Ph.D. DISSERTATION

ACOUSTIC SIGNAL PROCESSING APPLICATIONS FOR INDOOR MULTIPLE SOUND SOURCE ENVIRONMENT

실내 다중 음원 환경에 적용 가능한
음향 신호 처리 기법과 그 응용

BY

PARK JOO HYUN

AUGUST 2022

DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Ph.D. DISSERTATION

ACOUSTIC SIGNAL PROCESSING APPLICATIONS FOR INDOOR MULTIPLE SOUND SOURCE ENVIRONMENT

실내 다중 음원 환경에 적용 가능한
음향 신호 처리 기법과 그 응용

BY

PARK JOO HYUN
AUGUST 2022

DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

ACOUSTIC SIGNAL PROCESSING APPLICATIONS FOR INDOOR MULTIPLE SOUND SOURCE ENVIRONMENT

실내 다중 음원 환경에 적용 가능한
음향 신호 처리 기법과 그 응용

지도교수 김 성 철
이 논문을 공학박사 학위논문으로 제출함

2022년 8월

서울대학교 대학원

전기·정보공학부

박 주 현

박주현의 공학박사 학위 논문을 인준함

2022년 8월

위 원 장:	김 남 수	(인)
부위원장:	김 성 철	(인)
위 원:	심 병 효	(인)
위 원:	이 웅 희	(인)
위 원:	최 정 식	(인)

Abstract

Recently, research on acoustic signal processing is increasing. This is because meaningful information can be obtained and utilized usefully from acoustic signal processing. Therefore, this paper deals with the acoustic signal processing techniques for sound recorded in the indoor environment.

First, we introduce a method for estimating the location of a sound source under indoor environment where there are high reverberation and lots of noise. In the case of existing methods such as interaural level difference (ILD) based localization, time difference of arrival (TDoA) based localization, and steered response power phase transformation (SRP-PHAT) based localization, the accuracy is lowered when applied under recordings from indoor environment with high reverberation. However in this paper, we define a new cost function that can find an optimal combination of microphone pair which results in highest performance. The microphone pair with the lowest value of cost function was chosen as an optimal pair, and the source location was estimated with the optimal microphone pair. It was confirmed that the distance error was reduced compared to existing methods.

Next, a technique for recovering the lost sample value from the recorded signal called sketching and stacking with random fork (SSRF) is introduced. In this technique, the target sound source is a superposition of several sinusoidal signals. It is assumed that there are multiple sound sources in the anechoic chamber, but there is only one microphone. It is trivial that a sinusiodal wave can be transformed into an exponential function based on Euler's formula. If some of the terms of the exponential function follow a geometric sequence, those values can be obtained using SSRF. To solve this problem, the concept of a random fork is newly introduced. Comparing the recovery error based on SSRF with existing methods such as compressive sensing based technique and deep neural network (DNN) based technique, the accuracy of

SSRF based signal recovery was higher.

Finally, this paper introduces a blind source separation (BSS) technique for based on the previously introduced SSRF technique. In this technique, as before, it is assumed that the sinusoidal waves are superposed. In addition, while the previous technique assumed a situation where all sinusoidal waves were emitted simultaneously, this technique assumed a situation where different sound sources were separated by different distances from the microphone and arrived at the microphone with different time delays. Under these assumptions, a new BSS method for separating single signals from the mixture based on SSRF is introduced. The SSRF BSS is mainly composed of three steps: estimation of the number of sound sources, estimation of time delay, and signal separation. While the existing BSS methods require information on the source number to be known *a priori*, SSRF BSS does not require source number. Whereas existing BSS methods can only be applied to signals without time delay, SSRF BSS method has the advantage in that it can be applied to the mixture of signals with different time delays. It was confirmed that SSRF BSS produces more accurate separation results compared to the existing independent component analysis (ICA) BSS and Yu Gang (YG) BSS.

keywords: Acoustic Signal Processing, Acoustic Localization, Acoustic Signal Recovery, Acoustic Blind Source Separation, Sketching and Stacking with Random Fork

student number: 2016-20904

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

Research on acoustic signal processing techniques has been conducted for several decades. Acoustic signal processing techniques, which started decades ago, have mainly focused on analyzing sound waves in terms of waves [1]. However, nowadays, by using a microphone, sound is recorded as a discrete acoustic signal like the received signal in the radio wave processing. This means that the existing radio signal processing technique or localization method can also be applied to acoustic signal processing [2, 3].

In addition to performing acoustic signal processing in terms of sound waves, modern acoustic signal processing is utilized in much more diverse fields. In particular, indoor and outdoor localization techniques for finding the location of users or devices are being extensively studied [4, 5, 6]. Likewise in the field of acoustics, the most widely used technique is to find the location of the sound source. This is called acoustic localization. Although acoustic localization methods stem from existing radio wave based localization technique [7], the same performance cannot be expected with the same signal processing method, because radio waves and sound waves have different characteristics. Unlike radio waves whose frequency band is several tens of kHz, the audible frequency band is $20 - 20000$ Hz, which is very low compared to radio waves. Sound waves have a longer wavelength than radio waves due to their lower frequen-

cies. As a result, sound waves are often diffracted, and sound waves generate a lot of reverberation and noise in an indoor space compared to radio waves [8, 9]. Therefore, if the existing localization technique targeted for radio waves is applied to sound wave, the performance will be deteriorated.

The recorded sound is mixed with various noises and reverberations. In order to suppress noises and reverberations, most of the acoustic localization techniques convert the time domain acoustic signal into frequency domain [10]. This method is widely used because the effects of reverberation and noise can be slightly reduced in the process of converting recorded sound into a frequency domain. This is called a phase transformation (PHAT) technique [10]. Among the PHAT based acoustic localization techniques, the most widely used method is SRP-PHAT [11]. This is an acoustic localization technique that calculates the steered response power (SRP) and determines that the sound source exists at the point with the highest value of SRP. However, SRP-PHAT method has a drawback in that it is difficult to apply to real-time sound source tracking due to the large amount of computation. Another drawback of SRP-PHAT method is that when applied to the the actual recording, the accuracy of the angle of arrival (AoA) is high whereas the range estimation accuracy was low. This is because SRP-PHAT is kind of a beamforming method [12]. Since the localization accuracy of the SRP-PHAT method in the anechoic chamber was not high, the localization accuracy was even lower when applied to the sound recorded in an actual indoor environment with noise and reverberation. Therefore, a novel acoustic localization method is introduced in Chapter 2 which can be applied to noisy and reverberant recordings. The acoustic localization method proposed in Chapter 2 defines a new cost function to find the optimal microphone pair which can show highest localization accuracy. The detailed description proceeds in Chapter 2.

Next, we move on to acoustic signal restoration, which is another field of acoustic signal processing. Signals recorded with a microphone have a high probability of samples being lost. This could be a problem with the microphone hardware, or it could

be due to a loss in the process during saving the recording file. The acoustic signal itself has very high noise and reverberation, and if a loss occurs, the performance of the acoustic signal processing deteriorates even more. Therefore, it is important to restore the lost sample values in order to improve the performance of the acoustic signal processing. Therefore, in this paper, we propose a technique for recovering lost signal sample values. The sound targeted in this study is a mixture of sinusoidal signals. This is because most of the acoustic signals that exist in nature are expressed as the sum of the sinusoidals. In this paper, the concept of random fork is introduced to recover the lost signal samples. Assuming that the number of sound sources is already known, by using random fork, the elements constituting the sinusoidal signals can be obtained. Detailed explanation on signal recovery based on sketching and stacking with random fork (SSRF) is introduced in Chapter 3.

For the next, the signal separation technique that separates each signal from a mixture of multiple sound sources recorded with only one microphone is introduced. In the field of acoustics, a technique for separating individual sound sources from a multi-source mixture is called BSS [14]. A large number of BSS techniques assume that the number of sound sources is *a priori* [15, 16, 17]. However, there are few BSS techniques that can separate signal without knowing the source number [18, 19]. It is more difficult to separate the mixed signal in a situation where there is no information about the number of sound sources. Also, most of the BSS techniques use larger number of microphones than the number of sound sources [20, 21]. This is because the solution is determined only when the number of equations exceeds the number of sound sources. In this sense, the case of using fewer microphones than the number of sound sources is called an underdetermined BSS problem [18, 19]. Conventionally, it is more difficult to solve underdetermined problem than solving determined problem. In the SSRF based BSS method introduced in Chapter 4, only one microphone is used to solve the BSS problem when there is no information about the number of sound sources. In Chapter 4, based on the SSRF technique, the source number is obtained from the mixture

of sinusoidals, then time delays of each source are obtained, and finally, the signal is separated.

In this paper acoustic signal processing techniques including acoustic localization, signal recovery using SSRF and acoustic BSS based on SSRF are introduced. In chapter 2, acoustic localization technique which can be applied to noisy and reverberant condition is introduced. In chapter 3, signal recovery method called SSRF which can recover the lost sample values from the mixture of sinusoidal signals is introduced. Then, signal separation using SSRF is introduced in chapter 4. This signal separation method can separate mixed sinusoidal signals with time delays. Chapter 5 concludes this paper.

Chapter 2

IMPROVING ACOUSTIC LOCALIZATION PERFORMANCE BY FINDING OPTIMAL PAIR OF MICROPHONES BASED ON COST FUNCTION

2.1 Motivation

The demand for acoustic signal processing is increasing. Study on acoustic signal processing are utilized in various fields. Among the numerous acoustic signal processing applications, the most widely used field is sound source localization [22, 23, 24]. Locating the position of a sound source is called acoustic localization.

Let us find out why acoustic localization is widely studied. First, almost all beings around us generate sound. This includes both living and non-living things. Living things generate various sounds during their life. For example, humans speak with their voices, and birds chirp with their voices. In addition to these voices, in the case of living things, there are also sounds generated by movement. The example sounds above are generated from bio-energy not using electric power or devices. For the case of a machines in a factory, continuous and constant mechanic sound is emitted. Machines such as automobiles also make various noise sounds while driving. In particular, for the case of a cell phones, the sounds generated from cell phones are more varied. The

most typical example is a ringtone that occurs when a call is received. The notification sounds also frequently occur when receiving a text messages. Given this, if the location of the sounds source in daily life can be identified, the applications are limitless.

For the case of acoustic localization, the basic theory is same as radio wave based localization. An example is trilateration, which finds the distance from several nodes to a target and estimates the location by finding the intersecting position of three distances [25]. In radio wave localization, distance information is extracted from radio signals, where as in acoustic localization, distance information is extracted from sound wave. Advantages of acoustic localization will be explained from now on. First, the advantage of acoustic localization is that any living or no-living things making sound can be a target. For the case of radio wave based localization the location of the target can be found only when the target generates radio waves. In other words, a target that does not generate radio waves, localization cannot be performed. Since electric power is basically used to generate radio waves, an object that does not use electricity cannot be a target of radio wave localization techniques. However, the advantage of sound source localization is that it can target any object that generates sound. Examples of such sounds are the sound of falling objects, human speech, and howling of an animal. Second, the price of the microphone, which is a sensor necessary for acoustic localization, is cheaper than other sensors. Another advantage is that even with a low-priced microphone, a certain level of performance is guaranteed. It has a much more price advantage than the price of a transmitter or receiver used for radio wave based localization. Third, the fact that there are built-in microphones and speakers exist in a cellphone is one of the great advantages of acoustic localization. Since a number of microphones are already built-in in cell phones, these microphones can be utilized for acoustic localization without having to purchase and install additional microphones. In this case, the cost can be reduced even more.

From now on, the applications of acoustic localization will be discussed. As mentioned in the previous paragraph, sound source localization can target anything that

generates sound. This means that even under extreme condition when electricity is cut off, the position of a target can be found only by sound. A typical example is a disaster situation. In most cases, electric power is cut off at sites such as landslides caused by earthquakes or building collapses. In this extreme situation, it is important to locate and rescue the survivors before golden hour. Also under this harsh condition, the cell phone batteries usually run out, making radio based localization useless. Therefore, acoustic localization is important under harsh condition. Another example where acoustic localization can be applied is location based service (LBS). If the user's location can be identified in a space such as a shopping mall or a department store, its utilization is very wide. First of all, it is possible to obtain information such as which store is the best in business by finding out which floor is most crowded, and in which field visitors are most interested. Going one step further, based on the location information of the visitors, shopping information that is precisely targeted to their gender and age can be broadcasted to their location accurately. Like this, LBS is a service that accurately satisfies users' needs based on their location. If the LBS is based on the acoustic localization, unlike radio wave based localization, there is no need to go through the process of consenting to the collection of personal information, which is also a great benefit for the company. Next, an application such as abnormal noise detection in a factory is another application. Recently, a trend of transformation into smart factories of conventional factories have emerged due to the digital transformation. Smart factory refers to the evolution of a factory that extracts maximum efficiency with minimum manpower by automating most of the things in the factory with the introduction of cutting-edge technologies such as internet of things (IoT) and artificial intelligence (AI) [26]. In fact, even with the current level of technology, complete factory automation is impossible. When an unexpected situation such as a machine breakdown occurs, machines and AI become useless, where human resource is essential to solve the problem. Under this circumstance, acoustic localization can also be utilized for the detection of abnormal sound. In the factory, numerous machines generate repetitive and continu-

ous noise sound, and when an abnormal situation occurs, a completely different sound pattern is generated. If acoustic localization is utilized to detect such anomalies, the smart factory operation efficiency can be improved. Also abnormal sound detection and localization can be used for vehicle maintenance. An sudden abnormal sound often occurs that only the driver can sensitively hear while driving. However, there are many cases where the mechanics cannot find out where those small abnormal sound come from. In this case, using a microphone array to mechanically detect and localize the abnormal sounds can help vehicle maintenance.

In this chapter, a new technique for acoustic localization is introduced. Based on the existing sound source localization method, the newly defined cost function is introduced to find the optimal microphone pair that can produce the highest accuracy. In section 2.2, the existing sound source localization technique used in this technique will be described. In section 2.3, the design technique including the definition of the newly introduced cost function will be described. Results and discussion proceed in section 2.4. Finally, this chapter is summarized in section 2.5.

2.2 Conventional Acoustic Localization Methods

2.2.1 Interaural Level Difference

Humans use the difference between the two ears to estimate the direction and position of sound sources. The acoustic localization technique devised from this is an interaural level difference (ILD) based localization technique. It uses two microphones. It is a technique for estimating the location of a sound source using the difference in energy of sound between two microphones. This technique is similar to the received signal strength (RSS) based localization technique for radio wave.

The ILD acoustic localization method is applicable only to the sound of the high frequency band. This is because the wavelength in the high frequency band is shorter, so it can better reflect the path difference between the two microphones.

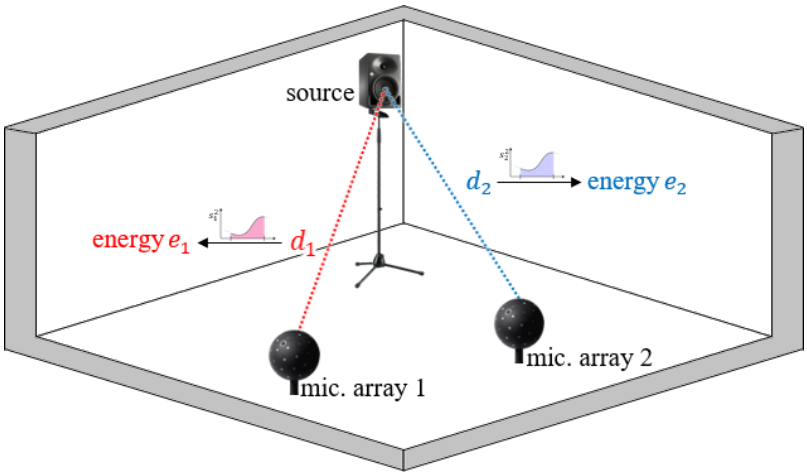


Figure 2.1: ILD acoustic localization concept.

The the energy level recorded with both microphones can be written as follows:

$$E_i = \int_0^T \left[\frac{s^2(t)}{d_i^2} + n_i^2(t)\right] dt \tag{2.1}$$

T, s_i, d_i and n_i stands for time length, recorded sound signal, distance between speaker and microphone and noise respectively. i stands for the microphone index. (2.1) can be expanded as below:

$$E_i = \frac{1}{d_i^2} \int_0^T s_i^2(t) dt + \int_0^T n_i^2(t) dt \tag{2.2}$$

After substituting two microphone indices $i = 1, 2$ to microphone index i in the above expression, the relationship between distance d_i and E_i is as follows:

$$E_1 d_1^2 = E_2 d_2^2 + \int_0^T \left[n_2^2(t) - n_1^2(t)\right] dt \tag{2.3}$$

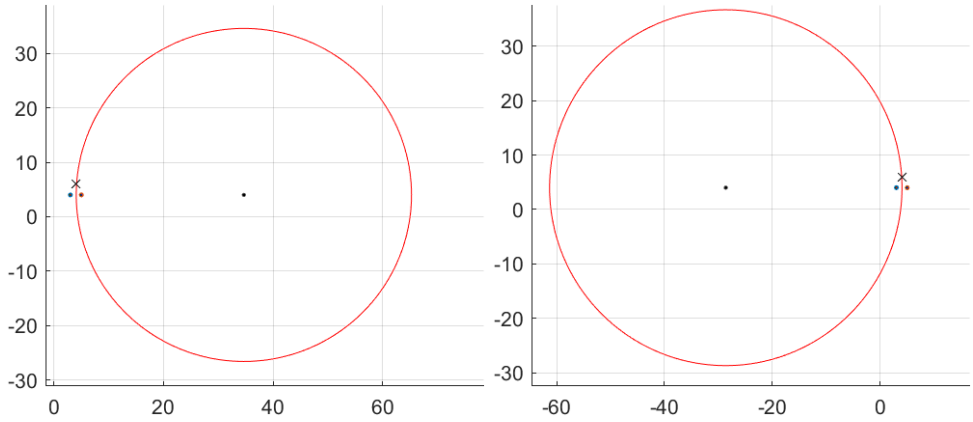
Assuming that the ambient noise is approximately equal so the noise term in (2.3) can be ignored, the relationship between the distance between the microphone and the sound source and the energy of the sound recorded with the microphones can be written as below:

$$E_1 d_1^2 = E_2 d_2^2 \tag{2.4}$$

From (2.4), it can be seen that the acoustic energy of the recorded signal and the square of the distance between the sound source and the microphone are inversely proportional. Physically, it means that the energy of the recorded sound decreases as the distance from the microphone increases. Therefore, if you calculate the ratio between the energy of the recorded signals, you can check how far away it is from each microphone. The concept of ILD acoustic localization is shown in Fig. 2.1.

