filters to improve the mask separation results and successfully resulted in significantly better separation performance.

Third, the thesis improves the l_1 minimization commonly used in audio separation algorithms. Speech separation can be converted to an l_1 minimization problem that aims to minimize the l_1 norm of the reconstructed signal. We derived and test a new weighted l_1

norm and showed that it can outperform the unweighted l_1 norm. The new algorithm can be solved using the same l_1 minimization solver but converges faster than the unweighted l_1 minimization. The improved l_1 minimization algorithm has been shown to work in the presence of reverberation and with more than two mixed speech sources.

Chapter 1

Introduction

1.1 Cocktail Party Problem

Imagine you are at a cocktail party and there are a lot of people around you talking. It may not come as a surprise to you that you can hear or communicate with specific people in such a noisy environment. Humans are born with this selective hearing ability. However, developing a digit algorithm to reconstruct speech signals from their mixtures is very challenging. This difficulty is referred to as the cocktail party problem [10]. As shown in Fig. 1.1, the mixed signal is very messy and the speech source signals are highly non-stationary, which can not be modeled by a simple random signal, e.g., the Gaussian signal. Their patterns are also elusive. It is so impressive that humans can do a real-time, highly robust separation of more sources even without any effort. For more than half a century, legions of scientists and engineers have been studying to overcome this problem. Not only is the problem itself challenging, but the solution to this problem has great application prospects in various real scenarios, including hearing aid devices, voice assistants, hands-free communication, conference call, humanoid robots with auditory systems, etc.

1.1.1 Significance of solving the cocktail party problem

Solutions to the cocktail party problem can be applied in a variety of scenarios (see Fig. 1.2). For example, using voice assistant and hand-free communication in a noisy environment, e.g., public transportation, restaurants, cafes, etc. In recent years, because of the pandemic of the Covid 19 virus, virtual meeting, like zoom, is becoming more and more popular. Due to its convenience, the virtual meeting may remain popular even if the

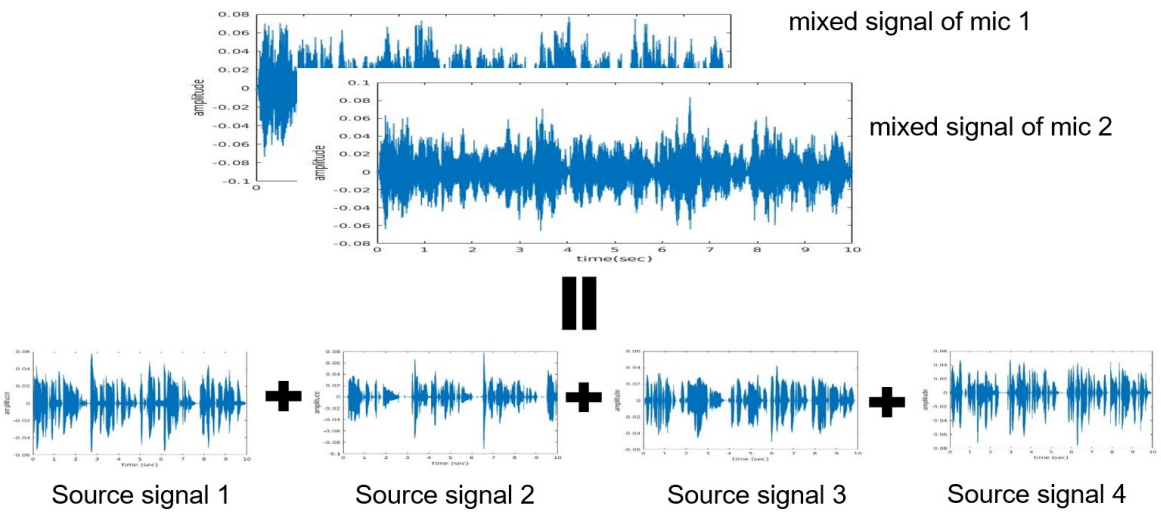

Figure 1.1: An illustration of speech signal separation. It is an inverse process to reconstruct speech source signals from their mixed signals. (a) Mixed signals observed by two microphones. (b1)-(b4) Four source signals.

virus pandemic is over. Solutions to the cocktail party problem can be applied in scenarios where multiple people have different virtual meetings in the same room, or having the same meeting but are assigned to different breakout rooms. A humanoid robot with an auditory system is another application.

1.1.2 Solutions - blind source separation algorithms

Blind source separation (BSS) is a technique of extracting one or more sources and canceling interference and/or noise without prior information on the sources' location or the mixing process. Generally speaking, BSS can be applied to solve not only speech separation but also image separation, e.g., separation of electroencephalogram (EEG) and magnetoencephalography (MEG) signals. In this thesis, when we talk about BSS, the separation object is a mixed speech signal.

A bulk of BSS techniques have been proposed and well documented in recent years. There are some review papers [11–13] to help people have a systematic knowledge of BSS. These techniques were developed into two different scenarios: monaural speech separation and multi-channel speech separation. Because more remarkable quality improvement can be achieved and a microphone array is much cheaper and more widely deployed than before, multi-channel techniques will become more popular and attractive. This thesis mainly talks about speech separation using more than one microphone. Multi-channel

(a) Voice assistant

(b) Hand-free communication

(c) Virtual meeting

(d) Humanoid robot

Figure 1.2: Some application scenarios of solution to the cocktail party problem. (a) Voice assistant. (b) Hand-free communication. (c) Virtual meeting. (d) Humanoid robot.

speech separation can be classified as:

1. Binary masking based on spatial cues, e.g., [4, 14].
2. Cancellation kernel via background learning,e.g., [3, 15].
3. Spatial filtering, or beamforming, e.g., [9, 16, 17].
4. Independent component analysis (ICA), e.g., [5, 18].
5. Non-negative matrix factorization (NMF),e.g., [7, 19].
6. l_1 minimization based sparse signal separation,e.g., [20–22].
7. Others, e.g., [8].

There are so many different kinds of algorithms, we have only listed some of them. All of these algorithms have their own strengths and limitations. In recent years, some people

have combined different methods to make up for each other's shortcomings and better performance has been obtained, for example, by combining ICA and NMF [6], binary masking and spatial filtering [23], etc.

Based on the existing algorithm, this thesis makes improvements by analyzing their advantages and disadvantages and combining the characteristics of speech signals. Compared with the existing algorithm, the new algorithm takes their essence, removes their dross, inherits their advantages, and improves their shortcomings.

1.1.3 Characteristics of speech signals

Let's observe the spectrograms of two speech signals. As shown in Fig. 1.3, although the two speech signals overlap in the time domain, their spectrograms do not overlap for a considerable part. This special structure can not only help us to obtain the information of the mixing process, in fact, there are many studies on how to estimate the mixing process by identifying the disjoint part [24–26] but also bearing this structure in mind can help us better reconstruct speech signals. However, many algorithms do not take this partially disjoint structure into account when reconstructing speech signals. For example, ICA only considers the signals to be independent of each other, or they over assume this structure, as in binary masking which over assumes that the signals are completely disjoint. This thesis explores how to exploit this partially disjoint relationship to separate speech signals.

1.2 Contributions

There are three major contributions of this thesis

- Improving cancellation kernel: the new algorithm greatly weakens the condition for using the original algorithm, has an analytical expression, and has comparable or better performance.
- Improving degenerate unmixing estimation techniques (DUET): the improved algorithm has a much better performance than the original algorithms and many other mainstream BSS algorithms.

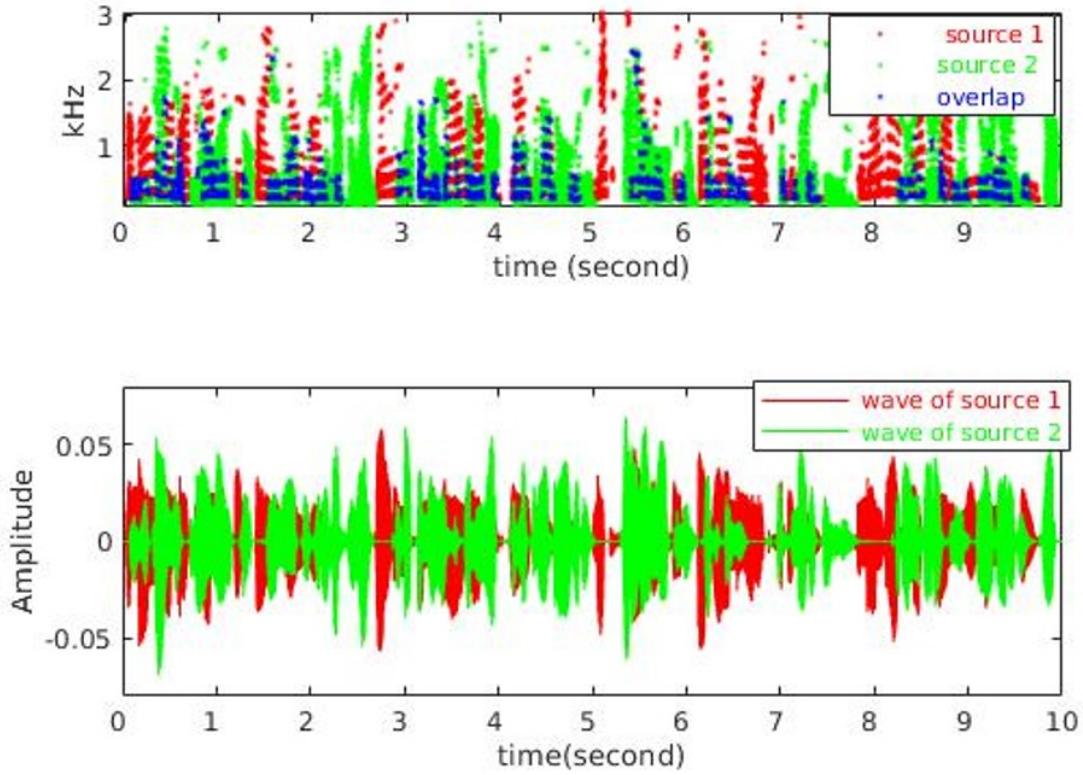

Figure 1.3: From top to bottom, waveform and spectrogram of two speech signals, respectively. Speech source 1 is colored red, and speech source 2 is colored green. In the spectrogram, the blue region is where the red and green region overlap.

- Improving l_1 minimization: the proposed algorithm not only improves the performance, but also reduces the time consumption.

Cancellation kernel is used to extract a very weak target signal from mixed signals. But computing the kernel is slow for real-time processing and requires an interval where the target signal is inactive. Our first contribution is to extend the cancellation kernel from the time domain to the time-frequency domain by applying the short-time Fourier transform (STFT). Compared to the existing cancellation kernel that only can be computed iteratively, the new cancellation kernel has analytic expression, making its computation more convenient and in-depth analysis possible. For the performance, the new version of the cancellation kernel has a comparable, even better performance than the previous one. Moreover, the proposed cancellation kernel can be computed even without the time interval where the target signal is inactive.

DUET is a binary masking-based method separating theoretically any number of

sources with two microphones. It is simple and cost-effective but the performance is not satisfactory. One reason is that binary masking can not completely separate the speech signal. The second contribution of this thesis is we propose to use spatial filter, which is essentially a weighted summation, as a post-filtering. The results show that our proposed method is not only far superior to DUET, but also faster and more effective than some mainstream blind source separation algorithms.

In our third contribution, we improved the l_1 minimization for speech separation. l_1 minimization is a convex optimization framework aiming to estimate the original signals by minimizing the l_1 norm of feasible signals under some fidelity constraints. The optimal solution of l_1 minimization is sparse, i.e., it only depends on a few non-zero coefficients. If the original solution is sparse enough, there is a good chance that they are equal. When the signals are speech signals, we find there is a better alternative than l_1 minimization, that is weighted l_1 minimization. We propose to use a weighted l_1 norm as the optimization objective. By different penalties for different coefficients in the solution, we can obtain a more sparse solution that may be closer to the original solution. Experiments show that our proposed weighted l_1 minimization indeed outperforms the plain l_1 minimization. This superiority is robust to disturbances on fidelity constraints. In addition, theoretically, the proposed weighted l_1 minimization has the same algorithm time complexity as plain l_1 minimization but in practice, it can approach convergence more quickly. Better performance can be obtained by introducing a reweighted scheme, that is, using the solution of previous weighted l_1 minimization as the current weights.

1.3 Outline

The rest of the thesis is divided into five chapters. Chapter 2 starts from the physics of sound and derives from the wave equation two principles of sound - superposition and reflection. Based on the two principles, we can have the time domain model describing the relationship between the source signals and the received signals by microphones. The time-domain model can be converted to a time-frequency domain model via the short-time Fourier transform. Chapter 3 to chapter 5 are the main body of this thesis. We introduce cancellation kernel, DUET, and l_1 minimization and their improvements, respectively. Chapter 6 summarizes this thesis, reiterates our contributions, and points out current

limitations and some interesting future work.

Chapter 2

Mathematical Models for BSS

In this chapter, we will cover some of the physics of sound, including how it is produced, how it propagates, and its equations of motion. Then we will introduce some mathematical models for blind source separation (BSS). These models are equations that assume the relationship between observed mixed signals and source signals. Directly solving these equations is literally impossible unless under some conditions, e.g., source signals are independent or have some sparse representation. Depending on whether the Short-time Fourier transform (STFT) is applied, these models can be divided into two categories, one that describes their relationship in the time domain and the other that describes their relationship in the time-frequency domain. Finally, we will compare the advantages and disadvantages of these two types of models.

2.1 Physics

2.1.1 Transmission of sound

Sound is produced by a fast vibration of an object. When an object vibrates, the surrounding air in the direction of motion is compressed. Note that, this vibration must be fast enough or the surrounding air will just flow over the object, not be compressed. The compressed air will generate extra pressure, pushing the air around it. The propelled air is compressed again, and so on, and the vibration of the air spreads out. When the state of local motion of the air caused by the extra pressure is propagated to the human ear, the eardrum is hit by the moving air, creating a perceptual stimulus that is finally perceived by the human brain as sound. We also commonly refer to airborne vibrations as sound. In addition to the propagation of sound in the gas, sound can also propagate in liquids or

even solids, and the principle is the same as that in the gas, except that the propagation speed is different. Generally, the sound propagation in a solid is the fastest, slower in a liquid, and slowest in a gas. The speed of sound in air is about 343 meters per second at $20°C$. It highly depends on the temperature but is nearly independent of frequency.

2.1.2 Wave equation

As we explained that sound is actually a vibration that propagates through air[1] , it is natural to ask how the air moves with the time given a certain space position, or how the movement of air varies with space at a given moment. We first consider the simplest situation where sound only propagates in one direction but this is enough to draw out some properties of sound. Denote $\chi(x,t)$ the displacement of air at space position x and time t. Assuming the extra pressure due to the vibration of air is relatively small compared to the air pressure at equilibrium, then it can be deduced from Newton's laws of motion that $\chi(x,t)$ satisfies the following second-order partial differential equation

$$\frac{\partial^2 \chi}{\partial x^2} = \frac{1}{\varkappa}\frac{\partial^2 \chi}{\partial t^2}, \tag{2.1.1}$$

where $\varkappa$ is the ratio of changing pressure P_ϵ and changing air density ρ_ϵ at the equilibrium state (equilibratory air will be compressed and its density will change when sound propagates)

$$\varkappa = \frac{P_\epsilon}{\rho_\epsilon} = \left(\frac{\mathrm{d}P}{\mathrm{d}\rho}\right)_0. \tag{2.1.2}$$

Now we examine by the motion equation (2.1.1) that whether the sound is a propagated vibration as we described before. If the status of air vibration propagates in a direction, then $\chi(x,t)$ has to satisfy the function formation $f(x-vt)$ where v is the speed of propagation. Now we substitute $\chi(x,t) = f(x-vt)$ into (2.1.1), then we realize that $f(x-vt)$ is a solution of (2.1.1) if $v = \pm\sqrt{\varkappa}$ where $+\sqrt{\varkappa}$ means sound propagates in the positive direction, and $-\sqrt{\varkappa}$ means sound propagates in the negative direction. Therefore, we verified that sound is a propagating vibration by its motion equation (2.1.1) and and have obtained a byproduct that the speed of sound depends on the property of the propagation medium, specifically, the response of its pressure to the change of density. Equation (2.1.1) is therefore named the wave equation, and sound is also called sound wave.

[1] Actually sound can travel the same way in a liquid, or in a solid, but we only mention air here since our topic is speech and speech normally propagates in air

2.1.3 Superposition & reflection of sound

If there are two sound waves, $\chi_1(x,t)$ and $\chi_2(x,t)$, meet at the same location, then they will be superposed there, resulting in a third sound wave $\chi_3(x,t) = \chi_1(x,t) + \chi_2(x,t)$. This is called the superposition principle. It can be easily proved that if $\chi_1(x,t)$ and $\chi_2(x,t)$ satisfy the wave equation (2.1.1), then $\chi_3(x,t)$ also satisfies the equation, hence, the superposition principle does not violate the laws of motion.

Sound is a wave just like light, and a sound wave also has the properties of reflection, refraction, and diffraction that a light wave has. In this thesis, we do an appropriate simplification and only focus on the reflection of sound. This is reasonable because for sound waves, reflection effects are the most common and obvious in real life, while refraction and diffraction are relatively negligible. We will use the following example to elicit the principle of reflection. Consider a sound wave hitting a rigid object with dimensions much larger than its wavelength, e.g., a wall. We consider a rigid object because we don't want to consider refraction (a refracted sound wave can not be created in a rigid object since such an object can not be compressed), and we consider objects with dimensions larger than sound wavelength because we want to avoid involving diffraction. In the one-dimensional case[2] , the general solution to the wave equation (2.1.1) is $\chi(x,t) = f(x-vt) + g(x+vt)$. Due to the displacement at the hitting point (assuming it is the origin point) is zero, substituting this boundary condition into the general solution of the wave equation, then we can have the motion of sound when reflection happens: $\chi(x,t) = f(x-vt) - f(-x-vt)$. If we reverse the superposition principle, we will realize the $\chi(x,t)$ is a superposition of two identical waves, except with opposite vibration and propagation direction. One wave is the incident wave, and the other wave is called the reflected wave. Normally, we call the wave moving toward the wall the incident wave, and the wave moving away from the wall the reflected wave. From the above discussion, we can conclude that the principle of reflection is to treat the wall as a mirror, and the reflected wave from the virtual sound source that is mirror-symmetrical with the real sound source continues to propagate forward through the hitting point. Note that, the reflected wave has the opposite vibration to the incident wave, or we say the reflected wave has a 180° phase difference with the incident wave. This is the discussion in the mathematical sense, and when we go back to the physical world, we can only observe the waveform on the air side while the waveform

[2]One-dimensional case is enough to draw out the reflection principle.

on the wall side can only be imagined.

Although we only discussed a simple case in one-dimension[3] , we already have enough principles and intuition to develop mathematical models for BSS.

2.2 Time Domain Model

The time-domain model assumes the relationship between source signals and mixed signals in the time domain. This model is very straightforward and easy to understand because building it only needs the superposition and reflection principles (see Sec. 2.1.3 for the two principles) which are very intuitive. When there is a point sound source at point j in a room, the sound field at another point i in the room is a superposition of the direct sound wave from this point sound source and reflected wave from the wall of the room. Note that the assumptions here, we assume that the sound source is a point sound source, and if the sound source is not a point sound source, e.g., wind, the time domain model is not valid anymore. Fortunately, if the room is quite large and the distance between point j and point i is far, it is reasonable to treat a human speaker as an ideal point sound source. Denote $s_j(t)$ as the sound source at point j, according to the superposition and reflection principles, the sound field at point i is

$$x_i(t) = \sum_{\tau=0}^{\infty} a_{ij}(\tau) s_j(t-\tau), \tag{2.2.1}$$

where $a_{ij}(\tau)$ is called the room impulse response (RIR) or acoustic impulse response (AIR) in literature, which can be a positive or negative real value. When there is no reverberation, RIR from point j to point i is

$$a_{ij}(t) = \frac{1}{\sqrt{4\pi} d_{ij}} \mathbb{1}_{t-d_{ij}/c}, \tag{2.2.2}$$

where d_{ij} is the distance between point i and point j, $\frac{1}{\sqrt{4\pi} d_{ij}}$ is the propagation attenuation since the energy of sound is radiated out in the form of a sphere, $\mathbb{1}_{t-d_{ij}/c} = 1$ when $t - d_{ij}/c = 0$ and 0 otherwise, d_{ij}/c is the propagation time of sound between point j and point i, and c is the velocity of sound, around 343 m/s at $20°C$. RIR like (2.2.2) only considers the direct path from the sound point to the microphone. In general, an RIR does not only contain the propagation attenuation and delay of the direct path, it

[3]For a more complicated case in three-dimension, we can always obtain the motion of sound by solving the wave equation under the boundary conditions.

also contains the attenuation, delay, and 180° phase difference of the reflected path. Note that, we assume that the wall in the room is not a completely rigid object now, so it will absorb the energy[4] of sound, hence, the reflected wave will be weaker than the incident wave. If define the absorption coefficients α the ratio of absorbed energy to the incident sound energy per unit of area of material, it is around the level of 0.04 on a natural brick wall, 0.02 on a painted brick wall, 0.1 on plywood, 0.07 on a wood floor and concrete floor[5] , etc. Note that, the absorption coefficient varies in different frequencies, the above values are an average overall frequency range.

Observe (2.2.1), if s_j is an impulse signal, i.e., $s_j(0) = 1$, and $s_j(t) = 0, \forall t > 0$, then $x_i(t) = a_{ij}(t)$. Therefore, one can measure $a_{ij}(t)$ by producing an impulse signal at point j and collecting the response signal at point i. Figure 2.1 depicts[6] a real room impulse response from *dEchorate* [1] dataset. As shown in the figure, an RIR can be divided into

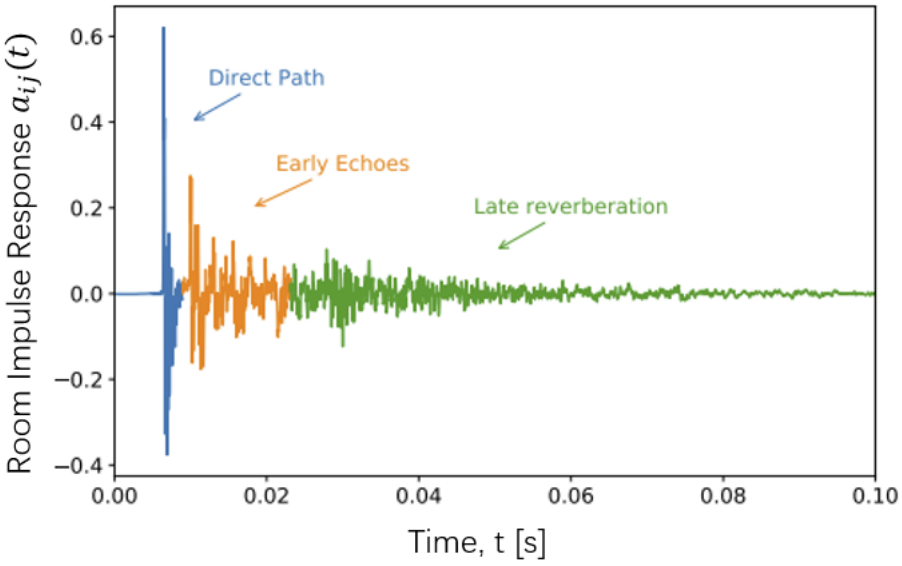

Figure 2.1: Depiction of a first 100 ms of real room impulse response from [1].

three parts. The first part that is colored blue, called direct path, is the impulse signal that travels from the source point i to the measurement point j via a straight path, so this part has the least time delay and the strongest amplitude. The second part which is colored yellow, called early echoes, consists of the earliest batches of echoes. At this time,

[4]The absorbed energy will exist in the form of a refracted wave. This will not happen in a completely rigid object because such an object can not be compressed, so a refracted wave can not be created.

[5]The above absorption coefficients of materials are all from this web page "https://www.acoustic-supplies.com/absorption-coefficient-chart/"

[6]The figure is copied from the paper [1] and modified by the author of this thesis.

these reflected sound waves can still be distinguished from each other due to their rarity. But after a while, as the number of reflections increases, the reflected sound waves become indistinguishable and the amplitude decays exponentially, and this is the third part that is colored green, called the late reverberation. The amplitude of the reflected sound wave in the late reverberation decays exponentially. There is a term called reverberation time 60 dB, RT_{60}, to quantify the time that it takes the sound to fade away. The definition of RT_{60} is the time it takes for the sound pressure level of the reflections to reduce by 60 dB compared to the first arrival sound. RT_{60} depends on the volume of the room, materials of the wall, furniture inside the room, etc. The RT_{60} is about 0 outdoors due to lack of planes to provide reflections, around 50 ms in a car, 0.3 s in a recording room (volume less than 50 m^3), 0.4s - 0.6s in a classroom (volume less than 200 m^3), 1.4 - 2.0 in a concert hall (volume less than 20'000 m^3), 2 - 10 s in a church, etc[7] .

Considering multiple sound sources and noise, by the superposition principle, (2.2.1) can be more generalized expressed as

$$x_i(t) = \sum_{j=1}^{J} (a_{ij} * s_j)(t) + n_i(t), \tag{2.2.3}$$

, where * is the convolution operator, J is the number of sources, $n_i(t)$ is noise signal. In this thesis, we only focus on the speech separation rather than noise deduction, hence, we assume $n_i(t) = 0$. We also only consider the separation when sound sources are static, i.e, not moving.

2.3 Time-frequency Domain Model

Although the time domain model is straightforward and accurately describes the relationship between the source signals and mixed signals, convolution is not mathematically easy to handle, in addition, an RIR in the presence of mild or moderate reverberation consists of hundreds or even thousands of coefficients, which greatly increases the difficulty of processing. By applying the Short-time Fourier transform (STFT) [27], this limitation can be addressed.

[7]Part of above values come from the web page "https://www.nti-audio.com/en/applications/room-building-acoustics/reverberation-time"

2.3.1 Multiplicative model

Applying STFT on the left hand and right hand of the equation (2.2.3), we can have a new equation that describes the relationship between the STFT coefficients of the sound field observed by microphone at the point i $x_i(t, f)$ and source signal at the point j $s_j(t, f)$:

$$x_i(t, f) = \sum_{j=1}^{J} a_{ij}(f) s_j(t, f) + n_i(t, f), \tag{2.3.1}$$

where $a_{ij}(f)$ called the acoustic transfer function (ATF) is the Fourier transform of RIR $a_{ij}(t)$, and $n_i(t, f)$ is the STFT coefficients of the noise signal $n_i(t)$. For example, when there is no reverberation, the ATF from point j to point i is the Fourier transform of (2.2.2):

$$a_{ij}(f) = \frac{1}{\sqrt{4\pi} d_{ij}} \exp -\imath 2\pi f d_{ij}/c, \tag{2.3.2}$$

where $\imath = \sqrt{-1}$. Note that, in order to transform (2.2.3) into (2.3.1) accurately, the window length of STFT must be longer than twice of the length of RIR. In literature [20, 22], they claim that (2.3.1) is valid when the STFT window length is longer than the length of RIR. A suitable STFT window length with respect to the length of RIR has been studied in [28]. We also investigate the effect of the STFT window length on our speech separation methods in this thesis. Two contributions have been made by (2.3.1). First, it turns convolution in (2.2.3) into multiplication, which provides convenience in math. Second, each frequency band can be dealt with independently, which greatly reduces the number of parameters or model complexity. However, there are two new problems introduced by (2.3.1). The first problem is called the scale ambiguity, that is the ambiguity in determining the variance (energy) of source signal $s_j(t, f)$ since any scale on $s_j(t, f)$ can be cancelled by dividing $a_{ij}(f)$ by the same scale. In independent component analysis (ICA), the scale ambiguity is solved by assuming $s_j(t, f)$ has a unit variance, i.e., $\sum_t s_j^2(t, f) = 1$ (assuming that $s_j(t, f)$ has zero mean). Another alternative is normalizing the amplitude of $a_{ij}(f)$, such as the relative transfer function (RTF) defined in [29]

$$\tilde{a}_{ij}(f) = a_{ij}(f)/a_{1j}(f), \quad \forall j. \tag{2.3.3}$$

Obviously, by choosing different reference microphone, there are different RTFs. To eliminate this difference, one can normalize the ATF by

$$\tilde{a}_{ij}(f) = a_{ij}(f)/\sqrt{\sum_i a_{ij}^2(f)}, \quad \forall j, \tag{2.3.4}$$

which is often used in the l_1 minimization based speech separation.

The second problem of the multiplicative model (2.3.1) is the permutation ambiguity, which is the ambiguity in determining the order of sources for each frequency band. For example, assume that we have two source signal estimations for two frequency bands, respectively: $\hat{s}_1(t, f_1)$ and $\hat{s}_1(t, f_2)$, but they may be corresponding to different sources since we dealt with the two frequency bands independently. A typical framework solving the permutation ambiguity is to apply some alignment methods [30, 31] to the separated frequency-wise signals.

2.3.2 Convolutive model

The multiplicative model (2.3.1) is valid when the STFT window length is larger than the RIR length. But a too long STFT window will reduce the performance of a separation algorithm (see the discussion in the next section). The authors in [32] made it possible to represent longer reverberations with shorter STFT windows by using a convolutional model:

$$x_i(t, f) = \sum_{j=1}^{J} \sum_{\tau} a_{ij}(\tau, f) s_j(t - \tau, f) + n_i(t, f), \tag{2.3.5}$$

where $a_{ij}(t, f)$ is called the convolutive transfer function (CTF) from the j-th source to the i-th microphone. This time-frequency domain convolution model was rarely used because it is more complex and requires more parameter estimation, but recently it has been slowly used to deal with high reverberation speech separation problems.

2.4 Limitation of Reverberation

The most tractable model of the above should be the multiplicative model in the time-frequency domain (see Sec. 2.3.1). This model is also called the narrowband approximation model in literature [33]. As the most prevalent model, the narrowband approximation model significantly simplifies the blind source separation (BSS) process: 1. The required convolution operation in the time domain model is converted to a multiplication in the frequency domain, which makes analytical processing more easier. 2. Each frequency band is treated independently to reduce the dimension of the problem. However, there is a limitation for the narrowband approximation model in the presence of reverberation. On the one hand, this model only holds when the window length of STFT is sufficiently

long, proportional to the reverberation time 60 dB (RT_{60}, see Sec. 2.2). on the other hand, a longer window length will add more frequency bands and reduce the number of time frames which can be used to train the parameters of each frequency band. Hence, it is not surprising that longer reverberation results in poor separation performance in the experimental part of some BSS algorithm literature [6, 8, 22]. In different scenarios, an optimal window length is determined by the trade-off between obtaining enough training data and model complexity [28]. This is a fundamental limitation of the narrowband approximation model. There are algorithms obtaining a better result with a shorter window length of STFT by using the convolutive model (see Sec. 2.3.2) [22, 34–36] and full-rank covariance model [8] in the time-frequency domain. Nevertheless, their performance is still underwhelming under a severe reverberant environment with RT_{60} longer than 500 ms, and they are more time consuming.

The best performing BSS models under reverberation would be the time domain model (see Sec. 2.2), also known as the wideband model [20, 28]. The wideband model is derived directly based on the superposition and reflection principle. It is the best fit model for the real physical world under reasonable simplification. Although a good performance has been demonstrated under a high reverberant condition (RT_{60}=890 ms), they assumed that the room impulse responses (RIRs) were known in advance, limiting the practical use. Since it is challenging to simultaneously estimate each source's RIR from a mixture signal, solving BSS using the wideband model is still a long-term goal.

In summary, reverberation is a difficulty in BSS. For the time-frequency domain BSS, there is a trade-off between the large number of parameters to cover a long reverberation and the lack of data to train these parameters as explained above. Even the observed signals are sufficiently long, high-latency processing and complex computations that accompany long window length are still problems in the real application. For the time domain BSS, accurately and robustly estimating the RIR of each source from a mixture signal remains a challenge. High computational cost is another limitation.

The above discussions are all non-deep learning based models. For the deep learning based model, it can be observed that reverberation still significantly reduces the final performance in the task of speech enhancement [37–40]. One of the reason may be the trade-off between the complexity of the deep learning model and the amount of training data. Even with sufficient data, over-fitting of a complicated model is also an issue when

there is no reverberation

Chapter 3

Improving cancellation kernel

There are various blind source separation (BSS) algorithms applied to solve the cocktail party problem according to different scenarios or conditions. They have their own advantages and limitations respectively. Cancellation kernel, first proposed by Wang and Zhou [15], has its uniqueness to extract the target signals when the target signals are much weaker than the interference (i.e., unwanted) signals. The idea is to "learn" a set of convolutive kernels from an interval where the target signals are inactive (hereafter, referred to as the silent interval). Once the kernels are calculated, the interference signals can be removed by a summation of convolutions between the mixed signals and the convolutive kernels, and only the target signals remain (therefore, the convolution kernels are called the cancellation kernels). Convolution can be done very fast, hence, real-time extraction is another advantage for the cancellation kernel. It is proved in [15] that there exist such cancellation kernels which can completely remove the interference signals by a summation of convolutions. They also formulated quadratic programming with non-negative constraints to compute the cancellation kernels. Although the extracted signal is distorted because of the convolution, its intelligibility is not compromised, and it is less annoying than an "unclean" signal generated by independent component analysis (ICA) based methods and binary masking based methods. Further, Yu et al. [3] proposed to use l_1 regularization to regulate the kernels instead of non-negative constraints and used a split Bregman algorithm to compute the kernels. By doing so, the kernels are more sparse (i.e., have more zero coefficients), hence, introducing less distortion. The computing speed is also improved by using the split Bregman algorithm. For more details please see Sec. 3.1.

