Blind Source Separation and Direction Estimation for Stereophonic Mixtures of Multiple Speech Signals Based on Time-Frequency Sparseness

March 2012

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Engineering

Keio University

Graduate School of Science and Technology
School of Integrated Design Engineering

Ning Ding

Preface

With the development of modern society, digital signal processing has become increasingly important because it is changing our daily lives in many ways. In the field of digital signal processing, human-machine interfaces are regarded as a key topic. Among the various applications of human-machine interfaces, speech information is very useful. We often encounter questions such as "Where is the speech source?" and "Can we obtain the desired speech from many simultaneous speeches?". The former is known as the direction-of-arrival (DOA) problem, while the latter is known as the blind source separation (BSS) problem, for which a sensor array technique is essential.

In this dissertation, several approaches to blind source separation and DOA estimation using a microphone array are described. In this study, without knowledge of source localization, active time or mixing process (blind), a pair of microphones is used to estimate the source directions and separate multiple speech signals even when the number of sources is two or more.

1. Speaker localization and source separation by Principal Component Analysis (PCA) and harmonic structure

In conventional methods data are treated as a whole in the time-frequency domain, and the difference between time frames is not distinguished. However, these methods suffer from low separation performance when the sources are closely located.

Since the ratio of the principal eigenvalues obtained by principal component analysis (PCA) indicates the degree of data spread around the first principal axis, in the author's approach, using the mathematical tool of PCA to analyze the phase difference versus frequency distribution data in a single time frame, The observated time frames are classified according to the activity pattern of multiple source frames to non-source active (NSA), single-source active (SSA) and double-source active (DSA) frames. SSA frames are used for DOA estimation. A new separation algorithm is explored for use in DSA frames. Depending on the frequency band, two methods are combined to obtain the separated signals in DSA frames: a DOA-based method and a harmonic-structure-based method.

2. Reliable cell selection

The common-sense approach that the use of reliable data guarantees reliable results is adopted. The problem is how to check the reliability of data from the observation. In the author's approach, the consistency with neighborhood data is utilized and cells are selected using a newly defined reliability index. The reliability index of a T-F cell's phase difference exploits the consistency of the time difference of arrival (TDOA) in the local window of the underlying cell. The consistency of the TDOA in a window is evaluated using the variance of the TDOAs for all T-F cells in the window.

3. Use of kernel density estimator for DOA estimation

A model of the propagation of the statistical error between the estimated phase difference and the consequent DOA is introduced. The model leads to a probability density function (PDF) of the DOA, then the DOA estimation problem is reduced to finding the most probable points for the DOA. Finally, the kernel density estimator is applied to selected cells to calculate the PDF and estimate the source direction.

Some experiments were performed to evaluate the proposed methods. The results show that the proposed source separation method is superior to the conventional method, and the proposed DOA estimation method outperforms other methods in terms of both accuracy in the case of real observed data and robustness in the case of simulation with additional diffused noise.

Acknowledgments

This dissertation is a summary of my studies as a PhD candidate at Keio University. First of all, I would like express my deepest thanks to my supervisor, Dr. Nozomu Hamada, Professor of Keio University. In the last three years, he has provided me with valuable guidance and constructive suggestions in every step of my research. Without his continuous encouragement, kindness and patience, I could not have completed this thesis. His keen academic observations have not only enlightened me during the production of this thesis but will also be invaluable in my future life.

I would also like to express my thanks to the members of thesis review committee, Dr. Masaaki Ikehara, Dr. Eiji Okada, Dr. Toshiyuki Tanaka, and Dr. Yasue Mitsukura, for their helpful advice and comments.

I extend my gratitude to all members of Hamada Laboratory, especially Mr. Hikaru Ouchi, Mr. Takayuki Shimada, Mr. Masashi Yoshida, Mr. Junya Ono, Mr. Jaime Lorenzo Trueba, and Mr. Ken Fujimoto, as well as Mr. Masashi Sekikawa in the Speech Signal Processing group, for their assistance with my study.

I would also like to thank Professor Wlodzimierz Kasprzak from Warsaw University of Technology for the cooperated research.

Last but not least, I would like to thank my family, especially my wife Mrs. Tianfeng Li. In order to support my study abroad, we had to live in different countries for almost three years. I really appreciate my family for their support.

Ning Ding
Yokohama, Japan
Feb. 2012

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 General background

With the explosive growth of digital communications and digital media, digital signal processing is more important than ever [3] [4]. Digital signal processing is concerned with the representation of discrete time signals by a sequence of numbers or symbols and the processing of these signals [5]. It has many applications such as audio and speech signal processing, sonar and radar signal processing, sensor array processing, digital image processing, signal processing for communications, the control of systems, biomedical signal processing, and seismic data processing [6].

Some of the most important applications of digital processing techniques have been in the area of speech processing. In fact, much of the theoretical background of digital signal processing has been derived from studies on speech. Digital processing has been applied to a wide range of problems in speech including speech recognition, speech synthesis, speech source separation, and speech source direction estimation.

Speech recognition, also known as automatic speech recognition or computer speech recognition, is a means of recognizing speech without targeting a single speaker. The first speech recognizer appeared in 1952 and consisted of a device that recognized single spoken digits [7]. For the past fifty years, speech recognition research has been characterized by the steady accumulation of small incremental improvements. There has also been a continued trend of focusing on increasingly difficult tasks owing to both progress in speech recognition performance and the availability of faster computers.

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer and can be implemented in software or hardware [8]. Speech synthesis has long been a vital auxiliary technology and its application is significant and widespread. Its longest application has been in the use of screen readers for people with visual impairment, but text-to-speech systems are now commonly used by people with dyslexia and other reading difficulties as well as by pre-literate children. Stephen Hawking is one of the most famous people using speech synthesis to communicate.

Speech source separation and speech source direction estimation, which are the focal points of the research reported in this thesis, are introduced in the following sections.

1.2 Blind source separation and source direction estimation

The speech source separation problem is to determine the original signals when several speech signals have been mixed together. A typical example is the "cocktail party problem", where a number of people are talking simultaneously in a room (such as at a cocktail party), and one is trying to follow one of the discussions. Although the human brain can handle this sort of auditory source separation problem, it is a very tricky problem in speech signal processing.

It is well documented that listeners with hearing loss have greater difficulty in understanding speech with background noise. Modern hearing aids improve the audibility of a speech signal and the comfort of noisy speech. However, the ability of hearing aids to improve the intelligibility of noisy speech is rather limited [9] [10]. Because of the ever-present nature of background noise, it is very important for hearing aid research to develop speech separation methods that have the potential to enhance speech intelligibility in noise.

In general, numerous mixed signals exist in real environments, including desired signals, undesired signals, and noise signals. The purposes of multiple speech signal processing are to localize and separate mixed signals, enhance target signals, reduce noise signals, and cancel echo. There are several types of separation problems based on various classification methods, some of which are shown in Tab. 1.1.

Blind source separation (BSS) is a typical speech source separation problem, in which the aim is to obtain the separated signals without any a priori information, such as the source position, mixing process, environment, and so forth. The use of BSS in the development of effective acoustic communication channels between humans and machines is widely accepted. Source direction estimation is the major means of acquiring a speaker’s location. Because the relative positions between the speaker and the microphones are different in each situation, locating and tracking the direction of the source are required. At the same time, source direction estimation can also provide useful information for BSS. BSS and direction estimation are widely applied in our daily lives. In the following, some examples of

Table 1.1: Classification of separation problems

Criterion	Classification	Remarks
Number of sensors	Monaural	
	Multiple sensors	Array processing
Echo	Echoic	Real mixture
	Anechoic	Approximated mixing process
Mixture model	Convolutive	Delay and intensity differences
	Instantaneous	Intensity difference
Source location	Moving	Short time interval
	Stationary	Fixed position
Number of sources and sensors	Overdetermined	number of sources < number of sensors
	Underdetermined	number of sources > number of sensors
Sensor configuration	Known	
	Unknown	
Representations	Time-frequency domain	
	Time domain	
Nature of sources	Sparseness in T-F domain	
	Statistical independence	

application are given.

Video conference system

A video conference system (Fig. 1.1) is a set of interactive telecommunication technologies which allows two or more locations to interact via simultaneous two-way video and audio transmissions. When multiple speakers utter simultaneously, a separation system is used to obtain the desired speaker's voice. At the same time, camera manipulation is necessary as well as acoustic processing to capture the active speaker's face properly. For this purpose, speaker direction estimation is also essential.

Hands-free system

The term "hands-free system" describes equipment that can be used without the use of the hands, for example, via voice commands, or in a wider sense, equipment which requires only limited use of the hands so that the hands can be employed for another task such as driving.

Figure 1.1: Video conference [1]

For instance, in a driving hands-free system (Fig. 1.2), the signals received by the cell phone or GPS navigation system in the car are not only the driver' voice, but also the noise from the engine or from the environment. Separating the driver's voice from other signals is very helpful for speech recognition. This type of system is convenient and safe for drivers.

1.3 Main contributions of this research

This dissertation focuses on the BSS and direction estimation problems. The key features of the problems and the methods in this thesis that are mainly considered and/or utilized are as follows:

- Only the data from a pair of omnidirectional microphones are used.
- Analysis is based on time-frequency (T-F) sparseness of speech signals.
- It is possible to solve underdetermined cases in which the number of sources is greater than the number of sensors.
- Basic data for processing are treated in phase difference versus frequency (PD-F) space.
- The distance between microphones is sufficiently small to avoid spatial aliasing, where the typical settings are a distance of 4 cm between sensors and a sampling frequency of 8 kHz.

Figure 1.2: Hands-free system [2]

The main contributions of this dissertation are as follows.

1. Speaker localization and source separation using principal component analysis and harmonic structure

The approach in this dissertation is based on the framewise analysis of the PD-F data plot. First, the PD-F distribution is investigated by principal component analysis (PCA) at individual time frames, and a set of single-source active frames is selected. The ratio of the principal eigenvalues for this set is used as a confidence measure for accurately estimating the source direction. Second, the separation of multiple-source active frames is developed. Because the source location attributes are not reliable for separation in the low-frequency band, initially separated signals in the medium-frequency band are obtained in accordance with the directions estimated in the first step. Then, to cluster the remaining T-F cells in the low-frequency band, the harmonic structures observed in the spectrograms of the initially separated individual sources are associated with the frequency components of the mixed signals.

2. Reliable cell selection

In this approach, a novel cell selection method based on a reliability index is proposed. This idea is originated from an observation derived from the following assumption: when a single source appears in a given set of T-F windows, the time differences of arrival (TDOAs) should take almost the same values in the windows, and these values are considered to be reliable. The selected cells are solely utilized for direction of arrival (DOA) estimation.

3. DOA estimation using kernel density estimator

A statistical model relating the phase difference and direction angle is constructed. By employing this model and the sparseness in the T-F domain, the DOA estimation problem is reduced to obtaining the local peaks of the probability density function of the DOA. The final stage is to estimate the source direction using the kernel density estimator. This approach can be applied in underdetermined cases in which there are three sources but only two sensors. Experimental results show that this approach has the advantages of high direction angle resolution and robustness against noise.

4. Methodological difference between the proposed and conventional approaches

The conventional T-F masking methods use the attenuation between received signals and the delay as the features. Unlike the conventional methods, the PD-F distribution is utilized as the feature. The significant difference is that the new feature contains the frequency axis. Based on this novelty, frame-by-frame approach is proposed in Chapter 3. From phase difference error distribution, the DOA error distribution can be derived, and its relationship is used for the kernel density estimator in Chapter 4.

1.4 Overview of dissertation

This dissertation is organized as follows.

In Chapter 2, the foundations of speech signal processing using a microphone array are summarized. First, the foundation of speech signal processing are introduced in Sec. 2.2. T-F analysis, one of the most useful methods for speech signal processing, is described in Sec. 2.3. Speech signal processing using a microphone array is outlined in Sec. 2.4. The latter part of Chapter 2 focuses on the applications of speech signal processing: BSS in Sec. 2.6 and DOA estimation in Sec. 2.8. The typical BSS method of T-F masking is discussed in Sec. 2.7.

In Chapter 3, a T-F domain masking method for separating stereophonic audio mixtures is proposed. The approach is based on the framewise analysis of the PD-F data plot and the harmonic structure in the low-frequency band. In Sec. 3.3, the BSS problem and T-F masking method are reviewed briefly. Sec. 3.4 discusses the proposed method in detail. In Sec. 3.8, some experiments that were performed to verify the proposed method are reported. Sec. 3.9 gives a summary of this chapter.

Chapter 4 addresses DOA estimation methods. The DOA information is introduced in Sec. 4.2. Sec. 4.3 describes DOA estimation by Hough transform. The proposed novel cell selection is discussed in Sec. 4.4. Then the DOA error distribution model is builted in Sec.

4.5. Finally, the source directions are estimated by the kernel density estimator in Sec. 4.6. Sec. 4.7 and Sec. 4.8 are the experiments. Sec. 4.9 is the summary.

Finally, Chapter 5 concludes the dissertation.

Chapter 2

Foundations of speech signal processing using microphone array

2.1 Introduction

The first part of this chapter mainly introduces the foundations of speech signal processing using a microphone array. First, the foundations of speech signal processing are reviewed in Sec. 2.2. Time-frequency (T-F) analysis on which the proposed methods are based is explained in Sec. 2.3. Then in Sec. 2.4, the signal processing using a microphone array is discussed. Many speech signal processing methods have been proposed that employ a microphone array. In Sec. 2.6 and Sec. 2.8, the blind source separation (BSS) and direction of arrival (DOA) problems are introduced and previous approaches are explained. Sec. 2.9 is a summary of this chapter.

2.2 Foundations of speech signal processing

Speech communication is one of the basic and most essential capabilities possessed by human beings. Speech can be considered to be the most important method through which people can convey information without the use of tools. Although we passively receive more stimuli from outside through the eyes than through the ears, mutually communication is almost entirely through speech.

Speech conveys several types of information including linguistic information that indicates the meaning the speaker wishes to impart, individual information representing who is speaking, and emotional information depicting the emotion of the speaker. Needless to say, the first type of information is the most important.

The basic unit for constructing a sentence is the word, and each word is composed of syllables. Each syllable consists of phonemes, which can be classified as vowels or consonants. The number of vowels and consonants depends on the language or the classification, but broadly speaking, English has 12 vowels and 24 consonants, whereas Japanese has 5 vowels and 20 consonants [11]. The speech production process involves three subprocesses: source generation, articulation, and radiation.

The mechanism of vocal vibration is very complicated. In principle, however, the Bernoulli effect [12] associated with the airflow and the stability produced by the elasticity of the muscles draw the vocal cords toward each other. When the vocal cords are strongly strained and the pressure of the air rising from the lungs is high, the open-and-close period becomes short and the pitch of the sound source becomes high. Conversely, a low air pressure produces a lower-pitched sound. This vocal cord vibration period is called the fundamental period and its reciprocal is called the fundamental frequency.

Statistical analysis of the temporal variation in the fundamental frequency during conversational speech for a large number of speakers indicates that the mean and standard deviation for female voices are roughly twice those for male voices [13]. The fundamental frequency distribution of speakers on a logarithmic frequency scale can be approximated by two normal distribution functions which correspond to male and female voices. The mean and standard deviation for male voices are 125 and 20.5 Hz, respectively, whereas those for female voices are two times larger.

Frequency analysis of the temporal pattern of the fundamental frequency, in which silent periods are smoothly connected, shows that the frequency of temporal variation is less than 10 Hz/s. This implies that the rate of temporal variation in the fundamental frequency is relatively low.

Among the various types of information contained in speech, the features in the temporal, spectral, and spatial signal domains are the three most important features. An example of a speech signal is shown in Fig. 2.1. It is clear that the power of the speech is concentrated in a number of time intervals and that the power in the other intervals is almost zero. In a certain sense, voice activity detection is directly related to temporal features.

The spectrogram is a three-dimensional (3D) plot illustrating how the frequency spectrum varies with time. A spectrogram of the speech signal in Fig. 2.1 is shown in Fig. 2.2. The vertical axis represents the frequency and the horizontal axis represents time. The third dimension is given by the color of the plot and indicates the power at a particular T-F point. Specifically, the color of the plot is proportional to the logarithm of the power.

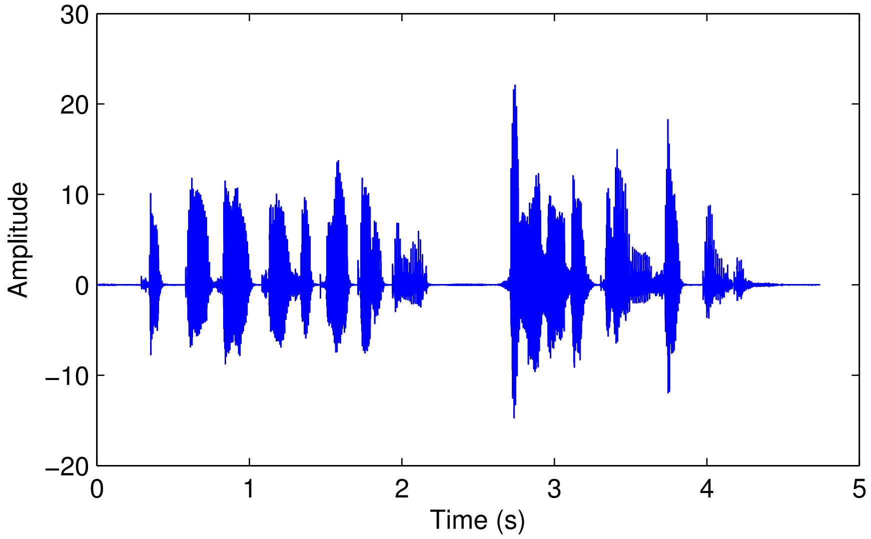


Figure 2.1: Example of speech signal (male speaking the Japanese sentence "Chotto osoi chuushokuwo torutame famiriresutoranni haittanodesu")

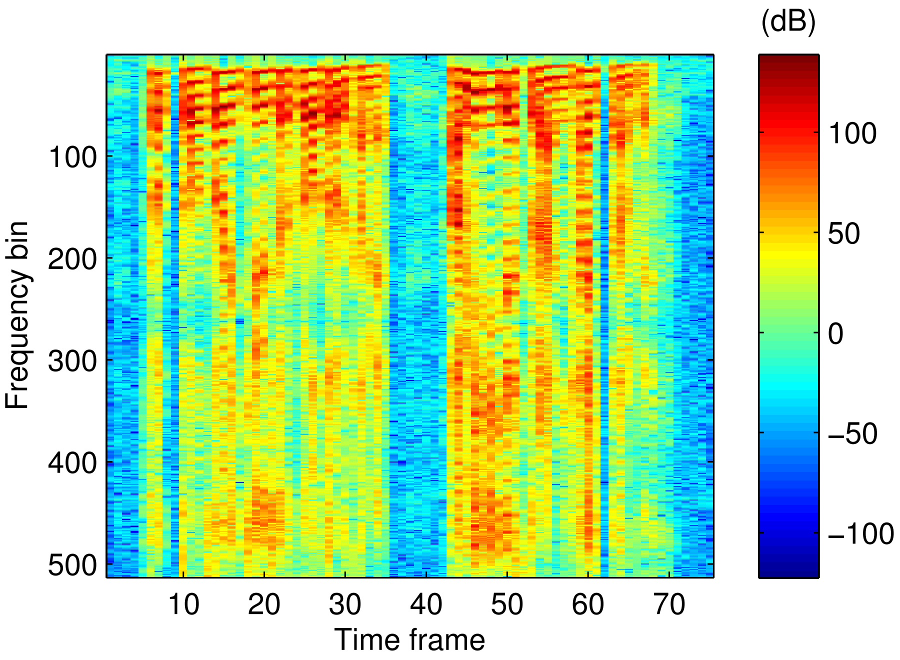


Figure 2.2: Spectrogram of speech signal in Fig. 2.1

The third feature of speech information is the spatial feature, which is closely related to the speaker direction and speaker localization.

2.3 Time-frequency analysis

The most basic requirement in the analysis of speech is to convert an analogue speech signal into a digital format, in which it is represented by a sequence of numbers. To convert the signal, sampling is necessary.

Sampling is the process of obtaining values of the analogue signal at discrete intervals of time$T$, where T is known as the sample period. The number of samples per second or the sampling frequency f_s in Hz is equal to the reciprocal of the sample period, that is, $f_s = 1/T$.

The sampling frequency that should be used in a given situation is determined by Nyquist's sampling theorem, which states that if the highest-frequency component present in the signal is f_h Hz, then the sampling frequency must be at least twice this value, that is,

$$f_s \geq 2f_h, \tag{2.1}$$

in order that the signal may be properly reconstructed from the digital samples. If fewer samples are used, then a phenomenon known as aliasing occurs, where a signal of a certain frequency may appear as a lower frequency upon reconstruction. If aliasing occurs in a complex signal such as speech, unwanted frequency components are inserted, which distort the signal.

Speech signals received by a microphone array are signals that are sampled so that their spectral features can be analyzed by carrying out a discrete Fourier transform [14] [15]. The Fourier transform is a mathematical operation that decomposes a signal into its constituent frequencies. Thus, the time-varying features of the Fourier transform of a musical chord are a mathematical representation of the amplitudes of the individual notes that comprise the chord. The original signal depends on time and is therefore called the time-domain representation, whereas the Fourier transform depends on frequency and is called the frequency-domain representation of the signal.

The use of a Fourier transform for the analysis of a speech signal has two major drawbacks: one is the computation cost, the other is that the temporal information of the speech signal is lost. A signal as a function of time may be considered as a representation with perfect time resolution. In contrast, the magnitude of the Fourier transform of the signal may be considered as a representation with perfect spectral resolution but with no time information because the magnitude of the Fourier transform conveys frequency content but fails to convey the time at which different events occur in the signal.

The T-F method provides a bridge between these two representations in that it provides some temporal information and some spectral information simultaneously. Thus, it is useful for the representation and analysis of signals containing multiple time-varying frequencies. The short-time Fourier transform (STFT) [16] is an effective means of transferring the signal from the time domain to T-F domain for analysis. In this section, a brief overview of the STFT and T-F analysis is given.

2.3.1 Short-time Fourier transform

The STFT is a Fourier-related transform used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time. Simply, in the continuous-time case, the function to be transformed is multiplied by a window function that is nonzero for only a short period of time. The Fourier transform (a one-dimensional (1D) function) of the resulting signal, which is taken as the window, is slid along the time axis, resulting in a 2D representation of the signal. Mathematically, the STFT is written as

$$STFT\{x(t)\} \equiv X(\tau,\ \omega) = \int_{-\infty}^{\infty} x(t)win(t-\tau)e^{-j\omega t}dt, \tag{2.2}$$

where $win(t)$ is the window function, which is commonly a rectangular Hanning or Hamming window centered around zero, and $x(t)$ is the signal to be transformed. $X(\tau,\ \omega)$ is the Fourier transform of $x(t)win(t-\tau)$, which is a complex function representing the phase and magnitude of the signal over time and frequency, τ is the time axis, and ω is the frequency axis.

In the discrete time case, the data to be transformed can be broken up into frames (which usually overlap each other to reduce artifacts at the boundary). Each frame is Fourier-transformed, and the complex-valued result is added to a matrix that records the magnitude and phase for each point in time and frequency. This can be expressed as

$$STFT\{x[n]\} \equiv X[k,l] = \sum_{n=0}^{N-1} x[n]win[n-k]e^{-j2\pi nl/N} \quad 0 \leq l < N, \tag{2.3}$$

with signal $x[n]$ and window $win[n]$. The square of the magnitude of the STFT yields the spectrogram of the function:

$$spectrogram\{x(t)\} \equiv |X[k,l]|^2. \tag{2.4}$$

The STFT has a fixed resolution. The width of the windowing function is related to how the signal is represented. It determines whether there is a good frequency resolution (frequency components close together can be separated) or good time resolution (the time at which frequencies change). A wide window gives a better frequency resolution but poor time resolution. A narrower window gives a good time resolution but poor frequency resolution. STFT with such windows are called narrowband and wideband transforms, respectively.

Window type

The role of the window in the STFT is to avoid the distortion that occurs at both ends of segmented signals and alters the spectra of signals. Thus, it is necessary to discuss the selection of the window used in the STFT. Three typical windows used for speech signal processing are introduced here.

Rectangular

$$win[n] = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & otherwise \end{cases} \tag{2.5}$$

Hanning

$$win[n] = \begin{cases} \frac{1}{2}(1 - \cos\frac{2\pi n}{N-1}) & 0 \leq n \leq N-1 \\ 0 & otherwise \end{cases} \tag{2.6}$$

Hamming

$$win[n] = \begin{cases} 0.54 - 0.46\cos\frac{2\pi n}{N-1} & 0 \leq n \leq N-1 \\ 0 & otherwise \end{cases} \tag{2.7}$$

The rectangular window quarries the target signal without weighting. This window does not modify the original signal, but the discontinuities at the end of the extracted frame have an undesirable effect on the spectrum. The Hanning and Hamming windows suppress this effect by eliminating the discontinuities in their temporal features.

The main lobe of both the Hamming and Hanning windows is twice as wide as that of the rectangular window, but the attenuation is much greater than that of the rectangular window. The secondary lobe of the Hanning window is 31 dB below the main lobe, whereas for the Hamming window it is 44 dB below the main lobe. On the other hand, the attenuation of the Hanning window decays with frequency rather rapidly, which is not the case for the Hamming window, whose attenuation remains approximately constant for all frequencies.

Inverse STFT

The time-domain output signal $y(n)$ is obtained by the inverse operation of the STFT:

$$y[n] = \sum_{k} win_s[n-k] \sum_{l=0}^{L-1} X[k,l] e^{-j2\pi nl/N}, \tag{2.8}$$

where $win_s[n-k]$ is a synthesis window that is nonzero only in the L-sample interval. To realize a perfect reconstruction, the analysis and synthesis windows should satisfy the condition

$$\sum_{k} win_s[n-k] win[n-k] = 1 \tag{2.9}$$

for any time n. This condition is satisfied by pairing a Hamming window for $win[n-k]$ and a rectangular window for $win_s[n-k]$. A synthesis window that tapers smoothly to zero at each end is also preferred in terms of mitigating the edge effect. In such a case, the above condition is satisfied, for instance, by using a square-root version of a Hamming window for both $win[n-k]$ and $win_s[n-k]$.

2.3.2 Harmonic structure

The fundamental distinction between sound types in speech is the voiced/voiceless distinction. Voiced sounds, including vowels, have a roughly regular pattern in their time and frequency structure that voiceless sounds. Voiced sounds typically contain more energy [17].

When the vocal folds vibrate during phoneme articulation, the phoneme is considered to be voiced; otherwise it is unvoiced. Vowels are voiced throughout their duration. The distinct vowel timbres are created by using the tongue and lips to shape the main oral resonance cavity in different ways. The vocal folds vibrate at various rates, from as low as 80 Hz for a large man to as high as 300 Hz for a woman or small child. The rate of cycling (opening and closing) of the vocal folds in the larynx during the phonation of voiced sounds is called the fundamental frequency. After the application of the STFT, the spectrum of a vowel has a harmonic structure. An example of such a spectrum is shown in Fig. 2.3.

2.4 Microphone array signal processing

A microphone array consists of a set of microphones positioned such that the spatial information is accurately captured. The main objective of microphone array signal processing is to estimate some parameters or to extract some signals of interest, depending on the appli-

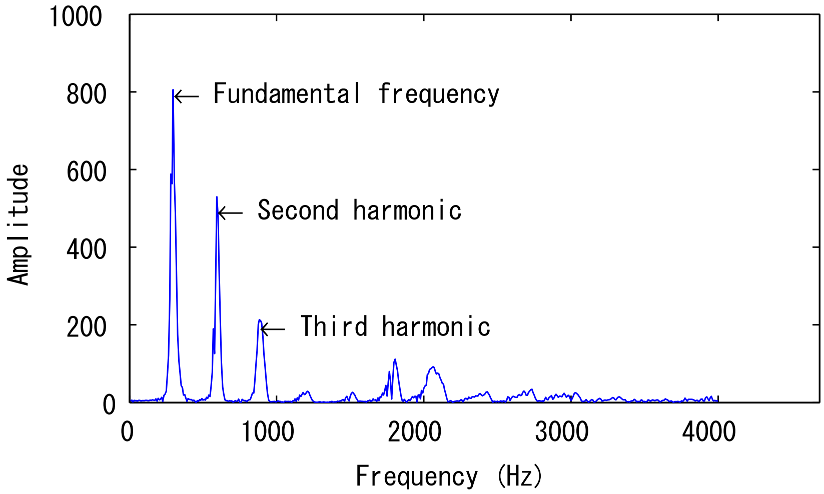


Figure 2.3: Spectrum of a vowel. The peaks are the harmonics of the vowel, while the fundamental frequency is approximately 297 Hz.

cation, by using the spatiotemporal and frequency information available at the output of the microphone array.