The ILD based acoustic localization for indoor experiment is shown in Fig. 2.2. In both Fig. 2.2 (a) and (b), microphone arrays are shown with tiny red and blue hexagons on $(3, 4)$ and $(4, 4)$. The position of the sound source is marked with an '$\times$'. The ILD result if shown with red solid line circle in both Fig. 2.2 (a) and (b). As shown in Fig. 2.2, the red ILD circle passes the '$\times$' mark.

However, the ILD technique has its drawbacks. First of all, since the ILD technique is based on the energy of the recorded sound, it is vulnerable to changes in the signal-to-noise ratio (SNR) of the microphone. The smaller the SNR value, the greater the noise. If the level of noise increases due to the surrounding environment, the calculation of energy of the recording will be inaccurate inevitably. In particular, the noise term ignored in (2.3) cannot be ignored anymore if the SNR is low. Next, there is a disadvantage in that the accuracy of the ILD localization is lowered under high reverberation condition. Reverberation of sound refers to a fake sound that con-



(a) ILD acoustic localization result. (b) ILD acoustic localization result.

Figure 2.2: Examples of ILD acoustic localization result.

tinues to remain due to reflection and diffraction in space even after emission of sound from source is over. If the reverberation is high, the reverberation is also added to the recording, and results in the distorted sound which is far from the original sound. If the sound energy is calculated from the recording with high-reverberation, the energy of the reverberation sound is also added, which lowers the accuracy of the ILD acoustic localization method. Third, the value varies greatly depending on the recording sampling rate. The higher the sampling frequency, the higher the accuracy of the energy calculation value. However, high sampling frequency has a disadvantage in that the size of the recorded file is too large and the amount of computation is also too large. However, for low sampling frequency, the advantage is that the recording file size is small. However, if the sampling rate is too low and sparsely sampled, the quality of the recorded sound file is deteriorated, so the accuracy of the calculation also gets low. The final disadvantage is that the accuracy of the ILD localization is low for sudden and unexpected burst of sounds. A sound such as gunshot that occurs suddenly has a similar waveform to the impulse sound, where time duration of impulse is very short. For a sudden burst of sound with such short time duration, the calculated energy value is not distinctively different for each microphone, which results in low localization accuracy.

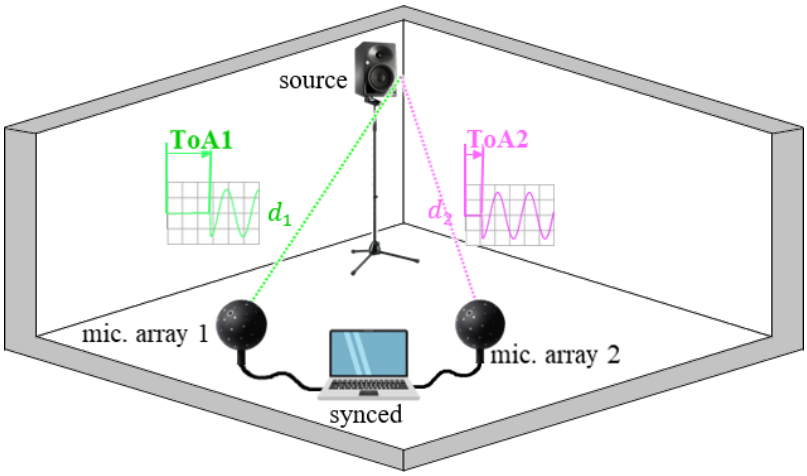


Figure 2.3: TDoA acoustic localization concept.

2.2.2 Time Difference of Arrival

In this section, the existing method for estimating sound source location based on TDoA is introduced. TDoA means the difference in arrival time due to the difference in the distance between the node and the target. This technique is widely used not only for acoustic localization but also for radio wave based localization. A method of obtaining the cross correlation of two signals is used for TDoA measurement and calculation. However, unlike time of arrival (ToA), which means absolute time of arrival, TDoA means only the difference in relative time of arrival.

TDoA based sound source localization estimation is mathematically expressed in this paragraph. The cross correlation of two signals is expressed as below:

$$f * g = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau \tag{2.5}$$

The $f(t)$ and $g(t)$ represents the continuous signal in time domain. The time when peak value of the cross correlation appears is chosen to be the TDoA of each signal.

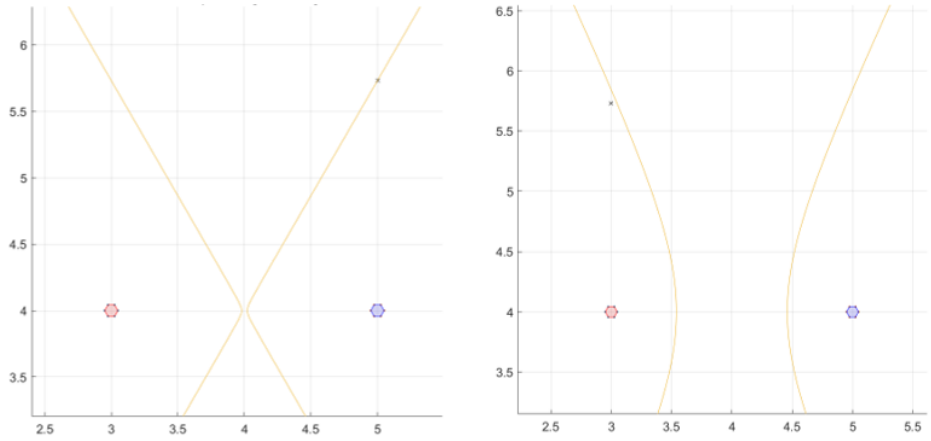
If the correlation peak index is i, the distance difference obtained from TDoA can be written as follows:

$$d_1 - d_2 = \frac{i}{f_s} \cdot c \tag{2.6}$$

d_1 and d_2 mean the distance from the sound source to microphone 1 and microphone 2 respectively. f_s is the sampling frequency and c is speed of sound. From this, all points with the same distance from two microphones can be selected as source position candidates. Mathematically, a set of points having the same distance difference from two points is defined as a hyperbola. Therefore, the location of the sound source can be estimated by drawing and finding the intersecting point of two hyperbolas using two or more microphone pairs. The concept of TDoA acoustic localization is shown in Fig. 2.3.

The experimental result of TDoA acoustic localization is shown in Fig. 2.4. In Fig. 2.4, two microphone arrays are pointed on $(3, 4)$ and $(-3, 4)$. Since microphone arrays used for experiment are in the shape of hexagon, each microphone array is drawn as a hexagon. Source position is marked with an '$\times$' mark. TDoA acoustic localization result is marked with orange solid line in both Fig. 2.4 (a) and (b). It can be seen that the hyperbola passes through the location of the sound source.

However, the TDoA based sound source localization method has its drawbacks. First, the performance of TDoA based acoustic localization is greatly affected by ambient reverberation and noise. This, in fact, generally stands true for most of other acoustic localization techniques. There is no environment without reverberation and noise in everyday life except for an anechoic chamber. If recording is performed in an environment where reverberation and noise exist, the recorded signal is contaminated, which inevitably lowers the TDoA based acoustic localization accuracy. Second, in order for TDoA based acoustic localization to show high accuracy, the sampling rate



(a) TDoA acoustic localization result. (b) TDoA acoustic localization result.

Figure 2.4: Examples of TDoA acoustic localization result.

must be high. A high sampling rate is related to a small sampling time interval. Since the recording is stored as discrete samples, if the sampling time interval is large, then the time interval between indices is widened. In other words, no matter how accurate the algorithm is, an error is bound to occur as much as the extended time interval. Therefore, the sampling rate should be high in order to have high cross correlation time accuracy. Finally, as you can see in (2.6), this method is affected by the speed of sound. However, the speed of sound is greatly affected by the surrounding environment. The speed of sound varies greatly depending on climatic factors such as temperature, humidity, and wind direction. Therefore, it is possible to obtain a more accurate TDoA by reflecting the change in the speed of sound, which varies greatly even with small climatic factors in real time. However, in reality, it is difficult to estimate the exact speed of sound in that particular measurement environment at every moment, and as a result, the accuracy of TDoA based acoustic location estimation is lowered.

2.2.3 Steered Response Power Phase Transformation

The previously introduced acoustic localization technique was to analyze the recorded sound signal in the time domain. However, it is very important to remove the reverber-

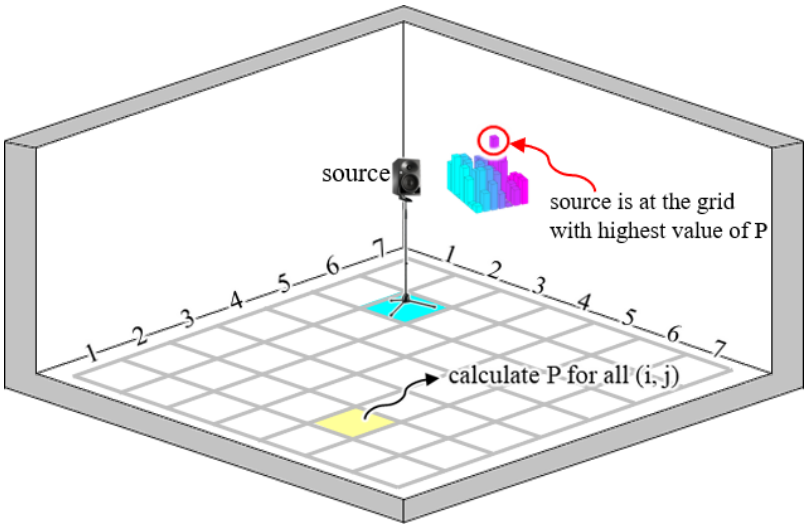


Figure 2.5: SRP-PHAT acoustic localization concept.

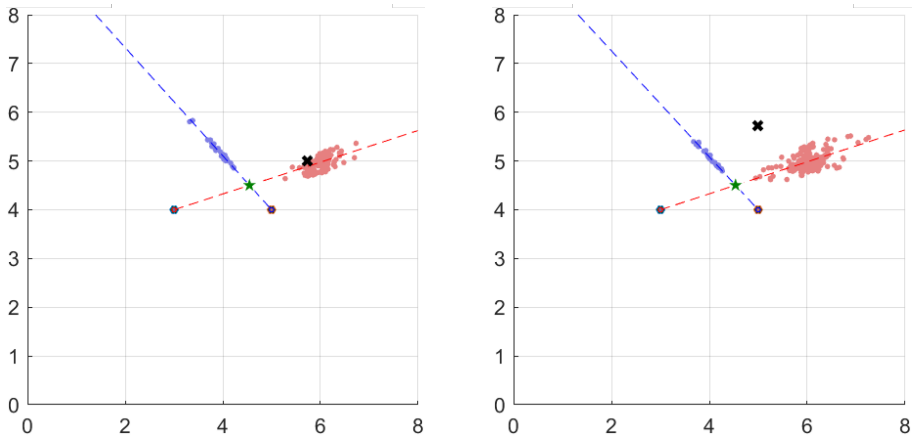
ation and noise from the recorded signal due to the characteristics of acoustic signal processing techniques that are vulnerable to reverberation and noise.

Removing reverberation and noise through a filter can be a solution, or compressive sensing can be another. However, in fact, it is almost impossible to accurately remove only noise and reverberation from the recorded signal. Therefore, in most cases, acoustic signal processing is performed by transforming a time domain signal into frequency domain. This method of transforming acoustic signal into frequency domain and then processing is called PHAT.

This section explains SRP-PHAT technique among other PHAT based acoustic localization techniques that have been widely studied for a long time. First SRP need to be defined. The definition of SRP can be written as follows:

$$P_n := \int_0^T \left| \sum_{i=1}^{M} w_i m_i(t - \tau(x, i)) \right|^2 dt \tag{2.7}$$

T is the observing time window, w_i is the weighting factor for i-th microphone, $\tau(x, i)$



(a) SRP-PHAT acoustic localization result. (b) SRP-PHAT acoustic localization result.

Figure 2.6: Examples of TDoA acoustic localization result.

is the direct path delay between position x and i-th microphone and M is the number of total microphones. P_n is a type of cost function. P_n has the largest value when position x is closest to the actual source.

For the suppression of noise and reverberation, P_n is converted into frequency domain which is written as below:

$$P_n^{'} := \sum_{k=1}^{M} \sum_{l=k+1}^{M} \int_{-\infty}^{\infty} W_k(\omega) W_l^*(\omega) M_k(\omega) M_l^*(\omega) e^{j\omega(\tau(x,l)-\tau(x,k))} \tag{2.8}$$

where M is the number of total microphones, W_i is the weighting factor where w_i in (2.7) converted into frequency domain and M_i is the i-th microphone signal converted into frequency domain. Usually weighting factor W_i is set to be $\frac{1}{|M_k||M_l^*|}$ so that the microphone signal power can be normalized. SRP-PHAT localization method is a widely used because the effect of noise and reverberation is reduced when converting time domain acoustic signal into frequency domain. The concept of SRP-PHAT is shown in Fig. 2.5.

The experimental result of SRP-PHAT acoustic localization is shown in Fig. 2.6.

The sound source was recorded in an extremely reverberant indoor condition. The source used was a real gunshot sound, which has very short time duration. In Fig. 2.6 (a) and (b), SRP-PHAT result of microphone array 1 and microphone array 2 are marked with red and blue dots respectively. As shown, the angle estimation accuracy of SRP-PHAT has a tendency, but the range estimation accuracy is low. As a result, two AoA results were combined and the intersecting point was chosen to be the estimated position. The localization result is marked with a green star in both figure. As shown in Fig. 2.6, the SRP-PHAT localization accuracy is low under reverberant indoor condition.

However, the SRP-PHAT acoustic localization also has its drawbacks. First of all, the accuracy is lowered for wideband signals. Considering that most of the sound signals are mainly wideband signal, SRP-PHAT cannot always achieve high accuracy. Also, when only two microphones are installed in a multi-source situation, the accuracy of the SRP-PHAT acoustic localization method is even lowered. In addition, there is a disadvantage in that computation complexity is too large because the value of the cost function must be calculated and compared for all directions or positions in space.

2.3 System Model

In this section, newly devised acoustic localization method to overcome the influence of high noise and reverberation is introduced. The system model is explained from now on.

2.3.1 Experimental Scenarios

This section describes an experimental scenario of the devised method. The study in this chapter assumes an indoor environment where high noise and reverberation exists. It refers to a general office environment that is far from an ideal anechoic chamber. The actual experiment was conducted in a seminar room within the Institute of New

Media and Communication at Seoul National University. In the seminar room, there is no sound absorbing material at all, so it can be said that the sound is highly resonant.

Two microphone arrays were installed. The shape of microphone array was in a from of hexagon, and had six microphones located at each vertex. The length of each side of the hexagon was 5 cm. Each microphone array was positioned at $(-1, 0)$ and $(1, 0)$ on the x-axis respectively. That is, the distance between the two microphone arrays was 2 m. After that, the speaker was positioned on the five points $30°$, $60°$, $90°$, $120°$, and $150°$ from the x-axis on a circle with a radius of 2 m from the origin.

The sound source is a recording of a gunshot. The gunshot sound is literally made when a gun is fired, which has a very short time interval, and a very high-energy occurs within that short time interval. This gunshot sound is similar to the shape of an impulse. When the gunshot sound source is converted to the frequency domain,]it is a wideband signal, which makes it difficult to localize.

2.3.2 Definition of Cost Function

In this section, a new cost function is defined. The purpose of this cost function is to find the optimal pair of microphones with the best performance. Therefore, learning process to find out microphone pair has highest accuracy is necessary. This learning process uses acoustic localization result from existing acoustic localization methods as a test set. The goal is to find the optimal microphone pair that produces the lowest error through the learning stage.