However, there are two problems with the current cancellation kernel method. First, even for the computationally efficient version - the l_1 regularized kernel, the computation of

the kernel is rather slow. Moreover, any movement of the interference signal source or the change of the reverberation environment requires recalculation of the cancellation kernel, which severely limits the real-time use. The second problem is that the computation of kernels requires a silent interval and this is very inconvenient and limits practical applications. In this thesis, a new version of the cancellation kernel, called interference suppression response (ISR), is proposed to be faster and more convenient to use. We use the Short-time Fourier transform (STFT) to transform the time-series signals into time-varying, multi-frequency bands signals. As different frequency bands are assumed to be independent of each other, we deal with them separately. For each frequency band, the mixing model is simpler than that in the time domain, and the length of ISR can be much shorter than previous cancellation kernels. This enables us to compute the kernels by formulating quadratic programming having an analytic solution. Finally, the inverse STFT will be used to transform the time-varying, multi-frequency band signal back to the time-series signal. Compared to the previous cancellation kernel, the proposed ISR has two extra advantages. First, due to the existence of an analytic solution, the time of computing ISR is negligible. The only time consuming part is the STFT and the inverse STFT but both of them can be done in real-time. Second, it is not necessary to compute ISR by finding a silent interval. Because speech signals will not completely overlap in the time-frequency domain (see Sec. 1.1.3), sufficient information can be obtained to compute ISR from the partially-disjoint region, which we will demonstrate in our second work in chapter 4. The high efficiency makes ISR more suitable in a dynamic environment, the analytic solution makes deeper analysis possible, and with no requirement of silent interval makes ISR more practical..

3.1 Cancellation Kernel

Let us consider the simplest case for signal separation, that is, there are two microphones and two signal sources. Denote $x_1(t)$, $x_2(t)$ the mixed signals received by the two microphones respectively, and $s_1(t)$, $s_2(t)$ the two source signals, respectively. According to the multi-source time domain model (2.2.3), assuming there is no noise, they satisfy

$$\begin{aligned} x_1(t) &= (a_{11} * s_1)(t) + (a_{12} * s_2)(t), \\ x_2(t) &= (a_{21} * s_1)(t) + (a_{22} * s_2)(t), \end{aligned} \tag{3.1.1}$$

where $*$ is the linear convolution operation, a_{ij} for $i = 1, 2$ and $j = 1, 2$ are the room impulse responses (RIRs, see Sec. 2.2) from j-th source to the i-th microphone. If $s_2(t)$ is the target signal in which we are interested, $s_1(t)$ is the interference signal which we want to get rid of, we can have two cancellation kernels $b_1 = a_{21}$ and $b_2 = -a_{11}$ such that the interference signal s_1 can be removed by $(b_1 * x_1) + (b_2 * x_2) = (b_1 * a_{12} + b_2 * a_{22}) * s_2$. Although the target signal is subject to another convolution, the intelligibility of the target signal is not affected according to our perception. In addition, the convolution is moderate if we can have a short and not dense cancellation kernel, and a human auditory system seems less sensitive to moderate convolution than artifacts generated by ICA or binary masking-based methods. In general, if there are I microphones, J signal sources, from 1 to P are interference signal sources, is there a group of cancellation kernels $b_1, \ldots, b_I$ such that

$$\sum_{i=1}^{I} b_i * x_i = \sum_{j=P+1}^{J} (\sum_{i=1}^{I} b_i * a_{ij}) * s_j. \tag{3.1.2}$$

It has been already proved in [15] that such non-all-zero cancellation kernels exist as long as the number of microphones is strictly larger than the number of interference signal sources, i.e., $I > P$.

3.1.1 Background learning

The next question is how to compute the cancellation kernels. For the simplest case, i.e., two microphone and two signal sources, if s_1 is the target signal, then you can set $b_1 = a_{22}$ and $b_2 = -a_{12}$. Generally, if there are I microphones, J signal sources, from 1 to P are interference signal sources, cancellation kernels should satisfy

$$\sum_{i=1}^{I} b_i * a_{ij} = 0, \quad \forall j = 1, \ldots, P. \tag{3.1.3}$$

But the problem is that we do not know a_{ij}. The solution given in [15] is to find a time interval $\alpha \leq t \leq \omega$ where only the target signals $s_P, \ldots, s_J$ are inactive or silent, that is, in the time interval $\alpha \leq t \leq \omega$, $s_P(t) = \cdots = s_J(t) = 0$, and all the interference signals $s_1, \ldots, s_P$ are non-silent, cancellation kernels can be calculated by minimizing the cost function

$$F(b_1, \ldots, b_I, \alpha, \omega) = \sum_{\alpha \leq t \leq \omega} |\sum_{i=1}^{I} (b_i * x_i \mathbb{1}_{\alpha \leq t \leq \omega})(t)|^2, \tag{3.1.4}$$

where $x_i \mathbb{1}_{\alpha \leq t \leq \omega}(t) = x_i(t)$ for $\alpha \leq t \leq \omega$ and 0 otherwise. The time interval $\alpha \leq t \leq \omega$ where the target signals are silent is called the silent interval. Fig. 3.1 depicts a

silent interval when there are only one target signal and one interference signal. This thesis assumes the silent interval can be and has been successfully detected. The design of the detection algorithms is closely related to practical applications. Depending on the subjective definition of the target signal and the interference signal, a number of completely different silent interval detection algorithms can be designed to detect the silent interval by detecting direction of the signal, speech or non-speech, male or female voice, etc. One can design a detection algorithm using the different angles of the target signal and the interference signal [3], or based on some speech activity detection algorithms [41, 42].

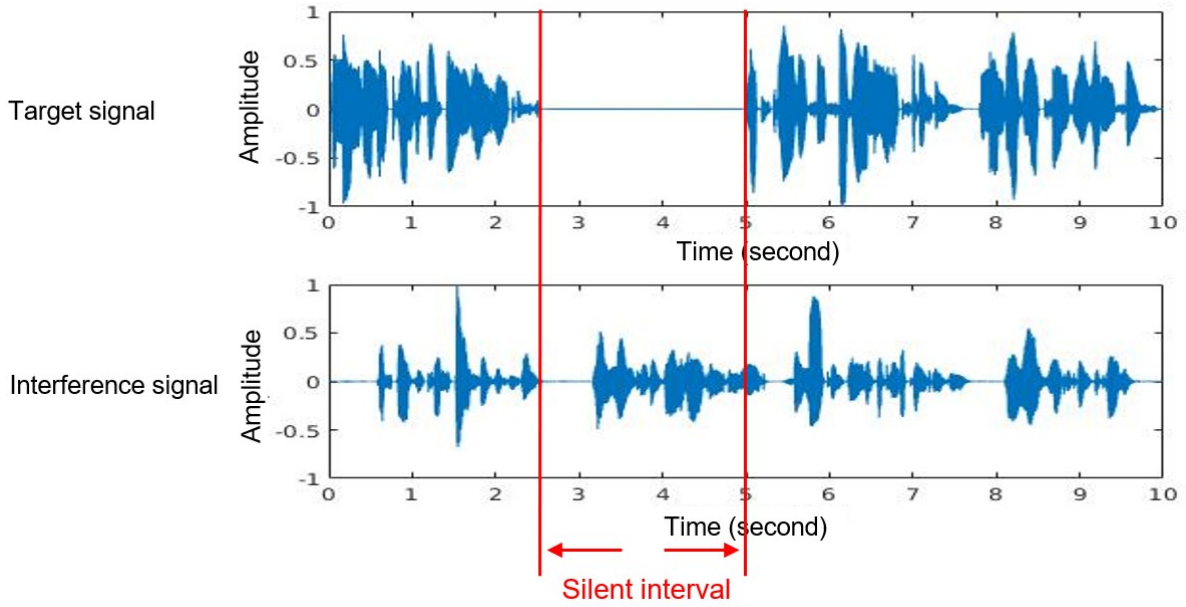

Figure 3.1: Depiction of a silent interval of one target signal. From top to bottom: the target signal, the interference signal. The range (around 2.5 s to 5 s) between the two red line is the silent interval for the target signal.

Minimizing the cost function (3.1.4) without any constraint will not get you ideal cancellation kernels. There are at least three issues. First, you may get trivial all zero solutions, that is, $b_1, \ldots, b_I = 0$. Second, solutions are not unique. For example, if $b_1^*, \ldots, b_I^*$ are solutions, then $cb_1^*, \ldots, cb_I^*$ are also solutions where c is an arbitrary constant. Third, you may obtain dense cancellation kernels which results in an over-fitting problem and seriously distorted the target signal as we mentioned in Sec. 3.1, the extracted target signals are subject to the further convolution of cancellation kernels. Therefore, aiming at the three issues, some constraints of cancellation kernels must be imposed to avoid trivial

all zero solutions, to obtain unique and sparse solutions. We will introduce two different versions of the cancellation kernel from different constraints or regulations in Sec. 3.1.2 and Sec. 3.1.3, respectively. However, both versions of the cancellation kernel are less useful in practice. It can be seen from (3.1.3) that any change in a_{ij} (e.g., interference source movement or change of reverberant environment) will change b_i, so in practice it is often necessary to recalculate b_i, but both versions of the cancellation kernel take some time to calculate and do not support real-time processing. This led us to invent the ISR, a time-frequency version of the cancellation kernel and having analytical expression that allow it to update in real-time in a dynamic environment. We will introduce ISR in Sec. 3.2.

3.1.2 Non-negative cancellation kernel

Imposing constraints is necessary for computing cancellation kernels as we just explained. In [15], the authors proposed a non-negative constraint, that is, $b_1 \geq 0, \ldots, b_I \geq 0$. We call such cancellation kernels non-negative cancellation kernels. The authors also imposed a constraint to confirm the uniqueness of the solution, that is, $\sum_{i=1}^{I} b_i(1) = 1$ where $b_i(1)$ is the first coefficient of the i-th cancellation kernel. Non-negative cancellation kernels can be calculated via quadratic programming in the case of two microphones, i.e., two mixed signals, and two source signals. Suppose the length of each cancellation kernel is L, define $t' = t + 1$ for brevity, rewrite the cost function (3.1.4) as

$$\begin{aligned} F(b_1, \ldots, b_I, \alpha, \omega) &= \sum_{\alpha \leq t \leq \omega} |\sum_{i=1}^{I} (b_i * x_i \mathbb{1}_{\alpha \leq t \leq \omega})(t)|^2 \\ &= \sum_{\alpha \leq t \leq \omega} \left(\sum_{l=1}^{L} (b_1(l) x_1(t'-l) \mathbb{1}_{t'-l \geq \alpha} + b_2(l) x_2(t'-l) \mathbb{1}_{t'-l \geq \alpha}) \right)^2 \\ &= b_1^{\mathsf{T}} \mathbf{A} b_1 + b_2^{\mathsf{T}} \mathbf{B} b_2 + 2 b_1^{\mathsf{T}} \mathbf{C} b_2, \end{aligned} \tag{3.1.5}$$

where $(\cdot)^{\mathsf{T}}$ is the transpose, $\mathbf{A} = [A_{ij}]$, $\mathbf{B} = [B_{ij}]$,and $\mathbf{C} = [C_{ij}]$ with $i, j = 1, \ldots, L$ are $L \times L$ matrices given by

$$\begin{aligned} A_{ij} &= \sum_{t=\alpha}^{\omega} x_1(t'-i) x_1(t'-j), \\ B_{ij} &= \sum_{t=\alpha}^{\omega} x_2(t'-i) x_2(t'-j), \\ C_{ij} &= \sum_{t=\alpha}^{\omega} x_1(t'-i) x_2(t'-j). \end{aligned} \tag{3.1.6}$$

Finally, the non-negative cancellation kernels can be computed by quadratic programming

$$\begin{aligned} \min_{b_1,b_2} \quad & b_1^{\mathsf{T}}\mathbf{A}b_1 + b_2^{\mathsf{T}}\mathbf{B}b_2 + 2b_1^{\mathsf{T}}\mathbf{C}b_2 \\ \text{s.t.} \quad & b_1 \geq 0, \\ & b_2 \geq 0, \\ & b_1(1) + b_2(1) = 1. \end{aligned} \tag{3.1.7}$$

Fig. 3.2 depicts two non-negative cancellation kernels, which is copied from [15] for self-containment.

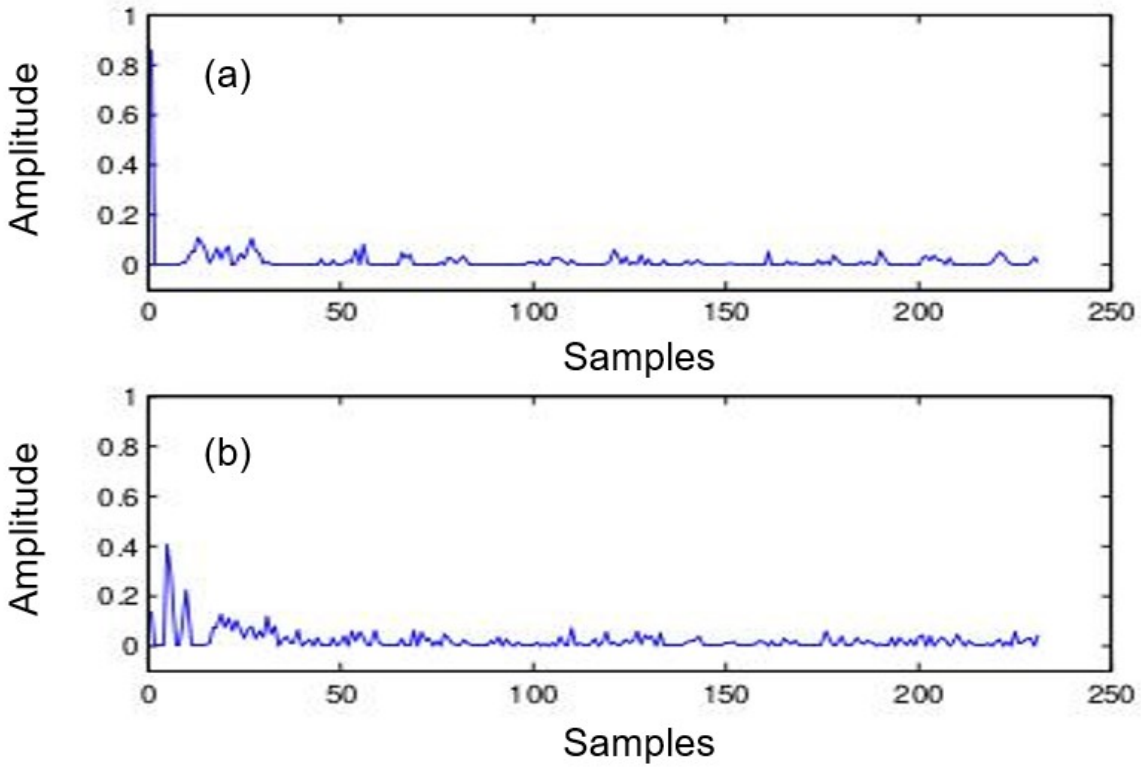

Figure 3.2: Depiction of two computed non-negative cancellation kernels. (a) The kernel b_1 that will be convolved with the mixed signal received by the first microphone. (b) The kernel b_2 that will be convolved with the mixed signal received by the second microphone.

3.1.3 l_1 regularized cancellation kernel

In [3], the authors propose sparse cancellation kernels whose sparsity is pursued by minimizing the l_1 norm of the kernels. Consider the simplest case where there are two microphones. Denote the mixed signals during a silent interval as d_1 and d_2, respectively.

They formulated a non-constraint convex optimization problem for computing the sparse cancellation kernels:

$$b_1, b_2 = \arg\min_{b_1,b_2} \frac{1}{2}\|b_1 * d_1 + b_2 * d_2\|_2^2 + \frac{\eta^2}{2}(b_1(1) + b_2(1) - 1)^2 + \mu(\|b_1\|_1 + \|b_2\|_1), \quad (3.1.8)$$

where η and μ are trade-off parameters for cross-channel cancellation and regularization, and $\|\cdot\|_1$ is the l_1 norm. The first regularization term $(b_1(1) + b_2(1) - 1)^2$ is to prevent all-zero solution, and the second regularization term $\|b_1\|_1 + \|b_2\|_1$ is to pursue the sparsity of b_1 and b_2. In this thesis, we call this version of cancellation kernel as l_1 regularized cancellation kernel. Yu et al. [3] also propose to use a split Bergman method to solve the optimization problem 3.1.8. It is claimed in [43] that compared to the non-negative cancellation kernel [15], the l_1 regularized cancellation kernel [3] is computationally efficient and can suppress a little bit more of the interference signals. Fig. 3.3 depicts two l_1 regularized cancellation kernels.

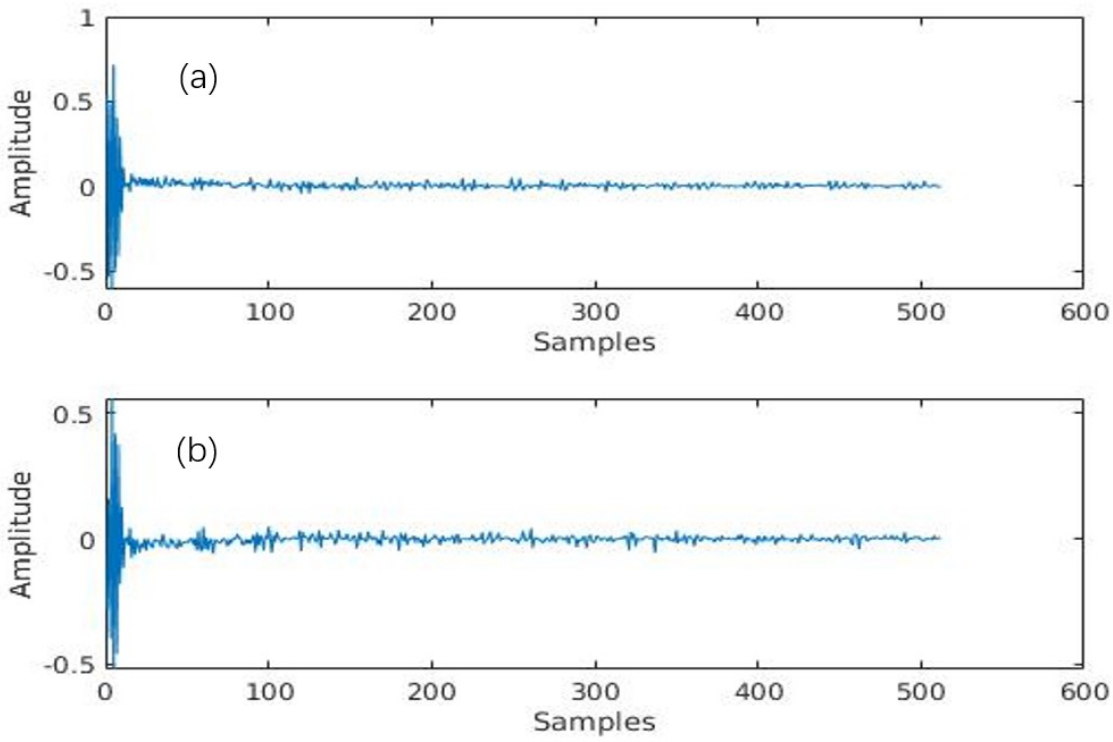

Figure 3.3: Depiction of two computed l_1 regularized cancellation kernels. (a) The kernel b_1 that will be convolved with the mixed signal received by the first microphone. (b) The kernel b_2 that will be convolved with the mixed signal received by the second microphone.

3.2 Improved Method

3.2.1 Multiplicative ISR

Let's start with the simplest case where there are two microphones and two speech sources. The difference with the cancellation kernel is that instead of considering the time domain model (2.2), we use the time-frequency domain multiplicative model (2.3.1):

$$\begin{aligned} x_1(t,f) &= a_{11}(f)s_1(t,f) + a_{12}(f)s_2(t,f), \\ x_2(t,f) &= a_{21}(f)s_1(t,f) + a_{22}(f)s_2(t,f), \end{aligned} \tag{3.2.1}$$

where $x_1(t,f)$ and $x_2(t,f)$ are the STFT coefficients of two mixed signals received by two microphones at time frame t and frequency bin f, respectively; $s_1(t,f)$ and $s_2(t,f)$ are the STFT coefficients of two speech source signals; $a_{ij}(f)$ for $i = 1,2$ and $j = 1,2$ are the acoustic transfer functions (ATF) (see Sec. 2.3.1) from the j-th source to the i-th microphone at the frequency bin f. Note that, here we assume there is no noisy signals. If $s_2(t,f)$ is the target signal and $s_1(t,f)$ is the interference signal, we can have two cancellation coefficients $w_1(f) = \overline{a}_{21}(f)$ and $w_2(f) = -\overline{a}_{11}(f)$ such that the interference signal can be removed by $\overline{w}_1(f)x_1(t,f) + \overline{w}_2(f)x_2(t,f) = (\overline{w}_1(f)a_{12}(f) + \overline{w}_2(f)a_{22}(f))s_2(t,f)$, where $\overline{(\cdot)}$ means the conjugation.

In general, if there are I microphones, J signal sources, from 1 to P are interference signal sources, denote $\mathbf{x}(t,f) = [x_1(t,f),\ldots,x_I(t,f)]^{\mathsf{T}}$, $\mathbf{s_u}(t,f) = [s_1(t,f),\ldots,s_P(t,f)]^{\mathsf{T}}$, and $\mathbf{s_t}(t,f) = [s_{P+1}(t,f),\ldots,s_J(t,f)]^{\mathsf{T}}$ the STFT coefficients of the observed signals (i.e., mixed signals), interference signals, and target signals, respectively, we extend (3.2.1) and rewrite it as the form of vector for compactness:

$$\mathbf{x}(t,f) = \mathbf{A_u}(f)\mathbf{s_u}(t,f) + \mathbf{A_t}(f)\mathbf{s_t}(t,f), \tag{3.2.2}$$

where the j-th column of $\mathbf{A_u}(f)$ is $[a_{1j}(f),\ldots,a_{Ij}(f)]^{\mathsf{T}}$ for $j = 1,\ldots,P$, and the j-th column of $\mathbf{A_t}(f)$ is $[a_{1(j+P)}(f),\ldots,a_{I(j+P)}(f)]^{\mathsf{T}}$ for $j = 1,\ldots,J-P$. $\mathbf{A_u}(f)$ and $\mathbf{A_t}(f)$ are also called the mixing matrix of interference signals and the target signals, respectively. To remove the interference signals, the cancellation coefficients $\mathbf{w}(f) = [w_1(f),\ldots,w_I(f)]^{\mathsf{T}}$ should satisfy:

$$\mathbf{w}^{\mathsf{H}}(f)\mathbf{A_u}(f) = 0, \tag{3.2.3}$$

where $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose. There are always non-all-zero solutions as long as the number of microphones is strictly larger than the number of interference signal

sources, i.e., $I > P$. The proof is omitted here as it is a fundamental knowledge of Linear Algebra. Then we can estimate the target signals by

$$\begin{aligned}\widehat{\mathbf{s}}_{\mathbf{t}}(t,f) &= \mathbf{w}^{\mathsf{H}}(f)\mathbf{x}(t,f) \\ &= (\mathbf{w}^{\mathsf{H}}(f)\mathbf{A}_{\mathbf{t}}(f))\mathbf{s}_{\mathbf{t}}(t,f).\end{aligned} \tag{3.2.4}$$

Since $\mathbf{A}_{\mathbf{u}}(f)$ is unknown, we adopt the background learning for computing $\mathbf{w}(f)$ (see Sec. 3.1.1). A constraint is also needed for avoiding an all-zero solution. Here, we use a singleton linear constraint $w_1(f) = 1$. For the non-negative cancellation kernel (see Sec. 3.1.2) and the l_1 regularized cancellation kernel (see Sec. 3.1.3), the number of parameters of each kernel is very large, so it is necessary to introduce sparsity for a good performance, but leads to the difficulty of calculation. However, for the proposed cancellation kernel in the time-frequency domain, i.e., $\mathbf{w}(f)$, different frequency bands are dealt with independently and for each frequency band, there is only one parameter for each kernel. Hence, there is no need to pursue a sparsity. This is a definite advantage over the time domain cancellation kernel. Denote $\mathbf{d}(t,f) = [d_1(t,f),\ldots,d_I(t,f)]^{\mathsf{T}}$ the STFT coefficients of the mixed signals during the silent interval, $\mathbf{w}(f)$ can be computed by solving the below quadratic convex optimization problem:

$$\begin{aligned}\min_{\mathbf{w}(f)} \quad & \sum_t |\mathbf{w}^{\mathsf{H}}(f)\mathbf{d}(t,f)|^2 \\ \text{s.t.} \quad & w_1(f) = 1.\end{aligned} \tag{3.2.5}$$

The optimal solution of the above convex optimization problem has an analytic expression which will be given in Sec. 3.2.2. To differentiate this time-frequency version cancellation kernel from the time domain cancellation kernel introduced in Sec. 3.1, we call the time-frequency version cancellation kernel the interference suppression response (ISR).

Computing frame by frame

Observe the optimization problem (3.2.5), the proposed ISR can be computed frame by frame following a least mean square (LMS) adaptive filter optimization criteria. Here, we directly give the frame-wise adaptive solution:

$$w_i^{(t)}(f) = \begin{cases} 1 & \text{if} \quad i = 1 \\ w_i^{(t-1)}(f) - \frac{\mu\overline{e}(t,f)d_i(t,f)}{\sum_{j=2}^{J}\overline{d}_i(t,f)d_i(t,f)} & 2 \le i \le I \end{cases}, \tag{3.2.6}$$

where $w_i^{(t)}(f)$ is ISR updated at t-th time frame, $e(t,f) = \mathbf{w}^{(t)}(f)^{\mathsf{H}}\mathbf{d}(t,f)$, and μ is learning rate and for normalized least mean square filter (NLMS) the optimal value $\mu_{\text{opt}} = 1$.

Distortionless response

If there is only one target signal (denote its ATF by $\mathbf{a}(f)$), the distortion on the target signal (see (3.2.4)) can be eliminated by the modified ISR $\mathbf{w}' = (\mathbf{a}^{\mathsf{H}}\mathbf{w})^{-1}\mathbf{w}$, where frequency index f is omitted for compactness.

3.2.2 Convolutive ISR

The time-frequency domain multiplicative model (2.3.1) becomes less accurate when the room impulse response (RIR) length is longer than the STFT window length. Nevertheless, an STFT window that is too long will make ISR introduce more distortion on the target signal, leading to poor performance. Therefore, it is not difficult to foresee that the above ISR will not work well in a high reverberation environment. Here, we will apply the time-frequency domain convolutive model (2.3.5) and design a convolutive version of the ISR to cope with high reverberation environments.

Starting with the simplest case, that is, two microphones and two speech sources, continuing the notations from the previous section, the time-frequency convolutive model (2.3.5) can be rewritten as:

$$\begin{aligned} x_1(t,f) &= \sum_{\tau} a_{11}(\tau,f) * s_1(t-\tau,f) + \sum_{\tau} a_{12}(\tau,f) * s_2(t-\tau,f), \\ x_2(t,f) &= \sum_{\tau} a_{21}(\tau,f) * s_1(t-\tau,f) + \sum_{\tau} a_{22}(\tau,f) * s_2(t-\tau,f), \end{aligned} \tag{3.2.7}$$

where $a_{ij}(t,f)$ for $i = 1,2$ and $j = 1,2$ is the convolutive transfer function (CTF) from the j-th source to the i-th microphone. Since we deal with each frequency independently, we omit the frequency index f in the rest of the section. Observe (3.2.7) and (3.1.1), they have exactly the same expression if you open the convolution operation in (3.1.1). Therefore, we derive the convolutive ISR following the same recipe of the cancellation kernel. The two convolutive ISRs applied to two channels of mixed signal are denoted as $\mathbf{w}_1 = [w_1(1), \ldots, w_1(K)]^{\mathsf{T}}$ and $\mathbf{w}_2 = [w_2(1), \ldots, w_2(K)]^{\mathsf{T}}$, respectively, and the STFT coefficients of two channels of mixed signal during the silent interval of target signals are $\mathbf{d}_1 = [d_1(1), \ldots, d_1(L_D)]^{\mathsf{T}}$ and $\mathbf{d}_2 = [d_2(1), \ldots, d_2(L_D)]^{\mathsf{T}}$, respectively, where K is the length of the convolutive ISR and L_D is the length of the silent interval. Since we only consider a small K (in fact, a small K is enough as shown in our experiments), the pursuit of sparsity as in [3] is redundant. Here, we propose a simple least squares formulation to

$$\mathbf{A}_d^T := \begin{pmatrix} d_1(1) & \dots & \dots & d_1(K) & \dots & d_1(L_D) & 0 & \dots & 0 & \eta_1 \\ 0 & d_1(1) & \dots & d_1(K-1) & \dots & d_1(L_D-1) & d_1(L_D) & 0 & \dots & 0 \\ \vdots & & \ddots & & & & & \ddots & 0 & 0 \\ 0 & \dots & 0 & d_1(1) & \dots & d_1(L_D-K+1) & \dots & \dots & d_1(L_D) & 0 \\ d_2(1) & \dots & \dots & d_2(K) & \dots & d_2(L_D) & 0 & \dots & 0 & \eta_2 \\ 0 & d_2(1) & \dots & d_2(K-1) & \dots & d_2(L_D-1) & d_2(L_D) & 0 & \dots & 0 \\ \vdots & & \ddots & & & & & \ddots & 0 & 0 \\ 0 & \dots & 0 & d_2(1) & \dots & d_2(L_D-K+1) & \dots & \dots & d_2(L_D) & 0 \end{pmatrix}, \tag{3.2.9}$$

compute $\mathbf{w}_1$ and $\mathbf{w}_2$:

$$\begin{aligned} (\mathbf{w}_1, \mathbf{w}_2) = \arg\min_{\mathbf{w_1},\mathbf{w_2}} & \|\mathbf{w}_1 * \mathbf{d}_1 + \mathbf{w}_2 * \mathbf{d}_2\|_2^2 \\ & + \eta^2 (\frac{\eta_1}{\eta} w_1(1) + \frac{\eta_2}{\eta} w_2(1) - 1)^2, \end{aligned} \tag{3.2.8}$$

where $w_1(1)$ and $w_2(1)$ are the first entry of $\mathbf{w}_1$ and $\mathbf{w}_2$, respectively, and η, η_1, η_2 are parameters for generalization. The more elegant formulation can be given by defining a matrix $\mathbf{A}_d \in \mathbb{C}^{(L_D+K)\times 2K}$ (see (3.2.9)). Concatenate $\mathbf{w}_1$ and $\mathbf{w}_2$ into $\mathbf{w} := [\mathbf{w}_1, \mathbf{w}_2]^\mathsf{T} \in \mathbb{C}^{2K\times 1}$ and define $\mathbf{f} := [0, \dots, 0, \eta]^T \in \mathbb{C}^{(L_D+K)\times 1}$, then (3.2.8) will be formulated more compactly:

$$\mathbf{w} = \arg\min_{\mathbf{w}} \|\mathbf{A}_d\mathbf{w} - \mathbf{f}\|_2^2. \tag{3.2.10}$$

The solution of the least squares (3.2.10) is given below

$$\mathbf{w} = \lim_{\epsilon\to 0} (\mathbf{A}_d^\mathsf{H}\mathbf{A}_d + \epsilon\mathbf{I})^{-1}\mathbf{A}_d^\mathsf{H}\mathbf{f}, \tag{3.2.11}$$

To generalize the convolutive ISR to any number of microphones is trivial, just extend $\mathbf{w}$, $\mathbf{A}_d$, and $\mathbf{f}$. In fact, the convolutive ISR is a generalization version of the multiplicative ISR: if we set $\eta = \eta_1 \to \infty$, $\eta_2 = 0$ and $K = 1$, the least squares (3.2.8) converges to the quadratic optimization problem (3.2.5). Hence, the solution of (3.2.5) can also be approximated by (3.2.11).

3.3 Experiment

Simulation experiments were conducted in Matlab R2019a and the system was Ubuntu 18.04.4 LTS with Intel Core i9-9900K CPU @ 3.60GHz X 16 and 62.7GiB memory. All simulation experiments in this thesis use this setting. Ten groups of speech signals were

collected from the data packages dev1 and dev2 in SiSEC2011 [44]. These data packages were widely used in literature [22, 45] and have 16 male and 16 female vocals in different languages. Although the voice files sound less emotional (unlike natural speech), our results still apply to natural speech because the evaluation metrics we use do not measure emotion. Each speech signal was sampled at 16kHz and had 10 seconds duration, containing four simple sentences. These speech signals were convolved by room impulse responses (RIRs), and then summed to obtain 10 groups of mixed speech signals. The RIRs were generated by the software *pyroomaoustics* [2] using the image method. The software simulated a virtual room with the size of dimension 4.45 m $\times$ 3.55 m $\times$ 2.5 m. Two microphones with a spacing distance of 5 cm or three microphones with a spacing distance of 2.5 cm were set at the center of the room with a height of 1.4 m. There are also two or three loudspeakers were put at the height of 1.6 m and 1 m away from the microphones. For two loudspeakers, they were put at $-50°$ and $45°$ azimuth angle. For three loudspeakers, they were put at $-50°$, $-10°$, and $45°$ azimuth angle. The azimuth angle is defined as follows: The front of the mid-point of microphones was defined $0°$. Azimuth angle increased anticlockwise and decreased clockwise. For example, the right hand of the mid-point is $-90°$, and $90°$ represents the left hand of the mid-point. We do not specify the angle of the target signal and the interfering signal, the end result is the average of all possible cases. This avoids introducing bias due to the choice of the target signal angle. Another advantage of this average is to avoid introducing a bias due to the selection of gender voice as the target signal. Figure 3.4 shows a top view of the virtual room. There are three microphones and three loudspeakers inside the room. Note that, the figure is just a schematic without the true scale. A RIR generated by the software *pyroomacoustics* is shown in figure 3.5. Since we are only interested in evaluating the separation performance of the ISR, not including silent interval detection, we therefore skip the silent interval detection and tell the algorithm where the silent interval is.