Depending on the nature of the application, the geometry of the microphone array may play an important role in the formulation of processing algorithms. For example, in source localization, the array geometry must be known to localize a source properly; moreover, sometimes a regular geometry will even simplify the problem of estimation, which is why uniform linear and circular arrays are often used [18]. Although these two geometries currently dominate the market more sophisticated 3D spherical arrays are becomingly increasingly widespread as they can capture the sound field better [19]. However, in some other crucial problems such as noise reduction and source separation, the geometry of the array may have little importance, depending on the algorithm. In this case, the system can be considered as a multiple microphone system instead of a microphone array.

The problems encountered in microphone arrays may appear easy to tackle because similar problems in narrowband antenna arrays have been tackled for a long time. However, this is misleading deceiving because microphone arrays work differently from antenna arrays, which are used for applications such as radar and sonar, for the following reasons [20]:

a) Speech is a wideband signal.

b) The reverberation of a room is high.

c) The environment and signals are highly nonstationary.

d) Noise can have the same spectral characteristics as the desired speech signal.

e) The number of sensors is usually restricted.

Because of these reasons, it is not surprising that for some problems, many existing algorithms do not perform well.

The main problems that have the potential to be solved using microphone arrays are as follows.

a) Noise reduction

b) Echo reduction

c) Dereverberation

d) Localization of a single source

e) Estimation of the number of sources

f) Source separation (Chapter 3)

g) Localization of multiple sources (Chapter 4)

2.5 Source mixing model

2.5.1 Instantaneous mixture

In instantaneous mixing, N unknown source signals s_i, $1 \leq i \leq N$, are combined to yield M measured sensor signals x_m, $1 \leq m \leq M$,

$$x_m(\tau) = \sum_{i=1}^{N} a_{mi} s_i(\tau), \tag{2.10}$$

where a_{mi} are the coefficients of the linear time-invariant mixing system represented by an $M \times N$ matrix.

The goal of BSS for instantaneous mixtures is to adjust the coefficients of the $N \times M$ separation or demixing matrix such that

$$y_i(\tau) = \sum_{m=1}^{M} b_{im}(\tau) x_m(\tau) \tag{2.11}$$

contains an estimate of a single original source.

2.5.2 Convolutive mixture

Let $s_1, s_2, \dots, s_N$ be N source signals. The observation x_{mi} at microphone m that originates only from source i is described by the convolutive model

$$x_{mi}(\tau) = \sum_{l=0}^{P} h_{mi}(l) s_i(\tau - l), \tag{2.12}$$

where τ represents the discrete time (a multiple of $\tau_s = 1/f_s$ with f_s being the sampling frequency) and h_{mi} is the impulse response from source i to microphone m modeled with $P + 1$ samples. A situation where the N sources are simultaneously active is considered. Hence, the observation x_m at microphone m is modeled by the convolutive mixture model as follows.

$$x_m(\tau) = \sum_{i=1}^{N} x_{mi}(\tau) = \sum_{i=1}^{N} \sum_{l=0}^{P} h_{mi}(l) s_i(\tau - l) \tag{2.13}$$

The convolutive mixture model is shown in Fig. 2.4.

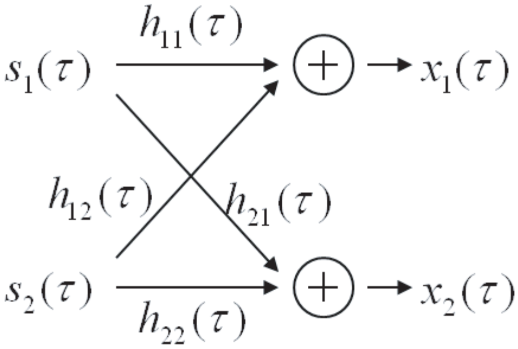


Figure 2.4: Convolutive mixture model.

The goal is to obtain separate signals $y_1, y_2, \dots, y_N$, each of which corresponds to one of the source signals $s_1, s_2, \dots, s_N$. The task should be performed with only M observed mixtures $x_1, x_2, \dots, x_M$, and without information on the sources s_i or the impulse response h_{mi}.

First, each of the time-domain microphone observations $x_m(t)$ is converted into frequency-domain time-series signals $X_m[k, l]$ by an STFT with an L-sample frame and its S-sample shift:

$$X_m[k,l] \leftarrow \sum_\tau x_m(\tau) win(\tau + kS) e^{-j\frac{2\pi l}{L}\tau}, \tag{2.14}$$

for all discrete frequencies bin l, and for frame index k. Here $win(\tau)$ is the analysis window, such as a Hanning window.

If the frame size L is sufficiently long to cover the main part of the impulse response h_{mi}, the convolutive model can be approximated as an instantaneous model at each frequency:

$$X_{mi}[k,l] = H_{mi}[l] S_i[k,l], \tag{2.15}$$

where $H_{mi}[l]$ is the frequency response from source i to microphone m, and $S_i[k,l]$ are the frequency-domain time-series signals of $s_i(\tau)$. Consequently, the convolutive mixture model can be treated as an instantaneous mixture model:

$$X_m[k,l] = \sum_{i=1}^{N} X_{mi}[k,l] = \sum_{i=1}^{N} H_{mi}[l] S_i[k,l]. \tag{2.16}$$

2.5.3 Evaluation of separation performance

The separation performance of i-th source is evaluated in terms of the signal-to-interference ratio (SIR) improvement, which is defined as

$$SIR\ \ improvement = OutputSIR_i - InputSIR_i, \tag{2.17}$$

where

$$InputSIR_i = 10\log_{10}\frac{\sum_\tau |\sum_l h_{Ji}(l) s_i(\tau - l)|^2}{\sum_\tau \sum_{n\neq i} |\sum_l h_{Jn}(l) s_n(\tau - l)|^2}(\mathrm{dB}) \tag{2.18}$$

$$OutputSIR_i = 10\log_{10}\frac{\sum_\tau |y_{ii}(\tau)|^2}{\sum_\tau |\sum_{n\neq i} y_{in}(\tau)|^2}(\mathrm{dB}). \tag{2.19}$$

Here $J \in 1, \cdots, M$ is the index of a selected reference sensor, and $y_{in}(\tau)$ is the component of $s_n(\tau)$ that appears at output $y_i(\tau)$: $y_i(\tau) = \sum_{n=1}^{N} y_{in}(\tau)$.

2.6 Blind source separation

We are surrounded by sounds. A noisy environment makes it difficult to understand desired speech and to converse easily. This makes it important to be able to separate and extract a target speech signal from noisy observations to enable both human-machine and human-human communication.

The technique for estimating individual source components from their mixtures at multiple sensors is known as BSS. The estimation is performed blindly, i.e., without possessing information about the mixing of the sources, such as the source location and its active time periods.

One well-recognized BSS application is the separation of audio sources that have been mixed and then captured by multiple microphones in a real room environment. A simple flow of BSS using a microphone array is shown in Fig. 2.5. The signals received by the microphones are mixed signals. Then, using the separation system, the separated signals are derived.

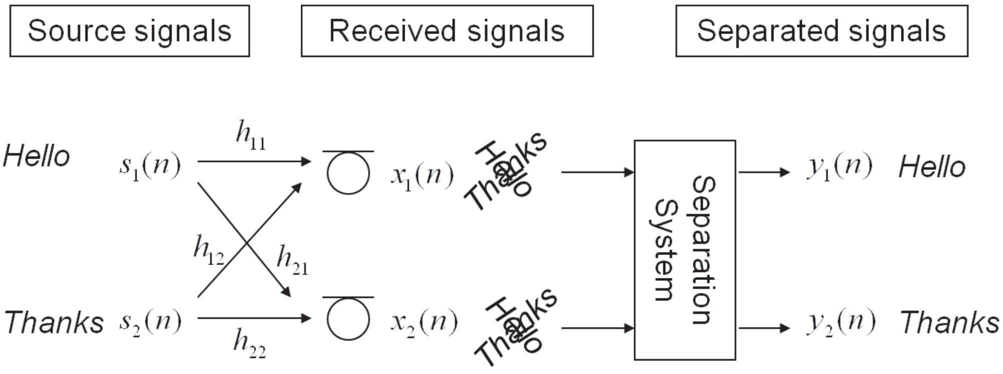


Figure 2.5: Blind source separation using microphone array

The difficulty of BSS lies in the fact that the mixed system is not simply instantaneous but convolutive, with delay and reflections. Such a mixing situation is generally modeled using the impulse responses from the sound sources to the microphones. In a practical room situation, such impulse responses can have thousands of taps even with an 8 kHz sampling rate, and this makes the convolutive problem difficult to solve.

Another challenge in BSS is to determine the number of sources, especially when there are more sources than sensors, because the mixing matrix is not invertible. Thus, the demixing method involving the estimation of the inverse mixing matrix does not work. At the same time, the reverberation and noise under real acoustic conditions make the received mixed data ambiguous.

Many methods have been proposed to resolve BSS problems involving speech signals. Among them, the most widely applied approaches are

1. the independent-component-analysis (ICA)-based approach [21]

2. the sparseness of source signals in the T-F-domain-based approach [22–25].

ICA relies on the statistical independence of speech signals and performs well even under a reverberant condition. However, it has difficulty in solving underdetermined cases in which the number of sources is greater than the number of sensors. The sparseness-based methods are applied to observed signals transformed from the time domain to the T-F domain by an STFT and use the sparse representation of speech in the T-F domain [26]. They are valid even in underdetermined cases. One sparseness-based approach is to estimate the mixing matrix at the first stage of the separation process [27]. The other approach utilizes a T-F binary mask. This group of methods is based on an assumption known as the W-disjoint orthogonality (WDO) of speech signals [22]. That is, although the observed signal is a mixture of several sources, its T-F cell contains at most one of the components of a source signal. A similar but weaker assumption than that of WDO has been proposed and used for the BSS problem [28] [29].

2.6.1 Independent component analysis

ICA is a statistical method for extracting mutually independent sources from their mixtures, and it relies on the statistical independence of speech signals. If the sources are to be separated blindly, they should have some distinct characteristics, such as non-Gaussianity, nonstationarity, or nonwhiteness. ICA, which is sometimes regarded as synonymous with BSS, relies on non-Gaussianity. Many textbooks have been published on BSS and ICA [30–33]. In this section the mathematical nortations of variables are followed by the reference [33], thus there are different symbols used in other sections.

In the frequency domain, the signals observed by microphones can be modeled using a T-F representation computed by an STFT. According to the convolutive model for the observed mixture, each T-F component can be considered as a linear combination of the T-F components of the original source signals. In matrix notation, one can write

$$X(t, f) = H(f)S(t, f), \tag{2.20}$$

where $X(t, f)$ are the observed mixtures, $S(t, f)$ are the original signals, t is the time instant at which each frequency is evaluated with the shift of the time frame, f is the frequency bin index, and $H(f)$ is the mixing matrix. A complex-valued ICA is applied to the time series of each frequency. Then the original components can be retrieved by computing a demixing matrix $W(f)$, which is an estimate of the matrix $H^{-1}(f)$ up to scaling and permutation ambiguities:

$$\hat{S}(t, f) = A(f)P(f)W(f)X(t, f), \tag{2.21}$$

where $A(f)$ and $P(f)$ are a complex-valued scaling matrix and a permutation matrix, respectively.

Many ICA algorithms for calculating the separation matrix W have been reported [30–33]. Although the source separation algorithms are different, their principles can be summarized by the following four approaches:

1. The most popular approach exploits some measure of signal independence, non-Gaussianity, or sparseness as the cost function. When the original sources are assumed to be statistically independent without a temporal structure, higher-order statistics are essential (implicitly or explicitly) to solve the BSS problem. In such a case, the method does not allow more than one Gaussian source.

2. If the sources have temporal structures, then each source has a nonvanishing temporal correlation and less restrictive conditions than statistical independence can be used, namely, second-order statistics are sufficient to estimate the mixing matrix and sources. Several methods have been developed along this line. Note that these methods based on second-order statistics methods do not allow the separation of sources with identical power spectra or independent and identically distributed sources.

3. The third approach exploits nonstationarity properties and second-order statistics. Mainly, we are interested in second-order nonstationarity in the sense that source variances vary in time. Nonstationarity was first considered by Matsuoka et al. [34], and it was shown that a simple decorrelation technique is able to perform the BSS task. In contrast to other approaches, nonstationarity-information-based methods allow the separation of colored Gaussian sources with identical power spectra. However, they do not allow the separation of sources with identical nonstationarity properties. There have been some studies on nonstationary source separation [35–37].

4. The fourth approach exploits the various diversities of signals, typically time, frequency (spectral or "time coherence") and/or T-F diversities, or more generally, joint space-time-frequency diversity.

Although ICA is a good method for speech separation, the permutation ambiguity of an ICA solution is a serious problem. The ambiguities should be aligned suitably so that the separated frequency components that originate from the same source are grouped together.

Various approaches have been proposed to solve the permutation problem. [38, 39, 39] by making the separation matrix W smooth in the frequency domain. This can be realized simply by windowing the separation filters in the time domain. However, this operation makes the separation matrix W different from the ICA solution and generally degrades the separation performance. In [40–42] information related to the source location is estimated, such as the DOA or time difference of arrival (TDOA). The beamforming approach analyzes the directivity patterns formed by the separation matrix W to identify the DOA of each source. However, analysis of the directivity patterns is only practically possible for a two-source case and becomes intractable when there are more sources. [43] [44] exploit the dependence of separated signals across frequencies. The advantage is that they are less affected by the mixing system under unfavorable conditions such as severe reverberations or closely located sources.

2.7 Time-frequency masking

The T-F masking method is an important technique for speech source separation [45]. Usually the received signals in the time domain are transfered to the T-F domain by an STFT. Then, on the basis of the features contained in the T-F domain, a T-F mask is designed that can be used to obtain the separated signals. Finally, the separated signals in a waveform representation are synthesized from the T-F representation by an inverse STFT.

The degenerate unmixing estimation technique (DUET) [22], sound source Segregation based on estimation incident Angle of each Frequency component of Input signals Acquired by multiple microphones (SAFIA) [23] and [46] are typical conventional approaches essentially based on the WDO assumption [47] [48], which means that although the observed signal is a mixture of several sources, its T-F cell contains at most one of the components of a source signal. These geometric parameters of the sources are estimated using the phase difference and intensity difference between two sensor observations. If the WDO assumption holds and these parameters are accurately estimated, the histogram in terms of the frequency-normalized phase difference, or time arrival delay, and the attenuation ratio at the mixed T-F cells from clusters corresponding to individual sources. Then, the essential problem in the separation becomes the development of a clustering algorithm. The preliminary clustering method adopted in Refs. [22] and [46] is to first find the peaks corresponding to the sources in the histogram, then each T-F cell in the mixed signal is associated with one peak depending on the distance in the cell's feature space. Finally, the reconstruction of the source signals is performed by masking the STFT-domain spectrogram of a mixture. In subsequent studies, the kernel density method and a maximum-likelihood (ML)-based method were proposed to enable real-time operation [49] . In addition, the *k*-means algorithm or hierarchical clustering has been applied to realize automatic and simplified separation [24] [25]. The method called Multiple sENsor dUET (MENUET) [24] applies the k-means algorithm to a vector space of the signal level ratio and the frequency-normalized phase difference with appropriately weighted terms to ensure effective clustering. Generally, k-means clustering is performed by minimizing a cost function in the spatial feature space. The optimization problem has been solved by developing an efficient iterative update algorithm. In another study [25], the k-means algorithm was also applied for an arbitrary sensor array configuration, even one with a greater distance between sensors at which spatial aliasing may occur.

2.7.1 DUET

The DUET [22] is a well-known method of solving the BSS problem by T-F masking. In this section, the DUET is outlined. The method is valid when sources exhibit WDO, that is, when the supports of the windowed Fourier transform of the signals in the mixture are disjoint. For anechoic mixtures of attenuated and delayed sources, the method allows one to estimate the mixing parameters by clustering relative attenuation-delay pairs extracted from the ratios of the T-F representations of the mixtures. The estimates of the mixing parameters are then used to partition the T-F representation of one mixture to recover the original sources. The technique is valid even in the case when the number of sources is larger than the number of mixtures.

The DUET separates degenerate mixtures by partitioning the T-F representation of one of the mixtures. In other words, the DUET assumes that the sources are already ’separate’ in the sense that, in the T-F plane, the sources are disjoint. The ’demixing’ process is then simply the partitioning of the T-F plane. Although the assumption of disjointness may appear unreasonable for simultaneous speech, it is approximately true in the sense that the T-F points that contain significant contributions to the average energy of the mixture are very likely to be dominated by a contribution from only one source. In other words, two people rarely produce the same frequency at the same time.

Consider mixtures of N source signals, $i = 1, \cdots, N$, being received at a pair of microphones, where only the direct path is allowed. The two anechoic mixtures can thus be expressed as

$$x_1(\tau) = \sum_{i=1}^{N} s_i(\tau), \tag{2.22}$$

$$x_2(\tau) = \sum_{i=1}^{N} h_i s_i(\tau - \delta_i), \tag{2.23}$$

where N is the number of sources, δ_i is the difference in the arrival time between the sensors, and h_i is a relative attenuation factor corresponding to the ratio of the attenuations of the paths between the sources and sensors.

Two functions $s_i(\tau)$ and $s_j(\tau)$ are called W-disjoint orthogonal if, for a given windowing function $win(\tau)$, the supports of the windowed Fourier transforms of $s_i(\tau)$ and $s_j(\tau)$ are disjoint. The windowed Fourier transform of $s_j(t)$ is $\hat{S}_i[k,l]$. The WDO assumption can be stated concisely as

$$\hat{S}_i[k,l]\hat{S}_j[k,l] = 0, \forall k, l, \quad \forall i \neq j. \tag{2.24}$$

This assumption is the mathematical idealization of the condition that every T-F point in the mixture with significant energy is dominated by the contribution of one source. WDO is crucial to the DUET because it allows the separation of a mixture into its component sources using a binary mask. An example of the WDO property is shown in Fig. 2.6.

The assumption of anechoic mixing allows the mixing equations (2.22) and (2.23) in the T-F domain to be written as

$$\begin{bmatrix} \hat{X}_1[k,l] \\ \hat{X}_2[k,l] \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ h_1 e^{-j\frac{2\pi l}{L}\delta_1} & \cdots & h_N e^{-j\frac{2\pi l}{L}\delta_N} \end{bmatrix} \begin{bmatrix} \hat{S}_1[k,l] \\ \vdots \\ \hat{S}_N[k,l] \end{bmatrix} \tag{2.25}$$

Because of the assumption of WDO, at most one source is active at every $[k,l]$, and the mixing process can be described as

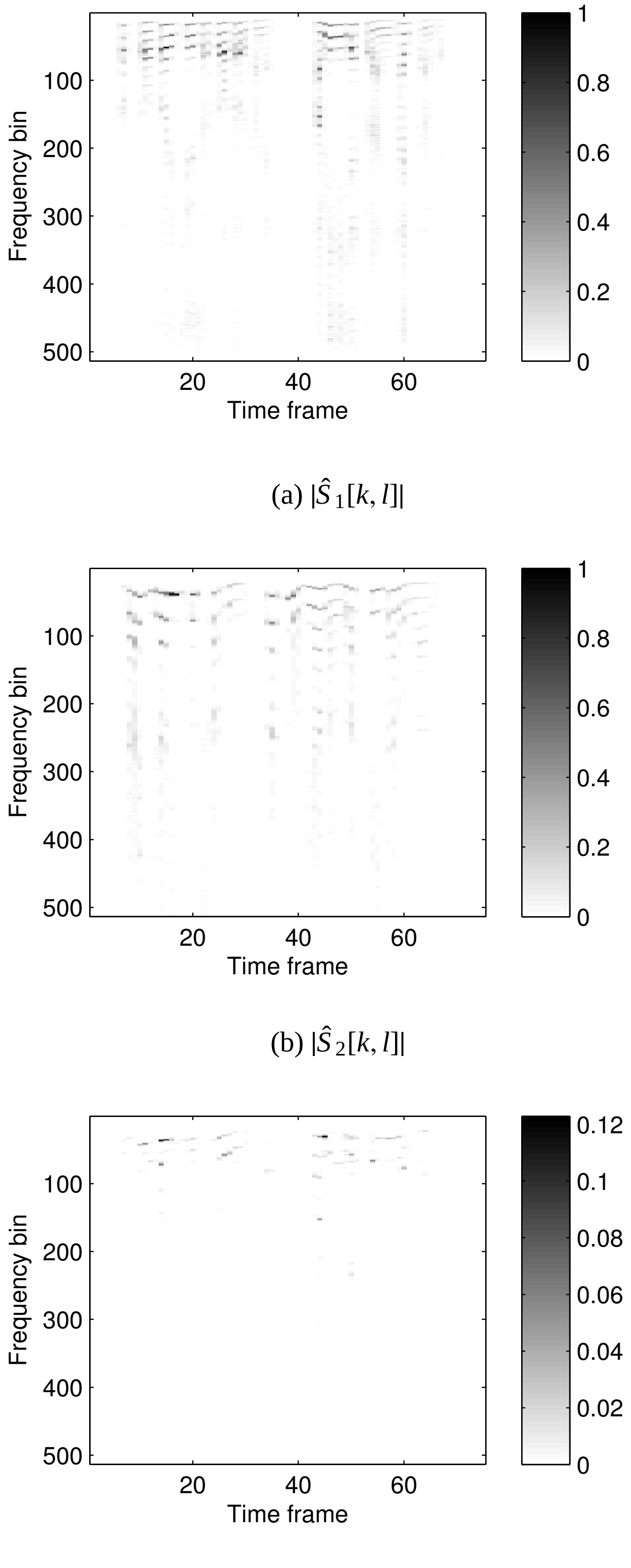
$$\textit{for each } [k,l], \quad \begin{bmatrix} \hat{X}_1[k,l] \\ \hat{X}_2[k,l] \end{bmatrix} = \begin{bmatrix} 1 \\ h_i e^{-j\frac{2\pi l}{L}\delta_i} \end{bmatrix} \hat{S}_i[k,l] \quad \textit{for some i.} \tag{2.26}$$

The main observation that the DUET leverages is that the ratio of the T-F representations of the mixtures does not depend on the source components but only on the mixing parameters associated with the active source component:

$$\forall [k,l] \in \Omega_i, \quad \frac{\hat{X}_2[k,l]}{\hat{X}_1[k,l]} = h_i e^{-j\frac{2\pi l}{L}\delta_i}, \tag{2.27}$$

where

$$\Omega_i := \{[k,l] : \hat{S}_i[k,l] \neq 0\}. \tag{2.28}$$



(a) $|\hat{S}_1[k,l]|$

(b) $|\hat{S}_2[k,l]|$

(c) $|\hat{S}_1[k,l] \cdot \hat{S}_2[k,l]|$

Figure 2.6: Example of WDO property. s_1 is a male source located at 0° and s_2 is a female source located at 50°.

The local attenuation estimator $\tilde{\alpha}[k,l]$ of the mixing parameters and the local delay estimator $\tilde{\delta}[k,l]$ associated with each T-F point can be calculated as

$$\tilde{\alpha}[k,l] := |\hat{X}_2[k,l]/\hat{X}_1[k,l]|, \tag{2.29}$$

$$\tilde{\delta}[k,l] := (-\frac{L}{2\pi l})\angle(\hat{X}_2[k,l]/\hat{X}_1[k,l]). \tag{2.30}$$

The set of points that contribute to a given location in the histogram is defined by

$$I(\alpha,\delta) := \{[k,l] | |\tilde{\alpha}[k,l]-\alpha| < \Delta_\alpha, |\tilde{\delta}[k,l]-\delta| < \Delta_\delta\}, \tag{2.31}$$

where Δ_α and Δ_δ are the smoothing resolution widths. The 2D smoothed weighted histogram is constructed as

$$H(\alpha,\delta) := \sum_k \sum_l |\hat{X}_1[k,l]\hat{X}_2[k,l]|^p l^q, \quad [k,l] \in I(\alpha,\delta), \tag{2.32}$$

where p and q are parameters, and N peaks corresponding to the N sources are clearly visible in the histogram. Given the histogram peak centers $\tilde{\alpha}_i, \tilde{\delta}_i, i = 1, \cdots, N$, the symmetric attenuation is converted back to the attenuation via

$$\tilde{a}_i = \frac{\tilde{\alpha}_i + \sqrt{\tilde{\alpha}_i^2 + 4}}{2}, \tag{2.33}$$

with a peak assigned to each T-F point via

$$J[k,l] := \arg\min_j \frac{|\tilde{a}_j e^{-j\tilde{\delta}_j \frac{2\pi l}{L}} \hat{X}_1[k,l] - \hat{X}_2[k,l]|^2}{1+\tilde{a}_j^2}. \tag{2.34}$$

Then each T-F point is assigned to an estimate of the mixing parameter via

$$M_i[k,l] := \begin{cases} 1, & \text{if } J[k,l] = i \\ 0, & \text{otherwise .} \end{cases} \tag{2.35}$$

M_i is used to separates $\hat{S}_i$ from the mixture via

$$\hat{S}_i[k,l] = M_i[k,l]\hat{X}_1[k,l]. \tag{2.36}$$

The sources can be reconstructed from their T-F domain by converting them back into the time domain.

The DUET algorithm is summarized as follows:

1. Construct T-F representations $\hat{X}_1[k,l]$ and $\hat{X}_2[k,l]$ from mixtures $x_1(\tau)$ and $x_2(\tau)$, respectively.

2. Calculate $(|\frac{\hat{X}_2[k,l]}{\hat{X}_1[k,l]}| - |\frac{\hat{X}_1[k,l]}{\hat{X}_2[k,l]}|, \frac{-L}{2\pi l}\angle(\frac{\hat{X}_2[k,l]}{\hat{X}_1[k,l]}))$.

3. Construct a 2D smoothed weighted histogram using (2.32).

4. Locate the peaks and peak centers that determine the estimates of the mixing parameter.

5. Construct T-F binary masks for each peak center using (2.35).

6. Apply each mask to the appropriately aligned mixtures using (2.36).

7. Convert each estimated source T-F representation back into the time domain.

To measure the WDO of a given mask, two criteria, the preserved-signal ratio (PSR) and the SIR, are introduced.

$$PSR_M := \frac{||M_i[k,l]\hat{S}_i[k,l]||^2}{||\hat{S}_i[k,l]||^2} \tag{2.37}$$

$$SIR_M := \frac{||M_i[k,l]\hat{S}_i[k,l]||^2}{||M_i[k,l]\hat{Y}_i[k,l]||^2} \tag{2.38}$$

Here, $y_i(\tau)$ is the sum of the sources interfering with the ith source.

$$y_i(\tau) := \sum_{j=1, i\neq j}^{N} s_j(\tau) \tag{2.39}$$

Then, PSR_M and SIR_M are combined into a single measure of WDO, WDO_M:

$$WDO_M := \frac{||M_i[k,l]\hat{S}_i[k,l]||^2 - ||M_i[k,l]\hat{Y}_i[k,l]||^2}{||\hat{S}_i[k,l]||^2} \tag{2.40}$$

$$= PSR_M - \frac{PSR_M}{SIR_M}.$$

$WDO_M = 1$ implies that mask M perfectly separates the ith source from the mixture.

2.7.2 MENUET

Similar to the DUET, Araki et al. proposed the binary mask approach MENUET [24], which employs the k-means clustering algorithm. The novel feature of this method is that it utilizes the level ratios and phase differences between multiple observations. To realize level ratio and phase difference variances of a comparable level, they proposed a method of weighting the phase term used for clustering. Moreover, their method does not require sensor location information. This allows freely arranged multiple sensors to be employed. Therefore, the method can separate signals that are distributed two- or three-dimensionally.

Feature extraction

If the sources $S_i[k,l]$ are sufficiently sparse, separation can be realized by accumulating the T-F points $[k,l]$ where only one signal s_i is estimated to be dominant. To estimate such T-F points, some features $\Theta[k,l]$ are calculated using the frequency-domain observation signals $\hat{X}[k,l]$.

The novel feature employed in MENUET is expressed as

$$\Theta[k,l] = [\Theta^L[k,l], \Theta^P[k,l]]^T, \tag{2.41}$$

where

$$\Theta^L[k,l] = \left[\frac{|\hat{X}_1[k,l]|}{A[k,l]}, \cdots, \frac{|\hat{X}_M[k,l]|}{A[k,l]}\right] \tag{2.42}$$

is the observation-level information,

$$\Theta^P[k,l] = \left[\frac{1}{\alpha_1 l'} \arg\left[\frac{\hat{X}_1[k,l]}{\hat{X}_J[k,l]}\right], \cdots, \frac{1}{\alpha_M l'} \arg\left[\frac{\hat{X}_M[k,l]}{\hat{X}_J[k,l]}\right]\right] \tag{2.43}$$

is the phase difference information, $l' = \frac{2\pi l}{L}$, $A[k,l] = \sqrt{\Sigma_{m=1}^{M} |\hat{X}_m[k,l]|^2}$, J is the index of one of the sensors, and $\alpha_m (m = 1, \cdots, M)$ is a positive weighting constant. By changing α_m, the weights for the level ratio and the normalized phase difference information of the observed signals can be controlled. A larger value increases the weight of the level ratio and a smaller value emphasizes the phase difference.

