A Research on Voice Activity Detection and Speaker Direction Estimation Using Microphone Array

March 2005

Yusuke Hioka

Preface

Due to the recent progress of digital processors and the rapid spread of broadband data transmitting channels, the application of digital signal processing to multimedia signals is now part of our daily lives. Moreover, the utilization of speech for human-machine interface is in the spotlight, because it simplifies the complicated procedures in the operation of systems. In various applications using speech, the information about the speaker is useful. In particular, the voice activity detection (VAD) identifies "When the speaker utters?" and the speaker direction estimation identifies "Where the speaker's location is?". Both of these are often used as preliminary data. This dissertation describes the research results achieved using specific methods to acquire this data using microphone array signal processing.

1. **Voice activity detection using speech features in multiple signal domains (Chapter 3)**
   Conventional VAD methods suffer from the performance degradation due to nonstationary interference that often occurs in practical environments. Added to this, the discrimination capability of VAD is reduced when the speech and interference arise from close directions. This research proposes a VAD method achieved by array signal processing in the wavelet domain. Since the method used the speech signal features in temporal, spectral, and spatial domains, it succeeded correctly in discriminating the segments of speech from interference correctly.

2. **Speaker direction estimation with uniform accuracy for omni-direction (Chapter 4)**
   The methods proposed for speaker direction estimation provided uniform accuracy for omni-direction. This method employed equilateral-triangular

microphone array and achieved uniform estimation accuracy for omni-direction. This was achieved through the integrated use of the data extracted from microphone pairs on each of the three sides. The estimation accuracy was successfully improved by selecting the harmonics of voiced sound consisting of major speech components. In addition, a method is proposed to estimate both azimuth and elevation angles using tetrahedral microphone array.

3. **Tracking of moving speaker direction (Chapter 5)**
   The proposed speaker direction tracking method is a refinement of 2. above. In this method, the performance index was changed during the adaptation to achieve fast and accurate global convergence to cope with abrupt movement of speaker direction.

In each situation, the effectiveness of the methods employed was verified through experiments in the real acoustic environment.

Yusuke Hioka

Acknowledgement

This dissertation is a summary of studies for six years carried out at Graduate School of Integrated Design Engineering, Keio University, Japan. I would like to express sincere thanks to my advisor, Dr. Nozomu Hamada, Professor of Department of System Design Engineering, Keio University, for his valuable guidance and continuous encouragement. Without his support, this research would not have been completed. I would also like to express my thanks to the members of thesis review committee, Dr. Masaaki Ikehara, Dr. Shinji Ozawa, and Dr. Akira Sano, for their advices and comments.

I am grateful to all members in Hamada Laboratory for their assistances and supports in my research. Especially, I wish to acknowledge to Dr. Tateo Yamaoka, Dr. Toshihiko Fukue, Mr. Masahiro Furukawa, Ms. Yoko Koizumi, Mr. Takuro Ema, and Mr. Masao Matsuo and other members in "Array Signal Processing Group". Their assistance, comments, and encouragement helped me a lot to complete this dissertation.

I would also like to express my gratitude to Mrs. & Mr. Balfour. Their help for checking and correcting English improved this dissertation much more.

I also appreciate to my sister Masako, and to all of my friends especially Ms. Nanase Nishimura for her constant encouragement. Finally, I would like to thank my parents, Tatsuki and Kumiko for their supports and patients in daily life and the financial aid to carry on the study.

Yusuke Hioka

February 2005

Title of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Backgrounds and purpose of the research

1.1.1 Global backgrounds

Digital signal processing [1] is one of the most important fields in science and technology. It is widely applied to areas, such as communication, acoustics, seismology, biology, and so on. In the last decade, digital processor has progressed amazingly in its performance even though its production cost has been remarkably reduced, and now various advanced and complicated functions can be easily installed even in products of personal use around us. Furthermore, the broadband data transmitting channels such as FTTH (Fibre To The Home) and ADSL (Asymmetric Digital Subscriber Line) have prevailed not only in offices but also in our houses.

With such progress, the application of digital signal processing to multimedia signals is now being expected to become reality. Speech signal processing has become a very important area of research, because in the future, speech will be the main interface between the human and machines.

Fig. 1.1 summarises the research area that deals with speech. The major topics involved in speech processing are the analysis, recognition, coding, and synthesis [2]. The researches for the recognition [2] have been carried out since the 1950s, and it has now reached the level of practical products, such as ViaVoice® [3]. On the other hand, the history of speech synthesis [2] is much older; there was an

attempt to achieve this, in the 18th century, using a mechanical device. After the development of electrical engineering in the 20th century, it made much progress and it is now adopted in several applications. The technology of speech coding was developed along with communication technologies, and it is now still continuing to progress.

Besides these, speech analysis itself is not directly applied to practical applications, but it plays a role in helping other speech processing techniques by extracting information that is inherent to an individual speaker. For example, the spectral envelope of a given speech signal is indispensable information for speech recognition and synthesis, and the voice activity segment is valuable for speech coding and recognition. Thus, extracting speaker information by speech analysis is an important factor in realising various speech processing applications. Within the information of the speaker, the temporal and spatial information, i.e. **"When the speaker utters?"** and **"Where the speaker's location is?"** respectively, is often required in various applications. In order to derive each of these features, the following schemes are available.

- Voice activity detection (VAD)
  Generally speaking, the speech signal component appears at particular time segments, and the rest of the segments are usually silent. VAD is the technology to discriminate the segments occupied by the speech components.

- Speaker direction estimation
  Speaker direction estimation is the major means of acquiring the speaker's location. Because the relative position between the speaker and the microphones is different in each situation, locating and tracking the speaker's direction is required.

On the other hand, the acoustical processing [4] is another important area firmly connected with the speech processing technologies. For this reason, the performance of speech processing is generally affected by the surrounding acoustic environment. For example, the received speech should be clear enough to be utilized for its recognition, and proper sound reproduction technique is necessary for playing the synthesized speech. The acoustical processing consists of the recording system and the reproduction system. In the recording process,

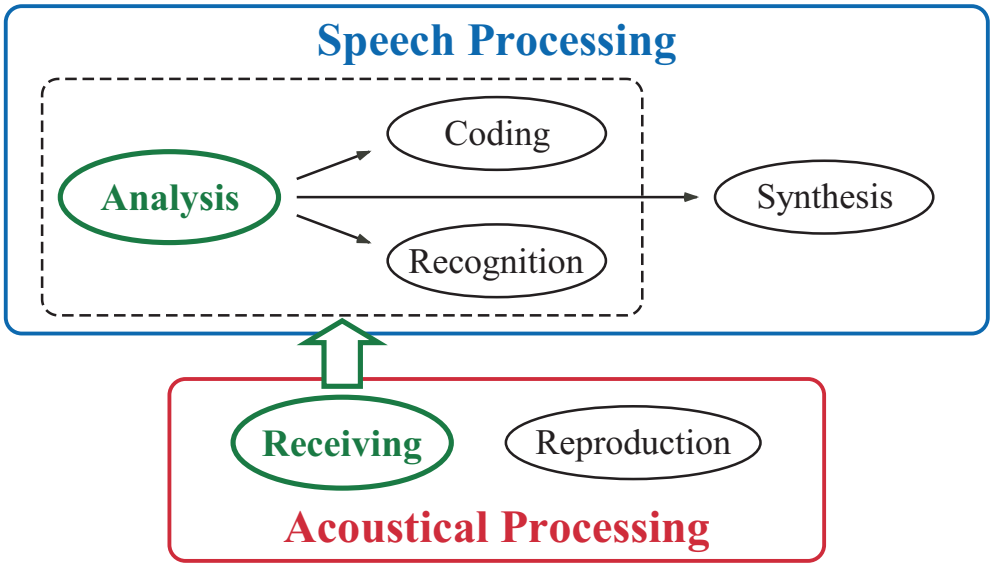

Figure 1.1: Position of our research target

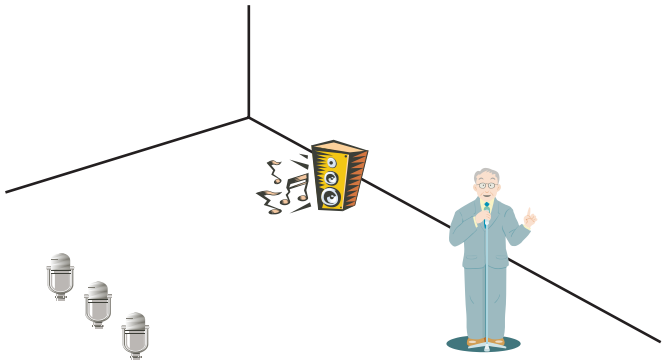

Figure 1.2: Problem of sound recording

a problem of receiving sounds emanating from spatially distant sources is described in Fig. 1.2. The sound recording is a helpful device for subsequent speech processing, especially for speech analysis. Microphone array [4][5] is one of the effective sound recording systems. It is able to handle the spatial information such as sound source location and its movement, and so it is appropriate for the acoustical processing.

The research subjects involved in this dissertation are speech analysis, especially voice activity detection and speaker direction estimation, using a micro-

phone array. In the following, the backgrounds to VAD and speaker direction estimation are given, with an introduction to related applications. The purpose of this research is also stated below.

1.1.2 Context of VAD and speaker direction estimation

First of all, we introduce two research topics, speech recognition and enhancement, that are closely connected with the VAD and speaker direction estimation. After that, some practical applications are referred to which are related to those technologies.

Speech recognition

Above all, speech recognition is a core technology to achieve the speech based human-machine interface. Although it has already reached the level of practical use, the recognition rate is severely degraded in environments where various interfering sounds exist. To solve this problem, the speech signal extraction and enhancement are available.

For extracting desired speech signal, VAD can be utilized. Because the silent segments in speech signal contain only interferences, they might be harmful to recognition performance. Thus, the non-speech segments are usually avoided by VAD to improve the recognition rate.

Speech enhancement

Even though the estimated voice activity segments contain the desired speech components, they might be contaminated by interferences. In this case, the application of speech enhancement is required to retrieve the quality of speech signal. As summarised in Table 1.1, many conventional techniques of speech enhancement have been proposed, most of which rely on speaker information given initially by either the VAD or the speaker direction estimation. In this table, the conventional speech enhancement methods are classified into two groups by the number of required microphones. The major difference between the monaural and multichannel speech enhancement methods is that the latter enables the use

Table 1.1: Classification of speech enhancement method

Number of Microphones	Method	*A priori* Information	
		Prerequisite	Subsidiary
Single =Monaural	Spectrum Subtraction		VAD
	Adaptive processing		VAD
Multiple =Microphone Array	Fixed Beamformer	speaker direction	
	Adaptive Beamformer	speaker direction	VAD
	Subspace		VAD
	ICA		speaker direction

of spatial information of the sound source. A brief introduction to these methods with the relation to VAD and speaker direction estimation is given below.

The methods using a single microphone mainly rely on the temporal and spectral features in speech signal. The spectral subtraction [6] is one of the most important speech enhancement methods applicable to the single microphone case. In this method, the spectrum of interference, which is known in advance, is subtracted from the input signal. To calculate the power spectrum of the interferences, estimated non-speech segments are useful because they do not contain the speech components that should not be eliminated. On the other hand, the adaptive signal processing is also adopted for monaural speech enhancement subject. Some studies use the information about voice activity segments to modify the adaptation rule depending on the existence of speech components [7].

In contrast to the monaural case, a microphone array makes it possible to utilize the spatial information about the speaker. One major function of array signal processing is the spatial filtering called beamforming [8]. The fixed beamforming, typified by the delay-and-sum beamformer [4], is able to emphasize or suppress sounds arriving from specific directions. The procedure in determining the property of beamformer requires the speaker direction to which the gain or null of its spatial response is pointed.

On the other hand, adaptive beamformer alters its response automatically, but usually, the direction of desired signal is given *a priori*. For example, a pop-

ular adaptive beamformer, LCMV [9], adjusts its spatial response by minimizing the power of the output signal with the constraint that the transfer function for the desired speaker direction is restricted to be unity. Hoshuyama *et al.* adopted this strategy for their speech enhancement system using a microphone array [10][11][12][13]. Furthermore, the VAD is known to be helpful to control the adaptive process [14][15][16].

Besides these beamforming techniques, there are some speech enhancement methods that are free of prerequisite speaker direction information. Asano *et al.* proposed an automatic beamforming method based on the subspace analysis of array covariance matrix [17]. Because the method determines the beamformer weights by the principal component analysis (PCA), it requires only the number of speakers that corresponds to the number of selected principal components. Because the desired principal components explicitly appear only at the voice activity segments, it is important that the subspace method is introduced into the VAD. Furthermore, the blind source separation (BSS) using independent component analysis (ICA) has been well-discussed recently [18][19][20][21]. This method separates signals with very weak assumptions in advance that the source signals are statistically independent of each other. Although BSS is able to separate the mixed signals under the "blind" condition, its performance is improved by providing the speaker direction as initial settings of the learning process [20][21].

In the following, we show some more practical applications are shown which relate to VAD and speaker direction estimation.

Hands-free System

The Hands-free system is one leading interface that uses speech. It releases us from the nuisances of wearing the headset or the microphone. But despite its obvious advantages, the following problems should be resolved in order to achieve an effective outcome.

1. Speech enhancement and noise reduction
   Usually, the speech signal received by the hands-free devices are severely contaminated by several kinds of ambient interfering noises due to the low

input SNR caused by a large distance between the speaker and the microphone. To improve the quality of speech, speech enhancement is necessary. As mentioned above, the VAD and speaker direction estimation help to improve the performance of speech enhancement schemes.

2. Acoustic echo canceller
   Another problem that should be considered in realising the hands-free system is the acoustic echo caused by the acoustic coupling of loudspeaker and microphone. In the case of adopting the hands-free system for a telephone receiver for example, the microphone receives not only the speaker's voice but also the sounds emitted from the loudspeaker. As this phenomenon annoys the speaker on the receiving end of the call, such echo components should be eliminated. Various studies of the acoustic echo canceller have already been carried out. Furthermore, some have tried to insert the microphone array in order to aim at the combination of noise suppression and acoustic echo cancelling [22]. Thus, the information about speaker direction is also helpful for the system.

Video conference system

Another major application that requires speech analysis is the video conference system. In putting this into effect, the following key technologies are required.

1. Speech enhancement
   The video conference system often utilizes the hands-free system as the sound recording devices. Hence, the speech enhancement processing is essential to the received signal.

2. Speech coding and compression
   Usually, speech coding accompanies the function of data compression. In some speech coding procedures, the VAD is adopted for achieving high compression rate. For example, the G.723.1 vocoder recommended as ITU-T standard includes both the silence compression scheme and the VAD [23]. This vocoder is one of the typical coding schemes used in the H.323 family of standards, which is ratified for the VoIP (Voice over IP).

3. Camera manipulation for speaker

   In video conference systems, the camera manipulation is necessary as well as the acoustic processing mentioned above in order to capture the active speaker's face properly. For this purpose, the estimated speaker directions using the speech signals are effective.

1.1.3 Research purpose

This dissertation deals with the following three subjects. The first one relates to the VAD problem, and the latter two topics concern the speaker direction estimation.

Voice activity detection using speech features in different signal domains (Chapter 3)

The aim of this topic is detecting voice activity segments under the condition where nonstationary interferences exist. Most of the existing VAD methods assume that the target speech signal is recorded in a sufficiently quiet environment. But in the practical situations, the VAD system has to cope with various interferences that occur sporadically. In the proposed method, the features of speech signal are classified by the signal domain to which each feature belongs. They are the temporal, spectral, and spatial domains. Then to utilize them in combination, the proposed method uses the strategy of array signal processing in the wavelet domain. Furthermore, it also tries to detect the unvoiced sound segment that has been difficult in conventional studies.

Speaker direction estimation with omni-directionality (Chapter 4)

Concerning the issue of speaker direction estimation, our purpose is the achievement of omni-directionality in the accuracy and discriminability. As the microphone array is sometimes surrounded by the speakers, the system has to specify the direction without any ambiguities, and uniform accuracy for omni-direction is preferred. Furthermore, the scale of the microphone array should be small enough to be adapted to practical applications. To solve these problems, we use an equilateral-triangular microphone array and exploit the harmonic structure in

the speech spectrum. In addition, the method includes a proposal to estimate not only the azimuth angle but also the elevation angle as a further subject.

Tracking of abruptly moving speaker directions (Chapter 5)

The purpose of this topic is the refinement of the above speaker direction estimation method for speaker tracking problems. The existing tracking methods of speaker direction assume the speakers' movements to be moderate. However, at a video conference, when numerous delegates speak in turn, the speaker direction can be abruptly changed. In this research, the proposed method achieves the proper tracking of the directions of speakers moving both gradually and abruptly. In the method, use is made of the different performance indices depending on the frequency band to avoid any local optimal trapping problem.

1.2 Overview of this dissertation

This dissertation is organized as follows. In the following Chap. 2, the conventional technologies of speech signal processing using the microphone array are summarised. For preliminary information, reference is first made to the speech signal with its production mechanism to understand the speech signal features. Then in Sec. 2.2, those speech signal characteristics are classified into three categories from the signal processing point of view. They are the features in temporal, spectral, and spatial domains. Signal processing schemes for time-frequency analysis are also mentioned in Sec. 2.3. The latter part of Chap. 2 consists of the introduction about the conventional speech signal processing methods using the microphone array. Following the principle of the microphone array signal processing given in Sec. 2.4, the typical functions of the microphone array, namely, the beamforming (Sec. 2.5), the direction estimation (Sec. 2.6), and the direction tracking (Sec. 2.7) respectively, are introduced.

Chap. 3 proposes a VAD method that has discriminability for nonstationary interferences. It uses the wavelet packet analysis and the microphone array in combination. This aims at achieving the connected use of the speech signal features in three signal domains. Sec. 3.2 recalls the speech signal features used in this method intensively, and the details of our proposal are explained in Sec. 3.3.

Some simulation results in Sec. 3.4 reveal the effectiveness of the method.

Chap. 4 and Chap. 5 are dedicated to the topic of speaker direction estimation. Chap. 4 deals with a new method to estimate the speaker direction that consists of three major topics. Sec. 4.2 firstly proposes a scheme to generate multi-dimensional array input data from a pair of microphones. In the method, a new idea to exploit the harmonic structure in speech spectrum for the purpose of estimation accuracy improvement is explained. Sec. 4.3 then puts forward a new algorithm for speaker direction estimation that utilizes the multi-dimensional data extracted from the three different sides of equilateral-triangle. The method employs the integrated use of this data to realise the omni-directionality. The last part of Chap. 4 further proposes an idea to extend the method in Sec. 4.4 to estimate both azimuth and elevation angles with less calculation load.