The definition of the cost function devised in this chapter is as follows:

$$P(\vec{x}) := \sum_{i<j} w_{ij}(\vec{x}) \left(\frac{|\vec{x} - \vec{x}_i|^2}{|\vec{x} - \vec{x}_j|^2} - \frac{\sum_0^T s_j^2(t)}{\sum_0^T s_i^2(t)} \right)^2 + \sum_{n<m} w'_{mn} \left(\left(|\vec{x} - \vec{x}_m| - |\vec{x} - \vec{x}_n| \right) - \tau_{mn} c \right)^2 \quad (2.9)$$

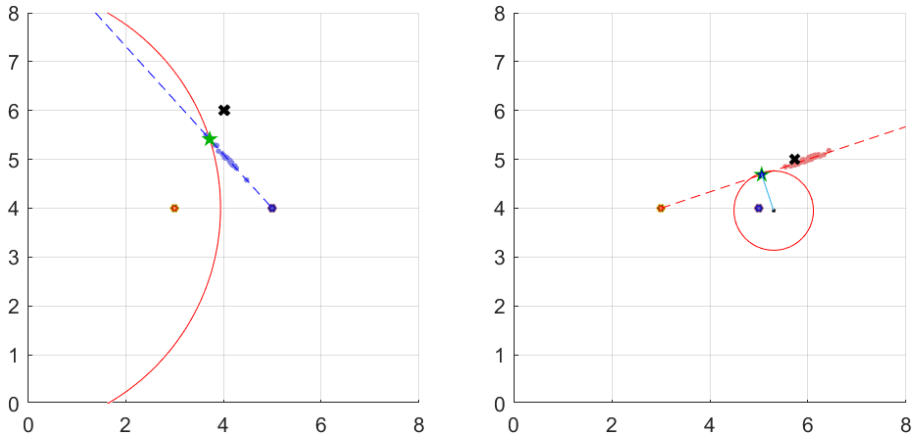
$\vec{x}$ is the input position in vector form. $\vec{x}_i$, $\vec{x}_j$, $\vec{x}_m$ and $\vec{x}_n$ are are position of i-th, j-th, m-th and n-th microphone respectively. s_i and s_j are the recorded signal of i-th and j-th microphone respectively. τ_{mn} is TDoA and c is the speed of sound. w_{ij} and w_{mn} are weighting factor for ILD method and weighting factor for TDoA method respectively.

Take a look at each term in (2.9). The first term $\frac{|\vec{x}-\vec{x}_i|^2}{|\vec{x}-\vec{x}_j|^2}$ in (2.9) is the ration between the distance between position $\vec{x}$ and microphone position $\vec{x_i}$ and $\vec{x_j}$. Take note that the energy of sound is inverse proportional to square of distance which was explained in 2.2.1. The $\frac{\sum_0^T s_j^2(t)}{\sum_0^T s_i^2(t)}$ is the ratio between the energy of recorded sound in j-th and i-th microphone. So $\big(\frac{|\vec{x}-\vec{x_i}|^2}{|\vec{x}-\vec{x_j}|^2} - \frac{\sum_0^T s_j^2(t)}{\sum_0^T s_i^2(t)}\big)^2$ in (2.9) is a term that reflects the accuracy of ILD results. If $\vec{x}$ is closer to the real sound source position, then this value gets smaller.

Next, take a look at the second term based on the TDoA method. $\big(|\vec{x}-\vec{x}_m| - |\vec{x}-\vec{x}_n|\big)$ in the second summation term of (2.9) is difference of distance between $\vec{x}$ and the position of i-th microphone and j-th microphone respectively. $\tau_{mn}c$ refers to the TDoA converted into distance difference by multiplying speed of sound c to TDoA τ_{mn}. That is, if the value of $\big(\big(|\vec{x}-\vec{x}_m| - |\vec{x}-\vec{x}_n|\big) - \tau_{mn}c\big)^2$ is low, that means that $\vec{x}$ is closer to the real source position. Summarizing the above, the value of the cost function newly defined in this section has a smaller value the closer the input $\vec{x}$ is to the actual sound source location. Therefore, the following expression holds:

$$\vec{x}_{\text{estimate}} = \operatorname{argmin}\big\{P(\vec{x})\big\} \tag{2.10}$$

Therefore, from the above equation, the microphone pair (i, j) or (m, n) with the smallest value of P is selected as the optimal microphone pair. After finding the optimal microphone pair, the localization result is derived by estimating the location of



(a) Source position $(4, 6)$. (b) Source position $(5.73, 5)$.

Figure 2.7: Cost function method acoustic localization result.

the sound source only with the corresponding optimal microphone pair.

2.4 Results and Discussion

For the localization result, the ILD result from the optimal microphone pair and AoA estimated from SRP-PHAT was combined. In the cost function, both ILD and TDoA were considered, but only the ILD results were chosen because the TDoA accuracy was low for gunshot recorded in an indoor environment with a lot of reverberation. Another reason for only using ILD result was that the individual microphone arrays were not synchronized. Microphones constituting an array are in sync, but individual arrays were out of sync. Therefore, only the ILD pair value was chosen as a result of cost function calculation.

The results of the acoustic localization method devised in this chapter were compared with the results of five existing acoustic localization methods: SRP-PHAT, inter-microphone time difference(ITD), generalized cross correlation - phase transformation (GCC-PHAT), fast least mean square (LMS), and adaptive eigenvalue decomposition (AEVD). The root mean square error (RMSE) values of each techniques is shown in

Table 2.1: RMSE of cost function and other acoustic localization methods (m)

Degree	Cost Function	SRP-PHAT	ITD	GCC-PHAT	fastLMS	AEVD
30	0.5019	1.2967	1.0718	5.8342	9.7481	38.0246
60	1.2034	1.3155	0.9262	30.2172	27.9031	18.1980
90	0.8171	1.5821	2.9323	3.5617	0.9874	7.3012
120	0.3014	1.9487	24.8606	0.9113	4.9519	8.0444
150	1.1370	2.3305	5.0913	28.2386	1.1063	8.2799

Table. 2.1. As shown in Table. 2.1, the RMSE of devised cost function method was smaller than the RMSE of other existing acoustic localization method. While RMSE was bigger than 1 m for SRP-PHAT method, the cost function method had smaller RMSE value of up to 1.2 m. Also the RMSE of devised cost function acoustic localization result was smallest compared to other acoustic localization methods. The graph of the acoustic localization result of each technique is shown in Fig. 2.7. In Fig. 2.7, the ILD result of the optimal microphone pair is drawn in red solid circle. The SRP-PHAT AoA result if shown in dotted blue line. The cross point of the localization result is marked with green star mark. In Fig. 2.7 (a) the microphone array 2 SRP-PHAT result was used which is marked with blue dots. In Fig. 2.7 (b) the microphone array 1 SRP-PHAT result is marked with red dots and red dotted line. Since range estimation accuracy is not high for SRP-PHAT method, AoA estimation results of two different microphone array was used. The localized point is marked with a green star. Considering that the existing sound source localization accuracy is not high when there is high reverberation, it can be said that the devised cost function acoustic localization method produces useful results even in an indoor environment with high reverberation.

2.5 Summary

In this chapter, a new localization method that can complement the existing acoustic localization method was introduced. Existing acoustic localization techniques vary greatly in their performance depending on the surrounding environment. The ILD based acoustic localization technique was inspired by the fact that the square of the distance and the sound energy coming into the two microphones are inversely proportional. However, this method greatly affected by reverberation and noise, and has a disadvantage in that the accuracy is lowered when the difference in the distance between the two microphones and the sound source is not large enough. The TDoA based acoustic localization technique calculates the cross correlation of the signals recorded with each microphone, and estimates the time index at which the cross correlation peak appears as TDoA. If this TDoA is converted into a distance, the distance difference between the sound source and each microphone can be obtained. However, there is a disadvantage in that the sampling frequency must be high to guarantee high accuracy. Finally, there is the SRP-PHAT method that first converts the time domain signal to the frequency domain, and then finds the cost function value for every point in space. Then the point having highest cost function value is chosen to be the estimated position. Since this technique converts a time domain signal into a frequency domain, it suppresses the effects of reverberation and noise, and has the advantage of high accuracy in a reverberation environment. However, there is a disadvantage in that the computation complexity is high because the value of the cost function for all directions or points must be obtained.

Therefore, in this chapter, a new acoustic localization method was devised to overcome the shortcomings of existing methods. New cost function different from SRP-PHAT was defined. This cost function consists of the sum of two terms. The first term corresponds to the difference between the ratio of the energy entering the two microphones obtained based on ILD and the ratio of the square of the distance between the microphones at the corresponding point. The closer the point is to the position of

the actual sound source, the smaller the first term value is. The second term is the difference between the distance difference between the two microphones and the corresponding point and the distance difference converted by dividing the ToA of the measured signal by the speed of sound. The second term also has a smaller value as the point is closer to the location of the actual sound source. Since the cost function is defined to have a smaller value as it has a true value, the pair of microphones whose cost function has the minimum value is chosen as an optimal microphone pair. As a result of estimating the location of the sound source using the optimal microphone pair found through the devised method, an error of within tens of centimeters occurred even in a general indoor seminar room environment with high reverberation, and the performance was better than that of existing methods.

Chapter 3

ACOUSTIC SIGNAL RECOVERY BASED ON SKETCHING AND STACKING WITH RANDOM FORK

3.1 Motivation

The performance and accuracy of acoustic signal processing techniques are affected by a number of factors: recording sampling rate, level of noise or reverberation, recording sound quality, etc. Among the many factors that affect performance, reverberation and noise can be reduced by using signal processing techniques such as noise suppression techniques. In the case of recording sound quality, one way to improve the record quality is to improve the performance or quality of the microphone used for recording. In other words, using an array of expensive, high-quality microphones will increase the quality of the recorded sound source, which is likely to further improve the acoustic signal processing performance. However, in reality, there is a limit in budget, so using expensive high-quality microphone arrays is not a good solution.

In addition to the type of microphone, what determines the quality of the recording sound source is whether the recorded sound source sample value is lost or corrupted. In fact, in the field of acoustics, a lossless sound source recorded at a high sampling rate is called free lossless audio codec (FLAC) format. The very existence of the term

FLAC shows that no loss of sample in recording is important in the field of acoustics and acoustic signal processing. However, the disadvantage of lossless recording file is that the size of a file is too large to handle, which also results in the increase of computation time. Therefore, most of the recording proceeds with the sampling rate lowered, resulting in sparse sampled recorded version. If we listen to this as music, in fact, our ears will not be able to detect whether some part of the recorded samples is lost if the loss is not severe. However, it is completely different from the point of view of acoustic signal processing. In most cases, the performance of the acoustic signal processing technique does not meet the expectations as the loss exists in the recorded sound. And the performance degradation is greater if degree of the loss in the recorded sound is greater. Because of this problem, research on acoustic signal processing techniques to restore lost or corrupted samples has been conducted.

Given that, this chapter introduces a technique designed to restore the lost sample value in the recorded acoustic signal when multiple sound sources are mixed and recorded by only one microphone. Let us assume the situation of everyday life. In everyday situations, there are sounds that occur abruptly such as human voices. On the other hand, there are also sounds that continuously spread in the background, such as ambient noise. An example of such ambient noise is a continuously generated machine sound. This machine sound is usually periodic and continuous of a certain frequency band. A sound with a certain frequency band is called single tone sinusoidal wave form. Therefore, it is assumed that there are several devices that produce continuous sound of such a constant frequency band in an indoor space.

In other words, the signal recovery technique introduced in this chapter assumes that several sinusoidal signals with different frequencies are mixed. Assume that some part of the sample is lost in this mixture of sinusoidal signals. This chapter introduces a technique to recover the lost sample values from the partially lost mixture of sinusoidal sound. A novel concept called random fork is introduced to recover a lost sample value from a signal samples that remains intact. This random fork is used to extract

information for restoration of a lost signal sample by extracting only a few samples among remaining samples. Therefore, this method is named to be SSRF. The rest of this chapter describes the SSRF technique.

3.2 SSRF Signal Model

3.2.1 Source Signal Model

As mentioned above, the SSRF technique assumes that there are k sources exist in a room. As a result, k sinusoidal signals with damping are superposed. And this mixed acoustic signal is recorded by only one microphone. The mixed sinusoidal sound source signal is expressed in continuous form as shown below:

$$v(t) = \sum_{i=0}^{k-1} V_l e^{-\gamma_l t} \cos\left(2\pi f_l t\right) \tag{3.1}$$

where $v(t)$ is acoustic signal, V_l is the peak amplitude, γ_l is damping factor and f_l is the frequency of l-th sinusoidal signal respectively. As shown in (3.1), the signal model is in the form of summation of k sinusoidal source signals. The exponential term represents damping of the signal.

3.2.2 Sampled Signal Model

Considering the sampling rate of microphone, the recorded or sampled signal $s(v)$ can be represented in discrete form as below:

$$\mathbf{s}(v) := \left\{ \sum_{l=0}^{k-1} V_l e^{-\gamma_l n\Delta t} \cos(2\pi f_l n\Delta t) \right\}_{n=0}^{P-1} \tag{3.2}$$

In (3.2), Δt is the sampling time interval. From (3.2), the sampled sequence of original

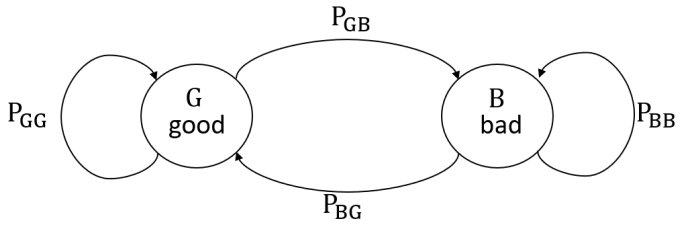


Figure 3.1: Two state Markov model for burst error.

signal with number of sample P can be expressed as a sequence having P elements.

Since only a small amount of information is required to restore a lost signal using SSRF technique, it is acceptable to set the sampling rate low. In other words, it means that SSRF based signal restoration is possible even for sparsely recorded acoustic samples.

3.2.3 Corrupted Signal Model

In many real-world environments, the recorded sound source signal has lost samples. This may be caused by a corruption of the recorded file, or it may be due to other reasons. Also, the loss of one sample affects adjacent samples in many cases. This case is often called burst error model.

The burst error model is a type of 2-state Markov chain model. In the burst error model, if the sample is not lost, then this is called a good state, which is represented as 'G'. For the opposite case when the sample is corrupted or lost, this is called a bad state 'B'. The two state Markov chain model is shown in Fig. 3.1. In Markov model, the probability is written in the form of transition matrix as below:

$$T = \begin{bmatrix} P_{GG} & P_{GB} \\ P_{BG} & P_{BB} \end{bmatrix} \tag{3.3}$$

P_{GG} is the probability of G going to G in the next state, which means that both adjacent

samples are not lost. P_{GB} is the probability of G going to B in the next state. In other words, P_{GB} is the probability that the next sample gets lost when the previous sample has been preserved. P_{BG} is the probability that the next sample doesn't get lost even if the previous sample has been lost. Likewise, P_{BB} is the probability of one sample getting lost when previous sample was also lost. The corrupted sequence according to burst error model is written in $\mathbf{c}$.

3.3 SSRF Problem Statement

Given the aforementioned system model, the SSRF problem of restoring the lost sample values can be simplified and defined mathematically as below:

- Recover the original signal $\mathbf{s}$ from the recorded signal sample with loss which is represented as $\mathbf{c}$.
- This is equivalent to finding amplitude V_l, damping factor γ_l and frequency f_l for all l.
- The source number k is known *a priori*.

This is a mathematical problem of extracting minimal information from the lossy signal $\mathbf{c}$. By using random fork and then estimating the value of the lost samples from the extracted information, one can recover the lost sample values. Therefore, the remainder of this chapter explains techniques for solving the above-mentioned mathematical problems based on SSRF technique.

3.4 SSRF Methodology

As mentioned above, the SSRF based acoustic signal restoration technique is capable of signal restoration even for sparse samples. In other words, SSRF based restoration technique is based on the hypotheses that if $\mathbf{s}$ is essentially a small amount of information, i.e. k is small, then the original signal $\mathbf{s}$ can be perfectly reconstructed by

extracting the intrinsic information of $\mathbf{s}$ from $\mathbf{c}$ using random fork. To support this hypotheses, we introduce the new concept, namely random fork which is written as ψ_m.

3.4.1 Geometric Sequential Representation

Take note that almost all sound generated indoor is has damping due to loss in the process of diffraction, scattering, or reflection. To reflect this damping, we used the exponential term when defining the signal. As the first process to solve the SSRF signal reconstruction problem, we focus on Euler's equation that all sinusoidal terms can be converted into exponential terms. Euler's equation is expressed as below:

$$\exp(ix) = \cos(x) + i\sin(x) \tag{3.4}$$

Since we previously defined the sinusoidal signal in the form of cosine when defining the signal, we can transform the cosine term exponentially by arranging the above equation. From Euler's equation, signal can be written as below where A and θ are variables:

$$A\cos\theta = \frac{A}{2}(e^{i\theta} + e^{-i\theta}) \tag{3.5}$$

Then in (3.5), substitute $\alpha_l = \frac{V_l}{2}$ and $\beta_l = e^{-\gamma_l \Delta t + i2\pi f_l \Delta t}$ for all l. As a result, the signal $\mathbf{s}$ can be rewritten as below:

$$\mathbf{s} := \sum_{l=1}^{k-1} \alpha_l \{\beta_l^n + (\beta_l^*)^n\}_{n=0}^{P-1} \tag{3.6}$$

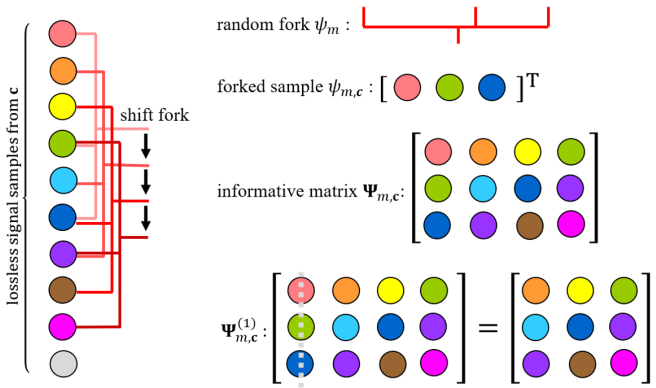


Figure 3.2: Concepts of random fork, forked sample and informative matrix.

The $(\cdot)^*$ operator means complex conjugate. As seen in the above equation, It can be seen that the number of unknowns that we need to find as a solution was three before, but has been reduced to two now. The mathematically written SSRF problem stated in previous section is reformed and simplified, and can be written as below:

- Calculate α_l and β_l for all l.
- Corrupted sample $\mathbf{c}$ and the number of acoustic sources k are *a priori*.