The performance of speech separation are measured by source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifact ratio (SAR) [46]. They are widely accepted metrics in speech separation. It involves two successive steps to compute them. First of all, a speech estimate $\hat{s}$ can be decomposed into three part:

$$\hat{s} = s_{target} + e_{interf} + e_{artif}, \tag{3.3.1}$$

where s_{target} is the target signal, e_{interf} is the interference error term coming from the

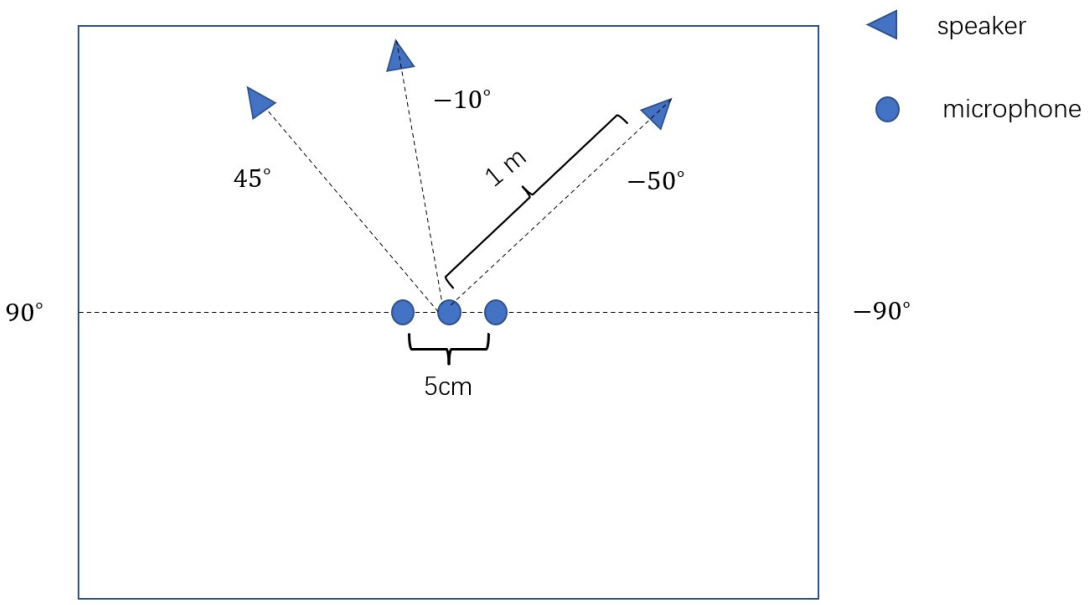

Figure 3.4: A schematic diagram (top view) of the virtual room used in experiments to evaluate speech separation algorithms.

interference signals, and e_{artif} is the artifact error term from other causes, like systematic distortion. Sometimes there is the fourth term called noise error term e_{noise} introduced by noise signal, like background noise, but we ignore it because the goal is to test the separation performance, not denoising, so there is no noise signal. The three components can be estimated by the projection of $\hat{s}$ on different subspaces expanded by original speech signals. Then, SDR, SIR, and SAR are computed by energy ratios in dB:

$$\begin{aligned} \text{SDR} &:= 10\log_{10}\frac{\|s_{target}\|_2^2}{\|e_{interf}+e_{artif}\|_2^2}, \\ \text{SIR} &:= 10\log_{10}\frac{\|s_{target}\|_2^2}{\|e_{interf}\|_2^2}, \\ \text{SAR} &:= 10\log_{10}\frac{\|s_{target}+e_{interf}\|_2^2}{\|e_{artif}\|_2^2}, \end{aligned} \tag{3.3.2}$$

where $\|\cdot\|_2^2$ denotes square of the l_2 norm. For all the three metrics, the larger, the better, and SDR is a more comprehensive measurement since it considers both interference error e_{interf} and artifact error e_{artif}. A larger SDR means the estimated signal is closer to the true signal, a larger SIR means the interference signals are removed cleanly, and a larger SAR means fewer artifacts distortion.

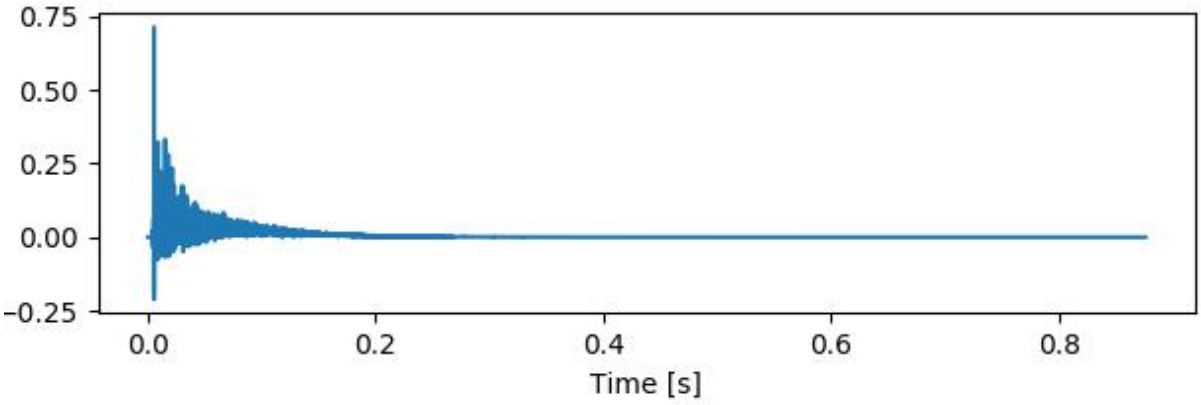

Figure 3.5: Depiction of a room impulse response (RIR) generated by the software *pyroomacoustics* [2].

3.4 Experiment Result

3.4.1 Effect of the STFT window length

We first examine the effect of the STFT window length (Fig. 3.6). The length of the STFT window varies from 128 to 1024. A 0.5-second silence interval is used to compute the ISR with lengths 1 (K=1) and 14 (K=14), respectively. Roughly speaking, as the window length increases, SDR decreases, SIR increases, and SAR decreases. When window length is longer, the cancellation model is more accurate, so SIR is larger. But a longer window length will result in more distortion on the target signal, so SAR decreases, and SDR, as a comprehensive measurement, also decreases along with SAR. Actually, this distortion sounds like the signal is being filtered by some kind of high pass filter, but the distortion is slight and doesn't affect the understanding of the sentence. For K=14, SIR decreases at window length 1024. We think this is because the data appears inadequate to compute ISR. Another observation is that an ISR of length 1 has a larger SAR than an ISR of length 14, so a shorter ISR will bring less distortion.

3.4.2 Effect of the length of silent interval

The second experiment tests the effect of the silent interval length (Fig. 3.7). The length of the STFT window is set to 128 and will keep the same in the rest of the experiments. The length of the silent interval varies from 0.1 s to 1 s. When the silent interval is short (0.1 s), a shorter ISR has a better performance. This is because there is not enough data to compute a longer ISR. As the silent interval length increases, long ISR has a better and

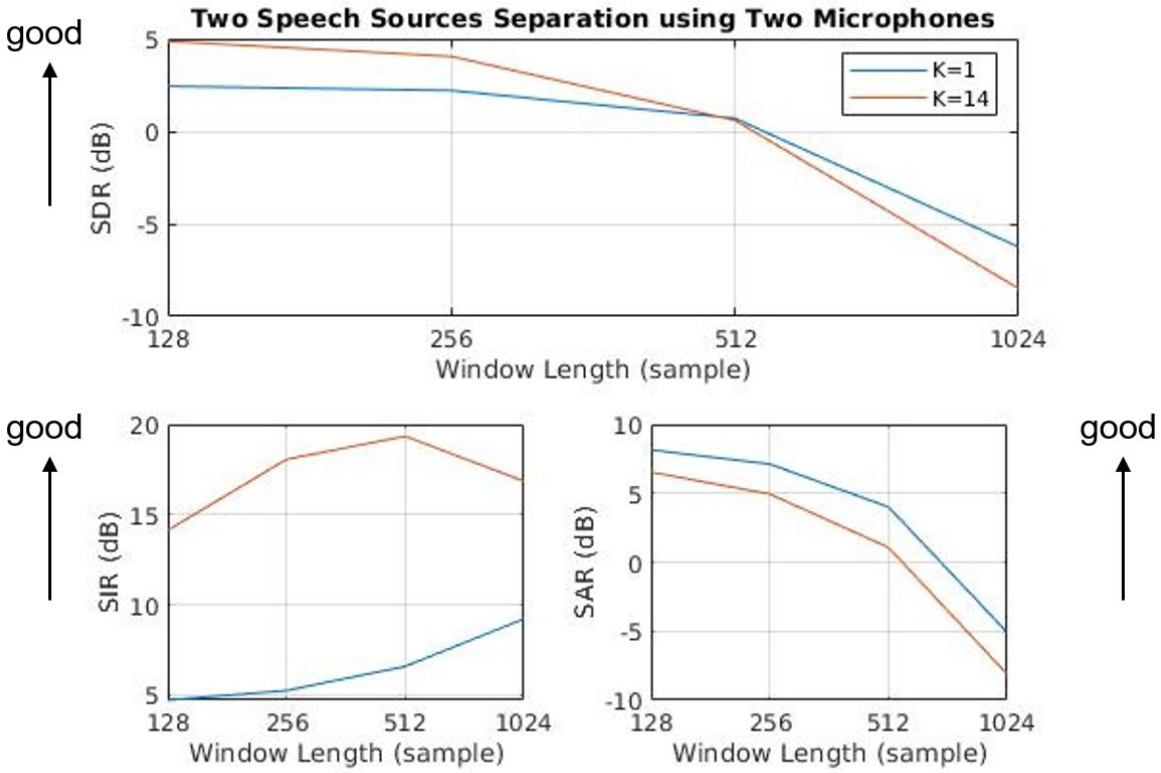

Figure 3.6: Average separation performance (output SDRs, SIRs, and SARs) as a function of the STFT window length in the presence of $RT_{60} = 300$ ms reverberation. There are two microphones and two speech sources. The length of ISR is 1 (K=1) for the blue and 14 (K=14) for the red.

better performance. It seems that the optimal silent length is 0.5 s, and a longer silence interval does not bring significant benefits. The optimal length of ISR is 14. Although a longer ISR has a larger SIR, it also has a smaller SAR, i,e, introduces more distortion. This experiment double confirms that a longer ISR will bring more distortion.

3.4.3 Weak target extraction

This experiment evaluates the weak target signal extraction ability of ISR (Fig. 3.8). We put the target signal at the azimuth angle of $-50°$ and the interference signal at the azimuth angle of 45°. The silent interval is 0.5 s and will keep the same in the rest of the experiments. The length of ISR is 14 (K=14). The input SIR is the SIR before applying the extraction algorithm. A small input SIR means the target signal is weak. When the input SIR is -10 dB, which means the target signal is very weak, and barely perceived, the SIR improvement is about 20 dB in the presence of 200 ms reverberation ($RT_{60} = 200$ ms),

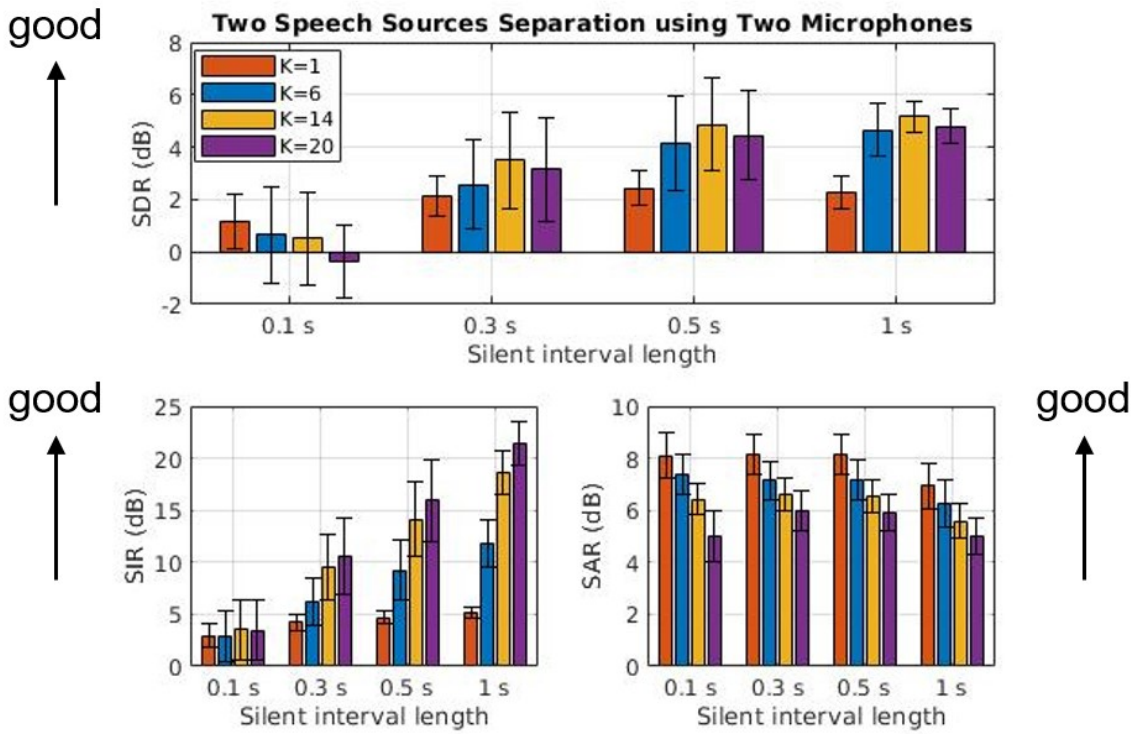

Figure 3.7: Average separation performance (output SDRs, SIRs, and SARs) as a function of the length of silent interval in the presence of RT_{60} = 300 ms reverberation. There are two microphones and two speech sources. The length of ISR is 1 (K=1), 6 (K=6), 14 (K=14), and 20 (K=20) for the red, blue, yellow, and purple, respectively.

which means now the situation reverses, and the interference signal is quite suppressed. This proves that our algorithm successfully extracts weak target signal. It should be note that reverberation will affect the extraction performance of the algorithm.

3.4.4 Performance comparison

In this experiment, we compare the speech separation performance between we proposed ISR and the l_1 regularized cancellation kernel [3]. The length of ISR is 14 (K=14) and the length of l_1 regularized cancellation kernel is 512 (L=512). We have tried other cancellation kernels with different length, e.g., 128, 256, 1024, and 2048. It seems that 512 is the optimal length and a longer kernel did not have a better performance but significantly increased the computational time. As shown in Fig. 3.9, in the case of two microphones and two speech sources, ISR and l_1 regularized cancellation kernel have a comparable separation performance, while in the case of three microphones and three speech sources, ISR has a better SDR and SIR than l_1 regularized cancellation kernel.

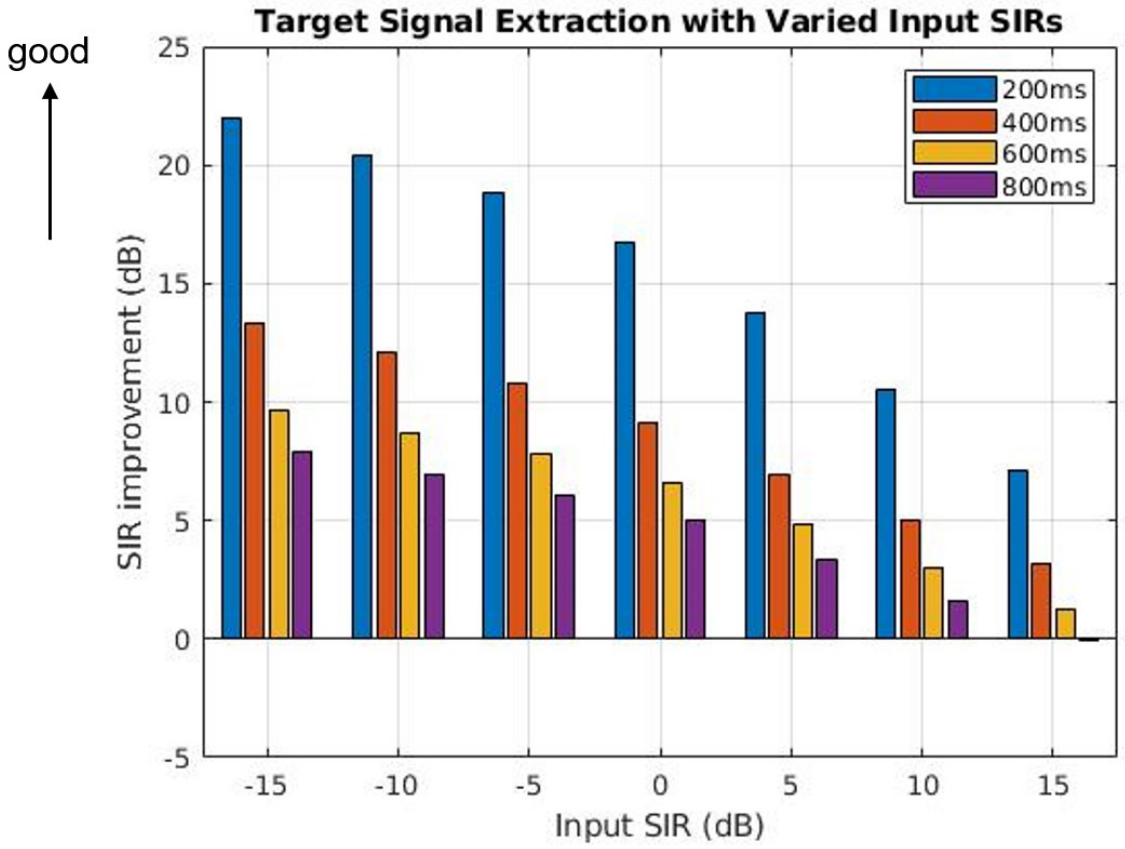

Figure 3.8: Average SIR improvement with varied input SIRs in the presence of different reverberations. There are two microphones, one target signal, and one interference signal. The length of ISR is 14 (K=14). The reverberation time RT_{60} is 200 ms, 400 ms, 600 ms, and 800 ms for the blue, red, yellow, and purple, respectively.

3.5 Summary

In the first work, we propose interference suppression response (ISR) - a time-frequency version of the cancellation kernel. ISR can extract very weak target signals just like traditional time-domain version of the cancellation kernel. Compared to the time domain cancellation kernel, the biggest advantage of ISR is that it has analytical expressions. For speech separation performance, ISR has a comparable or even better separation performance. About 2 dB SDR improvement in the case of three microphones and three sources. Moreover, ISR can be computed without silent interval, which we will show it in our second work (chapter 4). The limitation of this work is that ISR still can not avoid distortion on the target signal. Hence, eliminating distortion is worth future research.

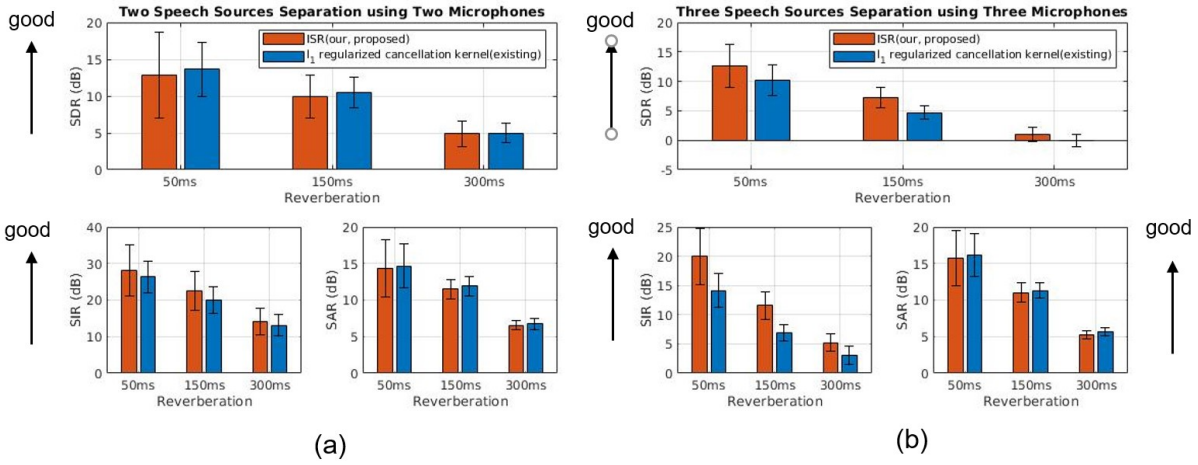

Figure 3.9: Average separation performance comparison between the proposed ISR (red) and the l_1 regularized cancellation kernel [3] in the presence of different reverberations. The length of ISR is 14 (K=14) and the length of the l_1 regularized cancellation kernel is 512 (L=512). (a) Two speech sources separation using two microphones. (b) Three speech sources separation using three microphones.

Chapter 4

Improving DUET

Degenerate Unmixing Estimation Technique (DUET) [4] is one of the most well-known blind source separation techniques based on binary masking. The algorithm is fast, robust to noise, and quite general, i.e. can separate more than two sources. But the separation performance of DUET is not satisfactory, which is the downside of binary masking. Binary masking separates signals by multiplying a set of binary matrices with a mixed signal's spectrogram. These binary matrices are also called binary masks. Fig. 4.1 shows an exaggerated illustration of binary masking. However, mixed speech signals can not be

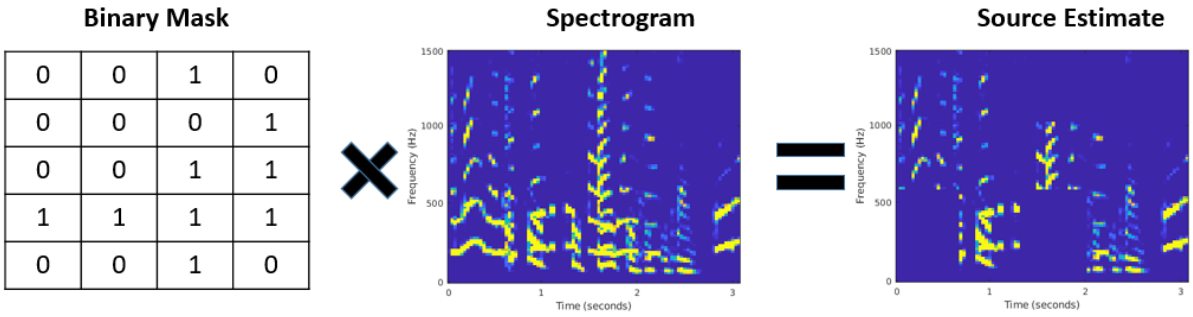

Figure 4.1: An exaggerated illustration of binary masking.

completely separated by binary masking. I will illustrate this problem using Fig. 4.2. Fig. 4.2 shows the spectrogram of two speech signals. There are three color regions shown in the figure, which are red, green, and blue regions. The time-frequency (TF) points in the red region are mainly contributed by source 1, the TF points in the green region are mainly contributed by source 2, and the TF points in the blue region are contributed by both sources. Obviously, the blue region in the spectrogram can not be separated by binary masking. Unfortunately, the blue region is quite common for speech signals. Hence, DUET fails to separate speech signals if there are too many sources.

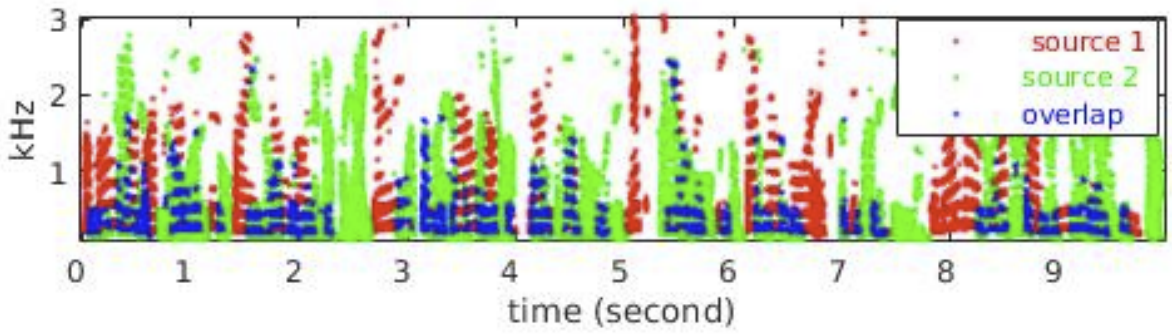

Figure 4.2: Spectrogram of two speech signals

In this work, to improve the separation performance of DUET, we replace binary masking with soft filtering, that is, a weighted summation of mixed signals. The weights are calculated using some linear spatial filtering (also called beamforming) criteria, e.g., the minimum variance distortionless response (MVDR) [9, 16] and our first work, interference suppression response (ISR). Unlike previous works computing parameters of beamformer using all the TF points, the novelty of this work is that we estimate parameters of beamformer by only utilizing information embedded in the TF points only contributed by a single source (single-source-points, SSPs), i.e., TF points in the red and green region in Fig. 4.2. Compared with conventional DUET, our method improves source-to-interference ratio (SIR) and source-distortion ratio (SDR) by 2 to 5 dB, respectively. This improvement level is comparable with past literature [23]. Weights computed by ISR criteria has obtained a much better performance than by MVDR. These results come from our speech separation experiments using real recordings.

In the rest of the chapter, we will introduce DUET and our improved method in section 4.1 and 4.2, respectively. Section 4.3 is the experiment protocol, and section 4.4 shows the results. This chapter is summarised in section 4.5.

4.1 Degenerate Unmixing Estimation Technique

In this section, we will introduce the degenerate unmixing estimation technique (DUET), a fast speech separation algorithm for separating multiple sources by only two microphones.

4.1.1 Local mixing parameters

DUET assumes that for each TF point, there is only one source active. Under this so-called the W-Disjoint Orthogonal assumption [47, 48], mixed speech signals can be separated by a set of binary masks that are constructed by clustering TF points. Since speech signal from different directions has a different time of arrival and attenuation between two microphones, TF points can be clustered based on the two features. Denote the arrival time, attenuation, and distance from the j-th source to the i-th microphone as t_{ij}, q_{ij}, and d_{ij}, respectively. They satisfy

$$\begin{aligned} t_{ij} &= d_{ij}/c, \\ q_{ij} &= \frac{1}{\sqrt{4\pi} d_{ij}}, \end{aligned} \tag{4.1.1}$$

where c is the speed of sound, and the latter equation is obtained since the sound is radiated out in the form of a sphere. Assume there are two microphones, the time difference (δ_j) and the attenuation ratio (r_j) between the j-th signal arriving at the two microphones are

$$\begin{aligned} \delta_j &:= t_{2j} - t_{1j} = d_{2j}/c - d_{1j}/c, \\ r_j &:= \frac{q_{2j}}{q_{1j}} = \frac{d_{1j}}{d_2 j}. \end{aligned} \tag{4.1.2}$$

Using (4.1.2) to calculate the arrival time difference δ_j and the attenuation ratio r_j requires knowing the distance from the j-th source to the two microphones. However, this knowledge is often not available in practice, so computations must rely on ways other than the definition. In fact, δ_j and r_j are encoded in the acoustic transfer functions (ATFs). When there is no reverberation, they can be computed from ATFs (2.3.2) by

$$\begin{aligned} \delta_j &= -1/(2\pi f) \angle (a_{2j}(f)/a_{1j}(f)), \\ r_j &= |a_{2j}(f)/a_{1j}(f)|, \end{aligned} \tag{4.1.3}$$

where $\angle(\cdot)$ is the angle of a complex number and $|\cdot|$ is the module of a complex number. Although ATF is also not available in practice, things will be different under the W-Disjoint Orthogonal assumption. Denote the STFT coefficients of two mixed signals observed by two microphones as $x_1(t, f)$ and $x_2(t, f)$, respectively, STFT coefficients of the j-th source signal as $s_j(t, f)$, and the acoustic transfer function (ATF) from j-th source to i-th microphones as $a_{ij}(f)$. Recap the time-frequency domain multiplicative

model (2.3.1), under the W-Disjoint Orthogonal assumption, there exists a $j \in \{1, \ldots, J\}$ where J is the total number of sources such that

$$\begin{aligned} x_1(t,f) &= a_{1j}(f)s_j(t,f) + n_1(t,f), \\ x_2(t,f) &= a_{2j}(f)s_j(t,f) + n_2(t,f), \end{aligned} \tag{4.1.4}$$

where $n_i(t,f)$ for $i = 1, 2$ is the noise signal on the i-th microphone. Combining (4.1.3) and (4.1.4), assuming there is no noise on the two microphones, i.e., $n_i(t,f) = 0$ for $i = 1, 2$, we can compute the arrival time difference and the attenuation ratio from samples of STFT coefficients of mixed signals by

$$\begin{aligned} \delta_j &= -1/(2\pi f)\angle(x_2(t,f)/x_1(t,f)), \\ r_j &= |x_2(t,f)/x_1(t,f)|. \end{aligned} \tag{4.1.5}$$

Note that, first, j here is a latent variable, i.e., we don't know what the j is (we only know that $j \in \{1, \ldots, J\}$), so classification is needed. Second, in practical, δ_j and r_j of the same j computed from different time-frequency points are slightly different. This is because of reverberation, noise, etc that breaks our assumptions. Hence, what (4.1.5) computes are time-frequency dependent, called local mixing parameters, and denoted as $\delta(t,f)$ and $r(t,f)$. Rather than calculating $r_j(t,f)$, a better alternative is to calculate

$$\begin{aligned} \alpha(t,f) &:= r(t,f) - \frac{1}{r(t,f)} \\ &= \left|\frac{x_2(t,f)}{x_1(t,f)}\right| - \left|\frac{x_1(t,f)}{x_2(t,f)}\right|. \end{aligned} \tag{4.1.6}$$

The reason is that $\alpha(t,f)$ is more symmetric than $r(t,f)$: if you swap the two microphones, $r(t,f)$ will becomes $1/r(t,f)$, but $\alpha(t,f)$ becomes $-\alpha(t,f)$. $\alpha(t,f)$ is called the symmetric attenuation ratio.

4.1.2 Classification by histogram

In the last section, we introduced two local mixing parameters (the arrival time difference $\delta(t,f)$ and the symmetric attenuation ratio $\alpha(t,f)$) computed from each TF point. If our assumptions are not seriously broken, then $\delta(t,f)$ and $\alpha(t,f)$ of the same source will only fluctuate within a certain range, significantly smaller than the difference between different sources. Hence, a histogram of $\delta(t,f)$ and $\alpha(t,f)$ can be built up to make clusters for TF points of different sources. We can imagine that if there are J sources from different

directions, then there are supposed to be J peaks in the histogram, and the peak locations reveal more accurate mixing parameters in a global view. A TF point belongs to the j-th cluster if its position of the local mixing parameters is closer to the j-th peak location.

First, we define a set containing all the TF points whose local mixing parameters are around given values:

$$I(\delta, \alpha) := \{(t, f) : |\delta(t, f) - \delta| < \triangle_\delta, |\alpha(t, f) - \alpha| < \triangle_\alpha\}, \tag{4.1.7}$$

where $\triangle_\delta$ and $\triangle_\alpha$ are relatively small values defining the resolution. Treating all TF points equally is not fair since for some TF points with small values are possibly noise. Hence, rather than directly calculate a normal histogram, a weighted histogram is constructed

$$H(\delta, \alpha) := \int\int_{(t,f)\in I(\delta,\alpha)} |x_1(t, f)x_2(t, f)|^p (2\pi f)^q dtdf, \tag{4.1.8}$$

where p and q are parameters motivated by the maximum likelihood (ML) estimators [49]. If $q = p = 0$, then the weighted histogram becomes normal counting histogram. Fig. 4.3 gives an example of a weighted histogram (p=1, q=0) computed from mixtures of four speech signals. It is recommended in [4] that by default, p=1, and q=0. If the source signals are not equal power, then a better choice is p=0.5, q=0 to prevent the small peaks covered by the dominant peaks.

4.1.3 Source separation

If a TF point is only or mainly contributed by a single source, the local mixing parameters, i.e., the arrival time difference $\delta(t, f)$ and the symmetric attenuation ratio $\alpha(t, f)$, are supposed to be around one of the peaks in the histogram. Then source separation is simple by just assigning this TF point to the corresponding source. Denote the j-th peak in the histogram as (δ_j, α_j) (we call them global mixing parameters). We convert the symmetric attenuation ratio back to the attenuation ratio by

$$r_j = \frac{\alpha_j + \sqrt{\alpha_j^2 + 4}}{2} \tag{4.1.9}$$

We measure the distance between a local mixing parameters and global mixing parameters via

$$\rho_j(t, f) := \frac{1}{1 + |\Lambda_j|^2} |\Lambda_j x_1(t, f) - x_2(t, f)|^2, \tag{4.1.10}$$

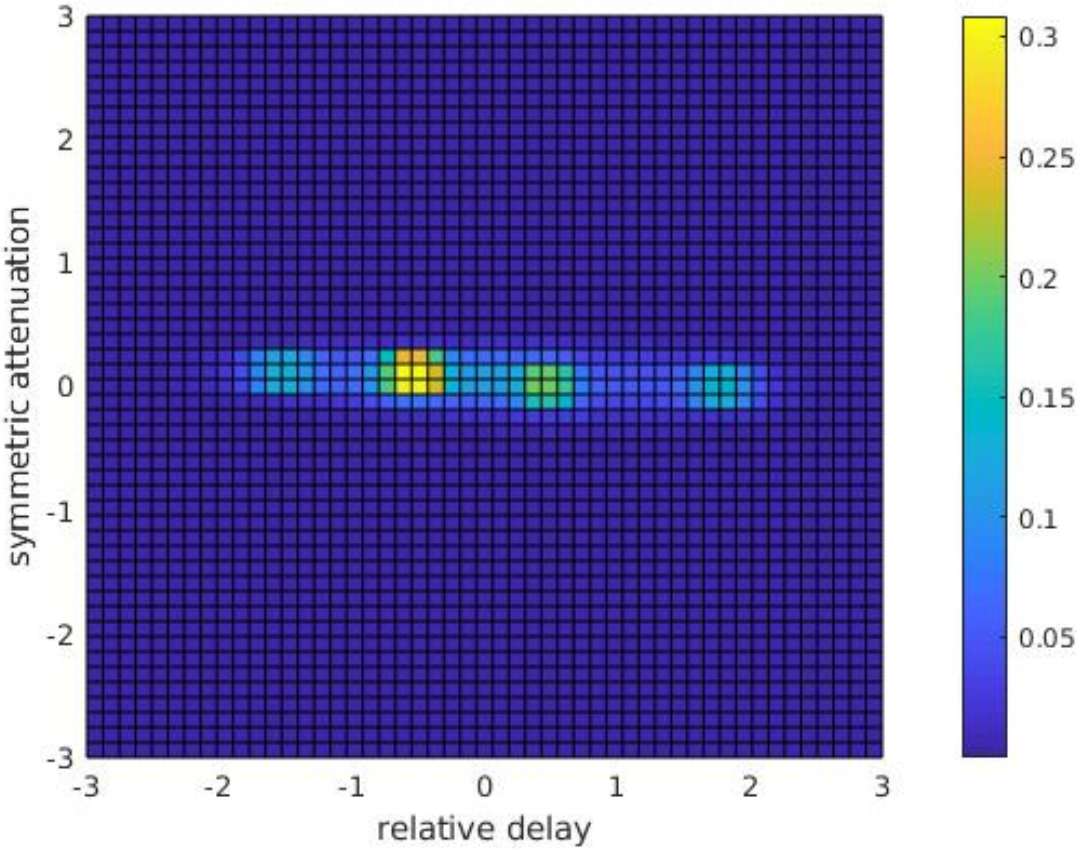

Figure 4.3: Top view of a weighted histogram (p=1, and q=0) of local mixing parameters computed from mixtures of four sources.

where $\Lambda_i = |r_i| \exp(-2\pi f \delta_i)$, and $j = 1, \ldots, J$. The smaller $\rho_j(t, f)$ is, the closer the local mixing parameters are to the global mixing parameters. Then the binary mask for separating j-th source signal can be constructed by

$$M_j(t, f) = \begin{cases} 1, & \text{if } \rho_j(t, f) < \rho_k(t, f), \forall k \neq j, \\ 0, & \text{otherwise.} \end{cases} \tag{4.1.11}$$

The j-th source signal can be estimated via masking

$$\hat{s}_j(t, f) = M_j(t, f) x_1(t, f). \tag{4.1.12}$$

4.2 Improved Method

The TF masking (4.1.12) can not separate mixed signal when the TF point is contributed by more than one source. In this case, we propose to do a weighted summation to remove

the interference contribution. The weights are calculated following some linear spatial filtering (or beamforming) criteria, for example, the minimum distortionless response (MVDR) and interference suppression response (ISR, our first work).