Chap. 5 deals with the strategy for applying the estimation mechanism to the speaker direction tracking. In the method, a scheme is proposed which estimates the direction by recursive calculation.

Finally Chap. 6 concludes this dissertation with further comments.

Chapter 2

Fundamental Technologies in Speech Signal Processing Using Microphone Array

2.1 Introduction

This chapter summarises conventional technologies of speech signal processing using the microphone array. First, we refer to the inherent features of speech signal with its production mechanism. Then in Sec. 2.2, we discuss the categorization of those speech signal characteristics into three groups from the signal processing point of view, namely the temporal, spectral, and spatial signal domains. There follows in Sec. 2.3, a discussion of time-frequency analysis that is able to represent the temporal and spectral features of a given signal. The next section explains the details of existing speech signal processing methods using the microphone array. After explaining the principle of the microphone array signal processing in Sec. 2.4, we introduce the typical methods that use the microphone array, i.e. beamforming (Sec. 2.5), direction estimation (Sec. 2.6) and its tracking (Sec. 2.7).

2.2 The production mechanism of speech signal and its features

Speech signals contain several kinds of information. They are linguistic information; namely, what the speaker wants to say, who is speaking, and emotional information about the speaker. In this section, especially from the signal processing point of view, we summarise the features of speech signal based on the production mechanism of the speech signal.

2.2.1 Modelling of speech signal production [2]

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 [2], whereas Japanese has 5 vowels and 20 consonants.

The mechanism of speech production consists of the following three steps, source generation, articulation and radiation. As summarised in figure [2][1], the human vocal system consists of the *lungs*, *trachea*, *larynx*, *pharynx*, *nasal* and *oral cavities*, connected together to form a tube. The upper part beginning with the larynx is called the *vocal tract*, which varies its shape by moving jaws, tongue, lips, and other internal parts. The abdominal muscle pushes up the *diaphragm*, the air is pressed out from the lung and passes through the *trachea* and *glottis*. The glottis, which is a gap between the left and right vocal chords, is usually open during breathing, and it gets narrower when the speaker intends to produce a sound. As the vocal chords open and close, the airflow passing through the glottis is periodically interrupted. This interrupted airflow is the source of speech.

When the vocal cords are strongly strained and the pressure of the airflow is high, the vocal cords vibrate more frequently and therefore the pitch of the sound source becomes high. In the low-pressure case in contrast, the airflow produces low-pitched sound. We call this vocal cord vibration period the *fundamental period* of speech, and its reciprocal is called the *fundamental frequency*. The

[1] see the Fig.2.2 on p.9 of [2]

accent and intonations are the results of fundamental frequency variation. The sound source generated by this vocal cord vibration consists of the fundamental component and its harmonics. The vocal tract to determine its timbre modifies their spectral envelope, and it results in each vowel.

Along with the vowel production and vocal cord vibration, there are two other mechanisms for changing the airflow from the lungs into speech sound. They are to produce the two kinds of consonants, *fricatives* and *plosives*. Fricatives are produced by turbulent flow, which occurs when the airflow passes through a narrow path in the vocal tract made by the tongue or lips. They sound like a noise, and the tonal difference between them depends on a fairly precisely located narrow path and vocal tract shape. Plosives are impulsive sounds that are produced by the sudden release of high-pressure air caused by trapping the airflow in the vocal tract by using the tongue or lips, and then expelling it. The tonal difference relates to the difference between this airflow trapping position and the vocal tract shape. The generating process of these phonemes is completely independent from the vocal tract vibration, and thus a sound produced with vocal cord vibration is classified as *voiced sound*, and for those without vocal cord vibration is called *unvoiced sound*. Fig. 2.1 and Fig. 2.2 show the waveforms and their spectra for both voiced and unvoiced sounds uttered by the same speaker.

On the other hand, the articulation changes the timbre of the source sound to make various linguistic sounds. It is produced by adjusting the vocal tract shape, because this adjustment changes the transmission characteristics of the vocal tract that works as a filter to put either emphasis or suppression on each frequency component.

Based on these principles of speech production mechanism, a speech sound is characterized by the types of sound source and the transmission characteristics of the vocal tract. Fig. 2.3 shows the linear equivalent system model of the speech production process.

2.2.2 Information in speech signal and its classification

Among the various kinds of information contained in speech, the speaker identification is important and useful information for many applications. So in this study, we observe the features of the speech signal, and classify them from the

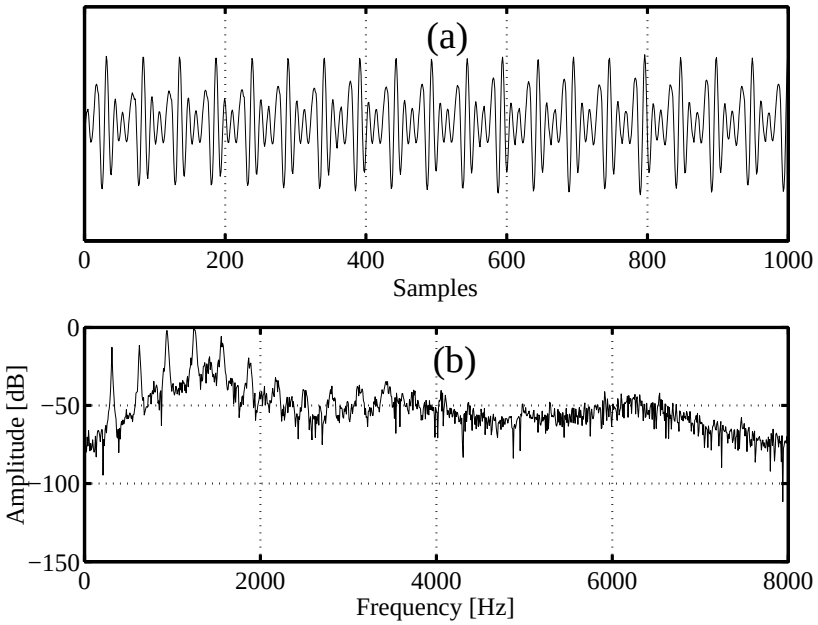

Figure 2.1: Phoneme /a/ : (a)waveform (b)spectrum

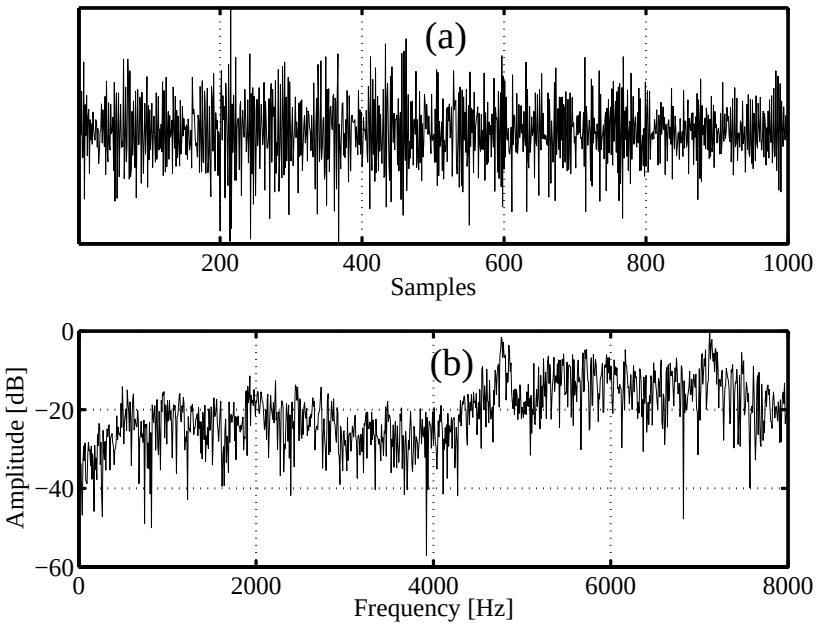

Figure 2.2: Phoneme /s/ : (a)waveform (b)spectrum

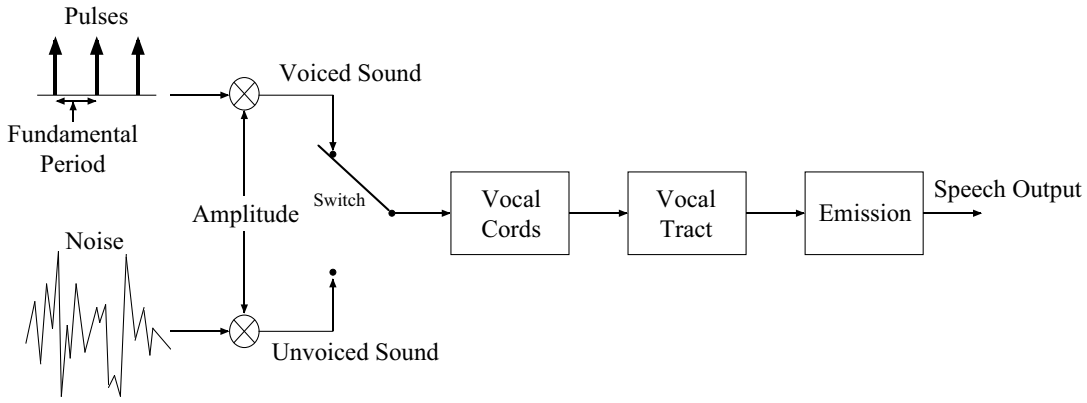

Figure 2.3: Linear equivalent system of the speech production process

signal processing view, i.e. the features in temporal, spectral and spatial signal domains. As aforementioned, a speech signal is classified into voiced and unvoiced sounds according to the speech production mechanism, therefore attention is paid to the differences between voiced and unvoiced sounds as well.

Temporal features

Looking at the speech signal waveform shown in Fig. 2.4, we immediately notice that the power of speech is concentrated in particular segments and there is no speech component in the rest of the segments. This power localization is one of the typical temporal features of a speech signal. Actually, the non-speech segments occupy more than half of the whole signal length, when we examine the signal over the sentence level unit. In particular, the unvoiced sounds are highly isolated and they disappear quickly within a few tens of milliseconds compared to the voiced sound that lasts for several hundreds of milliseconds.

As another temporal feature, a speech signal is generally conceived as a non-stationary signal because its spectrum characteristic varies at every moment. Nevertheless, the stationarity can be assumed for the voiced sound in a short period, which usually lasts for $30 - 40$[ms]. Furthermore, the signal energy in an unvoiced sound segment is much lower than that in a voiced sound segment because the driving power in the speech production process is different between voiced and unvoiced sounds.

Spectral features

The power spectrum of a speech signal spreads over wide band and its distribution typically depends on speakers, phonemes, and so on. The spectrum shape also largely depends on whether the sound is voiced or unvoiced. As shown in Fig. 2.1(b), the voiced sound has inherent harmonic structure, that is, its power is mostly concentrated on the specific harmonics in the lower frequency band. Usually, each harmonic component is assumed to be sufficiently narrowband. The fundamental frequency is time-varying between the limited range of $80-400$[Hz], and generally speaking, female speech signal shows higher fundamental frequency distribution in comparison with that of male speech signal. In contrast, the spectrum of unvoiced sound spreads over wide band as shown in Fig. 2.2(b). From the spectrogram of speech signal in Fig. 2.5, we can find the time-varying spectral features.

Spatial features

Besides the above two classifications, another category of information can be defined for speaker identification; the spatial characteristics of the received sound. This is acquired in receiving the sound.

In relation to spatial information for speech discrimination, the two features to consider are, the directionality of the received sound and the speaker location. The directionality of signal identifies whether the signal has information about the source position. As shown in Fig. 2.6, a sound with directionality shows high correlation if it is received at spatially different positions. But in the case of signals without directionality, such as sensor noise, it shows low correlation. Using such correlation degree, we measure the directionality of received signal. Furthermore, the speech signal uttered by each speaker has inherent position information. Hence, estimating the speaker position is an important technology in speech signal processing.

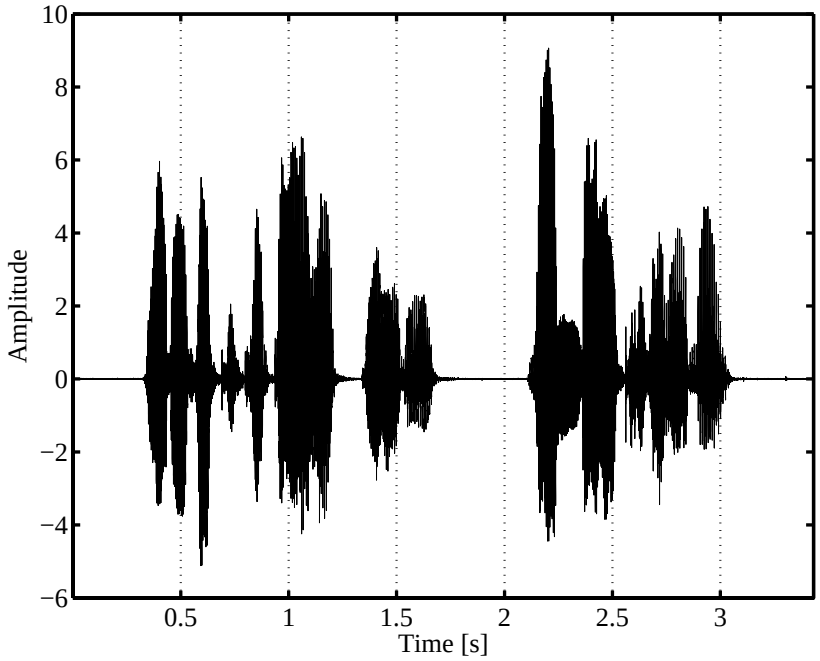

Figure 2.4: Example of speech signal (A male speaking a Japanese sentence /Yarubekikoto wa yatte ori nanra ochido wa nai/ [24])

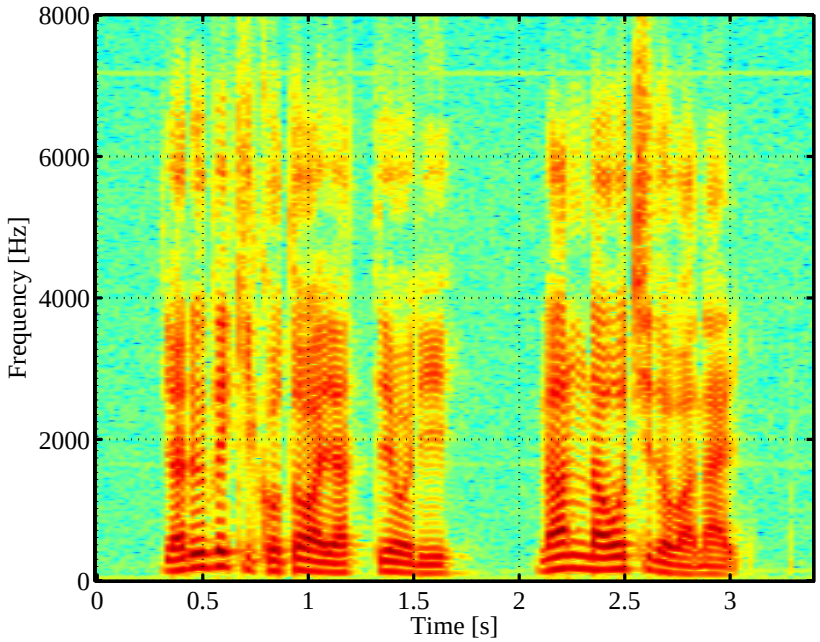

Figure 2.5: Spectrogram of speech signal in Fig. 2.4

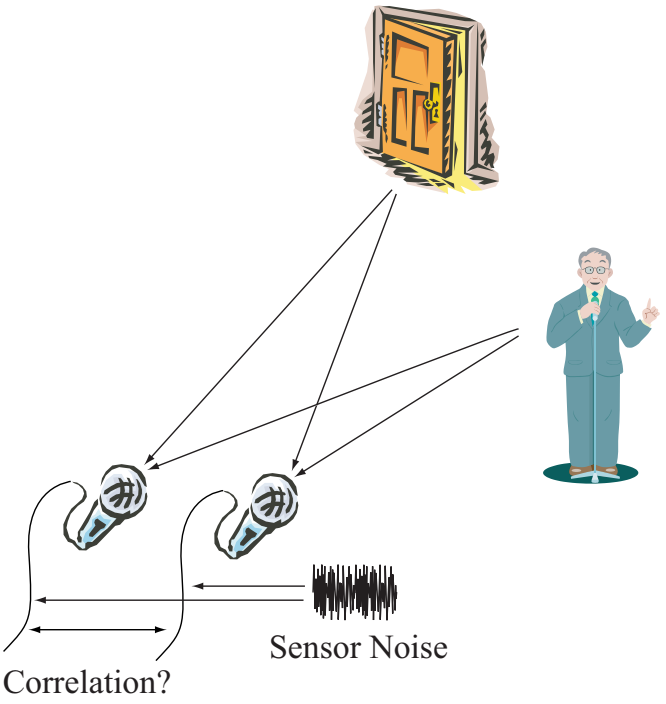

Figure 2.6: Directionality of the sound source

2.3 Time-frequency analysis [25][26]

A speech signal received at a microphone is usually represented by samples in the time domain. By performing spectrum analysis on a given signal, the spectral features can be exploited. Fourier transform [1] is the most popular and fundamental technology for spectrum analysis, and it offers spectrum distribution of a given signal with high precision. In the case of analysing a temporally nonstationary signal including a speech signal, Fourier analysis suffers from a drawback that it loses the temporal information of the input signal, and consequently, the target signal should be assumed as stationary. Thus, for the speech signals, we need time-frequency analysis. As one major method for that, the Short Time Fourier Transform (STFT) [26] divides the signal into several fixed-length frames with appropriate window function, and applies Fourier transform to each frame. This method is a compromise with the uncertainty principle, however, the spectral resolution is inflexible at different frequency bands because the frame length determines it. In contrast, wavelet analysis [26] is very powerful tool for speech

signal analysis because it can vary the window size according to the frequency band. This scheme is quite fruitful for time-frequency analysis of broadband signals. In this section, we have a brief view of STFT and wavelet analysis, which are the standard time-frequency analysing methods, with their fundamental principles and features. The differences between the wavelet packet analysis [26] and the conventional wavelet analysis are considered.