Therefore, from now on, we will discuss how to find the values of α_l and β_l. In this part, one more assumption is added to enable signal restoration. It is assumed that alpha and beta, which are unknown variables that we need to find, are an geometric sequence respectively. The information extracted from the random fork is used to obtain the α_l and β_l values, and the random fork will be defined mathematically in the next section.

3.4.2 Definition of Random Fork

As mentioned above, a concept of random fork $\psi_m \in \mathbb{N}^m$ is introduced in this section. Random fork ψ_m is defined as arbitrarily collected and lexicographically ordered m indices. In other words, ψ_m is a vector of m positive integers arranged in ascending

order. For example, $[1\ 3\ 5\ 9]^T$ is an possible example of fork ψ_m consisting of four indices. Unlike the example mentioned above, $[3\ 1\ 5\ 9]^T$ cannot be a random fork because the elements are not arranged in ascending order. ψ_m is used to extract m number of samples from the corrupted sample $\mathbf{c}$, and since it serves to pick up only a few samples like a fork, we named it a random fork. The mathematical representation of random fork is as below:

$$\psi_{m,\mathbf{c}} := \big(\mathbf{c}\big[\psi_m[0]\big], \ldots, \mathbf{c}\big[\psi_m[m-1]\big]\big)^T \tag{3.7}$$

By using random fork ψ_m, we can extract m number samples from corrupted sample $\mathbf{c}$. This forked sampled is $\psi_{m,\mathbf{c}} \in \mathbb{R}^{m\times 1}$ and its definition is in (3.7).

3.4.3 Informative Matrix

In this section, we discuss the small amount of information extracted using the random fork defined in the earlier section. This is names informative matrix. If only a few samples among sparsely recorded samples are extracted using random fork ψ_m, the sample value can be restored using the technique introduced in this chapter.

The informative matrix is written as $\mathbf{\Psi}_{m,\mathbf{c}}$ in bold. By stacking $\psi_{m,\mathbf{c}}$ defined in the previous section, we define informative matrices $\mathbf{\Psi}_{m,\mathbf{c}} \in \mathbb{R}^{m\times(m+1)}$ in (3.8).

$$\begin{aligned}\mathbf{\Psi}_{m,\mathbf{c}} := \big[\, \mathbf{c}(\psi_m) \mid \mathbf{c}(\mathbf{1}_m + \psi_m) \mid \mathbf{c}(2\mathbf{1}_m + \psi_m) \mid \cdots \\ \mid \mathbf{c}((m-1)\mathbf{1}_m\psi_m) \mid \mathbf{c}(m\mathbf{1}_m\psi_m) \,\big]\end{aligned} \tag{3.8}$$

In the above equation, $\mathbf{1}_m$ is the one-vector consisting of m number of ones, and $[\,\cdot \mid \cdot\,]$ is column matrix stacking operator. In other words, (3.8) means to stack the sample while shifting the fork one by one. If we define informative matrices $\mathbf{\Psi}_{m,\mathbf{c}}$ like

this, it becomes a rectangular matrix with m number of rows and $m+1$ number of columns. The reason why $\mathbf{\Psi}_{m,\mathbf{c}}$ is named informative matrix is that we will extract the information we need from $\psi_{m,\mathbf{c}}$ in the subsequent process. The concept of random fork and informative matrix is shown in Fig. 3.2.

3.4.4 Data Augmentation

Finally, before getting solutions for the SSRF problem, we define a column-collecting matrix that collects all columns except the j-th column. As mentioned in the previous section, the existing informative matrix was a rectangular matrix with one more column than a row. However, if you get only one matrix except for one column through column collecting, you get a square matrix as a result. We write this column-collecting symbol as $\Psi_{m,\mathbf{c}}^{(j)} \in \mathbb{R}^{m\times m}$. The definition is in (3.9):

$$\Psi_{m,\mathbf{c}}^{(j)} := \Psi_{m,\mathbf{c}}\mathbf{\Phi}_{m}^{(j)} \tag{3.9}$$

Let us take a look at example. The matrix below is an example of a column-collecting matrix that obtains a square matrix from a rectangular matrix such as an informative matrix. The matrix below is in the form of a square matrix in which all the values in the first row are zero, and in all other rows except for the first row, only the diagonal value is one and the rest are zero.

$$\mathbf{\Phi}_{m}^{(1)} := \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.10}$$

If the column-collecting matrix in the example above is multiplied by the informative

matrix, a square matrix consisting of the remaining columns except for the first column in the rectangular matrix is obtained.

(3.9) is calculated by multiplying the original informative matrix by $\mathbf{\Phi}_m^{(j)}$. Likewise, $\mathbf{\Phi}_m^{(j)}$ is a matrix consisting of zeros and ones, and serves to collect all columns except for the j-th column. As will be mentioned in a later section, the process of deliberately creating a square matrix is mainly intended to obtain the determinant from the square matrix. Now that the extraction of various information for solving the problem is completed, in the next section, we will solve the SSRF mathematical problem.

3.4.5 Solution of SSRF Problem

As mentioned in the previous section, the SSRF problem has been transformed into a problem of finding α_l and β_l for all l. This happened because signal format which was previously a multiplication of cosine and exponential terms changed into exponential term by using the Euler equation. In this section, we finally get the solution of the SSRF based acoustic signal restoration problem. To solve this problem, the following Theorem is applied.

Theorem 1. *If m is set to be $2k$ then $\{\beta_1, \ldots, \beta_k, \beta_1^*, \ldots, \beta_k^*\}$ is equal to the roots of the following polynomial equation.*

$$p(\beta) := \sum_{j=0}^{2k} c_j (-\beta)^j \tag{3.11}$$

$$c_j = \frac{\det\left(\mathbf{\Psi}_{2k,\mathbf{c}}^{(j)}\right)}{\det\left(\mathbf{\Psi}_{2k,\mathbf{c}}^{(0)}\right)} \tag{3.12}$$

$p(\beta)$ defined in (3.11) is the polynomial of variable β, and c_j term represents the

coefficient of the polynomial which is shown in (3.12). From now on, we prove the aforementioned theorem, and this will lead to the solution of the problem. The proof is as below:

Proof. To prove, let us write $\{\beta_1^*, \ldots, \beta_k^*\}$ to be $\{\beta_{k+1}, \ldots, \beta_{2k}\}$ for simplification. Then, we can decompose $\mathbf{\Psi}_{m,\mathbf{c}}$ into sub-matrices as (3.13). Matrix decomposition is a necessary process because it enables inverse operation to find a solution in the future. The decomposition is as below:

$$\mathbf{\Psi}_{m,\mathbf{c}} = \mathbf{U}\mathbf{\Sigma}_\alpha \mathbf{\Sigma}_\beta \mathbf{V}^T \tag{3.13}$$

The dimension of the matrices are $\mathbf{U} \in \mathbb{R}^{2k\times 2k}$, $\mathbf{V} \in \mathbb{R}^{(2k+1)\times 2k}$, $\mathbf{\Sigma}_\alpha \in \mathbb{R}^{2k\times 2k}$, and $\mathbf{\Sigma}_\beta \in \mathbb{R}^{2k\times 2k}$. The complete form of $\mathbf{U}$, $\mathbf{V}$, $\mathbf{\Sigma}_\alpha$ and $\mathbf{\Sigma}_\beta$ in (3.13) are shown in (3.14), (3.15), (3.16) and (3.17) respectively. In (3.14), ψ_{2k} has been replaced with ψ for simplicity.

$$\mathbf{U} = \begin{bmatrix} \beta_1^{\psi[0]} & \cdots & \beta_{2k}^{\psi[0]} \\ \vdots & \ddots & \vdots \\ \beta_1^{\psi[2k-1]} & \cdots & \beta_{2k}^{\psi[2k-1]} \end{bmatrix} \tag{3.14}$$

$$\mathbf{V} = \begin{bmatrix} \beta_1^0 & \cdots & \beta_1^{2k+1} \\ \vdots & \ddots & \vdots \\ \beta_{2k}^0 & \cdots & \beta_{2k}^{2k+1} \end{bmatrix} \tag{3.15}$$

$$\mathbf{\Sigma}_\alpha = \mathrm{diag}(\alpha_1, \ldots, \alpha_k, \alpha_1, \ldots, \alpha_k) \tag{3.16}$$

$$\mathbf{\Sigma}_\beta = \mathrm{diag}(\beta_1, \ldots, \beta_{2k}) \tag{3.17}$$

Since the rank of all sub-matrices $\mathbf{\Psi}^{(j)}_{m,\mathbf{c}}$ is equal to $2k$, c_j in (3.11) can be simplified as below:

$$\begin{aligned} c_j &= \frac{\det\left(\mathbf{\Psi}^{(j)}_{2k,\mathbf{c}}\right)}{det\left(\mathbf{\Psi}^{(0)}_{2k,\mathbf{c}}\right)} \\ &= \frac{\det\left(\mathbf{U}\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_\beta\mathbf{V}^T\mathbf{\Phi}^{(j)}_{2k}\right)}{\det\left(\mathbf{U}\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_\beta\mathbf{V}^T\mathbf{\Phi}^{(0)}_{2k}\right)} \\ &= \frac{\det\left(\mathbf{V}^T\mathbf{\Phi}^{j}_{2k}\right)}{\det\left(\mathbf{V}^T\mathbf{\Phi}^{0}_{2k}\right)} \\ &= \frac{\det\left(\mathbf{V}_j\right)}{\det\left(\mathbf{V}_0\right)} \end{aligned} \tag{3.18}$$

For simplification, write $\mathbf{V}^T\mathbf{\Phi}^j_{2k}$ as $\mathbf{V}_j$. That is, it can be seen that $\det\left(\mathbf{V}_j\right)$ need to be obtained from (3.18) to obtain c_j. From now on, we are going to apply the factor theorem to calculate the value of $\det\left(\mathbf{V}_j\right)$.

In $\mathbf{V}_j$ for any j, let us switch arbitrary β value, i.e. β_m and β_n where $m \neq n$. Switching the values of β_m and β_n is equivalent to switching two columns in $\mathbf{V}_j$. In this process $\det \mathbf{V}_j$ becomes zero, which means that the determinant of V_j has $(\beta_m - \beta_n)$ as a factor. If this process is repeated for all combinations of m and n, $\det\left(\mathbf{V}_j\right)$ has $(\beta_m - \beta_n)$ as a factor for all possible combinations of m and n. The operation of switching rows or columns in a matrix changes the sign of the determinant each time it is performed, but the sign of the determinant does not change for the case of $\det(\mathbf{V}_j)$ because m was set to be $2k$, that is, an even number.

From the above property, the determinant of $\mathbf{V}_j$ can eventually be written as the following expression by factor theorem:

$$\det\left(\mathbf{V}_j\right) = \prod_{n \le n < m \le 2k} \left(\beta_m - \beta_n\right) \sum_{1 \le i_1 < \cdots < i_l \le 2k} \left(\prod_{n=1}^{j} \beta_{i_n}\right) \tag{3.19}$$

Therefore, after reducing the common factors of the numerator and denominator in (3.12), the values of $\{c_j\}_{j=0}^{2k}$ can be finally written as follows:

$$c_j \in \left\{1, \ldots, \sum_{1 \le i_1 < \cdots < i_l \le L-1} \left(\prod_{n=1}^{l} \beta_{i_n}\right), \ldots, \prod_{n=1}^{2k} \beta_n\right\} \tag{3.20}$$

The above set in (3.20) is coefficients of polynomial whose roots are $\{-\beta_1, \ldots, -\beta_{2k}\}$. That is, if we find the solution of the equation of the polynomial equals zero, i.e. $p(\beta) = 0$, it means that the solution of the equation becomes the β_n value we wanted to find.

The solution for α_k is trivial. After obtaining β_k values, we can inversely calculate $\{\alpha_1, \ldots, \alpha_k\}$ values by solving simple linear equations. Through the above series of processes, we obtained the desired values of α_k and β_k. By substituting this obtained value into the exponential expression of $\mathbf{c}$, the lost sample value can also be obtained.

□

As above, the acoustic signal recovery process using SSRF has been completed. In the next section, we will analyze and discuss the acoustic signal restoration results using this technique. The algorithm of SSRF is summarized in Algorithm 1.

Algorithm 1 : Procedure for SSRF.

if corrupted sequence $\mathbf{c}$ meets **Condition then**
i) Select a ψ_{2k} satisfying Condition 1.
ii) Re-sample $\mathbf{c}$ using ψ_{2k} and construct $\Psi_{2k,\mathbf{c}}$.
iii) Extract $\{\Psi_{2k,\mathbf{c}}^{(j)}\}_{j=0}^{2k}$.
iv) Solve equation to obtain $\{\beta_l\}_{l=1}^{2k}$.
v) Extract $\{\alpha_l\}_{l=1}^{2k}$ by substitute $\{\beta_l\}_{l=1}^{2k}$.
vi) Recover $\mathbf{s}$ by obtained $\{\alpha_l\}_{l=1}^{2k}$ and $\{\beta_l\}_{l=1}^{2k}$.
else
i) Implement cubic interpolation.
end if

3.4.6 Reconstruction of Corrupted Samples

Note that the following condition should be met to obtain $\{\alpha_1, \cdots, \alpha_k\}$ and $\{\beta_1, \cdots, \beta_k\}$.

Condition 1. *Given* $\mathbf{c}$*, if there exist at least one* ψ_{2k} *satisfying the following relationship, then* $\mathbf{s}$ *can be perfectly retrieved by only using* $\mathbf{c}$*.*

$$\bigcup_{n=0}^{2k} \{\mathbf{s}(n\mathbf{1}_{2k} + \psi_{2k})\} \subseteq \mathbf{c}, \tag{3.21}$$

where $\{\cdot\}$ *is the operator making a set consisting of all elements of an input.*

3.5 Performance Analysis

In this section, we evaluate the performance of the proposed SSRF solution in terms of mean squared errors (MSE) of reconstruction.

3.5.1 Simulation Set-up

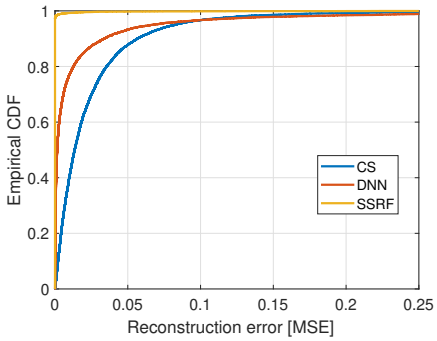
For a performance comparison, we choose the algorithms of compressive sensing (CS) and deep-neural network (DNN). The reconstruction method based on CS is implemented by the orthogonal matching pursuit (OMP) algorithm with the partial discrete cosine transform (DCT) matrix. The number of bases in the CS based method is 5000.

Next, the DNN based reconstruction method is utilized, where the number of training data set is 30000. In addition, the number of hidden layer and the number of perceptrons in each layer are 2 and 40 respectively. The sigmoid activation function is considered in neural network model. In addition, MSE is selected as the loss function in the DNN algorithm. For optimization, scaled conjugate gradient method is applied.

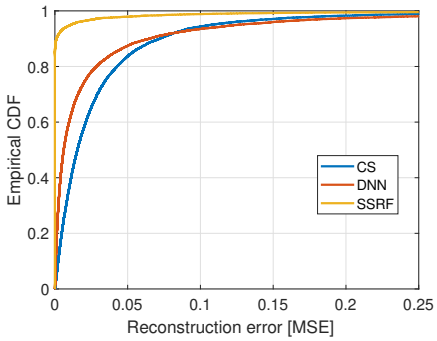
The simulation results are based on Monte Carlo simulation experiments with 10000 cases. The parameter settings for performance comparison is as follows. The peak amplitude V_l follows the Normal distribution with zero mean and $1/\sqrt{k}$ as variance. The damping factor γ_l follows uniform distribution $\mathcal{U}(0, 10^3)$. In addition, the frequency follows uniform distribution with $\mathcal{U}(0, 10\text{ kHz})$. The sampling time interval Δt is set to 0.5×10^{-4} second, and the number of samples P is set to be 30.

3.5.2 Reconstruction Error According to Bernoulli Parameter and Number of Signals

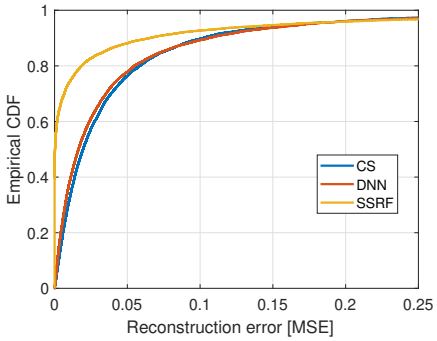
Fig. 3.3 shows the cumulative distribution function (CDF) of reconstruction error of each algorithm according to different k and p. The representative challenge of that CS method is that it requires large number of bases due to the generation of f_l and γ_l in the continuous domain, i.e. there is no guarantee of the orthogonality among superposed signals. In addition, the shortcomings of DNN based signal reconstruction is that it requires an enormous training data set for statistical inference because of the binary corruption affecting the samples. Here, we remark that the SSRF method is superior to CS and DNN even though there is no dictionary of bases and training data sets. For all cases of k and p value, the SSRF method shows less reconstruction error than CS and DNN algorithms. In addition, if condition 1 is satisfied, the SSRF method result in the exact signal recovery. As shown in Fig. 3.3 (a) and (d), it is interesting that the SSRF technique shows the capability of signal reconstruction close to perfection in an environment with high probability of satisfying condition 1.