Clustering

The features $\Theta[k,l]$ are grouped into N clusters $C_1, \cdots, C_N$, where N is the number of possible sources. In MENUET, the k-means clustering algorithm is used with a given number of sources N. The clustering criterion is to minimize the sum ξ of the Euclidean distances (EDs) between cluster members and their centroid $\mathbf{c}_i$,

$$\xi = \sum_{i=1}^{N} \xi_i, \tag{2.44}$$

where

$$\xi_i = \sum_{\Theta[k,l] \in C_i} ||\Theta[k,l] - \mathbf{c}_i||^2. \tag{2.45}$$

After setting an appropriate initial centroid $\mathbf{c}_i (i = 1, \cdots, N)$, ξ can be minimized by the following iterative updates:

$$C_i = \{\Theta[k,l] | i = \arg\min_i ||\Theta[k,l] - \mathbf{c}_i||^2\} \tag{2.46}$$

$$\mathbf{c}_i \leftarrow E[\Theta[k,l]]_{\Theta\in C_i}, \tag{2.47}$$

where $E[\cdot]_{\Theta\in C_i}$ is a mean operator for the members of cluster C_i. If the feature $\Theta[k,l]$ is suitably chosen, each cluster corresponds to an individual source.

Mask design

Next, the separated signals $Y_i[k,l]$ are estimated on the basis of the clustering results. A T-F domain binary mask is designed as follows that extracts the T-F points of each cluster

$$M_i[k,l] = \begin{cases} 1, & \Theta[k,l] \in C_i \\ 0, & \text{otherwise.} \end{cases} \tag{2.48}$$

Then, applying the binary mask as above to one of the observations $\hat{X}_J[k,l]$, the separated signals can be obtained as

$$Y_i[k,l] = M_i[k,l]\hat{X}_J[k,l], \tag{2.49}$$

where J is a selected sensor index. Finally, by employing an inverse STFT (ISTFT), the separated signals are obtained in the time domain.

2.8 DOA estimation

A major function of microphone array signal processing is the estimation of the location from which a source signal originates. Depending on the distance between the source and the array relative to the array size, this estimation problem can be divided into two subproblems, i.e., DOA estimation and source localization.

DOA estimation deals with the case where the source is in the array's far-field. In this situation, the source radiates a plane wave with a waveform that propagates through the nondispersive medium of air. The normal to the wavefront makes an angle θ with the line joining the sensors in the linear array, and the signal received at each microphone is a time-delayed version of the signal at a reference sensor. In other words, the DOA estimation problem is the same as the so-called TDOA estimation in the far-field case.

Although the incident angle can be estimated using two or more sensors, the distance between the sound source and the microphone array is difficult to determine if the source is in the array's far-field. However, if the source is located in the near-field, it becomes possible to estimate not only the angle from which the wave ray reaches each sensor but also the distance between the source and each microphone. A problem is called source localization. All the information regarding the source position relative to the array can be determined using the triangulation rule once the TDOA information is available. This basic triangulation process forms the foundation for most source localization techniques.

Regardless of whether the source is located in the far-field or near-field, the most fundamental step in obtaining information on the source origin is that of estimating the TDOA between different microphones. This estimation problem would be an easy task if the received signals were merely a delayed and scaled version of each other. In reality, however, the source signal is generally immersed in ambient noise since we live in an environment where the existence of noise is inevitable. Furthermore, each observation signal is reflected from boundaries and objects. This multipath propagation effect introduces echoes and spectral distortion into the observation signal, termed as reverberation, which severely deteriorates the source signals. In addition, the source may also move, resulting in a changing time delay. All these factors make DOA estimation and source localization a complicated and challenging problem.

A large number of DOA estimation methods have been proposed [50] [51]. Typical array-processing approaches are as follows.

1. The generalized cross-correlation (GCC) method [52].

2. Signal subspace approaches to determine the spatial covariance matrix of observed signals [53].

3. Multiple-source localization methods based on clustering in the T-F decomposition space known as histogram mapping [46] (DUET [22], DEMIXb [29] and others [54]).

4. ICA-based approaches [55] [56].

Their features of these methods are summarized in Tab. 2.1.

Table 2.1: Summary of typical DOA estimation methods

Typical DOA estimation methods	Features
Generalized cross-correlation (GCC)	Single-source model
Signal subspace (MUSIC et al.)	Number of sensors > Number of sources
Time-frequency sparseness	Unlimited sources
Independent component analysis (ICA)	Number of sensors ≥ Number of sources

2.8.1 Generalized cross-correlation method

The most widely used approach for estimating DOA/TDOA using a pair of microphones is the GCC algorithm [52]. It estimates the delay time that maximizes the cross-correlation function between the filtered outputs of the signals acquired at the microphones. Among the many variations of the GCC method, the phase transform (PHAT) method [57] is closely related to the time-frequency sparseness method. The PHAT method exploits the fact that the TDOA information is conveyed in the phase, therefore, the generalized cross-spectrum in the PHAT method is given by a delay operator component by neglecting the amplitude characteristic. Although GCC methods are usually successfully employed and are also computationally efficient, they basically employ a single-source model.

Among the four major methods of speaker direction estimation, the time-delay-estimation-based method possesses a significant computational advantage over the other methods, and it is used in many speaker direction estimation systems. The time delay estimation method has a two-step procedure. First, it estimates the time arrival difference between the signals relative to a pair of spatially separated microphones. Using these values in combination with the known microphone positions, the direction of the sound source is estimated.

Suppose that a plane sound wave is received by a pair of microphones. If the time arrival difference τ_s is estimated, the sound source direction θ_s is given by

$$\theta_s = \sin^{-1}(c\tau_s/d) \tag{2.50}$$

where c is the sound velocity and d is the distance between microphones. For the estimation of τ_s the GCC function is the most popular method defined as

$$R(\tau) = \int_{-\infty}^{\infty} \Psi(\omega) G_{x_0 x_1}(\omega) e^{j\omega\tau} d\omega, \tag{2.51}$$

where $G_{x_0 x_1}(\omega)$ is the cross spectrum of the received signals $x_0(t)$ and $x_1(t)$ and is given by

$$G_{x_0 x_1}(\omega) = X_0(\omega) X_1^*(\omega). \tag{2.52}$$

$\Psi(\omega)$ is a frequency weighting filter and $*$ denotes the complex conjugate. By searching for the value of τ_s that gives the largest peak of $R(\tau)$, the speaker direction can be estimated.

Since accurate and robust time delay estimation is the key to effective direction estimation in this area, several frequency-weighting filters $\Psi(\omega)$ are selected for use in the GCC function. There are two main interfering sources that degrade the estimation performance, which are nondirectional background noise and the multipath channel due to room reverberation. To cope with the former type of interference, an ML-based function has been proposed [52]. Because this weighting function is based on the signal-to-noise (SNR) ratio at each frequency, it is appropriate for reducing the effects of spatially uncorrelated white noise. However, in the case of room reverberation, these ML-based methods exhibit severe performance degradation.

In contrast, the basic approach to dealing with multipath channel distortions is to make the GCC function more robust by deemphasizing the frequency weightings. The PHAT, given by

$$\Psi_{PHAT}(\omega) = \frac{1}{|X_0(\omega)X_1^*(\omega)|} = \frac{1}{|G_{x_0x_1}(\omega)|}, \tag{2.53}$$

is one of the weighting functions that has received considerable attention. By placing equal emphasis on each frequency component, the resulting peak in the GCC-PHAT function that corresponds to the dominant delay can be clearly observed. Although the GCC-PHAT function is effective for reducing the degradation due to the multipath channel, it emphasizes the components of the spectrum with a poor SNR, particularly in the case of low reverberation.

Furthermore, other approaches for selecting the frequency-weighting function in adverse environments are available. Brandstein et al. utilized a criterion based on a speech-specific harmonic structure in the Fourier spectrum [58].

2.8.2 MUSIC

The second category of conventional DOA estimation algorithms is based on subspace analysis exploiting a statistical narrowband array model [59] [60]. MUltiple SIgnal Classification (MUSIC) is a typical method categorized into the subspace approach. It is an algorithm used for frequency estimation [61] and emitter location [53]. Schmidt was the first to correctly exploit the measurement model in the case of sensor arrays of arbitrary form, and he accomplished this by first deriving a complete geometric solution in the absence of noise, then cleverly extending the geometric concepts to obtain a reasonable approximate solution in the presence of noise. The resulting algorithm was named MUSIC and has been widely studied. For broadband signals such as speech, several frequency-domain approaches have been proposed. Among them, the coherent signal subspace method [62] is effective. However, subspace-based approaches with two microphones must overcome two drawbacks, one of which is the limited precision of DOA estimation, and the other is that it is unable to deal with underdetermined cases.

The broadband formulation of MUSIC algorithm is derived based on the eigenvalue decomposition of the spatial correlation matrix. The noise eigenvectors are used for calculating the MUSIC cost function and its peak points will correspond to the true TDOA.

2.8.3 Time-frequency clustering method

The third category of the DOA estimation algorithms closely related to the BSS approaches are based on the source sparseness assumption, known as WDO, and its weaker condition TIme-Frequency Ratio Of Mixtures (TIFROM) [28] [29]. These conditions are crucial properties of speech signals used to solve the DOA for underdetermined multiple sources. The BSS approach associated with these assumptions is a group of T-F masking frameworks, such as the classical methods of the DUET [22] and SAFIA [23] and the recent approaches of TIFROM [28] and MENUET [24]. The assumption of WDO implies that the mixed sources have essentially disjoint T-F supports, i.e., only one source is dominant in a T-F cell of the mixtures. The representative methods are the DUET [22], and others [46] [54]. The common approach of these methods is to estimate information about each source and use it to identify the attribute of each T-F cell. The clustered T-F cells that belong to one of the sources are used for the separation and DOA estimation of individual sources. In DUET-like methods [46] [54] [63] as well as SAFIA [23], the delay time or the frequency-normalized ratio of the frequency-domain observations at each T-F point is used to compute the TDOA. To obtain a global estimate of the mixing parameters (attenuation ratio and delay) from these local individual estimates, DUET-like methods use a weighted smoothed histogram. An alternative DOA estimation method proposed by Araki et al. [54], especially in the context of BSS, estimates the DOA as the centroid of each cluster of normalized observation vectors corresponding to an individual source. Note that their method can deal with arbitrary sensor configurations including 3D arrangements.

TIFROM [28] exploits a weaker assumption than WDO as follows. In the neighborhood of some T-F cells, only one source essentially contributes to the mixture. These T-F points provide a robust local estimation of the DOA corresponding to each source direction on the basis of the TIFROM concept, the Direction Estimation of Mixing matrIX (DEMIX) [29] algorithm introduces a statistical model to exploit a local confidence measure to detect the regions where robust mixing information is available. This algorithm is also based on a clustering algorithm that gives more weight to more reliable T-F regions according to the introduced confidence measure. However, the computational cost of DEMIX is high owing to the performance of principal component analysis for every local scatter plot of observation vectors at individual T-F points.

Araki et al. proposed a DOA estimation method for underdetermined cases involving

T-F decomposition [54]. Their method is based on the normalization and clustering of the observation vectors. Let $\mathbf{q}_i$ be the 3D vector of a unit norm representing the direction of source s_i. The location of sensor j is given by the 3D vector $\mathbf{d}_j$. The sensor observations are $\mathbf{x}(t, f)$. Using the azimuth θ_i and elevation ϕ_i, the DOA $\mathbf{q}_i$ can be written as

$$\mathbf{q}_i = [\cos\theta_i \cos\phi_i, \sin\theta_i \cos\phi_i, \sin\phi_i]^T. \tag{2.54}$$

An anechoic model is assumed, that is, the frequency response $h_{ji}(f)$ is expressed solely interms of the time delay $\tau_{ji} = \mathbf{d}_j^T \mathbf{q}_i / c$ with respect to the origin:

$$h_{ji}(f) \approx \exp[j2\pi f \mathbf{d}_j^T \mathbf{q}_i / c], \tag{2.55}$$

where c is the propagation velocity of the signals.

In the first step, Araki et al. applied unit-norm normalization to all observation vectors $\mathbf{x}(t, f)$,

$$\mathbf{x}(t, f) \leftarrow \mathbf{x}(t, f) / ||\mathbf{x}(t, f)||. \tag{2.56}$$

Then, in the clustering step, the normalized vectors $\mathbf{x}(t, f)$ are clustered into N clusters $C_1, \cdots, C_N$. The clustering criterion is to minimize the sum ζ of the squared distances between the cluster members and their centroid:

$$\zeta = \sum_{k=1}^{M} \zeta_k, \ \zeta_k = \sum_{\mathbf{x}(t,f) \in C_k} ||\mathbf{x}(t, f) - \mathbf{c}_k||^2. \tag{2.57}$$

After setting appropriate initial centroids $\mathbf{c}_k$ ($k = 1, \cdots, N$), this ζ can be minimized by the k-means clustering algorithm with a given number of sources N. Because each cluster corresponds to an individual source, centroid $\mathbf{c}_k$ represents the geometry of the source s_k as

$$\arg[\{\mathbf{c}_k\}_j] = 2\pi c^{-1} (\mathbf{d}_j - \mathbf{d}_J)^T \mathbf{q}_k, \tag{2.58}$$

where J is the reference sensor.

2.8.4 ICA-based method

The fourth category is related to the use of frequency-domain independent component analysis (ICA) for the BSS problem. This is because the demixing matrix obtained by the ICA algorithm contains a propagation model of the sources. The idea of Sawada et al. [55] has recently been generalized to a state coherence transform (SCT) approach by Nesta et al. [56].

Their method, which is based on a cumulative SCT, achieves joint multipath TDOA estimation using an SCT without being affected by spatial aliasing. Although the cumulative SCT method for a stereophonic sensor can be applied to the underdetermined DOA estimation problem, it is necessary to detect independent time blocks where only two sources are dominant prior to the application of ICA demixing. Therefore, the application of an ICA-based approach is basically limited to cases where the number of sources is equal to the number of microphones.

Nesta et al. proposed a method for TDOA estimation based on frequency-domain ICA and the SCT [56]. ICA is performed and the obtained demixing matrices are used to generate observations of the propagation model of the sources and estimate the DOA by the SCT. The explanation and the notation of the following ICA approach is based on the reference [56].

In the ideal case, neglecting reverberation, we can assume the sources to be under free-field conditions. Thus, the signals observed at the microphones can be considered to be the sum of delayed and scaled versions of the original source signals depending on the relative position of the sources to the microphones. For the case of two channels, in the frequency domain each mixing matrix can be modeled as

$$H(f) = \begin{pmatrix} |h_{11}(f)|e^{-j\phi_{11}(f)} & |h_{12}(f)|e^{-j\phi_{12}(f)} \\ |h_{21}(f)|e^{-j\phi_{21}(f)} & |h_{22}(f)|e^{-j\phi_{22}(f)} \end{pmatrix} \tag{2.59}$$

$$\phi_{iq}(f) = 2\pi f_s f \delta_{iq}/L, \tag{2.60}$$

where δ_{iq} is the propagation time from the qth source to the ith microphone, f_s is the sampling frequency, and L is the window length. The elements of each row of the demixing matrix $W(f) = H^{-1}(f)$ can be directly used to obtain the observations of the ideal propagation model. These are expressed for the two sources as the ratios computed as follows:

$$r_1(f) = -\frac{w_{12}(f)}{w_{11}(f)}, \quad r_2(f) = -\frac{w_{22}(f)}{w_{21}(f)}. \tag{2.61}$$

Such ratios are scaling-invariant and their phase is expected to vary linearly with the frequency, depending on the TDOAs of the sources. Neglecting wave attenuation, the ideal propagation model of each source can be represented as

$$c(t, f) = e^{-j2\pi f_s f \delta/L}, \tag{2.62}$$

where δ is the TDOA of the source. Thus, for each frequency, an observation of the ideal propagation model can be obtained by normalizing the ratios $r_i(f)$ by their magnitude:

$$\overline{r_i}(f) = \frac{r_i(f)}{||r_i(f)||}. \tag{2.63}$$

A joint multiple TDOA estimation can be performed by using an SCT which is formulated as follows:

$$SCT(\delta) = \sum_t \sum_{i=1}^{N} \left[1 - g(\frac{||c(t, f) - \overline{r_i}(f)||}{2}) \right], \tag{2.64}$$

where N is the number of observed states for each frequency and $g(\cdot)$ is a function of the Euclidean distance. They selected the nonlinear function

$$g(x) = \tanh(\alpha \cdot x) \tag{2.65}$$

where α is chosen according to the distance between the microphones.

Finally, the peaks of the SCT envelope that correspond to the TDOAs and source directions are extracted.

2.9 Summary

This chapter introduces the foundations of speech signal processing using a microphone array. In Sec. 2.2, the foundations of speech signal processing were explained. Then in Sec. 2.4, microphone array signal processing was discussed. In Sec. 2.3 it was shown that T-F analysis can represent speech signal features in the temporal and spectral domains. In the second half of the chapter, two applications of speech signal processing using a microphone array were explained: BSS in Sec. 2.6 and DOA estimation in Sec. 2.8. Some conventional methods and recent studies were also mentioned.

Chapter 3

Speaker localization and source separation using PCA and harmonic structure

3.1 Introduction

This chapter describes the proposed method for speaker localization and source separation using principal component analysis (PCA) and the harmonic structure. An overview of the prosed method and its advantages is given in Sec. 3.2. In Sec. 3.3, the blind source separation (BSS) problem and time-frequency (T-F) masking method are reviewed briefly. From Sec. 3.4 the proposed method is discussed in detail. In Sec. 3.8, some experiments performed to verify the proposed method are reported. Sec. 3.9 is a summary.

3.2 Overview

3.2.1 Phase difference versus frequency distribution

This study focuses on the T-F binary masking approach using a pair of microphones [64–66]. As the T-F cell features depend on the spatial location of the sources, framewise, namely, time-sequential PD-F data are exploited here. Since the conventionally utilized features associated with the time delay at each T-F cell can be estimated by the frequency normalization of PD-F data, the proposed method disregards the other conventional features of the signal level and the attenuation ratio between two sensors. This is because the sensor distance in the setup is smaller than half the minimum wavelength of interest to avoid the spatial aliasing assumed in many studies [22] [24]. Actually, the setup is a pair of typical microphones with 4 cm spacing and 8 kHz sampling frequency. Under these conditions, the signal level difference between observations should be very small. Therefore, the attenuation ratio is less distinctive than the phase difference. That is, the attenuation ratio would not be effective for clustering. This has been observed in many actual two-dimensional (2D) histograms of the attenuation ratio and delay [22] [24]. As in Refs. [23] and [25], a setup with a large distant sensor array violating the nonaliasing condition can be used, but the separation algorithm for such a setup would be complex. For example, in Ref. [25], the clustering procedure is divided into two steps, one of which is applicable to the nonaliasing or low-frequency band and the other is applicable to the remaining aliasing at frequency band.

On the other hand, the PD-F plot itself is not new. It has appeared in several papers, such as in Ref. [25]. However, to the author's knowledge, the mathematical analysis of the plot and the idea of using it in a frame-by-frame manner have not been reported so far. The advantages of using the time-frame PD-F distribution are as follows.

- The PD-F data that can be reliably estimated is located along a specific line through the origin; thus, the PD-F plot directly illustrates the tendency of the PD estimation error and gives an intuitive insight into the dependence of PD error distribution on the frequency.

- By observing the variance of PD-F data at a specific frame, it can be determined whether or not a single source is active at that time.

To be more precise, PCA is applied to the 2D PD-F data space at each frame, and the ratio of the two principal eigenvalues is used to determine the source activity in a frame-by-frame manner, namely, to determine whether a single source is active or multiple sources are simultaneously active. The obtained source activity condition in each frame is effectively used for source direction estimation and separation, as described in a later section.

The feature of the PD estimation error in the proposed approach is significantly different from that in the delay-histogram approach. In particular, for real-life acoustic data, the delay estimation error in the low-frequency band tends to be very large. In fact, owing to the frequency normalization of the estimated PD, the estimated delay in low-frequency bins tends to be an outlier, even for a comparable phase estimation error over the whole frequency range. Therefore, the delay histogram for real data will have a long-tailed distribution. Additionally, the number of data within a few frames is too small to obtain accurate peaks for the delay histogram.

Among the proposed features of the T-F cells, the attenuation ratio and its modifications do not exhibit any distinctive differences for closely located microphones. In the experimental setup in this study, the distance between the microphones is 4 cm to avoid spatial aliasing for the 8 kHz sampling rate. Thus, the features associated with the signal level difference are disregarded. On the other hand, although T-F masking solely based on the delay histogram is effective, it gives rise to highly misestimated delay data in the low-frequency band [67–70]. The clustering using the delay causes the results to exhibit low performance. Here framewise PD-F data are employed to classify each time frame into three cases associated with the number of active sources. In this stage, the estimated PD is adopted without applying frequency normalization. The second step is to perform clustering for two frequency bands. In the high-frequency band, the PD-F data plot for one frame is divided into two clusters by determining a separation line through the origin. Namely, this initial separation adopts the delay between sensors as the feature at each frame without the application of peak finding or the *k*-means algorithm. Although the delay is essentially used as the separation feature, its frame-by-frame usage is crucial in this study. The separation in the low-frequency band utilizes the speaker's harmonic structure with nonspatial features. The spectrogram in this band is integrated with the results of initial separation in the high-frequency band.

3.2.2 Harmonic structure as a means of separation

In addition to source location attributes, such as the above-mentioned delay and attenuation ratio, sound source attributes such as harmonic structure are also useful for segregation [71] [72]. Early separation approaches using the harmonic structure [72] are applied to monaural mixture signals. The first process is to estimate the fundamental frequency (f_0) of each speaker, then the local frequency spectrum components that are assigned to a speaker are selected according to their harmonic relation with the estimated f_0. In essence, the harmonic-structure-based approach is valid for vowels and vowel-like sound intervals.

In this study, the harmonic structure is utilized as a means of restrictive separation in the low-frequency band where the attributes of the source location are less reliable for characterizing each T-F cell. In Ref. [71], both the harmonic structure and the source direction are exploited for clustering. The main difference between the proposed method and previous approaches is that the harmonic structure of a monaural signal in the low-frequency band is related to that of the initially separated signals by means of T-F masking.

3.2.3 Advantages

In this study, the PD-F distribution at each frame is first investigated by applying PCA to classify the PD-F data in one frame into three cases: a) non-source active (NSA), b) single-source active (SSA[1]), and c) double-source active (DSA). Since the ratio of the principal eigenvalues in PCA indicates the degree of data spread around the first principal axis, it is utilized to detect SSA frames. Next, the selected SSA frames are used to estimate source directions. From these directions, the active source at each SSA frame is identified. This means that all T-F cell components in an SSA frame are associated with the identified source. The third step is to seperate the DSA frames via two substeps. The first clustering step is performed in a high-frequency band, denoted by B_{high}, in which the PD estimation error is relatively small; therefore, PD-F data are much more reliable. In this clustering procedure, the delay value is adopted as the cell's feature. For the T-F cells in the remaining low-frequency band, or the complement of B_{high}, denoted by B_{low}, a harmonic structure relationship between the initially separated spectrogram in B_{high} and the spectrogram in B_{low} is effectively used.

The novel features of this study are summarized as follows.

1. Framewise PD-F analysis is used to detect the SSA frame by means of PCA. The results are used to accurately estimate source directions by introducing a novel reliability degree.

2. To seperate the DSA frames, the relationship between the harmonic structures of initially separated source signals and the mixed signals in the low-frequency band is exploited.

3.3 BSS problem

3.3.1 Observation model

The mixing model in a discrete time domain and its transformed T-F domain description are described here. All discrete time signals are sampled versions of analog signals with sampling frequency f_s[Hz]. Suppose N source signals $s_1(\tau)$, $s_2(\tau)$,$\cdots$,$s_N(\tau)$ are mixed by time-invariant convolution and the observed signals $x_1(\tau)$, $x_2(\tau)$,$\cdots$,$x_M(\tau)$ at M sensors are described as

$$x_m(\tau) = \sum_{i=1}^{N} \sum_{l} h_{mi}(l) s_i(\tau - l) \tag{3.1}$$

where $h_{mi}(l)$ represents the impulse response from the i-th source to the m-th sensor . Signal transformation to the T-F domain is performed as follows. Observed signals $x_m(\tau)$ (m =

[1] The term "SSA" means that the PD-F distribution at the frame appears to result from a single directional source.

$1 \sim M$) are converted into T-F domain signals $X_m[k,l]$ by L-point windowed STFT. That is, $X_m[k,l]$ can be written as

$$X_m[k,l] = \sum_{r=-L/2}^{L/2-1} x_m(r+kS)win(r)e^{-j\frac{2\pi l}{L}r}, \qquad k = 0 \sim K, l = 0 \sim \frac{L}{2} \tag{3.2}$$

where $win(r)$ is a window, and S is the window shift length. Here, a half-window-size overlapping transformation is applied, namely, $S = \frac{L}{2}$ in (3.2). In addition, STFT without zero-padding is applied. Then the transformed T-F mixture model of Eq. (3.1) is described by the instantaneous mixtures at each time frame index k and frequency bin l.

$$X_m[k,l] = \sum_{i=1}^{N} H_{mi}[l]S_i[k,l] \tag{3.3}$$

Here, $H_{mi}[l]$ is the frequency response (DFT) of $h_{mi}(\tau)$, and $S_i[k,l]$ is the T-F domain representation of the i-th source signal $s_i(\tau)$.

An anechoic mixing model is adopted as used in Ref. [22]. In this model, the source signals to recover are alternatively redefined as the observed signals at the first sensor. Namely, the following mixing model in the T-F domain is henceforth discussed in this paper.

$$X_1[k,l] = \sum_{i=1}^{N} S_i[k,l] \tag{3.4}$$

$$X_m[k,l] = \sum_{i=1}^{N} H_{mi}[l]S_i[k,l] \quad (m = 2,\cdots,M) \tag{3.5}$$

where $S_i[k,l]$ is the i-th source signal observed at the first sensor location, and $H_{mi}[l]$ ($m = 2,\cdots,M$) eventually represents the DFT domain operation of the subsample delay of the i-th source signal, which is caused between the m-th sensor and the first sensor. In a later discussion and experiments in this study, the cases of two sources (N=2) and two sensors (M=2) are considered without loss of generality.

3.3.2 WDO in T-F masking

The assumption on which the proposed separation algorithm is based is WDO. This property being satisfied between source signals is commonly assumed in the T-F masking algorithm, and it is approximately satisfied for speech signals. The WDO means that sources have disjoint T-F supports. This inherently stems from the sparseness of the T-F domain component distribution of speech signals. The definition and its use in separation for the two-source and two-sensor case are described as follows.

Consider two source signals $s_i(\tau)$ $(i = 1, 2)$, and define the T-F supports Ω_i of their T-F domain representations $S_i[k, l]$ by

$$\Omega_i := \{[k, l] | S_i[k, l] \neq 0\} \quad i = 1, 2 \tag{3.6}$$

In practice, the above nonzero condition is replaced by $|S_i[k, l]| < e$ where e is a sufficiently small positive value. Then, the WDO assumption between two source signals $s_1(\tau)$ and $s_2(\tau)$ can be represented by

$$\Omega_1 \cap \Omega_2 = \phi \; (\textit{empty set}) \tag{3.7}$$

The following null component domain, denoted by Ω_N, is also introduced.

$$\Omega_N = \overline{\Omega_1 \cup \Omega_2}, \; \overline{\cdot}; \textit{complementary set} \tag{3.8}$$

Therefore, the WDO stipulated that the T-F domain representation of the mixed signal $X_1[k, l]$, given by Eq. (3.4), can be decomposed into the following three parts with no overlap.

$$X_1(k, l) = \begin{cases} S_1[k, l] & [k, l] \in \Omega_1 \\ S_2[k, l] & [k, l] \in \Omega_2 \\ 0 & [k, l] \in \Omega_N \end{cases} \tag{3.9}$$

The T-F binary masking separation utilizes the above disjoint separation assumption. Its essential process is to separate the support of $X_1[k, l]$ into two subregions Ω_1 and Ω_2, and to obtain $S_1[k, l]$ and $S_2[k.l]$ shown in Eq. (3.9). In order to perform this separation, a pair of $X_1[k, l]$ and $X_2[k, l]$ is used to introduce the spatial feature of the T-F cell at $[k, l]$, and the clustering process is performed in the estimated feature space.