2.3.1 Short time Fourier transform

Short time Fourier transform

As shown in Fig. 2.7 and Fig. 2.8, Fourier transform acquires the spectrum at the expense of the temporal information of the original signal. Short Time Fourier Transform overcomes this problem by dividing the signal into short frames and performs ordinary Fourier transform to the signals in each frame. To investigate the signal features at time instant t, we first multiply a window $h(\tau)$ with the signal $s(\tau)$ to extract a segmented signal, given by

$$s_t(\tau) = s(\tau)h(\tau - t). \tag{2.1}$$

Applying Fourier transform to this segmented signal, the short time spectrum around the time instant t is given by

$$S_t(\omega) = \frac{1}{\sqrt{2\pi}} \int e^{-j\omega\tau} s_t(\tau) d\tau \tag{2.2}$$

$$= \frac{1}{\sqrt{2\pi}} \int e^{-j\omega\tau} s(\tau)h(\tau - t) d\tau. \tag{2.3}$$

Thus, the power spectral density at t is derived as

$$P_{PSD}(t, \omega) = |S_t(\omega)|^2 = \left| \frac{1}{\sqrt{2\pi}} \int e^{-j\omega\tau} s(\tau)h(\tau - t) d\tau \right|^2, \tag{2.4}$$

and by calculating P_{PSD} for all t and ω, we have the spectral distribution in the time-frequency domain called *Spectrogram*.

Uncertainty principle

In using STFT, we encounter a problem about the determination of frame size. Improving the temporal resolution, which is equal to shortening the frame length,

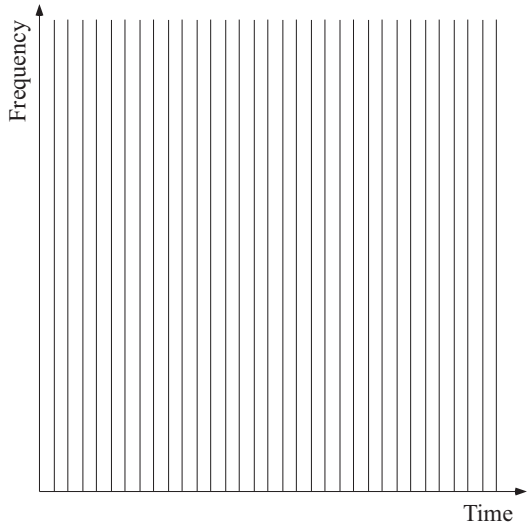

Figure 2.7: Segmentation of time-frequency plane in time domain signal

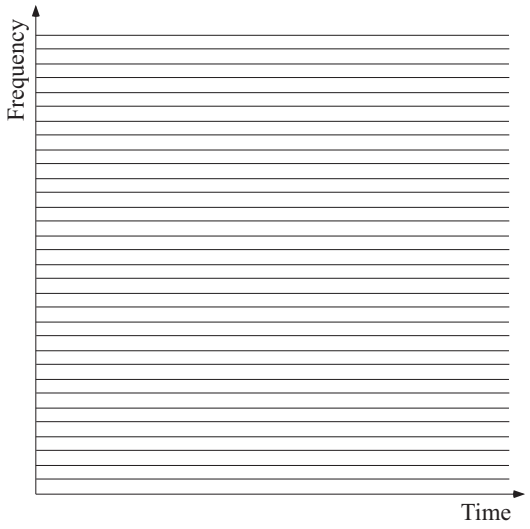

Figure 2.8: Segmentation of time-frequency plane by Fourier transform

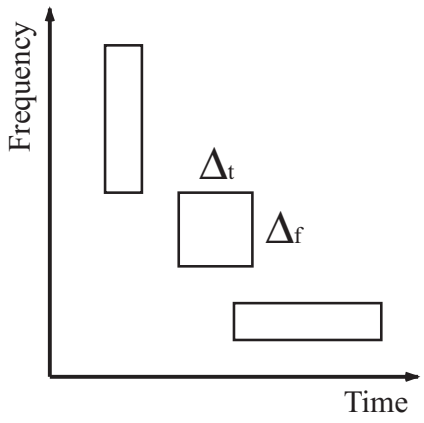

Figure 2.9: Segmentation of time-frequency plane under restriction of uncertainty principle

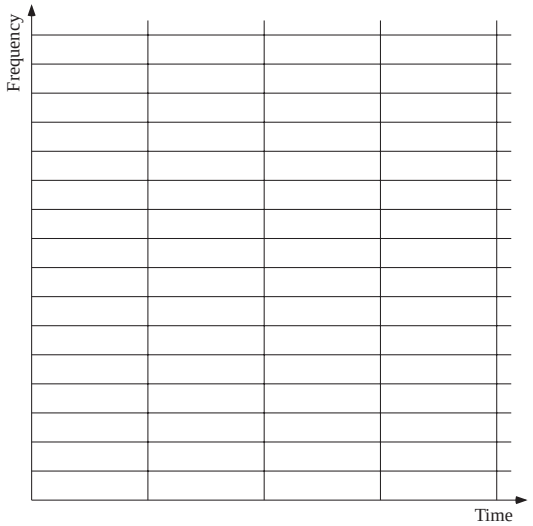

Figure 2.10: Segmentation of time-frequency plane by STFT

degrades the spectral resolution. In vice versa, expanding the frame length helps to improve the spectral resolution but it degrades the temporal resolution. Thus, there is a tradeoff between the temporal and spectral resolution called the *Uncertainty Principle*, and in applying time-frequency analysis, appropriate sets of temporal and spectral resolution should be chosen in keeping with the limitation by the uncertainty principle.

In general, a finite-length signal $f(t)$ occupies an area on the temporal axis with its range Δ_t. On the other hand, from the spectral point of view, the Fourier transform of $f(t)$, $F(\omega)$, occupies the frequency domain with the range Δ_f. As we can conjecture from the above features in STFT, decreasing both Δ_t and Δ_f simultaneously is impossible beyond the following rule, which is determined by the uncertainty principle.

$$\Delta_t \Delta_f > 2 \tag{2.5}$$

Thus, we have to take care of choosing the best tiling within the uncertainty principle to achieve efficient time-frequency analysis as shown in Fig. 2.9.

Fig. 2.10 shows the tiling of the time-frequency domain achieved by STFT. Because STFT divides the signal into frames at the beginning, the temporal resolution of that analysis is determined by this frame size, and thus the whole frequency band is analysed by the same spectral resolution. However, such fixed resolution is not suitable for speech analysis, because the voiced and unvoiced sounds in speech signal have different temporal and spectral features. Wavelet analysis that gives a solution to this problem is introduced in the next Sec. 2.3.2.

Window selection

Although the role of window in STFT is to avoid the distortion that occurs at both ends of segmented signal, it also alters the spectrum of signal. Thus, we need some discussions for the window selection used in STFT. We introduce three typical windows used for speech signal processing.

[Rectangular]

$$w(n) = \begin{cases} 1 & 0 \leq n \leq M-1 \\ 0 & \text{else} \end{cases} \tag{2.6}$$

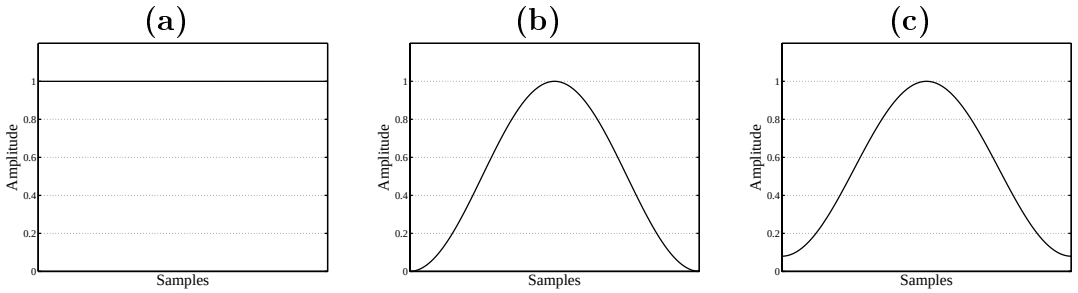

Figure 2.11: Windows for frame segmentation : (a)Rectangular, (b)Hanning, (c)Hamming

[Hanning]

$$w(n) = \begin{cases} \frac{1}{2}\left(1 - \cos\frac{2\pi n}{M-1}\right) & 0 \leq n \leq M-1 \\ 0 & \text{else} \end{cases} \tag{2.7}$$

[Hamming]

$$w(n) = \begin{cases} 0.54 - 0.46\cos\frac{2\pi n}{M-1} & 0 \leq n \leq M-1 \\ 0 & \text{else} \end{cases} \tag{2.8}$$

The rectangular window (Fig. 2.11(a)) quarries the target signal without weighting. This window does not perform any modification to the original signal, but the discontinuities at the end of the extracted frame have an undesired influence on the spectrum. Other windows (Fig. 2.11(b)(c)) suppress this effect by eliminating the discontinuities in their temporal features.

2.3.2 Wavelet analysis

Wavelet analysis is a powerful scheme for time-frequency analysis that can solve the fixed resolution problem in STFT. It realises an efficient time-frequency analysis within the uncertainty restriction, by applying different spectral resolution suitable for each frequency band. "Wavelet" is a unit to quarry a part of signal given by

$$\psi\left(\frac{x-b}{a}\right), \tag{2.9}$$

where $\psi(x)$ is a function called *Mother Wavelet*, a and b are the scaling (determines the frequency) and translating (determines the time instant) parameters, respectively. In the Wavelet transform, the scaled and translated mother wavelet extracts the time-frequency characteristics of a given signal. In the following, we first explain the continuous wavelet transform, which is the fundamental principle of wavelet analysis. Then, the discrete wavelet transform and its expansion, wavelet packet analysis, are described. These schemes are suitable for the discrete signals.

Continuous wavelet transform

In the Continuous Wavelet Transform (CWT), the scale and translate parameters a and b are continuous, so the result will be continuous as well. The CWT for a signal $f(x)$ with mother wavelet $\psi(x)$ is denoted as

$$(W_\psi f)(b,a) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{|a|}} \psi^* \left(\frac{x-b}{a} \right) f(x) dx, \tag{2.10}$$

where $*$ means the complex conjugate. Through plotting the transformed result $(W_\psi f)(b,a)$ on a $(b, 1/a)$ plane, we have the time-frequency representation of the signal. This $(W_\psi f)(b,a)$ shows how the signal $f(x)$ matches with the scaled mother wavelet $\psi(x/a)$ around $x = b$.

In contrast, the original signal $f(x)$ can be reconstructed from the wavelet transform, i.e. the inverse wavelet transform. It is defined as

$$f(x) = \frac{1}{C_\psi} \int \int_{\mathbf{R}^2} (W_\psi f)(b,a) \frac{1}{\sqrt{|a|}} \psi \left(\frac{x-b}{a} \right) \frac{\mathrm{d}a\mathrm{d}b}{a^2}. \tag{2.11}$$

Here, this equation holds if and only if the following admissible condition is satisfied.

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} \mathrm{d}\omega < \infty, \tag{2.12}$$

where $\hat{\psi}$ means the Fourier transform of ψ. Instead of this general admissible condition, usually the following condition is utilized.

$$\int_{-\infty}^{\infty} \psi(x) \mathrm{d}x = 0 \tag{2.13}$$

This equation means that the mother wavelet is oscillatory.

Discrete wavelet transform

Although the CWT is an effective tool to obtain time-frequency features of signals, it is not really an efficient way to be performed by numerical calculation. One reason is that it is difficult to realise the integral in the right side of Eq.(2.10) numerically. Furthermore, $(W_\psi f)(b, a)$ includes a lot of redundant information. If the following two points on the time-frequency plane, $(b, 1/a)$ and $(b', 1/a')$, are sufficiently close, $(W_\psi f)(b, a)$ and $(W_\psi f)(b', a')$ are no longer independent. From these facts, it is better to extract discrete points on the time-frequency plane but without loss of information. This is achieved by sampling the coordinate axes of b and $1/a$, which is usually performed by binary sampling as $(b, 1/a) = (2^{-j}k, 2^j)$.

Motivated by these facts, Discrete Wavelet Transform (DWT) is derived by sampling the signal of CWT in Eq.(2.10) as

$$(W_\psi f)(2^{-j}k, 2^j) \equiv d_k^{(j)} = 2^j \int_{-\infty}^{\infty} \psi^*(2^j x - k) f(x) \mathrm{d}x, \tag{2.14}$$

and its inverse (corresponds to Eq.(2.11)) is

$$f(x) = \sum_j \sum_k d_k^{(j)} \psi(2^j x - k). \tag{2.15}$$

Fig. 2.12 shows an example of DWT tiling on the time-frequency domain.

In the inverse DWT, the left and right sides of Eq.(2.15) are equal if and only if $\psi(2^j x - k)$ is a basis of the space to which $f(x)$ belongs. Thus in DWT, we need to select appropriate mother wavelet.

Multiresolution analysis

Now let us rewrite the summation about k in Eq.(2.15) as

$$g_j(x) \equiv \sum_k d_k^{(j)} \psi(2^j x - k). \tag{2.16}$$

Using this $g_j(x)$, we can denote

$$f_j(x) = g_{j-1}(x) + g_{j-2}(x) + \cdots, \tag{2.17}$$

where the integer j is the *level* of the subband. Assuming that $f(x) = f_0(x)$, Eq.(2.15) can be denoted as

$$f_0(x) = g_{-1}(x) + g_{-2}(x) + \cdots. \tag{2.18}$$

This corresponds to the decomposition of signal $f_0(x)$ into the wavelet coefficients $g_{-1}(x), g_{-2}(x), \cdots$. Now the Eq.(2.17) can be rewritten as the recursive form of $f_j(x)$ as

$$f_j(x) = g_{j-1}(x) + f_{j-1}(x). \tag{2.19}$$

Using this relation, a signal is decomposed into two subbands with one level decrement, and the bandwidth of the decomposed signal is half of its original one.

At the decomposition in Eq.(2.17), no ambiguity about $g(x)$'s is allowed, and the reconstruction must be correctly performed as well. To achieve these, the mother wavelet ψ should be carefully selected to be the basis of the signal. The appropriate mother wavelet is derived by the hierarchic structure called *Multiresolution Analysis*. Now we have a function $\phi(x)$ satisfying the equation called *two-scale relation*

$$\phi(x) = \sum_k p_k \phi(2x - k), \tag{2.20}$$

and such $\phi(x)$ that satisfies Eq.(2.20) is called *scaling function*. From *scaling function*, the mother wavelet is also defined as

$$\psi(x) = \sum_k q_k \phi(2x - k). \tag{2.21}$$

Recalling Eq.(2.19), the function $f_j(x)$ can be denoted by the linear combination of the scaling functions

$$f_j(x) = \sum_k c_k^{(j)} \phi(2^j x - k), \tag{2.22}$$

where $\phi(x)$ is invariable for every level j. Lead by this fact, we reach to the following decomposition algorithm in practice.

$$\begin{cases} c_k^{(j-1)} = \frac{1}{2} \sum_{l \in \mathbf{Z}} g_{2k-l} c_l^{(j)} \\ d_k^{(j-1)} = \frac{1}{2} \sum_{l \in \mathbf{Z}} h_{2k-l} c_l^{(j)} \end{cases}, \tag{2.23}$$

where $\mathbf{Z}$ means a set of natural number. The coefficients for decomposition $\{g_k\}$ and $\{h_k\}$ are determined from the two-scale relation, so they are also invariable for the decomposition level.

Subband decomposition

The decomposition algorithm given by Eq.(2.23) can be recognised as filtering and downsampling of discrete signal $\{c_k^{(j)}\}$. In other words, the decomposition coefficients $\{g_k\}$ and $\{h_k\}$ play a role of FIR filter coefficients, therefore the values of these progressions decide the frequency response at the decomposition. Fig. 2.14 shows the frequency response of a typical wavelet, Daubechies3. The decomposition coefficients $\{g_k\}$ and $\{h_k\}$ relate to the low-pass filter (LPF) and high-pass filter (HPF), respectively. According to the sampling theory, the maximum frequency of the digital signal is determined by the sampling frequency ω_S that corresponds to the normalized discrete frequency 2π, thus the signal $\{c_k^{(j)}\}$ possesses the spectrum components within the frequency band $[0, \pi]$. The LPF and HPF divide the band at $\omega = \frac{\pi}{2}$, and therefore the outputs $\{c_k^{(j-1)}\}$ and $\{d_k^{(j-1)}\}$ relate to the lower and higher subband components of the signal $\{c_k^{(j)}\}$. Thus, such division of a given signal into two subbands is called *Subband Decomposition.*

Conversely, it is also able to reconstruct the original signal $\{c_k^{(j)}\}$ from the subband signals $\{c_k^{(j-1)}\}$ and $\{d_k^{(j-1)}\}$. In the reconstruction, the coefficients $\{p_k\}$ and $\{q_k\}$, which have appeared in the two-scale relation, play a role of interpolation filters. After upsampling the subband signals, we apply these filters and sum up them to derive the reconstructed signal.

As shown in Fig. 2.15, there is a set of 4 filters called *filter bank* in the subband decomposition and its reconstruction. Arranging the filters in line with the *Quadrature Mirror Filter* theory is a popular method of achieving perfect reconstruction subband decomposition. In practical DWT whose example is shown as a tree-structured filter bank in Fig. 2.16, we recursively decompose the lower band signal. Through this process, we can decompose the signal into subbands.

Wavelet packet analysis

Wavelet Packet Analysis (WPA) extends the ability of wavelet analysis. In WPA, there is a choice of decomposing the higher band as well as the lower band to improve the flexibility in setting the multiresolution on the time-frequency plane.