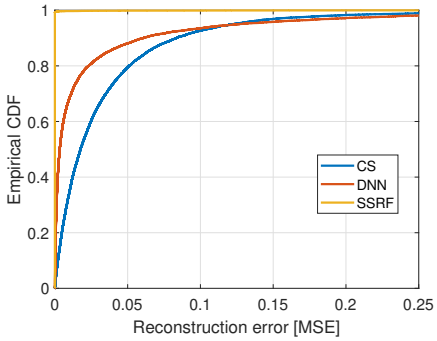
(a) $k = 3$, $p = 0.05$.



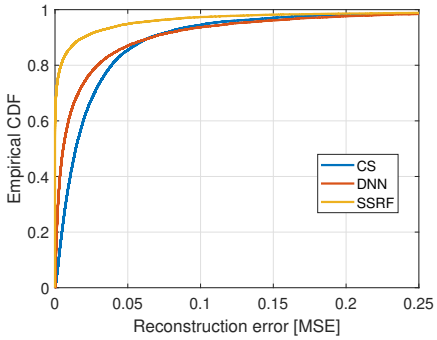
(b) $k = 3$, $p = 0.1$.



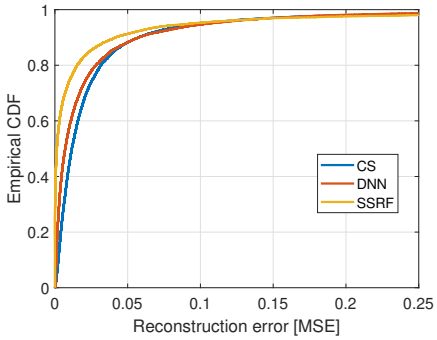
(c) $k = 3$, $p = 0.2$.



(d) $k = 2$, $p = 0.1$.

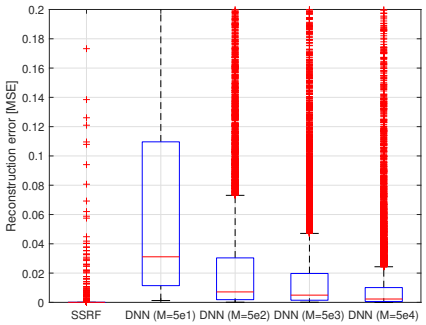


(e) $k = 4$, $p = 0.1$.

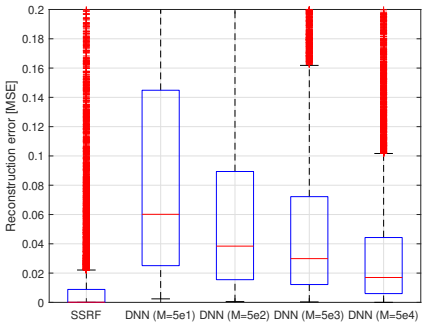


(f) $k = 6$, $p = 0.1$.

Figure 3.3: Reconstruction error according to k and p.

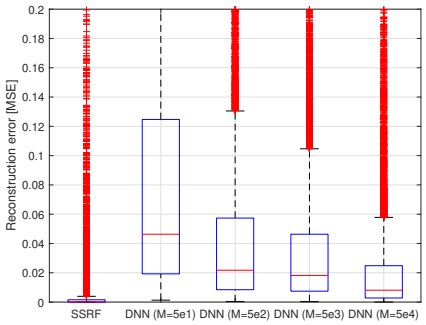


(a) $p = 0.05$.

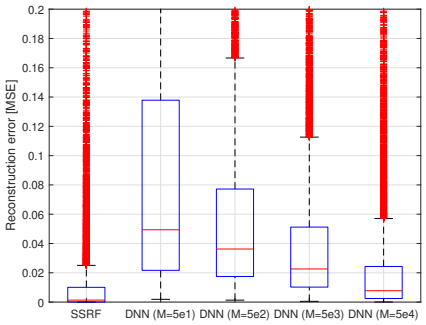


(b) $p = 0.2$.

Figure 3.4: Comparison of reconstruction error with DNNs according to M ($k = 3$).



(a) $k = 4$.

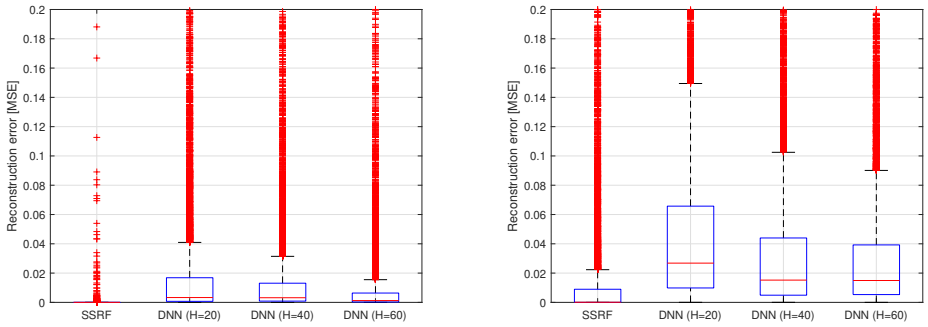


(b) $k = 6$.

Figure 3.5: Comparison of reconstruction error with DNNs according to M ($p = 0.1$).

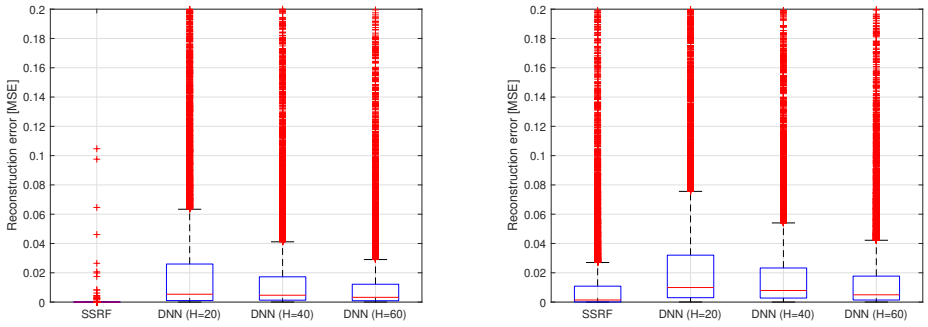
3.5.3 Detailed Comparison between SSRF and DNN

More intensive comparison between SSRF reconstruction and DNN methods according to the number of training data (M) is shown in Fig. 3.4 and 3.5, where the number of perceptrons in each layer (H) is fixed to 40. To describe the statistical aspects, the box plot is utilized. In each figure, the top and bottom line of the blue box represents the first and third quartiles. In addition, the red horizontal line in the box represents the median value. The dotted vertical line represents the range of data excluding outliers. For all sub-figures in Fig. 3.4 and 3.5, as the number of training data gets bigger, the median value of the reconstruction error gets smaller. With more training data, the



(a) $p = 0.05$ (b) $p = 0.2$.

Figure 3.6: Comparison of reconstruction error with DNNs according to H ($k = 3$).



(a) $k = 4$. (b) $k = 6$.

Figure 3.7: Comparison of reconstruction error with DNNs according to H ($p = 0.1$).

accuracy of DNN increases. However, given the extremely low reconstruction error of SSRF as seen in Fig. 3.4 and 3.5, increasing M cannot excel the SSRF performance. With same k value in Fig. 3.4 (a) and (b), reconstruction error increases as p gets larger both for SSRF and DNN based reconstruction. Then for a fixed p as in Fig. 3.5 (a) and (b), reconstruction error gets slightly higher as k gets bigger.

Next, comparison between SSRF and DNN methods for reconstruction according to number of perceptrons in each layer (H) is shown in Fig. 3.6 and 3.7, where the number of training data set is fixed to 3×10^4. As shown in Fig. 3.6 and 3.7, reconstruction error decreases as the number of perceptrons increases. However, the increase

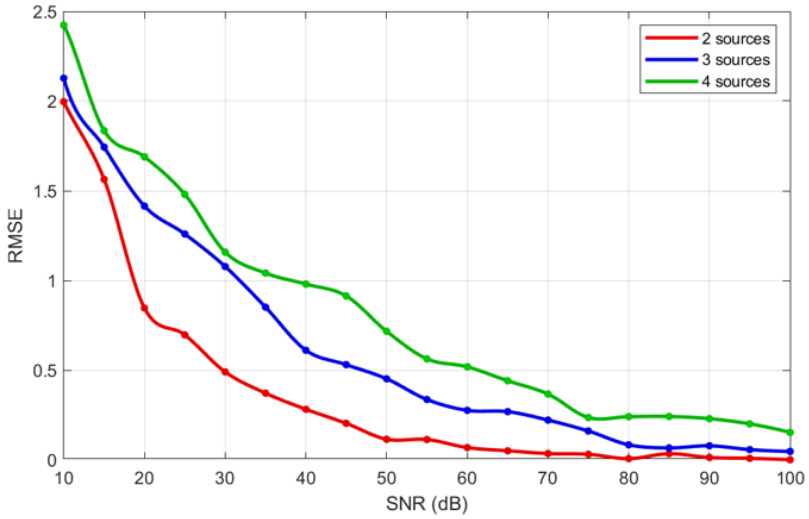


Figure 3.8: SNR-RMSE of SSRF result for $k = 2, 3$ and 4.

in the number of perceptrons does not dramatically decrease the reconstruction error. Even in the case of Fig. 3.6 (b), DNN with $H = 40$ and $H = 60$ show almost similar performance in terms of each median, the first and third quartiles. As a result, SSRF results in higher performance in terms of the distribution of reconstruction error regardless of k, p and H.

3.5.4 SSRF Result for Signal with Additive White Gaussian Noise

In order to apply the SSRF technique to a more realistic environment, consider a signal with noise. The noise that best mimics the real environment is additive white Gaussian noise (AWGN). AGWN has same power distributed to all frequencies. This is the reason why this noise is called white. In this case, it is considered that AWGN was added during measurement, in other words, recording. RMSE of recovered sample value was calculated for different SNR. The relationship between SNR and RMSE is shown in Fig. 3.8. As shown in Fig. 3.8, the RMSE gets smaller as SNR gets larger. Also, as the number of sinusoidal signals k gets larger, RMSE also gets larger.

3.6 Summary

In this study, we conducted a study on how to recover the lost signal samples when it is sparsely sampled. We did not assume a simple situation with a single sound source, but assumed more complex situation where multiple sound sources produced sound simultaneously. In this study, mixture of sinusoidal signals with decay was a target. An example of such such sound is a continuously occurring machine sound.

The process of SSRF based signal recovery is as follows. First, in order to recover the lost signal sample, the signal defined as cosine, a sinusoidal term, was transformed into exponential term using Euler's formula for simplification. In the process, the number of unknown variables that used to be three was reduced to two. Then, from recorded signal, re-sampling using random fork was conducted. After that, the SSRF problem was solved under the assumption that each unknown variable is part of a geometric sequence.

In order to verify the performance of the devised technique, the performance was compared with the existing lost signal recovery technique, including compressive sensing, and deep learning based signal restoration technique. As a result, the SSRF based acoustic signal restoration technique newly devised in this study showed higher signal restoration accuracy than the existing technique with less computation.

The contribution of this study can be summarized as follows. First, unlike other acoustic signal processing techniques that target only a single source signal, signal recovery technique that can be applied to a mixture of multiple sources was devised. Second, it showed higher recovery performance with lower amount of computation than compressive sensing and deep learning based acoustic signal recovery techniques. Lastly, the advantage is that it can be applied to continuous and repetitive sounds of a certain frequency band that are always present in life.

Chapter 4

SINGLE CHANNEL ACOUSTIC SOURCE NUMBER ESTIMATION AND BLIND SOURCE SEPARATION BASED ON SKETCHING AND STACKING WITH RANDOM FORK

4.1 Motivation

As demand for acoustic signal processing applicable to various environments increases, the research demand for acoustic signal processing techniques is also rapidly increasing [27, 28]. However, most of the existing acoustic signal processing techniques target a single sound source [29, 30, 31, 32]. Unlike the anechoic chamber environment where the number and location of sound sources can be intentionally manipulated, it is difficult to know the sound source number in advance in the real environment. In order to apply existing acoustic signal processing technique for a single sound source to multiple sound source environment, it is essential to separate mixed multiple source signal into individual ones. The technique of estimating the number of sound sources and separating them into individual signals in a situation where there is almost no information is called BSS [33]. Acoustic BSS has been extensively studied since the past because it is a core technology that can expand the application of existing laboratory

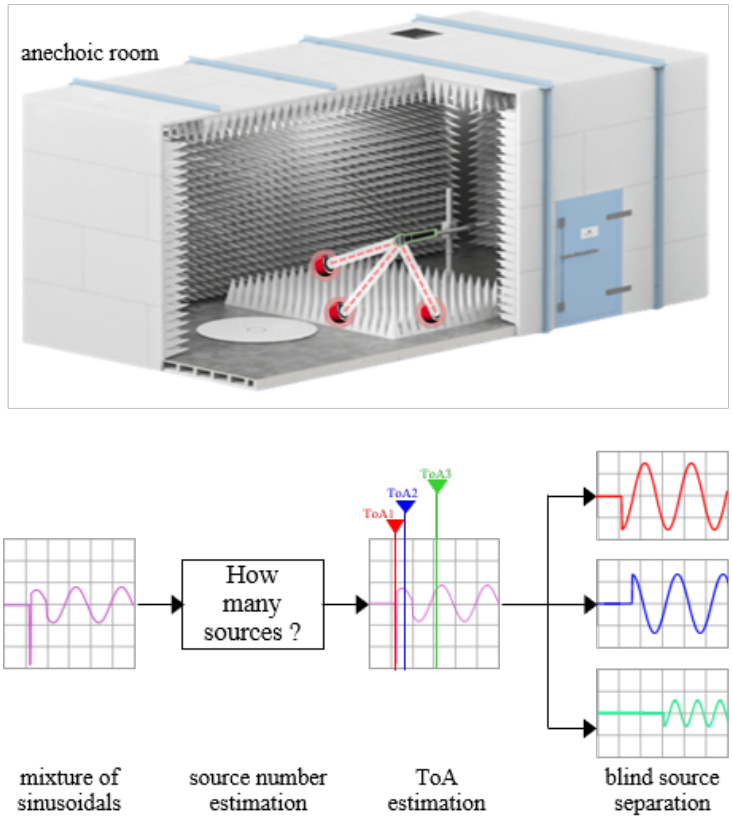


Figure 4.1: Concept of source number estimation and BSS based on SSRF.

based acoustic signal processing technique to the real world [34, 35, 36].

The BSS technique can be roughly divided into two types: single channel BSS where there is only one microphone [37], and multi channel BSS where multiple microphones record sound from one scene, e.g. microphone arrays [38]. In the case of using multiple microphones, a beamforming technique is widely used [16, 39]. In this case, the direction of arrival (DoA) can be extracted and then each signal is separated based on DoA information. [34], [35] and [40] are representative BSS techniques. In [34], a real-time acoustic BSS technique based on second-order statistics which is robust under noisy condition is introduced. This method has advantage in that it can be applied real time even under noisy condition. However, there is a disadvantage in

that computational complexity increases depending on the window length of the filter. [35] has combined frequency-domain independent component analysis (FDICA) and time-domain independent component analysis (TDICA) which is robust in reverberant condition. Considering that the multistage ICA based BSS is based on the ICA technique in the frequency domain, this technique has a problem in that the time resolution is lowered when frame length increases, and as a result, the BSS performance is lowered. That is, there is a trade-off between TDICA and FDICA, which means that there is a limit on improving BSS performance.

However, in the case of aforementioned multi channel BSS technique, there is a disadvantage that multiple microphones must be implemented to increase accuracy. The higher the number of microphones, the higher the accuracy, but inefficient in terms of cost. [36] and [41] are commonly well-known single channel BSS. [36] is the latest BSS technique that combines BSS with a neural network. There have been several attempts to perform blind decomposition by applying a deep neural network (DNN) to a single channel acoustic signal, but it has not been able to achieve automatic classification, and the standard for classification had to be based on the standard established by humans. [41] introduces an single channel acoustic BSS technique for sound detection of falls that occur frequently in the elderly. It is argued that the efficiency can be increased through the BSS process, which separates the ambient noise and the falling sound, because the fall sound mainly comes in mixed with the surrounding noise. [41] assumes a single channel situation, and non-negative matrix factorization (NMF) BSS technique is applied to separate interference noise and falling sound. However, distinguishing the target sound from the ambient noise is similar to the noise suppression technique which is far from BSS in the true sense.

Although BSS has a long history, most of the existing BSS methods assume that the number of source is known *a priori*. Also in the case of the ICA technique, which is the most widely used among BSS techniques, it is assumed that there should be no time delay in the individual mixed signals. In addition, most of the BSS is aims to

separate human voice which is not applicable for mixture of machine sounds. So, it would be meaningful to estimate the ToA of each sound source from the mixed sound signal. This is because estimated ToA can be applied to acoustic localization since by multiplying speed of sound to ToA, ToA turns into distance information. The accuracy of localization increases if the distance between the sound source and the microphone gets more accurate. Moreover, estimating ToA accurately can also improve the performance of BSS. However, there are few studies on the estimating ToA from the mixed signal in the field of acoustic signal processing. Existing studies related to ToA in acoustic signals mostly focus on how to increase the accuracy of ToA estimation for only single sound source [42], which is not applicable to mixed signals. In [43], ToA estimation method is introduced but it corresponds to the multichannel technique. That is, multiple microphones are used in [43] which is not cost-efficient. Although [44] conducts research related to multi source localization and tracking, [44] make use of the ToA estimation from the mote sensor, and has not developed technique for increasing the ToA estimation accuracy from multiple mixed source signals. Moreover, there is no research on accurate ToA estimation among existing studies on acoustic BSS techniques.