The above clustering results Ω_i generate the separation masks, $M_i[k, l]$, as the binary functions defined in the T-F domain.

$$M_i[k, l] = \begin{cases} 1 & [k, l] \in \Omega_i \\ 0 & \textit{otherwise.} \end{cases} \quad (i = 1, 2) \tag{3.10}$$

The final stage of the separation process is to obtain time-domain-separated signals $\hat{s}_i(\tau)$ $(i = 1, 2)$ by applying the inverse STFT to

$$\hat{S}_i[k, l] = M_i[k, l] X_1[k, l] \;\; (i = 1, 2) \tag{3.11}$$

3.3.3 Clustering feature

As stated in the previous section, the objective of the separation algorithm is to classify T-F cells composing the support of $X_1[k,l]$ into either Ω_1 or Ω_2. Therefore, the first process is to introduce the appropriate feature of the T-F cell at $[k,l]$ by utilizing $X_1[k,l]$ and $X_2[k,l]$. Commonly used features are the signal level or attenuation ratio and the frequency-normalized PD between $X_1[k,l]$ and $X_2[k,l]$. Features conventionally used in the clustering are summarized in Ref. [24], and these are evaluated from the separation performance point of view. The basic features associated with the attenuation ratio α and the delay δ between sensors are respectively defined as

$$\alpha = \frac{|X_1[k,l]|}{|X_2[k,l]|} \tag{3.12}$$

$$\delta[k,l] = \frac{L}{2\pi f_s l}\phi[k,l] \tag{3.13}$$

where $\phi[k,l]$ is the PD between $X_1[k,l]$ and $X_2[k,l]$, as defined by

$$\phi[k,l] = \angle X_1[k,l] - \angle X_2[k,l] \tag{3.14}$$

One example of the phase difference versus frequency data plot is shown in Fig. 3.1.

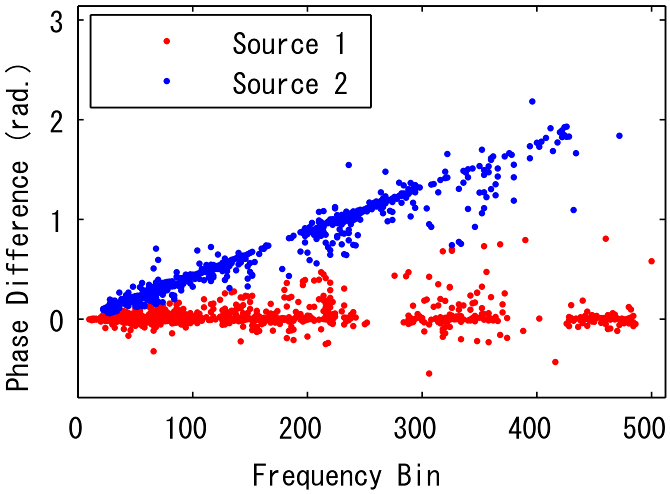


Figure 3.1: Example of phase difference versus frequency data plot. The mixture condition is a male source s_1 located at 0° and a female source s_2 located at 50°.

3.4 PD-F distribution data

In the k-th frame, the PD-F data is defined as a collection of two-dimensional vectors.

$$\left\{ \begin{bmatrix} l \\ \phi[k,l] \end{bmatrix}, \quad l = 0 \sim \frac{L}{2} \right\} \tag{3.15}$$

Figure 3.2 illustrates an example of the time series of the PD-F data plot.

Before discussing the proposed method in detail, four frequency bands are introduced as follows. According to the low-frequency limit in human voice signals, the proposed separation process is applied to the following range:

$$B_{full} := \{l | l_1 < l < L/2\} \tag{3.16}$$

where $l_1 = \lfloor (f_1 \cdot L/f_s) \rfloor$, $\lfloor \ \rfloor$ is the Gauss floor function, and f_1 is the analog frequency, which is set to 80 Hz in this study. In the above full frequency range, the following three frequency intervals, denoted B_{high}, B_{low}, and B_{mid} are defined below.

$$B_{high} := \{l | l_2 < l < L/2\} \tag{3.17}$$

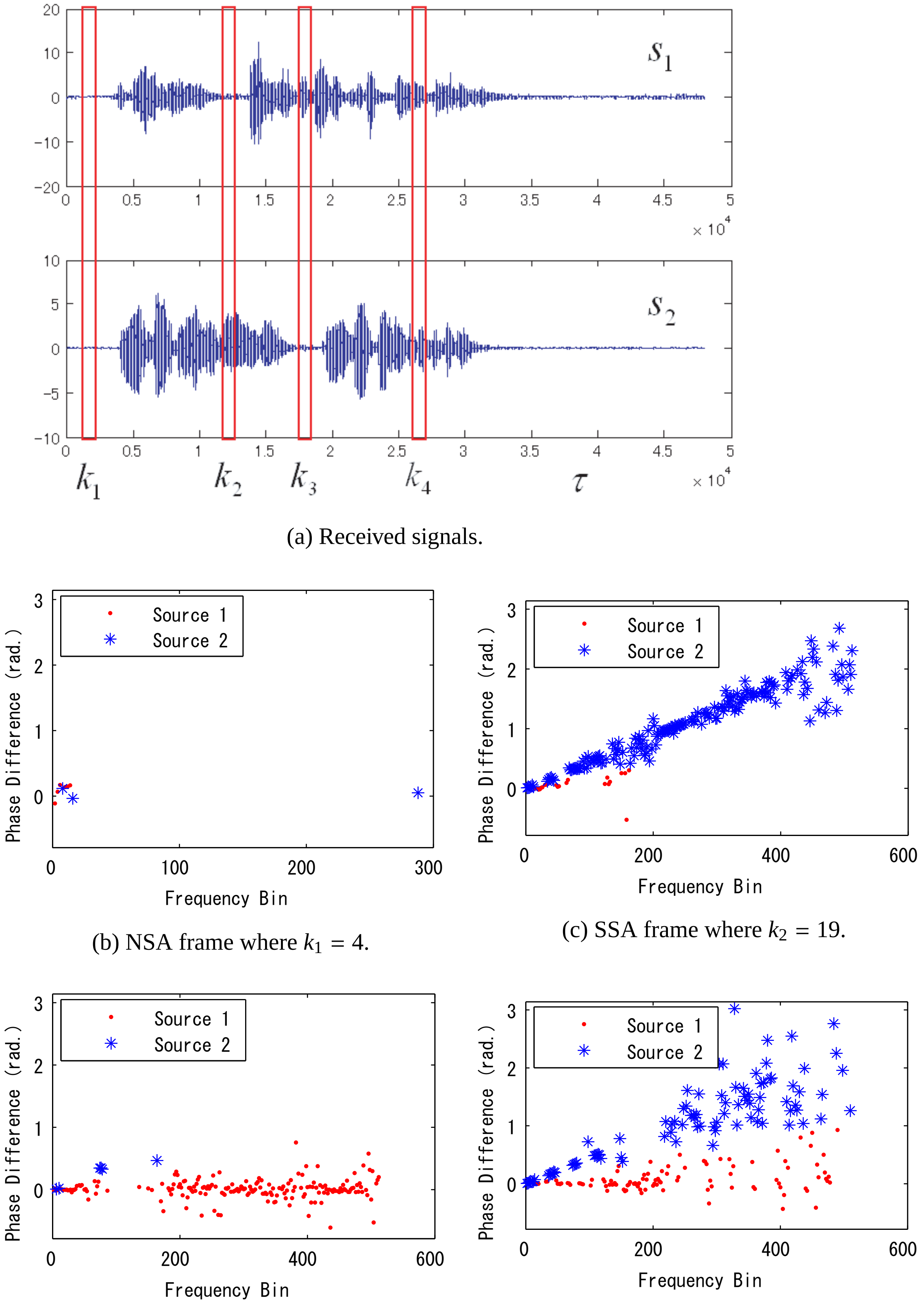
Here, $l_2 = \lfloor (f_2 \cdot L/f_s) \rfloor$, and f_2 is set at 400 Hz empirically. For source signals from any direction, PD in the frequency range lower than l_2 becomes too small to determine source direction exactly. The first part of the proposed PD-based separation algorithm is applied to T-F cells in the B_{high} range. The attribute of T-F cells used for clustering in the frequency range lower than l_2 should adopt other features that are not associated with source location, such as delay and attenuation ratio. Therefore, the third frequency range is introduced as

$$B_{low} := \{l | l_1 < l < l_2\} \tag{3.18}$$

For the separation of T-F cells in B_{low}, we the utilize harmonic relationship between the spectrum $|X_1[k,l]|$ in B_{low} and the spectra of initially separated signals in the relatively low part of the B_{high} range, denoted by B_{mid}. Therefore the following range for estimating fundamental frequency is introduced.

$$B_{mid} := \{l | l_2 < l < l_3\} \tag{3.19}$$

Here, $l_3 = \lfloor (f_3 \cdot L/f_s) \rfloor$, and f_3 is set at 1 kHz in the method. Using this interval, the fundamental frequency is estimated in order to employ the harmonic structure relation with the spectrum in B_{low}. The various frequency band is shown in Fig. 3.3. The detailed algorithm will be given later.



(a) Received signals.

(b) NSA frame where $k_1 = 4$.

(c) SSA frame where $k_2 = 19$.

(d) SSA frame where $k_3 = 24$.

(e) DSA frame where $k_4 = 35$.

Figure 3.2: Sequence of PD-F distribution. The mixture condition is a male source s_1 located at 4° and a female source s_2 located at 51°.

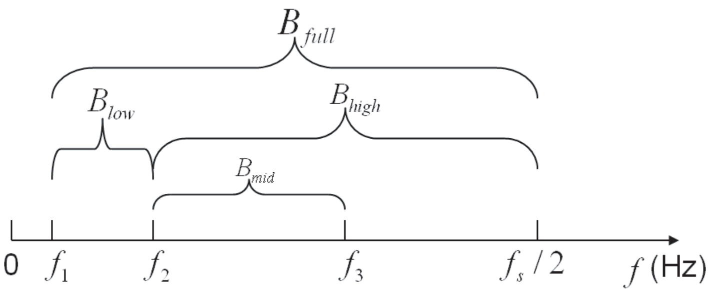


Figure 3.3: Various frequency band

The normalized PD-F data set at the k-th frame, denoted by P_k, is manipulated hereafter for simplifying the analysis:

$$P_k \triangleq \left\{ \mathbf{p}_k(l) = \begin{bmatrix} p_k^1(l) \\ p_k^2(l) \end{bmatrix} = \begin{bmatrix} l/(\frac{L}{2}) \\ \phi[k,l]/\pi \end{bmatrix}, \quad l \in B_{high} \right\} \tag{3.20}$$

For normalized PD-F, the gradient $\beta[k,l]$ is defined as

$$\beta[k,l] = \frac{\phi[k,l]/\pi}{l/(\frac{L}{2})} = \frac{L}{2\pi} \cdot \frac{\phi[k,l]}{l} \tag{3.21}$$

3.4.1 Outline of the method

The proposed method requires the following two assumptions.

(A1) Although the received signals are mixture signals, for each source $s_i(\tau)$, there exist some time frames in the T-F domain where only $s_i(\tau)$ is active. These frames can be used to detect the source direction of $s_i(\tau)$.

(A2) Since this research mainly focuses on the source separation algorithm, the noise level is assumed to be sufficiently low with respect to the level of the sources. Therefore, its noise robustness is not concerned, like other BSS separation studies.

The outline of the proposed method is shown in Fig. 3.4 and summarized as follows.

(1) NSA frame detection: local power at a frame is used to determine whether the frame may be classified as NSA or Source Active (SA).

(2) PCA is applied to the PD-F plot of each SA frame, and the principal eigenvalue ratio is used to separate individual SA frames into SSA and DSA.

(3) Source directions are estimated by some selected reliable SSA frames. Then, the results can determine whether either of the source signals is active at each SSA frame.

(4) The separation algorithm with two subseparation steps is applied to each DSA frame using PD-F scattered plots.

(5) Integrating the above separation processes, two supports, Ω_1 and Ω_2 are obtained. The masks $M_1[k, l]$ and $M_2[k, l]$ are generated by Eq. (3.10), and finally, separated source signals are estimated.

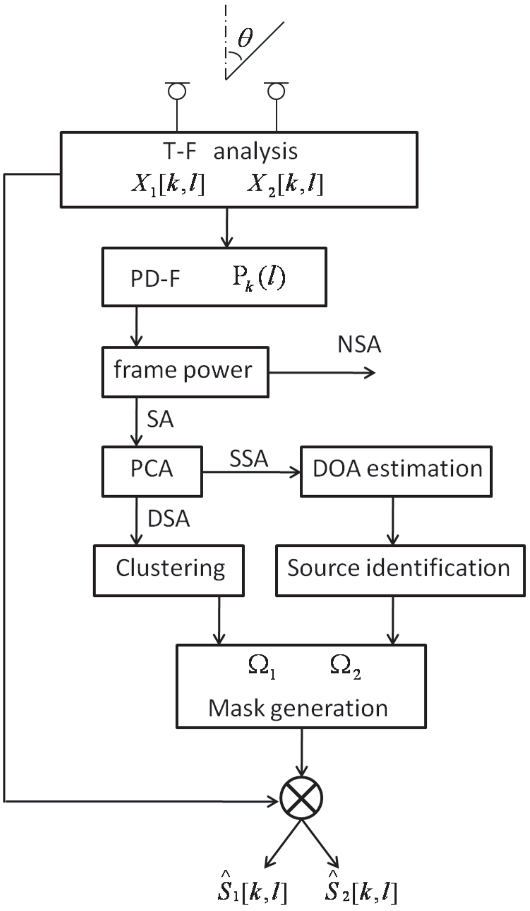


Figure 3.4: System flow

3.5 Non-source active (NSA)

(A2) stipulates that environmental noise power would be sufficiently small. Under this assumption, the average power at one frame is utilized to indicate the possible presence of speech in the frame. Thus, the preliminary threshold operation of frame power, known as a basic voice activity detection algorithm, is valid.

The average local power of frame k is defined as

$$E(k) := \frac{1}{L/2+1} \sum_{l=0}^{\frac{L}{2}} |X_1[k,l]|^2 \tag{3.22}$$

NSA is judged by

$$if \quad E(k) < Th_1, \text{ then } k\text{-th frame} \in NSA \tag{3.23}$$

In this study, Th_1 is determined beforehand by a pre-experiment during no utterance.

$$Th_1 = E_0 + 2\sigma_E \tag{3.24}$$

E_0 is the average noise power, and σ_E is the standard deviation.

$$E_0 = \frac{1}{N} \sum_{k=1}^{N} E(k) \quad k \in NSA \tag{3.25}$$

$$\sigma_E = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (E(k) - E_0)^2} \quad k \in NSA \tag{3.26}$$

3.6 Single source active (SSA)

As observed from the typical PD-F distribution of the SSA frame in Fig. 3.2 (c)(d), whether the given frame is in the SSA state or not would be reflected in a scattering feature along a constant gradient line, and the gradient indicates the source direction. The relationship between the gradient β in a normalized PD-F plane defined in Eq. (3.20) and the source direction θ (as shown in Fig. 3.4) is

$$\beta = f_s \cdot \frac{d}{c} \sin\theta \tag{3.27}$$

The following will describe (i) how to identify the SSA frame, and (ii) how to use SSA frames to estimate source directions.

Because the PD estimation in low-power-level T-F components is unreliable, the subset $\tilde{P_k}$ of P_k are defined as

$$\tilde{P_k} = \{\mathbf{p}_k(l) \mid \frac{|X_1[k,l]|^2}{B(k)} > Th_2\} \tag{3.28}$$

$$B(k) = \frac{2}{L} \sum_{y=0}^{L/2-1} |X_1[k,y]|^2 \tag{3.29}$$

Th_2 is set to 0.05 empirically.

PCA is applied to $\tilde{P}_k$ by computing the following 2×2 principal component covariance matrix $\mathbb{R}_k$:

$$\mathbb{R}_k := \frac{1}{L/2 - l_0} \sum_{l=l_0}^{L/2-1} \mathbf{p}_k(l)\mathbf{p}_k^{\top}(l)$$

$$= \begin{bmatrix} R_{11}(k) & R_{12}(k) \\ R_{21}(k) & R_{22}(k) \end{bmatrix} \tag{3.30}$$

Denoting the eigenvalues of the covariance matrix of $\mathbb{R}_k$ by $\lambda_1(k)$ and $\lambda_2(k)$ (assume $\lambda_1(k) \geq \lambda_2(k)$ without loss of generality), their corresponding eigenvectors are represented, respectively, as

$$e_1(k) := \begin{bmatrix} \cos\beta(k) \\ \sin\beta(k) \end{bmatrix} \tag{3.31}$$

$$e_2(k) := \begin{bmatrix} \cos\gamma(k) \\ \sin\gamma(k) \end{bmatrix} \tag{3.32}$$

where $\beta(k)$ and $\gamma(k)$ (rad.) are the gradients of the principal axes in the k-th frame. Theoretically, $|\beta(k)| \leq 1$ is satisfied, so $\beta(k)$ is redefined if $|\beta(k)| > 1$:

$$\beta(k) = \begin{cases} \beta(k), & |\beta(k)| \leq 1 \\ 1, & \beta(k) > 1 \\ -1, & \beta(k) < -1 \end{cases} \tag{3.33}$$

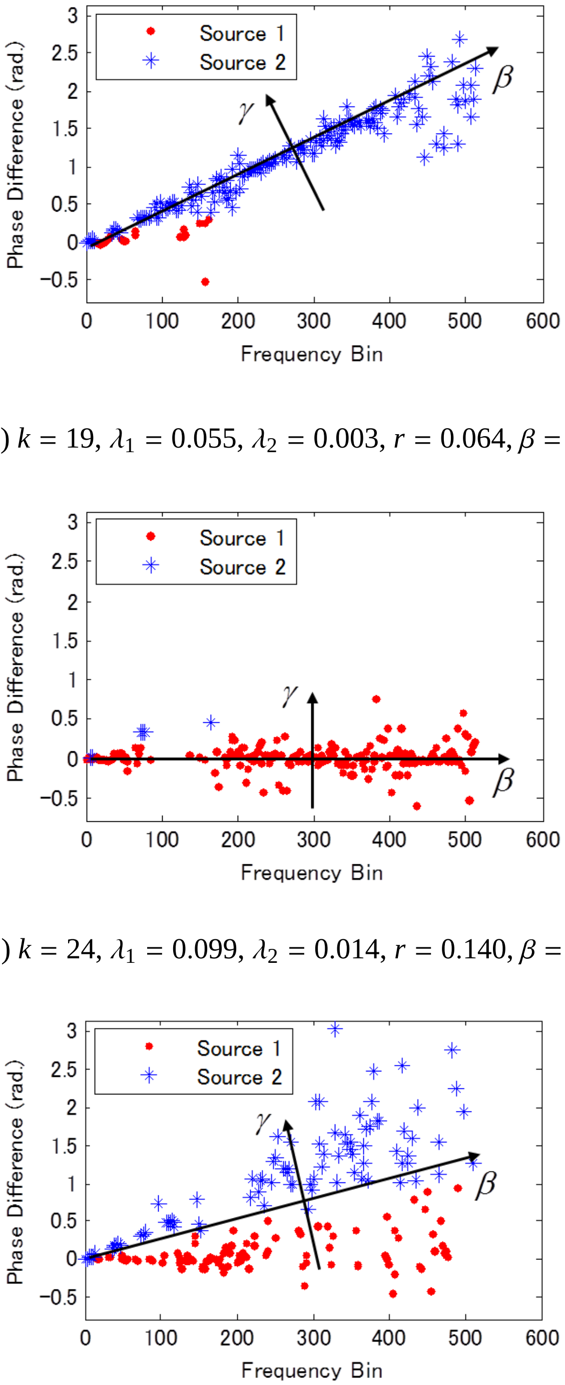
From Eq. (3.27), source direction $\theta(k)$ with respect to $\beta(k)$ is given by

$$\theta(k) = \sin^{-1}[\beta(k)/(f_s \cdot \frac{d}{c})] \tag{3.34}$$

Next, the ratio of the principal eigenvalue defined by

$$r(k) := \frac{\lambda_2(k)}{\lambda_1(k)} \tag{3.35}$$

is introduced to determine the SSA frame from others. If the PD-F data in the k-th time frame are distributed solely along the first principle axis and not along the second axis, $r(k)$

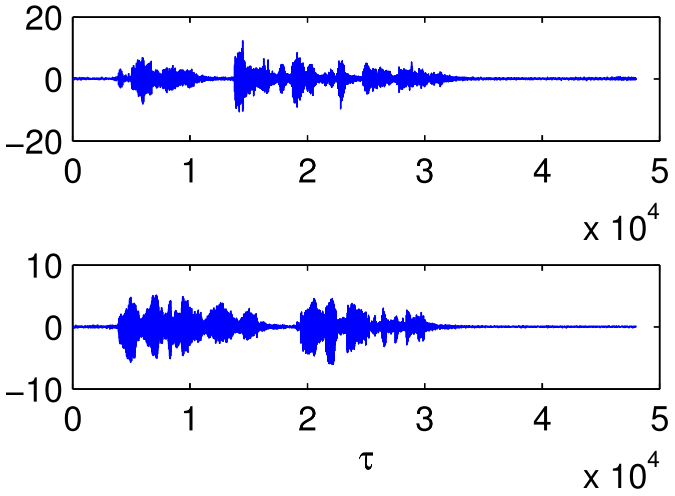


(a) $k = 19$, $\lambda_1 = 0.055$, $\lambda_2 = 0.003$, $r = 0.064$, $\beta = 0.72$

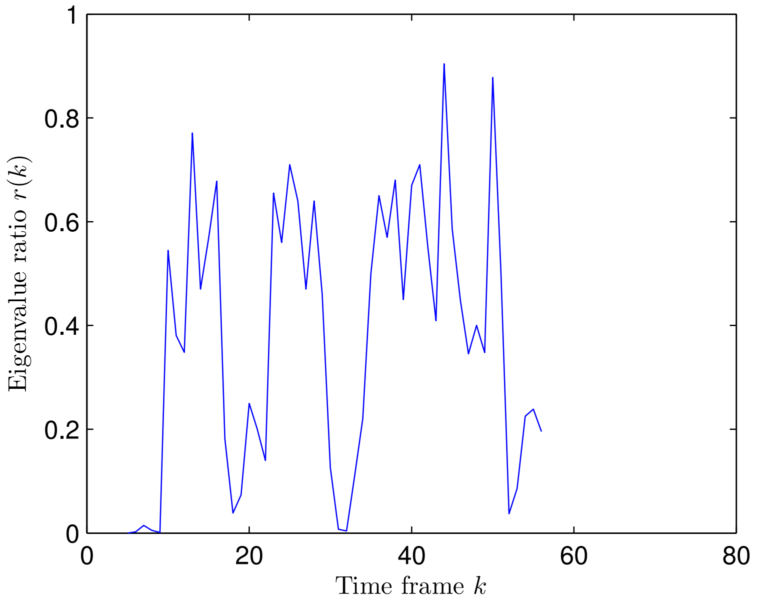
(b) $k = 24$, $\lambda_1 = 0.099$, $\lambda_2 = 0.014$, $r = 0.140$, $\beta = 0.02$

(c) $k = 35$, $\lambda_1 = 0.092$, $\lambda_2 = 0.057$, $r = 0.618$, $\beta = 0.39$

Figure 3.5: A set of examples applying PCA to several PD-F distribution frames. The mixture condition is the same as in Fig. 3.2.



(a) Received Signals. The mixture condition is a male source s_1 located at 4° and a female source s_2 located at 51°.



(b) Eigenvalue ratio.

Figure 3.6: One result of eigenvaule ratio

is small. Therefore, the frame will be classified into SSA. A set of examples applying PCA to several PD-F distribution frames is shown in Fig. 3.5. One result of eigenvaule ratio is shown in Fig. 3.6.

For each SSA frame, the $\theta(k)$ obtained by Eq. (3.34) indicates the direction of the source that is active at the frame. These observations lead to the following steps for estimating the source directions and identifying the active source at every SSA frame.
(1) The following criterion is applied to determine whether the k-frame is SSA.

$$r(k) < Th_{SSA} \tag{3.36}$$

Later, Th_{SSA} is determined to be 0.3 experimentally. A set of the frame indices satisfying the condition

$$T := \{k | r(k) < Th_{SSA}\} \tag{3.37}$$

is defined.
(2) The SSA frames cannot be in a single frame, but in several continuous frames. Let T_j ($j \in J$) be a time interval in T corresponding to the j-th SSA period, and define the smallest eigenvalue ratio in each T_j by taking

$$r_j := \min_{k \in T_j} r(k) \tag{3.38}$$

and its source direction is θ_j calculated by Eq. (3.34).
(3) The first source direction θ_1 is provided by the minimum r_j in $j \in J$.

$$\theta_1 = \theta_u, \quad u = \arg\min_{j \in J} r_j \tag{3.39}$$

(4) The other source direction θ_2 is determined as the direction θ_z which has the next smallest r_j ($j \neq u, j \in J$). That is,

$$\theta_2 = \theta_z, \quad z = \arg\min_{j \neq u} r_j \tag{3.40}$$

(5) Identify the active source at each SSA time frame, and classify all T-F cells as either Ω_1 or Ω_2 as follows:

$$\begin{cases} if \ \ |\theta(k) - \theta_1| \leq |\theta(k) - \theta_2| \ \ k \in T, \ \{ [k, l] \mid \forall l\} \in \Omega_1 \\ otherwise \ \ k \in T, \ \{ [k, l] \mid \forall l\} \in \Omega_2 \end{cases} \tag{3.41}$$

3.7 Double source active (DSA)

Now the set of DSA frames satisfying the condition $r(k) > Th_{SSA}$ is considered in this section. The problem that must be solved is the clustering of the PD-F data in the DSA frame into two sets. In theory, when two sources are active simultaneously and the WDO assumption holds, accurately estimated PD-F data fall along two lines through the origin. The gradients of these lines correspond to θ_1 and θ_2. However, in practical circumstances, the PD-F data distribution is more or less spread around these lines because of phase estimation error. Assuming independent identically distributed (i.i.d.) estimation error, clustering of the PD-F data in the low-frequency band is inherently difficult. This means that the T-F cell feature corresponding to the spatial location of sources is not suitable for source separation in the low-frequency band. On the basis of the experimental results obtained with the microphone array setup, the low-frequency band is set as B_{low}.

The first separation process in the DSA frame is the initial separation in the B_{high} band accomplished by applying the nearest neighbor approach between the PD-F data and the lines corresponding to the directions θ_1 and θ_2 obtained. Next, the separation in B_{low} utilizes the harmonic structure relationship between the spectrum of observation $X_1[k,l]$ in B_{low} and the spectra of initially separated signals in B_{high}. Finally, by integrating the initial separation masks in B_{high} and B_{low}, signals in the DSA frame are separated. One example of separation strategy in DOA frame is shown in Fig. 3.7.

3.7.1 Initial separation

The source directions have been estimated as θ_1 and θ_2, and their corresponding gradients in the normalized PD-F plane are β_1 and β_2 defined in Eq. (3.27). All the points in these two lines can be expressed as

$$\phi_n(l) = \beta_n \cdot l \quad (n = 1, 2) \tag{3.42}$$

In the k-th frame, the nearest neighbor method gives the binary mask $\tilde{M}_i[k,l]$ in B_{high}, which is defined as

$$\tilde{M}_i[k,l] = \begin{cases} 1, \; if \;\; i = \arg\min\limits_n |\phi[k,l] - \phi_n(l)|, \;\; l \in B_{high} \\ 0, \; otherwise \end{cases} \tag{3.43}$$

Therefore, the separated individual signals $\tilde{S}_i[k,l]$ $(i = 1, 2)$ are represented by

$$\tilde{S}_i[k,l] = \tilde{M}_i[k,l] X_1[k,l], \;\; l \in B_{high} \tag{3.44}$$

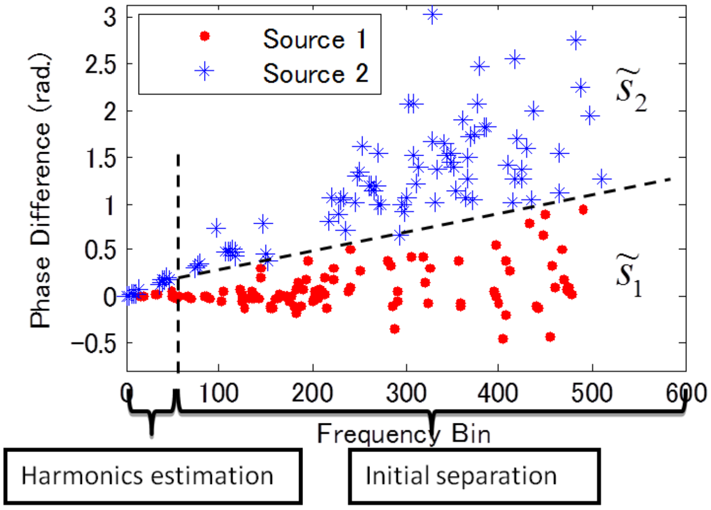


Figure 3.7: One example of separation strategy in DOA frame where time frame k = 35. The mixture condition is the same as in Fig. 3.2.

3.7.2 Local maximum in B_{mid}

The final task in the separation process is to generate individual masks applied on the T-F cells in the B_{low} range. In this final separation process, the observed amplitude spectrum given by $|X_1[k,l]|$ with $l \in B_{low}$ is compared with the initially separated spectra $\tilde{S}_1[k,l]$ and $\tilde{S}_2[k,l]$ with $l \in B_{mid}$ in terms of harmonic relationships.

First, with the help of local maximum frequencies of $|\tilde{S}_i[k,l]|$, the harmonic structure in B_{mid} is estimated for each separation spectra. Smoothing with the 5-point running average is applied for interpolation.

$$V_i[k,l] = \frac{1}{5}\sum_{j=-2}^{2} |\tilde{S}_i[k,l+j]|,\ l \in B_{mid} \tag{3.45}$$

The local maximum frequencies of $V_i[k,l]$ which satisfy the following two conditions are selected.