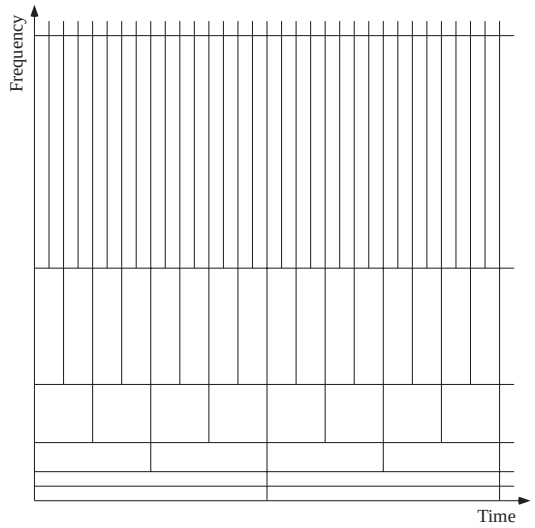

Figure 2.12: Segmentation of time-frequency plane by DWT

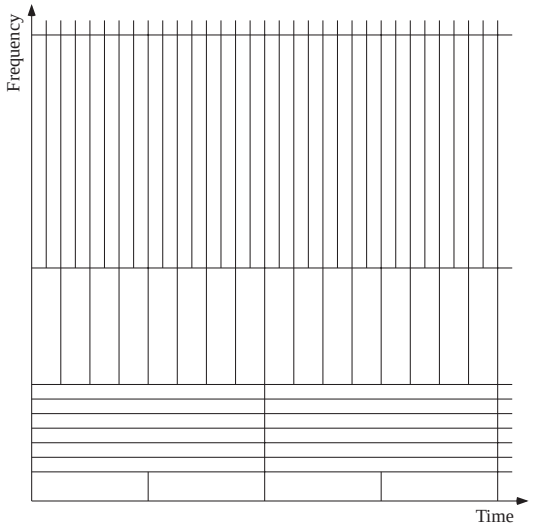

Figure 2.13: Segmentation example of time-frequency plane by WPD

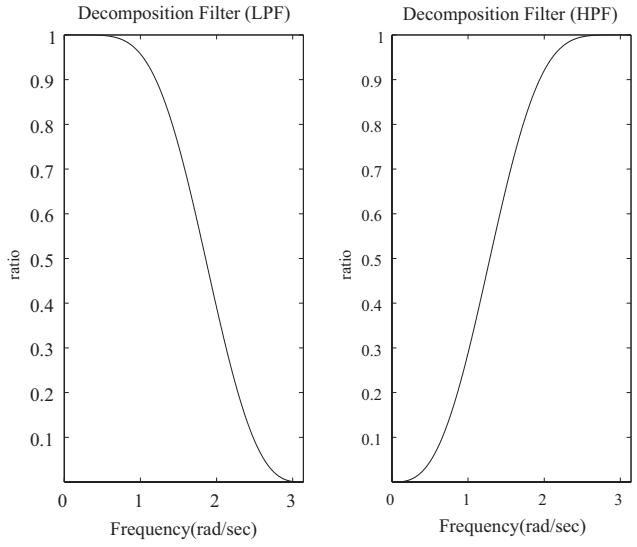

Figure 2.14: Decomposition filters of Daubechies $N = 3$

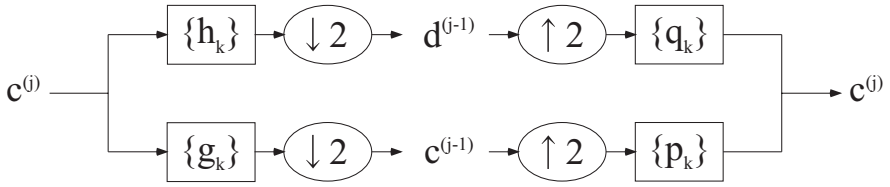

Figure 2.15: Subband decomposition & reconstruction with Quadrature Mirror Filter

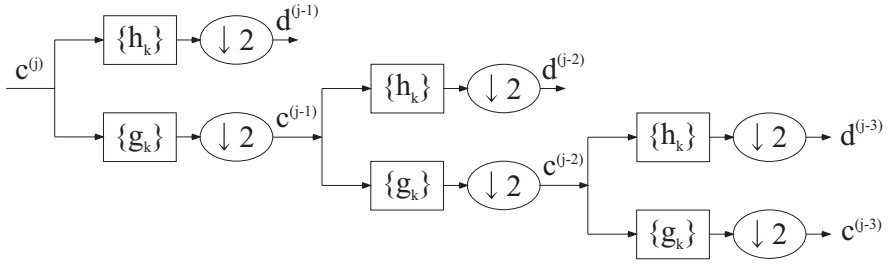

Figure 2.16: DWT filter bank tree (decomposition level : 3)

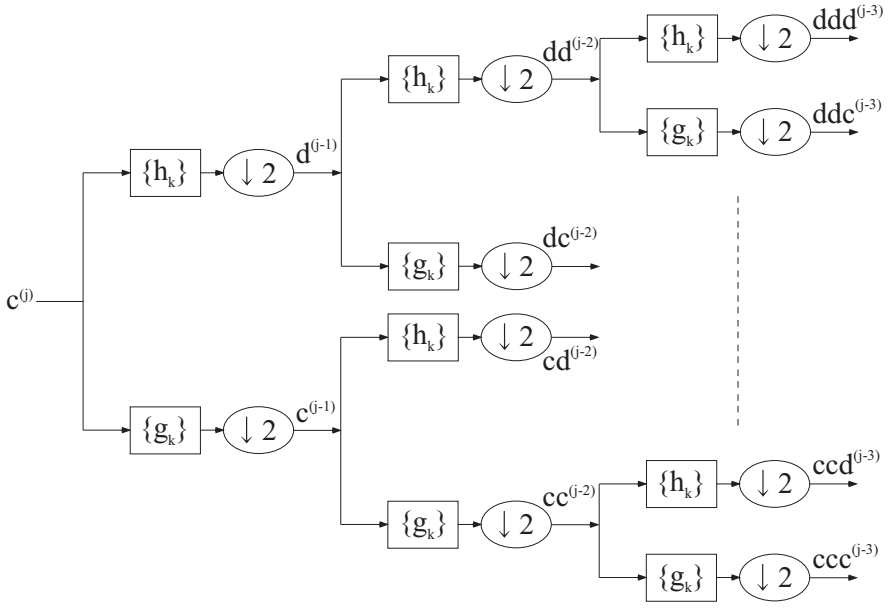

Figure 2.17: WPA filter bank tree (decomposition level : 3)

An example WPA filter bank tree is shown in Fig. 2.17. Thus, WPA can give arbitrary segmentation of a time-frequency plane as shown in Fig. 2.13 that is appropriate for the given signal as long as the decomposition does not exceed the given level.

For speech signal, the voiced and unvoiced sounds have different features on the time-frequency domain, so the WPA is suitable for the analysis of speech signal. In Chap. 3, we utilize such advantages of WPA to the voice activity detection problem.

2.4 Microphone array signal processing [4]

2.4.1 Signal propagation model at indoor environment

In speech signal processing, spatial information such as, direction-of-arrival (DOA) angle, source localization, etc. is useful. In open air space, a sound is emitted from the source and it reaches a receiver directly, so the model of sound propagation is quite simple. However, most of the practical situations that the speech signal processing is required for are indoor environments. When recording a speech signal in such a situation, the propagation is modelled as a dynamic linear system.

In a general indoor environment, the microphone receives not only the direct sound but also some reflections and reverberations. In the example as shown in Fig. 2.18, we can find 1st and 2nd order reflections that arrive from different directions. The sound reflection repeats until the sound energy vanishes due to the wall absorption. Fig. 2.19 shows a model of impulse response between the sound source and the microphone. In this example, the direct sound is received at first, and then the low order reflection arrives at the microphone successively. These low order reflections are called *early reflections*. Following the early reflections, high order reflections called *reverberation* reach the microphone.

In denoting the impulse response as a sample sequence $\{a(0), a(1), a(2), \cdots\}$, the transfer function $A(z)$ in Fig. 2.18 corresponds to the z-transform of this

sequence

$$A(z) = \sum_{i=0}^{\infty} a(i) z^{-i}, \tag{2.24}$$

and thus the room transfer function $A(z)$ involves the spatial information between the sound source and the microphone. Fig. 2.20 shows the system function of Fig. 2.18.

For a more general case, Fig. 2.21 shows the N-input-M-output sound propagation model. The z-transform of input and output signals are denoted in vector form as

$$\mathbf{U}(z) = \begin{bmatrix} U_1(z) & U_2(z) & \cdots & U_N(z) \end{bmatrix}^T \tag{2.25}$$

$$\mathbf{Y}(z) = \begin{bmatrix} Y_1(z) & Y_2(z) & \cdots & Y_M(z) \end{bmatrix}^T. \tag{2.26}$$

Using the transfer function matrix $\mathbf{A}(z)$ whose ij element $A_{ij}(z)$ is the transfer function between i-th source and j-th microphone

$$\mathbf{A}(z) = \begin{bmatrix} A_{11}(z) & A_{21}(z) & \cdots & A_{N1}(z) \\ A_{12}(z) & A_{22}(z) & \cdots & A_{N2}(z) \\ \vdots & \vdots & \ddots & \vdots \\ A_{1M}(z) & A_{2M}(z) & \cdots & A_{NM}(z) \end{bmatrix}, \tag{2.27}$$

the relation among $\mathbf{A}(z)$, $\mathbf{U}(z)$ and $\mathbf{Y}(z)$ is denoted as

$$\mathbf{Y}(z) = \mathbf{A}(z)\mathbf{U}(z). \tag{2.28}$$

Thus, the acoustic characteristic of a general indoor environment can be modelled as linear time-invariant Multi-Input-Multi-Output (MIMO) system.

2.4.2 Acquisition of spatial features

The acquisition of spatial information about the speaker is realised by several sensing schemes, such as using ultrasonic, optics, infrared rays, and so on. Among these, utilizing the speech signal received by microphones is convenient and low cost. In the high SNR recording systems, we design a system that emphasizes the sound signal arriving from the speaker direction, which can be estimated by the received sound signal. Now there are two major well-known strategies to exploit the spatial features of speech signals, (A)directional microphone and (B)microphone array.

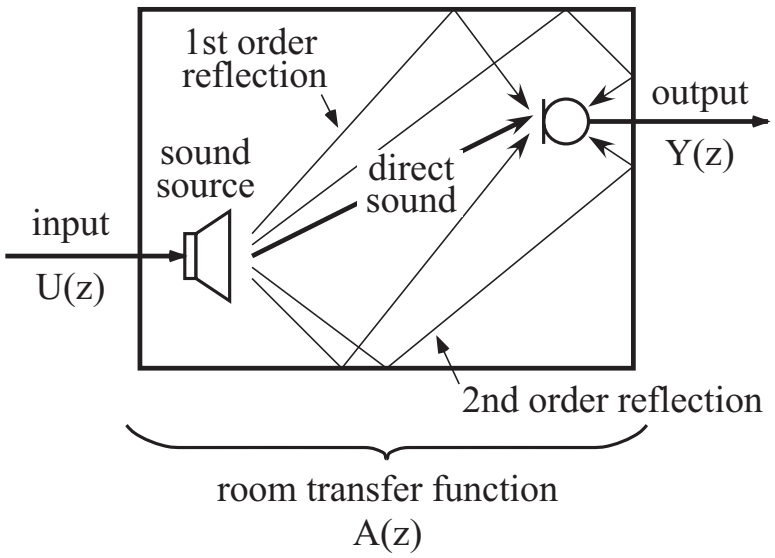

Figure 2.18: Sound propagation model in an indoor environment

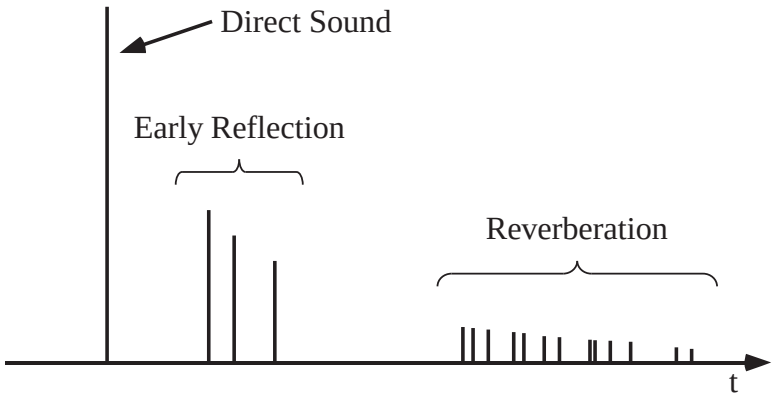

Figure 2.19: Impulse response model for indoor transfer function

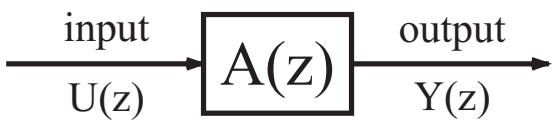

Figure 2.20: System function $A(z)$

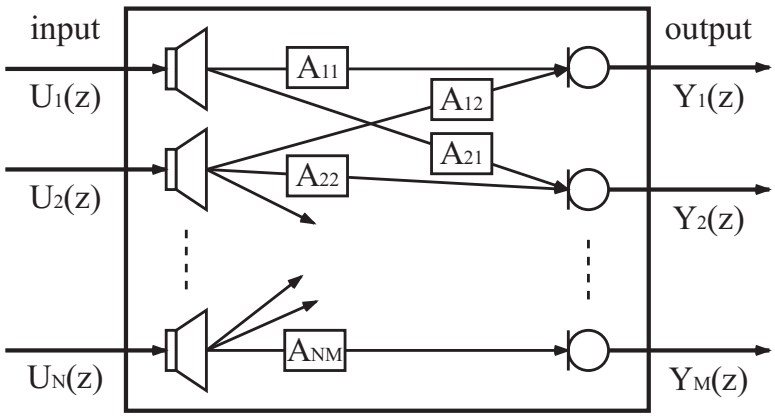

Figure 2.21: Sound propagation model for MIMO system

Directional microphone

As shown in the Fig. 2.22(a), the directional microphone has different sensitivity depending on the direction, so that only the sound arriving from the high sensitivity direction is received and others are suppressed. But in some cases including the example in Fig. 2.22(a), the noise cannot be suppressed due to the spatial response at the noise direction is not efficiently low. To overcome this problem, a super-directional microphone is used because it has a sharper directional beam in the direction of the desired signal. In the case of Fig. 2.22(b), the noise signal will be effectively suppressed. However, the super-directional microphone has a problem with its size. Because any of the conventional super-directional microphones are more than several tens of centimetres [4], their use in practice is restricted. Furthermore, if the desired or undesired signal directions frequently vary, we need to move the aspect of the microphone mechanically. This is the most crucial and inherent drawback of the directional microphones because its spatial response is fixed and never to be variable.

Microphone array

Microphone array is an effective solution to realise directional characteristic. Its directional beampattern can be modified electronically and it does not require much space to realise a sharp directional pattern. Fig. 2.22(c) shows a spatial response of an adaptive microphone array, which is a typical scheme to adaptively

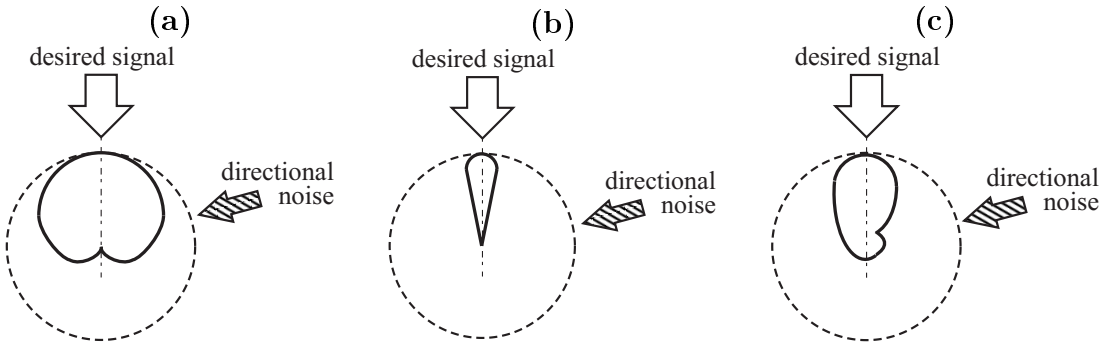

Figure 2.22: Spatial response of directional microphone : (a)Directional microphone (b)Super-directional microphone (c)Adaptive microphone array

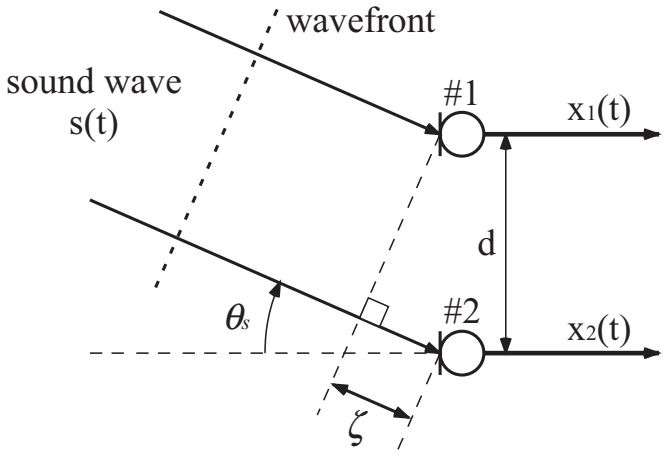

Figure 2.23: Sound signal received by spatially scattered microphones

specify the function of it. In the figure, a deep null is steered to the noise direction while the large gain is pointed in the target direction.

2.4.3 Problem settings of microphone array signal processing

On receiving a sound using these microphones, we can find each microphone signal with their amplitude and phase. The array signal processing is a scheme to realise the spatial signal processing by manipulating these signals.

Now using Fig. 2.23, we explain the mechanism of how the microphone array extracts the direction of sound signal. We assume that the sound is received at

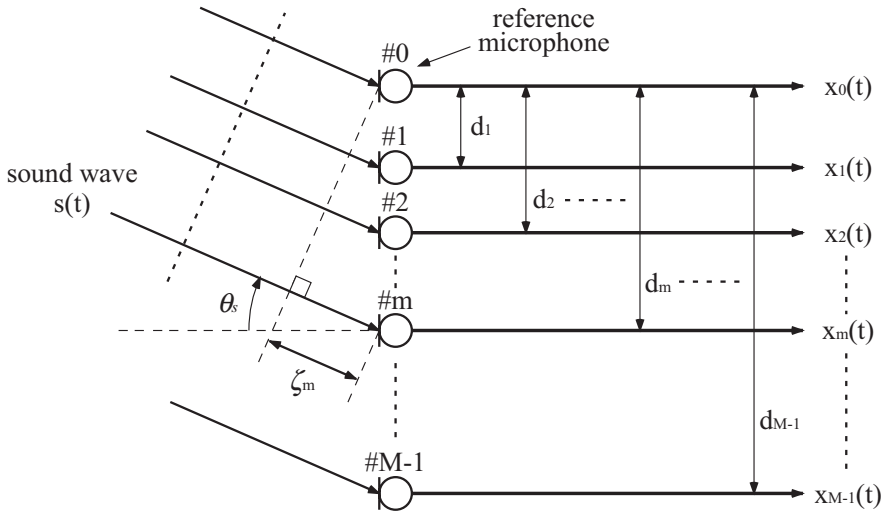

Figure 2.24: M-sensors linear microphone array

an open field as a plane wave, and the microphones are omni-directional (uniform response to omni-direction). In this example, the sound wave arrives from the direction θ_s, so it first arrives at the microphone #1 and then reaches the microphone #2 after propagating a distance ζ. From the figure, the distance ζ is denoted as

$$\zeta = d \sin\theta_s, \tag{2.29}$$

and therefore the signal received at #2 is the τ_s delayed signal of that received at #1, i.e.

$$x_2(t) = x_1(t - \tau_s) \tag{2.30}$$

$$\tau_s = \frac{\zeta}{c} = \frac{d \sin\theta_s}{c}, \tag{2.31}$$

where c is the sound speed. Measuring or manipulating this time arrival difference, we can obtain the direction of the received sound.