4.2.1 Spatial filtering

A linear spatial filter, also called beamformer, is a set of frequency-dependent weights $\mathbf{w}(f) = [w_1(f), \ldots, w_I(f)]^{\mathsf{T}}$ whose the number of coefficients is equal to the number of microphones. The signal from the desired direction can be estimated by applying the weights to the STFT of mixed signals $\mathbf{x}(t, f)$ observed by the microphone array:

$$\hat{s}_t(t, f) = \mathbf{w}^H(f)\mathbf{x}(t, f), \tag{4.2.1}$$

where $\hat{s}_t(t, f)$ is the estimation of STFT of the signal from desired direction. Different beamformers have been proposed according to different objectives. There are so-called fixed beamformers whose weights are pre-fixed and only can be used to extract signals from directions set in advance. The advantage of fixed beamformers is they only rely on the direction of arrival (DOA) and the geometry of the microphone arrays while the disadvantage is lacking of flexibility in choosing a direction and the microphone arrays. There is another class of beamformers which are data-dependent and more flexible in practice, called adaptive beamformers. Since these beamformers are adaptive to the statistical characteristics of signals, they usually perform better than the fixed beamformers. In this thesis, we introduce two types of adaptive beamformers assuming the time-frequency multiplicative model (see section 2.3.1). Following the equation (2.3.1), denote $\mathbf{a}_j f = [a_{1j}(f), \cdots, a_{Ij}(f)]^{\mathsf{T}}$, STFT of signals satisfy

$$\mathbf{x}(t, f) = \mathbf{A}(f)\mathbf{s}(t, f) + \mathbf{u}(t, f), \tag{4.2.2}$$

where $\mathbf{x}(t, f) = [x_1(t, f), \cdots, x_I(t, f)]^{\mathsf{T}}$ are STFT coefficients of mixed signals, $\mathbf{s}(t, f) = [s_{P+1}(t, f), \cdots, s_J(t, f)]^{\mathsf{T}}$ are STFT coefficients of the target signals, $\mathbf{u}(t, f) = \sum_{j=1}^{P} \mathbf{a}_j s_j(t, f)$ are contributions from the interference signals, and $\mathbf{A} = [a_{P+1}(f), \cdots, a_J(f)]$ are mixing matrix containing ATFs of the target signals. We omit frame index t and frequency index f for brevity as long as there is no ambiguity. Assuming $\mathbf{s}$ and $\mathbf{u}$ are independent random signals with zero mean, denote the covariance matrix of signals as $\mathbf{\Sigma}_{\mathbf{x}} = \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathsf{H}}\right]$, $\mathbf{\Sigma}_{\mathbf{s}} = \mathbf{A}\mathbb{E}\left[\mathbf{s}\mathbf{s}^{\mathsf{H}}\right]\mathbf{A}^{\mathsf{H}}$, and $\mathbf{\Sigma}_{\mathbf{u}} = \mathbb{E}\left[\mathbf{u}\mathbf{u}^{\mathsf{H}}\right]$, respectively, they have the below relationship

$$\mathbf{\Sigma}_{\mathbf{x}} = \mathbf{\Sigma}_{\mathbf{s}} + \mathbf{\Sigma}_{\mathbf{u}}. \tag{4.2.3}$$

The covariance equation (4.2.3) is often used in spatial signal processing by the microphone array. The energy of the filtered signal can be expressed as

$$\begin{aligned} \|\mathbf{w}^{\mathsf{H}}\mathbf{x}\|_2^2 &= \mathbf{w}^{\mathsf{H}}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{w} \\ &= \mathbf{w}^{\mathsf{H}}\boldsymbol{\Sigma}_{\mathbf{s}}\mathbf{w} + \mathbf{w}^{\mathsf{H}}\boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{w}, \end{aligned} \tag{4.2.4}$$

where the first term $\mathbf{w}^{\mathsf{H}}\boldsymbol{\Sigma}_{\mathbf{s}}\mathbf{w}$ is the energy of the filtered target signals and the second term $\mathbf{w}^{\mathsf{H}}\boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{w}$ is the energy of the filtered interference signals. A common metric called source-to-interference ratio (SIR) is defined to measure the performance of a linearly spatial filter:

Definition 4.1 (SIR of a Linear Spatial Filter)**.** *The source-to-interference ratio (SIR) of a linear spatial filter is the ratio of the energy of the filtered target signals and the interference signals, which is*

$$SIR = \frac{\mathbf{w}^{\mathsf{H}}\boldsymbol{\Sigma}_{\mathbf{s}}\mathbf{w}}{\mathbf{w}^{\mathsf{H}}\boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{w}}. \tag{4.2.5}$$

Minimum variance distortionless response

Assume there is only one target direction, then $\mathbf{w}^{\mathsf{H}}\mathbf{x} = \mathbf{w}^{\mathsf{H}}\mathbf{a}s + \mathbf{w}^{\mathsf{H}}\mathbf{u}$ where s is the target signal and $\mathbf{a}$ is the corresponding ATF. The criterion of the MVDR is to minimize the energy of the filtered interference signal, i.e., $\min_{\mathbf{w}} \mathbf{w}^{\mathsf{H}}\boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{w}$ meanwhile introduce no distortion on the target signal, i.e, $\mathbf{w}^{\mathsf{H}}\mathbf{a}$=1. Hence, MVDR filter can be obtained by solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^{\mathsf{H}}\boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^{\mathsf{H}}\mathbf{a} = 1. \end{aligned} \tag{4.2.6}$$

The solution of the above problem is given by

$$\mathbf{w}_{\text{MVDR}} = \frac{\boldsymbol{\Sigma}_{\mathbf{u}}^{-\mathbf{1}}\mathbf{a}}{\mathbf{a}^{\mathsf{H}}\boldsymbol{\Sigma}_{\mathbf{u}}^{-\mathbf{1}}\mathbf{a}}. \tag{4.2.7}$$

We substitute (4.2.7) into (4.2.4), the filtered signal energy by MVDR is

$$\|\mathbf{w}_{\text{MVDR}}^{\mathsf{H}}\mathbf{x}\|_2^2 = \sigma_s^2 + \frac{1}{\mathbf{a}^{\mathsf{H}}\boldsymbol{\Sigma}_{\mathbf{u}}{}^{-1}\mathbf{a}}, \tag{4.2.8}$$

where σ_s^2 is the variance of the target signal $\sigma_s^2 = \mathbb{E}[s^2]$. The SIR of the MVDR is

$$\text{SIR}_{\text{MVDR}} = \sigma_s^2\mathbf{a}^{\mathsf{H}}\boldsymbol{\Sigma}_{\mathbf{u}}^{-\mathbf{1}}\mathbf{a}. \tag{4.2.9}$$

Under the distortionless constraint $\mathbf{w}^{\mathsf{H}}\mathbf{a} = 1$, minimizing $\mathbf{w}^{\mathsf{H}}\mathbf{\Sigma_u}\mathbf{w}$ is equivalent to minimize $\mathbf{w}^{\mathsf{H}}\mathbf{\Sigma_x}\mathbf{w} = \sigma_s^2 + \mathbf{w}^{\mathsf{H}}\mathbf{\Sigma_u}\mathbf{w}$ since σ_s^2 is a constant. Hence, the MVDR is equivalent to the so-called minimum power distortionless response (MPDR) criterion

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^{\mathsf{H}}\mathbf{\Sigma_x}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^{\mathsf{H}}\mathbf{a} = 1. \end{aligned} \tag{4.2.10}$$

The MPDR beamformer is

$$\mathbf{w}_{\text{MPDR}} = \frac{\mathbf{\Sigma_x^{-1}}\mathbf{a}}{\mathbf{a}^{\mathsf{H}}\mathbf{\Sigma_x^{-1}}\mathbf{a}}. \tag{4.2.11}$$

The advantage of MPDR is that $\mathbf{\Sigma_x}$ is always available but the disadvantage is that MPDR is more sensitive to the misalignment error of the ATF of the target signal, i.e., $\mathbf{a}$.

Interference suppression response

Recap our first work, the multiplicative interference suppression response (ISR) (see Sec. 3.2.1) is also a type of adaptive beamformer and can be derived from a similar criterion of MVDR

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^{\mathsf{H}}\mathbf{\Sigma_u}\mathbf{w} \\ \text{s.t.} \quad & w(1) = 1, \end{aligned} \tag{4.2.12}$$

where $w(1)$ is the first coefficient of $\mathbf{w}$. The resulting ISR beamformer

$$\mathbf{w}_{\text{ISR}} = \frac{\mathbf{\Sigma_u^{-1}}\mathbf{a_0}}{\mathbf{a_0}^{\mathsf{H}}\mathbf{\Sigma_u^{-1}}\mathbf{a_0}}, \tag{4.2.13}$$

where $\mathbf{a_0} = [1, 0, \cdots, 0]^{\mathsf{T}} \in \mathbb{R}^{I\times 1}$. The output power of ISR is

$$\|\mathbf{w}_{\text{ISR}}^{\mathsf{H}}\mathbf{x}\|_2^2 = \frac{\mathbf{a_0}^{\mathsf{H}}\mathbf{\Sigma_u^{-1}}\mathbf{\Sigma_s}\mathbf{\Sigma_u^{-1}}\mathbf{a_0}}{(\mathbf{a_0}^{\mathsf{H}}\mathbf{\Sigma_u^{-1}}\mathbf{a_0})^2} + \frac{1}{\mathbf{a_0}^{\mathsf{H}}\mathbf{\Sigma_u^{-1}}\mathbf{a_0}}. \tag{4.2.14}$$

The SIR of the ISR is

$$\text{SIR}_{\text{ISR}} = \frac{\mathbf{a_0}^{\mathsf{H}}\mathbf{\Sigma_u^{-1}}\mathbf{\Sigma_s}\mathbf{\Sigma_u^{-1}}\mathbf{a_0}}{\mathbf{a_0}^{\mathsf{H}}\mathbf{\Sigma_u^{-1}}\mathbf{a_0}}. \tag{4.2.15}$$

Compared to MVDR, ISR extract the target signal only assume the covariance matrix of the interference signals. So, ISR is completely not sensitive to the misalignment error of the ATF of the target signal. Although ISR introduce distortion on the target signal, the effect sounds like a mild high-pass filter that the human auditory system doesn't seem to be sensitive to.

Generalized ISR

ISR can be generalized by modifying the constraint in (4.2.12) to $\mathbf{w}^{\mathsf{H}}\mathbf{a_0} = 1$ where $\mathbf{a_0}$ is an arbitrary vector with the same dimension of $\mathbf{w}$. The generalized ISR (GISR) criterion is given by

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^{\mathsf{H}}\mathbf{\Sigma_u}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^{\mathsf{H}}\mathbf{a}_0 = 1. \end{aligned} \tag{4.2.16}$$

If $\mathbf{a_0}$ is the ATF of the target signal, then (4.2.16) is equivalent to the MVDR criterion (4.2.6), and if $\mathbf{a_0} = [1, 0, \cdots, 0]^{\mathsf{T}}$, (4.2.16) is equivalent to the ISR criterion (4.2.12). Hence, both MVDR and ISR are special cases of GISR. The solution of (4.2.16) has the same expression as (4.2.13). Explicit criteria should be given to determine the explicit value of the vector $\mathbf{a_0}$. For example, let us consider what $\mathbf{a_0}$ maximizes the SIR. The SIR of the GISR is

$$\mathrm{SIR}_{\mathrm{GISR}} = \frac{\mathbf{a_0}^{\mathsf{H}}\mathbf{\Sigma_u^{-1}}\mathbf{\Sigma_s}\mathbf{\Sigma_u^{-1}}\mathbf{a_0}}{\mathbf{a_0}^{\mathsf{H}}\mathbf{\Sigma_u^{-1}}\mathbf{a_0}}. \tag{4.2.17}$$

Interestingly, such $\mathbf{a}_0$ which maximizes the SIR (4.2.17) is exactly the ATF of the target signal. Hence, the GISR which maximizes the SIR is exactly the MVDR. We will discuss it below.

Definition 4.2 (Maximum Eigenvector)**.** *The maximum eigenvectors of a matrix is the eigenvectors corresponding to the maximum eigenvalues. There may be multiple independent maximum eigenvectors as there are multiple maximum eigenvalues.*

Proposition 4.3. *If* $\mathbf{a_0}$ *maximizes the SIR* (4.2.17)*, then* $\mathbf{a_0}$ *is the maximum eigenvector of* $\mathbf{\Sigma_s}\mathbf{\Sigma_u^{-1}}$*.*

Proof. According to the relevant knowledge of complex analysis, $\mathbf{a_0}$ maximizing the SIR (4.2.17) has to satisfy the below stationary equation:

$$\frac{\partial}{\partial \mathbf{a_0}} \frac{\mathbf{a_0}^{\mathsf{H}}\mathbf{\Sigma_u^{-1}}\mathbf{\Sigma_s}\mathbf{\Sigma_u^{-1}}\mathbf{a_0}}{\mathbf{a_0}^{\mathsf{H}}\mathbf{\Sigma_u^{-1}}\mathbf{a_0}} = 0. \tag{4.2.18}$$

Denote the maximum point by $\mathbf{a_0^*}$, then $\mathbf{a_0^*}$ should satisfy

$$(\mathbf{a_0^*}\mathbf{\Sigma_u^{-1}}\mathbf{\Sigma_s}\mathbf{\Sigma_u^{-1}}\mathbf{a_0^*})\mathbf{a_0^*} = (\mathbf{a_0^*}^{\mathsf{H}}\mathbf{\Sigma_u^{-1}}\mathbf{a_0^*})\mathbf{\Sigma_s}\mathbf{\Sigma_u^{-1}}\mathbf{a_0^*}. \tag{4.2.19}$$

Denote $\mathbf{a_0^*}\mathbf{\Sigma_u^{-1}}\mathbf{\Sigma_s}\mathbf{\Sigma_u^{-1}}\mathbf{a_0^*}$ by β and $\mathbf{a_0^*}^{\mathsf{H}}\mathbf{\Sigma_u^{-1}}\mathbf{a_0^*}$ by α, the stationary equation (4.2.19) can be simply expressed as

$$\mathbf{\Sigma_s}\mathbf{\Sigma_u^{-1}}\mathbf{a_0^*} = \frac{\beta}{\alpha}\mathbf{a_0^*}. \tag{4.2.20}$$

It is not difficult to see that $\mathbf{a_0^*}$ is actually a eigenvector of $\mathbf{\Sigma_s \Sigma_u^{-1}}$, and $\frac{\beta}{\alpha}$ is the corresponding eigenvalue. $\frac{\beta}{\alpha}$ is exactly the SIR that we want to maximize. Therefore, $\mathbf{a_0^*}$ is the maximum eigenvector of $\mathbf{\Sigma_s \Sigma_u^{-1}}$. Proof is complete. □

When there is only one target signal, $\mathbf{\Sigma_s} = \sigma^2 \mathbf{a}\mathbf{a}^{\mathsf{H}}$ where σ^2 is the energy (variance) of the target signal and $\mathbf{a}$ is the ATF. Since $\mathbf{\Sigma_s}$ is a rank-1 matrix, $\mathbf{\Sigma_s \Sigma_u^{-1}}$ is also rank-1 matrix and $\mathbf{a}$ is the only one eigenvector corresponding to the positive eigenvalue ($\mathbf{\Sigma_u}$ is a positive definite matrix). Therefore, the maximum eigenvector of $\mathbf{\Sigma_s \Sigma_u^{-1}}$ is $\mathbf{a}$, the ATF of the target signal, and MVDR is a special case of GISR which maximizes the SIR (4.2.17). The above discussion can be concluded in the following theorem:

Theorem 4.4 (Optimal Linearly Constrained Filter)**.** *Considering one target direction, MVDR* (4.2.7) *is the filter that maximizes SIR in any linearly constrained linear spatial filter trying to minimize the energy of interference signals.*

The above discussion gives us an example to design specific filter according to a specific criteria. Although MVDR seems to be the optimal filter, in fact, it is not robust to the error of the ATF or covariance matrix. Perhaps a robust MVDR can be driven by the objective of maximizing SIR while being robust. This is an interesting future work.

4.2.2 Parameters estimation

In the last section, we introduced some spatial filtering criteria. We will use these criteria to compute weights for source separation. However, these criteria assume some parameters, e.g., the ATF of the target signal and the covariance matrix of the interference signals, which are not available in the blind source separation. In this section, we will introduce a method to estimate these parameters given mixed signals.

Recap the local mixing parameters and global mixing parameters we introduced in Sec. 4.1.1 and Sec. 4.1.2. If our assumptions strictly hold, then the local mixing parameters are equivalent to the corresponding global mixing parameters. When a TF point is contributed by two sources, the local mixing parameters will deviate from the global mixing parameters. If the two contributions are equal, maybe the local mixing parameters are in the middle of the global mixing parameters of the two sources. When one of the sources is dominant, it is possible that the local mixing parameters will be closer to that

source's global mixing parameters. Based on this hypothesis, we identify single-source-points (SSPs), which are dominantly contributed by a single source, by comparing the closeness measurements for all sources $\{\rho_j(t,f), \forall j\}$ (4.1.10): Given a TF point, if the smallest closeness is smaller than a threshold, then we regard this point as an SSP of the corresponding source. We denote the STFT of the n-th identified SSP of the k-th source as $\hat{c}_k(n)$, $\forall k \in \{1, \cdots, J\}$, and assume the j-th source is the target source, the ATF of the target source and the covariance matrix of the interference signals can be estimated by

$$\begin{aligned} \hat{\mathbf{a}}_j &= \frac{1}{N}\sum_{n=1}^{N} \hat{\mathbf{c}}_j(n)/\hat{c}_{1j}(n), \\ \hat{\mathbf{\Sigma}}_{\mathbf{u}} &= \frac{1}{N}\sum_{n=1}^{N}(\sum_{k\neq j}\hat{\mathbf{c}}_j(n))(\sum_{k\neq j}\hat{\mathbf{c}}_j(n))^{\mathsf{H}}, \end{aligned} \tag{4.2.21}$$

where N is the total number of SSPs of each source, and note that, instead of estimating the ATF, we actually estimate the relative transfer function (RTF) (2.3.3), a normalized ATF, to prevent the scale ambiguity (see Sec. 2.3.1 for more detail). Now we can calculate the weights for separating the j-th source according to the spatial filtering criteria that we introduced in the last section.

4.2.3 Source separation

Instead of binary masking (4.1.11), we propose a weighted summation whose weights are calculated according to the spatial filtering criteria, e.g., the criteria of MVDR (4.2.6) and ISR (4.2.12). Denote the computed weights for separating the j-th source as $\mathbf{w}_j(f)$, the source separation is given by

$$\hat{s}_j(t,f) = \begin{cases} \mathbf{w}_j^{\mathsf{H}}(f)\mathbf{x}(t,f), & \text{if } \rho_j(t,f) < \rho_k(t,f), \forall k \neq j, \\ 0, & \text{otherwise.} \end{cases} \tag{4.2.22}$$

The improved DUET is detailed in Algorithm 1. In this paper, we focus on the two-microphone case, the minimum required number of microphones. In fact, more microphones can be used to calculate more weights, so the separation effect will be better.

Algorithm 1: Improved DUET by soft-filtering [50]

Input: two channels of the mixed signals, the number of sources J, a closeness threshold μ

Output: J separated source signals

Implement STFT on the mixed signals;
Calculate the local mixing parameters, the arrival time difference $\delta(t, f)$ (4.1.5) and the symmetric attenuation ratio $\alpha(t, f)$ (4.1.6), for each TF point;
Build up the histogram described in Sec. 4.1.2;
Find J peaks on the histogram and calculate the closeness between the local mixing parameters of each TF point and each peak by (4.1.10). Denote the closeness $\rho_j(t, f), j = 1, \ldots, J$;
$\forall j \in \{1, \ldots, J\}$, label the TF point the j-th source's SSP if $\forall k \neq j$, $\rho_j(t, f) < \rho_k(t, f)$ and $\rho_j(t, f) < \mu$;
for $j = 1$; $j \leq J$; $j = j + 1$ **do**
    Estimate the ATF of the j-th source and the covariance matrices of the interference signals by (4.2.21);
    Calculate the weights by (4.2.11) or (4.2.13);
    Separate mixed signals via (4.2.22)
end
Return the separated source signals in the time domain by implementing the inverse STFT.

4.3 Experiment

We use the standard audio source separation metrics: source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifact ratio (SAR) (which we have introduced in Sec. 3.3) to evaluate the performance of the proposed algorithm. The mixed signals are real recordings using two microphones. The recordings are from the SiSEC2011 [44] competition and recorded in a conference room with 130 ms or 250 ms reverberation. The conference room has a dimension of 4.45 m $\times$ 3.55 m $\times$ 2.5 m. The four speakers are located 1 m from the two microphones, and the two microphones are separated by 5 cm. The azimuth angles of the speakers are $-50°$, $-10°$, $15°$, and $45°$. All recordings have a duration of 10-second.

We have tested our algorithm by separating two mixed signals of two, three, and four sources. The benchmark methods include DUET [4], the independent vector analysis (IVA) [5], the independent low rank matrix analysis (ILRMA) [6], the multi-channel non-negative matrix factorization (MULTINMF) [7], and the full-rank covariance model (FULLRANK) [8]. The parameter settings of the above algorithms refer to their original literature. For each algorithm, we tested many STFT windows of different lengths and

finally chose the window length that gave the largest SDR. Our final choice was a 3/4 overlapping Hann window of 1024 points long (64 ms) for the proposed method, DUET and IVA. The semi-overlapping 4096-point length (256 ms) Hann window is used for ILRMA, and the literature [13, 16] also obtained the same optimal window length for ILRMA as ours. A 3/4 overlapping 2048 point long (128 ms) Hann window was applied to MULTINMF and ILRMA. Other window functions and overlap ratios with different window lengths do not bear higher SDR values. The threshold for identifying SSPs was set to 0.05. The number of bases of ILRMA was 2 which is appropriate for speech signals [6]. The number of components per source in MULTINMF was 10 as recommended in [7].

4.4 Experiment Result

4.4.1 Separation performance

2 mixtures of 2 sources

Average performance was obtained by separating 24 pairs of dual-channel mixed signals of two speech sources in the presence of 130ms and 250ms reverberation. Six different directions of arrival (DOA) combinations (two out of four azimuth angles) and four different gender combinations (male or female for each source) produced these 24 pairs of mixed signals. We compared our algorithm (weights were calculated according to ISR criterion) with DUET [4], IVA [5], and ILRMA [6]. As we can see in Fig. 4.4, the proposed method performs much better than the benchmark algorithms in terms of SDR, SIR, and SAR. The better performance than DUET implies that the weighted summation successfully removes the interference component in TF points where both the target source and the interference sources contribute. Both IVA and ILRMA are independent component analysis (ICA) based algorithms. ICA assumes that source signals are independent to each other, hence, ICA separates signals by maximizing some metrics measuring signals' independence, e.g., kurtosis and mutual information [51]. Although maximizing signals' independence has a certain separation effect, it is not as good as our method. This may be because independence is not a key property of separating speech signals, but a sparse property we mentioned earlier, that is, a partially disjoint relationship in the spectrogram.

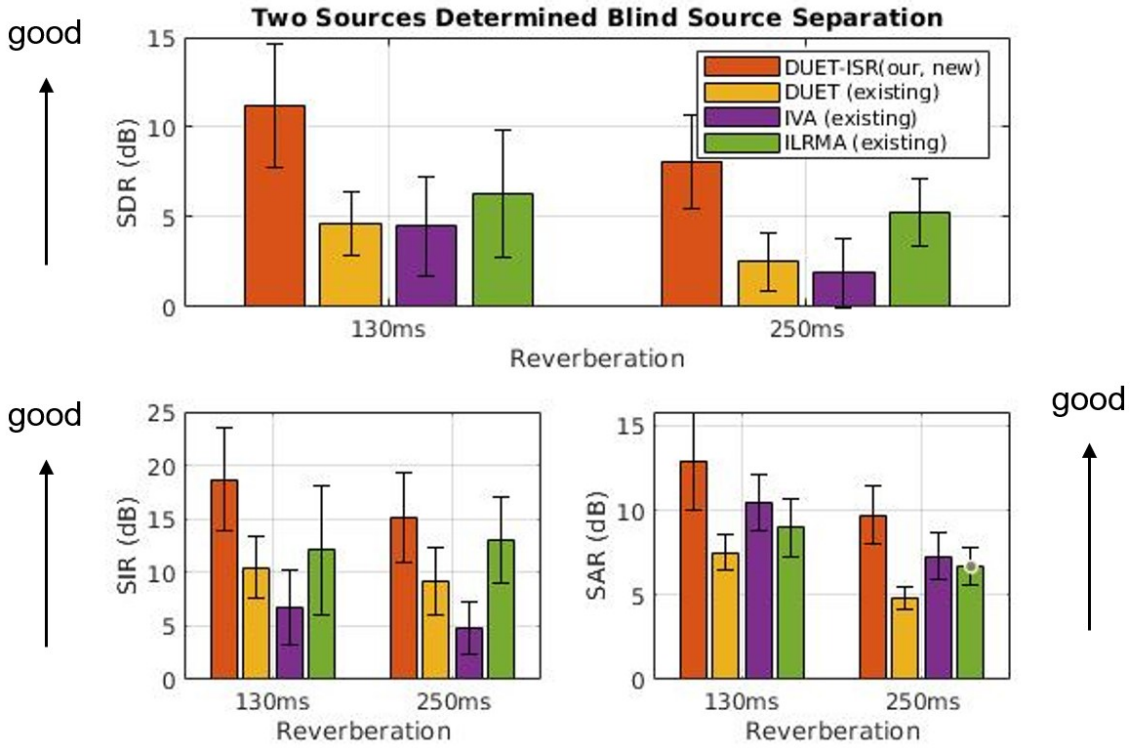

Figure 4.4: Average BSS performance comparison (output SDRs, SIRs, and SARs) for two mixtures of two sources in the presence of 130 ms and 250 ms reverberation time. The red, yellow, purple, and green are our proposed algorithm, DUET [4], IVA [5], and ILRMA [6], respectively.

2 mixtures of 3 sources

Average performance was obtained by separating 32 pairs of dual-channel mixed signals of three speech sources in the presence of 130ms and 250ms reverberation. Four different DOA combinations (three out of four azimuth angles) and eight different gender combinations (male or female for each source) resulted in these 32 pairs of mixed signals. We compared our proposed algorithms with DUET, MULTINMF [7], and FULLRANK [8]. We did not compare with IVA and ILRMA because they require the number of sources are equal to the number of microphones. The result is shown in the Fig. 4.5. MULTINMF and FULLRANK are generative models that assume the signal follows a Gaussian distribution. But speech signals are not Gaussian signals, they are more like random signals that follow a Laplacian distribution. If a Laplacian distribution is assumed, the reconstructed signals will be more sparse. Maybe MULTINMF and FULLRANK will have a better performance.

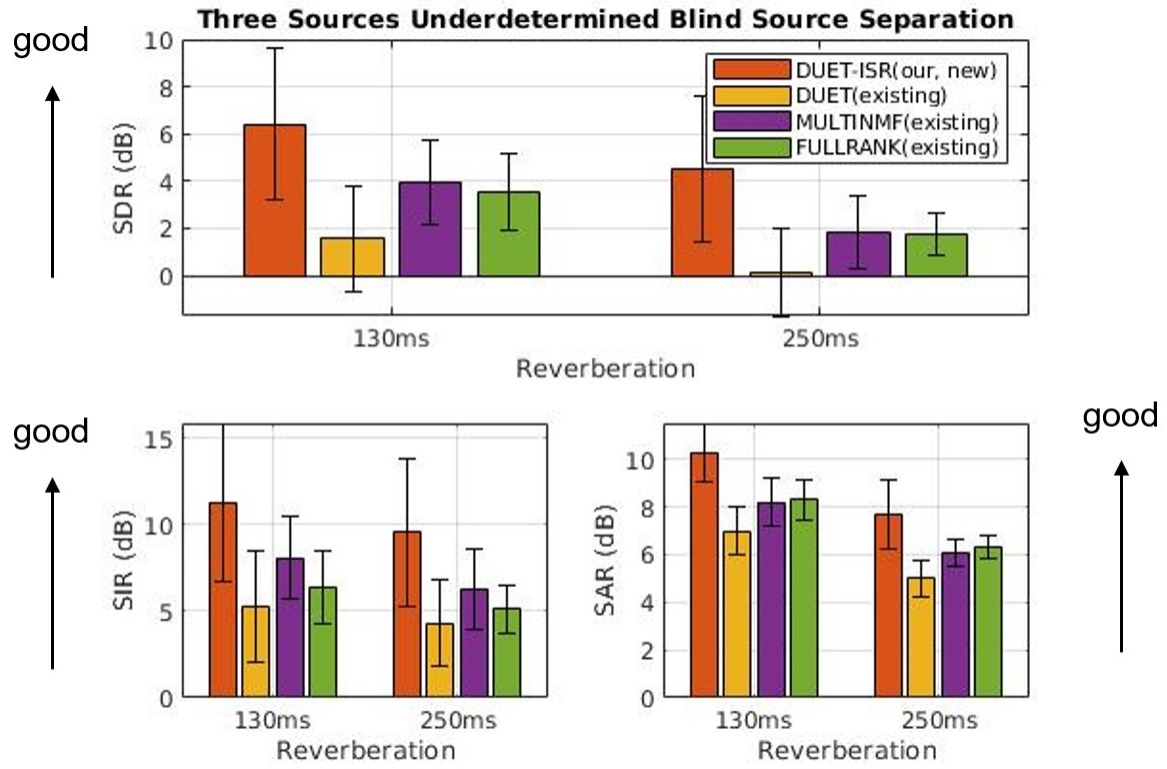

Figure 4.5: Average BSS performance comparison (output SDRs, SIRs, and SARs) for two mixtures of three sources in the presence of 130 ms and 250 ms reverberation time. The red, yellow, purple, and green are our proposed algorithm, DUET [4], MULTINMF [7], and FULLRANK [8], respectively.

2 mixtures of 4 sources

There are 16 pairs of dual-channel mixed signals of four speech sources to evaluate the average separation performance. The result is shown in Fig. 4.6.

Comparison between ISR and MVDR

We also compared the separation performance of different weights according to different spatial filtering criteria. Fig. 4.7 shows the comparison between ISR (red) and MVDR (blue). Clearly, ISR performs much better than MVDR in terms of SDR, SIR, and SAR. This implies that the destruction by the misalignment error of ATF for MVDR is more serious than the innate distortion for the ISR. According to our perceptual evaluation, MVDR introduces more low frequency noise coming from the interference signals. This may be caused by the fact that the estimated target source direction is deviated towards the direction of the interference sources due to the estimation error. ISR, on the other hand, moves the energy towards the high frequencies, thus the separated signals sound

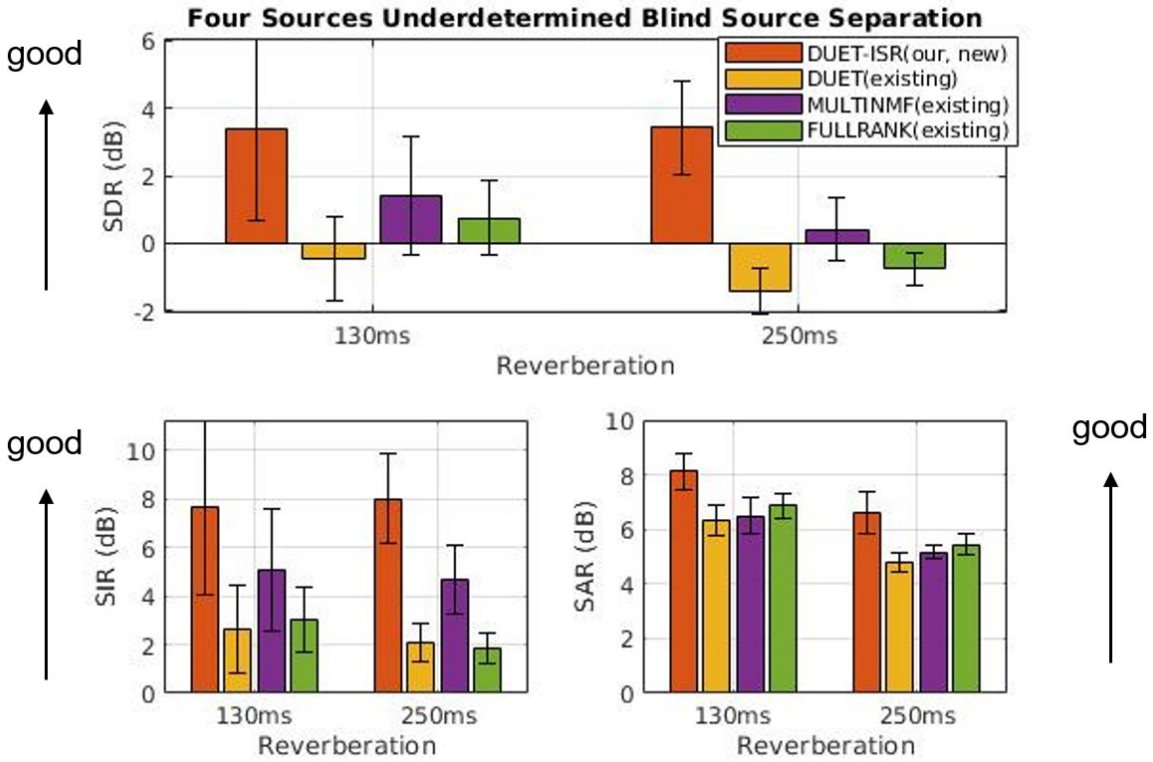

Figure 4.6: Average BSS performance comparison (output SDRs, SIRs, and SARs) for two mixtures of four sources in the presence of 130 ms and 250 ms reverberation time. The red, yellow, purple, and green are our proposed algorithm, DUET [4], MULTINMF [7], and FULLRANK [8], respectively.

like being processed by some high pass filter.