(1) Relative sufficient amplitude, where

$$\frac{|V_i[k,l]|}{\max_y |V_i[k,y]|} > Th_V \tag{3.46}$$

In later experiments, Th_V = 0.2 is adopted.

(2) The amplitude is maximum among the amplitude values at several adjacent frequency

bins. Because the fundamental frequency of the human voice is greater than 80 Hz,

$$l > \left\lfloor \frac{L}{f_s} \cdot 80 \right\rfloor = 10 \tag{3.47}$$

which means that it is only possible to have one harmonic frequency within at least 10 adjacent bins.

Under these conditions, the frequency bins of local maxima are obtained. The obtained local maximum frequencies of $|\tilde{S}_i[k, l]|$ are denoted as $b_{i1}(k), b_{i2}(k), \cdots$, and $q_i(k)$ denotes the number of local maxima in B_{mid}.

3.7.3 Harmonics estimation

The distance between adjacent local maxima $\Delta d_i(k)$ is defined as

$$\Delta d_i(k) = b_{i2}(k) - b_{i1}(k),\ q_i(k) \geq 2 \tag{3.48}$$

When $q_i(k) = 0$ or 1, let consider that there are no harmonic characteristics in the source $\tilde{S}_i[k, l]$ at frame k. The estimated harmonics $g_{in}(k)$ in B_{low} is

$$g_{in}(k) = b_{i1}(k) - \Delta d_i(k) \cdot n \tag{3.49}$$

where $n = 1, 2, 3, \cdots$, $g_{in}(k) \in B_{low}$, and $g_{in}(k)$ means the harmonic structure of source i at frame k.

There is a special situation in which both $q_1(k)$ and $q_2(k) = 0$ or 1. In this case, the harmonics is set to be the same as in the last previous frame as follows:

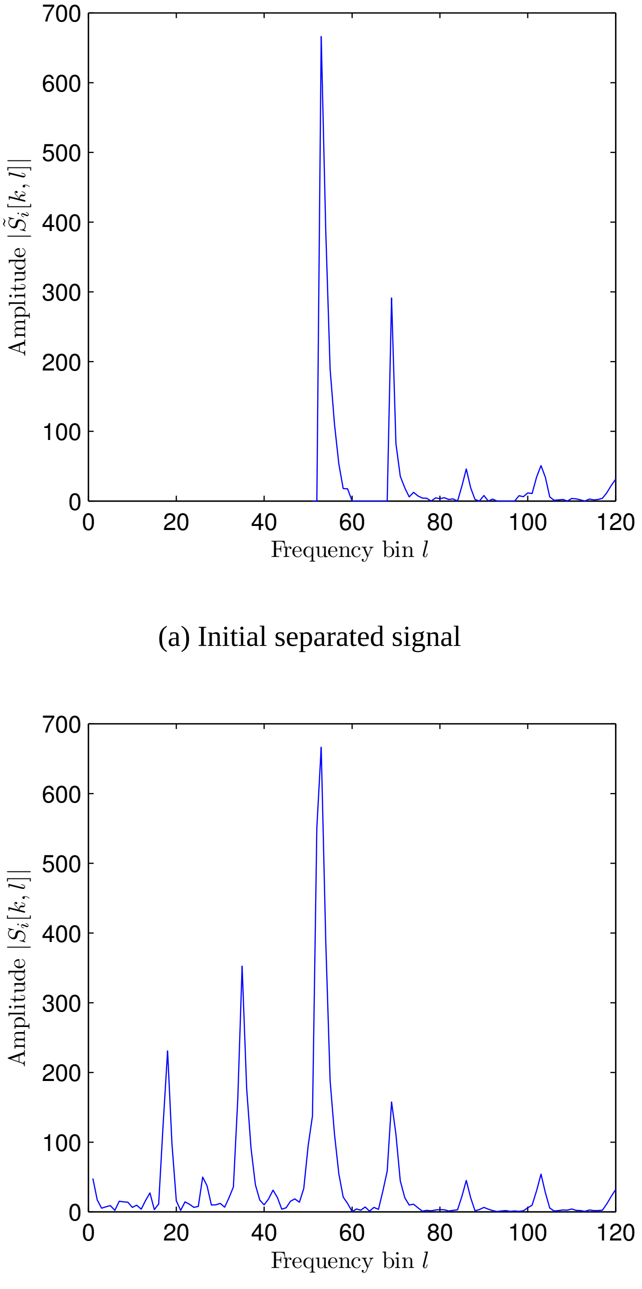
$$g_{in}(k) = g_{in}(k - y) \tag{3.50}$$

for the frames with the smallest $y > 0$ satisfying $q_i(k - y) \geq 2$ and $k - y \in$DSA.

One example of harmonics estimation is shown in Fig. 3.8 (a). The local maxima of the initial separated signal $|\tilde{S}_1[k, l]|$ are: $b_{11}(18) = 53, b_{12}(18) = 70$. Then the location of local maxima and their distance are used to estimate the harmonic structure in low-frequency band, and the estimated harmonics $g_{in}(k)$ are: $g_{11}(18) = 36, g_{12}(18) = 19$. Compared with the source signal in Fig. 3.8 (b), the estimated harmonics are almost the same as source signal.

3.7.4 Mask generation

Assume that the bandwidth of each harmonic in B_{low} is the same, and use 5 adjacent cells (i.e., 40 Hz) as the bandwidth in the T-F domain. The mask in B_{low} is defined as



(a) Initial separated signal

(b) Source signal

Figure 3.8: One example of harmonics estimation. The mixture condition is: a male source at 4° and a female source at 40°. The above signal is that of the male source. The time frame *k* is 18.

$$\bar{M}_i[k,l] = \begin{cases} 1, \ if \ \ g_{in}(k) - 2 < l < g_{in}(k) + 2 \ \ and \\ \qquad q_i(k) \geq 2, \ l \in B_{low}, \ n = 1, 2, 3, \cdots \\ 0, \ otherwise \end{cases} \tag{3.51}$$

The final mask is represented by

$$M_i[k,l] = \tilde{M}_i[k,l] + \bar{M}_i[k,l] \tag{3.52}$$

The separated signal is obtained as shown in Eq. (3.11).

3.8 Experiments

3.8.1 Experimental conditions

Some experiments are performed in a conference room to evaluate the proposed methods. The experimental environment is shown in Fig. 3.9. The experimental setup is shown in Fig. 3.10, and the experimental parameters are shown in Table 3.1. The experimental parameters are determined by the following reasons: the sampling frequency and the windows length determine the frequency resolution. Proper frequency resolution will concentrate the speech signal power spectrum on specific frequency components and minimize the degree of frequency component overlap between two speech signals. According to some researchers' investigation [23], a suitable frequency resolution is about 10 Hz. In this study, the sampling frequency is 8 kHz, and the windows length is 1024, so the frequency resolution is about 8 Hz. In order to avoid spatial aliasing, the delay between two microphones must be less than a sample. While the sampling frequency is 8 kHz, the maximum distance between two microphone is 4.25 cm, so the distance between microphones is set to 4 cm.

The speaker used in the experiments is SONY speaker system model NO. SRP-S320. The microphone is SONY electret condenser microphone ECM-77B. The Acoustic Society of Japan (ASJ) continuous speech corpus is used for research as the source signal. The mixture signals are combinations of the same sex or opposite sex from 10 male and 10 female sources, such as male & male, female & female, and male & female.

Note that a pair of microphones detects the signals where the sources are at the half-side with respect to the array axis, because the signals from the symmetrical positions of the axis are the same at the sensors. Thus the source direction range is 180°. The condition of the minimum difference in the source direction angle is 10°. This condition is a result of the limiting ability of the directivity resolution of the setup. One source is placed at the broad side (0°) and the location of the other source is varied from 0° to 80° at intervals of 10°.

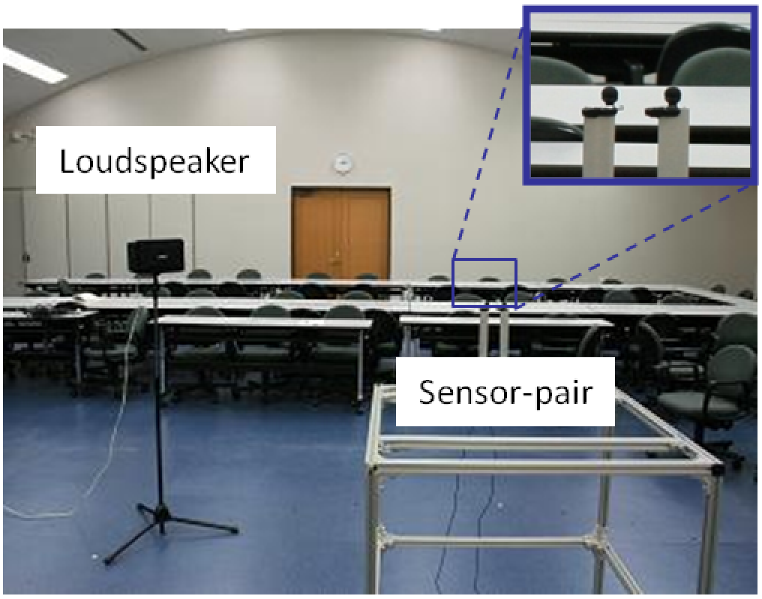


Figure 3.9: Experimental environment

Table 3.1: Experiment parameters

Source Signal	Speeches of 5 s
Sampling Frequency	8 kHz
Sound Velocity	340 m/s
Window	Hamming
STFT Frame Length	1024
Frame Overlap	512

3.8.2 Experimental results

The separation algorithm is based on the DOA estimation in SSA. the DOA estimation results are evaluated by estimation error θ_{error}, which is defined as

$$\theta_{error} = |\theta - \theta_{true}| \tag{3.53}$$

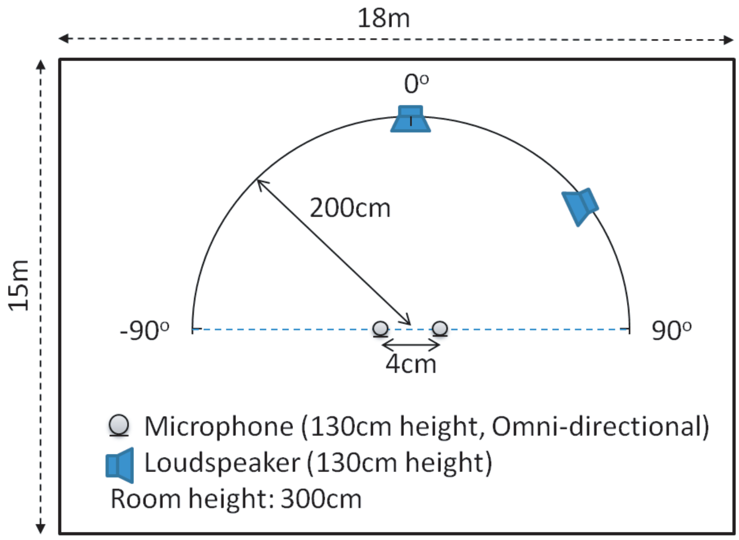

18m
15m
0°
200cm
-90°
90°
4cm
Microphone (130cm height, Omni-directional)
Loudspeaker (130cm height)
Room height: 300cm

Figure 3.10: Experimental setup

where θ is the estimated source direction, and θ_{true} is the true source direction. The DOA estimation results are shown in Fig. 3.11. It can be observed that the proposed method can properly detect the source direction. At the position of large source direction, the estimation error increases because of the low resolution near endfire (90°).

Fig. 3.12 shows the average SIR improvement results of the proposed and conventional delay-histogram methods. It is obvious that the proposed method exceeds the conventional method.

The effective separation of the proposed method is brought by integrating results of NSA, SSA, and DSA. Among them, the primary component is DSA separation, because it has many time frames, and contains a high power value that can influence the result to a large extent. The proposed method can match the component to the corresponding source on the basis of harmonic structure, but conventional method cannot. One example is shown in Fig. 3.13. The next contribution is obtained by SSA separation, which cannot improve the separation as well as DSA, but provides very important DOA information. The average improvement ratio for different types of time frames is shown in Tab. 3.2.

The accuracy of detecting SSA in this experiment is checked. Tab. 3.3 shows the results. The total number of SSA frames is estimated manually, and 75% of those frames are correctly detected by the proposed method. To the remaining 25% frames, which are undetected because of phase estimation error, the separation algorithm is applied. Tab. 3.4

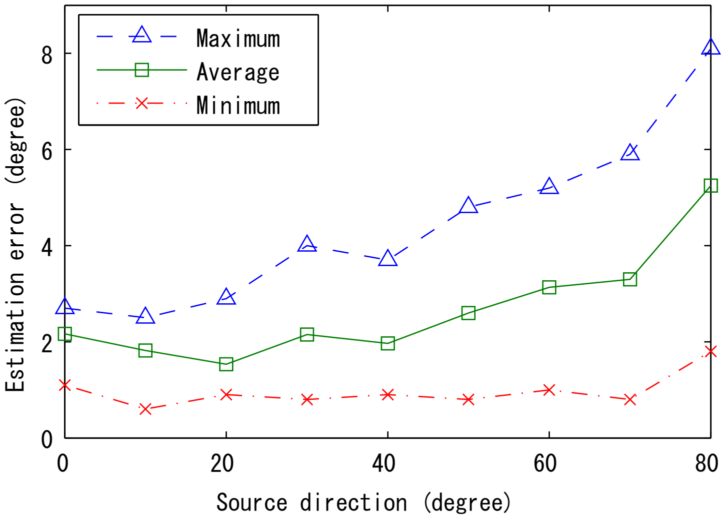


Figure 3.11: DOA estimation results in SSA

Table 3.2: Average improvement ratio

	SIR improvement (dB)	ratio
Total	6.22	100%
By NSA frame	0.58	9.3%
By SSA frame	1.36	21.9%
By DSA frame	4.28	68.8%

demonstrates the performance of harmonic structure detection in our approach. The rate of successfully detecting harmonic structure in B_{mid} relative to the DSA frames with vowel and/or vowel-like frames is shown. The initial separation ability will actually influence the estimation accuracy. One example of failure harmonics estimation caused by initial separation is shown in Fig. 3.14. The initial separation missed two harmonics in B_{mid}, so the proposed method cannot detect correct local maximum, and give wrong harmonics estimation in B_{low}.

3.9 Summary

A new time–frequency masking method was proposed for separating mixed speech signals utilizing phase difference versus frequency data in a frame-by-frame manner. The first

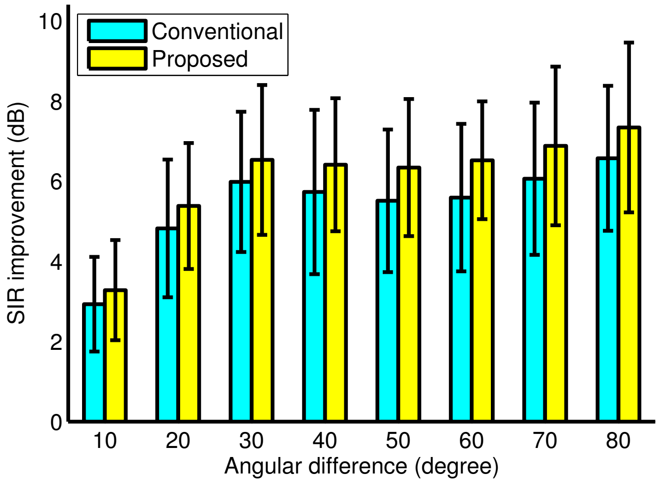


Figure 3.12: Experimental results: The error bar shows the standard deviation.

Table 3.3: SSA identification accuracy rate

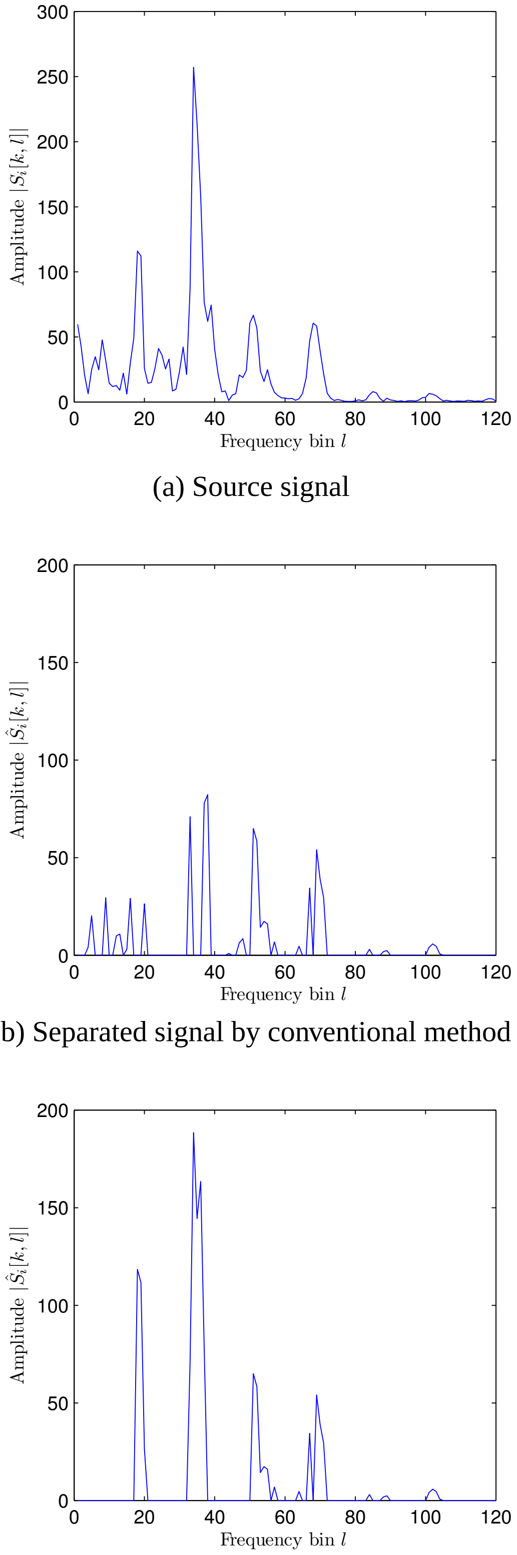
Total number of SSA frames	101
Correct identification by proposed method	75
Accuracy rate	74.2%

contribution of this study is the theory on stereophonic cases for estimating source directions by introducing a confidence measure at each time frame. The second contribution is the use of harmonic structure of initially separated signals appearing in the middle-frequency band through the consistency checking with the mixture spectrogram in the low-frequency band. The experiments were performed and the proposed method was evaluated by comparison with the conventional separation algorithm utilizing the delay feature. It was shown that enhancement was achieved.

Since the proposed separation method basically operates at each frame, the BSS prob-

Table 3.4: Harmonics estimation accuracy rate

Total number of DSA frames with harmonics	258
Correct estimation by proposed method	173
Accuracy rate	67.1%

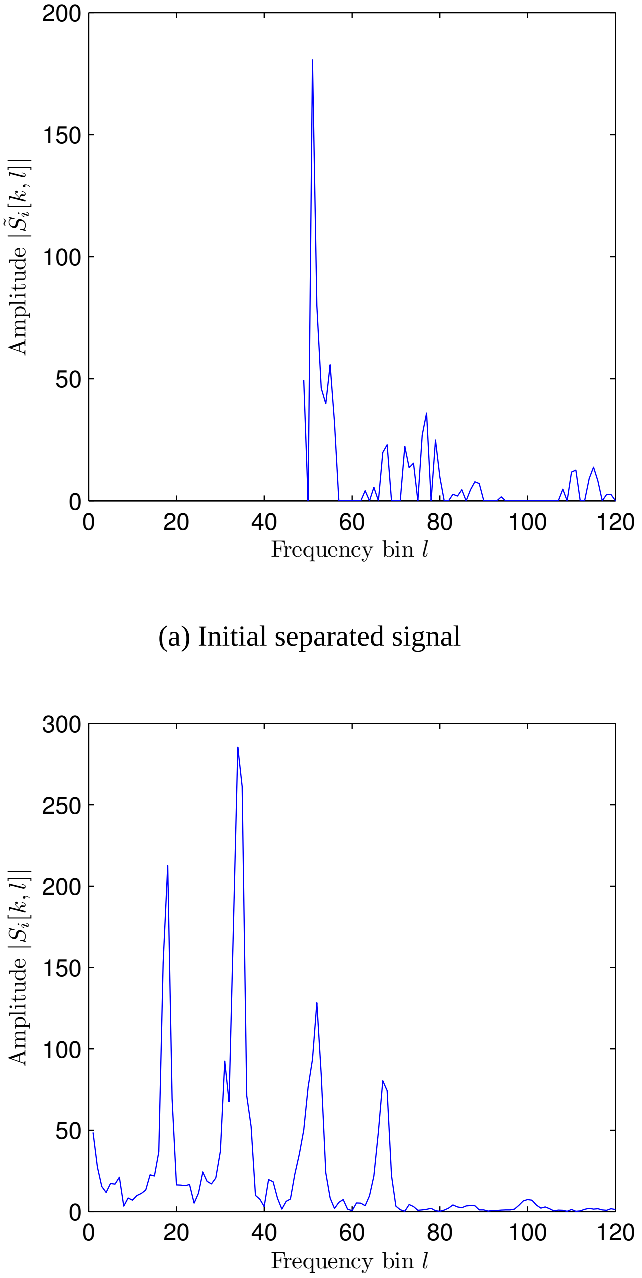


(a) Source signal

(b) Separated signal by conventional method

(c) Separated signal by proposed method

Figure 3.13: Comparison of separation results: The conditions of the mixture signal are: a male source at 0° and a female source at 60°. The above source signal is that of the male source.



(a) Initial separated signal

(b) Source signal

Figure 3.14: One example of failure harmonics estimation caused by initial separation. The mixture condition is: a male source at 14° and a female source at 50°. The above signal is that of the male source. The time frame $k = 28$.

lems for moving sources will be of interest in future studies. We compare our method to the previous one in which the delay or direction parameter is used without causing spatial aliasing. Therefore, the directional resolution is limited by small-distance sensor pairs. The comparison of the proposed method and the conventional methods on the basis of non-directional information, such as attenuation or harmonics, is another future issue.

Chapter 4

DOA estimation

4.1 Introduction

The underlying direction of arrival (DOA) estimation problems addressed in this chapter are listed as follows:

a) The use of a pair of microphones (doublet).

b) Multiple simultaneously uttered speech signal sources under the assumption that the number of sources is known a priori.

c) Underdetermined cases, where the sources outnumber the sensors.

d) The intersensor distance is bounded so as to avoid spatial aliasing. For instance, 4 cm spacing for an 8 kHz sampling rate.

As the doublet or stereophonic sensor investigated in this paper is the simplest array sensor system, its array processing capability is obviously limited. Nevertheless, the study of how to improve the accuracy of a DOA estimator implemented in a doublet is meaningful because any complex array configuration can be considered as a combination of doublets. In addition, an effective method for doublets could be generalized to more complicated multiple-doublet systems [54] [73].

As stated in b), the DOA estimation problem considered here deals with estimating the DOA of multiple speech signals uttered simultaneously. The same framework is commonly used in blind source separation (BSS) which means that the separation problem of speech mixtures can be solved by only considering observations at microphones. Therefore, obtaining an accurate DOA in the case of multiple sources is closely related to the BSS problem. The underdetermined situation c) means that the proposed method should be applicable to cases where the number of sources can exceed the number of sensors. This condition is important when dealing with practical audio conditions and has been the main focus of recent BSS approaches [22–25, 29, 46, 54–56, 63, 73].

Two novel DOA estimation methods are proposed involving (1) the Hough transform [74] and (2) reliability index and kernel density estimation [75–78]. The first method applies the Hough transform to the phase difference (PD) versus frequency (PD-F) distribution of received mixed signals and estimates the DOA. By introducing the bandwidth in the Hough parameter space, the errors in the real data are considered, and the stability and accuracy of DOA estimation are guaranteed. This approach can also be regarded as an attempt to combine sounds and images.

The DOA estimation using the reliability index and kernel density estimation is based on the following three novel approaches.

1) Inspired by the ideas of TIme-Frequency Ratio Of Mixtures (TIFROM)-like assumptions, a novel reliability index is introduced. Then, the selected cells with higher reliability are solely utilized for DOA estimation.

2) A statistical error propagation model relating PD estimation and the consequent DOA is introduced. The model leads to a probability density function (PDF) of the DOA, and then the DOA estimation problem is reduced to finding the most probable points of the PDF.

3) The final DOAs are determined using the kernel density estimator by utilizing a proposed bandwidth control strategy.

Comparing the proposed method with DEMIX, the proposed method utilizes the PD and its frequency-normalized quantity, i.e., the TDOA between sensors is used instead of using the steering vectors in DEMIX. The PD is focused by eliminating the amplitude factor of the steering parameter as in generalized cross-correlation phase transform (GCC-PHAT). The reliability or confidence measure of a time-frequency (T-F) cell's PD is used to exploit the consistency of the TDOA in the local window of the underlying cell. This idea was derived from the following assumption: when a single source occurs in a given set of T-F windows, TDOAs should take almost the same values in the windows and these values are considered to be reliable. The consistency of the TDOA in a window is evaluated by using the variance of the TDOAs for whole T-F cells in the window.

The second investigation based on the introduced statistical error propagation model shows that the DOA estimation problem can be altered to obtain the local maxima or peaks of a DOA PDF generated by an assumed error distribution in the PD. The use of kernel density estimation is the third novel point.

Starting from the time order, the DOA estimation method using the Hough transform is first proposed, then the method using the reliability index and kernel density estimator is proposed. Compared with the former method, cell selection using the reliability index and the DOA error distribution model in the second method is more novel and effective. This is why the second method is mainly discussed in this chapter, especially in the experimental part. However, some advantages of DOA estimation using the Hough transform should also be noted, for example, its outstanding performance in dealing with spatial aliasing.

4.2 DOA information

The observation model is the same as that in Chapter 3. When an anechoic model without a signal level difference between sensors is assumed, and the first sensor ($m = 1$) is used as the reference, according to the Eq. (2.16) and WDO assumption Eq. (2.24), the ratio between two received signals $X_m[k,l]$ is

$$\frac{X_2[k,l]}{X_1[k,l]} = \frac{H_{2n}[l]}{H_{1n}[l]} = \exp[j\frac{2\pi f_s l}{L} \cdot \frac{d}{c}\sin\theta], \tag{4.1}$$

where d is the distance between the sensors, c is the sound velocity, and θ is the source direction. $\theta = 0$ corresponds to the broadside direction. Using

$$X_m[k,l] = |X_m[k,l]| \exp[j\angle X_m[k,l]], \tag{4.2}$$

then the PD $\phi[k,l]$ between two observations $X_m[k,l]$ ($m = 1,\ 2$) is defined by

$$\phi[k,l] = \angle X_2[k,l] - \angle X_1[k,l]. \tag{4.3}$$

Finally,

$$\phi[k,l] = \frac{2\pi f_s l d}{Lc}\sin\theta = \Delta\omega T l \sin\theta, \tag{4.4}$$

where $T = \frac{d}{c}$ is the maximum delay time between sensors and $\Delta\omega = \frac{2\pi f_s}{L}$ is the unit frequency width in L-point short-time Fourier transform (STFT) analysis. From Eq. (4.1), the TDOA normalized by T, denoted by $\delta[k,l]$, can be obtained by the following frequency normalization.

$$\delta[k,l] = \sin\theta = \frac{\phi[k,l]}{T\Delta\omega l} \tag{4.5}$$

4.3 DOA estimation using delay histogram and Hough transform

4.3.1 Delay histogram

A conventional method of DOA estimation based on T-F clustering methods is the delay-histogram algorithm based on the DUET [22]. This method generates the histogram distribution of delays $\delta[k, l]$ that are generated by PD as given by Eq. (4.5). Then number of peaks in the obtained histogram is the same as the number of sources that are detected, and these peaks are used to determine the DOAs. Some examples of DOA estimation using a delay histogram are shown in Fig. 4.1.

4.3.2 Hough transform

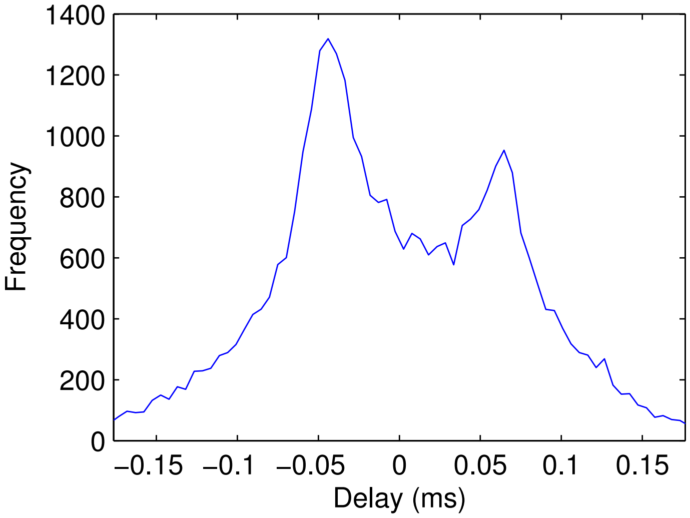
Because DOA estimation corresponds to finding a linear phase relationship in a PD-F distribution, the Hough transform can be applied as a line extraction technique. In our case of multiple DOA estimation, the problem is to fit a number of lines, each of which corresponds to an individual source. The Hough transform, named after Paul Hough who patented the method in 1962, is a feature extraction technique originally used in image analysis, computer vision, and digital image processing.