For more general case, Fig. 2.24 shows the M-sensors linear microphone array. By the same mechanism as we have explained, we have the delayed signals

$$x_m(t) = x_0\left(t - \frac{\zeta_m}{c}\right) = x_0(t - \tau_{s,m}) \tag{2.32}$$

$$\tau_{s,m} = \frac{\zeta_m}{c} = \frac{d_m \sin\theta_s}{c}, \tag{2.33}$$

where $x_0(t)$ is the signal received at the reference microphone #0.

Here the signal is assumed to be a complex sinusoidal wave with its angular frequency ω, so that the signal received at the reference microphone is denoted as

$$x_0(t) = X_0 e^{j\omega t} = s(t), \tag{2.34}$$

where X_0 is a complex amplitude of the signal. From Eq.(2.32), the signal at m-th microphone is derived as

$$\begin{aligned} x_m(t) &= X_0 e^{j\omega(t-\tau_{s,m})} \\ &= X_m e^{j\omega t} \qquad (2.35) \\ X_m &\equiv X_0 e^{-j\omega\tau_{s,m}}. \qquad (2.36) \end{aligned}$$

In Eq.(2.36), there is a phase difference $e^{-j\omega\tau_{s,m}}$ between X_m and X_0, which is easily expected from the relation in the Fourier transform. Substituting Eq.(2.33) and $\omega = 2\pi f$ for Eq.(2.36), we have

$$\begin{aligned} X_m &= X_0 e^{-j2\pi f \frac{d_m \sin\theta_s}{c}} \\ &= X_0 e^{-j2\pi f_{SP} d_m \sin\theta_s}, \qquad (2.37) \end{aligned}$$

where $f_{SP} = \frac{f}{c}$ means the spatial frequency of the sound wave. Writing in vector form, the input array data $\mathbf{x}(t)$ is defined as

$$\begin{aligned} \mathbf{x}(t) &= \left[\begin{array}{cccc} x_0(t) & x_1(t) & \cdots & x_{M-1}(t) \end{array} \right]^T \\ &= \left[\begin{array}{cccc} s(t) & s(t)e^{-j\omega\tau_{s,1}} & \cdots & s(t)e^{-j\omega\tau_{s,M-1}} \end{array} \right]^T \\ &= s(t) \left[\begin{array}{cccc} 1 & e^{-j\omega\tau_{s,1}} & \cdots & e^{-j\omega\tau_{s,M-1}} \end{array} \right]^T \\ &= s(t) \left[\begin{array}{cccc} 1 & e^{-j\omega\frac{d_1 \sin\theta_s}{c}} & \cdots & e^{-j\omega\frac{d_{M-1} \sin\theta_s}{c}} \end{array} \right]^T \\ &= s(t) \cdot \mathbf{s}(\theta_s). \qquad (2.38) \end{aligned}$$

The vector $\mathbf{s}(\theta)$ called *steering vector* contains arriving angle θ because it consists of the phase difference caused by the time arrival delay $\tau_{s,m}$. $\mathbf{s}(\theta)$ is given by

$$\mathbf{s}(\theta) = \left[\begin{array}{ccccc} e^{j\omega<\mathbf{r}_0\cdot\mathbf{l}_\theta>} & \cdots & e^{j\omega<\mathbf{r}_m\cdot\mathbf{l}_\theta>} & \cdots & e^{j\omega<\mathbf{r}_{M-1}\cdot\mathbf{l}_\theta>} \end{array} \right]^T, \tag{2.39}$$

where the notation $< \cdot >$ means the inner product. $\mathbf{r}_m$ is *array configuration vector* that determines the m-th microphone position, and $\mathbf{l}_\theta$ called *look direction vector* is a unit vector pointing at direction of signal arrival θ. For an example of linear microphone array in Fig. 2.24, the $\mathbf{r}_m$ for the microphone $\#m$ and $\mathbf{l}_\theta$ are defined as

$$\mathbf{r}_m = [d_m, 0, 0]^T \tag{2.40}$$

$$\mathbf{l}_\theta = [-\sin\theta, \cos\theta, 0]^T. \tag{2.41}$$

Apart from the subject of array input signals, let us consider a complex sinusoidal wave signal $Xe^{-j2\pi f_a t}$ whose frequency is f_a. When sampling this signal by the period T_S, we have a discrete signal denoted as

$$Xe^{-j2\pi f_a m T_S} \qquad (m = 0, 1, 2, \cdots) \tag{2.42}$$

If we substitute $f_a = f_{SP} \sin\theta_s$, $X = X_0$ and $mT_S = d_m$ into Eq.(2.42), the equation becomes Eq.(2.37) as far as the array configuration is equally-spaced linear array, i.e. $d_m = md (m = 0, 1, 2, \cdots)$. In other words, the complex amplitudes of equally-spaced linear array input signals correspond to the temporally sampled signal. From this fact, receiving a signal using the microphone array relates to spatial sampling. So as we state in detail later, the array signal processing works as a spatial FIR filter by summing up every microphone signal with appropriate weightings.

2.4.4 General features of microphone array

The scale of microphone array

Generally speaking, increasing either the number of microphones or the microphone array aperture size contributes to the enhancement of spatial resolution. However, the largest size of the microphone array is under the restriction of spatial sampling theorem given as follows.

[Spatial sampling theorem]

To avoid the spatial aliasing, the following relation should be held between the spatial sampling frequency determined by inter-microphone distance d and the maximum frequency f_{max} of the received signal.

$$\frac{1}{d} > 2\frac{f_{max}}{c} \tag{2.43}$$

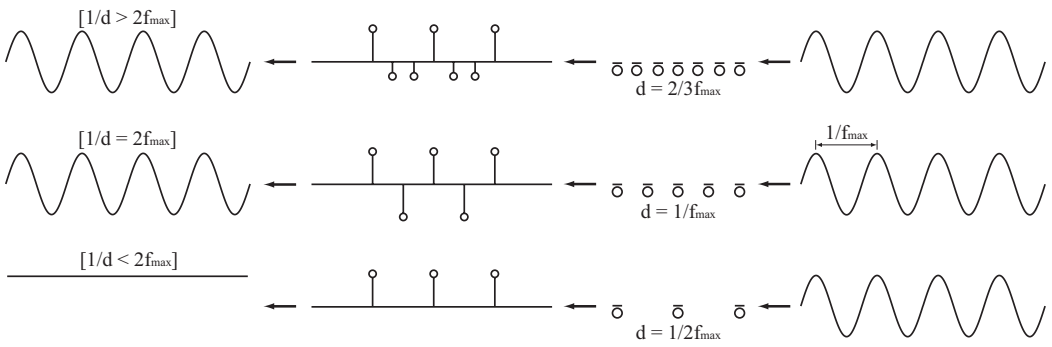

Figure 2.25: Spatial sampling theorem

As Fig. 2.25 shows, the spatial aliasing occurs if the inter-microphone distance d exceeds the half of the minimum wavelength λ_{min}, and thus rewriting the Eq.(2.43), we have the maximum restriction for d given by

$$d < \frac{\lambda_{min}}{2}. \tag{2.44}$$

In the beamforming, the spatial aliasing causes the glatinglobe in the beampattern. This problem is discussed in detail later.

Microphone arrangement

There are several microphone array arrangements. Fig. 2.26 shows some typical ones. The array configuration vector for each arrangement is given in the following. Here the look direction vector is defined as $\mathbf{l}_{\theta,\phi} = [\sin\theta\sin\phi, \cos\theta\sin\phi, \cos\phi]^T$ based on the situation in Fig. 2.27.

(a)Linear array

$$\mathbf{r}_m = [d_m, 0, 0]^T \tag{2.45}$$

(b)Rectangular array

$$\mathbf{r}_m = [md_x, nd_y, 0]^T \tag{2.46}$$

(c)Equilateral-triangular array

$$\mathbf{r}_m = \left[d\sin\frac{m\pi}{3}, d\cos\frac{m\pi}{3}, 0\right]^T \tag{2.47}$$

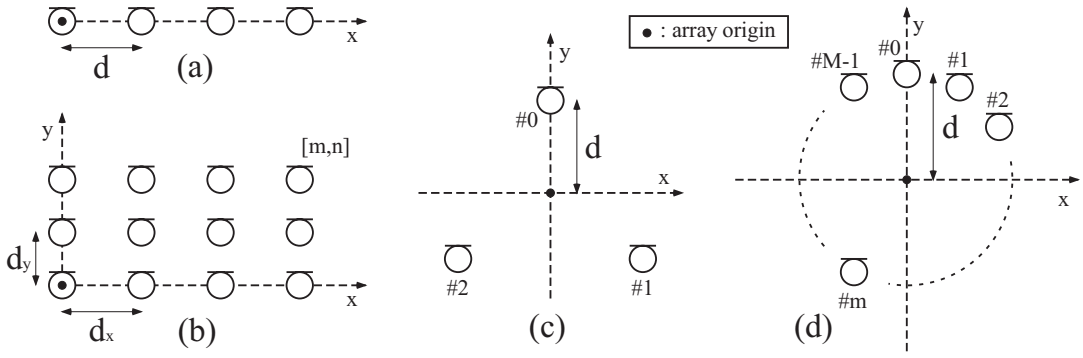

Figure 2.26: Typical microphone arrangement : (a)Linear (b)Rectangular (c)Equilateral-triangular (d)Circular

(d)Circular array

$$\mathbf{r}_m = \left[d \sin \frac{m\pi}{M}, d \cos \frac{m\pi}{M}, 0\right]^T \tag{2.48}$$

The linear arrangement has some typical disadvantages from the other arrangements. One of the drawbacks is the lack of discriminability for omni-direction. As explained by Fig. 2.28, the steering vector for the direction θ is coincident with that for the direction $\pi - \theta$, so from the phase difference, we cannot discriminate unique direction. This is because the microphones are located in line, it loses the ability to discriminate the elevation angle ϕ (direction around the array axis). In contrast, the another disadvantage is the nonuniform spatial resolution. As we will find in Fig. 2.34(a) and (b) later, the linear microphone array has the highest spatial resolution to its front side, and it decreases as the direction goes apart from the front direction. The details of this feature are discussed in Chap. 4.

Far field model & Near field model

So far for simplicity, we have had discussions about the microphone array with the assumption that the propagating signal is plane wave called *far field model*. But in some cases, this assumption does not hold, so that we have to consider another model, *near field model*, for the array input signal. Fig. 2.29 shows the wave front of sound signal received at the different distant position from the source. In the figure, the wave front becomes plane as the distance from the sound source is

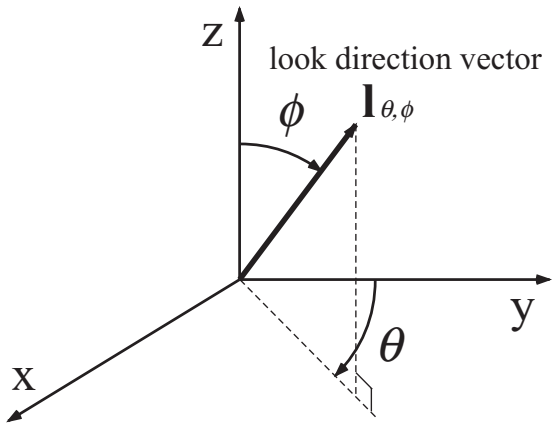

Figure 2.27: Look direction vector

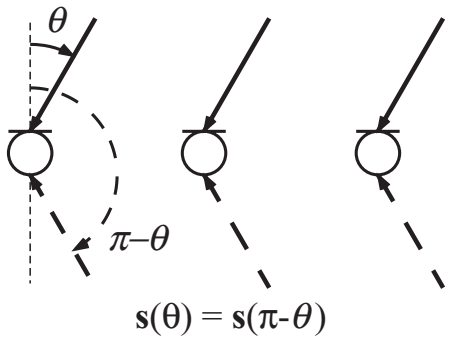

Figure 2.28: Lack of discriminability for omni-direction

enlarged. In general, the wave front of the sound signal is not plane wave because the sound wave is emitted spherically and omni-directionally from the source. In the near field model, the differences in the amplitude and phase are determined by the distance between source and microphone, R_m $(m = 1, \cdots, M-1)$. Thus the signal received at m-th microphone is derived as

$$x_m(t) = \frac{R_m}{R_0} x_0 \left(t - \frac{R_m - R_0}{c} \right) \tag{2.49}$$

If the microphone is sufficiently apart from the sound source, it is enough to assume the far field model. As we show in Fig. 2.30, the wave front gets spherical along with the enlargement of the inter-microphone distance.

Thus in designing a microphone array system, a consideration to the appropriate wave propagating model for the given problem is essential. The following

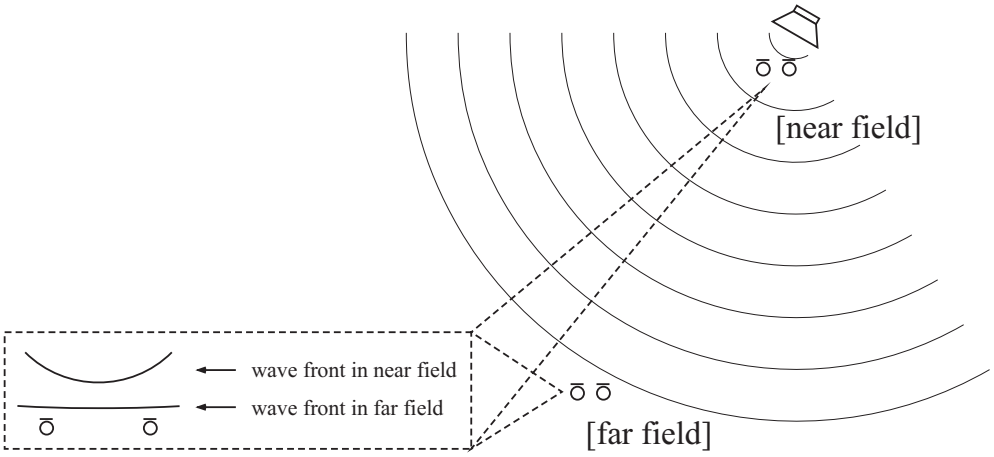

Figure 2.29: Difference of wave front shape at near field and far field

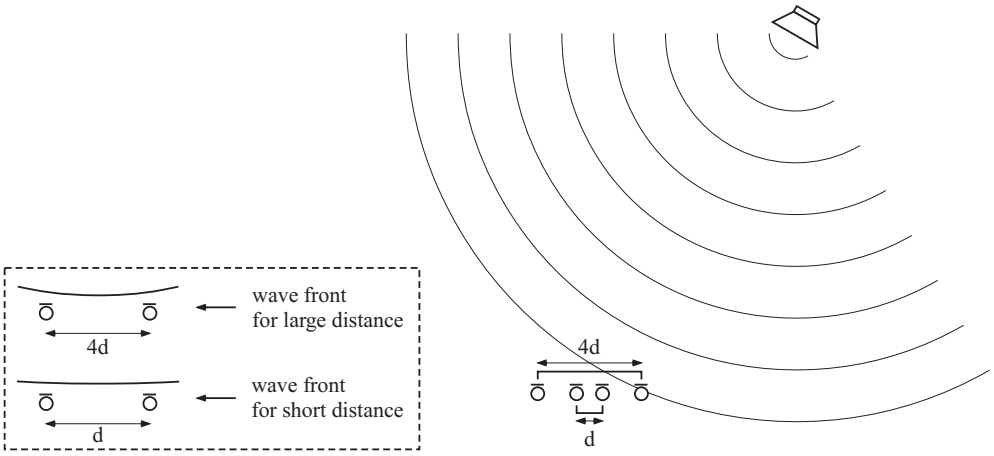

Figure 2.30: Difference of wave front shape for large and short inter-microphone space

decision rule [27] is known for selecting near field model.

$$R \leq \frac{2L^2}{\lambda} \tag{2.50}$$

where R is the distance between microphone and sound source, and L and λ are the array aperture and the wave length respectively. In the following part of the thesis, we adopt the far field model for the problem settings.

2.5 Beamforming

Microphone array works as a spatial FIR filter, called *beamforming*, by summing up every microphone input signal with appropriate weights. In this section, we introduce two types of beamforming, fixed beamforming and adaptive beamforming, including the mechanism of spatial filtering.

2.5.1 Fixed beamforming

The fixed beamforming is the most basic strategy to design a beamformer. In this type of beamformer, the weight for each microphone is not varied depending on the received signal. Here we first explain the mechanism of beamforming using the most basic and representative scheme, which is referred to as delay-and-sum beamforming. Following that, another fixed beamforming method is stated as well.

Delay and sum beamforming

Delay-and-sum beamforming is a method to steer a directional beam in the direction of the desired signal source. Now let us consider the problem that we receive a signal arriving from the direction θ_s using an equally-spaced linear microphone array, as shown in Fig. 2.31. The delay-and-sum beamformer has weights to delay the signal of m-th microphone with D_m defined as

$$D_m = D_r - m\tau_d \tag{2.51}$$

$$\tau_d = \frac{d \sin\theta_d}{c}, \tag{2.52}$$

where θ_d is the direction to which the beam should be steered, and D_r is the delay to ensure the causality. Summing up every delayed signal together, the output signal of the delay-and-sum beamformer is given by

$$y(t) = \sum_{m=0}^{M-1} x_m(t - D_m) \tag{2.53}$$

$$= \sum_{m=0}^{M-1} x_0(t - \tau_{s,m} - D_m) \tag{2.54}$$

$$= \sum_{m=0}^{M-1} x_0\left(t - m\frac{d}{c}\left(\sin\theta_s - \sin\theta_d\right)\right), \tag{2.55}$$

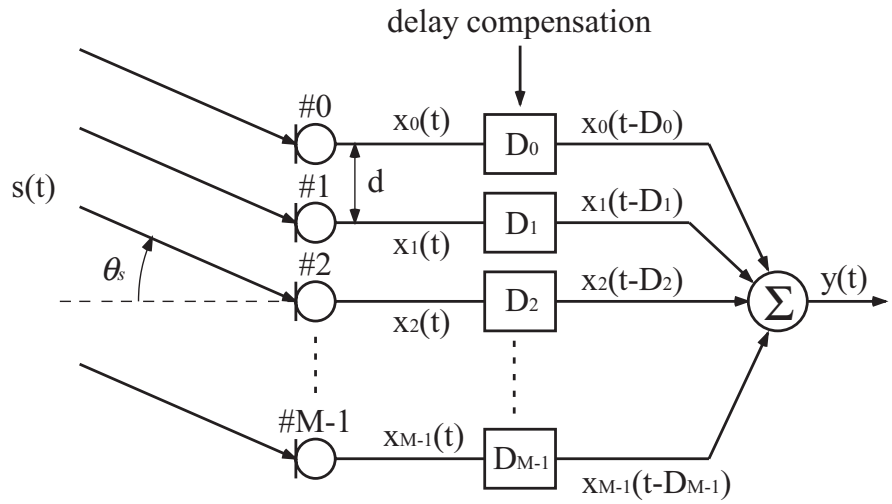

Figure 2.31: Delay-and-sum beamformer

where we substituted the Eq.(2.32) and Eq.(2.33) into Eq.(2.53).