Comparison between different thresholds

The threshold μ determines whether an SSP is strictly or not strictly selected. A smaller threshold means that a TF point should be recognized as an SSP point when its local mixing parameters are close enough to the global mixing parameters. If the threshold is infinity, then a TF point is identified as an SSP of the j-th source as long as its local mixing parameters are closer to the global mixing parameters corresponding to the j-th source. At this time, the TF point may not be contributed by a single source, but by multiple sources. Fig 4.8 shows a comparison between $mu = 0.05$ and $\mu = \infty$. When the number of sources is two or three, the results indicate that there seems not necessary to have a strict SSP detection. However, the number of sources is four, a strict SSP detection will lead to a better separation performance. This is because when the number of sources is larger, more TF points will be occupied by multiple sources, which will reduce the

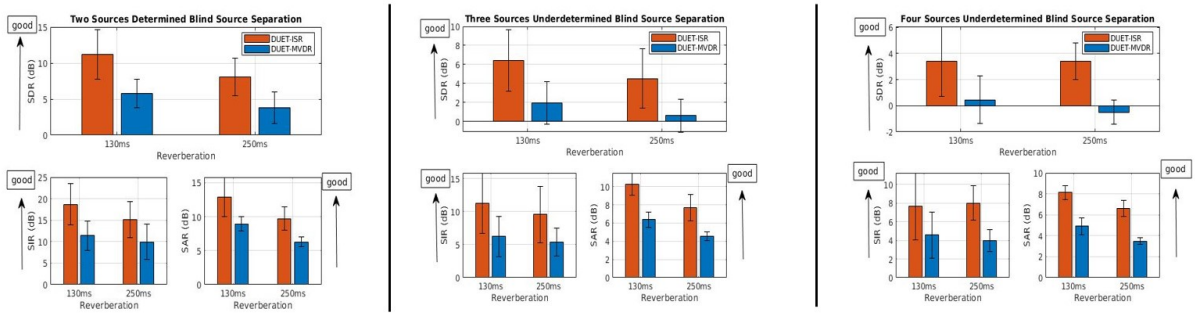

Figure 4.7: Average BSS performance comparison (output SDRs, SIRs, and SARs) between the soft filtering with the proposed ISR (red) and with the well-known MVDR [9] (blue) in the presence of 130 ms and 250 ms reverberation time.

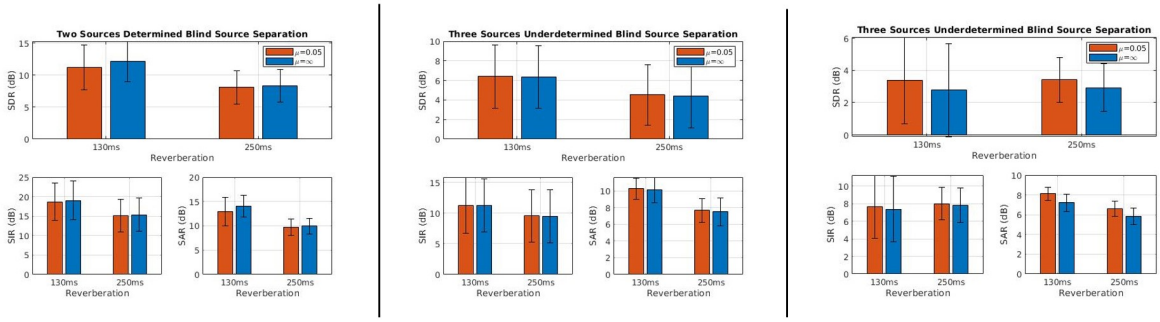

Figure 4.8: Average BSS performance comparison between a small μ ($\mu = 0.05$, strict selection of SSP) and a large μ ($\mu = \infty$, no selection of SSP) in the presence of 130 ms and 250 ms reverberations.

accuracy of parameter estimation. Therefore, strict SSP detection is needed when the number of sources increases.

4.4.2 Time consumption

Table 4.1 records the time consumption for each algorithm we performed above to separate one pair of mixed signals. Our proposed algorithm not only possesses the best performance, but also is computationally cheap, and can complete the separation of a 10-second length of speech signal within one second, far less than MULTINMF and FULLRANK. Although it is not fair to compare with the DUET as it almost failed in the separation of four sources, we still give the time of its separation for completeness.

	2 sources	3 sources	4 sources
DUET-MVDR	0.74 s	0.84 s	0.96 s
DUET-ISR	0.73 s	1.00 s	1.01 s
DUET [4]	0.1 s	0.14 s	0.18 s
IVA [5]	3.40 s	\	\
ILRMA [6]	8.39 s	\	\
MULTINMF [7]	\	32.85 s	43.54 s
FULLRANK [8]	\	575.32 s	585.39 s

Table 4.1: Time consumption of different algorithms for one pair of mixtures.

4.5 Summary

In this work, we modified the DUET algorithm and propose a computationally efficient and cost-effective algorithm for blind speech separation. The key idea is to exploit the partially-disjoint nature of speech in the spectrogram to compute weights according to spatial filtering criteria and do weighted summation instead of binary masking to separate signals. The proposed algorithm achieves more than 5 dB SIR improvement and 4 dB SDR improvement than DUET. It also outperforms the second best-performed algorithm, MULTINMF (2 to 3 dB SDR improvement), but only takes 1/40th of the time to compute in the task of separating four speech sources with two microphones. Improvement for the proposed algorithm can be achieved by applying more sophisticated single-source-point detection (or partially disjoint region detection), spatial filtering criteria, and more microphones.

Chapter 5

Improving l_1 Minimization

In many real applications of interest, we want to reconstruct an object $\mathbf{s_0} \in \mathbb{F}^{J\times 1}$ from data $\mathbf{x} \in \mathbb{F}^{I\times 1}$, where $\mathbb{F}$ is either the real numbers $\mathbb{R}$ or complex numbers $\mathbb{C}$. The object and the data often satisfy a linear relationship, i.e., $\mathbf{x} = \mathbf{As_0}$, and commonly there are fewer observations, i.e., $I < J$. Matrix $\mathbf{A}$ is called mixing matrix in signal separation or sensing matrix in compressed sensing. Given $\mathbf{A}$, solving the linear equations is an ill-posed problem since many candidates exist. It is well-known now, that if the object we wish to recover is sparse, that is, it only depends on a smaller number of parameters, exact recovery may be obtained by searching for the feasible solution with the minimal l_1 norm [52, 53]. It is also established [52, 54] that when $\mathbf{A}$ is chosen from a suitable distribution, with a high probability, $\mathbf{s_0}$ can be perfectly recovered if the number of non-zero entries is smaller than I/α with $\alpha = O(\log \frac{J}{I})$. Moreover, even $\mathbf{s_0}$ is only approximately sparse and the observed data has noise, the feasible solution with the minimal l_1 norm is still reasonably close to $\mathbf{s_0}$. Due to the stable recovery ability from significantly fewer observations, l_1 minimization has drawn a tremendous interest in the field of compressed sensing. Its application involves medical imaging, data compression, seismology, astronomy, sensor network, and so on.

l_1 minimization has also been widely used to solve speech separation [22, 55–57]. Compared with the famous independent component analysis (ICA) based speech separation algorithms, speech separation by l_1 minimization can handle situations where the number of sources is greater than the number of microphones. Since speech signals are not sparse in the time domain, in the context of speech separation, recall the linear model $\mathbf{x} = \mathbf{As_0}$, $\mathbf{x}$ and $\mathbf{s_0}$ are some appropriate representations of the mixed and source signals, respectively. Perhaps the most common representation for a speech signal is its coefficients of the Short-time Fourier transform (STFT). A speech separation algorithm based

on l_1 minimization is usually split into two successive steps. The first step is to estimate the mixing matrix $\mathbf{A}$. There are many articles addressing this problem, e.g., [24, 25, 58]. The second step is to estimate the source signal by l_1 minimization given the mixing matrix [20, 21]. The focus of this work is on the second step. Note that, because the problem is underdetermined, that is, there are more sources than the microphones, hence, even given the mixing matrix, the second step is not a trivial problem. Both steps can also be alternatively repeated in the iteration [22].

Although a lot of works demonstrate the validity of l_1 minimization in speech separation, is there a better alternative than pursuing the minimal l_1 norm? Arberet, Vandergheynst, and the rest [21] tried to use a weighted l_1 minimization where the weights are chosen according to a reweighted scheme [59], and the results show a superiority over the plain l_1 minimization, but it is not a fair comparison because the computational time of the reweighted l_1 minimization is several times that of the plain l_1 minimization.

In this work, we propose a better alternative than the l_1 minimization for speech separation, a novel weighted l_1 minimization called adaptive weighted l_1 minimization. We observe that, with proper normalization, the STFT coefficients of mixed signals are not arbitrarily distributed, but clustered together. This gives us an opportunity to roughly estimate the source signal and calculate the weights based on that. The weights can be easily computed, a definite advantage over the weights computed via the reweighed scheme. The weighted l_1 minimization can be solved by the same solver of the l_1 minimization, so no extra effort is required. The result shows a much better (up to 5 dB SDR) separation performance than that of the plain l_1 minimization. Further improvement can be obtained by utilizing the reweighted scheme. The improvement is robust to the estimation error of the mixing matrix, which is proved by manually adding Gaussian noise to the mixing matrix and real estimation error. Although the adaptive weighted l_1 minimization and l_1 minimization can be solved using the same convex optimization solver, in practice, the solution of adaptive weighted l_1 minimization converges faster and can save up to 35% of the time. Let's dive into more details in the following sections.

5.1 l_1 Minimization

Is it possible to reconstruct a signal when the measurement is incomplete? The Nyquist sampling theorem tells us it is possible when the sampling frequency is at least twice larger than the signal's frequency. A further question is: can the signal be compressed more? The answer is still yes under certain conditions. Milestone works have been done by Donoho [60], Candes and the rest [61]. The key condition is that the signal (or its representation) has to be sparse. This is intuitive because if the signal is sparse, there is natural room to compress the signal. The truly surprising is that the restoration can be done by a simple convex optimization, that is, to minimize the l_1 norm of the reconstructed signal:

$$\begin{aligned} \min_{\mathbf{s}} \quad & \|\mathbf{s}\|_1 \\ \text{s.t.} \quad & \mathbf{x} = \mathbf{A}\mathbf{s}, \end{aligned} \tag{5.1.1}$$

where $\mathbf{s} \in \mathbb{F}^{J\times 1}$ is the reconstructed signal, $\mathbf{x} \in \mathbb{F}^{I\times 1}$ is the observed data, and $\mathbf{A} \in \mathbb{F}^{I\times J}$ is a matrix with less rows than columns, i.e., $I < J$ (hence, the measurement is incomplete), and $\|\mathbf{s}\|_1$ is the one-norm of $\mathbf{s}$: $\|\mathbf{s}\|_1 := \sum_{j=1}^{J} |s_j|$. $\mathbb{F}$ is either the real numbers $\mathbb{R}$ or complex numbers $\mathbb{C}$.

5.1.1 Recovery condition

In this section, we discuss the exact recovery condition of l_1 minimization, i.e., in what condition the optimal solution of the l_1 minimization (5.1.1) is equal to $\mathbf{s_0}$. The discussion is on the real number domain, i.e., $\mathbb{F} = \mathbb{R}$ and partially refers to the book [62] but all the proofs are independently given by us. We first give a weak recovery condition based on mutual coherence for qualitative description, then give a strong recovery condition based on the restricted isometry property (RIP). The strong condition means we have a larger recovery range compared to the weak condition.

Recovery under incoherence

l_1 minimization (5.1.1) is a convex relaxation of l_0 minimization. Hence, we begin with the exact recovery condition of l_0 minimization, then the condition of l_1 minimization. l_0

minimization aims to find the sparsest solution satisfying the linear equations:

$$\begin{aligned} \min_{\mathbf{s}} \quad & \|\mathbf{s}\|_0 \\ \text{s.t.} \quad & \mathbf{x} = \mathbf{A}\mathbf{s}, \end{aligned} \tag{5.1.2}$$

where $\|\mathbf{s}\|_0$ is the zero-norm of $\mathbf{s}$: the number of non-zero entries in $\mathbf{s}$. Since l_0 minimization pursues the sparsest solution, it is not surprising that the optimal solution "hits" the original solution if the original solution is also sparse. The formal definition of the sparseness of a vector is:

Definition 5.1 (k-sparse)**.** *A vector* $\mathbf{s}$ *is* k*-sparse means at most* k *entries of* $\mathbf{s}$ *are non-zero, i.e.,* $\|\mathbf{s}\|_0 \leq k$*.*

The recovery condition of l_0 minimization (5.1.2) consists of two parts:

- The original solution $\mathbf{s_0}$ is sparse, i.e., $\|\mathbf{s_0}\|_0$ is small.
- The matrix $\mathbf{A}$ has a nice structure: the column vectors of $\mathbf{A}$ are as linearly independent as possible.

If $\mathbf{A}$ has a nicer structure, then $\mathbf{s_0}$ can be less sparse, and vice versa. The mathematical way to define the structure of $\mathbf{A}$ is the Kruska rank [63]:

Definition 5.2 (Kruskal Rank)**.** *The Kruskal rank of a matrix* $\mathbf{A}$*, denoted by krank(*$\mathbf{A}$*), is the maximum number* r *such that any* r *column vectors of* $\mathbf{A}$ *are linearly independent.*

If a matrix has a large Kruskal rank, its vectors in the null space are dense, which is the key for exact recovery by l_0 minimization.

Theorem 5.3 (l_0 Recovery)**.** *Suppose* $\mathbf{x} = \mathbf{A}\mathbf{s_0}$*, if* $\|\mathbf{s_0}\|_0 < \frac{1}{2}(krank(\mathbf{A}) + 1)$*, then* $\mathbf{s_0}$ *is the unique optimal solution of the* l_0 *minimization:*

$$\begin{aligned} \min_{\mathbf{s}} \quad & \|\mathbf{s}\|_0 \\ s.t. \quad & \mathbf{x} = \mathbf{A}\mathbf{s}. \end{aligned} \tag{5.1.3}$$

Proof. Suppose we have $\mathbf{s}'$ such that $\mathbf{x} = \mathbf{A}\mathbf{s}'$, then $\mathbf{A}(\mathbf{s}' - \mathbf{s_0}) = 0$, i.e., $(\mathbf{s}' - \mathbf{s_0}) \in \text{null}(\mathbf{A})$. According to the triangular inequality of l_0 norm, we have $\|\mathbf{s}'\|_0 + \|\mathbf{s_0}\|_0 \geq \|\mathbf{s}' - \mathbf{s_0}\|_0$. Note that $\|\mathbf{s}' - \mathbf{s_0}\|_0 \geq \text{krank}(\mathbf{A}) + 1$, thus $\|\mathbf{s}'\|_0 + \|\mathbf{s_0}\|_0 \geq \text{krank}(\mathbf{A}) + 1$. If $\frac{1}{2}(\text{krank}(\mathbf{A}) + 1) > \|\mathbf{s_0}\|_0$, then we have $\|\mathbf{s}'\|_0 \geq \text{krank}(\mathbf{A}) + 1 - \|\mathbf{s_0}\|_0 > \frac{1}{2}(\text{krank}(\mathbf{A}) + 1)$. Proof is complete. □

If entries of $\mathbf{A} \in \mathbb{R}^{I\times J}$ (assume $I \leq J$) are from i.i.d. distribution with a probability density function (pdf), then with probability 1, it is a full rank matrix and krank$(\mathbf{A}) = I$. This means if $\mathbf{s_0}$ has at most $I/2$ non-zero entries, it can be exactly recovered by l_0 minimization almost surely. In other words, for a k-sparse vector $\mathbf{s_0}$, generally speaking, only $2k$ linear measurements are sufficient to keep all the information. However, solving the l_0 minimization is NP-hard. This means when the dimension is large, finding the optimal solution is computationally intractable. Thus, people propose an alternative which can be computed more efficiently - the l_1 minimization, a convex relaxation of the l_0 minimization.

For the exact recovery of the l_1 minimization, compared to the l_0 minimization, we need a nicer structure. For the l_0 minimization, we require the column vectors of $\mathbf{A}$ to be as independent as possible, now, we require the column vectors are as dispersed as possible, or as orthogonal as possible. Mutual coherence is defined as a measure of the dispersion of column vectors in a matrix.

Definition 5.4 (Mutual Coherence)**.** *The mutual coherence of a matrix* $\mathbf{A} = [\mathbf{a}_1|\cdots|\mathbf{a}_J] \in \mathbb{R}^{I\times J}$ *without zero column vector, written as* $\mu(\mathbf{A})$*, is the maximum absolute value of the inner product of any two normalized column vectors in* $\mathbf{A}$*:*

$$\mu(\mathbf{A}) := \max_{i\neq j} \left| < \frac{\mathbf{a}_i}{\|\mathbf{a}_i\|_2}, \frac{\mathbf{a}_j}{\|\mathbf{a}_j\|_2} > \right|. \tag{5.1.4}$$

If each column vector in $\mathbf{A}$ is orthogonal to each other, then $\mu(\mathbf{A}) = 0$. However, this will not happen when $I < J$, and $\mu(\mathbf{A})$ will increase as J increases. The relationship between the Kruskal rank and the mutual coherence is given by the theorem:

Theorem 5.5 (Mutual Coherence and Kruskal Rank)**.** *Given a matrix* $\mathbf{A} \in \mathbb{R}^{I\times J}$*, its mutual coherence (*$\mu(\mathbf{A})$*) and Kruskal rank (*$krank(\mathbf{A})$*) satisfy:*

$$\frac{1}{\mu(\mathbf{A})} \leq krank(\mathbf{A}). \tag{5.1.5}$$

Proof. Without losing any generality, suppose each column vector of $\mathbf{A}$ has a unit l_2 norm (The normalization will not change krank$(\mathbf{A})$) nor $\mu(\mathbf{A})$). Suppose $\mathbf{A}_k$ is a submatrix consisting of any k column vectors of $\mathbf{A}$, $\mathbf{B} = \mathbf{A}_k^\mathsf{T}\mathbf{A}_k$, and b_{ij} is the entry in the i-th row and the j-th column of $\mathbf{B}$. Note that the diagonal entries of $\mathbf{B}$ is 1, according to the Gershgorin circle theorem, we have

$$1 + \max_j \sum_{i\neq j} |b_{ij}| \geq \sigma_{\max}(\mathbf{A}_k^\mathsf{T}\mathbf{A}_k) \geq \sigma_{\min}(\mathbf{A}_k^\mathsf{T}\mathbf{A}_k) \geq 1 - \max_j \sum_{i\neq j} |b_{ij}|. \tag{5.1.6}$$

According to the definition of the mutual coherence, we have $b_{ij} \leq \mu(\mathbf{A})$, $\forall i$, j and $i \neq j$. Hence, we have

$$1 + (k-1)\mu(\mathbf{A}) \geq \sigma_{\max}(\mathbf{A}_k^{\mathsf{T}}\mathbf{A}_k) \geq \sigma_{\min}(\mathbf{A}_k^{\mathsf{T}}\mathbf{A}_k) \geq 1 - (k-1)\mu(\mathbf{A}). \tag{5.1.7}$$

If $k = \text{krank}(\mathbf{A}) + 1$, then there must be a submatrix $\mathbf{A}_k$ which is not a full column rank matrix. Since (5.1.7) holds for any submatrix, we have

$$0 \geq 1 - \text{krank}(\mathbf{A})\mu(\mathbf{A}). \tag{5.1.8}$$

The proof is completed. □

Since computing a Kruskal rank of a matrix is an NP-hard problem, the Theorem 5.5 provides an computationally tractable way to check whether the l_0 minimization can exactly recover the original solution, that is, if $\|\mathbf{s_0}\| < \frac{1}{2}(\mu^{-1}(\mathbf{A}) + 1)$, then $\mathbf{s}_0$ is the unique minimizer of the l_0 minimization. Intriguingly, such $\mathbf{s_0}$ is also the unique optimal solution of the l_1 minimization if $\mathbf{A}$ has unit l_2 norm columns:

Theorem 5.6 (l_1 Recovery under Incoherence [64, 65])**.** *Suppose* $\mathbf{x} = \mathbf{A}\mathbf{s_0}$*, if* $\|\mathbf{s_0}\|_0 < \frac{1}{2}(\mu^{-1}(\mathbf{A}) + 1)$ *and each column of* $\mathbf{A}$ *has unit* l_2 *norm, then* $\mathbf{s_0}$ *is the unique optimal solution of the* l_1 *minimization:*

$$\begin{aligned} \min_{\mathbf{s}} \quad & \|\mathbf{s}\|_1 \\ s.t. \quad & \mathbf{x} = \mathbf{A}\mathbf{s}. \end{aligned} \tag{5.1.9}$$

Proof. Suppose $\mathbf{s}'$ is an optimal solution of the l_1 minimization and $\mathbf{s}' \neq \mathbf{s_0}$ for the sake of the contradiction. Let S be the support of $\mathbf{s_0}$ and denote the rest entries' location by S^c. For any vector $\mathbf{y}$, define $\mathbf{y}_S$ such that the i-th entry of $\mathbf{y}_S$ equals to the i-th entry of $\mathbf{y}$ if $i \in S$, otherwise 0. The whole proof can be split into two steps, and the final result can be obtained from the contradiction produced by combining the two sub-results. Denote $\mathbf{e} = \mathbf{s}' - \mathbf{s_0}$. First we shall prove that $\|\mathbf{e}\|_1 \leq 2\|\mathbf{e}_S\|_1$. According to our assumption, we have $\|\mathbf{s}'_S + \mathbf{s}'_{S^c}\|_1 \leq \|\mathbf{s_0}\|_1 \Rightarrow \|\mathbf{s}'_S\|_1 + \|\mathbf{s}'_{S^c}\|_1 \leq \|\mathbf{s_0}\|_1 \Rightarrow \|\mathbf{s_0}\|_1 - \|\mathbf{s}'_S\|_1 \geq \|\mathbf{s}'_{S^c}\|_1$. According to the triangular inequality, we have $\|\mathbf{s}'_S - \mathbf{s_0}\|_1 \geq \|\mathbf{s}'_{S^c}\|_1$. Note that $\mathbf{e}_S = \mathbf{s}'_S - \mathbf{s_0}$ and $\|\mathbf{e}_{S^c}\|_1 = \|\mathbf{e}\|_1 - \|\mathbf{e}_S\|_1$. Hence, we have

$$\|\mathbf{e}\|_1 \leq 2\|\mathbf{e}_S\|_1. \tag{5.1.10}$$

The second step is to prove that $|e_i| \leq \frac{\mu(\mathbf{A})}{1+\mu(\mathbf{A})}\|\mathbf{e}\|_1$, $\forall i$ (e_i is the i-th entry of $\mathbf{e}$). Suppose $\mathbf{B} = \mathbf{A}^{\mathsf{T}}\mathbf{A}$ and b_{ij} is the entry in the i-th row and j-th column of $\mathbf{B}$. Since $\mathbf{A}\mathbf{e} = 0$,

we have $\mathbf{Be} = 0$. This means $\forall i$, we have $e_i + \sum_{j,\, j\neq i} b_{ij}e_j = 0 \Rightarrow |e_j| = |\sum_{j,\, j\neq i} b_{ij}e_j| \leq \sum_{j,\, j\neq i} |b_{ij}||e_j|$. Note that $\mu(\mathbf{A}) = \max_{i,\, j,\, i\neq j} |b_{ij}|$. Thus, we have $|e_i| \leq \mu(\mathbf{A}) \sum_{j,\, j\neq i} |e_j| = \mu(\mathbf{A})(\|\mathbf{e}\|_1 - |e_i|)$. It is equivalent to say that

$$|e_i| \leq \frac{\mu(\mathbf{A})}{1+\mu(\mathbf{A})}\|\mathbf{e}\|_1, \quad \forall i. \tag{5.1.11}$$

If we do a summation of $|e_i|$ over the support S, according to (5.1.11), we have $\|\mathbf{e}_S\|_1 \leq \frac{\mu(\mathbf{A})}{1+\mu(\mathbf{A})}\|\mathbf{e}\|_1|S|$, where $|S|$ is the number of elements in S. Combining this inequality with (5.1.10), we have

$$\begin{aligned} \|\mathbf{e}_S\|_1 &\leq \frac{\mu(\mathbf{A})}{1+\mu(\mathbf{A})}\|\mathbf{e}\|_1|S| \\ &\leq \frac{2\mu(\mathbf{A})}{1+\mu(\mathbf{A})}\|\mathbf{e}_S\|_1|S|. \end{aligned} \tag{5.1.12}$$

From (5.1.12), we can easily have the inequality $\|\mathbf{s_0}\| = |S| \geq \frac{1}{2}(1 + \mu^{-1}(\mathbf{A}))$, a contradiction to our condition in the theorem. Thus, there is no optimal solution $\mathbf{s}'$ such that $\|\mathbf{s}'\|_1 \leq \|\mathbf{s_0}\|_1$ and $\mathbf{s}' \neq \mathbf{s_0}$. The proof is completed. □

Theorem 5.6 tells us the recovery ability of l_1 minimization depends on the incoherence of the matrix $\mathbf{A}$: a smaller $\mu(\mathbf{A})$ allows exact recovery for a denser $\mathbf{s_0}$. Here we give lower and upper bounds on the incoherence of a generic matrix $\mathbf{A}$.

Theorem 5.7 (Mutual Coherence Lower Bound [66])**.** *For any matrix $\mathbf{A} \in \mathbb{R}^{I\times J}$ with $I \leq J$, the lower bound of the mutual coherence of $\mathbf{A}$ is given by:*

$$\mu(\mathbf{A}) \geq \sqrt{\frac{J-I}{I(J-1)}}. \tag{5.1.13}$$

Proof. Define $\mathbf{B} := \mathbf{A}^\mathsf{T}\mathbf{A} \in \mathbb{R}^{J\times J}$, $\mathrm{tr}(\mathbf{B})$ as the trace of $\mathbf{B}$, and $\lambda_i(\mathbf{B})$ as the i-th eigenvalue of $\mathbf{B}$. Suppose the rank of $\mathbf{A}$ is r ($r \leq I$), so we have r non-zero eigenvalues for $\mathbf{B}$, denoted as $\lambda_1(\mathbf{B}), \cdots, \lambda_r(\mathbf{B})$. According to the Cauchy–Schwarz inequality, we have $\mathrm{tr}^2(\mathbf{B}) = (\sum_{i=1}^r \lambda_i^2(\mathbf{B}))^2 \leq r\sum_{i=1}^r \lambda_i^2(\mathbf{B}) \leq I\sum_{i=1}^r \lambda(\mathbf{B})$. Hence we have

$$\sum_{i=1}^r \lambda_i^2(\mathbf{B}) \geq \frac{\mathrm{tr}^2(\mathbf{B})}{I} = J^2/I. \tag{5.1.14}$$

About the Frobenius norm of $\mathbf{B}$, we have $\|\mathbf{B}\|_F := \sum_{i,j} b_{ij}^2 = J + \sum_{i\neq j} b_{ij}^2 \leq J + J(J-1)\mu^2(\mathbf{A})$. A fact about the Frobenius norm is that $\|\mathbf{B}\|_F^2 \geq \sum_{i=1}^r \lambda_i^2(\mathbf{B})$, which can be proved by the Schur decomposition. Thus we have

$$J + J(J-1)\mu^2(\mathbf{A}) \geq \sum_{i=1}^r \lambda_i^2(\mathbf{B}). \tag{5.1.15}$$

Combining (5.1.14) and (5.1.15), we can easily have the lower bound of $\mu(\mathbf{A})$. The proof is completed. □

Theorem 5.8 (Mutual Coherence Upper Bound)**.** *For any matrix* $\mathbf{A} = [\mathbf{a}_1|\mathbf{a}_2|\dots|\mathbf{a}_J] \in \mathbb{R}^{I\times J}$ *with column* $\mathbf{a}_j \sim uni(S^{I-1})$ *is independently chosen from a uniform distribution on the sphere. Then with a high probability (can be arbitrarily close to 1),* $\mu(\mathbf{A}) \leq O(\sqrt{\frac{\log(2J)}{I}})$.

Proof. See [62]. □

According to the Theorem 5.8 (the upper bound theorem), if $\mathbf{A}$ is chosen from a suitable distribution and $\|\mathbf{s_0}\|_0 \leq O(\sqrt{\frac{I}{\log(2J)}})$, then with a high probability, the success of the Theorem 5.6 is guarantied. According to the Theorem 5.7 (the lower bound theorem), if I is proportional to the J and I is large, then the success of the Theorem 5.6 requires at least the guarantee that $\|\mathbf{s}_0\|_0 \leq O(\sqrt{I})$. In fact, there are lots of successful examples of l_1 minimization when $\|\mathbf{s}_0\|_0$ is proportional to the I. This indicates that the mutual coherence is too rough to represent the "nice" structure of $\mathbf{A}$. We need a more sharpened theorem for the recovery of l_1 minimization.

Recovery under RIP

In the last section, we discussed how the recovery ability of the l_1 minimization is sufficiently determined by the mutual coherence. The fact that many examples not guaranteed by mutual coherence are successful cases indicates that mutual coherence is a "rough" description of the nice structure of $\mathbf{A}$. Intuitively, finding a better fine description can be benefited from below observation, that is, the inequality (5.1.7) we derived from the Theorem 5.5, which we restated below:

$$1 + (k-1)\mu(\mathbf{A}) \geq \sigma_{\max}(\mathbf{A}_k^\mathsf{T}\mathbf{A}_k) \geq \sigma_{\min}(\mathbf{A}_k^\mathsf{T}\mathbf{A}_k) \geq 1 - (k-1)\mu(\mathbf{A}), \tag{5.1.16}$$

where $\mathbf{A}_k$ is any submatrix of $\mathbf{A}$ consisting of k columns. Note that $\sigma(\mathbf{A}_k^\mathsf{T}\mathbf{A}_k) = \sigma^2(\mathbf{A}_k)$. Since a "nice" $\mathbf{A}$ for l_1 minimization is implied by a small $\mu(\mathbf{A})$, whether it also can be implied by that the singular values of $\mathbf{A}_k$ concentrates around 1? Or equivalently[1] to say for any k-sparse vector $\mathbf{s}$, there exists a small number δ such that

$$(1-\delta)\|\mathbf{s}\|_2^2 \leq \|\mathbf{A}\mathbf{s}\|_2^2 \leq (1+\delta)\|\mathbf{s}\|_2^2. \tag{5.1.17}$$

[1]The equivalence can be easily seen by keeping in mind that for any k sparse vector $\mathbf{s}$, $\mathbf{A}\mathbf{s} = \mathbf{A}_k\mathbf{s}$ where $\mathbf{A}_k$ is the submatrix of $\mathbf{A}$ consisting of the corresponding k columns.

The answer is positive! And what beyond the mutual coherence is that for a "generic" matrix, its δ is small enough if I, J, and k are proportional. We give a formal definition of δ:

Definition 5.9 (Order-k Restricted Isometry Constant)**.** *δ_k is an order-k restricted isometry constant of a matrix $\mathbf{A} \in \mathbb{R}^{I \times J}$ ($k \leq J$) if it is the smallest number such that*

$$(1 - \delta_k)\|\mathbf{s}\|_2^2 \leq \|\mathbf{A}\mathbf{s}\|_2^2 \leq (1 + \delta_k)\|\mathbf{s}\|_2^2 \tag{5.1.18}$$

holds for any k-sparse vector $\mathbf{s}$.

If $\mathbf{A}$ has a small δ_k, it means the mapping $\mathbf{s} \to \mathbf{A}\mathbf{s}$ nearly preserves the amplitude with the restriction that $\mathbf{s}$ is a k-sparse vector. We call this property of $\mathbf{A}$ the restricted isometry property:

Definition 5.10 (Order-k Restricted Isometry Property [67])**.** *A matrix $\mathbf{A}$ satisfies the restricted isometry property of order k if its order-k restricted isometry constant $\delta_k \in [0, 1)$.*

The success of l_0 or l_1 minimization is guaranteed as long as $\mathbf{A}$ has a small restricted isometry constant.