A useful parameterization for a straight line in (x, y) plane is to consider its shortest distance from the origin ρ and its orientation θ:

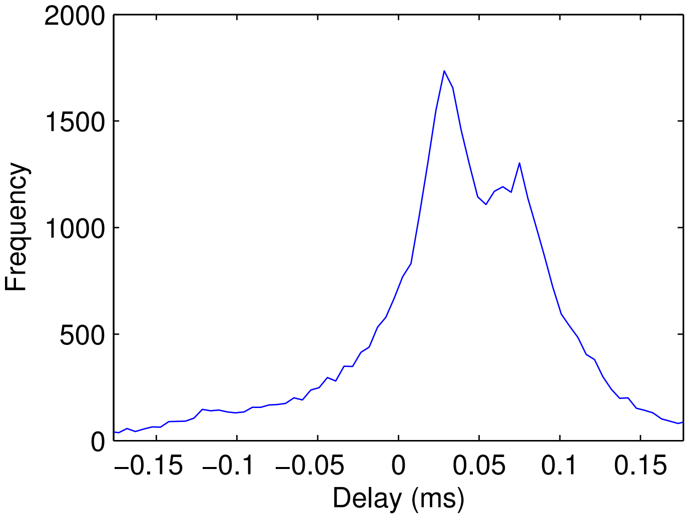
$$\rho = x \cos\theta + y \sin\theta. \tag{4.6}$$

For a normalized PD-F distribution, the orientation θ is equivalent to β as defined by Eq. (3.21).

The first study in which the source direction was estimated using the Hough transform was that of Suzuki et al. [79]. The main purpose of their study was to develop an omnidirectional acoustic sense with which a robot can localize and recognize multiple sounds from an unlimited number of directions even in a noisy environment. They first investigated the relationship between source direction and PD, then used the Hough transform to detect straight lines from the PD-F space for the detection and localization of sound sources. In their Hough transform approach, a histogram of the gradients given by the PD-F data is generated. The peaks of this histogram are then used to estimate the DOAs. Therefore, the Hough method with the constraint $\rho = 0$ is identical to the delay-histogram method.



(a) The mixture condition is a male source located at −23° and a female source located at 34°. The estimation results are −22° and 33°.



(b) The mixture condition is a male source located at 14° and another male source located at 42°. The estimation results are 14° and 40°.

Figure 4.1: Some examples of DOA estimation using a delay histogram

Another Hough transform approach for the blind localization of several sound sources from two binaural signals was proposed by Marchand et al. [80] in 2009. First, the binaural signals are organized as two-dimensional (2D) data, where each sound source appears as a line. Second, the Hough transform is used to recognize these lines. The slopes of the lines give the mixing coefficients and directions of arrival (azimuths). On the basis of only one of the interaural levels or time differences, two variants of their methods are given. Since this new contribution is based on source sparseness, the method can deal with underdetermined cases, which means that three source directions can be estimated using two received signals. This method as well as that of Suzuki et al. are applicable in cases involving spatial aliasing.

Unlike [79] and [80], the proposed method in this thesis introduces the concept of bandwidth in the Hough transform, which reflects the phase estimation error under real acoustic conditions and obtains accurate results. In the next section, the proposed Hough transform with a bandwidth approach is introduced, which is followed by the T-F cell selection process.

4.3.3 Hough transform in PD-F distribution

An example of a PD-F data set defined by the vectors $\{l, \phi[k,l]\}$ in a 2D plane is shown in Fig. 4.2(a). In the proposed method, the following frequency band is used:

$$B_{high} := \{l | l > l_1\}, \tag{4.7}$$

where $l_1 = \lfloor (f_1 \cdot L/f_s) \rfloor$, $f_1 = 400$ Hz, and $\lfloor \rfloor$ is the Gauss floor function, which maps a real number to the largest previous integer.

Cell selection and normalization

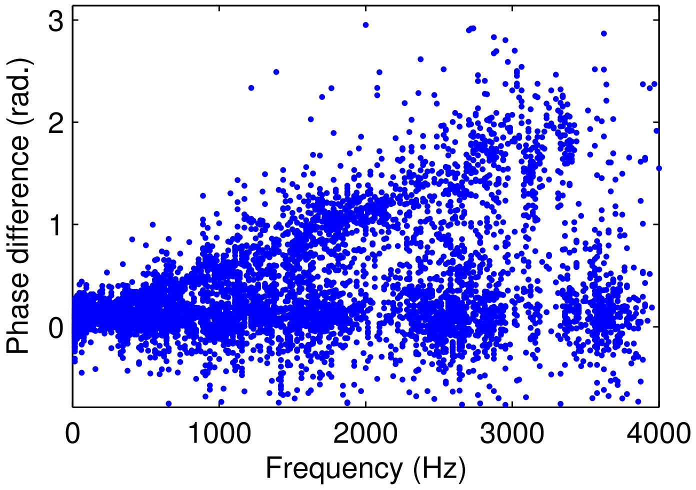
A set of T-F cells, that satisfy the following two conditions is selected:

(1) Because the PD in the low-frequency band ($l \neq B_{high}$) is too small for accurate estimation, l is restricted such that $l \in B_{high}$.

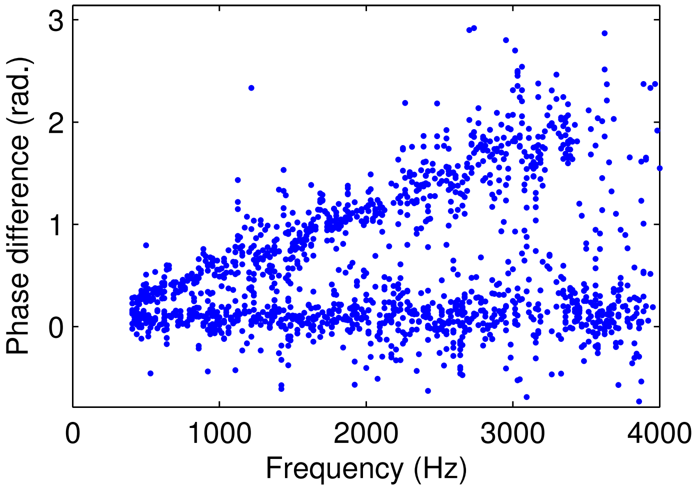
(2) The maximum amplitude value $A(l)$ in each frequency bin is defined by $A(l) = \max_{k \in [0,K]} |X_1[k,l]|$, and then the T-F cells $[k,l]$ satisfying

$$\gamma[k,l] = \frac{|X_1[k,l]|}{A(l)} \geq Th_1 \tag{4.8}$$

are selected, where $\gamma[k,l]$ is used as the weight factor in the Hough transform, and $Th_1 = 0.5$ is set on the basis of the result of experiments. This criterion is used to detect the line in the PD-F distribution. Owing to the use of the relative power for cell selection, the points in the PD-F distribution will be distributed in every frequency bin, rather than in only



(a) All T-F cells [duration 5 seconds]



(b) Selected T-F cells [duration 5 seconds]

Figure 4.2: PD-F distribution. $f_s = 8$ kHz. The mixture condition is a male source located at 4° and another male source located at 51°.

some frequency bins, which will be helpful for obtaining the line direction using the Hough transform. The selected cells $[k, l]$ are denoted as Ω_1. An example of a PD-F distribution of Ω_1 is shown in Fig. 4.2(b).

For the analysis, all vectors in Ω_1 are normalized by

$$[y(l), z_k(l)] := [l/(L/2), \phi[k, l]/\pi]. \tag{4.9}$$

Angle range

The gradient of a line from the origin in the normalized PD-F plane, denoted by α, is related to the actual DOA θ (degrees) by the equation $\theta = \arcsin[\frac{Lc}{2\pi f_s d} \cdot \tan\alpha]$. In addition, the theoretical limit of α is given by $|\alpha| \leq \arctan|\frac{2\pi f_s d}{Lc}|$, where d is the distance between sensors and c is the sound velocity. From this inequality, α is restricted within the interval $|\alpha| < \alpha_{limit}$.

Calculation of Hough transform

By transforming the 2D grid index $[k, l] \in \Omega_1$ into an arbitrary one-dimensional (1D) alignment integer index n, the corresponding formula is obtained

$$[x(l), y_k(l)] \to [x_n, y_n], \gamma[k, l] \to \gamma_n, [k, l] \in \Omega_1. \tag{4.10}$$

The Hough transform is calculated by

$$\rho_n(\alpha) = x_n \cdot \cos\alpha + y_n \cdot \sin\alpha, \ \ |\alpha| < \alpha_{limit}, \tag{4.11}$$

where ρ is the shortest distance from the origin to the line.

4.3.4 Hough transform with bandwidth

In theory, the DOA corresponds to the gradient angle α for $\rho = 0$. However, to consider the phase estimation error at each frequency, the interval $|\rho_n(\alpha)| \leq \epsilon_0 \cos\alpha$ at each α is combined into a unit rectangular cell for the Hough voting procedure. The bandwidth interval of Hough transform $|\rho_n|$ is related with the allowable error range of phase difference estimation ϵ_0 and orientation α, which is shown in Fig. 4.3. Throughout the experiments in this study, ϵ_0 is set to 0.03.

Assume that T-F cells with low power are not confidence, therefore the relative amplitude ratio is utilized as the criterion to select reliable cells. At the same time, the selected cells should play different roles in the voting process of Hough calculation depending on their confidence ability. The relative amplitude ratio is therefore used as weight factor,

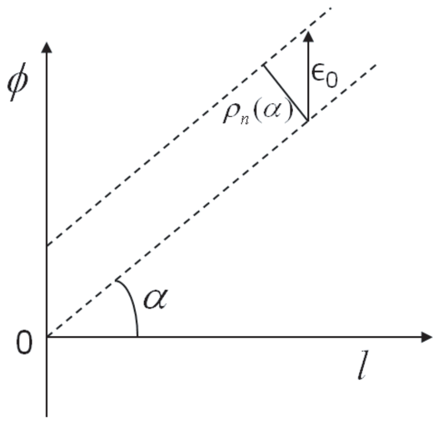


Figure 4.3: The reason why $|\rho_n|$ is related with α.

which means the high power cells will devote more in Hough calculation. The intersection value at angle α (denoted by $IV(\alpha)$) with weight γ_n is calculated as

$$IV(\alpha) = \sum_{\Omega_1} \gamma_n, \quad if\ |\rho_n(\alpha)| \leq \epsilon_0 \cos\alpha. \tag{4.12}$$

As demonstrated in Fig. 4.4, the obtained value in experiments diverges from the theoretical value at the same frequency. If this value is still regarded as voting for the line through the origin, it will give a false DOA estimation. However, all the lines with the same slope but various values of ρ within the bandwidth $\pm\rho_n(\alpha)$ are considered, the effect of errors will be diminished significantly, and the allowable error can also vote for the correct PD distribution line.

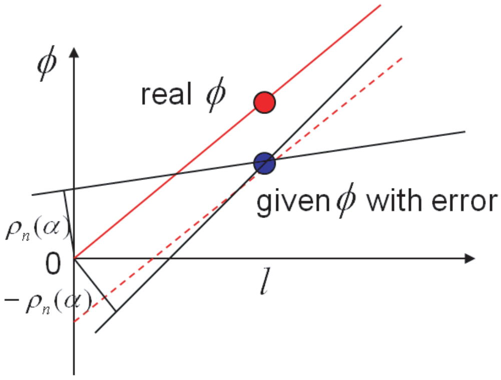


Figure 4.4: Demonstration of bandwidth in Hough transform. The mixture condition is a male source located at 4° and another male source located at 51°.

In practice, $IV(\alpha)$ is evaluated at sampled α values, such as integer values within the interval $[-\alpha_{limit}, \alpha_{limit}]$. The DOA estimation is performed as follows. The α_1 that maximizes $IV(\alpha)$ gives the first source direction θ_1 using the relationship between α and θ. Next, the DOA θ_2 is obtained using the α_2 at which $IV(\alpha_2)$ is a local maximum taking a submaximum value and θ_2 is more than 10° apart from the estimated DOA θ_1. Some examples of intersection-value histograms are shown in Fig. 4.5.

4.4 Reliable T-F cell selection

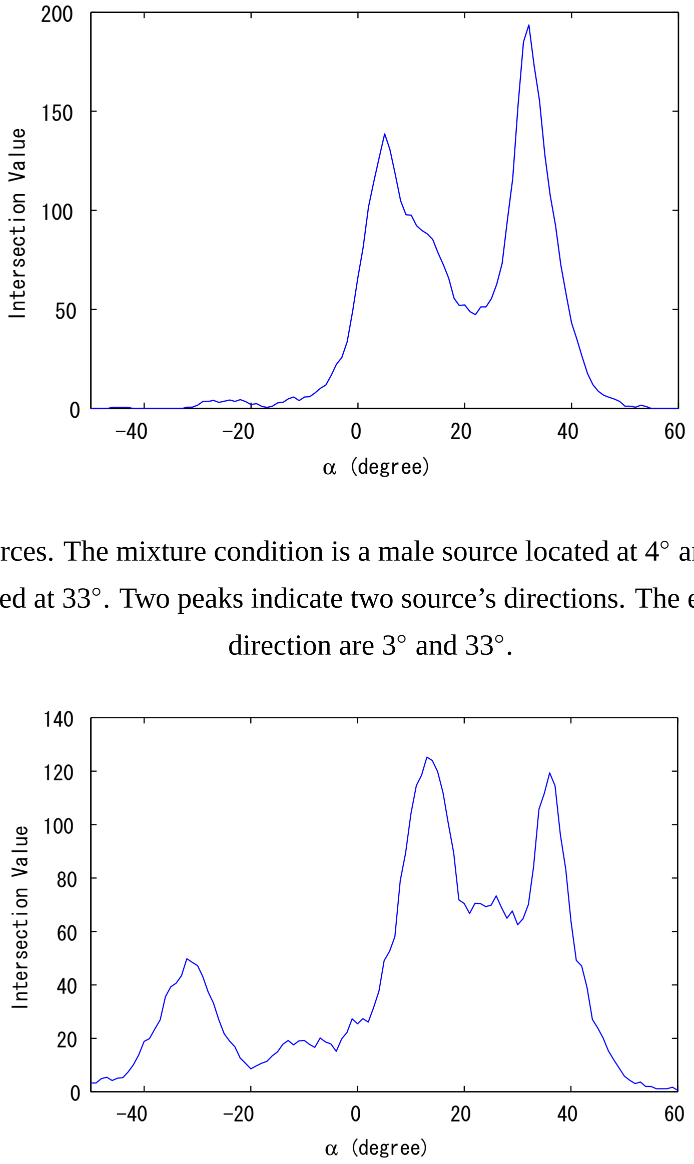
4.4.1 Preselection of T-F cells

Before selecting a set of reliable T-F cells, as discussed in the next section, the preselection of T-F cells on the basis of their local T-F power is performed. In this approach, T-F cells whose amplitudes have significantly small values, namely, T-F cells for which the inequality

$$|X_m[k, l]| < Th_1 \tag{4.13}$$

is satisfied, are deleted to avoid unnecessary computation because the PD estimation errors at these T-F points are relatively large. To investigate this, an experiment was performed using real observed data from ten speakers with known DOAs. Since a known DOA gives a real PD value, the PD estimation error can be obtained. Fig. 4.6 shows the obtained relationship between the PD estimation error and the amplitude $|X_m[k, l]|$. All the observed signals in this experiment were normalized by scaling to satisfy $\max_t |s(t)| = 1$. The ten relationships between the PD estimation error and $|X_m[k, l]|$ for individual speakers are shown by dotted lines in the figure. The solid line in the figure shows the average relationship over all speakers. From this figure, it can be seen that the PD estimation error is only related to the amplitude at significantly small amplitudes.

Concerning the local power of T-F cells, the DUET [22] utilizes a cell's attenuation ratio and delay to enhance separation ability. However, the result shown in Fig. 4.6 indicates that the local power of a cell has no specific relation to the magnitude of the PD estimation error at the cell unless it has significantly small power. In this setup, $Th_1 = 2.0$ is set on the basis of the experimental results in Fig. 4.6. Further observations on the determination of Th_1 will be discussed in Sec. 4.8.1.



(a) Two sources. The mixture condition is a male source located at 4° and another male source located at 33°. Two peaks indicate two source's directions. The estimated source direction are 3° and 33°.

(b) Three sources. The mixture condition is a female source located at −42°, a male source located at 14°, and another female source located at 51°. Three peaks indicate three source's directions. The estimated source direction are: −40°, 15° and 48°.

Figure 4.5: Two examples of intersection-value histograms

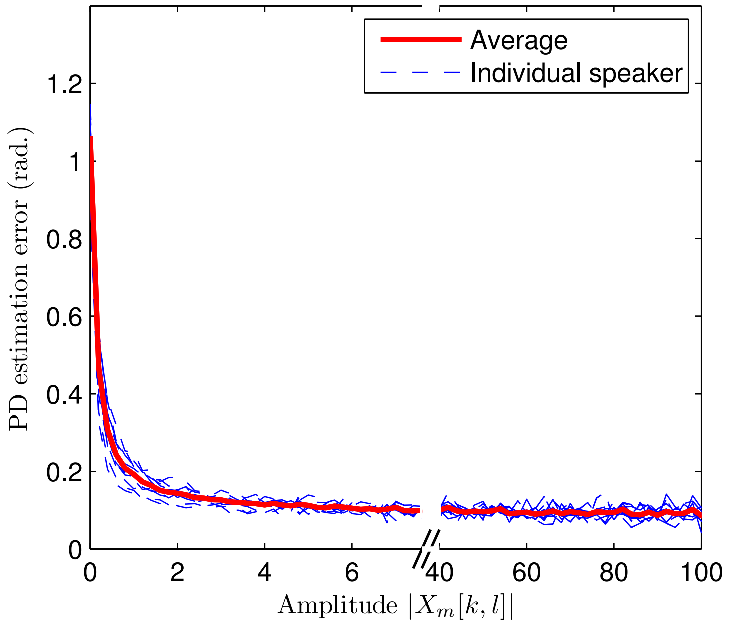


Figure 4.6: Relationship between PD estimation error and amplitude. Each individual speaker is one of 10 speakers: 5 males and 5 females.

4.4.2 Reliable cell selection

The PD estimation is subjected to unavoidable error. The success of the proposed method of DOA estimation is expected if more reliable PD data are selected and outliers are eliminated. Similarly to in [28], the following assumption is employed. When one source components is dominant in a set of cells, all delays in it will take almost the same value; hence, the delay and obviously the PD data in this set are expected to be reliable.

Conventionally, the confidence measure is obtained from the results of applying principal component analysis to a set of steering vectors in individual horizontal and vertical T-F regions in [28]. In [81], the entropy of the estimated DOAs in a rectangular T-F neighborhood region was employed. Unlike these methods, the normalized delays given by Eq. (4.5) are used to evaluate the feature consistency of the T-F cells in various regions, and a novel reliability measure is introduced.

According to the above assumption, two types of T-F regions around cell $[k,l]$ are considered: a temporal neighborhood $\Gamma_t[k,l]$ and a frequency neighborhood $\Gamma_f[k,l]$,

$$\Gamma_t[k,l] := \{[k+y,l] \mid |y| \leq Y\} \tag{4.14}$$

$$\Gamma_f[k,l] := \{[k,l+z] \mid |z| \leq Z\}, \tag{4.15}$$

where integers Y and Z determine the numbers of cells in these regions, as denoted by $|\Gamma_t[k,l]| := 2Y + 1$ and $|\Gamma_f[k,l]| := 2Z + 1$. For each $\Gamma_t[k,l]$ and $\Gamma_f[k,l]$, the standard deviations of the normalized delays $\sigma_{\Gamma_t}[k,l]$ and $\sigma_{\Gamma_f}[k,l]$ are calculated by

$$\sigma_\Gamma[k,l] = \sqrt{\frac{1}{|\Gamma|}\sum_{[p,q]\in\Gamma}(\delta[p,q]-\mu_\Gamma[k,l])^2}, \tag{4.16}$$

where

$$\mu_\Gamma[k,l] = \frac{1}{|\Gamma|}\sum_{[p,q]\in\Gamma}\delta[p,q], \quad \Gamma = \Gamma_t,\ \Gamma_f. \tag{4.17}$$

The reliability index $\eta[k,l]$ is calculated by

$$\eta[k,l] = \exp\{-\min(\sigma_{\Gamma_t}[k,l],\ \sigma_{\Gamma_f}[k,l])\}, \tag{4.18}$$

where $\eta[k,l]$ is a normalized index satisfying $0 < \eta \leq 1$. When $\sigma_{\Gamma_t}[k,l]$ and/or $\sigma_{\Gamma_f}[k,l]$ at $[k,l]$ is sufficiently small, $\eta[k,l]$ approaches unity, consequently the corresponding delay value $\delta[k,l]$ is considered to be reliable.

To verify the validity of the introduced reliability index, the relationship between the PD estimation error and the reliability index η is observed with Y and Z both set to 1 for the experimental data with a known DOA. The experimental data used in this investigation are the same as those used in Fig. 4.6. Fig. 4.7 shows both the results of individual speakers and their averaged characteristic. The figure indicates that the PD error decreases as the reliability index increases. Then, the cell group is selected with the highest reliability index $\eta[k,l] > \eta_{th}$ for subsequent DOA estimation. In this paper, η_{th} is set to 0.96. The reason for using this value and related remarks are given in Sec. 4.8.1.

To present quantitative evidence of the assignment $Y = 1$ and $Z = 1$, as discussed above, the correlation between the absolute PD estimation error $|\Delta\phi|$ and the reliability index η is investigated. Since a definite negative cross-correlation between these is appropriate, it is evaluated for several combinations of Y and Z, then the values of Y and Z giving the largest negative correlation are used as optimal values. For each pair of positive integers Y and Z, the sampled cross-correlation Q is calculated for the given data $\{|\Delta\phi_i|,\ \eta_i | i = 1, 2, \ldots, I\}$ by

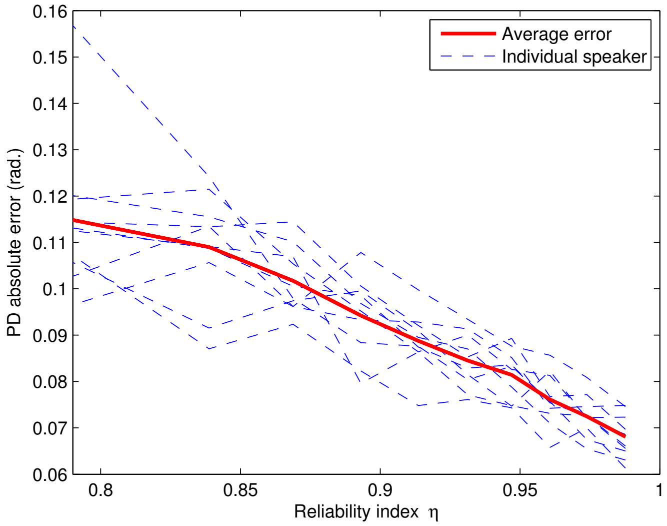


Figure 4.7: Relationship between reliability index and PD estimation error setting $Y = 1$ and $Z = 1$. The individual speaker is varied among 10 speakers: 5 male and 5 female.

$$Q = \frac{\frac{1}{I-1}\sum_{i=1}^{I}(|\Delta\phi_i| - \overline{|\Delta\phi|})(\eta_i - \overline{\eta})}{\sqrt{\sum_{i=1}^{I}(|\Delta\phi_i| - \overline{|\Delta\phi|})^2}\sqrt{\sum_{i=1}^{I}(\eta_i - \overline{\eta})^2}}, \tag{4.19}$$

where $\overline{|\Delta\phi|}$ and $\overline{\eta}$ are the sample averages of $\{|\Delta\phi_i|\}$ and $\{\eta_i\}$, respectively.

Among the values of Q for various Y and Z, the case of $Y = 1$ and $Z = 1$ has the largest negative cross-correlation value with $Q(Y = 1) = -0.85$ and $Q(Z = 1) = -0.83$. In other cases, for example, $Y = 2$ and $Z = 2$, $Q(Y = 2) = -0.68$ and $Q(Z = 2) = -0.64$.

For each selected reliable T-F cell, the direction θ is computed using Eq. (4.5). Here the set of computed directions is denoted as follows:

$$\left\{\theta_i^{[l_i]} | i = 1, 2, \ldots, I\right\}, \tag{4.20}$$

where i is the numbering integer of the selected cells, I is the total number of data, and l_i is the frequency bin in which the ith cell is located.

4.5 DOA error distribution model

In the problem considered in this study, the reliable T-F cells selected in previous section consist of the components of multiple sources. Even so, each of the selected T-F cells corresponds to the component of one source signal. Consider a T-F cell at which the *n*th source dominates and is located in the unknown direction θ_n. From Eq. (4.4), the theoretical PD in the cell is given by

$$\phi_n[l] = \Delta\omega T l \sin\theta_n = B_n l, \tag{4.21}$$

where $B_n = \Delta\omega T \sin\theta_n$. Since k is not essential in this section, henceforth the frame index k is omitted. In the lth frequency bin, the observed $\phi_n[l]$ is distributed around its mean value $B_n l$,

$$\phi_n[l] = B_n l + \Delta\phi[l], \tag{4.22}$$

where $\Delta\phi[l]$ is a random variable representing the PD estimation error.

Then, assume that the random variable $\Delta\phi[l]$ is an independent identical Gaussian distribution with zero mean and constant variance σ_ϕ^2. The constant variance means that $\Delta\phi[l]$ is independent of the frequency bin l; this assumption is represented as follows:

$$\Delta\phi[l] \sim N(0, \sigma_\phi^2). \tag{4.23}$$

Since main concern is with the first- and second-order statistics, the type of distribution is not essential in developing the algorithm. The use of a Gaussian distribution is motivated from the simplicity of theoretical manipulation. To verify that this assumption is really satisfied in a statistical sense, it was checked experimentally. For the observed speech signals of ten speakers (5 males and 5 females) from a known DOA which is varied from 0° to 70° at 10° intervals, PD estimation errors in individual frequency bins are calculated. The averaged value and the standard deviation around it are calculated for each l . The PD estimation error for the ten speakers from the individual DOAs are shown in Fig.4.8 as dotted lines. The average value and the standard deviation around it for each l are also calculated and illustrated in the figure. From Fig. 4.8, it can be verified that the average PD error $\Delta\phi[l]$ is approximately zero and the standard deviation of PD error σ_ϕ is almost constant. This means that assumption (4.23) is approximately satisfied.

Under assumption (4.23), the source direction θ_n estimated by Eq. (4.5) is also considered as a random variable, which is denoted by $\theta_n^{[l]}$. Even though $\Delta\phi[l]$ is an independent random variable with respect to l, the estimated source direction has a variance depending on l. Now, the following proposition can be proved.

Proposition: If the random variable $\Delta\phi[l]$ is given by (4.23) and σ_ϕ is sufficiently small, the PDF of $\theta_n^{[l]}$ is given by

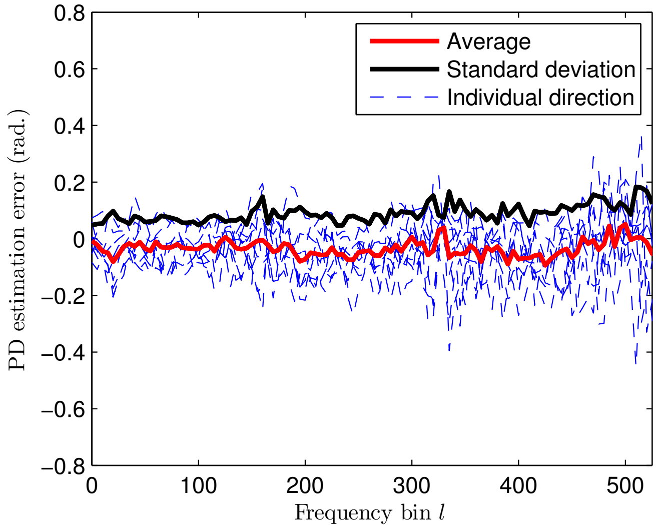


Figure 4.8: Mean and standard deviation of PD error vs frequency observed from ten individual speakers

$$\theta_n^{[l]} \sim N(\theta_n,\ \sigma_{\theta_n}^2[l]), \tag{4.24}$$

where

$$\sigma_{\theta_n}[l] = \frac{1}{T\Delta\omega l \cos\theta_n}\sigma_\phi. \tag{4.25}$$

The proof of this is as follows.