Now looking at the Eq.(2.55), it can be seen that every delay compensated signal is in phase if the beam-steered direction is coincident with the signal arrival direction, i.e. $\theta_s = \theta_d$, and it results in amplifying the signal.

$$y(t)|_{\theta_s=\theta_d} = \sum_{m}^{M-1} x_0(t) \tag{2.56}$$

$$= Mx_0(t). \tag{2.57}$$

In contrast, if an undesired sound arrives from the direction $\theta_u \neq \theta_d$, the compensated signals are not in phase with each other and therefore they are not amplified by the summation.

Beampattern

To quantitatively evaluate the beamforming from spatial filtering point of view, the spatial response called *beampattern* is calculated. Here we explain the procedure to calculate the beampattern of delay-and-sum beamformer using an equally-spaced linear microphone array.

Now let us assume a complex sinusoidal wave arrives from the direction θ_s and its frequency is ω. As described in Eq.(2.36), the received signal at m-th

microphone is denoted as

$$x_m(t) = X_0 e^{-j\omega\tau_{s,m}} e^{j\omega t}. \tag{2.58}$$

The output of the delay-and-sum beamformer is given by

$$\begin{aligned} y(t) &= \sum_{m=0}^{M-1} x_m(t - D_m) \\ &= \sum_{m=0}^{M-1} X_0 e^{-j\omega m(\tau_s - \tau_d)} e^{j\omega t} \\ &= M X_0 e^{j\omega t} \sum_{m=0}^{M-1} e^{-j\omega m(\tau_s - \tau_d)}. \end{aligned} \tag{2.59}$$

The beampattern or the spatial amplitude response of a beamformer is defined as the amplitude ratio between the input and output signals.

$$G(\theta_s) = \left|\frac{y(t)}{s(t)}\right| = \frac{|y(t)|}{|x_0(t)|} \tag{2.60}$$

$$= M \left| \sum_{m=0}^{M-1} e^{-j\omega m(\tau_s - \tau_d)} \right| \tag{2.61}$$

$$= M \left| \sum_{m=0}^{M-1} e^{-j\omega m \frac{d}{c}(\sin\theta_s - \sin\theta_d)} \right| \tag{2.62}$$

$$= M \left| \frac{1 - e^{-j\Omega M}}{1 - e^{-\mathrm{j}\Omega}} \right| \tag{2.63}$$

$$= M \left| \frac{\sin(\Omega M/2)}{\sin(\Omega/2)} \right| \tag{2.64}$$

$$\Omega = \omega \frac{d}{c}(\sin\theta_s - \sin\theta_d) \tag{2.65}$$

Fig. 2.32 shows an example beampattern of a delay-and-sum beamformer. This beamformer steers its *mainlobe* to $\theta_d = 0°$, and getting apart from this direction, the gain decreases to the first *null*, which is analytically derived as

$$\theta_{null(1)} = \sin^{-1}(c/fdM). \tag{2.66}$$

The pass band of the beamformer, called *beamwidth*, is defined as the difference between the first nulls in both ends of the mainlobe[2]. As sharper as the

[2]There is another beamwidth definition using the direction that gives the -3dB gain of the mainlobe [28].

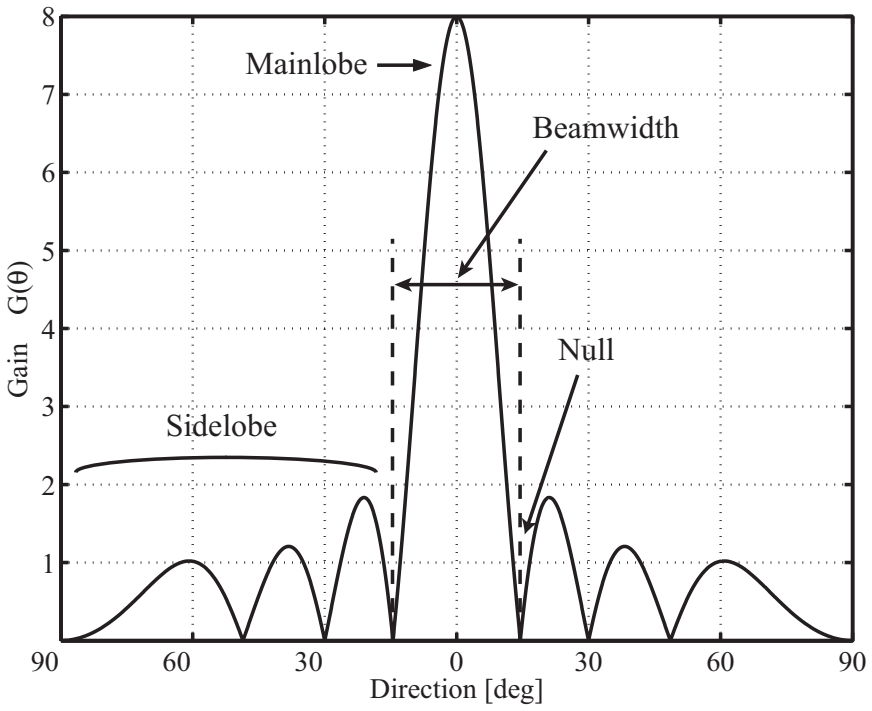

Figure 2.32: Beampattern of delay-and-sum beamformer ($M = 8, \theta_d = 0°, d = \frac{\lambda}{2}$)

beamwidth is, the spatial resolution of the beamformer becomes high. Outside the first nulls, there are some *sidelobes* that should be suppressed as low as possible.

Reconsideration of spatial sampling theorem

As stated in Sec. 2.4.4, the spatial sampling theorem gives the no-aliasing condition between ω and d. In Fig. 2.33, we show the beampattern of delay-and-sum beamformer where this condition is not satisfied. Comparing to the result in Fig. 2.32, the beampattern of Fig. 2.33 has the spatial aliasing called *glatinglobe*. The mechanism of glatinglobe occurrence is due to the loss of unique correspondence between the phase difference and the direction of signal arrival. In the case of 2-microphones ($M = 2$) delay-and-sum beamformer, for example, the beampattern of it is derived from Eq.(2.62), given by

$$G(\theta_s) = \left| 1 + e^{-j\omega \frac{d}{c}(\sin\theta_s - \sin\theta_d)} \right|, \tag{2.67}$$

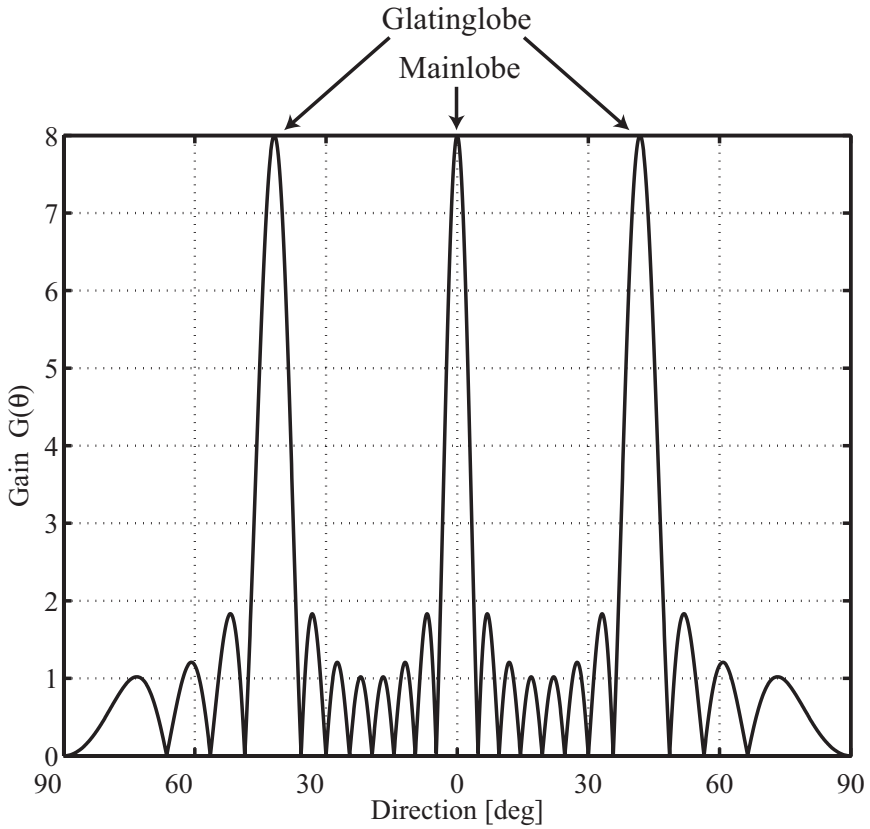

Figure 2.33: Effect of spatial aliasing for beampattern ($M = 8$, $\theta_d = 0°$, $d = \frac{3\lambda}{2}$)

which takes its maximum $G(\theta_s) = 2$ when the condition

$$e^{-j\omega\frac{d}{c}(\sin\theta_s - \sin\theta_d)} = 1 \tag{2.68}$$

is satisfied at $\omega\frac{d}{c}(\sin\theta_s - \sin\theta_d) = 2\pi l$ $(l = 0, \pm 1, \pm 2, \cdots)$. Because $\sin\theta_s$ varies between -1 and 1, and $\sin\theta_d$ is fixed, the condition holds uniquely at $\theta_s = \theta_d$ if $\omega\frac{d}{c} \leq \pi$, and otherwise, it holds at several θ_s's not equal to θ_d that causes the glatinglobe. Rewriting this relation by substituting $\omega = 2\pi f$, we have $d \leq \frac{\lambda}{2}$ that is exactly same as Eq.(2.44). Thus, the spatial aliasing causes the glatinglobes that emphasize the signals arriving from unexpected directions in the case of beamforming.

Effects of parameter setting for beampattern

The characteristic of beamforming highly depends on the parameters settled in its design. The beampattern is determined by the four parameters, that are the number of microphones M, direction of desired signal arrival θ_d, inter-microphone distance d, and frequency of the received signal ω. Fig. 2.34 shows the beampatterns of delay-and-sum beamforming using different parameters given in Table 2.1.

[DOA of desired signal]

In the case (b), the direction to steer the mainlobe is different to that of case (a). The mainlobe is wider than that of case (a) caused by the effect of non-uniform spatial resolution as explained in Sec. 2.4.4.

[Frequency of received signal]

In the case (c) and (d), the input signal frequencies are lower and higher frequencies than that of case (a), respectively. As the frequency increases, the mainlobe is sharpened and spatial resolution is improved. However, it must be remarked that the spatial aliasing occurs in higher frequency as shown in the case (d).

[Inter-microphone distance]

The case (e) and (f) show the beampatterns for different inter-microphone distance. The spatial resolution is improved as the distance is widened, but we need to be careful for the spatial aliasing as well as increasing the frequency.

[Number of microphones]

Finally the case (g) and (h) show the beampatterns for different number of microphones. As increasing it, the spatial resolution improves without the occurrence of spatial aliasing. Naturally, due to increasing the number of microphones extends the scale of the microphone array, it might be an obstruction to apply the array to small applications.

Beamforming for broadband signals

Generally, the speech signal should be dealt as a broadband signal because its spectrum spreads over wide frequency band as we have mentioned in Sec. 2.2. In application of beamforming to the broadband signals, the weightings cannot be achieved by the simple amplification and time-shifting of the received signal because the weights vary depending on the frequency. For this reason, we deal

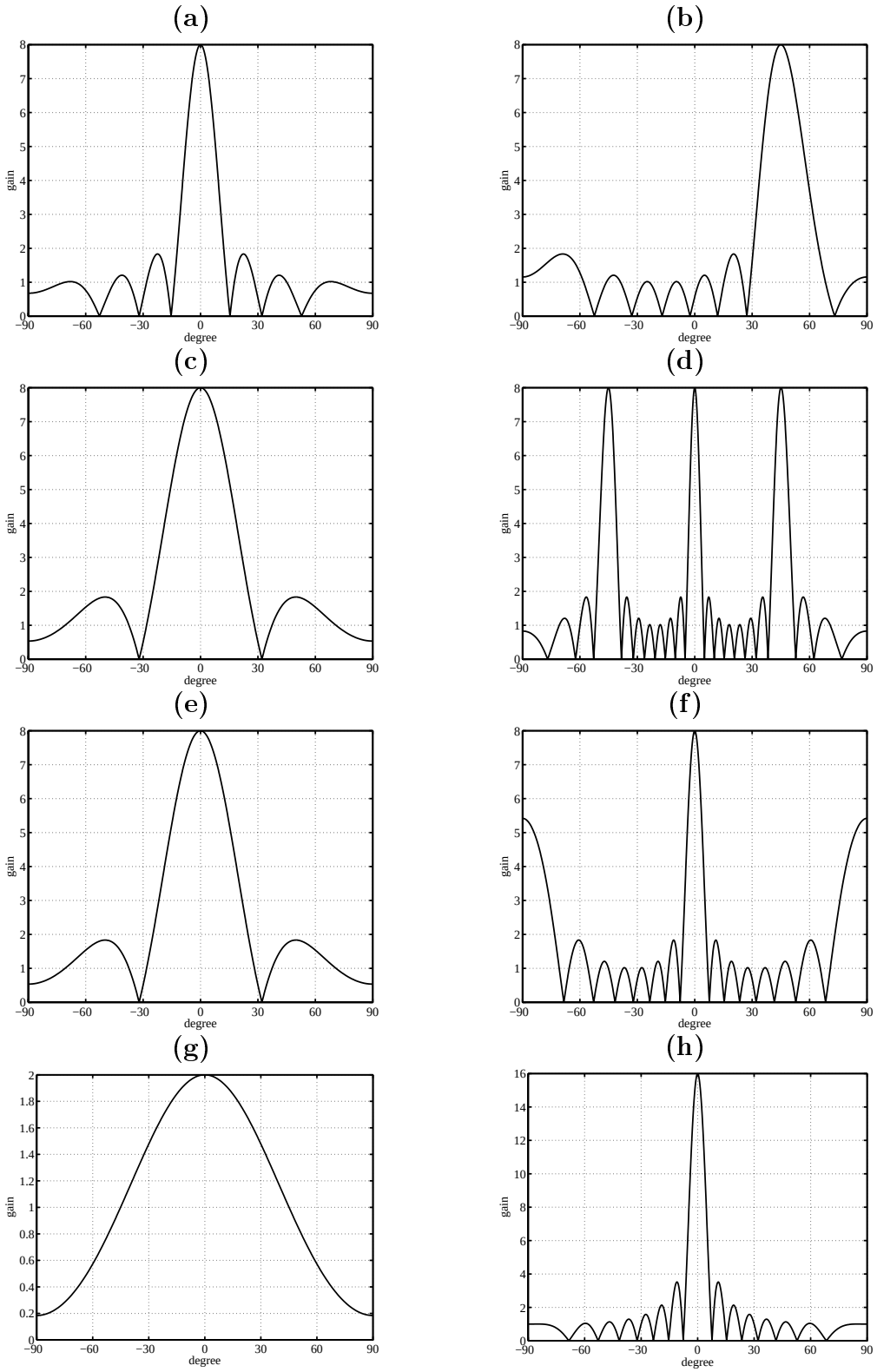

Figure 2.34: Parameter effects for delay-and-sum beamformer

Table 2.1: Parameter settings for Fig. 2.34

Case	d[cm]	f[Hz]	M	θ_d[deg]	Comments
(a)	8	2000	8	0	Standard setting
(b)	8	2000	8	45	Different θ_d
(c)	8	1000	8	0	Lower frequency f
(d)	8	6000	8	0	Higher frequency f
(e)	4	2000	8	0	Shorter inter-microphone distance d
(f)	16	2000	8	0	Larger inter-microphone distance d
(g)	8	2000	2	0	Smaller number of microphones M
(h)	8	2000	16	0	Larger number of microphones M

with the signals in the frequency domain. As shown in Fig. 2.35, the weight for m-th microphone is substituted by a transfer function $W_m(\omega)$, and the beamformer output is given by

$$Y(\omega) = \sum_{m=0}^{M-1} W_m(\omega) X_m(\omega), \tag{2.69}$$

where $X_m(\omega)$ and $Y(\omega)$ are the Fourier transforms of the received signal at m-th microphone and the output signal, respectively. This formula is given in vector form as

$$Y(\omega) = \mathbf{W}^T(\omega)\mathbf{X}(\omega), \tag{2.70}$$

where

$$\mathbf{W}(\omega) \equiv \left[\begin{array}{cccc} W_0(\omega) & W_1(\omega) & \cdots & W_{M-1}(\omega) \end{array} \right]^T \tag{2.71}$$

$$\mathbf{X}(\omega) \equiv \left[\begin{array}{cccc} X_0(\omega) & X_1(\omega) & \cdots & X_{M-1}(\omega) \end{array} \right]^T. \tag{2.72}$$

Thus in the case of beamforming for broadband signals, we need to apply FIR filter as shown in Fig. 2.36 whose impulse response is derived by the inverse Fourier transform of $\mathbf{W}(\omega)$. On the other hand, the steering vector for broadband signals is also represented by the transfer functions between the sound source and m-th microphone $A_m(\omega, \theta)$, which is exactly equal to the room transfer function

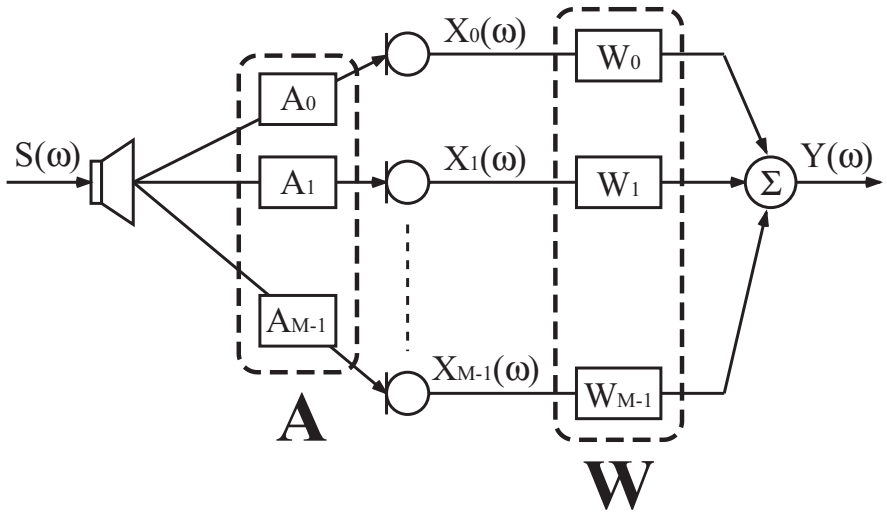

Figure 2.35: Transfer functions for broadband beamforming

as mentioned in Sec. 2.4.1. The beampattern for a broadband signal is derived by

$$G(\omega,\theta) = \left|\mathbf{W}^T(\omega)\mathbf{A}(\omega,\theta)\right| \tag{2.73}$$

$$\mathbf{A}(\omega,\theta) \equiv \left[\begin{array}{cccc} A_0(\omega,\theta) & A_1(\omega,\theta) & \cdots & A_{M-1}(\omega,\theta) \end{array}\right]^T. \tag{2.74}$$

For the delay-and-sum beamforming, the Fourier transform of Eq.(2.53) is given by

$$Y(\omega) = \sum_{m=0}^{M-1} X_m(\omega) e^{-j\omega D_m}, \tag{2.75}$$

and thus,

$$W_m(\omega) = e^{-j\omega D_m}. \tag{2.76}$$

From Eq.(2.76), the delay compensation corresponds to a linear-phase all pass filter for the broadband beamforming. Fig. 2.37 shows the beampattern of a delay-and-sum beamforming.