Theorem 5.11 (l_0 Recovery under RIP)**.** *Suppose $\mathbf{x} = \mathbf{A}\mathbf{s_0}$, if k=$\|\mathbf{s_0}\|_0$ and $\delta_{2k} < 1$, then $\mathbf{s_0}$ is the unique optimal solution of the l_0 minimization:*

$$\begin{aligned} \min_{\mathbf{s}} \quad & \|\mathbf{s}\|_0 \\ s.t. \quad & \mathbf{x} = \mathbf{A}\mathbf{s}. \end{aligned} \tag{5.1.19}$$

Proof. If $\delta_{2k} < 1$, then the null space of $\mathbf{A}$ does not have a $2k$-sparse vector according to the definition. This means for any feasible solution $\mathbf{s}'$ satisfying $\|\mathbf{s}'\|_0 + \|\mathbf{s_0}\| \geq \|\mathbf{s}' - \mathbf{s_0}\|_0 > 2k$. Hence $\|\mathbf{s}'\|_0 > k$. □

Theorem 5.12 (l_1 Recovery under RIP)**.** *Suppose $\mathbf{x} = \mathbf{A}\mathbf{s_0}$, if k=$\|\mathbf{s_0}\|_0$ and $\delta_{2k} < \sqrt{2} - 1$, then $\mathbf{s_0}$ is the unique optimal solution of the l_1 minimization:*

$$\begin{aligned} \min_{\mathbf{s}} \quad & \|\mathbf{s}\|_1 \\ s.t. \quad & \mathbf{x} = \mathbf{A}\mathbf{s}. \end{aligned} \tag{5.1.20}$$

Theorem 5.12 is given without a proof as we will give a more general theorem considering the noisy recovery when $\mathbf{s_0}$ is any vector. Considering there is an unknown noise on the observation, i.e., $\mathbf{x} = \mathbf{A}\mathbf{s_0} + \mathbf{n}$ with $\|\mathbf{n}\|_2 \leq \epsilon$, one may search a feasible solution satisfying $\|\mathbf{x} - \mathbf{A}\mathbf{s}\| \leq \epsilon$. Then what can we have from the following convex optimization problem

$$\begin{aligned} \min_{\mathbf{s}} \quad & \|\mathbf{s}\|_1 \\ \text{s.t.} \quad & \|\mathbf{x} - \mathbf{A}\mathbf{s}\| \leq \epsilon. \end{aligned} \tag{5.1.21}$$

Moreover, what if $\mathbf{s_0}$ is not a sparse but a general vector? Amazingly, the difference between $\mathbf{s}^\star$, the minimizer of (5.1.21), and $\mathbf{s_0}$ can be bounded by the ϵ and the difference between $\mathbf{s_0}$ and its best k-sparse approximation $\mathbf{s}_{\mathbf{0},k}$ which is the vector $\mathbf{s_0}$ with all but the k-largest entries set to zero.

Theorem 5.13 (Noisy l_1 Recovery [52, 53])**.** *Suppose* $\mathbf{x} = \mathbf{A}\mathbf{s_0} + \mathbf{n}$, $\delta_{2k} < \sqrt{2} - 1$, *and* $\|\mathbf{n}\|_2 \leq \epsilon$, *then* $\|\mathbf{s}^\star - \mathbf{s_0}\|_2 \leq C_0 k^{-1/2} \|\mathbf{s_0} - \mathbf{s}_{\mathbf{0},k}\|_2 + C_1 \epsilon$, *where* $C_0 > 0$ *and* $C_1 > 0$ *are some constants,* $\mathbf{s}_{\mathbf{0},k}$ *is the best* k*-sparse approximation of* $\mathbf{s_0}$, *and* $\mathbf{s}^\star$ *is the solution to*

$$\begin{aligned} \min_{\mathbf{s}} \quad & \|\mathbf{s}\|_1 \\ \textit{s.t.} \quad & \|\mathbf{x} - \mathbf{A}\mathbf{s}\| \leq \epsilon. \end{aligned} \tag{5.1.22}$$

Proof. Suppose $\mathbf{s}^\star - \mathbf{s_0} = \mathbf{h}$. Denote the support of $\mathbf{s}_{\mathbf{0},k}$ as T_0. Decompose $\mathbf{h}$ into $\mathbf{h}_{T_0}$, $\mathbf{h}_{T_1}$, $\mathbf{h}_{T_2}, \cdots$, each of sparsity at most k. T_1 corresponds to the position of the k largest absolute coefficients of $\mathbf{h}_{T_0^C}$, T_2 corresponds to the position of the k largest absolute coefficients of $\mathbf{h}_{(T_0 \cup T_1)^C}$, and so on. The following calculation rules are often used in the proof: for any two disjoint sets T, $T' \subseteq \{1, 2, \cdots, N\}$, $\mathbf{s}_T + \mathbf{s}_{T'} = \mathbf{s}_{T \cup T'}$ and $\|\mathbf{s}_T\|_1 + \|\mathbf{s}_{T'}\|_1 = \|\mathbf{s}_{T \cup T'}\|_1$. There are four useful facts

1. For $j \geq 2$, $\|\mathbf{h}_{T_j}\|_2 \leq \sqrt{k}\|\mathbf{h}_{T_j}\|_\infty \leq k^{-\frac{1}{2}}\|\mathbf{h}_{T_{j-1}}\|_1$. This is a simple application of the norm inequality.

2. $\|\mathbf{s_0} + \mathbf{h}\|_1 \leq \|\mathbf{s_0}\|_1 \Rightarrow \|\mathbf{h}_{T_0^C}\|_1 \leq 2\|\mathbf{s}_{\mathbf{0}T_0^C}\|_1 + \|\mathbf{h}_{T_0}\|_1$. This can be obtained from $\|\mathbf{s_0} + \mathbf{h}\|_1 = \|\mathbf{s}_{\mathbf{0}T_0} + \mathbf{s}_{\mathbf{0}T_0^C} + \mathbf{h}_{T_0} + \mathbf{h}_{T_0^C}\|_1 \geq \|\mathbf{s}_{\mathbf{0}T_0} + \mathbf{h}_{T_0^C}\|_1 - \|\mathbf{s}_{\mathbf{0}T_0^C} + \mathbf{h}_{T_0}\|_1 = \|\mathbf{s}_{\mathbf{0}T_0}\|_1 + \|\mathbf{h}_{T_0^C}\|_1 - \|\mathbf{s}_{\mathbf{0}T_0^C}\|_1 - \|\mathbf{h}_{T_0}\|_1$.

3. $\|\mathbf{h}_{T_0}\|_1 \leq \sqrt{k}\|\mathbf{h}_{T_0}\|_2$. This is a simple application of the Cauchy-Schwarz inequality.

4. Lemma: for any $\mathbf{s}$, $\mathbf{s}'$ supported on disjoint sets T, $T' \subseteq \{1, 2, \cdots, n\}$ with $|T| \leq k$, $|T'| \leq k'$, then $|\langle \mathbf{A}\mathbf{s}, \mathbf{A}\mathbf{s}' \rangle| \leq \delta_{k+k'}\|\mathbf{s}\|_2\|\mathbf{s}'\|_2$. This is an application of the Parallelogram identity. $|\langle \mathbf{A}\mathbf{s}, \mathbf{A}\mathbf{s}' \rangle| = \frac{1}{4}(\|\mathbf{A}\mathbf{s} + \mathbf{A}\mathbf{s}'\|_2^2 - \|\mathbf{A}\mathbf{s} - \mathbf{A}\mathbf{s}'\|_2^2) \leq \frac{1}{4}((1 + \delta_{k+k'})\|\mathbf{s} +$

$\mathbf{s}'\|_2^2 - (1+\delta_{k+k'})\|\mathbf{s}-\mathbf{s}'\|_2^2)$). Note that $\|\mathbf{s}+\mathbf{s}'\|_2^2 = \|\mathbf{s}-\mathbf{s}'\|_2^2 = \|\mathbf{s}\|_2^2 + \|\mathbf{s}'\|_2^2$ since $\mathbf{s}$ and $\mathbf{s}'$ are disjoint. Then we have $|\langle \mathbf{As}, \mathbf{As}'\rangle| \le \frac{1}{2}\delta_{k+k'}(\|\mathbf{s}\|_2^2 + \|\mathbf{s}'\|_2^2)$. The upper bound can be tighter by replacing $\mathbf{s}$, $\mathbf{s}'$ with $\frac{\mathbf{s}}{\|\mathbf{s}\|_2}$, $\frac{\mathbf{s}'}{\|\mathbf{s}'\|_2}$. Then we have the lemma.

Now we are ready to start the proof of the theorem. The whole proof can be split into two steps. First, we shall prove that $\|\mathbf{h}_{(T_0\cup T_1)^C}\|_2$ is bounded by $\|\mathbf{h}_{T_0\cup T_1}\|_2$, then we prove that $\|\mathbf{h}_{T_0\cup T_1}\|_2$ is small.

For the first step, according to the fact 1, we have

$$\|\mathbf{h}_{(T_0\cup T_1)^C}\|_2 = \|\sum_{j\ge 2}\mathbf{h}_{T_j}\|_2 \le \sum_{j\ge 2}\|\mathbf{h}_{T_j}\|_2 \le \sqrt{k}\sum_{j\ge 2}\|\mathbf{h}_{T_j}\|_\infty \le k^{-\frac{1}{2}}\|\mathbf{h}_{T_0^C}\|_1. \tag{5.1.23}$$

This can be further bounded according to the fact 2 and the fact 3:

$$\begin{aligned}\|\mathbf{h}_{(T_0\cup T_1)^C}\|_2 &\le k^{-\frac{1}{2}}\|\mathbf{h}_{T_0}\|_1 + 2e_k \\ &\le \|\mathbf{h}_{T_0}\|_2 + 2e_k \\ &\le \|\mathbf{h}_{T_0\cup T_1}\|_2 + 2e_k,\end{aligned} \tag{5.1.24}$$

where $e_k := k^{-\frac{1}{2}}\|\mathbf{s_0} - \mathbf{s_0}_{T_0}\|_1$.

For the second step, according to the definition of restricted isometry property, we have

$$\begin{aligned}(1-\delta_{2k})\|\mathbf{h}_{T_0\cup T_1}\|_2^2 &\le \|\mathbf{Ah}_{T_0\cup T_1}\|_2^2 \\ &= \langle \mathbf{Ah}_{T_0\cup T_1}, \mathbf{Ah}\rangle - \left\langle \mathbf{Ah}_{T_0\cup T_1}, \mathbf{Ah}_{(T_0\cup T_1)^C}\right\rangle \\ &\le |\langle \mathbf{Ah}_{T_0\cup T_1}, \mathbf{Ah}\rangle| + |\left\langle \mathbf{Ah}_{T_0\cup T_1}, \mathbf{Ah}_{(T_0\cup T_1)^C}\right\rangle|.\end{aligned} \tag{5.1.25}$$

Then we try to bounded the two inner products. For the first inner product, since $\|\mathbf{Ah}\|_2 = \|\mathbf{As}^\star - \mathbf{As_0}\|_2 \le \|\mathbf{As}^\star - \mathbf{x}\|_2 + \|\mathbf{x} - \mathbf{As_0}\|_2 \le 2\epsilon$, we have

$$\begin{aligned}|\langle \mathbf{Ah}_{T_0\cup T_1}, \mathbf{Ah}\rangle| &\le \|\mathbf{Ah}_{T_0\cup T_1}\|_2\|\mathbf{Ah}\|_2 \\ &\le 2\epsilon\sqrt{1+\delta_{2k}}\|\mathbf{h}_{T_0\cup T_1}\|_2.\end{aligned} \tag{5.1.26}$$

For the second inner product, according to the fact 4, we have

$$\begin{aligned}|\left\langle \mathbf{Ah}_{T_0\cup T_1}, \mathbf{Ah}_{(T_0\cup T_1)^C}\right\rangle| &= |\left\langle \mathbf{A}(\mathbf{h}_{T_0}+\mathbf{h}_{T_1}), \mathbf{A}\sum_{j\ge 2}\mathbf{h}_{T_j}\right\rangle| \\ &\le \sum_{j\ge 2}|\left\langle \mathbf{Ah}_{T_0}, \mathbf{Ah}_{T_j}\right\rangle| + \sum_{j\ge 2}|\left\langle \mathbf{Ah}_{T_1}, \mathbf{Ah}_{T_j}\right\rangle| \\ &\le (\|\mathbf{h}_{T_0}\|_2 + \|\mathbf{h}_{T_1}\|_2)\delta_{2k}\sum_{j\ge 2}\|\mathbf{h}_{T_j}\|_2.\end{aligned} \tag{5.1.27}$$

Since $\|\mathbf{h}_{T_0}\|_2 + \|\mathbf{h}_{T_1}\|_2 \leq \sqrt{2}\|\mathbf{h}_{T_0 \cup T_1}\|_2$ for T_0 and T_1 are disjoint, we have

$$|\left\langle \mathbf{A}\mathbf{h}_{T_0 \cup T_1}, \mathbf{A}\mathbf{h}_{(T_0 \cup T_1)^C} \right\rangle| \leq \sqrt{2}\delta_{2k}\|\mathbf{h}_{T_0 \cup T_1}\|_2 \sum_{j \geq 2} \|\mathbf{h}_{T_j}\|_2. \tag{5.1.28}$$

It follows from (5.1.25), (5.1.26), (5.1.28), and $\delta_{2k} < \sqrt{2} - 1$, we have

$$\|\mathbf{h}_{T_0 \cup T_1}\|_2 \leq \alpha e_k + \rho\epsilon, \tag{5.1.29}$$

where $e_k = k^{-\frac{1}{2}}\|\mathbf{s_0} - \mathbf{s_0}_{T_0}\|_1$, $\alpha = \frac{2\sqrt{2}\delta_{2k}}{1-(\sqrt{2}+1)\delta_{2k}}$, and $\rho = \frac{2\sqrt{1+\delta_{2k}}}{1-(\sqrt{2}+1)\delta_{2k}}$.

Now we can conclude this proof from (5.1.24) and (5.1.29) that

$$\begin{aligned} \|\mathbf{h}\|_2 &\leq \|\mathbf{h}_{T_0 \cup T_1}\|_2 + \|\mathbf{h}_{(T_0 \cup T_1)^C}\|_2 \\ &\leq C_0 e_k + C_1 \epsilon. \end{aligned} \tag{5.1.30}$$

where $C_0 = 2(\alpha + 1)$ and $C_1 = 2\rho$. □

Theorem 5.12 is a special case of Theorem 5.13 when $\epsilon = 0$ and $\mathbf{s_0}$ is a k-sparse vector. Theorem 5.13 mainly states that if $\mathbf{A}$ has a nice structure, the reconstruction of l_1 minimization has the same performance with the best sparse approximation of $\mathbf{s_0}$ (the performance is measured by the 2-norm distance). Another fact makes the l_1 minimization so useful is than for a generic matrix, its restricted isometry constant is usually small.

Theorem 5.14 (RIP of Gaussian Matrices [61, 68])**.** *Suppose $\mathbf{A} \in \mathbb{R}^{I \times J}$ is a Gaussian matrix with each entry follows i.i.d $\mathcal{N}(0, 1/I)$, then with a high probability, $\delta_k \leq \delta$ if*

$$I \geq Ck \log(J/k)/\delta^2, \tag{5.1.31}$$

where C is some positive constant.

Proof. Please see [61, 62, 68] □

Theorem 5.14 shows the superiority of the restricted isometry constant over the mutual coherence. According to the previous result, the success of the recovery under coherence requires at least $\|\mathbf{s}_0\|_0 = k \leq O(\sqrt{I})$ if I is proportional to J, but now it only requires that (k, I, J) to scale proportionally. For practical purposes, we need to know the exact constant factor C in (5.1.31). However, there is a gap in the smallest number of measurements between the theoretical guarantee and practical good reconstruction by l_1 minimization, i.e., the practical requirement of the minimal number of measurements is far less than that of the theoretical requirement. One of the best theoretical requirement

is given in [69], but still can not explain the good performance of l_1 minimization in speech separation, where there are only two or three microphones ($I = 2$ or 3). The theoretically required minimal number of measurements for signal reconstruction by l_1 minimization in low-dimensional problems,e.g., speech separation, remains an open problem in this thesis. More general, given Gaussian measurements, how is the performance of l_1 minimization in a low dimensional problem.

l_1 minimization v.s. l_2 minimization

One may also consider a l_2 minimization problem

$$\begin{aligned} \min_{\mathbf{s}} \quad & \|\mathbf{s}\|_2 \\ \text{s.t.} \quad & \mathbf{x} = \mathbf{A}\mathbf{s}, \end{aligned} \tag{5.1.32}$$

where $\|\cdot\|$ is the 2-norm: $\|\mathbf{s}\|_2 := \sqrt{\sum_{j=1}^{J} s_j^2}$. Compared to the l_1 minimization, l_2 minimization is more easier to solve since the 2-norm is differentiable everywhere. However, the solution of the l_2 minimization is not sparse. This can be illustrated in Fig. 5.1 by a simple 2-dimensional example. The tangent point $(s_1^{l_2}, s_2^{l_2})$ of the solid circle and the black line is exactly the solution to the l_2 minimization

$$\begin{aligned} \min_{s_1,s_2} \quad & \sqrt{s_1^2 + s_2^2} \\ \text{s.t.} \quad & x = a_1 s_2 + a_2 s_2. \end{aligned} \tag{5.1.33}$$

Both $s_1^{l_2}$ and $s_2^{l_2}$ are non-zero, i.e., the solution is not sparse. For the l_1 minimization, it can be seen in Fig. 5.2, the point $(s_1^{l_1}, s_2^{l_1})$ where the solid 1-norm ball intersects the black line is exactly the solution of the l_1 minimization

$$\begin{aligned} \min_{s_1,s_2} \quad & |s_1| + |s_2| \\ \text{s.t.} \quad & x = a_1 s_2 + a_2 s_2. \end{aligned} \tag{5.1.34}$$

The solution is sparse as $s_1^{l_1} = 0$.

5.1.2 Reweighted l_1 minimization

In the previous section, we have seen the "magical" recovery ability of the l_1 minimization. The robustness to the measurement noise and the recovery guarantee to the nearly sparse signals makes l_1 minimization practically powerful. We also qualitatively give the smallest number of measurements (5.1.31) that l_1 minimization needs for theoretical guarantee

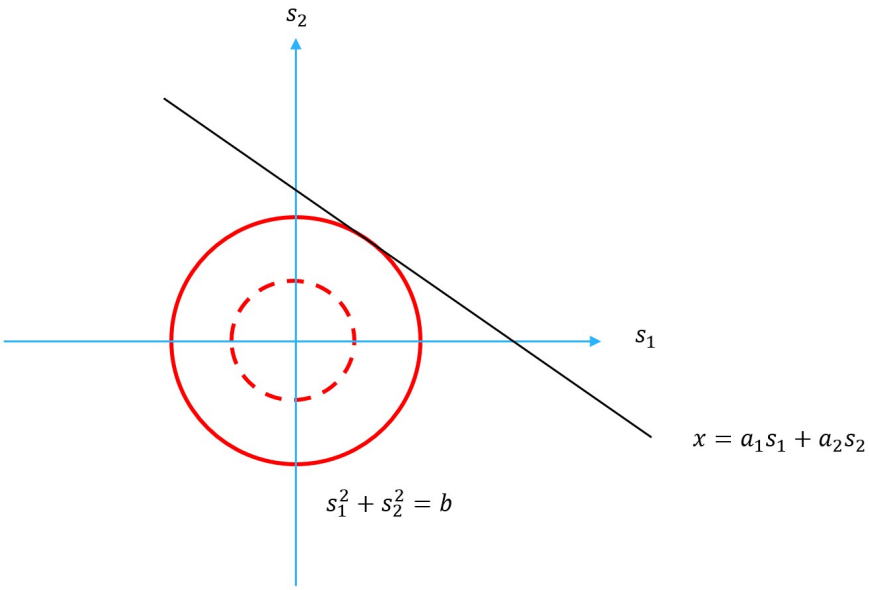

Figure 5.1: The solution of l_2 minimization is usually not sparse. All points on the black line satisfy the equation $x = a_1 s_1 + a_2 s_2$, and all points on the red circle (2-norm ball) satisfy the equation $s_1^2 + s_2^2 = b$. The b corresponding to the dotted circle is smaller, and the b for the solid circle is larger. The solid circle is exactly tangent to the black line.

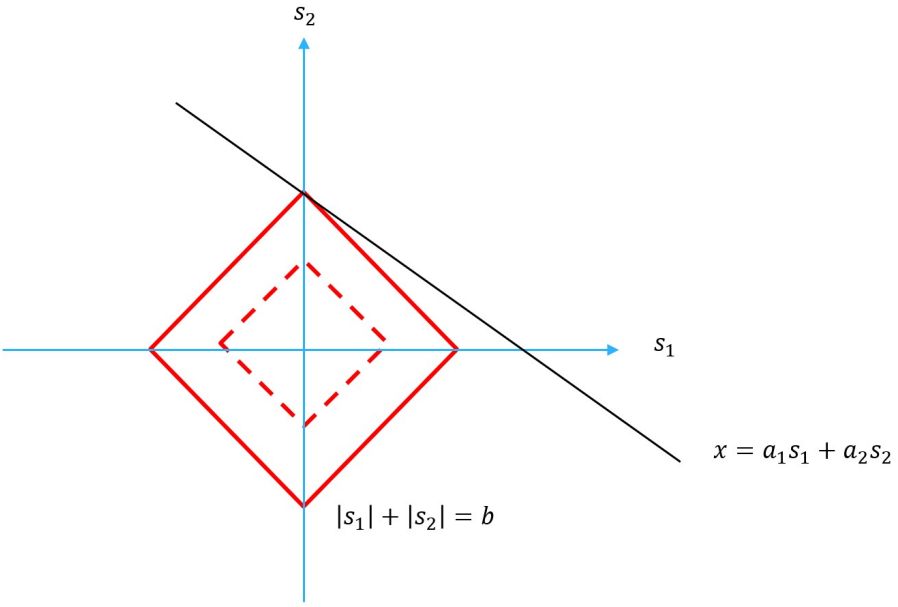

Figure 5.2: The solution of l_1 minimization is usually sparse. All points on the black line satisfy the equation $x = a_1 s_1 + a_2 s_2$, and all points on the red 1-norm ball satisfy the equation $|s_1| + |s_2| = b$. The b corresponding to the dotted 1-norm ball is smaller, and the b for the solid 1-norm ball is larger. The solid 1-norm ball intersects the black line at only one point on the axis.

through RIP of Gaussian matrix, (k, I, J) to scale proportionally. As a convex relaxation of the l_0 minimization, l_1 minimization needs more measurements to find the correct solution. Is there another convex relaxation for l_0 minimization, but with less measurement requirements than l_1 minimization or better performance with the same measurements? The answer is positive, and the better alternative is to "cleverly" put some weights to punish different coefficients in the l_1 norm, so-called weighted l_1 minimization.

Definition 5.15 (Weighted l_1 Minimization)**.** *Given a vector* $\mathbf{s} \in \mathbb{R}^{J\times 1}$ *and a set of positive weights* $\{w_1 > 0, w_2 > 0, \ldots, w_J > 0\}$*, the weighted* l_1 *norm is defined as:*

$$\|\mathbf{s}\|_{\mathbf{w},1} := \sum_{j=1}^{J} |w_j s_j|. \tag{5.1.35}$$

The corresponding weighted l_1 *minimization is defined as:*

$$\begin{aligned} \min_{\mathbf{s}} \quad & \|\mathbf{s}\|_{\mathbf{w},1} \\ s.t. \quad & \mathbf{x} = \mathbf{A}\mathbf{s}. \end{aligned} \tag{5.1.36}$$

When each entry of $\mathbf{s}$ has an equal weight, i.e., $w_1 = w_2 = \cdots = w_J$, the weighted l_1 minimization is equivalent to the l_1 minimization. Hence, weighted l_1 minimization is a generalization of l_1 minimization. With a smart weight design, weighted l_1 minimization is supposed to outperform l_1 minimization. Recap that the definition of the l_1 norm of a vector $\mathbf{s} \in \mathbb{R}^{J\times 1}$ is $\|\mathbf{s}\|_1 := \sum_{j=1}^{J} |s_j|$, and the l_0 norm is the number of non-zero coefficients in $\mathbf{s}$. Compared to the l_1 norm, l_0 norm is more "democratic" - the penalty for different coefficients (except 0) is the same. Hence a suitable weight design is to punish different coefficients in the l_1 norm to simulate this democracy of l_0, that is, a smaller weight for a larger coefficient and a larger weight for a smaller coefficient. An example is given in Fig 5.3 for illustration. In this example, we wish to recover $\mathbf{s_0}$ from measurements $\mathbf{x} = \mathbf{A}\mathbf{s_0}$. The red line is the set of all the feasible solutions. Among all the feasible solutions, there is one solution $\mathbf{s}' \neq \mathbf{s_0}$ with a smaller l_1 ball, i.e., $\|\mathbf{s}'\|_1 < \|\mathbf{s_0}\|$ (see Fig. 5.3 (b)). Hence, the solution of l_1 minimization is $\mathbf{s}'$ rather $\mathbf{s_0}$. If we consider a smaller weight on the direction of $\mathbf{s_0}$, then we will have a weighted l_1 ball ($\|\mathbf{s_0}\|_{\mathbf{w},1}$) in Fig. 5.3 (c). Since there is no feasible solution with a smaller weighted l_1 ball, the weighted l_1 minimization will find the correct solution.

The above example shows the superiority of weighted l_1 minimization over l_1 minimization. In practice, this requires us to have prior knowledge or estimation of $\|\mathbf{s_0}\|$.

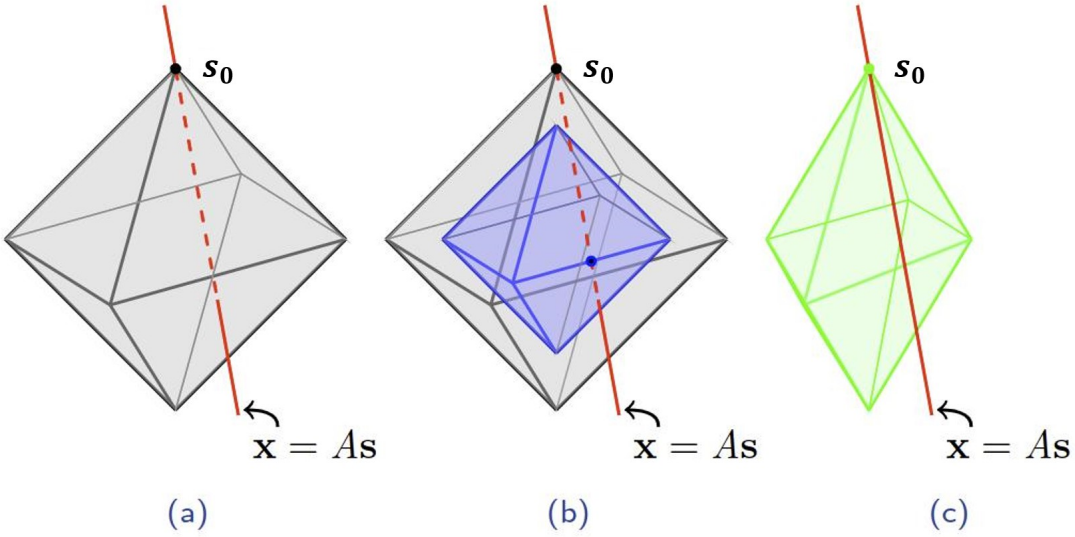

Figure 5.3: Weighted l_1 minimization to improve sparse signal recovery. (a) Sparse signal $\mathbf{s_0}$, feasible set $\mathbf{x} = \mathbf{As}$, and l_1 ball of radius $\|\mathbf{s_0}\|_1$. (b) There exists an $\mathbf{s}' \neq \mathbf{s_0}$ for which $\|\mathbf{s}'\|_1 < \|\mathbf{s_0}\|_1$. (c) Weighted l_1 ball. There exists no $\mathbf{s} \neq \|\mathbf{s_0}\|$ for which $\|\mathbf{s}\|_{\mathbf{w},1} \leq \|\mathbf{s_0}\|_{\mathbf{w},1}$

.

Candes and the rest [59] propose to compute the weights based on a simple iterative algorithm between estimating $\mathbf{s_0}$ and refining the weights. The algorithm is given below:

1. Set the iteration counter $l = 0$, maximum iteration number $l_{\max}$, and initialize the weights $w_1^{(0)} = w_2^{(0)} = \cdots = w_J^{(0)} = 1$.

2. Solve the weighted l_1 minimization

$$\begin{aligned} \min_{\mathbf{s}} \quad & \|\mathbf{s}\|_{\mathbf{w}^{(l)},1} \\ \text{s.t.} \quad & \mathbf{x} = \mathbf{As}. \end{aligned} \tag{5.1.37}$$

   Denote the solution by $\mathbf{s}^{(l)}$.

3. Update the weights: for each $j \in 1, 2, \cdots, J$,

$$w_j^{(l)} = \frac{1}{|s_j^{(l)}| + \epsilon}. \tag{5.1.38}$$

4. End the iteration if the number of iterations exceeds the maximum number of iterations, i,e, $l > l_{\max}$ or if the solution converges. Otherwise, increase the iteration number and return to the step 2.

The parameter ϵ in step 3 is introduced to keep the iteration process stable and allow a nonzero estimate in the next iteration if the current estimate of a component in $\mathbf{s}^{(l)}$ is

zero. According to the empirical demonstration in [59], ϵ should not be set too small or too large but the performance is quite robust to the choice of ϵ in a suitable range, around 10% of the standard deviation of the nonzero coefficients of $\mathbf{s_0}$.

5.1.3 l_1 minimization for speech separation

l_1 minimization is widely studied in the context of speech separation. Suppose we have I microphones, J speech sources, and $I < J$. Denote the observed signal of the i-th microphone by $x_i(t)$, the speech signal of the j-th source by $s_j(t)$. Their relationship is given by the convolutive model (Sec. 2.2):

$$x_i(t) = \sum_{j=1}^{J} (a_{ij} * s_j)(t), \quad 1 \leq i \leq I, \tag{5.1.39}$$

where a_{ij} is the room impulse response encoding the propagation process from the j-source to the i-th microphone, and $*$ is convolution. One may simplify the convolution model by applying the Short-time Fourier transform (STFT) (see Sec. 2.3.1):

$$x_i(t, f) = \sum_{j=1}^{J} a_{ij}(f) s_j(t, f), \quad 1 \leq i \leq I, \tag{5.1.40}$$

where $x_i(t, f)$ and $s_j(t, f)$ are STFT coefficients of $x_i(t)$ and $s_j(t)$ at the time frame t and frequency band f, respectively; $a_{ij}(f)$ is the Fourier transform of $a_{ij}(t)$, called the acoustic transfer function. When the number of microphones is less than the number of sources ($I < J$), even if the mixing parameters ($\{a_{ij}(t), \forall i, j\}$ or $\{a_{ij}(f), \forall i, j\}$) are given, reconstructing the speech signals $\{s_j(t), j = 1 \cdots, J\}$ or their STFT coefficients $\{s_j(t, f), j = 1 \cdots, J\}$ from the observed signals $\{x_i(t), i = 1 \cdots, I\}$ or their STFT coefficients $\{x_i(t, f), i = 1 \cdots, I\}$ is not a trivial problem. This is where the l_1 minimization comes in as l_1 minimization allows the reconstruction of sparse signals in the lack of measurements. Recap the spectrogram of speech signal in Fig 4.2, most of the STFT coefficients of a speech signal are zero or not significantly deviated from zero. In fact, for most of the time-frequency (TF) points, only a few sound sources have contributed. We demonstrate this claim through empirical experiments. Ignoring the TF points that no source contributes, if a source contributes more than 10% energy, we say the source is active for this TF point. We investigated ten groups of speech signals mixed by four sound sources, and record the percentage of TF points for different numbers of active sources in the Table 5.1. In this experiment, about half of the TF points were contributed more

Group index\number of active sources	1	2	3	4
1	44.88%	35.80%	15.72%	3.60%
2	41.77%	38.65%	16.36%	3.23%
3	47.99%	39.06%	11.63%	1.32%
4	47.74%	40.05%	11.20%	1.00%
5	40.71%	38.55%	17.18%	3.55%
6	41.72%	38.82%	16.13%	3.33%
7	47.05%	39.54%	12.06%	1.35%
8	44.76%	40.80%	13.15%	1.28%
9	49.52%	37.47%	11.69%	1.31%
10	51.71%	36.60%	10.60%	1.09%
Avg.	45.79%	38.53%	13.57%	2.11%

Table 5.1: The percentage of time-frequency point of different number of active sources. For a time-frequency point, if a source contributes more than 10% energy, this source is an active source for this time-frequency point.

than 70% of the energy by a single source, 85% of the TF points were contributed more than 80% of the energy by two sources. Hence, l_1 minimization can be used to reconstruct the speech sources when the number of microphones is inadequate.

For the sake of conciseness, , we introduce some notations and operators. Denote multichannel signals by a matrix $\mathbf{S} \in \mathbb{R}^{J\times T}$ where J is the number of channels and T is the number of samples. For any matrix $\mathbf{Z}$, $\|\mathbf{Z}\|_1$ is the summation of the entry-wise absolute values: $\|\mathbf{Z}\|_1 := \sum_{i,j} |z_{ij}|$. $\boldsymbol{\mathcal{A}} : \mathbb{R}^{J\times T} \to \mathbb{R}^{I\times T}$ is the convolution operator defined by

$$[\boldsymbol{\mathcal{A}}(\mathbf{S})]_{it} = \sum_{j=1}^{J} (a_{ij} * s_j)(t), \tag{5.1.41}$$

where $s_j(t)$ is the j-th channel signal in the $\mathbf{S}$. $\boldsymbol{\Psi} \in \mathbb{C}^{T\times B}$ is the discrete STFT matrix where $B = QF$, Q is the number of time frames, and F is the number of frequency bands. $\boldsymbol{\Psi}$ transforms multichannel signals $\mathbf{S} \in \mathbb{R}^{J\times T}$ into a matrix $\tilde{\mathbf{S}} \in \mathbb{C}^{J\times B}$ containing the STFT coefficients:

$$\tilde{\mathbf{S}} = \mathbf{S}\boldsymbol{\Psi}. \tag{5.1.42}$$

Denote the adjoint operator of $\boldsymbol{\Psi}$ by $\boldsymbol{\Psi}^*$. $\boldsymbol{\Psi}^* \in \mathbb{C}^{B\times T}$ is actually the inverse discrete STFT matrix, thus the time domain signals can be obtained from the STFT coefficients by

$$\mathbf{S} = \tilde{\mathbf{S}}\boldsymbol{\Psi}^*. \tag{5.1.43}$$

Given mixed signals $\mathbf{X} \in \mathbb{R}^{I\times T}$ and the mixing process $\boldsymbol{\mathcal{A}} : \mathbb{R}^{J\times T} \to \mathbb{R}^{I\times T}$, under the time

domain model (5.1.39), the source signals $\mathbf{S_0} \in \mathbf{R}^{J\times T}$ can be estimated by the solution of the following optimization problem:

$$\begin{aligned} \min_{\mathbf{S}} \quad & \|\mathbf{S}\mathbf{\Psi}\|_1 \\ \text{s.t.} \quad & \mathbf{X} = \boldsymbol{\mathcal{A}}(\mathbf{S}). \end{aligned} \tag{5.1.44}$$

Considering the weighted l_1 minimization and the measurement noise on $\mathbf{X}$, the optimization problem (5.1.44) can be generalized as

$$\begin{aligned} \min_{\mathbf{S}} \quad & \|\mathbf{S}\mathbf{\Psi}\|_{\mathbf{W},1} \\ \text{s.t.} \quad & \|\mathbf{X} - \boldsymbol{\mathcal{A}}(\mathbf{S})\|_F \leq \epsilon, \end{aligned} \tag{5.1.45}$$

where $\mathbf{W} \in \mathbb{R}^{J\times B}$ is a matrix with positive entries w_{ij}, $\|\mathbf{Z}\|_{\mathbf{W},1} := \sum_{ij} w_{ij}|z_{ij}|$ is weighted l_1 norm, and ϵ is the upper bound of the l_2 norm of the measurement noise. Assuming i.i.d Gaussian noise with zero mean and variance σ_e^2, $\frac{1}{\sigma_e^2}\|\mathbf{X} - \boldsymbol{\mathcal{A}}(\mathbf{S})\|_F^2$ follows a χ^2 distribution with IT degrees of freedom. Thus set $\epsilon^2 = (IT + 2\sqrt{2IT})\sigma_e^2$. The choice of the ϵ provides a likely bound for $\|\mathbf{X} - \boldsymbol{\mathcal{A}}(\mathbf{S})\|_F$, since the probability that the l_2 norm term exceeds ϵ is the probability that a χ^2 random variable with IT degrees of freedom exceeds its mean IT by at least two times the standard deviation $2\sqrt{2IT}$, which is very small. When there is no prior on the noise nor design on the weight, we may set $\epsilon = 0$ and $w_{ij} = 1$, $\forall i, j$. Then the optimization problem (5.1.45) becomes (5.1.44).