To compensate for these shortcomings, this chapter introduces an SSRF BSS which can accurately estimate ToA even for mixture of multiple sinusoidal signals with time delays. The conceptual diagram of the SSRF BSS is shown in Fig. 4.1. SSRF is a technique to recover the lost sample values in a mixture of several sinusoidal signals without time delay [45]. In this chapter, it is assumed that there is only one microphone in the anechoic chamber environment and multiple sinusoidal sound sources exist. SSRF which targets a mixture of sinusoidal signals has the advantage in that it can be applied to a sound with a continuous single tone such as a mechanical sound. It is assumed that the distance between the microphone and each sound source is different. In other words, the signals have time delay, ToA in this chapter. And it is assumed that there is a difference between ToAs' are greater than sample duration. That is, it

means that each signal is distinguishable under a certain sampling frequency. It is also assumed that a microphone and multiple sound sources in the anechoic chamber are synchronized. In other words, it means that the index extracted using this technique is ToA, not TDoA. Since TDoA is a relative arrival time difference, it does not mean the absolute distance from the microphone, whereas the absolute distance of the sound source away from the microphone is known from the ToA.

4.2 SSRF based BSS System Model

In the previous chapter, detailed descriptions and introductions of SSRF techniques were provided. In the previous section, SSRF technique was for recovery of the lost signal samples. However, in this chapter, we apply SSRF technique for separating mixed signal when number of sources are unknown. In the field of BSS, it is called under-determined BSS problem, which means that there is no information about the source number. From now on, I will discuss the system models in this study. Detailed explanations on the simulation scenarios and assumptions are followed.

4.2.1 Simulation Scenarios

It is assumed that up to ten sources exist in an anechoic chamber. Although the number of sound sources is limited to a maximum of 10 in the simulation, it is also possible to increase the maximum number of sound sources. The anechoic chamber is surrounded by sound absorbing material, so it is an ideal environment where there is almost no reverberation. There is one microphone in the anechoic chamber. That is, it corresponds to the case of single channel BSS. Also, it is assumed that one microphone and multiple sound sources are all connected to one array and synchronized. From this, it can be said that the time difference we obtain is ToA, not TDoA. While TDoA means the difference in arrival time between two sound sources, ToA means accurate arrival time, so the range between each source from the microphone can be accurately calculated

from ToA. The simulation scenario setting is shown in Fig. 4.2.

Additionally, it is assumed that the distances between the microphone and each sound source are all different. Let us say that are k sound sources in the anechoic chamber and the distance between each sound source and the microphone is d_k. Then for i and j where $i \neq j$, $d_i \neq d_j$ for all i and j. To be more precise, if the sampling rate at which the microphone records sound is f_s and the speed of sound is c, the distance difference of i-th and j-th sound source $|d_i - d_j| > c/f_s (i \neq j)$ must be satisfied to accurately obtain the ToA of each sound source.

The sampling frequency was set to $f_s = 44100$ Hz. This is because the frequency is widely used as a recording frequency as a standard for music work. The maximum value of the number of sound sources was fixed at 10, and the distance from each sound source to the microphone was set randomly. This is shown in Fig. 4.4. In Fig. 4.4 (a), the sound recorded from ten sound sources separated by a random distance is shown. Each sound source generates a continuous sinusoidal signal. In Fig. 4.4 (b), the recorded signal sampled at a sampling rate of f_s is shown.

4.2.2 System Model

As mentioned before, in this chapter, it is assumed that the mixed signals generated from k multiple sound sources enter one microphone. In this case, let the time delay caused by the distance between each sound source and the microphone be τ_i. The signal collected by the microphone can then be written as (4.1):

$$\mathbf{s}(t) = \sum_{i=1}^{k} s_i(t - \tau_i) \tag{4.1}$$

Each signal of a i-th single sound source $s_i(t)$ is sinusoidal with a decay as mentioned in precious section. If the time delay of each signal source is τ, the signal expressed in continuous sinusoidal form can be written as:

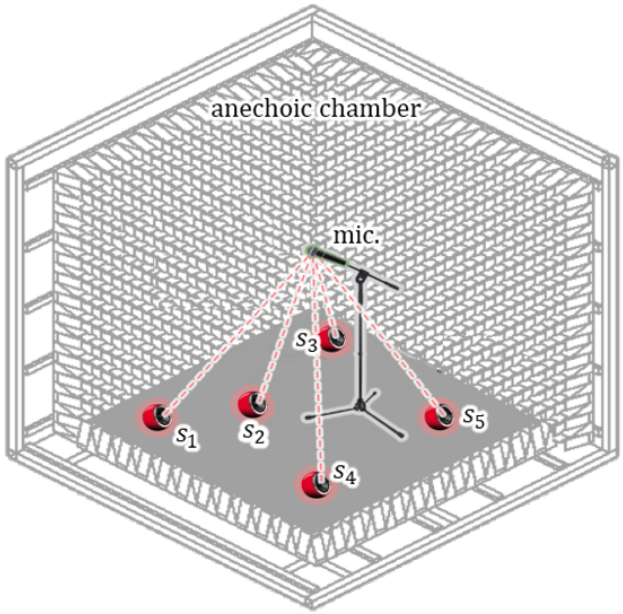


Figure 4.2: Simulation Scenario Setting.

$$v(t) = \sum_{l=0}^{k-1} V_l e^{-\gamma_l t} \cos\left(2\pi f_l (t - \tau_l)\right) \tag{4.2}$$

The SSRF technique can also be applied to sinusoidal signals out of phase. This means that the phase between each sinusoidal signals are different. For this the phase term is added in (4.3)

$$v(t) = \sum_{l=0}^{k-1} V_l e^{-\gamma_l t} \cos\left(2\pi f_l (t - \tau_l + \theta_l)\right) \tag{4.3}$$

θ_l in (4.3) represents phase of l-th sinusoidal signal.

Then, the sampled version of the above equation can be written as below:

$$\mathbf{s}(v) := \left\{ \sum_{l=0}^{k-1} V_l e^{-\gamma_l (n-\delta_l)\Delta t} \cos\left(2\pi f_l \{(n-\delta_l)\Delta t + \theta_l\}\right) \right\}_{n=0}^{P-1} \tag{4.4}$$

δ_l is the delay index obtained by dividing τ_l with f_s.

According to the Euler equation introduced in the previous chapter, the cosine can be rewritten into exponential form. The exponential transformation of (4.4) can be written as below:

$$\begin{aligned} \mathbf{s}(v) &= \left\{ \sum_{l=0}^{k-1} \frac{1}{2} V_l \left\{ e^{i2\pi f_l \{(n-\delta_l)\Delta t + \theta_l\}} + e^{-i2\pi f_l \{(n-\delta_l)\Delta t + \theta_l\}} \right\} \right\}_{n=0}^{P-1} \\ &= \left\{ \sum_{l=0}^{k-1} \frac{1}{2} V_l \left\{ e^{i\theta_l} \cdot (\beta_{l,\delta})^n + e^{-i\theta_l} \cdot (\beta_{l,\delta}^*)^n \right\} \right\}_{n=0}^{P-1} \end{aligned} \tag{4.5}$$

Take note that $e^{-i\theta_l}$ and $e^{i\theta_l}$ are complex conjugate to each other. Also $e^{-i\theta_l} \times e^{i\theta_l} = 1$ which means that the two terms are reciprocal to each other. For α_l, which was previously defined as $\alpha_l = V_l/2$, it changes into $\alpha_{l,\theta} = (V_l/2)\mathrm{e}^{i\theta_l}$ this time. As a result, (4.5) can be rewritten as below:

$$\mathbf{s}(v) = \left\{ \sum_{l=0}^{k-1} \left\{ \alpha_{l,\theta} \cdot (\beta_{l,\delta})^n + \alpha_{l,\theta}^* \cdot (\beta_{l,\delta}^*)^n \right\} \right\}_{n=0}^{P-1} \tag{4.6}$$

Remember, in the previous section, we introduced α_l and β_l to convert expressions to exponential form. So the time delay term is added in the previously defined β_l. The $\beta_{l,\delta}$ which was introduced in (4.5) is written as below:

$$\beta_{l,\delta} = e^{-\gamma_l(\Delta t - \delta_l) + i2\pi f_l(\Delta t - \delta_l)} \tag{4.7}$$

In the above expression, it can be seen that the τ term, which means time delay, has been added. Expanding the above expression for β, it can be arranged as follows:

$$\begin{aligned}
\beta_{l,\delta} &= e^{-\gamma_l(\Delta t-\delta_l)+i2\pi f_l(\Delta t-\delta_l)} \\
&= e^{\gamma_l\delta_l-i2\pi f_l\delta_l} \cdot e^{-\gamma_l\Delta t+i2\pi f_l\Delta t} \\
&= e^{(\gamma_l\delta_l-i2\pi f_l\delta_l)\Delta t/\Delta t} \cdot \beta_l \\
&= (\beta_l)^{\delta_l/\Delta t} \cdot \beta_l \\
&= (\beta_l)^{\delta_l f_s} \cdot \beta_l
\end{aligned} \tag{4.8}$$

Based on this, a mathematical approach to SSRF technique for signal separation is described in next section. Before signal separation, ToA extraction will be explained first in the next section.

4.3 SSRF based BSS Methodology

This section describes how to find the ToA from a mixed acoustic signal based on the SSRF technique introduced in 3.

4.3.1 Source Number and ToA Estimation based on SSRF

The entire SSRF technique is introduced in 3. In particular, technique of estimating source number and ToAs' from mixed sound signal pay attention to the function $p(\beta)$ in (3.11) and the coefficient c_j of β in (3.12) defined for signal restoration.

The key for estimating ToA is to find when the value of c_j changes. It is determined that the point at which the value of c_j suddenly changes is the time when the signal from the corresponding sound source arrives at the microphone. It is much more beneficial in terms of computational complexity and execution time to obtain only the

coefficient c_j than to apply the entire SSRF technique by solving k-th order equation.

In the simulator, the maximum number of sound sources is limited to ten, but this only sets the possible maximum value, and we start the simulation without knowing how many sound sources are in the anechoic chamber. Since we do not know how many sound sources exist in space, we start with two indices of fork, i.e. $k = 2$, in SSRF technique. And the interval between fork indices is fixed to one. In other words, while the interval between fork indices was random in the existing SSRF technique, this time, a fork with consecutive indices at intervals of one is used. This is to increase the ToA estimation accuracy. This is because if the interval between fork indices becomes a random number exceeding one, then the size of the fork increases, making it difficult to check the ToA index exactly at which another source sound reaches the microphone. This is explained in Fig. 4.3. If the interval between fork indices is greater than one and the entire fork width is widened, the length of the transition part becomes longer, making it unclear at which index the new signal was added. In Fig. 4.3 when only the sound from the first sound source reaches the microphone, it is called state 1. After that, the sound from the second sound source arrives at the microphone, and the state in which the signals of the two sound sources are mixed is called state 2. Consequently, the state in which the signal from the j-th sound source arrives at the microphone and j signals are mixed is called state j.

The algorithm for finding ToA is as follows. As mentioned above, first, the number k of indices in fork is fixed to $m = 2$. After that, starting from sample index 1, the value of c_1 is obtained while shifting fork one by one. c_1 means the coefficient of the

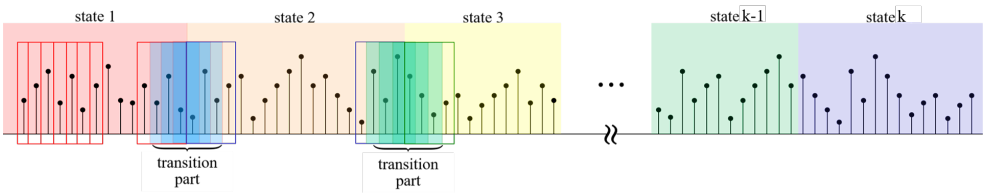


Figure 4.3: Concept of SSRF source number and ToA estimation.

Algorithm 2 BSS based on SSRF for mixed sinusoidal signal with time delay.

i) Set random fork ψ_{2k} with an interval of 1 and indices with $m = 4$.
while ψ_{2k} reaches end **do**
  ii) Starting with the first sample of the recorded signal, shift ψ_{2k} one by one and calculate c_1.
  iii) Find where c_1 value abruptly changes and set that index to be delay index.
  iv) Increase the fork indices by 2, in other words $m = m + 2$.
end while
v) The number of change in delay is same as source number k.
vi) For the signal where all source signals are mixed, solve the SSRF equation with the samples forked with ψ_{2k} where $m = 2k$.

first order term in the polynomial $p(\beta)$.

When a signal emitted from the sound source s_1, which is the closest sound source from the microphone, a sudden change will occur in the c_1 value. The index at which this sudden change appears is set as ToA of s_1. After that, the fork indices are increased to $m = 4$ to obtain the ToA of state2 where the signal is received from s_2. Since SSRF is applied to an even number of fork indices, increase m to an even number such as $m = 2, 4, 6, \ldots$. Then, as before, find the index at which a sudden burst of change appears in the value of c_1 and this is the ToA of s_2. In this process, the number of fork indices is increased to an even number, and ToA is repeatedly found one after another until the recorded signal ends. If the number of remaining signal samples is less than the number of fork indices, this process is terminated.

This process is summarized in algorithm 2. n refers to the number of sound sources which is unknown. The reason for $n \leq 10$ is that the maximum number of sound sources in the space is limited to 10. k, τ, j are the number of fork indices, ToA and source number index respectively.

Through the above algorithm, we can extract information about the number of sound sources and the distance each sound source is from the microphone in a situation where there is no information about the number of sound sources in the anechoic chamber. Unlike the existing SSRF technique, which solves the equation by calculating

the coefficients of all the terms of the polynomial $p(\beta)$, the devised method calculates only one coefficient without performing the process of solving the equation to extract the number of sound sources and ToA information. Therefore, there is an effect that the amount of calculation is reduced and the computation time is shortened.

4.3.2 Signal Separation

In this section, we solve the mathematical problem required to separate the signals by applying the SSRF technique when multiple signals coming in with time delays. In the previous section, $\mathbf{\Psi}_{mc}$ matrix decomposition was performed to solve the SSRF mathematical problem. Similarly, in order to obtain the component values of the signal to which the time delay is reflected, this is also written in matrix form and then matrix decomposition is performed.

As in (3.13), the $\mathbf{\Psi}_{mc}$ of the signal with delay can also be decomposed. The matrix decomposition of the mixed signal with time delay can be written as below:

$$\mathbf{\Psi}_{mc} = \mathbf{U}\mathbf{\Sigma}_{\alpha}\mathbf{\Sigma}_{\beta}\mathbf{\Theta}\mathbf{\Delta}\mathbf{V}^{T} \tag{4.9}$$

The definition of the matices $\mathbf{U}, \mathbf{\Sigma}_{\alpha}, \mathbf{\Sigma}_{\beta}$ and $\mathbf{V}$ are as same as the definition in (3.14), (3.15), (3.16) and (3.17) respectively. In this section, $\mathbf{\Theta}$ and $\mathbf{\Delta}$ is newly introduced. $\mathbf{\Theta}$ is the matrix that includes the phase term. $\mathbf{\Theta}$ can be written as below:

$$\mathbf{\Theta} = \mathrm{diag}(e^{i\theta_1}, e^{i\theta_2}, \ldots, e^{i\theta_k}, e^{-i\theta_1}, e^{-i\theta_2}, \ldots, e^{-i\theta_k}) \tag{4.10}$$

$\mathbf{\Delta}$ is a matrix that reflects the ToA of each sound source. As introduce in previous section, since the delayed term is expressed in the form of a power of β_l, the elements of the matrix can be written as below:

$$\mathbf{\Delta} = \mathrm{diag}(\beta_1^{\delta_1 f_s}, \ldots, \beta_k^{\delta_k f_s}, \beta_1^{\delta_1 f_s}, \ldots, \beta_k^{\delta_k f_s}) \tag{4.11}$$

written as below:

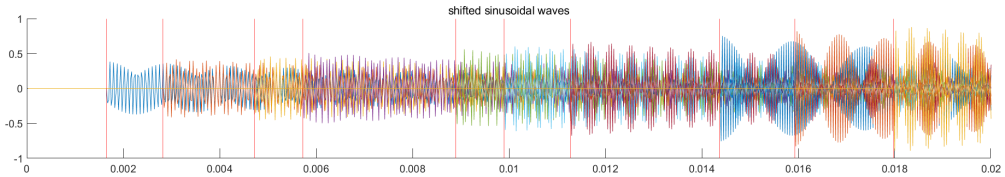
Then, the previously defined coefficients are arranged as follows: written as below:

$$\begin{aligned} c &= \frac{\det\left(\mathbf{\Psi}_{2k,\mathbf{c}}^{(j)}\right)}{\det\left(\mathbf{\Psi}_{2k,\mathbf{c}}^{(0)}\right)} \\ &= \frac{\det\left(\mathbf{U}\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_\beta\mathbf{\Theta}\mathbf{\Delta}\mathbf{V}^T\mathbf{\Phi}_{2k}^{(j)}\right)}{\det\left(\mathbf{U}\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_\beta\mathbf{\Theta}\mathbf{\Delta}\mathbf{V}^T\mathbf{\Phi}_{2k}^{(0)}\right)} \\ &= \frac{\det\left(\mathbf{V}^T\mathbf{\Phi}_{2k}^{j}\right)}{\det\left(\mathbf{V}^T\mathbf{\Phi}_{2k}^{0}\right)} \\ &= \frac{\det\left(\mathbf{V}_j\right)}{\det\left(\mathbf{V}_0\right)} \end{aligned} \tag{4.12}$$

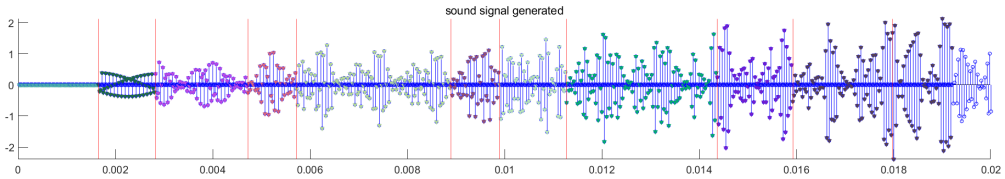
As in (4.12), $\det(\mathbf{\Delta})$ is the common term in the numerator denominator and so it is reduced. Therefore, it can be confirmed that the coefficient of the polynomial presented in the SSRF solution is constant regardless of the delay term. Therefore, by ignoring the delay term and by solving the equation in (3.11) to be zero, we can also get the solution for $\beta_1, \cdots, \beta_{2k}$ for signal with time delay τ_l.