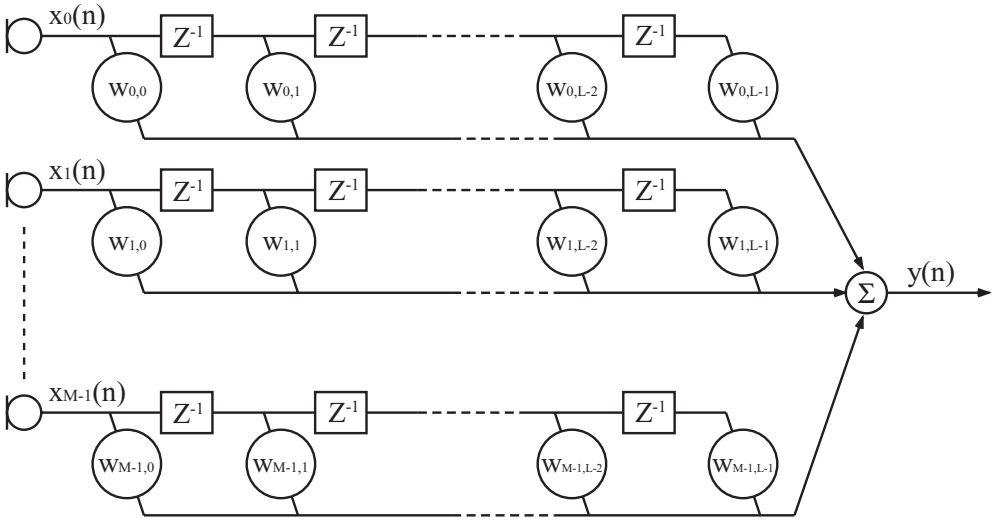

Figure 2.36: Beamforming for broadband signal

Null steering beamforming

As an extension of delay-and-sum beamformer, the null steering beamformer (delay-and-differential beamformer), which is described in Fig. 2.38, is defined by

$$\begin{aligned} y(t) &= x_1(t - D_1) - x_0(t - D_0) \\ &= x_0(t) - x_0\left(t - \frac{d}{c}\left(\sin\theta_s - \sin\theta_d\right)\right). \end{aligned} \tag{2.77}$$

In this beamformer, we use only two microphones, and calculate the difference between two microphones after the delay compensation. When $\theta_s = \theta_d$, the signal arrives from θ_d is in phase, and therefore it is suppressed by the subtraction. In Fig. 2.39, we show a beampattern of null steering beamformer for broadband signal.

2.5.2 Adaptive beamforming

Adaptive beamforming is another strategy for the beamformer design that determines its spatial response automatically by adapting to the condition of signal reception. Among several kinds of conventional adaptive beamforming methods, many of them are basically based on the mechanism of null steering beamforming. Fig. 2.40 gives an example of adaptive beamformer using a pair of microphones.

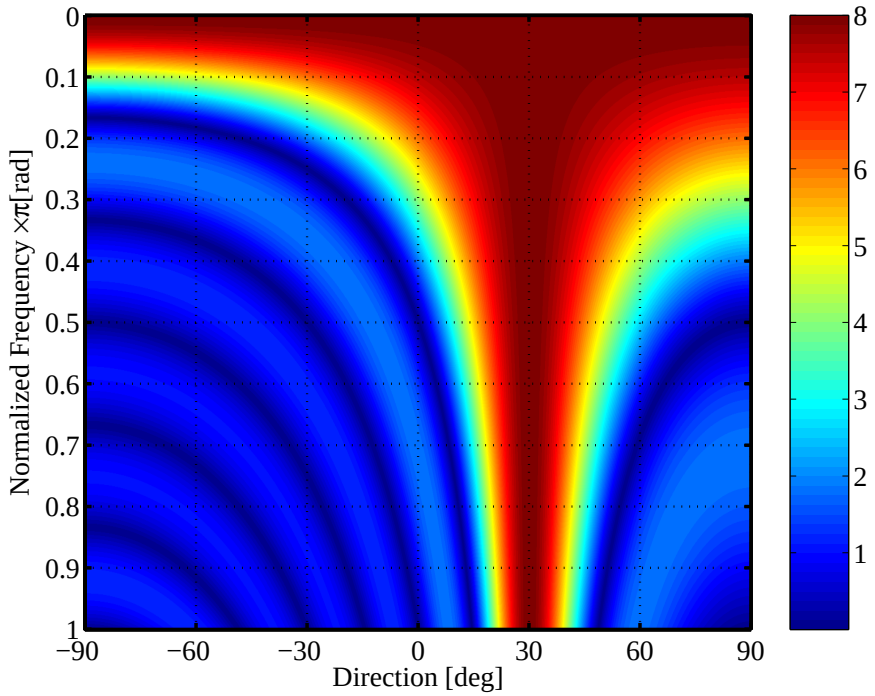

Figure 2.37: Beampattern of delay-and-sum beamformer for broadband signal ($M = 8$, $\theta_d = 30°$)

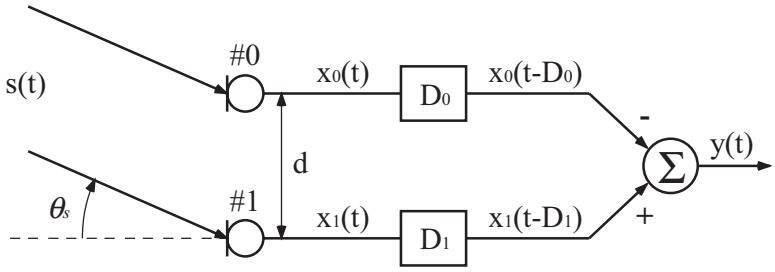

Figure 2.38: Null steering beamforming

This beamformer decides its characteristics by adjusting the delay compensation to minimize the output signal power. In other words, it searches the best θ_n to steer the null to suppress the interfering signal. From this fact, the biggest difference of the adaptive beamforming from the fixed beamforming is its dependence on the received signals. In this section, we introduce a major adaptive beamforming method, Generalized Sidelobe Canceller (GSC). GSC has been used for

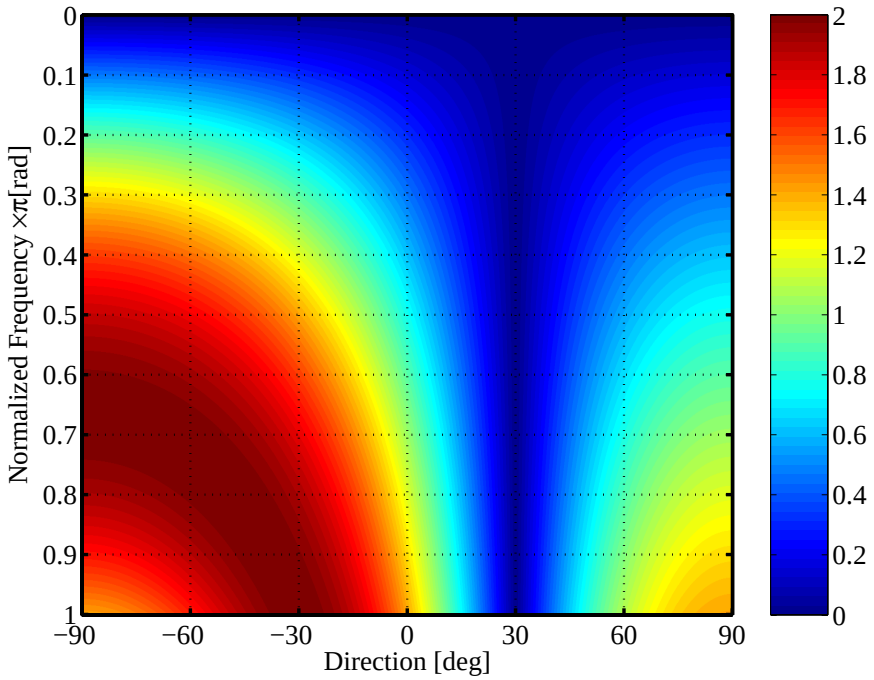

Figure 2.39: Beampattern of null steering beamformer for broadband signal

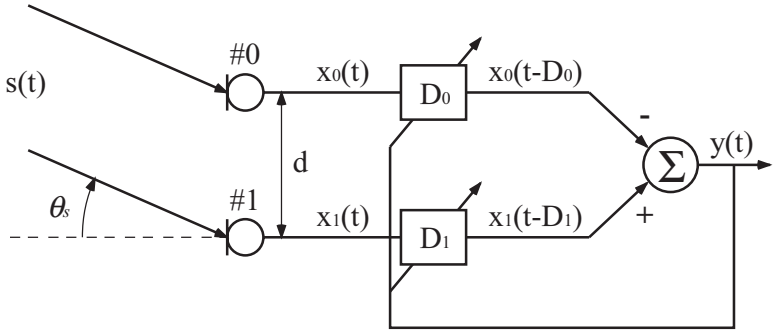

Figure 2.40: Two microphones adaptive beamforming

several speech signal processing studies.

Generalized sidelobe canceller

The GSC was first proposed by Griffith and Jim [29] for the adaptive beamforming of narrow band signals. It is an alternative scheme to realise the Linearly Constrained Minimum Variance (LCMV) beamformer, which is the most basic

adaptive beamforming technology proposed by Frost [9]. For the microphone array, several conventional works adopted the GSC structure for the purpose of sound source separation [10][30]. A brief summary of the mechanisms of GSC is given below.

Let us suppose a problem extracting a sound interfered with by some directional noises using a linear equi-spaced microphone array. We assume that the DOA of desired signal θ_d is known *a priori*. As shown in Fig. 2.41, the GSC consists of $M-1$ null steering beamformers whose outputs are followed by $M-1$ adaptive filters. Because the signal arriving from direction θ_d is suppressed by the null steering beamformers, the outputs of the null steering beamformers $z_m(t)$ contain only the undesired interferences. In contrast, the received signal $x(t)$ holds not only the interferences but also the desired signal. Based on these facts, GSC minimizes the output $y(t)$ that is derived by subtracting $z_m(t)$ from $x(t)$ because it results in suppressing the interference while retaining the desired signal. The calculation of GSC is given by

$$\mathbf{h}_{opt} = \arg\min_{\mathbf{h}} |y(t)|^2 \tag{2.78}$$

$$y(t) = x(t) - \sum_{m=0}^{M-2} z_m(t) \otimes h_m(t) \tag{2.79}$$

$$z_m(t) = x_{m+1}(t - D_{m+1}) - x_m(t - D_m), \tag{2.80}$$

where

$$\mathbf{h}(t) = \left[\begin{array}{cccc} h_0(t) & h_1(t) & \cdots & h_{M-2}(t) \end{array} \right]^T. \tag{2.81}$$

$\mathbf{h}_{opt}$ is the optimal $\mathbf{h}$, and $\otimes$ denotes the convolution. In practical calculation, the minimization is achieved by the recursive update of $\mathbf{h}$ using the steepest descent method, given by

$$\mathbf{h}(t+1) = \mathbf{h}(t) + \mu \mathbf{z}(t) y(t), \tag{2.82}$$

where

$$\mathbf{z}(t) = \left[\begin{array}{cccc} z_0(t) & z_1(t) & \cdots & z_{M-2}(t) \end{array} \right]^T, \tag{2.83}$$

and μ is the stepsize.

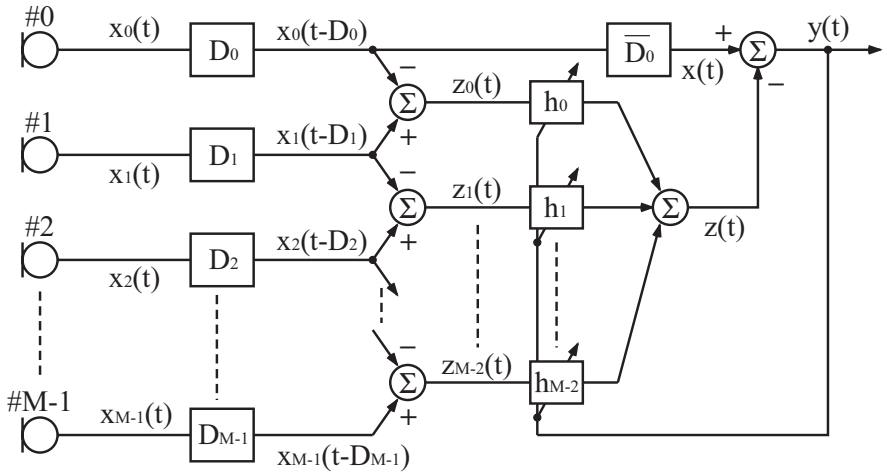

Figure 2.41: Generalized sidelobe canceller

2.6 Speaker direction estimation

The time arrival difference between two microphones corresponds to a unique direction of signal arrival. Measuring this time arrival difference connects the direction estimation of sound source.

Existing direction estimation methods can be roughly classified into three general categories [31]; those employing time delay estimation (TDE) calculated from the Generalized Cross Correlation (GCC) [32], the approaches based upon maximizing the steered response power of beamformer, and techniques adopting subspace analysis to achieve high-resolution estimation. The first group includes schemes, which calculate source locations from a set of delay estimates measured across various combinations of microphones. The second refers to any situation where the location estimate is derived directly from the output of beamforming. Methods in the last category apply the subspace analysis of the signal covariance matrix. We explain their mechanisms and the features respectively.

2.6.1 Time delay estimation using the generalized cross correlation function

Among the three major methods of speaker direction estimation, the TDE-based method possesses a significant computational advantage over the others. Actually,

a lot of passive speaker direction estimation systems in use today are based on TDE. The TDE method adopts a two-step procedure. First of all, we estimate the time arrival difference between the signals relative to a pair of spatially separated microphones. Using these values with knowledge of the microphone positions in combination, the direction of sound source is estimated.

The mechanism of direction estimation using TDE is quite simple. Suppose that a plane sound wave is received by a pair of microphones as shown in Fig. 2.38 . If the time arrival difference τ_s is estimated, the sound source direction θ_s is derived by

$$\theta_s = \sin^{-1}(c\tau_s/d). \tag{2.84}$$

For the estimation of τ_s, the Generalized Cross Correlation (GCC) function is the most popular method defined as

$$R(\tau) = \int_{-\infty}^{\infty} \Psi(\omega) G_{x_0 x_1}(\omega) e^{j\omega\tau} d\omega. \tag{2.85}$$

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

$$G_{x_0 x_1}(\omega) = X_0(\omega) X_1^*(\omega), \tag{2.86}$$

$\Psi(\omega)$ is a frequency weighting filter, and $*$ denotes the complex conjugate. Searching τ_s that gives the maximum peak of $R(\tau)$ and substituting it into Eq.(2.84), we have the estimated speaker direction.

Since accurate and robust TDE is the key to the effectiveness of direction estimation in this area, several frequency weighting filters $\Psi(\omega)$ are selected in the GCC function. There are two major interfering sources that degrade the estimation performance, which are non-directional background noise and multi-path channel due to room reverberation. To cope with the interferences like the former kind, the ML (Maximum Likelihood)-based function $\Psi_{ML}(\omega)$ was proposed [32]. Because this weighting function is based on the signal SNR at each frequency, it is appropriate to reduce the effects of spatially uncorrelated white noise. However, in the existence of room reverberations, these ML-based methods exhibit severe performance degradations.

In contrast, basic approach to dealing with multi-path channel distortions makes the GCC function more robust by deemphasizing the frequency weightings.

The Phase Transform (PHAT), given by

$$\Psi_{PHAT}(\omega) = \frac{1}{|X_0(\omega)X_1^*(\omega)|} = \frac{1}{|G_{x_0x_1}(\omega)|}, \tag{2.87}$$

is one of the weighting functions, which has received considerable attention recently. By placing equal emphasis on each frequency component, the resulting peak in the GCC-PHAT function that corresponds to the dominant delay stands out explicitly. Although the GCC-PHAT function is effective to reduce the degradations due to multi-path, it emphasizes the components of the spectrum with poor SNR, particularly under low reverberation.

Furthermore, other approaches for the selection of frequency weighting function in adverse environments are available. Brandstein *et al.* utilize a criterion based on the speech specific harmonic structure in the spectrum [33].

2.6.2 Fixed beamformer based method

The second group of direction estimation adopts the output of fixed beamforming. The simplest type of fixed beamforming is obtained using the output of a delay-and-sum beamformer. As already explained in Sec. 2.5.1, the delay-and-sum beamformer steers a sharp directional beam in a particular direction. We estimate the arrival direction of speech signal by measuring the directions that give the explicitly large power of beamforming output. In the situation of estimating single speaker direction, the angle that gives the maximum output power is recognised as the speaker direction given by

$$\overline{\theta} = \arg\max_{\theta_d} P_y(\theta_d), \tag{2.88}$$

where $\overline{\theta}$ is the estimated speaker direction and $P_y(\theta_d)$ is the power of beamformer output derived from Eq.(2.59).

$$P_y(\theta_d) = |y(t,\theta_d)|^2 = M^2X_0^2\left|\sum_{m=0}^{M-1} e^{-jm\Omega(\theta_d)}\right|^2 \tag{2.89}$$

$$\Omega(\theta_d) \equiv \omega\frac{d}{c}(\sin\theta - \sin\theta_d) \tag{2.90}$$

Because $P_y(\theta_d)$ is the estimate of received signal power, it is desirable to be as sharp as possible. Having a look at Eq.(2.89) in comparison with Eq.(2.64),

we can find that they are similar. Thus the characteristics of beampattern as mentioned in Sec. 2.5.1 hold for the relation between the beamformer output, the number of microphones M, the inter-microphone distance d, and the signal frequency f.