Another l_1 formulation is under the time-frequency domain model (5.1.39), rather estimating the source signals directly, we can estimate their STFT coefficients:

$$\begin{aligned} \min_{\tilde{\mathbf{S}}} \quad & \|\tilde{\mathbf{S}}\|_{\mathbf{W},1} \\ \text{s.t.} \quad & \|\tilde{\mathbf{X}} - \tilde{\boldsymbol{\mathcal{A}}}(\tilde{\mathbf{S}})\|_F \leq \tilde{\epsilon}, \end{aligned} \tag{5.1.46}$$

where $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{\Psi}$ is the matrix containing the STFT coefficients of mixed signals, $\tilde{\mathbf{S}}$ is the estimated STFT coefficients of source signals, $\tilde{\boldsymbol{\mathcal{A}}} : \mathbb{C}^{J\times B} \to \mathbb{C}^{I\times B}$ is the multiplication operator defined by the equation (5.1.40), and $\tilde{\epsilon}$ is the upper bound of the l_2 norm of noise. Assuming i.i.d complex Gaussian noise with variance $\sigma_{\tilde{e}}^2$, the l_2 norm term $\frac{2}{\sigma_{\tilde{e}}^2}\|\tilde{\mathbf{X}} - \tilde{\boldsymbol{\mathcal{A}}}(\tilde{\mathbf{S}})\|_F^2$ follows a χ^2 distribution with $2IB$ degrees of freedom. Thus set $\tilde{\epsilon}^2 = (IB + 2\sqrt{IB})\sigma_{\tilde{e}}^2$ according to the same argument of the setting of ϵ in (5.1.45). Compared to the time domain optimization problem (5.1.45), (5.1.46) has a more practical meaning because there is enough research work on estimating $\tilde{\boldsymbol{\mathcal{A}}}$, but for $\boldsymbol{\mathcal{A}}$, as far as we know, there is almost no related work that can estimate $\boldsymbol{\mathcal{A}}$ in a multi-source and reverberated

environment. However, in the presence of high reverberation, the solution to (5.1.46) is not a good estimate but the solution to (5.1.45) is. For the reason please refer to Sec. 2.4. The work in this paper does not involve the estimation of $\boldsymbol{\mathcal{A}}$, but aims to propose a new objective for sparse optimization-based speech separation methods.

5.2 Improved Method

In this work, we introduce a new objective to pursue the sparsity of speech signals, a weighted l_1 norm with a novel weighting design, called adaptive weighting. We find out how to roughly estimate the STFT coefficients of the source signal and design the weights accordingly to penalize small coefficients more heavily and penalize less heavily for large coefficients. By doing so, the weighted l_1 minimization is closer to l_0 minimization than the normal l_1 minimization while can be solved using the same solver of l_1 minimization with the same computational complexity (including the computation of the weights). Three major contributions are made through this work: 1) The proposed weighted l_1 minimization has a better reconstruction performance than l_1 minimization, up to 5 dB improvement in SDR. They have the same theoretical time complexity but in practice, the weighted l_1 minimization converges faster, up to 35% faster. 2) Further improvement can be achieved by combining the reweighted scheme [21, 59]. 3) The improvement is robust to the estimation error of the mixing process in the time domain, i.e., $\boldsymbol{\mathcal{A}}$ or in the STFT domain, i.e., $\tilde{\boldsymbol{\mathcal{A}}}$.

5.2.1 Clustering phenomenon

The key question is how to make a quick estimate of the STFT coefficients of source signals? Let us have some inspiration from an observation. In this experiment, we have two microphones and three sources. Denote the STFT coefficients of mixed signals by $\mathbf{x}(t,f) = [x_1(t,f), x_2(t,f)]^{\mathsf{T}}$ and the acoustic transfer functions (ATFs) from the j-th source to the two microphones by $\mathbf{a}_j(f) = [a_{1j}(f), a_{2j}(f)]^{\mathsf{T}}$. We do an amplitude and phase normalization as following:

$$\mathbf{x} \leftarrow \frac{\mathbf{x}}{\|\mathbf{x}\|_2} e^{-\theta(x_1)}, \tag{5.2.1}$$

where we omit (t,f) for the sake of conciseness, and $\theta(x_1)$ is the angle of the complex number x_1. The same normalization is also applied on the ATFs $\mathbf{a}(f)$. The amplitude and

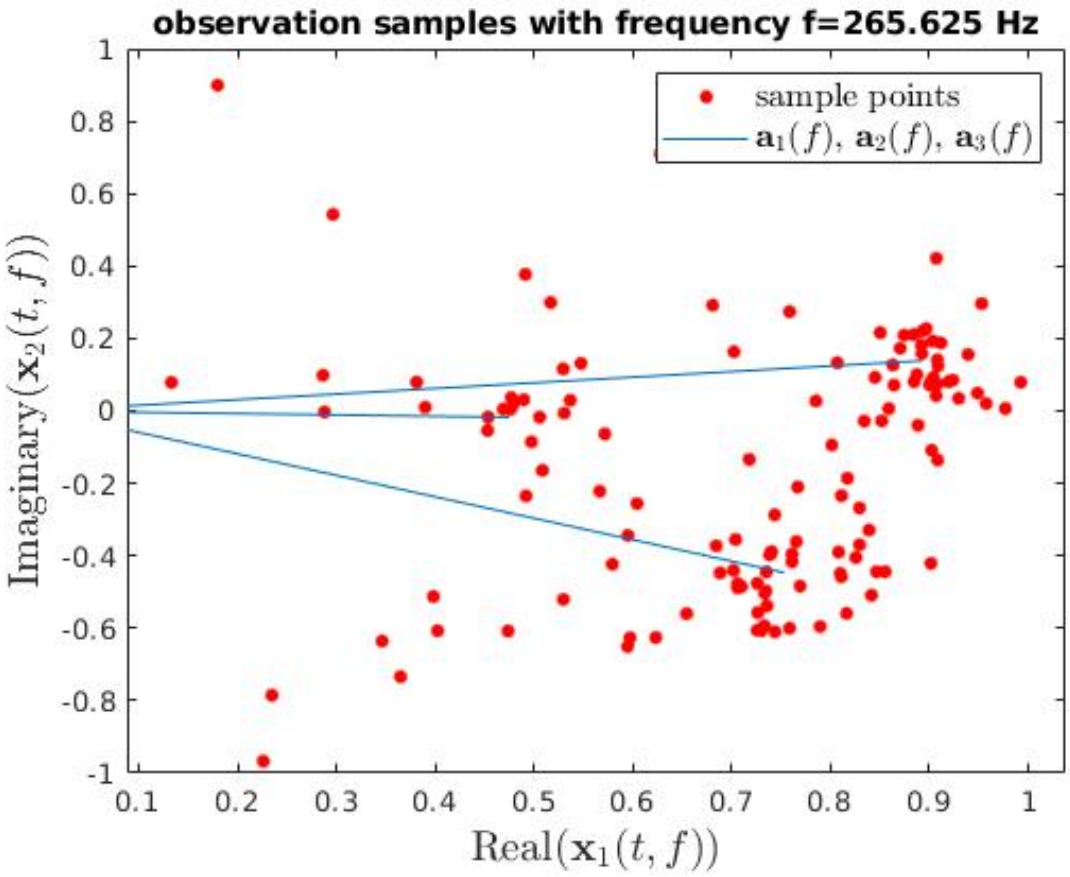

Figure 5.4: Clustering phenomenon of the STFT coefficients of mixed signals. We have two microphones and three sources. The x-axis is the real value of the STFT coefficients of the mixed/observed signal of the first microphone, and the y-axis is the imaginary value of the STFT coefficients of the mixed signal of the second microphone. The red points are sample points evaluated at frequency $f = 265.625$ Hz, and the blue lines are mixing vectors containing mixing parameters corresponding to the three sources, respectively.

phase normalization (5.2.1) is important, otherwise we can not have the following observation. Then we draw a scatter plot of the normalized $\mathbf{x}(t, f)$ and draw the three vectors $\{\mathbf{a}_1(f), \mathbf{a}_2(f), \mathbf{a}_3(f)\}$ on it. Since all the numbers are complex numbers, for the sake of visualization, the plot only considers the real value of the first entry and the imaginary value of the second entry. Figure 5.4 shows the distribution of $\mathbf{x}(t, f)$ at $f = 265.625$ Hz. Obviously, the samples assemble around $\mathbf{a}_j(f),\ j = 1, 2, 3$. This clustering phenomenon echoes the fact that a large number of time-frequency points are majorly contributed by a single source which is consistent with the table 5.1. The normalization (5.2.1) is to eliminate the phase and amplitude differences caused by the sound source at different times.

5.2.2 Short-time spectral amplitude estimator

Inspired by the Fig. 5.4, we believe that for each time-frequency point, the amplitude of the j-th source's STFT coefficient $|s_j(t,f)|$ can be roughly estimated by the amplitude of the projection of $\mathbf{x}(t,f)$ on the $\mathbf{a}_j(f)$. We only focus on the amplitude estimator since the weights for the weighted l_1 minimization are positive numbers. The projection is defined in the Euclidean vector space, so vectors are real number vectors, which helps us build geometric intuitions. For the sake of conciseness, new notations are introduced first:

$$\begin{aligned}\mathbf{h} &:= \begin{pmatrix}\Re(\mathbf{x})\\ \Im(\mathbf{x})\end{pmatrix} \in \mathbb{R}^{2I}, \quad \mathbf{u}_j := \begin{pmatrix}\Re(\mathbf{a}_j)\\ \Im(\mathbf{a}_j)\end{pmatrix} \in \mathbb{R}^{2I},\\ \mathbf{v}_j &:= \begin{pmatrix}-\Im(\mathbf{a}_j)\\ \Re(\mathbf{a}_j)\end{pmatrix} \in \mathbb{R}^{2I}, \quad j = 1,\dots,J,\end{aligned} \tag{5.2.2}$$

where $\Re\{\cdot\}$ and $\Im\{\cdot\}$ are the real and the imaginary part of a complex number, respectively, and $\mathbf{x}$ and $\mathbf{a}_j$ are the STFT coefficients of mixed signal and ATFs of the j-th sources, respectively (the time and frequency indices are omitted here for compactness). Remark that, for all j, $\mathbf{u}_j$ and $\mathbf{v}_j$ are orthogonal to each other and have the same amplitude, i.e., $\mathbf{u}_j \perp \mathbf{v}_j$ and $\|\mathbf{u}_j\|_2 = \|\mathbf{v}_j\|_2$, $\forall j \in \{1,\cdots,J\}$.

Rewrite the time-frequency multiplication model (5.1.40) using the real vector notations, we have

$$\mathbf{h} = \sum_{j=1}^{J} \Big(\mathbf{u}_j \Re(s_j) + \mathbf{v}_j \Im(s_j)\Big), \tag{5.2.3}$$

where s_j is the j-th source's STFT coefficient. As can be seen from (5.2.3), $\mathbf{h}$ is a summation of J vectors in the J subspaces spanned by the J orthogonal bases $\{\mathbf{u}_1,\mathbf{v}_1\},\dots,\{\mathbf{u}_J,\mathbf{v}_J\}$, respectively. Denote by $\mathbf{p}_j$ the projection of $\mathbf{h}$ to the subspace spanned by $\mathbf{u}_j$ and $\mathbf{v}_j$, the amplitude of $|s_j| = \sqrt{\Re^2(s_j) + \Im^2(s_j)}$ can be estimated by

$$\begin{aligned}\hat{c}_j &= C \cdot \|\mathbf{p}_j\|_2 / \sqrt{\|\mathbf{u}_j\|_2 \|\mathbf{v}_j\|_2}\\ &\propto \sqrt{(\mathbf{h}^{\mathsf{T}}\mathbf{u}_j)^2 + (\mathbf{h}^{\mathsf{T}}\mathbf{v}_j)^2} / (\|\mathbf{u}_j\|_2 \|\mathbf{v}_j\|_2)\\ &= |\mathbf{a}_j^{\mathsf{H}}\mathbf{x}| / (\mathbf{a}_j^{\mathsf{H}}\mathbf{a}_j),\end{aligned} \tag{5.2.4}$$

where C is a normalization factor which is independent of j. The denominator in (5.2.4) is to eliminate the contribution of the basis vectors. To evaluate the power of this short-time spectral amplitude estimator, we conduct an experiment where there are two microphones and three source signals. The distance between two microphones is one meter, and there

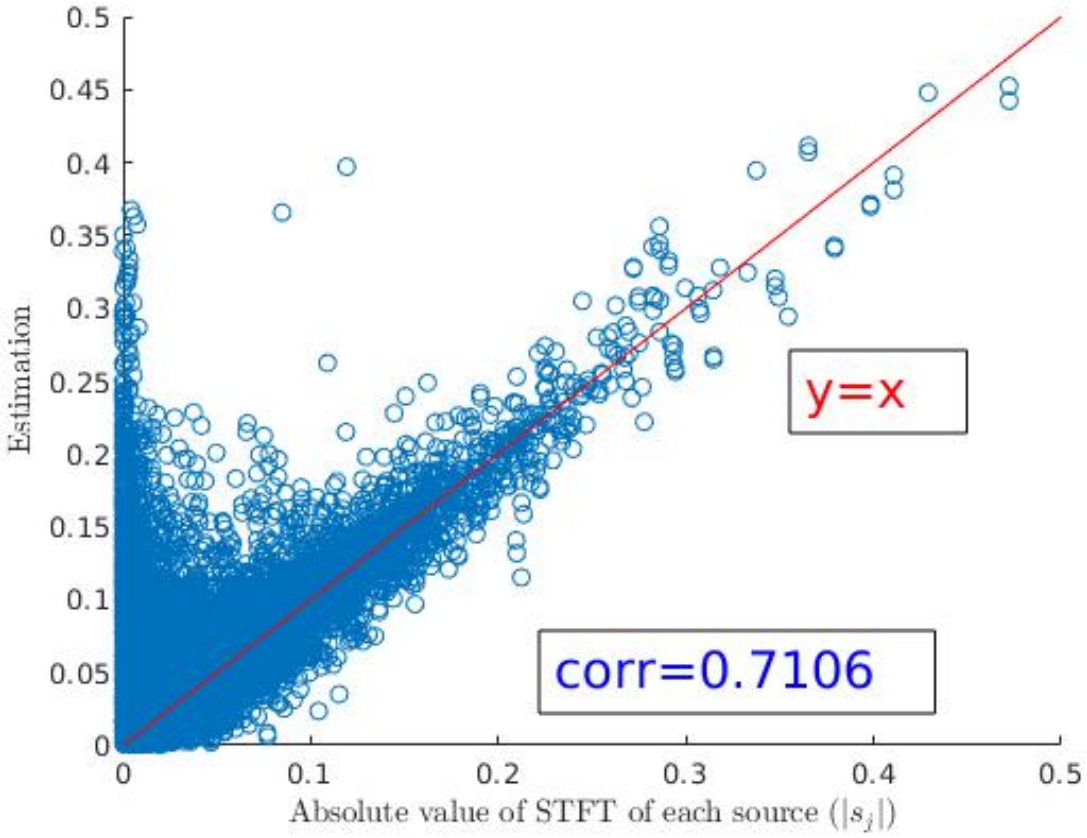

Figure 5.5: Source signal estimation by the short-time spectral amplitude estimator (5.2.4). The x-axis is the ground truth of the amplitude of the STFT coefficient of each source ($|s_j|$); The y-axis is the corresponding estimation by (5.2.4). The red line ($y = x$) is for reference.

is no reverberation in the environment. The experiment involves ten groups of mixed signals (each group of mixed signals is from different source signals) and the result is shown in Fig. 5.5. In general, the estimation is not bad, but there are some samples' amplitude that is over-estimated. More specifically, there are some samples' amplitude that is close to zero, but whose estimate is significantly larger than zero. This is because the J orthogonal bases $\{\mathbf{u}_1, \mathbf{v}_1\}, \cdots .\{\mathbf{u}_J, \mathbf{v}_J\}$ are not orthogonal to each other. However, the estimation is only used as a pre-estimation to calculate the weight calculation, and the final estimation will be refined by the weighted l_1 minimization. Hence, it does not matter that there are parts of overestimation for the pre-estimation.

On the calculation of weight, since for a small $|s_j|$, we should put a larger weight, one could calculate the weight by

$$w_j = \frac{1}{\hat{c}_j + \epsilon}, \tag{5.2.5}$$

where ϵ is introduced to keep stability and to avoid infinity weight if $\hat{c}_j = 0$. However, we find out that the weighted l_1 minimization based on the weights calculated by (5.2.5) does not have a good performance when the distance of the microphone is too close, e.g., 5 cm. This is may because we did not set the ϵ well. We consider an alternative to calculating the weight:

$$\begin{aligned} w_j &= \|\mathbf{a}_j\|_2 \sqrt{1 - (\frac{\hat{c}_j \|\mathbf{a}_j\|_2}{\|\mathbf{x}\|_2})^2 + \epsilon} \\ &= \|\mathbf{a}_j\|_2 \sqrt{1 - (\frac{\mathbf{a}_j^{\mathsf{H}} \mathbf{x}}{\|\mathbf{a}_j\|_2 \|\mathbf{x}\|_2})^2 + \epsilon}, \end{aligned} \tag{5.2.6}$$

where ϵ is an arbitrary small number to prevent $w_j = 0$. Remark that w_j in (5.2.6) is actually multiplication of $\|\mathbf{a}_j\|_2$ and the sine of the angle between $\mathbf{x}$ and $\mathbf{a}_j$, where $\|\mathbf{a}_j\|_2$ is to eliminate the effect that $\mathbf{x}$ is offset in the direction of $\mathbf{a}_j$ due to the magnitude of $\mathbf{a}_j$.

5.2.3 Algorithm

Our proposed weighted l_1 minimization, called adaptive weighted l_1 minimization, can be obtained by substituting the weights (5.2.6) in (5.1.46) or (5.1.45). In addition, the weights (5.2.6) can be the initialization weights in the reweighted scheme introduced in Sec. 5.1.2, and a more general problem, called the adaptive reweighted l1 minimization, can be formalized. Predecessors [21] have already proposed an algorithm to estimate speech signals using the reweighted l_1 minimization, here we give a modified algorithm (Algorithm 2) according to the proposed weights. The parameters and notations in the algorithm are introduced below. l is the iteration counter, ρ is our defined difference between two successive solutions to check the convergence of the reweighted process, and $l_{\max}$ and μ are the maximum number of iteration and threshold for convergence, respectively. $B = QF$ where Q is the number of time frames and F is the number of frequency bins (We reshape the two dimensions, time and frequency, in one dimension for convenience). $\triangle(\mathbf{X}, \boldsymbol{\mathcal{A}}, \mathbf{W}^{(l)}, \epsilon)$ is the solution of the weighted l_1 minimization (5.1.45). $\delta^{(l)}$ is a parameter to keep the stability of each weighting process and will decrease with a rate β along with the process until a threshold $\delta_{\tilde{s}}$. $\delta_{\tilde{s}}$ is the standard deviation of the noise in the STFT domain and is computed as $\delta_{\tilde{s}} = \delta_e \sqrt{IT/2JB}$, and $\epsilon^2 = (IT + 2\sqrt{2IT})\sigma_e^2$. For any matrix $\mathbf{Z}$, $\|\mathbf{Z}\|_F := \sqrt{(\sum_{i,j} Z_{ij}^2)}$ is the Frobenius norm. In Algorithm 2, we estimate the source signals by iteratively solving the time domain weighted l_1 minimization

Algorithm 2: Adaptive reweighted l_1 minimization for source estimation

Input: $\mathbf{X}$, $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\Psi}$, ϵ

Initialize: $l = 1$, $\rho = 1$, compute $W_{ij}^{(0)}$ according to (5.2.6), for $i = 1, \cdots, J$, $j = 1, \cdots, B$;

Solve the weighted l_1 minimization (5.1.45): $\mathbf{S}^{(0)} = \triangle(\mathbf{X}, \boldsymbol{\mathcal{A}}, \mathbf{W}^{(0)}, \epsilon)$;

$\delta^{(0)} := \text{std}(\mathbf{S}^{(\mathbf{0})}\boldsymbol{\Psi})$;

while $\rho > \eta$ *and* $l < l_{max}$ **do**

  Update the weight matrix: $W_{ij}^{(l)} = \delta^{(l-1)}/(\delta^{(l-1)} + |\tilde{S}_{ij}^{(l-1)}|)$, for $i = 1, \cdots, J$, $j = 1, \cdots, B$ with $\tilde{\mathbf{S}} = \mathbf{S}\boldsymbol{\Psi}$;

  Solve the weighted l_1 minimization (5.1.45): $\mathbf{S}^{(l)} = \triangle(\mathbf{X}, \boldsymbol{\mathcal{A}}, \mathbf{W}^{(l)}, \epsilon)$;

  Update $\delta^{(l)} := \max(\beta\delta^{(l-1)}, \delta_{\tilde{s}})$;

  Update $\rho := \|\mathbf{S}^{(l)} - \mathbf{S}^{(l-1)}\|_F / \|\mathbf{S}^{(l-1)}\|_F$;

  $l = l + 1$;

end

Return $\mathbf{S}^{(l-1)}$

problem (5.1.45), and the reader can estimate the STFT coefficients of source signals, i.e., $\tilde{\mathbf{S}}$, by iteratively solving the time-frequency domain weighted l_1 minimization problem (5.1.46) and the source signals can be obtained by applying the inverse STFT. This is just a trivial modification of the Algorithm (5.1.45). It is worth noting some minor changes that $\delta^{(0)} := \text{std}(\tilde{\mathbf{S}})$, $\delta_{\tilde{s}} = \delta_{\tilde{e}}\sqrt{I/J}$, and $\tilde{\epsilon}^2 = (IB + 2\sqrt{IB})\delta_{\tilde{e}}^2$. In the experiment, we set $l_{\max} = 200$ (the algorithm is usually convergent around $l = 10$), $\eta = 0.001$, $\beta = 0.1$, and $\epsilon = 1\text{e}{-4}$. Note that we only consider the noise-free case, thus ideally ϵ should be set to 0 but it will take infinite iterations for solving the weighted l_1 minimization, so we set a small tolerance.

5.2.4 Weighted l_1 minimization solver

In this section, we will introduce how to solve the weighted l_1 minimization by applying the proximal splitting methods. The proximal splitting methods solve the optimization problem of the form

$$\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x}), \tag{5.2.7}$$

where f and g are convex function. The proximal splitting methods do not require f and g differentiable: instead of calculating the gradient, they calculate the proximal point. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a closed, convex, and proper (CCP) function. The proximal operator of f is defined as:

$$\text{prox}_f(\mathbf{z}) = \arg\min_{\mathbf{x}}(f(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2). \tag{5.2.8}$$

An example is when $f(x) = |x|$, then $\text{prox}_f(z) = \arg\min_x(|x| + \frac{1}{2}(x-z)^2)$, which is the soft-thresholding function:

$$\text{prox}_f(z) = \begin{cases} z-1, & z > 1 \\ 0, & -1 \le z \le 1 \\ z+1 & z < -1. \end{cases} \tag{5.2.9}$$

A fixed point of prox_f is $\mathbf{z}$ such that $\text{prox}_f(\mathbf{z}) = \mathbf{z}$, and this point is also the minimizer of f, i.e., $\{z \,|\, \text{prox}_f(\mathbf{z}) = \mathbf{z}\} = \{\arg\min_{\mathbf{x}} f(\mathbf{x})\}$. Another property of the proximal operator is that it is firmly non-expansive, that is, $\|\text{prox}_f(\mathbf{x}) - \text{prox}_g(\mathbf{y})\|_2^2 \le \left\langle \mathbf{x}-\mathbf{y} \,|\, \text{prox}_f(\mathbf{x}) - \text{prox}_g(\mathbf{y}) \right\rangle$. The properties imply that one can approach the minimizer of f by iteratively applying prox_f. This is how we find the minimizer of f without calculating its gradient if we know $\text{prox}_f(\mathbf{y})$ for every $\mathbf{y}$ analytically or numerically. The Douglas-Rachford (DR) algorithm [70] is an optimization algorithm based on the proximal operator to solve the optimization problem (5.2.7). In experiment, we set $l_{\max} = 200$, $\eta_{\text{dr}} = 0.01$, $\alpha = 1$, and $\gamma = 0.1$.

Algorithm 3: The Douglas-Rachford algorithm

```
Initialize: l = 0, ρ = 1, z, α ∈ (0, 2), γ > 0;
while ρ > η_dr and l < l_max do
    x^(l) = prox_{γg}(z^(l));
    z^(l+1) = z^(l) + α(prox_{γf}(2s^(l) − z^(l)) − s^(l));
    Update ρ := ‖x^(l) − x^(l−1)‖/‖x^(l−1)‖;
    l = l + 1;
end
Return x^(l−1)
```

We convert the constrained problem (5.1.45) to an unconstrained problem by introducing an indicator function $g_\epsilon : \mathbb{R}^{J\times T} \to \mathbb{R}$ defined as

$$g_\epsilon(\mathbf{S}) = \begin{cases} 0, & \|\mathbf{X} - \boldsymbol{\mathcal{A}}(\mathbf{S})\|_F \le \epsilon \\ +\infty, & \text{otherwise.} \end{cases} \tag{5.2.10}$$

The constrained problem (5.1.45) can be rewritten as

$$\min_{\mathbf{S}} \|\mathbf{S}\boldsymbol{\Psi}\|_{\mathbf{W},1} + g_\epsilon(\mathbf{S}), \tag{5.2.11}$$

which can be solved by the DR algorithm. The proximal operators of the two functions in (5.2.11) can be obtained by applying the relevant theorems of the proximal operator

under mapping of frame [70, 71]. The STFT transform $\mathbf{\Psi}$ is a tight frame whose the adjoint frame $\mathbf{\Psi}^*$ is the inverse STFT, and $\boldsymbol{\mathcal{A}}$ is a general frame whose the adjoint frame $\boldsymbol{\mathcal{A}}^* : \mathbb{R}^{I\times T} \to \mathbb{R}^{J\times T}$ is the convolution operator defined by

$$[\boldsymbol{\mathcal{A}}^*(\mathbf{X})]_{jt} = \sum_{i=1}^{I}(a_{ji}^* * x_i)(t), \tag{5.2.12}$$

where $a_{ji}^*(t) := a_{ij}(-t)$.

The proximal operator of the weighted l_1 norm is

$$\text{prox}_{\|\cdot\mathbf{\Psi}\|_{\mathbf{W},1}}(\mathbf{Z}) = \mathbf{Z} + (\text{prox}_{\|\cdot\|_{\mathbf{W},1}} - \mathbf{I})(\mathbf{Z}\mathbf{\Psi})\mathbf{\Psi}^*, \tag{5.2.13}$$

where $\text{prox}_{\|\cdot\|_{\mathbf{W},1}}$ is the element-wise soft-thresholding function given by

$$\begin{aligned}\left[\text{prox}_{\|\cdot\|_{\mathbf{W},1}}(\mathbf{Z})\right]_{ij} &= \text{prox}_{w_{ij}\|\cdot\|}(z_{ij}) \\ &= \frac{z_{ij}}{|z_{ij}|}\max(0, |z_{ij}| - w_{ij}).\end{aligned} \tag{5.2.14}$$

The proximal operator of g_ϵ can be iteratively approached by

$$\begin{aligned}\mathbf{U}^{(l+1)} &= \mu_l(\mathbf{I} - \text{prox}_{\|\cdot\|_F\leq\epsilon})(\mu_l^{-1}\mathbf{U}^{(l)} + \boldsymbol{\mathcal{A}}(\mathbf{P}^{(l)}) - \mathbf{X}) \\ \mathbf{P}^{(l+1)} &= \mathbf{Z} - \boldsymbol{\mathcal{A}}^*(\mathbf{U}^{(l+1)}),\end{aligned} \tag{5.2.15}$$

where $\mu_l \in (0, 2/v)$, v is the upper bound of frame $\boldsymbol{\mathcal{A}}$, and

$$\text{prox}_{\|\cdot\|_F\leq\epsilon}(\mathbf{U}) = \min(1, \frac{\epsilon}{\|\mathbf{U}\|_F})\mathbf{U}. \tag{5.2.16}$$

Then $\mathbf{P^{(l+1)}}$ linearly approaches $\text{prox}_{g_\epsilon}(\mathbf{Z})$. The upper bound of $\boldsymbol{\mathcal{A}}$ is defined by a positive number v such that $\|\boldsymbol{\mathcal{A}}(\mathbf{S})\|_F^2 \leq v\|\mathbf{S}\|_F$ satisfies for all $\mathbf{S}$. The tightest possible upper bound is the operator norm of $(\boldsymbol{\mathcal{A}}^*\boldsymbol{\mathcal{A}})$, which can be computed by the well-known power iteration algorithm.

Algorithm 4: The power iteration algorithm for the calculation of v

```
Initialize: V ∈ R^{J×T};
repeat
    W = A*(A(V));
    v = ||W||_∞;
    V = W/v;
until convergence;
```

5.3 Experiment

We convolve the room impulse responses (RIRs) with the source signal and add them together to get the mixed signals whose duration are four seconds. The RIRs were generated by the software *pyroomacoustics* [2], with a room size of dimension $3.55\,m \times 4.45\,m \times 2.5\,m (l \times w \times h)$. The number of microphones was two ($I = 2$). The number of sources was in a range of three to six ($3 \leq J \leq 6$). The reverberation is set to be 0 ms (no reverberation) and 250 ms, and the distance between two microphones are set to be 5 cm and 1 m. No artificial noise was added since the scope is to evaluate the source reconstruction from mixtures rather than denoising. The source-to-distortion ratio (SDR) is used as an objective evaluation. A larger SDR indicates a better reconstruction result. We evaluate the reconstruction ability of four optimizations: l_1 minimization, reweighted l_1 minimization [21], the adaptive weighted l_1 minimization, and the adaptive reweighted l_1 minimization. The first two are benchmarks and the last two are our proposed optimizations with the novel objective. The four optimization problems are solved by the weighted l_1 minimization solver we introduced in the Sec. 5.2.4.

5.4 Experiment Result

5.4.1 Impact of the STFT window length

Different lengths of windows will have different resolutions in the time and frequency. For the same signal, a longer window will result in a smaller number of time frames and more frequency bins. Consequently, the time frame will cover a longer time period, and frequency will be more subdivided. Therefore, a long STFT window will have a low resolution in the time and a high resolution in the frequency, and vice visa. Correspondingly, the sparsity of the STFT representation of the signals will vary. The two extreme cases are there is no resolution in time, i.e., spectrum, or no resolution in frequency, i.e., waveform. Since both the extreme cases will reduce the sparsity, there may exist an optimal window length between them that results in the most sparse. To find the optimal window length, in this experiment, we evaluated the effect of the STFT window length on the separation performance. As shown in Fig. 5.6, regardless of the presence or absence of reverberation, regardless of which way to pursue sparseness, the best window length is 64

ms (1024 sample points as the sampling rate is 16 kHz). The worse performance on the shortest window length, i.e., 8 ms, and the longest window length, i.e., 256 ms, is also in line with our expectations. There is another "byproduct" in this experiment. In the presence of reverberation, the adaptive weighted l_1 minimization (our proposed) shows a poorer performance than the l_1 minimization when the window length is short (less than 16 ms) and a better performance when the window length is long (longer than 64 ms). However, in the absence of reverberation, the adaptive weighted l_1 minimization always outperforms the l_1 minimization. This is reasonable since the weights are calculated based on the time-frequency domain multiplication model which is only appropriate when the window length is comparable to the length of reverberation.