For the case of solving α, it is different from the previous SSRF solution that did not consider time delay term. In the original SSRF introduced in 3, there was no time delay term, so α_l could be obtained by solving a simple linear equation or using an inverse matrix. However, in this section, we have to add the time delay term δ_l when solving the linear equation. This time, time delay τ_l is added to the matrix $\mathbf{\Sigma}_\alpha$. The elements of $\mathbf{\Psi}_m c$ can be written as below:

$$\mathbf{\Psi}_{mc}(i,j) = \sum_{i=1}^{k} \alpha_i \left\{ \beta_i^{(\psi_{mc}(j)-\delta_i)} + (\beta_i^*)^{(\psi_{mc}(j)-\delta_i)} \right\} \tag{4.13}$$



(a) Ten sinusoidal signals with different ToAs'.



(b) Superposition of ten sinusoidal signals.

Figure 4.4: Generated sound signal when $k = 10$.

Keeping the above equation in mind, you need to establish an equation by substituting different time delay values for the multipliers according to each beta value. If we obtain the alpha value that matches each beta value through the above process, we can perform signal separation by SSRF BSS technique for mixture of sinusoidal signals arriving at one microphone with different ToAs' even if we do not have information on the number of sound sources.

4.4 Results and Discussion

4.4.1 Source Number and ToA Estimation Results

In this chapter, it is assumed that the information on the number of sound sources in the anechoic chamber is unknown. Therefore, it is necessary to start by finding the number of sound sources.

For source number estimation performance analysis, YG BSS was used as a comparison method [46]. This is because YG BSS also includes estimating the number of sound sources when there is no prior information on the number of sound sources By increasing the number of sound sources from 2 to 10, simulations were performed to

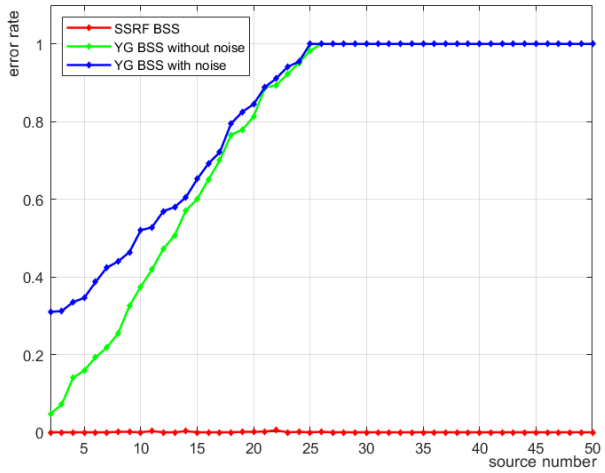


Figure 4.5: Source number estimation error rate.

estimate the number of sound sources 1000 times for each method. The result is shown in Fig. 4.5. As indicated by the red solid line in Fig. 4.5, the SSRF method has an error rate of 0% for all numbers of sound sources estimation. The YG BSS method simulated two situations with and without noise, which is the blue and green solid line in Fig. 4.5 respectively. As shown in the Fig. 4.5, the error rate for estimating the number of YG BSS sound sources is higher than the SSRF technique, and the error rate for YG BSS with noise is higher.

Next, the ToA estimation result will be discussed. The mixture of multiple sinusoidal signals with time delay generated are shown in Fig. 4.4. In Fig. 4.4 (a), mixture of sinusoidals with different ToAs' from ten different sound sources are shown. The red vertical line in Fig. 4.4 (a) indicates previously set ToAs'. The red vertical line in Fig. 4.4 (b) are the ToAs' calculated using SSRF BSS method. As shown in Fig. 4.4 (a) and (b), the ToA estimation result of SSRF BSS is correct. Existing BSS methods do not target signals with delay, so this chapter could not compare the ToA estimation results with other BSS techniques.

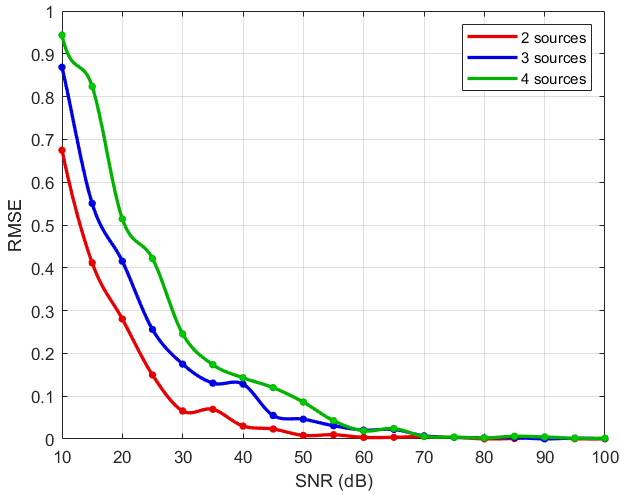


Figure 4.6: RMSE of signals with AWGN noise for $k = 2, 3$, and 4.

4.4.2 Separation of the Signal

In this section, we discuss the the BSS results of the proposed method. We check the result of whether the signal generated by each source is well separated, and compare the result with the comparison technique.

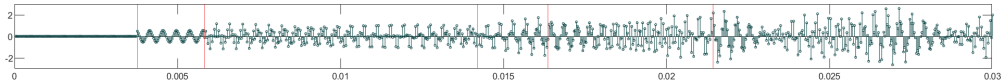
The separation result when $k = 5$, where there are five sound sources, is shown in Fig. 4.7. The mixed signal recorded with microphone is shown in Fig. 4.7 (a). Separated signals from source 1 to source 5 are shown in Fig. 4.7 (b) to Fig. 4.7 (f). The separation result when $k = 7$ is shown in Fig. 4.8. The mixed signal is shown in Fig. 4.8 (a). The separated signals are shown in Fig. 4.8 (b) to (g). The red vertical line added to each graph means time delay. As shown in Fig. 4.7 and 4.8, it can be confirmed that using the technique devised in this chapter, it is possible to separate the multi-source sine wave signal arriving with a time delay.

For signals with AWGN added, the RMSE for different SNR was calculated. This result is shown in Fig. 4.6. As shown in Fig. 4.6, the RMSE gets bigger as source number gets larger. For low SNR, RMSE is larger than the RMSE with higher SNR. For SNR over 40 dB, the RMSE is smaller than 0.1.

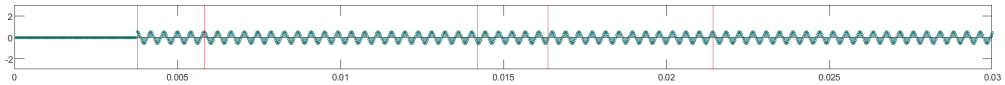
From now on, separation result is analyzed by comparing the performance with other BSS techniques. The techniques used to compare BSS results are independent component analysis (ICA) BSS and YG BSS. YG BSS is named after the name of the author. ICA BSS is the most widely used among BSS techniques. ICA BSS is under three conditions. First, all source signals are independent from each other. Second, each source signal has non-Gaussian distributions. Last, most of the ICA BSS assume that there is no time delay. YG BSS is established on the clustering features of time-frequency (TF) transform of modal response signals.

Although SSRF BSS can perform BSS on more than 10 mixed signals, the number of sound sources is limited to $k = 3$ for comparison with other BSS techniques. In the case of ICA BSS, the number of sound sources is *a priori*. YG BSS does not know the number of sound sources. Therefore, the result of performing three different BSS on the mixed sinusoidal signals with time delay is shown in Fig. 4.9. In Fig. 4.9 (a), the mixed sinusoidal signals with time delay is shown. The red vertical line is the ToA of each signal. The red stem signals in a box refers to the random-forked samples for SSF BSS method. In Fig. 4.9 (b), BSS result for each BSS method is shown. The separated signals on the left,middle and right side of Fig. 4.9 (b) are BSS results of SSRF BSS, ICA BSS and YG BSS respectively. As shown in Fig. 4.9 (b), SSRF BSS methos clearly separates three sinusoidal signals with exact ToA. However, the ICA BSS separation result shown in the middle of Fig. 4.9 (b) is not similar to the original sinusoidal signals. Even the separated results got noise. It is estimated that this is because the signal targeted by the ICA BSS is a signal without a time delay. Since source number $k = 3$ is *a priori*, it is meaningless to compare the results of estimation of the number of sources. YG BSS result is shown on the right side of Fig. 4.9 (b). As shown in Fig. 4.9 (b), YG BSS has estimated source number $k = 6$, and separated the original signal into six signals. Since the estimation of the number of sound sources is wrong, the separation result is also inaccurate.

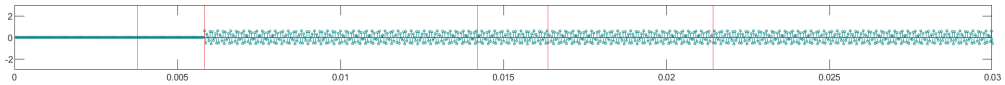
In summary, SSRF BSS can successfully separate each signals from the mixture of



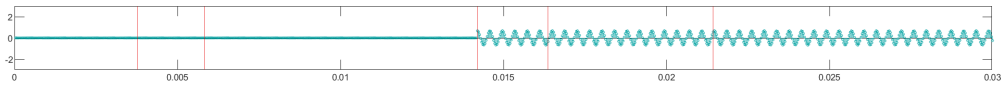
(a) mixed sinusoidal signals with time delay.



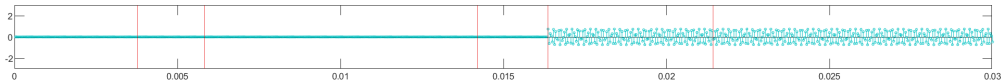
(b) Separated signal for source 1.



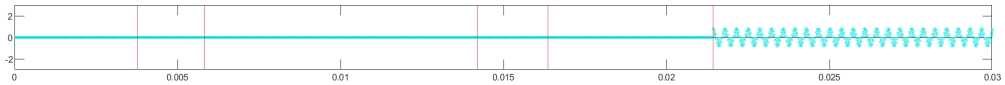
(c) Separated signal for source 2.



(d) Separated signal for source 3.



(e) Separated signal for source 4.



(f) Separated signal for source 5.

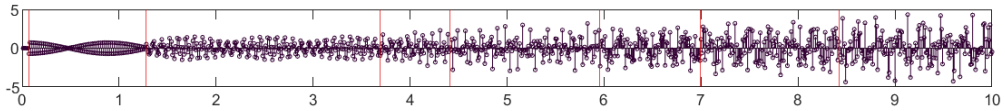
Figure 4.7: Signal separation when $k = 5$.

sinusoidals with time delay. SSRF does not need source number information *a priori*.

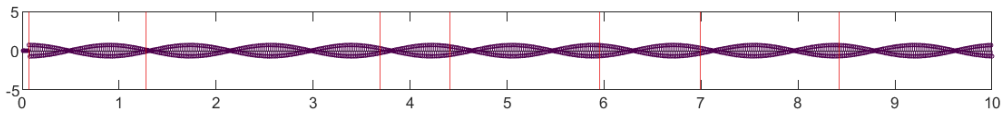
4.5 Summary

In this chapter, based on the SSRF technique, we introduced a technique for separating each signal after extracting each ToA from a signal with up to ten sound sources. It was assumed that up to ten sound sources were located at different distances from the microphone in the anechoic chamber, and each sound source generated different sinusoidal signals. Although the maximum number of sound sources is limited up to ten for the simulation, this technique can be applied to a larger number of sound sources.

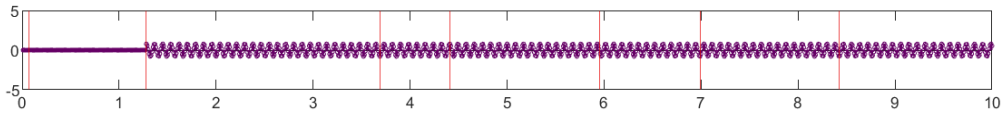
The devised method consist of two procedures. First one is to find the source num-



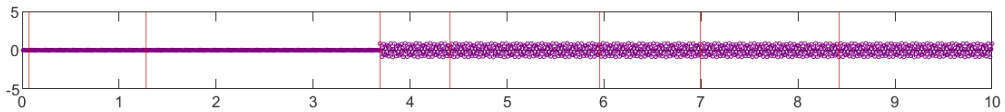
(a) mixed sinusoidal signals with time delay.



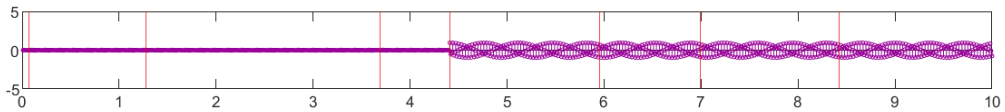
(b) Separated signal for source 1.



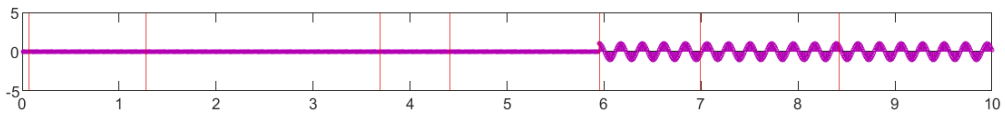
(c) Separated signal for source 2.



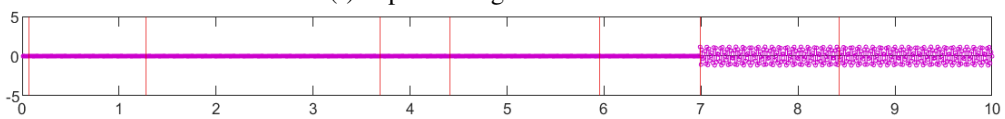
(d) Separated signal for source 3.



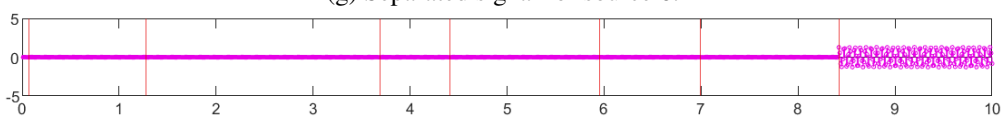
(e) Separated signal for source 4.



(f) Separated signal for source 5.

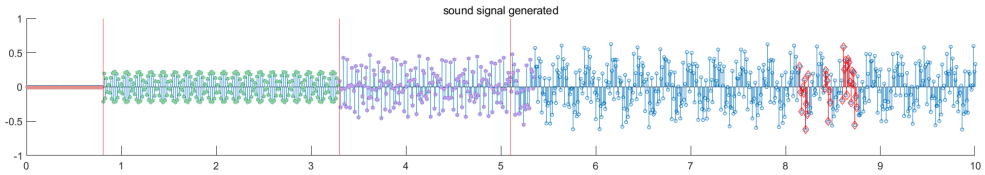


(g) Separated signal for source 6.

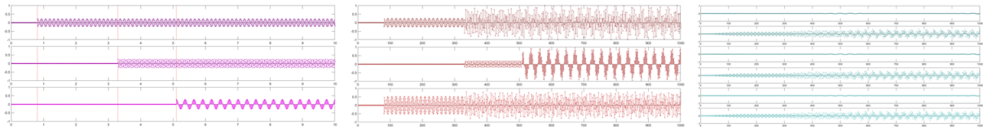


(h) Separated signal for source 7.

Figure 4.8: Signal separation when $k = 7$.



(a) Original mixed sinusoidal signal.



(b) Separation result of SSRF BSS on the left, ICA BSS in the middle and YG BSS on the right.

Figure 4.9: Comparison between different BSS methods.

ber and ToA, which both are unknown. And the second part is the separation of signal. For the first part, the value of only one coefficient of the polynomial was obtained by calculating the determinant using a part of the informative matrix, and changes in the coefficient value was concluded as ToAs'.

Then, based on the ToA obtained earlier, SSRF was applied for signal separation. The performance of SSRF BSS was compared to the separation results of ICA BSS and YG BSS. Since ICA BSS method has information of source number *a priori*, the ICA BSS separated the signals as much as the number of sound sources. But the separated signal from ICA BSS was very noisy and ToA of each signal was not clear due to high noise. The reason that the signal separation result of ICA BSS is not accurate and noisy is because ICA BSS assumes a signal without time delay. YG BSS method was even worse because YG BSS couldn't get the right source number. In comparison, SSRF BSS accurately found the number of sound sources and each ToAs', and was able to separate the same signal as the original signal based on this. As a result, SSRF BSS can be applied to separation of mixed sinusoidals with time delay. For future works, based on the ToA obtained through this technique, we plan to study the sound source localization estimation in a multi-channel multi-source situation.