Denoting $\Delta\theta_n$ to represent the deviation of the estimation error from the real value θ_n, the following relationship between $\Delta\theta_n$ and $\Delta\phi[l]$ is obtained.

$$\sin(\theta_n + \Delta\theta_n) = \frac{B_n l + \Delta\phi}{T\Delta\omega l} \tag{4.26}$$

From the assumption that σ_ϕ is sufficiently small, $\Delta\theta_n$ takes a small value. Therefore, the left-hand side of Eq.(4.26) can be replaced by the first-order approximation with respect to $\Delta\theta_n$ as

$$\sin\theta_n + \cos\theta_n \cdot \Delta\theta_n = \frac{B_n}{T\Delta\omega} + \frac{\Delta\phi}{T\Delta\omega l}. \tag{4.27}$$

By substituting Eq. (4.21), the relationship is satisfied

$$\Delta\theta_n = \frac{1}{T\Delta\omega l \cos\theta_n}\Delta\phi. \tag{4.28}$$

From the assumption that $\Delta\theta_n \sim N(0,\ \sigma^2_{\theta_n})$, Eq. (4.25) can be derived. The DOA error distribution model in shown in Fig. 4.9.

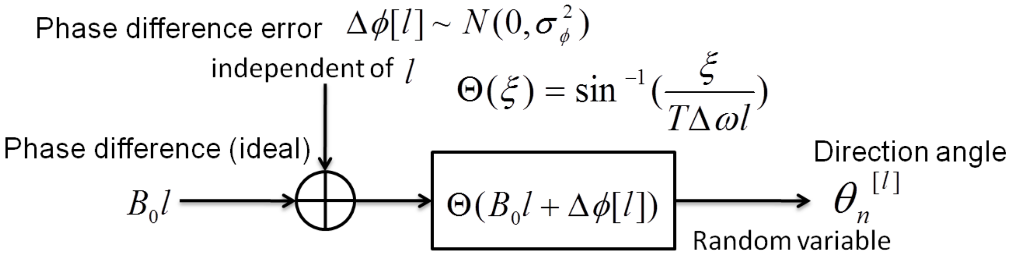

Phase difference error $\Delta\phi[l] \sim N(0, \sigma_\phi^2)$
independent of l
$\Theta(\xi) = \sin^{-1}(\frac{\xi}{T\Delta\omega l})$
Phase difference (ideal)
$B_0 l$
$\Theta(B_0 l + \Delta\phi[l])$
Direction angle
$\theta_n^{[l]}$
Random variable

Figure 4.9: DOA error distribution model

4.6 DOA estimation using kernel density estimator

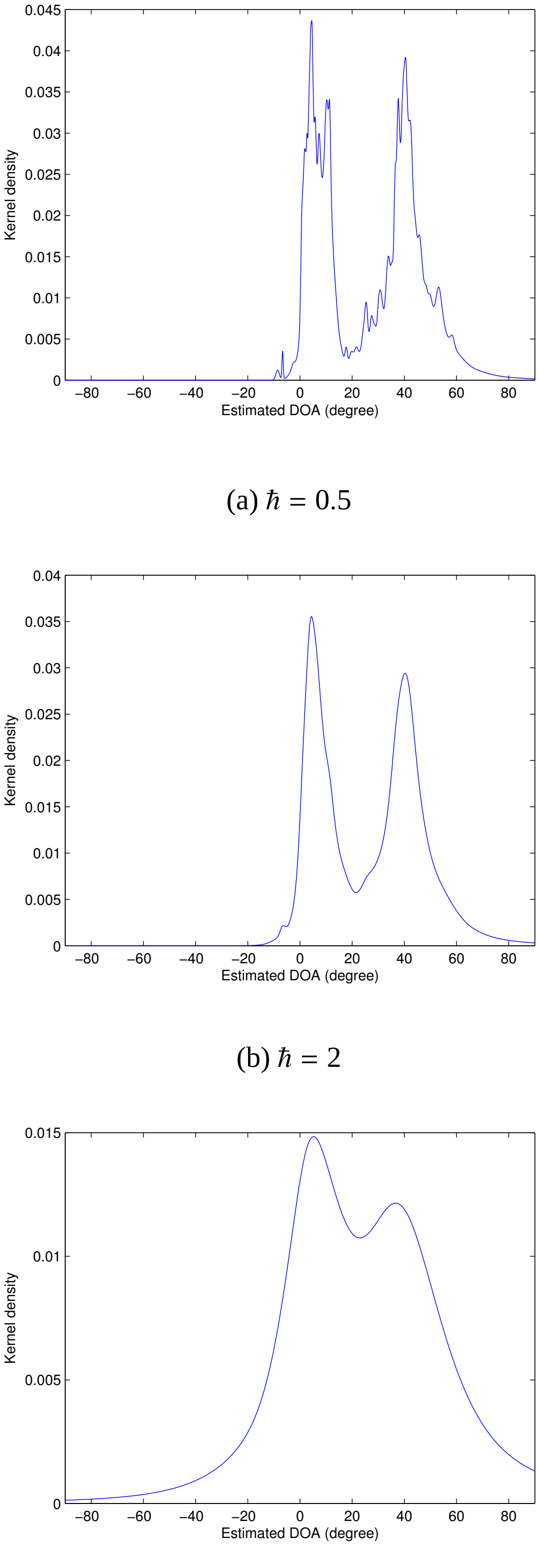
The kernel density estimation algorithm [82] or Parzen window approach [83] is useful for statistical estimation even for a multiple-source problem. The algorithm is used to obtain the PDF of $\theta^{[l]}$, which is theoretically described by Eq.(4.24), by using only observed samples.

If a large number of reliable observations $\theta^{[l]}$ can be obtained, the estimated PDF becomes reliable. This means that the maximum PDF point, namely, the mode of the PDF, can be considered as the optimal estimate of θ_n in the sense of the most probable value. In the kernel density estimator approach, the DOA estimation problem is reduced to the approximate estimation of the PDF of $\theta^{[l]}$.

It is necessary to generalize the above theoretical investigation to multisource and multifrequency cases. The theoretical PDF formulation of θ in the case of multiple sources should be a Gaussian mixture with the same number of local modes (local peaks), each of which corresponds to an individual source. For the selected reliable data used in Eq. (4.20), the kernel density estimator is applied to estimate the multimodel PDF as follows:

$$\hat{p}(\theta) = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{\epsilon[l_i]}K(\frac{\theta - \theta_i^{[l_i]}}{\epsilon[l_i]}), \tag{4.29}$$

where $K(\theta)$ is a kernel function, for which a Gaussian function is adopted in this study. $\epsilon[l]$ is the bandwidth of the kernel. The idea behind applying the kernel density estimator is to reflect the theoretical result obtained by the above proposition in the determination of bandwidth. Since the variance of $\theta^{[l]}$ depends on l and θ_n as indicated in Eq. (4.25), the bandwidth is selected using



(a) $\hbar = 0.5$

(b) $\hbar = 2$

(c) $\hbar = 5$

Figure 4.10: Estimated PDF for various $\hbar$. The mixture condition is a male source located at 4° and another male source located at 42°.

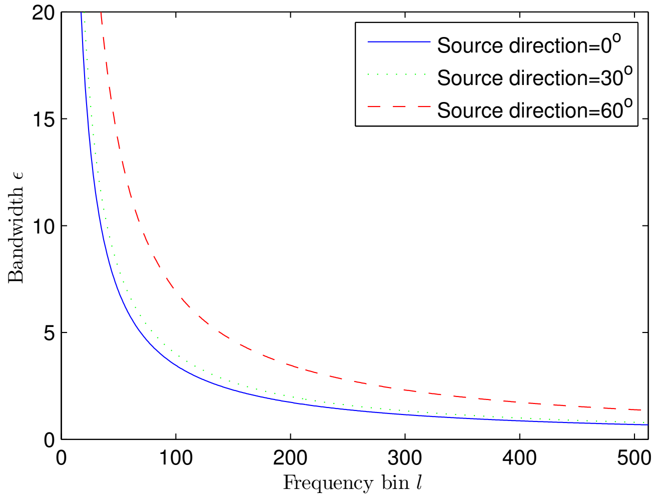


Figure 4.11: The relation between kernel bandwidth ϵ and frequency bin l with several source direction θ.

$$\epsilon[l_i] = \frac{1}{T\Delta\omega l_i \cos\theta_i^{[l_i]}}\hbar, \tag{4.30}$$

where $\hbar$ is the control parameter and the observed $\theta_i^{[l_i]}$ is substituted in place of a real unknown θ_n in Eq. (4.25). Accordingly, the dependence of the bandwidth on θ_n is indirectly controlled. The control parameter $\hbar$ is predetermined experimentally. Fig. 4.10 shows three examples of estimated PDFs for a two-source case with different $\hbar$. As discussed in Sec. 4.8.1, $\hbar = 2$ is hereafter used. The relation between kernel bandwidth ϵ and frequency bin l with several source direction θ is shown in Fig. 4.11.

Finally, by finding the same number of local modes (peaks) as the number of pre-assigned source numbers, the source directions are estimated. The peaks of the estimated PDF used to determine θ_n are found by an exhaustive search. Because the estimated PDF is 1D, an exhaustive search is very effective. In the computation, MATLAB 7.12.0 is used with a PC, an Intel Core2 Quad CPU (2.83 GHz) and a Windows XP OS. The computation time required to search for the peaks is 0.6 ms.

4.7 Experimental conditions

Some experiments were performed in a conference room to evaluate the proposed methods. The experimental setup and parameters are the same as those in Chapter. 3. The amplitude information obtained from the original observation features is neglected. This type of simplification is mentioned in [54].

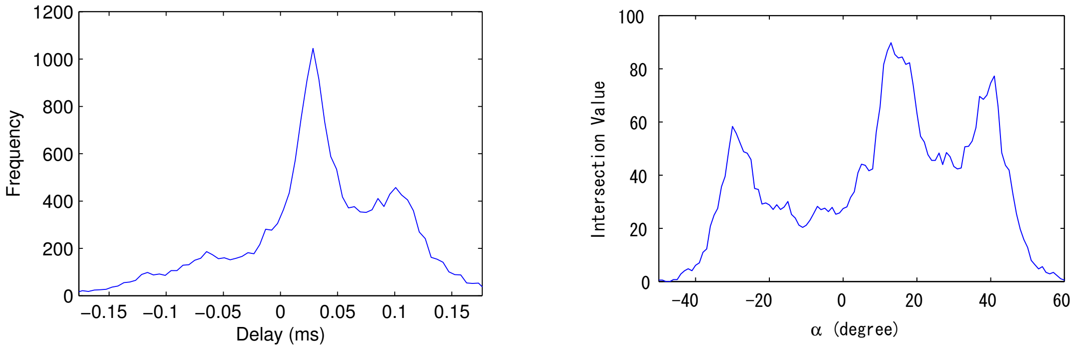
The true source direction is determined as follows. Because the conventional methods [22] [54] and the proposed method can estimate the DOA accurately in the case of a single source, only one source is made active in the experiment, and the source direction is estimated by the conversional and proposed methods. All methods gave the same rounded integer angles; thus, the obtained values were determined to be the true source directions.

4.7.1 Experimental results obtained using delay histogram and Hough transform with bandwidth

Table 4.1: Estimation results obtained using Hough transform

Case	Number of sources	Source	Sources direction	Conventional method	Proposed method
1	2	Male1 Female1	4° 14°	3° *Fail*	4° 14°
2	2	Female2 Male2	51° 70°	55° *Fail*	48° 67°
3	2	Male3 Male3	23° 42°	20° 40°	25° 40°
4	2	Female4 Female4	4° 59°	2° 64°	4° 63°
5	3	Female5 Male4 Female6	−34° 14° 59°	*Fail* 14° 59°	−34° 15° 58°
6	3	Male5 Male6 Male7	−14° 14° 34°	−11° 14° *Fail*	−14° 14° 30°

Tab. 4.1 shows the experimental results obtained using the Hough transform. Here, four typical cases are considered and compared with the conventional DUET [22]. In cases 1 and 2, two sources are closely located in different directions. The proposed method can estimate the directions of the sources but the conventional histogram-mapping method cannot. *Fail* in the table means that the method cannot identify the direction of the sources. In cases 3 and 4, both the proposed method and the conventional method can estimate the directions of the two sources, but the proposed method can estimate the DOA more accurately than the conventional method. Cases 5 and 6 are underdetermined situations, and the corresponding results are shown in Fig. 4.12 and Fig. 4.13. The results show that the proposed method is superior to the conventional method.



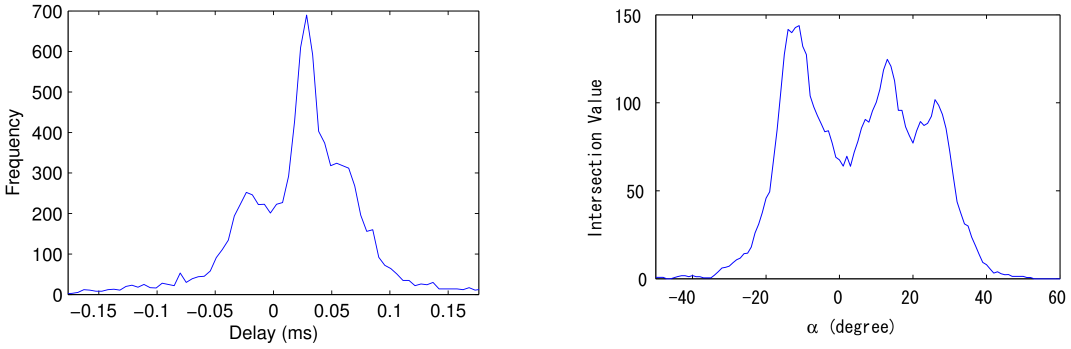
(a) By delay histogram.

(b) By Hough transform.

Figure 4.12: Result figures of case 5 in Table 4.1. The mixture condition is: a female source located at −34°, a male source located at 14°, and another female source located at 59°. The estimation results by delay histogram are *Fail*, 14° and 59°, while the estimation results by Hough transform are −34°, 15° and 58°.

Compared with the conventional method, there are two significant advantages of the proposed method. One is that the conventional method searches for points in the distribution coinciding with the theoretical value. From the viewpoint of the Hough transform, the conventional method is limited to the case when $\rho = 0$, which does not allow an error to exist. In the Hough transform, by using a bandwidth of ρ, the deviation is considered, and the analysis is more stable.

The other advantage is in the calculation of delay. Since the PD is divided by the frequency bin l, a small error in the PD will cause a large delay value in the low-frequency band, which affects the delay histogram. However, in the Hough transform, this problem does not arise.



(a) By delay histogram.

(b) By Hough transform.

Figure 4.13: Result figures of case 6 in Table 4.1. The mixture condition is: three male sources are located at −14°, 14° and 34° . The estimation results by delay histogram are −11°, 14° and *Fail*, while the estimation results by Hough transform are −14°, 14° and 30°.

4.8 Experimental results obtained using kernel density estimator

Two conventional methods are used for comparison. One is an independent component analysis (ICA)-based approach for the two-microphone case proposed by Nesta et al. [56]. The cumulative state coherence transform (SCT) generates a likelihood function whose local peaks correspond to the TDOA. The other is the method based on the *k*-means clustering algorithm proposed by Araki et al. [54], which is applicable to underdetermined DOA estimation for an arbitrary array configuration, but in this case, their algorithm is restrictively applied to a pair of microphones for the sake of comparison.

The case of two sources is first tested, then the underdetermined case, in which there are three sources, is considered as an extension. A noise-added case is also investigated.

4.8.1 Tuning parameters

Before presenting the experimental results, some remarks are given on the determination of the three tuning parameters in the proposed method.

(1) The threshold Th_1 in the preselection is used to eliminate the meaningless tiny power cells and to reduce the computation cost. Th_1 is determined from the relationship between the PD error and the amplitude $|X_m[k,l]|$ as shown in Fig. 4.6. Th_1 is set to 2.0. For values of Th_1 between 0 and 4.0 (0 means without preselection), little difference was observed in the final DOA estimation results.

(2) The reliable index threshold η_{th} is determined from the relationship between the reliability index η and the PD estimation error as shown in Fig. 4.7. The following decision process is used. In the experiments, the standard deviation of the PD estimation error is 0.1° as shown in Fig. 4.8. By including 0.1° or a similar value (0.08° in the present case) in the vertical axis of Fig. 4.7, the solid line in this figure yields the corresponding reliability index of approximately 0.96. This value is used as the threshold η_{th} in the paper.

(3) The control parameter $\hbar$ in the kernel density estimator is used to determine the fundamental bandwidth of the kernel. Various $\hbar$ were considered, as shown in Fig. 4.10, and $\hbar = 2$ is set in this paper. In the experiments, $\hbar$ had a very small effect on DOA estimation. For example, the obtained DOAs for $\hbar$ between 1.0 and 5.0 are almost identical. The effect of $\hbar$ on DOA estimation results is shown in Fig. 4.14. The effect of Th_1 is similar to that of $\hbar$.

As discussed above, the three tuning parameters are known before by performing preparatory experiments.

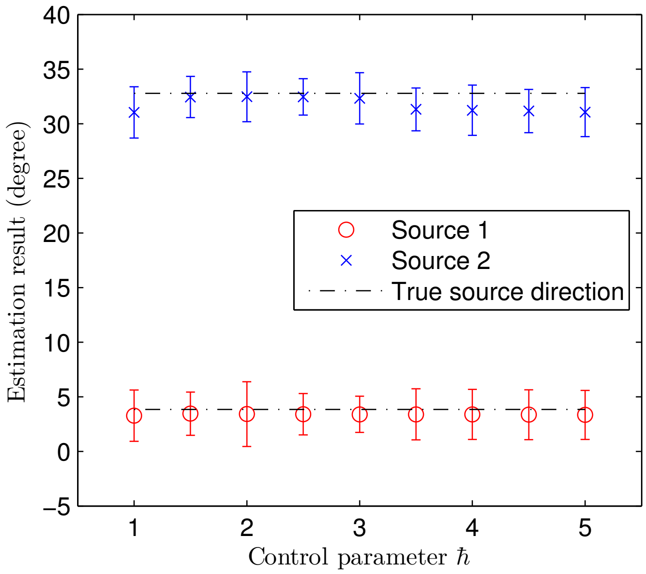


Figure 4.14: Effect of $\hbar$ on DOA estimation results

4.8.2 Two sources

In the case of two sources, the experiment compared two situations: (1) two sources and located at symmetrical positions with respect to the broadside of the microphone array, and

(2) two sources located on one direction side with one source placed at the broadside (near 0°) and the location of the other source varied from 20° to 60° at intervals of 10°. The results are shown in Fig. 4.15 and Fig. 4.16.

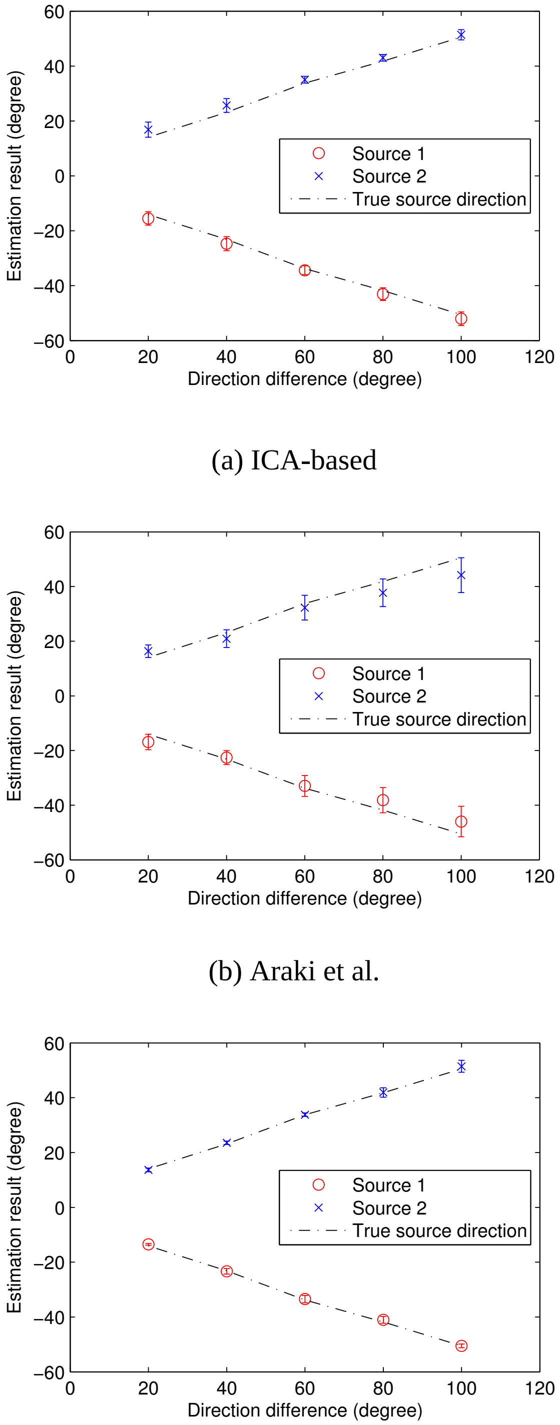
The ICA-based method for symmetric source positions always gives very accurate results as shown in Fig. 4.15(a). However, the results obtained by the ICA-based method for nonsymmetric source positions, as shown in Fig. 4.16(a), tend to be biased with relatively large deviations. In the ICA-based method proposed by Nesta et al., their DOA estimation results are mainly based on the SCT function with a sigmoid transform for a pair of normalized states of two sources. Unlike the Gaussian kernel in the proposed method, the SCT function is essentially a nonsymmetric function with respect to the TDOA except in the case of symmetric source positions. This explains why the resulting likelihood given by the accumulation of STC functions causes biased peak positions.

The method of Araki et al. based on the *k*-means algorithm is feasible because it avoids the peak search process, but the estimated results are slightly biased as shown in Fig. 4.15(b) and Fig. 4.16(b). This is caused by the hard clustering or the assignment of the *k*-means algorithm. In fact, when two sources are closely located, the hard clustering may eventually create a nonsymmetric data distribution. Thus, both centroids may be forced to move in outward directions. The opposite deviational behavior can be seen when the two sources are too far apart because the normalized feature values related to the DOA reach the boundary of their extent. This eventually causes the opposite outcome.

On the other hand, the proposed method gives a non biased estimation as shown in Fig. 4.15(c) and Fig. 4.16(c). It works well in both situations and outperforms the conventional methods. The combination of cell selection and the symmetric kernel function of θ is considered to result in accurate and nonbiased estimation.

4.8.3 Two sources with added diffuse noise

Additional experiments with two sources in a diffuse noise environment are performed to evaluate the robustness according to [63] [84]. In the diffuse noise, there is equal probability of energy flow in all directions. The noise appears to have no single source and is correlated between sensors. A two-channel diffuse noise is generated and added using the theoretical frequency-dependent covariance matrix to computer-generate directly propagating speech signals. The diffuse noise $\mathbf{N}[k,l] = (N_1[k,l],\ N_2[k,l])^T$ is assumed to be independent of the source signals with the correlation matrix

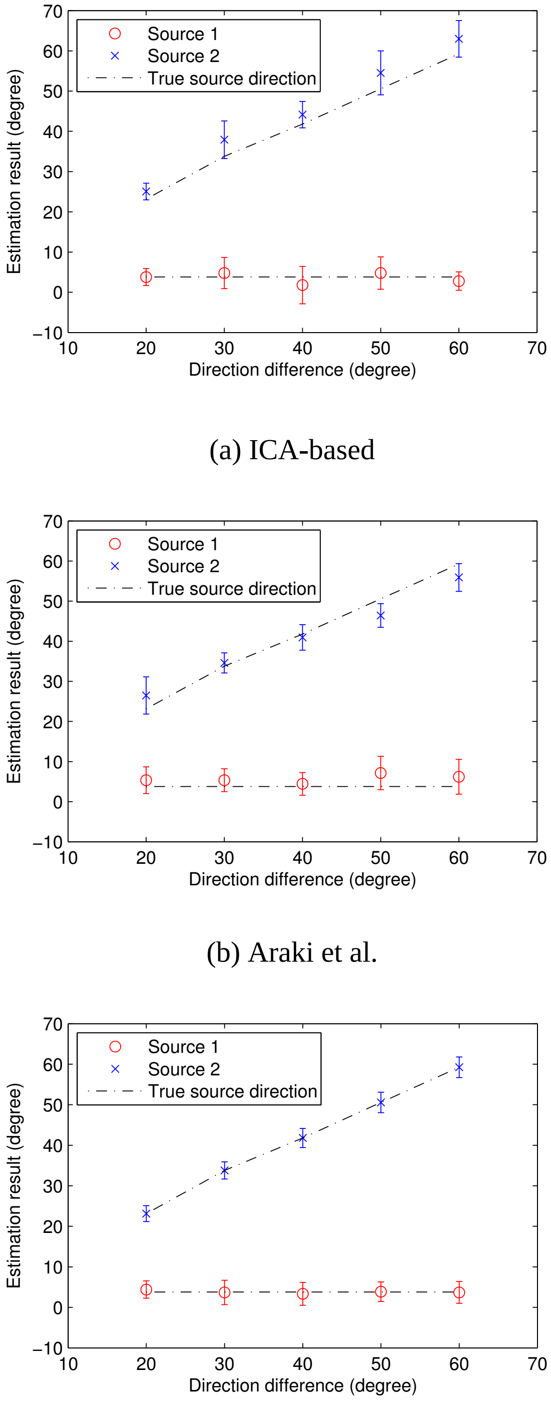


(a) ICA-based

(b) Araki et al.

(c) Proposed

Figure 4.15: DOA estimation results for two sources located at symmetrical positions. The horizontal axis is the source direction difference, and the vertical axis is the estimation result. The error bars show the standard deviation.



(a) ICA-based

(b) Araki et al.

(c) Proposed

Figure 4.16: DOA estimation results for two sources located on one direction side. The horizontal axis is the source direction difference, and the vertical axis is the estimation result. The error bars show the standard deviation.

$$\mathbf{V} = E[\mathbf{N}\mathbf{N}^H] = \sigma^2 \begin{pmatrix} 1 & \mathrm{sinc}(\Delta\omega T l) \\ \mathrm{sinc}(\Delta\omega T l) & 1 \end{pmatrix}, \tag{4.31}$$

where σ^2 is the power of the noise, $T = d/c$ is the maximum delay time between senors, and $\Delta\omega = 2\pi f_s/L$ is the unit frequency width in L-point STFT analysis. Matrix (4.31) is factorized to generate the diffuse noise in the frequency domain. The generation of diffuse noise is described in Appendix A. Source 1 is fixed at 0°, source 2 is varied from 20° to 60° at intervals of 20°, and noise with various signal-to-noise ratios (SNRs) is added.

The method of diffuse noise generation is verified and the results are shown in Fig. 4.17. The figure shows that the generation method is correct.

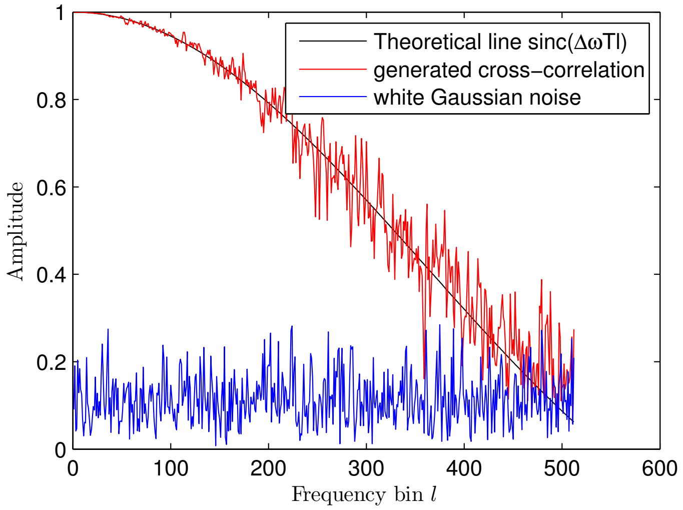


Figure 4.17: The verification of diffuse noise generation

The DOA estimation results using the estimation error $\theta_{error} = |\theta - \theta_{true}|$ are evaluated and its standard deviation σ_θ, where θ is the estimated source direction and θ_{true} is the true source direction. The results are shown in Tab. 4.2. Because the estimation results for source 1 do not vary significantly, the estimation results for source 2 are shown only. The estimation error θ_{error} is the mean value for 10 random mixtures.

The results show that the proposed method can estimate source directions stably and accurately even under a low-SNR condition (SNR=5 dB), while the conventional methods can only work when SNR=20 dB.

Table 4.2: DOA estimation results for two sources with added noise

(a) ICA-based

Direction of source 2	20		40		60	
(degree)	θ_{error}	σ_{θ}	θ_{error}	σ_{θ}	θ_{error}	σ_{θ}
SNR=20 dB	1.6	1.0	1.9	0.9	2.1	1.7
SNR=10 dB	11.7	8.9	2.7	1.8	2.6	2.5
SNR=5 dB	11.2	7.2	8.1	10.2	16.3	14.9

(b) Araki et al.