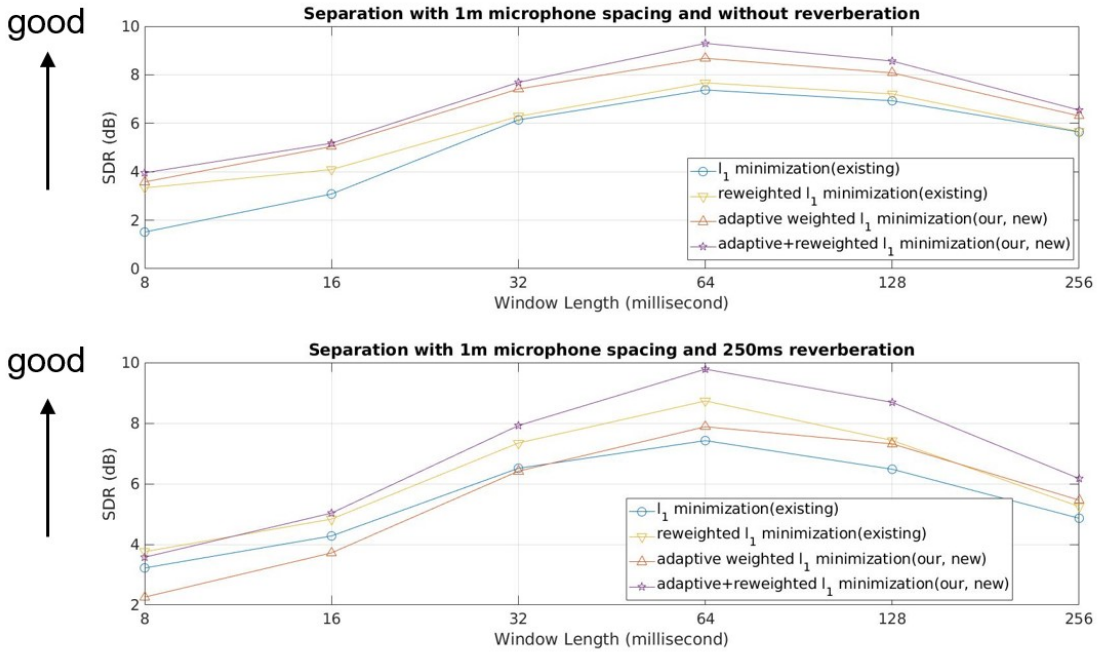

Figure 5.6: Average SDR as a function of the STFT window length. There are two microphones with a one-meter spacing and four speech sources. (a) Without reverberation. (b) In the presence of $RT_{60} = 250$ ms reverberation.

5.4.2 Impact of number of sources

In this experiment, we mainly evaluate the separation performance in the different number of sources. In addition, we evaluate the impact of different distances between two microphones. The window length is set to be 64 ms, which is the best window length for the performance from the last experiment. The results are shown in Fig. 5.7 for 1

m microphone spacing and Fig. 5.8 for 5 cm microphone spacing. In general, the performance degrades linearly as the number of sources increases. The adaptive reweighted l_1 minimization (the proposed) always outperforms the other methods, and the adaptive weighted l_1 minimization always outperforms the l_1 minimization. The superiority of the proposed algorithms is more significant when the microphone spacing is 5 cm and there is no reverberation, at least a 5 dB improvement. It is worth noting that in this case, i.e., 5 cm microphone spacing and no reverberation, the reweighted l_1 minimization performs abnormally worse than the l_1 minimization. The same result also appears in [21]. Finding out why is still an open question for this thesis.

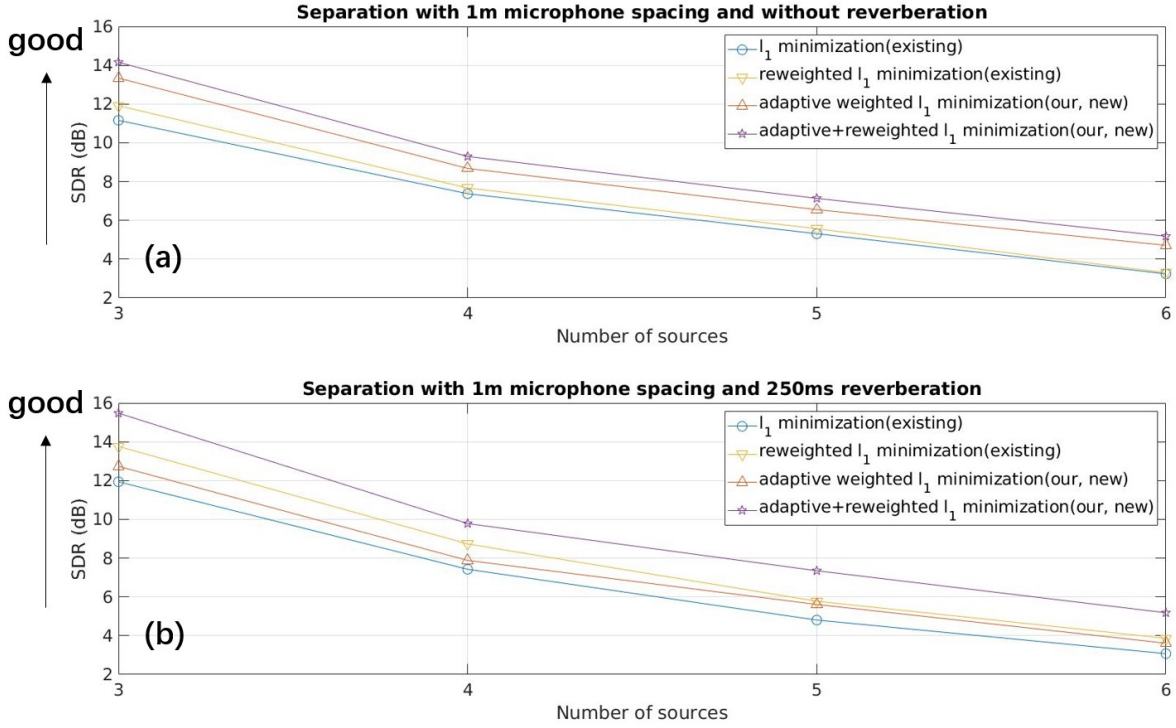

Figure 5.7: Average SDR as a function of the number of sources. There are two microphones with a one-meter spacing. (a) Without reverberation. (b) In the presence of $RT_{60} = 250$ ms reverberation.

5.4.3 Robustness

In the practice, we only know the mixed signals observed by microphones and don't know the mixing process, i.e., $\boldsymbol{\mathcal{A}}$ or $\tilde{\boldsymbol{\mathcal{A}}}$. Estimating the mixing process is another topic and is out of the focus of this work. However, it is supposed to investigate the robustness of the proposed algorithms to the estimating error of the mixing process. We manually add different degrees of Gaussian white noise on the mixing parameters in $\boldsymbol{\mathcal{A}}$. For the mixing

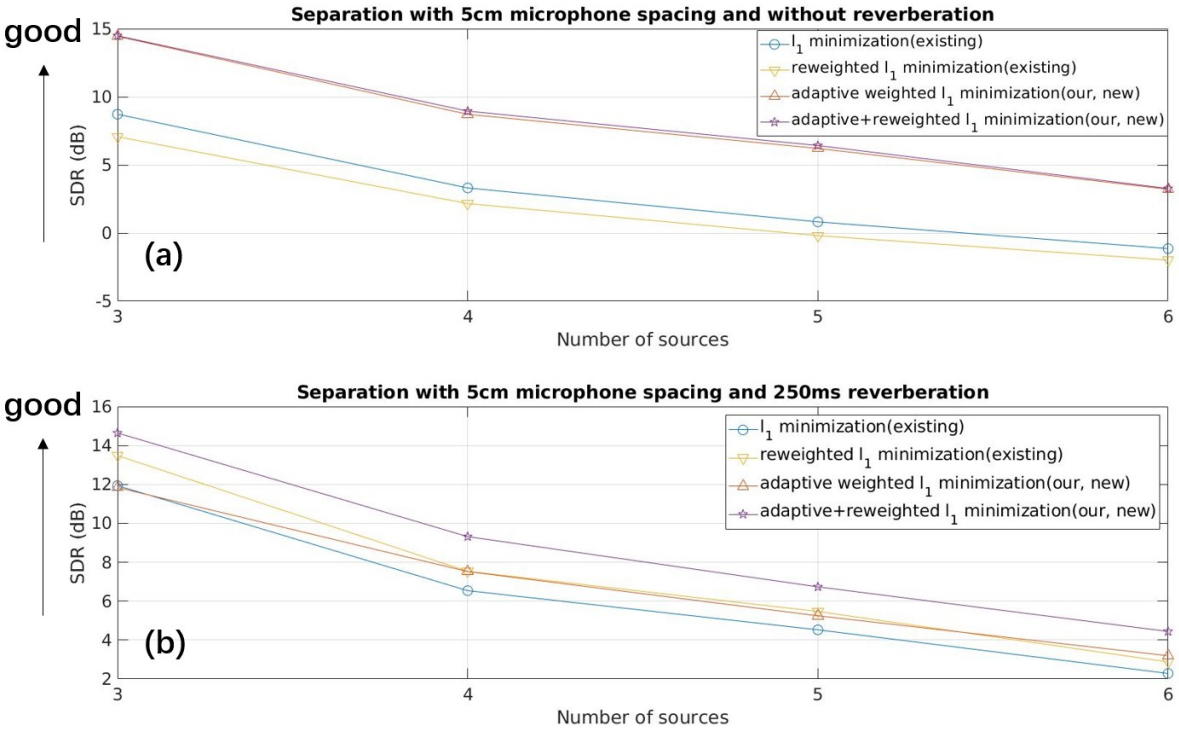

Figure 5.8: Average SDR as a function of the number of sources. There are two microphones with a five-centimeter microphone spacing. (a) Without reverberation. (b) In the presence of $RT_{60} = 250$ ms reverberation.

process from the j-th source to the i-th microphone, we define the mixing parameters to the noise ratio (MPNR) as the ratio of the sum of squares of the mixing parameters from the j-th source to the i-th microphones, i.e., $a_{ij}(t)$, and the noise, i.e., $n(t)$:

$$\text{MPNR}_{ij} := 10 \log \frac{\sum_t a_{ij}^2(t)}{\sum_t n^2(t)}. \tag{5.4.1}$$

We assume for each $i = 1, \cdots, I$ and $j = 1, \cdots, J$, the MPNRs are the same, and test the performance of the separation algorithms under different MPNRs from 50 dB to 0 dB. The window length of the STFT is set to be 64 ms and the microphone spacing is 1 m. As we can see from Fig. 5.9, the superiority of the proposed algorithms is clearly maintained until the MPNR is below 5 dB. Moreover, the adaptive weighted l_1 minimization seems more robust than the reweighted l_1 minimization.

To make the claim that our proposed algorithms are robust more convincing, we conduct another experiment in which we blindly estimate the mixing parameters of $\tilde{\boldsymbol{\mathcal{A}}}$, i.e., the acoustic transfer functions (ATFs), and based on the estimation, we estimate the STFT coefficients of the source signals. The estimation of the ATFs is exactly we used in our second work and is introduced in Sec. 4.2.2. In short, we estimate the source

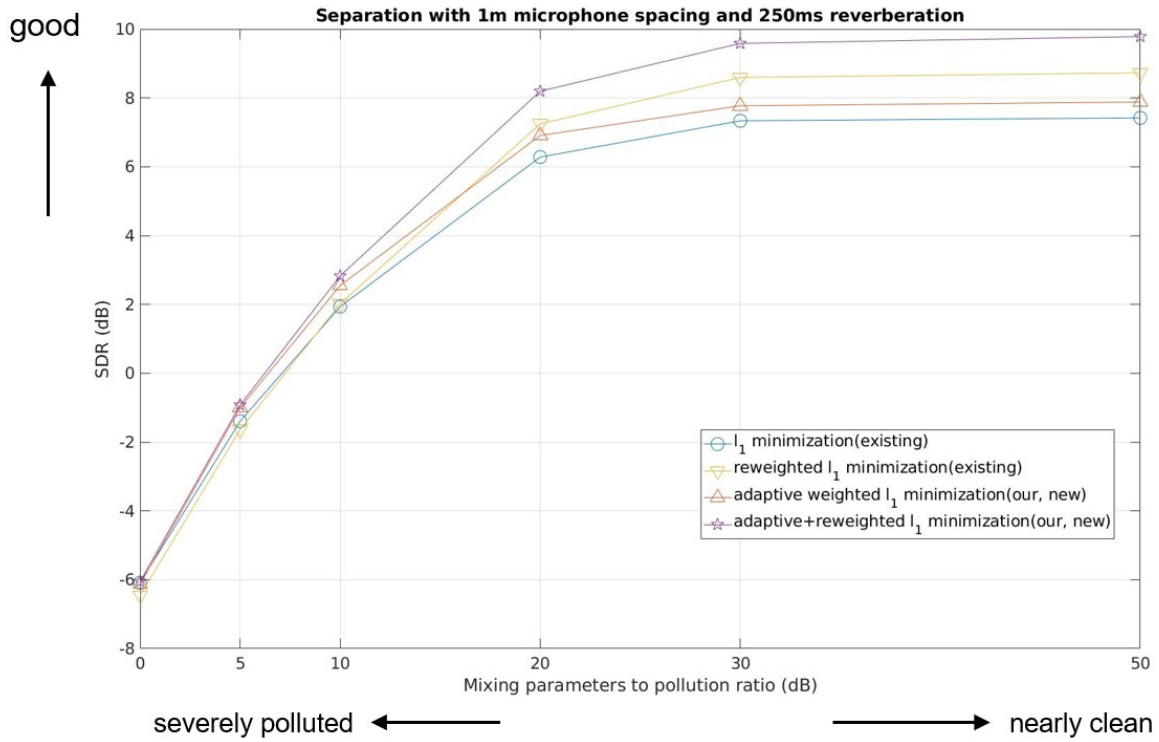

Figure 5.9: Average SDR as a function of the mixing parameters to noise ratio. There are two microphones with a one-meter spacing and four speech sources.

signals assuming only known the mixed signals. The result is shown in Fig. 5.10. In the absence of reverberation, the proposed methods outperform the benchmark methods, showing consistent robustness with the previous experiment. However, in the presence of reverberation, due to the difficulty of estimating ATF, the separation results of all methods are similarly bad. This is the restriction of the current ATF estimation method.

5.4.4 Time consumption

In this experiment, we evaluate the speed of the four algorithms. Theoretically, the adaptive reweighted l_1 minimization has equal time complexity to the reweighted l_1 minimization, several times of that of the normal l_1 minimization, and the adaptive weighted l_1 minimization has equal time complexity to the normal l_1 minimization. However, as we can see in Fig. 5.11, using the same l_1 minimization solver, the l_1 minimization with the adaptive weights always converges faster than without adaptive weights. The more sources, the difference is more obvious, up to 35% faster in the case of six sources and 5 cm microphone spacing. We guess the reason is that the adaptive weights indicate

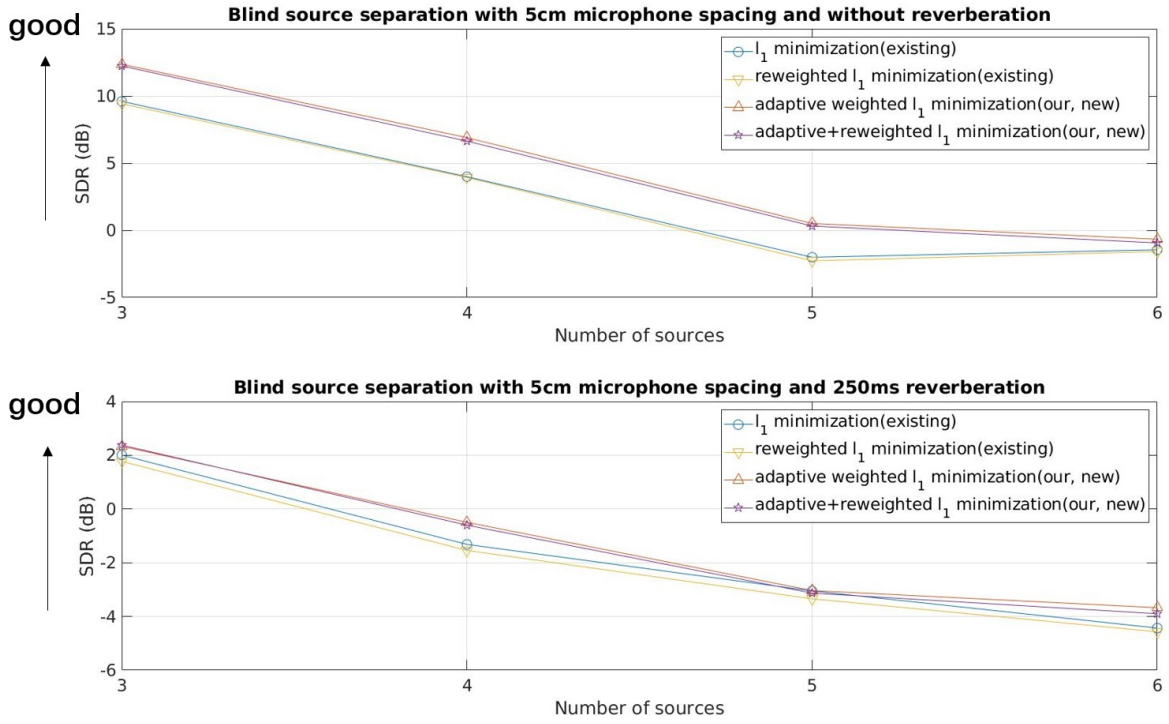

Figure 5.10: Average SDR as a function of the number of sources. There are two microphones with a five-centimeter spacing. (a) Without reverberation. (b) In the presence of $RT_{60} = 250$ ms reverberation. In this case, the mixing parameters are estimated so it is a blind source separation.

the correct sparse structure of the solution from the beginning of the iteration while l_1 minimization has no such priory information.

5.5 Summary

In this work, we propose a novel optimization framework for reconstructing source signals in the presence of reverberation and the number of microphones being fewer than the number of sources. More specifically, we propose a weighted l_1 minimization with weights adaptively changing with the mixed signals. The novel weights design constructs a closer connection between the source estimation and the mixing process. Compared to the normal l_1 minimization, the proposed adaptive weighted l_1 minimization shows a better separation performance, up to 5 dB improvement in SDR. Further improvement can be achieved by a reweighted scheme (the final result is also superior to the normal reweighted l_1 minimization). The improvement is robust to the estimation error of the mixing process, which is proved by two experiments, the first experiment is manually adding white noise

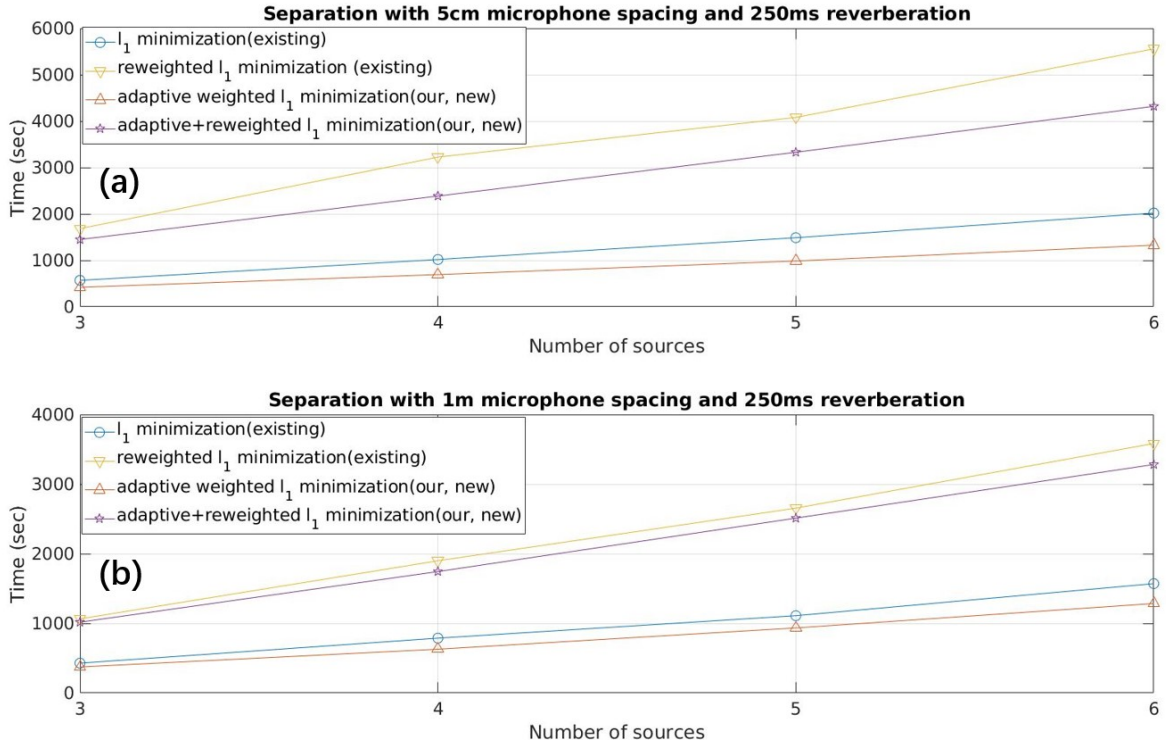

Figure 5.11: Average time consumption as a function of the number of sources in presence of $RT_{60} = 250$ ms reverberation. (a) Two microphones with a five-centimeter spacing. (b) Two microphones with a one-meter spacing.

to the mixing parameters, and the second experiment is a true blind estimation of the mixing parameters. An additional surprise is that the adaptive weighted l_1 minimization converges faster than the normal l_1 minimization using the same solver. The more sources, the difference is more obvious, up to 35% faster in the six sources case.

For future study, one may combine the mixing process estimation and the proposed source estimation into one optimization framework. Recently, there is one work [22] trying to estimate the mixing process and source signals simultaneously. This work is inspiring but the remaining challenge is how to add the update of the weights. Another future work is to keep following the convex optimization community for potentially faster algorithms.

Chapter 6

Conclusion

This thesis studies multi-microphone signal processing to enable speech separation in the presence of competitive talkers, In particular,

1. We proposed a time-frequency filtering technique to extract a very weak target signal by suppressing the interference signals: the interference suppressing response (ISR). In the presence of 200 ms reverberation (RT_{60} = 200 ms), ISR can boost a target signal from being 10 dB weaker than the interference signal to 10 dB stronger than the interference signal. ISR still works in highly reverberant environment (RT_{60} = 800 ms).

2. Based on the filtering technique, we proposed an algorithm to separate multiple sources using two microphones without knowing the sources' direction. The algorithm adopts the framework of binary masking followed by post-filtering. We use a well-known binary masking-based algorithm, the degenerate unmixing estimation technique (DUET), for the preliminary separation. Then, we construct a set of filters based on the separation results and use these filters to refine the preliminary separation. The new algorithm has a better performance compared to some mainstream blind source separation algorithms and a much faster processing speed (40 times faster than the non-negative matrix factorization-based algorithm).

3. When the environment has reverberation and the number of microphones is less than the number of sound sources, we proposed to reconstruct signals by minimizing the weighted l_1 norm of their STFT coefficients. Compared to normal l_1 minimization, experiments show a consistent performance improvement of our weighted l_1 minimization, up to 5 dB SDR improvement. The reconstruction speed is also faster,

up to 35%. Further performance improvement can be achieved by combining a reweighted scheme to iteratively refine the result.

6.1 Key findings

The key findings of this thesis are

1. If the time-domain filter is too long and the pursuit of sparsity is considered, it is a better alternative to filter in the time-frequency domain. Because each frequency is processed separately, this reduces the length of the filter and avoids the computational complexity caused by the pursuit of sparsity. This is demonstrated in our first work (Chapter 3).

2. The different speech signals in the STFT domain are neither completely disjoint nor completely overlapped. The completely disjoint assumption is too strong for source separation, resulting in the bad performance of binary masking. A better way is to use some information obtained from disjoint regions to help separate overlapped regions. This is demonstrated by our second work (Chapter 4), where spatial filters are computed using information from the disjoint region to take out interference components in the overlapped region.

3. The mixed signal's STFT coefficients are not arbitrarily distributed, with proper normalization, these STFT coefficients will cluster. If there are N sources, there will be N clusters. This phenomenon is more obvious when there is no reverberation, which can be utilized to estimate the mixing process or to roughly estimate the STFT coefficients of sources. The latter is demonstrated in our third work (Chapter 5) where we utilized it to compute the weights for the weighted l1 minimization.

4. Too long or too short STFT window length degrades the signal reconstruction performance of l_1 minimization, implying the reduction of sparsity. In our experiments, the optimal window length is 64 ms or 1024 samples under a 16 kHz sampling rate.

6.2 Limitations

Here we discuss the limitations of this thesis as well as some questions for further consideration.

1. The proposed ISR filter innately introduces distortion to the target signal although the distortion is mild (as if the target signal being processed by a high-frequency pass filter) and does not affect the understanding of semantics.

2. The proposed BSS algorithm in our second work assumes the number of sources, which may cause problems in practical applications. However, this is a general assumption in most BSS algorithms.

3. Predictably the proposed BSS algorithm in our second work will fail in high reverberation environments. One reason is that reflections produce some "fake" spatial cues (each reflection produces a virtual sound source), and algorithms have difficulty for source separation without knowing the direction of the real sound source. So far, speech separation in a highly reverberant environment still remains to be a challenge.

4. In a highly reverberant environment, the longer the STFT window, the more effective our proposed weights are. However, as mentioned in our key findings, if the STFT window is too long, sparsity is reduced and the results worsen.

5. Although we analyzed the recovery ability of l_1 minimization for sparse signals, our conclusions on this matter are somewhat qualitative and not sharpened enough to explain the good performance of l_1 minimization in speech separation where the dimension is extremely low, with only a few microphones and sources. The recovery ability of l_1 minimization in low-dimensional problems is expected to be explored, e.g., the distribution of the restricted isometry constant of a Gaussian matrix with small number of rows and columns.

6. In the case of 5 cm microphone spacing and the absence of reverberation, the reweighted l_1 minimization shows a worse-than-expected performance. This still remains an open question in this thesis.

6.3 Future works

In this section, we discuss some natural extensions of the works concluded in this thesis and some directions for further exploration.

1. **Robust MVDR**. This thesis generalizes ISR and minimum variance distortionless response (MVDR), pointing out that they come from the same type of filter: both try to suppress interference signals under some linear constraint that is based on different criteria. For example, maximizing the signal-to-interference ratio will derive MVDR. There is a large space to explore different filters obtained from different criteria. Most notable among these is robust MVDR, which is robust to the target direction estimation.

2. **Real-time blind source separation**. Design a real-time version for our BSS algorithm. The key is how to construct the binary masks and update the parameters of the filters in real-time.

3. **Blind sparse reverberant speech separation**. Currently, when we reconstruct the source signal using weighted l_1 minimization, the mixing process is assumed to be known. If the mixing process is not available, we can attempt to simultaneously estimate the mixing process, weights, and source signals. If the mixing process is available in some practical applications, we may borrow the idea from the Bayesian inference to modify some current blind source separation algorithms in which the mixing process is assumed not available. This should make for better separation results. Or the mixing process can be somehow translated into information about the current reverberant environment. This will be useful in civil engineering. Furthermore, the signal reconstruction by current l_1 minimization solver is time-consuming. We should keep following the convex optimization community for a potentially faster l_1 minimization solver.

References

[1] D. D. Carlo, P. Tandeitnik, C. Foy, N. Bertin, A. Deleforge, and S. Gannot, "dechorate: a calibrated room impulse response dataset for echo-aware signal processing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–15, 2021.

[2] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2018.

[3] M. Yu, W. Ma, J. Xin, and S. Osher, "Multi-channel l_1 regularized convex speech enhancement model and fast computation by the split bregman method," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 661–675, Feb 2012.

[4] S. Rickard, *The DUET Blind Source Separation Algorithm.* Dordrecht: Springer Netherlands, 2007, pp. 217–241. [Online]. Available: https://doi.org/10.1007/978-1-4020-6479-1_8

[5] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ica to multivariate components," in *Independent Component Analysis and Blind Signal Separation*, J. Rosca, D. Erdogmus, J. C. Príncipe, and S. Haykin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 165–172.

[6] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[7] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[8] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[9] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[10] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[11] A. Hidri, S. Meddeb, and H. Amiri, "About multichannel speech signal extraction and separation techniques," *Journal of Signal and Information Processing*, vol. 03, no. 02, p. 238–247, 2012. [Online]. Available: http://dx.doi.org/10.4236/jsip.2012.32032

[12] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[13] Y.-m. Qian, C. Weng, X.-k. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers of Information Technology & Electronic Engineering*, vol. 29, pp. 40–63, 2018. [Online]. Available: https://doi.org/10.1631/FITEE.1700814

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

[15] Y. Wang and Z. Zhou, "Source extraction in audio via background learning," p. 283, 2013. [Online]. Available: http://aimsciences.org//article/id/cc742ecf-de15-4f7d-9fcb-cce5939f4b44

[16] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[17] S. Doclo and M. Moonen, "Design of far-field and near-field broadband beamformers using eigenfilters," *Signal Processing*, vol. 83, no. 12, pp. 2641 – 2673, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168403002123

[18] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287 – 314, 1994, higher Order Statistics. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0165168494900299

[19] S. U. Wood, J. Rouat, S. Dupont, and G. Pironkov, "Blind speech separation and enhancement with gcc-nmf," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 745–755, 2017.

[20] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1818–1829, 2010.

[21] S. Arberet, P. Vandergheynst, R. E. Carrillo, J.-P. Thiran, and Y. Wiaux, "Sparse reverberant audio source separation via reweighted analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1391–1402, 2013.

[22] F. Feng and M. Kowalski, "Underdetermined reverberant blind source separation: Sparse approaches for multiplicative and convolutive narrowband approximation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 442–456, 2019.

[23] K. Yamaoka, N. Ono, S. Makino, and T. Yamada, "Time-frequency-bin-wise switching of minimum variance distortionless response beamformer for underdetermined situations," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7908–7912.

[24] V. G. Reju, S. N. Koh, and Y. Soon, "An algorithm for mixing matrix estimation in instantaneous blind source separation," *Signal Processing*, vol. 89, no. 9, pp. 1762–1773, 2009.

[25] Q. Guo, G. Ruan, and P. Nan, "Underdetermined mixing matrix estimation algorithm based on single source points," *Circuits, Systems, and Signal Processing*, vol. 36, no. 11, pp. 4453–4467, 2017.

[26] W. Cheng, Z. Jia, X. Chen, L. Han, G. Zhou, and L. Gao, "Underdetermined convolutive blind source separation in the time–frequency domain based on single source points and experimental validation," *Measurement Science and Technology*, vol. 31, no. 9, p. 095001, 2020.

[27] E. Sejdić, I. Djurović, and J. Jiang, "Time–frequency feature representation using energy concentration: An overview of recent advances," *Digital signal processing*, vol. 19, no. 1, pp. 153–183, 2009.

[28] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 109–116, 2003.

[29] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[30] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2010.

[31] ——, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain bss," in *2007 IEEE international symposium on circuits and systems*. IEEE, 2007, pp. 3247–3250.

[32] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 546–555, 2009.

[33] P. Smaragdis, "Blind separation of convolved mixtures in the frequency

domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231298000472

[34] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation," *IEEE Transactions on Signal Processing*, vol. 40, no. 8, pp. 1862–1875, 1992.

[35] Y. Avargel and I. Cohen, "Adaptive system identification in the short-time fourier transform domain using cross-multiplicative transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 162–173, 2008.

[36] F. Feng and A. Begdadi, "Reverberant audio blind source separation via local convolutive independent vector analysis," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020, pp. 1–6.

[37] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 774–784, 2006.

[38] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 5, pp. 1075–1084, 2017.

[39] Z.-Q. Wang and D. Wang, "All-neural multi-channel speech enhancement." in *Interspeech*, 2018, pp. 3234–3238.

[40] Y. Zhao, Z.-Q. Wang, and D. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 53–62, 2019.

[41] L. Ferrer, M. Graciarena, and V. Mitra, "A phonetically aware system for speech activity detection," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5710–5714.

[42] F. Grondin and F. Michaud, "Robust speech/non-speech discrimination based on pitch estimation for mobile robots," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 1650–1655.

[43] X. Wang, *A Novel Approach to Blind Source Separation and Extraction in Audio*. Michigan State University. Applied Mathematics, 2015.

[44] S. Araki, F. Nesta, E. Vincent, Z. Koldovský, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (sisec2011): - audio source separation -," in *Latent Variable Analysis and Signal Separation*, F. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 414–422.

[45] S. Mogami, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, K. Kondo, and N. Ono, "Independent low-rank matrix analysis based on time-variant sub-gaussian source model for determined blind source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 503–518, 2019.

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

[47] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, vol. 5, 2000, pp. 2985–2988 vol.5.

[48] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2002, pp. I–529–I–532.

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

[50] Y. He, H. Wang, Q. Chen, and R. H. So, "Harvesting partially-disjoint time-frequency information for improving degenerate unmixing estimation technique," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 506–510.

[51] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[52] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 8, pp. 1207–1223, 2006.

[53] E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes rendus mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008.

[54] D. Donoho and J. Tanner, "Counting faces of randomly projected polytopes when the projection radically lowers dimension," *Journal of the American Mathematical Society*, vol. 22, no. 1, pp. 1–53, 2009.

[55] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

[56] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and-norm minimization," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–12, 2006.

[57] Y. Li, S.-I. Amari, A. Cichocki, D. W. Ho, and S. Xie, "Underdetermined blind source separation based on sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 2, pp. 423–437, 2006.

[58] Y. Li, W. Nie, F. Ye, and Y. Lin, "A mixing matrix estimation algorithm for underdetermined blind source separation," *Circuits, Systems, and Signal Processing*, vol. 35, no. 9, pp. 3367–3379, 2016.

[59] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted 1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5, pp. 877–905, 2008.

[60] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[61] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.

[62] J. Wright and Y. Ma, *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. Cambridge University Press, 2022.

[63] J. B. Kruskal, "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear algebra and its applications*, vol. 18, no. 2, pp. 95–138, 1977.

[64] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[65] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE transactions on Information theory*, vol. 49, no. 12, pp. 3320–3325, 2003.

[66] L. Welch, "Lower bounds on the maximum cross correlation of signals (corresp.)," *IEEE Transactions on Information theory*, vol. 20, no. 3, pp. 397–399, 1974.

[67] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[68] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[69] M. Rudelson and R. Vershynin, "On sparse reconstruction from fourier and gaussian measurements," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 61, no. 8, pp. 1025–1045, 2008.

[70] P. L. Combettes and J.-C. Pesquet, "A douglas–rachford splitting approach to nonsmooth convex variational signal recovery," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 564–574, 2007.

[71] M.-J. Fadili and J.-L. Starck, "Monotone operator splitting for optimization problems in sparse recovery," in *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2009, pp. 1461–1464.