Chapter 5

CONCLUSION

Sound is ubiquitous in our daily life. There are sounds that are generated intentionally , but there are also sounds that are continuously generated without intention. Regardless of whether the sound is intentionally generated or not, if useful information can be obtained through appropriate acoustic signal processing even for an unintentional sound, it will be of great help to our lives. The study in this paper was conceived in this respect. I tried to obtain useful information by signal processing the recorded sounds that occur in our daily life.

In this paper, various techniques of acoustic signal processing and research on their applications were introduced. Among various acoustic signal processing techniques, this paper especially deals with methods for acoustic localization, restoration of lost acoustic signal samples, and BSS which separates mixed acoustic signal.

The most representative example of acoustic signal processing application is acoustic localization. If the location of a sound generated by a person or a device can be estimated, LBS can be provided based on this. Acoustic localization technique can also be utilized to locate survivors in an emergency or disaster situation.

In order to improve the accuracy and performance of acoustic signal processing techniques such as acoustic localization, the quality of the recording must be high. One way of getting high quality recording is to install expensive, high-performance

microphones and signal processors. However, in reality, there is budget limit so this is not a decent solution. In addition, there is also the burden of having to deliberately purchase and install a high-quality microphone for recording. However if we use built in microphones in cell phones, this burden will be much lessened. However, built in microphones in cell phones provide limited recording performance. Also the top priority of built in microphones in cell phones is to improve the call quality. In order to achieve that, beamforming is performed according to the position of a caller's mouth, so there is a directivity dependency. As a result improving the recording quality by upgrading hardware is difficult. So there must be a way of recovering the lost sample values of acoustic signals through acoustic signal processing.

In the real world, sounds are emitted simultaneously from multiple sources and the recording is a mixture of those signals. A mixture of numerous sound signals such as human speech, machine sounds, cell phone sounds, music sounds, and background noise is all mixed in to a microphone. Therefore, in order to obtain a signal processing result that exactly follows the intention, it is necessary to separate the mixed signal into single signals.

With this in mind, we conducted studies on acoustic localization, recovery of lost signal samples, and BSS for separating mixed signals in this paper. Chapter 2 dealt with the study of acoustic localization. Chapter 3 discussed the lost signal recovery technique called SSRF. Chapter 4 introduced a SSRF BSS technique for separating mixture of multiple sound sources. In particular, the acoustic signal targeted in Chapters 3 and 4 was a mixture of several sinusoidal signals.

In chapter 2, novel acoustic localization method using a newly defined cost function is introduced. Existing acoustic localization methods have disadvantage in that their accuracy is greatly reduced under high reverberation indoor condition. To improve the acoustic localization performance under highly reverberant condition a new cost function is defined in this chapter. The newly defined cost function finds an optimal pair of microphones with the best performance for ILD and TDoA techniques.

Compared to other existing acoustic localization methods, the RMSE of devised cost function acoustic localization was lowest.

Chapter 3 dealt with signal restoration method bases on SSRF. Basically, it was assumed that the sound of multiple sound sources was mixed and recorded with one microphone. In this study, the source number is known *a priori*. Sinusoidal signal with a decay was assumed for the signal generated by each sound source. Assuming that some of the recorded samples were lost, we aimed to restore the lost sample values. For restoration, we defined a novel concept called random fork. Random fork is a tool that picks several samples with fixed random intervals like a fork. If the number of fork indices is set to double the number of sources, it is transformed into a mathematical problem that can recover the values of lost samples. Components of sinusoidal signals are solutions of n-th order equation and some other simple linear equations. The performance of SSRF signal recovery was compared with other existing signal recovery methods based such as DNN and compressive sensing. As a result, the signal recovery performance based on SSRF technique outperformed the other two existing methods.

In Chapter 4, the BSS problem was solved using the SSRF technique in a situation where the number of sound sources is unknown. The BSS problem in this chaper was single-channel underdetermined BSS. Since each sound source has different distance from microphone, it is assumed that they all have different ToAs'. Given this, source number and ToAs' were obtained first and follows by signal separation. It was confirmed that BSS result using SSRF BSS method was more accurate than other existing BSS methods.

Throughout this whole paper, acoustic signal processing and its applications were discussed. A study was conducted on acoustic localization using microphone arrays, recovering lost signal sample values, and BSS which separated mixed signal by using only one microphone. Cost function based acoustic localization method devised in this paper made it possible to improve the localization accuracy under highly reverberant condition. From SSRF signal recovery method, lost sample values of acoustic signal

were recovered. Finally, SSRF BSS separated mixture of sinusoidals with time delay into single ones.

Bibliography

[1] T. Alkhalifah, “An acoustic wave equation for anisotropic media,” *Geophysics*, vol. 65, no. 4, pp. 1028-1340, July 2000.

[2] B. Thoen, S. Wielandt and L. D. Strycker, “Fingerprinting method for acoustic localization using low-profile microphone arrays,” in *Proceedings of International Conference on Indoor Positioning and Indoor Navigation*, Nantes, France, September 2018. pp. 1-7.

[3] K. Lee, J. Ou and L. Wang, “Underwater acoustic localization by probabilistic fingerprinting in eigenspace,” in *Proceedings of Oceans 2009*, Mississippi, United States, October 2009. pp. 1977-1980.

[4] M. Nakamura, K. Fujimoto, H. Murakami, H. Hashizume and M. Sugimoto, “Indoor localization method for a microphone using a single speaker,” in *Proceedings of International Conference on Indoor Positioning and Indoor Navigation*, Lloret De Mar, Spain, November 2021. pp. 516-523.

[5] M. Miskovic, M. Eric, M. Stanojevic, M. Milosavljevic and Z. Mihailovic, “Experimental results of the outdoor near-field acoustic source location,” in *Proceedings of 19th Telecommunications Forum*, Belgrade, Serbia, November 2011. pp. 1056-1058.

[6] D. Salvati, C. Drioli, G. Ferrin and G. L. Foresti, “Acoustic source localization from multirotor UAVs,” *IEEE Transactions on Industrial Electronics*, vol. 67, no. 10, pp. 8618-8628, October 2020.

[7] X. Jiahui, C. Keyu and C. En, “An improved APIT localization algorithm for underwater acoustic sensor networks,” in *Proceedings of IEEE International Conference on Signal Processing, Communications and Computing*, Xiamen, China, October 2017. pp. 399-403.

[8] D. Kuryliak and V. Lysechko, “Diffraction of the sound wave by a finite soft (rigid) cone,” in *Proceedings of International Seminar/Workshop on Direct and Inverse Problems of Electromagnetic and Acoustic Wave Theory*, Tbilisi, Georgia, September 2014. pp. 160-162.

[9] Y. Wang, S. Yin, J. Sun, S. Hu and Y. Yu, “Characteristics of acoustic scattering from random fluctuation surface based on PM spectrum” in *Proceedings of International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, Shanghai, China, October 2017. pp. 1228-1232.

[10] A. Johansson, G. Cook and S. Nordholm, “Acoustic direction of arrival estimation, a comparison between root-music and SRP-PHAT,” in *Proceedings of IEEE Region 10 Conference TENCON*, Chiang Mai, Thailand, November 2004. pp. 629-632.

[11] H. Do and H. F. Silverman, “SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, Texas, March 2010. pp. 125-128.

[12] A. Levi, and H. F. Silverman, “An alternate approach to adaptive beamforming using SRP-PHAT,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, Texas, March 2010. pp. 2726-2729.

[13] B. Zou, J. Zhai, J. Xu, Z. Li and S. Gao, "Experimental study on underwater acoustic signal recovery technique based on chirp-modulation passive time reversal using multi-receive array," in *Proceedings of IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference*, Chongqing, China, March 2017. pp. 267-271.

[14] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066-1074, March 2007.

[15] S. N. Jain, and C. Rai. "Blind source separation and ICA techniques: a review," *International Journal of Engineering Science and Technology*, vol. 4, no. 4, pp. 1490-1503, April 2012.

[16] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 666-678, March 2006.

[17] T. Takatani, T. Nishikawa and H. Saruwatari, "Blind source separation based on binaural ICA," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, April 2003. vol. 5, pp. 321-324.

[18] Y. Li and B. Li, "A novel online algorithm for blind source separation with unknown number of sources," in *Proceedings of International Conference on Consumer Electronics, Communications and Networks*, Xianning, China, March 2011. pp. 2885-2888.

[19] Q. Lv and X. Zhang, "A unified method for blind separation of sparse sources with unknown source number," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 49-51, January 2006.

[20] F. Wang and J. Zhang, “Adaptive sparse factorization for even-determined and over-determined blind source separation,” in *Proceedings of International Conference on Computational Intelligence and Software Engineering*, Wuhan, China, December 2009. pp. 2075-2078.

[21] L. Wang, J. D. Reiss and A. Cavallaro, “Over-determined source separation and localization using distributed microphones,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1573-1588, September 2016.

[22] X. Sheng and Y. Hu, “Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks,” *IEEE Transactions on Signal Processing*, vol. 53, no. 1, pp. 44-53, January 2005.

[23] J. C. Chen, R. E. Hudson and K. Yao, “Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field,” *IEEE Transactions on Signal Processing*, vol. 50, no. 8, pp. 1843-1854, August 2002.

[24] J. C. Chen, K. Yao and R. E. Hudson, “Source localization and beamforming,” *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 30-39, March 2002.

[25] C. Klungmontri and I. Nilkhamhang, “Acoustic underwater positioning system using fast fourier transform and trilateration algorithm,” in *Proceedings of 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, Phuket, Thailand, June 2017. pp. 521-524.

[26] P. Osterrieder, L. Budde, and T. Friedli, “The smart factory as a key construct of industry 4.0: A systematic literature review,” *International Journal of Production Economics*, vol. 221, no. 1, pp. 107476, March 2020.

[27] R. F. Estrada and E. A. Starr, “50 years of acoustic signal processing for detection: coping with the digital revolution,” *IEEE Annals of the History of Computing*, vol. 27, no. 2, pp. 65-78, April 2005.

[28] R. Lienhart, I. Kozintsev, M. Yeung and S. Wehr, “On the importance of exact synchronization for distributed audio signal processing,” in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, United States, October 2003. vol. 4, pp. 840-843.

[29] L. Cheng, Z. Wang, Y. Zhang, W. Xu and J. Wang, “Acouradar: Towards single source based acoustic localization,” in *Proceedings of IEEE International Conference on Computer Communications*, Toronto, Canada, July 2020. pp. 1848-1856.

[30] W. Meng and W. Xiao, “Energy-based acoustic source localization methods: a survey,” *Sensors*, vol. 17, no.2, pp. 376-396, February 2017.

[31] J. Vermaak and A. Blake, “Nonlinear filtering for speaker tracking in noisy and reverberant environments,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Utah, United States, May 2001. vol. 5, pp. 3021-3024.

[32] H. Do, and H. F. Silverman, “Stochastic particle filtering: A fast SRP-PHAT single source localization algorithm,” in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, United States, October 2009. pp. 281-284.

[33] J. Cardoso, “Blind signal separation: statistical principles,” *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009-2025, October 1998.

[34] R. Aichner, H. Buchner, F. Yan and W. Kellermann, “A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments,” *Signal Processing*, vol. 86, no. 6, pp. 1260-1277, June 2006.

[35] T. Nishikawa, H. Saruwatari, and K. Shikano. “Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA,” *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 86, no. 4, pp. 846-858, April 2003.

[36] L. Drude and R. Haeb-Umbach, “Integration of neural networks and probabilistic spatial models for acoustic blind source separation,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 815-826, April 2019.

[37] E. S. Warner and I. K. Proudler, “Single-channel blind signal separation of filtered MPSK signals,” *IEE Proceedings - Radar, Sonar and Navigation*, vol. 150, no. 6, pp. 396-402, December 2003.

[38] A. Holobar and D. Zazula, “Multichannel blind source separation using convolution kernel compensation,” *IEEE Transactions on Signal Processing*, vol. 55, no. 9, pp. 4487-4496, September 2007.

[39] L. Parra and C. Fancourt, *Noise Reduction in Speech Applications*, CRC Press, Florida, 2018.

[40] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa and K. Shikano, “Blind source separation combining independent component analysis and beamforming,” *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 11, pp. 1135-1146, October 2003.

[41] Y. Li, K. C. Ho, and M. Popescu, “Efficient source separation algorithms for acoustic fall detection using a microsoft kinect,” *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 745-755, November 2013.

[42] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, “Calibration of distributed sound acquisition systems using TOA measurements from a moving acoustic source,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, May 2014. pp. 7455-7459.

[43] J. R. Jensen, U. Saqib, and S. Gannot, “An EM method for multichannel TOA and DOA estimation of acoustic echoes,” in *Proceedings of 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, United States, October 2019. pp. 120-124.

[44] A. Ledeczi, P. Volgyesi, M. Maroti, G. Simon, G. Balogh, A. Nadas, B. Kusy, S. Dora and G. Pap, “Multiple simultaneous acoustic source localization in urban terrain,” in *Proceedings of the Fourth International Symposium on Information Processing in Sensor Networks*, Idaho, United States, April 2005. pp. 491-496.

[45] J. H. Park, W. Cho, S. Kim and W. Lee, “Sketching and stacking with random fork based exact signal recovery under sample corruption,” *IEEE Access*, vol. 8, no. 1, pp. 171195-171202, September 2020.

[46] Y. Gang, “An underdetermined blind source separation method with application to modal identification,” *Shock and Vibration*, vol. 2019, no. 1, pp. 1-15, October 2019.

[47] C. Jutten and J. Karhunen, “Advances in blind source separation (BSS) and independent component analysis (ICA) for nonlinear mixtures,” *International Journal of Neural Systems*, vol. 14, no. 5, pp. 267-292, October 2004.

초 록

최근 음향 신호 처리에 대한 연구가 증가하고 있다. 음향 신호 처리를 통해 유의미한 정보를 얻어내 유용하게 활용할 수 있기 때문이다. 따라서 본 논문에서는 실내 환경에서 취득한 소리에 적용 가능한 음향 신호 처리 기법에 관한 내용을 다룬다.

처음으로는 잔향이 높고 잡음이 많은 실내 환경에서 녹음한 음원 신호로부터 음원 위치를 추정하는 기법을 소개한다. 기존 음원 위치 추정 기법인 에너지 기반 위치 추정, 시간 지연 기반 위치 추정 및 SRP-PHAT 기반 위치추정 기법의 경우 잔향이 높아 소리가 울리는 실내 환경에 적용하면 그 정확도가 떨어진다. 반면 본 논문에서는 여러개의 마이크로 구성된 마이크 어레이로 부터 최적의 성능을 낼 수 있는 마이크의 조합을 찾아낼 수 있는 비용 함수를 새로이 정의한다. 이 비용함수 값이 최저가 되는 마이크 조합을 찾아내 해당 마이크로 음원 위치 추정을 진행한 결과 기존 기법 대비 거리 오차가 줄어든 것을 확인하였다.

다음으로는 손실이 발생한 녹음 음원에서 손실된 값을 복원하는 기법을 소개한다. 본 기법에서 목표로 삼는 음원은 여러 개의 사인파형 신호가 합쳐져서 들어오는 음원이다. 무향실에는 여러개의 음원이 존재하지만 마이크는 단 한개만 있는 상황을 가정한다. 사인 파형은 오일러 공식에 기반해 지수 함수 꼴로 변형할 수 있고, 만약 지수함수 구성 항 중 일부가 등비수열을 따르는 경우 본 논문에서 소개하는 기법을 이용해 해당 등비수열의 구성값을 구할 수 있다. 본 문제를 풀기 위해 랜덤 포크라는 개념을 새로이 도입했다. 본 기법을 이용해 신호를 복원한 결과, 신호 복원 정확도는 기존의 압축 센싱 기반 복원기법 및 DNN 기반 복원 기법보다 그 정확도가 높았다.

마지막으로 본 논문에서는 이전에 소개한 SSRF 기법을 기반으로 합쳐진 신호

를 분리하는 기법을 소개한다. 본 기법에서는 이전과 같이 사인 파형의 신호가 합쳐져서 들어오는 상황을 가정한다. 거기에 더해 이전 기법에서는 모든 사인 파형이 동시에 재생되는 상황을 가정한 반면, 본 기법에서는 각기 다른 음원이 마이크로부터 각각 다른 거리만큼 떨어져 있어서 모두 다른 시간 지연을 가지고 마이크로 도달하는 상황을 가정한다. 이렇게 서로 다른 시간지연을 갖고 하나의 마이크로 도달하는 사인파형의 신호가 합쳐진 상황에서 각각의 신호를 분리한다. 본 논문에서 소개하는 기법은 크게 음원 갯수 추정, 시간 지연 추정 및 신호 분리의 세 개 단계로 구성된다. 기존의 음향 신호 분리 기법들이 음원의 갯수에 대한 정보를 미리 알아야 한다거나, 시간지연이 없는 신호에 대해서만 적용이 가능했다면, 본 기법은 사전에 음원 갯수에 대한 정보가 없어도 적용 가능하다는 장점이 있다. 해당 기법은 SSRF 기법을 기반으로 하는데, SSRF 문제를 푸는 과정에서 구해지는 방정식의 계수 값이 변하는 지점을 시간 지연으로 추정한다. 그리고 시간 지연 값의 변화가 몇 번 발생하는가에 따라 음원의 갯수를 추정한다. 마지막으로 모든 신호가 합쳐진 최종 구간에서 SSRF 문제를 풀어 개별 신호를 구성하는 값을 구해내 신호 분리를 완료한다. 본 기법은 여러 가정이 필요한 기존의 ICA 기반 음향 신호 분리 및 YG 음향 신호 분리에 비해 더 정확한 신호분리 결과를 내는 것을 확인하였다.

주요어: 음향 신호 처리, 음원 위치 추정, 음향 신호 복원, 음향 신호 분리, 랜덤 포크를 이용한 스케치 및 스택킹

학번: 2016-20904