Direction of source 2	20		40		60	
(degree)	θ_{error}	σ_{θ}	θ_{error}	σ_{θ}	θ_{error}	σ_{θ}
SNR=20 dB	1.4	0.4	1.4	0.9	3.3	1.7
SNR=10 dB	1.0	0.6	6.6	1.7	15.3	3.6
SNR=5 dB	1.1	0.5	10.6	2.6	23.3	4.7

(c) Proposed

Direction of source 2	20		40		60	
(degree)	θ_{error}	σ_{θ}	θ_{error}	σ_{θ}	θ_{error}	σ_{θ}
SNR=20 dB	0.3	0.1	0.9	0.4	1.4	0.6
SNR=10 dB	0.5	0.3	0.9	0.4	1.8	1.1
SNR=5 dB	0.6	0.4	1.7	0.5	2.5	0.9

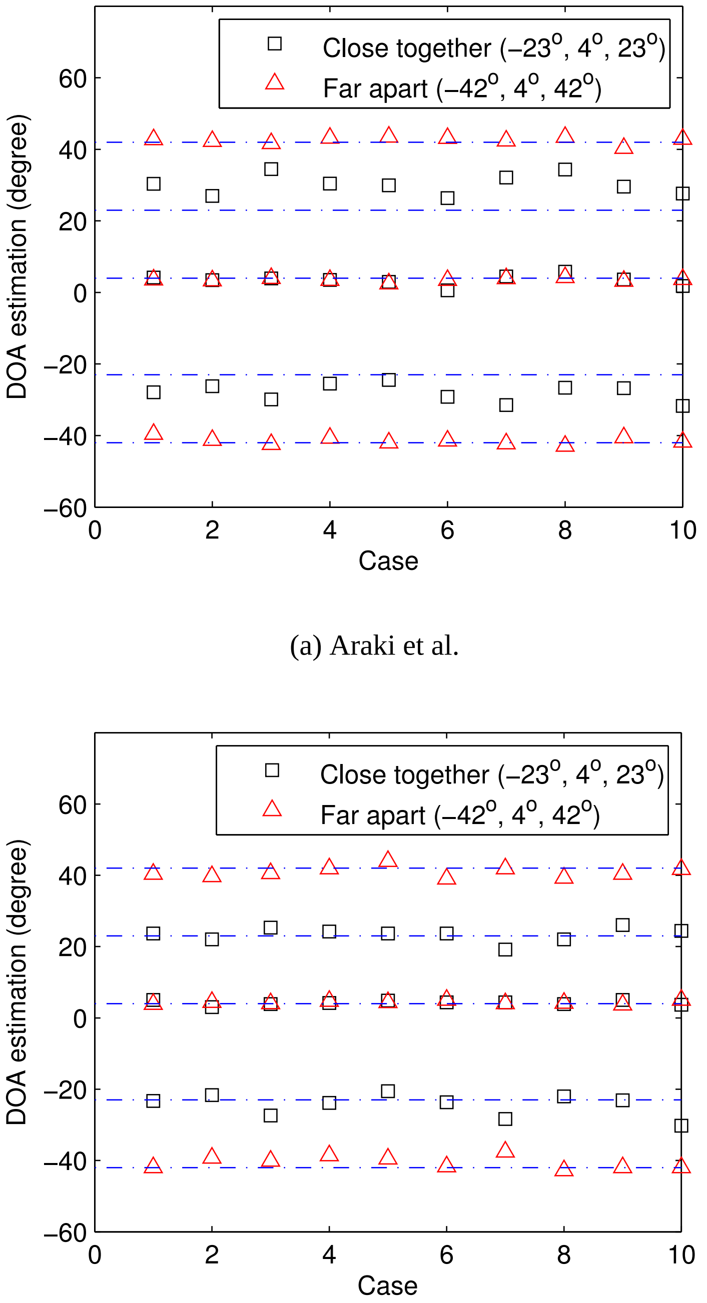
4.8.4 Three sources

Experiments for the underdetermined case of three sources were also performed. Three sources were set at close locations (−23°, 4°, and 23°) or further apart (−42°, 4°, and 42°). Fig. 4.18 shows the results. Since the ICA-based method cannot solve underdetermined situations theoretically, for the conventional methods, only the results for Araki et al.'s method are shown. When the sources are close together, $(\theta_{error},\ \sigma_\theta) = (7.1,\ 2.8)$ was obtained for the source at 23° by the conventional method, while $(\theta_{error},\ \sigma_\theta) = (1.5,\ 1.1)$ was obtained by the proposed method. When the sources were far apart, $(\theta_{error},\ \sigma_\theta) = (1.1,\ 0.5)$ was obtained for the source at 42° by the conventional method, while for the proposed method, $(\theta_{error},\ \sigma_\theta)$ was $(1.4,\ 1.0)$.

For widely spaced sources, both the conventional method and the proposed method can estimate the source directions very well. However, when three sources are closely located, the proposed method provides much more accurate and stable DOA estimation than the conventional method.

4.9 Summary

In this chapter, the author's research on DOA estimation was discussed. The DOA estimation problem was described in Sec. 4.2. Then two new methods for estimating the DOA were proposed: the Hough transform in Sec. 4.3 and the reliability index and kernel density estimator in Sec. 4.4. The experiments in Sec. 4.7 show that the improvement in direction accuracy and noise robustness are achieved by the proposed methods.



(a) Araki et al.

(b) Proposed

Figure 4.18: DOA estimation results for three sources. The horizontal axis ’Case’ refers to random three-source mixtures comprising the same sex or opposite sexes.

Chapter 5

General conclusion

In this chapter, a general review of the thesis is first given. Then some future areas of research are proposed.

5.1 Review of the thesis

This dissertation is a summary of the author's research on blind source separation (BSS) and estimating of direction of arrival (DOA) using a pair of microphones. The main purpose of the research was to obtain information on speech sources such as "Where is the speech source?" or "Can we obtain the desired speech from many simultaneous speeches?".

In the following, the proposed methods for solving the BSS and DOA estimation problems are reviewed, the reasons why they are successful are analyzed, and the difference between them are compared.

5.1.1 Speaker localization and source separation using PCA and harmonic structure

Many methods have been proposed for solving source direction estimation and source separation problems. However, in most of these methods, all of the data are treated in the time-frequency (T-F) domain, and the difference between time frames is not specified. In this study, the phase difference versus frequency (PD-F) distribution at each time frame is investigated, and the PD-F data of each frame are classified into three cases: a) non-source active (NSA), b) single-source active (SSA), and c) double-source active (DSA).

The advantages of using a time-frame PD-F distribution are as follows. 1) The PD-F data are located along a specific line through the origin and can thus be reliably estimated. Therefore, the PD-F graph directly illustrates the PD estimation error tendency and gives an intuitive insight into the dependence of the PD estimation error distribution on the frequency. 2) By observing the variance of the PD-F data at a specific frame, it can determine whether or not a single source is active at that time frame. To be more precise, principle component analysis (PCA) is applied to the two-dimensional PD-F data space at each time frame.

Since the ratio of the principal eigenvalues in PCA indicates the degree of data spread around the first principal axis, PCA is applied to detect SSA frames. Then, the detected SSA frames are used to estimate source directions.

To separate the DSA frames, the relationship between the harmonic structures of the initially separated source signals and the mixed signals in the low-frequency band is exploited. The separation process contains two substeps. The first step, which involves clustering, is performed in a high-frequency band, denoted by B_{high}, in which the PD estimation error is relatively small; therefore, the PD-F data are much more reliable. In this clustering procedure, the delay value is adopted as the cell's feature. For the T-F cells in the remaining low-frequency band, or the complement of B_{high}, denoted by B_{low}, the harmonic structure relationship between the initially separated spectrogram in B_{high} and the spectrogram in B_{low} is effectively used.

Experiments were performed to evaluate the proposed method by comparison with the conventional separation algorithm utilizing the delay feature. It was shown that the average improvement in the signal to interference ratio (SIR) obtained by the proposed method exceeded that obtained by the conventional method.

5.1.2 DOA estimation using kernel density estimator

A DOA estimation method for multiple speech sources from a stereophonic mixture in an underdetermined case was proposed, where the number of sources exceeds the number of sensors. The method relies on the sparseness of speech signals in the T-F domain representation. First, a set of T-F cells providing reliable spatial information is selected by using a newly proposed reliability index, which is defined as the estimated interaural phase difference at each T-F cell. Then, a statistical model for the propagation of the error between the phase difference at the T-F cell and its consequent DOA is introduced. By employing this model and the sparseness in the T-F domain, the DOA estimation problem is reduced to obtaining the local peaks of the probability density function of the DOA. Finally, a kernel density estimator approach based on the proposed statistical model is applied.

The performance of the method was assessed experimentally. The method outperforms other methods in terms of both its accuracy for real observed data and its robustness in the case of simulation with additional diffused noise.

5.1.3 Comparison between PCA, Hough transform, and kernel density estimator for DOA estimation

Three methods for DOA estimation are proposed in this thesis: by PCA, and using a Hough transform and a kernel density estimator.

The PCA method is different from the other two methods because it is applied to single time-frame data, to estimate the source direction. The PCA method can determine whether or not the time frame is SSA, while the other two methods cannot. On the other hand, the PCA method can only estimate one source direction, while the other two methods can estimate multiple source directions.

The common feature of the methods involving the Hough transform and kernel density estimator is that both of them consider the error obtained from a real environment. However, their details of application are different. To detect the lines in the PD-F distribution, the Hough transform applies the relative power for cell selection. The points in the PD-F are distributed in every frequency bin, rather than in only some frequency bins, which is helpful for obtaining the line direction by the Hough transform. Meanwhile, the kernel density estimator applies a novel reliability index for cell selection, and only the reliable cells are used for source direction estimation.

To obtain the error from a real environment, the Hough transform utilizes the bandwidth, where it is assumed that the error is distributed around the theoretical value, and the bandwidth is determined empirically. In the kernel density estimator a statistical model for the propagation of the error between the estimated phase difference and the consequent DOA is introduced. This model leads to a probability density function of the DOA, and DOA estimation is reduced to finding the most probable points.

As stated previously, starting from the time order, the method using the Hough transform was proposed, then the method using the kernel density estimator was proposed. Compared with the former method, cell selection using the reliability index and DOA error distribution model in the kernel density estimator is more novel and effective. This is why the kernel density estimator was mainly discussed, especially in the experimental part. Note that cell selection using the reliability index can also be applied in the Hough transform method.

However, some advantages of DOA estimation using the Hough transform should also be noted, for example, its outstanding performance in dealing with spatial aliasing. Although the Hough transform is a feature extraction technique used in image analysis, computer versions, and digital image processing, here it was utilized for speech signal processing. This can also be regarded as an attempt to combine sound and an image. Because sometimes, speech signal processing involves not only speech signal processing. Borrowing ideas from other fields may lead to a new approach.

5.2 Possible topics for future research

As discussed in Chapter 3, the proposed separation method basically operates at each time frame; thus, it has the potential to solve problems involving moving sources and real-time problems.

In this thesis, a pair of microphones are used as the sensors because this is the simplest array sensor system from the viewpoint of cost, and any complex array system can be regarded as a combination of numerous pairs of microphones. However, the estimation ability of a pair of microphones is limited. A pair of microphones acquires the sources from only half side of the array axis, because the signals from the symmetrical positions of the axis are the same as the sensors. Thus, how to extend the microphone configuration to deal with three-dimensional situations should be another interesting work, especially in the case of an arbitrary microphone array.

Another factor worth investigating is the distance between microphones in the array. In this thesis, the inter sensor distance is restricted to avoid spatial aliasing. However, increasing the distance will also bring some benefits, such as the attenuation ratio becoming clear. In this sense, consideration of the situation in which spatial aliasing occurs both in DOA estimation and source separation is another possible topic for future research.

Appendix A

Diffuse noise generation

In the following, the methods of generating diffuse noises $n_1(t)$ and $n_2(t)$ from independent white Gaussian noises $w_1(t)$ and $w_2(t)$ with zero means and unit variances is described.

In STFT domain, let denote the white Gaussian noises in the time-frequency(T-F) domain by $W_1[k,l]$ and $W_2[k,l]$, and the diffuse noises are $N_1[k,l]$ $N_2[k,l]$. The next formulation is derived

$$\mathbf{A}(l)\mathbf{W} = \mathbf{N}, \tag{A-1}$$

where

$$\mathbf{A}(l) := \begin{bmatrix} A_{11}(l) & A_{12}(l) \\ A_{21}(l) & A_{22}(l) \end{bmatrix} \tag{A-2}$$

$$\mathbf{W} := \begin{bmatrix} W_1[k,l] \\ W_2[k,l] \end{bmatrix} \tag{A-3}$$

$$\mathbf{N} := \begin{bmatrix} N_1[k,l] \\ N_2[k,l] \end{bmatrix}. \tag{A-4}$$

Next, it is necessary to determine the matrix $\mathbf{A}(l)$.

The correlation matrix of white Gaussian noise is

$$\mathbf{V}_w[k,l] = E[\mathbf{W}\mathbf{W}^H] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{A-5}$$

where H is the Hermitian operator. The correlation matrix of diffuse noise is

$$\mathbf{V}_n[k,l] = E[\mathbf{N}\mathbf{N}^H] = \sigma^2 \begin{bmatrix} 1 & R(l) \\ R(l) & 1 \end{bmatrix} \tag{A-6}$$

$$= E[\mathbf{A}(l)\mathbf{W}\mathbf{W}^H\mathbf{A}^H(l)]$$
$$= \mathbf{A}(l)E[\mathbf{W}\mathbf{W}^H]\mathbf{A}^H(l)$$
$$= \mathbf{A}(l)\mathbf{A}^H(l),$$

where

$$R(l) = \mathrm{sinc}(\Delta\omega T l) = \frac{\sin(\Delta\omega T l)}{\Delta\omega T l}. \tag{A-7}$$

Let use the eigenvalue and eigenvector method to solve the following matrix factorization.

$$\mathbf{V}_n(l) = \sigma^2 \begin{bmatrix} 1 & R(l) \\ R(l) & 1 \end{bmatrix} = \mathbf{A}(l)\mathbf{A}^H(l) \tag{A-8}$$

Let us write

$$\mathbf{C}(l) = \begin{bmatrix} 1 & R(l) \\ R(l) & 1 \end{bmatrix}. \tag{A-9}$$

Assume that λ_1 and λ_2 are the eigenvalues of $\mathbf{C}$ and that $\mathbf{b}_1$ and $\mathbf{b}_2$ are the corresponding eigenvectors.

$$\mathbf{C}(l)\mathbf{b}_1 = \lambda_1\mathbf{b}_1 \tag{A-10}$$

$$\mathbf{C}(l)\mathbf{b}_2 = \lambda_2\mathbf{b}_2 \tag{A-11}$$

$$\mathbf{C}(l) = \begin{bmatrix}\mathbf{b}_1\mathbf{b}_2\end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \mathbf{b}_1^T \\ \mathbf{b}_2^T \end{bmatrix} \tag{A-12}$$

$$\mathbf{V}_n(l) = \sigma^2\mathbf{C}(l) = \sigma\begin{bmatrix}\mathbf{b}_1\mathbf{b}_2\end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} \begin{bmatrix} \mathbf{b}_1^T \\ \mathbf{b}_2^T \end{bmatrix} \sigma = \mathbf{A}(l)\mathbf{A}^T(l) \tag{A-13}$$

This can derive the matrix $\mathbf{A}(l)$

$$\mathbf{A}(l) = \sigma\begin{bmatrix} \sqrt{\lambda_1}\mathbf{b}_1, & \sqrt{\lambda_2}\mathbf{b}_2 \end{bmatrix}, \tag{A-14}$$

and using Eq. (A-1), $\mathbf{N}$ can be obtained. Finally, by performing an ISTFT, the diffuse noises $n_1(t)$ and $n_2(t)$ are obtained.

Bibliography

[1] http://www.softdistrict.com/wp-content/uploads, Mar. 2011.

[2] http://www.carsanjay.com/wp-content/uploads, Jun. 2011.

[3] B. Porat, A Course in Digital Signal Processing, Wiley, 1997.

[4] R.G. Lyons, Understanding Digital Signal Processing, Prentice Hall, 2004.

[5] J.Y. Stein, A Computer Science Perspective, Wiley, 2000.

[6] J.G. Proakis and D.G. Manolakis, Digital Signal Processing- Principles, Algorithms and Applications, Prentice-Hall, 1996.

[7] K.H. Davis, R. Biddulph, and S. Balashek, "Automatic speech recognition of spoken digits," Acoustical Society of America, vol.24, pp.637–642, 2002.

[8] J. Allen, M.S. Hunnicutt, and D. Klatt, From Text to Speech: The MITalk system, Cambridge University Press, 1987.

[9] H. Dillon, Hearing aids, Thieme, 2001.

[10] B.C. Moore, Cochlear hearing loss (2nd ed.), Wiley, 2007.

[11] S. Furui, Digital Speech Processing, Synthesis and Recognition, Marcel Dekker Inc., 2001.

[12] O. Castro-Orgaz and H. Chanson, "Bernoulli theorem, minimum specific energy and water wave celerity in open channel flow," Journal of Irrigation and Drainage Engineering, vol.135, pp.773–778, 2009.

[13] S. Saito, K. Kato, and N. Teranishi, "Statistical properties fo fundamental frequencies of japanese speech voices," Journal of Acoustic Society of Japan, vol.14, pp.111–116, 1958.

[14] K. Bochner, S. & Chandrasekharan, Fourier Transforms, Princeton University Press, 1949.

[15] A.V. Oppenheim, R.W. Schafer, and J.R. Buck, Discrete-Time Signal Processing, Prentice Hall, 1998.

[16] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.32, no.2, pp.236–243, 1984.

[17] X. Huang, A. Acero, and H. Hon, Spoken language processing, Prentice Hall, 2001.

[18] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," IEEE Signal Processing Magazine, vol.13, no.4, pp.67–94, 1996.

[19] J. Meyer and G. Elko, “A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield,” Proc. IEEE Int Acoustics, Speech, and Signal Processing (ICASSP) Conf, 2002.

[20] D. Van Compernolle, “Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings,” Proc. Int Acoustics, Speech, and Signal Processing ICASSP-90. Conf, pp.833–836, 1990.

[21] S. Makino, T.W. Lee, and H. Sawada, eds., Blind speech separation, Springer, 2007.

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

[23] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, “Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones,” Acoust. Sci. & Tech, vol.22, pp.149–157, 2001.

[24] S. Araki, H. Sawada, R. Mukai, and S. Makino, “Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors,” Signal Processing, vol.87, pp.1833–1847, 2007.

[25] H. Sawada, S. Araki, R. Mukai, and S. Makino, “Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation,” IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.5, pp.1592–1604, 2007.

[26] M.D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M.E. Davies, “Sparse representations in audio and music: From coding to source separation,” Proceedings of the IEEE, vol.98, no.6, pp.995–1005, 2010.

[27] J.M. Peterson and S. Kadambe, “A probabilistic approach for blind source separation of underdetermined convolutive mixtures,” Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP ’03), 2003.

[28] F. Abrard and Y. Deville, “A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources,” Signal Processing, vol.85, pp.1389–1403, 2005.

[29] S. Arberet, R. Gribonval, and F. Bimbot, “A robust method to count and locate audio sources in a multichannel underdetermined mixture,” IEEE Transactions on Signal Processing, vol.58, no.1, pp.121–133, 2010.

[30] T.W. Lee, Independent Component Analysis - Theory and applications, Kluwer Academic Publishers, 1998.

[31] S. Haykin, ed., Unsupervised Adaptive Filtering, John Wiley & Sons, Inc., 2000.

[32] A. Hyvarinen, J. Karhumen, and E. Oja, Independent Component Analysis, John Wiley & Sons, Inc., 2001.

[33] A. Cichocki and S. Amari, Adaptive Blind Signal and Image Processing, John Wiley & Sons, Inc., 2002.

[34] K. Matsuoka, M. Ohya, and M. Kawamoto, “A neural net for blind separation of nonstationary signals,” Neural Networks, vol.8, pp.411–419, 1995.

[35] S. Choi and A. Cichocki, “Blind separation of nonstationary sources in noisy mixtures,” Electronics Letters, vol.36, pp.848–849, 2000.

[36] D.T. Pham and J.F. Cardoso, “Blind separation of instantaneous mixtures of nonstationary sources,” IEEE Transactions on Signal Processing, vol.49, no.9, pp.1837–1848, 2001.

[37] S. Choi and A. Cichocki, “Blind separation of nonstationary and temporally correlated sources from noisy mixtures,” IEEE Workshop on Neural Networks for Signal Processing, NNSP’2000, 2000.

[38] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” Neurocomputing, vol.22, pp.21–34, 1998.

[39] S. Amari and A. Cichocki, “Adaptive blind signal processing-neural network approaches,” Proceedings of the IEEE, vol.86, no.10, pp.2026–2048, 1998.

[40] H. Saruwatari, S. Kurita, and K. Takeda, “Blind source separation combining frequency-domain ica and beamforming,” Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP ’01), pp.2733–2736, 2001.

[41] M.Z. Ikram and D.R. Morgan, “A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation,” Proc. IEEE Int Acoustics, Speech, and Signal Processing (ICASSP) Conf, 2002.

[42] H. Sawada, R. Mukai, S. Araki, and S. Makino, “A robust and precise method for solving the permutation problem of frequency-domain blind source separation,” IEEE Transactions on Speech and Audio Processing, vol.12, no.5, pp.530–538, 2004.

[43] N. Mitianoudis and M.E. Davies, “Audio source separation of convolutive mixtures,” IEEE Transactions on Speech and Audio Processing, vol.11, no.5, pp.489–497, 2003.

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

[45] D.L. Wang, “Time-frequency masking for speech separation and its potential for hearing aid design,” Trends in amplification, vol.12, no.4, pp.332–353, 2008.

[46] J. Huang, N. Ohnishi, and N. Sugie, “A biomimetic system for localization and separation of multiple sound sources,” IEEE Trans. on Instrumentation and Measurement, vol.44, pp.733–738, 1995.

[47] M. Yoshida, D. Ning, and N. Hamada, “Blind speech separation by integrating three pairs of phase differences of equilateral triangular microphone array,” Proc. Int Intelligent Signal Processing and Communication Systems (ISPACS) Symp, pp.1–4, 2010.

[48] M. Sekigawa, M. Yoshida, N. Ding, and N. Hamada, “Speech separation and localization by exploiting constraints on delay-time vector introduced by array configuration,” NCSP’11, pp.239–242, 2011.

[49] S. Rickard, R. Balan, and J. Rosca, “Real-time time-frequency based blind source separation,” ICA, pp.651–656, 2001.

[50] J. Benesty, J. Chen, and Y. Huang, Microphone Array Signal Processing, Springer, 2008.

[51] E.D.D. Claudio and R. Parisi, Microphone Arrays, ch. Multi-Source Localization Strategies, pp.181–201, Springer-Verlag, 2001.

[52] C.H. Knapp and G.C. Carter, “The generalized correlation method for estimation of time delays,” IEEE Trans. on Acoust. Speech Signal Process., vol.ASSP-24, pp.320–327, 1976.

[53] R.O. Schmidt, “Multiple emitter location and signal parameter estimation,” IEEE Trans. on Antennas and Propagation, vol.34, pp.276–280, 1986.

[54] S. Araki, H. Sawada, R. Mukai, and S. Makino, “DOA estimation for multiple sparce sources with arbitrarily arranged multiple sensors,” Journal of Signal Processing Systems, vol.63, pp.265–275, 2009.

[55] H. Sawada, R. Mukai, and S. Makino, “Direction of arrival estimation for multiple source signals using independent component analysis,” Proc. Seventh Int Signal Processing and Its Applications Symp, pp.411–414, 2003.

[56] F. Nesta, P. Svaizer, and M. Omologo, “Cumulative state coherence transform for a robust two-channel multiple source localization,” Proc. ICA, pp.290–297, 2009.

[57] M. Omologo and P. Svaizer, “Use of the cross power-spectrum phase in acoustic event location,” IEEE Trans. on Audio, Speech, and Language Processing, vol.5, pp.288–292, 1997.

[58] M.S. Brandstein, “Time-delay estimation of reverberated speech exploiting harmonic structure,” Journal of the Acoustic Society of America, vol.105, pp.2914–2919, 1999.

[59] J. Capon, R.J. Greenfield, and R.J. Kolker, “Multidimensional maximum-likelihood processing of a large aperture seismic array,” Proceedings of the IEEE, vol.55, no.2, pp.192–211, 1967.

[60] J. Burg, “The relationship between maximum entropy spectra and maximum likelihood spectra,” Geophysics, vol.37, pp.375–376, 1972.

[61] Hayes and H. Monson, Statistical Digital Signal Processing and Modeling, John Wiley & Sons, Inc., 1996.

[62] H. Wang and M. Kaveh, “Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources,” IEEE Trans. on Acoust. Speech Signal Process, vol.33, pp.823–831, 1985.

[63] Y. Izumi, N. Ono, and S. Sagayama, “Sparseness-based 2CH BSS using the EM algorithm in reverberant environment,” Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics, pp.147–150, 2007.

[64] N. Ding, M. Yoshida, J. Ono, and N. Hamada, “Speech mixture separation and doa estimation utilizing sequence of phase difference distribution,” NCSP 10, pp.341–344, 2010.

[65] N. Ding and N. Hamada, “Speaker localization and speech separation using phase difference versus frequency distribution,” EUSIPCO 2011, pp.250–253, 2011.

[66] N. Ding, M. Yoshida, J. Ono, and N. Hamada, “Blind source separation using sequential phase difference versus frequency distribution,” Journal of Signal Processing, vol.15, no.5, pp.375–385, 2011.

[67] W. Kasprzak, N. Ding, and N. Hamada, “Relaxing the wdo assumption in blind extraction of speakers from speech mixtures,” Journal of Telecommunications and Information Technology, vol.4, pp.50–58, 2010.

[68] N. Ding, T. Shimada, M. Yoshida, W. Kasprzak, and N. Hamada, “A consideration on time-frequency masking methods for speech separation,” 24th Signal Processing Symposium, 2009.

[69] W. Kasprzak, N. Ding, and N. Hamada, “Blind localization and separation of two speakers based on two mixtures,” Proceedings of 2nd IEEE Workshop on Bio-Inspired Signal and Image Processing (BISIP), 2010.

[70] W. Kasprzak, N. Ding, and N. Hamada, “Speaker localization and speech separation in two echoic mixtures,” Science-Future of Lithuania, vol.3, no.1, pp.43–49, 2011.

[71] T. Nakatani, M. Goto, and H.G. Okuno, “Localization by harmonic structure and its application to harmonic sound stream segregation,” Proc. Conf. IEEE Int Acoustics, Speech, and Signal Processing ICASSP-96, pp.653–656, 1996.

[72] T.W. Parsons, “Separation of speech from interfering speech by means of harmonic selection,” Journal of the Acoustic Society of America, vol.60, pp.209–222, 1976.

[73] B. Berdugo, J. Rosenhouse, and H. Azhari, “Speakers direction finding using estimated time delays in the frequency domain,” Signal Processing, vol.82, pp.19–30, 2002.

[74] N. Ding and N. Hamada, “DOA estimation by Hough transform in Phase Difference distribution versus Frequency,” Proc. Int Intelligent Signal Processing and Communication Systems (ISPACS) Symp, pp.1–4, 2010.

[75] J. Ono, N. Ding, and N. Hamada, “Sound source localization using phase-difference vs frequency plots,” APSIPA ASC 2010, p.46, 2010.

[76] N. Ding and N. Hamada, “DOA estimation in underdetermined stereophonic mixture by kernel density estimator,” Proc. Int Intelligent Signal Processing and Communication Systems (ISPACS) Symp, 2011.

[77] N. Ding, K. Fujimoto, and N. Hamada, “Kernel density estimator approach for solving underdetermined source localization problem from arbitrary microphone configuration,” 26th Signal Processing Symposium, 2011.

[78] N. Ding and N. Hamada, “DOA estimation of multiple speech sources from a stereophonic mixture in underdetermined case,” IEICE Trans. on Fundamentals, 2012. (accepted for publication).

[79] K. Suzuki, T. Koga, J. Hirokawa, H. Ogawa, and N. Matsuhira, “Clustering of sound-source signals using hough transformation, and application to omni-directional acoustic sense for robots,” 22th Artificial Intelligence Challenge, pp.53–58, 2005. (in Japanese).

[80] S. Marchand and A. Vialard, “The hough transform for binaural source localization,” Proc. of the 12th Int. Conference on Digital Audio Effects (DAFx-09), 2009.

[81] J.E. Rubio, K. Ishizuka, H. Sawada, S. Araki, T. Nakatani, and M. Fujimoto, “Two-microphone voice activity detection based on the homogeneity of the direction of arrival estimates,” Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing ICASSP 2007, pp.385–388, 2007.

[82] C.M. Bishop, Pattern recognition and machine learning, Springer, 2006.

[83] R. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, John Wiley & Sons, 2001.

[84] N.Q.K. Duong, E. Vincent, and R. Gribonval, “Spatial covariance models for under-determined reverberant audio source separation,” Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics WASPAA ’09, pp.129–132, 2009.