Although the beamforming based method does not take as much calculation as the following subspace analysis, it has a drawback that its spatial resolution is limited by the selected beamformer characteristics especially when numerous speakers talk simultaneously. Furthermore, the response of a beamformer is highly dependent on the spectral content of the source signal.

2.6.3 High-resolution spectral-estimation-based method

The last group of speaker direction estimation methods includes the beamforming schemes adapted from the field of high-resolution spectral analysis that are linear prediction method, minimum variance method, and MUSIC. Each of these techniques is based upon the spatiospectral covariance matrix derived from the signal received at the microphones. Here we explain the mechanism of MUSIC (MUltiple SIgnal Classification), which is the most popular method in this category, and refer to its advantages and disadvantages.

MUSIC

Let us suppose a M-microphones array as shown in Fig. 2.42 receiving N different speech signals ($M > N$) that are mutually uncorrelated. The received signal vector is given by

$$\mathbf{x}(t) = \left[\begin{array}{cccc} x_0(t) & x_1(t) & \cdots & x_{M-1}(t) \end{array} \right]^T \tag{2.91}$$

$$= \left[\begin{array}{c} \sum_{n=0}^{N-1} s_n(t) e^{-j\omega\tau_{n,0}} + n_0(t) \\ \sum_{n=0}^{N-1} s_n(t) e^{-j\omega\tau_{n,1}} + n_1(t) \\ \vdots \\ \sum_{n=0}^{N-1} s_n(t) e^{-j\omega\tau_{n,M-1}} + n_{M-1}(t) \end{array} \right] \tag{2.92}$$

$$= \sum_{n=0}^{N-1} s_n(t)\mathbf{s}_n + \mathbf{n}(t), \tag{2.93}$$

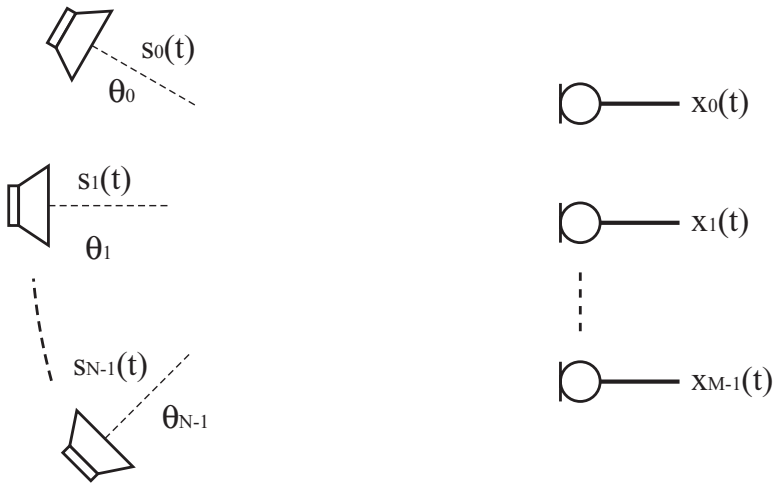

Figure 2.42: Reception of N speech signals using M-microphone array

where $\mathbf{s}_n$ denotes the steering vector for direction θ_n and

$$\mathbf{n}(t) = [n_0(t)\ n_1(t), \cdots, n_{M-1}(t)]^T \tag{2.94}$$

is the vector of Gaussian noise assumed to be spatially uncorrelated. We also assume that the mean value of each $x_m(t)$ is normalized to 0. Hence, the covariance matrix of $\mathbf{x}(t)$ is derived as

$$\begin{aligned} \mathbf{R} &= E[\mathbf{x}\mathbf{x}^H] && (2.95)\\ &= \sum_{n=0}^{N-1} E\left[\left|s_n(t)\right|^2\right] \mathbf{s}_n\mathbf{s}_n^H + E[\mathbf{n}(t)\mathbf{n}(t)^H] && (2.96)\\ &= \sum_{n=0}^{N-1} P_n\mathbf{s}_n\mathbf{s}_n^H + \delta^2\mathbf{I}, && (2.97) \end{aligned}$$

where $E[\cdot]$ means the expectation substituted by the temporal average, δ^2 and $\mathbf{I}$ are the noise power and unit matrix, respectively. Note that the terms relating to the cross-correlation between microphones are eliminated due to the uncorrelatedness assumption for the source signals.

Performing the eigenvalue decomposition, the covariance matrix $\mathbf{R}$ is decom-

posed into the M eigenvalues λ_m and the corresponding eigenvectors $\mathbf{v}_m$.

$$\mathbf{R} = \mathbf{E}\Lambda\mathbf{E}^H \tag{2.98}$$

$$\mathbf{E} \equiv \left[\begin{array}{cccc} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_M \end{array} \right] \tag{2.99}$$

$$\Lambda \equiv \mathrm{diag}\left[\begin{array}{cccc} \lambda_1 & \lambda_2 & & \lambda_M \end{array} \right] \tag{2.100}$$

$$\text{subject to} \quad \lambda_1 > \lambda_2 > \cdots > \lambda_M. \tag{2.101}$$

If every speaker direction θ_n is different, the eigenvalues are categorized into two groups by their amount.

$$\lambda_m = \begin{cases} \nu_m + \delta^2 & m = 1, 2, \cdots, N \\ \delta^2 & m = N+1, N+2, \cdots, M \end{cases}. \tag{2.102}$$

Thus the covariance matrix is rewritten as

$$\begin{aligned} \mathbf{R} = \mathbf{E} & \left[\begin{array}{ccc|ccc} \nu_1 & 0 & & \cdots & & 0 \\ 0 & \ddots & & & & 0 \\ & & \nu_N & & & \\ \hline \vdots & & & 0 & & \vdots \\ & & & & \ddots & \\ 0 & & & \cdots & & 0 \end{array} \right] \mathbf{E}^H \\ + \mathbf{E} & \left[\begin{array}{cccc} \delta^2 & 0 & \cdots & 0 \\ 0 & \delta^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \delta^2 \end{array} \right] \mathbf{E}^H. \end{aligned} \tag{2.103}$$

If the SNR between $s_n(t)$ and $n_m(t)$ is sufficiently high, the eigenvalues can be classified into two groups depending on their size as shown in Fig. 2.43. The eigenvectors corresponding to the largest N eigenvalues are the normalized orthogonal bases of the subspace called *Signal Subspace* $\mathcal{S}$. The rest of eigenvectors span the *Noise Subspace* $\mathcal{N}$, which is the orth-complement space of $\mathcal{S}$.

$$\mathcal{S} = span\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_N\} \tag{2.104}$$

$$\mathcal{N} = span\{\mathbf{e}_{N+1}, \mathbf{e}_{N+2}, \cdots, \mathbf{e}_M\} \tag{2.105}$$

$$\mathcal{S} \perp \mathcal{N} \tag{2.106}$$

Furthermore, because the steering vectors span the signal subspace as far as they are linearly independent,

$$\mathbf{s}_n \in \mathcal{S}, \tag{2.107}$$

thus, $\mathbf{s}_n$ consists of the linear combination of the eigenvectors in $\mathcal{S}$ using arbitrary scalar complex α_l

$$\mathbf{s}_n = \sum_{l=1}^{N} \alpha_l \mathbf{e}_l. \tag{2.108}$$

From these facts, each steering vector corresponding to one of the speaker directions is orthogonal to the eigenvectors in $\mathcal{N}$. Consequently, we estimate the speaker direction $\overline{\theta}$ by the following MUSIC spectrum, defined as

$$\overline{\theta} = \arg\max_{\theta} P_{MUSIC}(\theta) \tag{2.109}$$

$$P_{MUSIC}(\theta) = \frac{\mathbf{s}^H(\theta)\mathbf{s}(\theta)}{\sum_{l=N+1}^{M} |\mathbf{e}_l^H \mathbf{s}(\theta)|^2}. \tag{2.110}$$

The most major advantage of MUSIC is its spatial resolution, because it does not depend on the beamspace-based discriminability. One of its disadvantages is that the signal coherence which is caused by the reverberation conditions is detrimental to the performance due to the covariance matrix being rank-deficient. Another demerit is the cost of the calculation load of the eigenvalue decomposition.

Coherence signal subspace for broadband signal

To apply MUSIC to the direction estimation of broadband signals, a preliminary process is required because the MUSIC is originally developed for the DOA estimation of narrowband signals. Coherence Signal Subspace (CSS) method [34] is one major solution to this subject that has been adopted in many conventional studies of speaker direction estimation [35][36][37].

To deal with the signal as narrowband, we first calculate the Fourier transform of the array input signal $X_m(\omega)$. In the CSS method, we employ the average of the covariance matrices over the frequency domain whose index range of $[k_0 - K/2, k_0 + K/2]$ with the centre frequency ω_{k0},

$$\mathbf{R}_{CSS} = \sum_{k=k_0-K/2}^{k=k_0+K/2} \mathbf{T}_k \mathbf{R}_k \mathbf{T}_k^H, \tag{2.111}$$

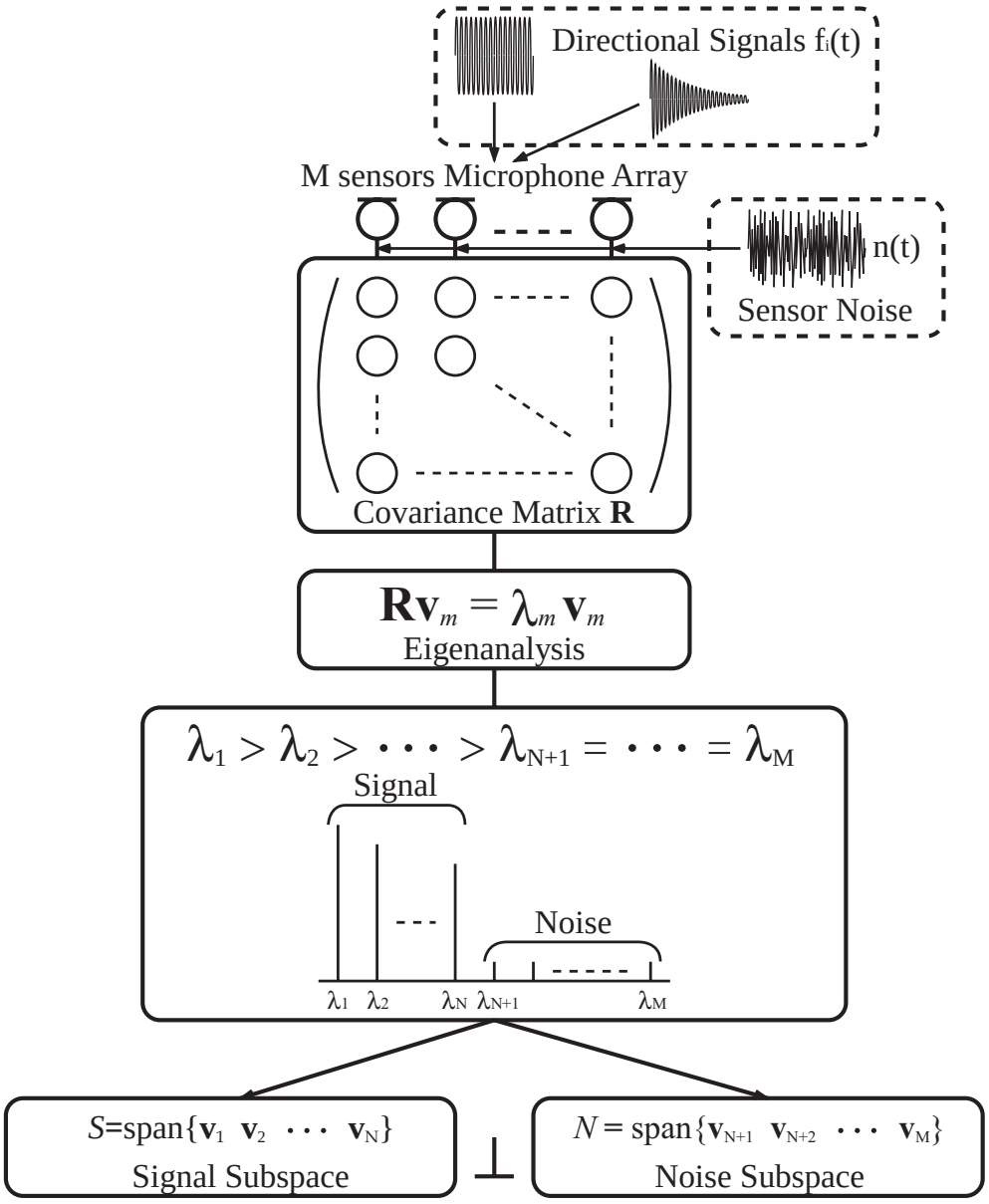

Figure 2.43: Features in subspace of the array covariance matrix

where $\mathbf{T}_k$ is termed *focusing matrix*, and $\mathbf{R}_k$ is the signal covariance matrix in k-th frequency bin. The focusing matrix is used to adjust the direction of the steering vector $\{\mathbf{s}_1(\omega_k), \cdots, \mathbf{s}_N(\omega_k)\}$ that is part of $\mathbf{R}_k$, in different frequencies. Applying the transformation using the focusing matrix, $\mathbf{R}_k$ in different frequencies can be averaged with the signal subspace structure being preserved. Hence, the MUSIC is applicable to the broadband speech signals by performing the process stated

in the previous section to this frequency-averaged covariance matrix.

For the determination of focusing matrix, several procedures have been proposed and examined from the statistical point of view [38][39][40][41]. The most typical method is the RSS focusing matrix [38] derived by using the following singular value decomposition.

$$\mathbf{T}_k = \mathbf{V}_k \mathbf{U}_k^H \tag{2.112}$$

$$\text{subject to} \quad \mathbf{U}_k \Sigma_k \mathbf{V}_k^H = \mathbf{s}_k \mathbf{s}_{k_0}^H \tag{2.113}$$

For the steering vector in Eq.(2.113), we should know the directions of the sound sources, yet our purpose is to estimate unknown source directions. To solve this discrepancy, we usually perform a rough DOA estimation beforehand, but its accuracy affects the final estimation result.

2.7 Speaker direction tracking

For an advanced subject of direction estimation, the tracking of speaker direction is important. One solution for this subject is repeatedly performing the direction estimation methods mentioned in the above section [42], but they often meet the problem of heavy calculation load. Because the real-time process is desirable in most cases where the speaker direction tracking is required, such complicated calculation obstructs the method to be adopted for the speaker tracking system. Some methods have been designed for the speaker direction tracking problem. In the following of this section, we explain about two major schemes.

2.7.1 Adaptive beamforming method

One major stream of the speaker direction tracking is the adaptive-beamformer-based method [43][44][45]. As we stated in Sec. 2.5.2, the adaptive beamformer steers the nulls to the sound source direction by adaptively modifying its weights depending on the received signals. So by calculating the beampattern and searching the null-steered direction at every weight update, we can derive the direction of sound source. Fig. 2.44 shows the basic structure of the speaker tracking based on adaptive beamformer. The direction estimate at i-th time instant (or frame)

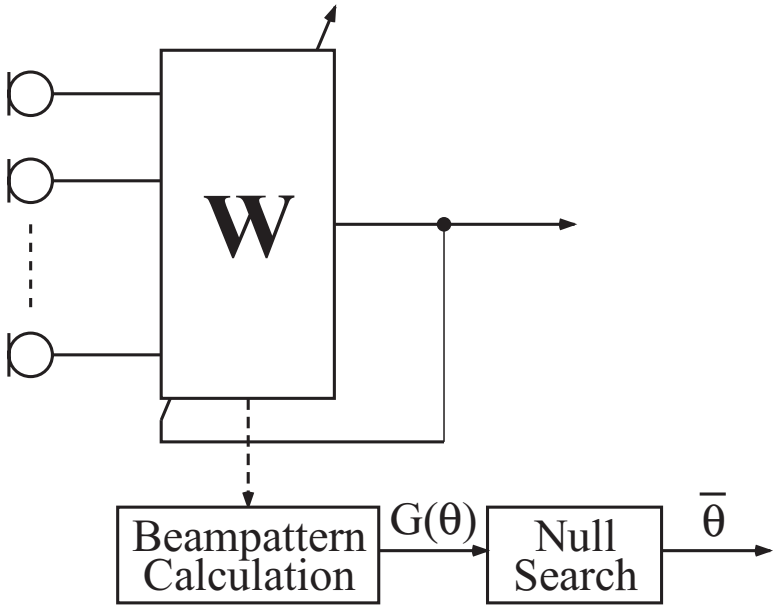

Figure 2.44: Structure of adaptive-beamformer-based speaker direction tracking

$\overline{\theta_i}$ is derived from the null steered direction of beamforming weight vector $\mathbf{W}_i$.

$$\overline{\theta_i} = \arg\min_{\theta} G(\theta, \mathbf{W}_i) \tag{2.114}$$

$$\text{subject to} \quad G(\theta, \mathbf{W}_i) = \mathbf{s}^H(\theta)\mathbf{W}_i \tag{2.115}$$

For the types of adaptive beamformer, several methods have been proposed. For example, Nokas *et al.* [43] proposed to use the LCMV, and Nagata *et al.* [44] adopted the GSC for the adaptive beamformer.

Although this kind of methods estimates the direction from the repeatedly updated beamformer weights, it still suffers from an excessive calculation load because the beampattern calculation is necessary at every update of $\mathbf{W}_i$. Furthermore, the resolution of the beampattern calculation affects the accuracy of the estimated direction. Another technique to avoid these problems that updates the direction estimate directly without beampattern calculation is introduced in the next section.

2.7.2 Direct update of null steering direction

A method for speaker direction tracking without beampattern calculation is proposed by Kawakami *et al.* [45]. In the method as shown in Fig. 2.45, they used the null steering beamformer and estimated the direction by updating the compensation delay τ_i to make the power of null steering beamformer output to be minimized. Because the power of beamformer output is given by

$$\begin{aligned} J &= E\left[\sum_{\omega}|Y(\omega)|^2\right] \\ &= E\left[\sum_{\omega}\left|X_1(\omega)-X_2(\omega)e^{-j\omega\tau}\right|^2\right] \\ &= E\left[\sum_{\omega}\left(|X_1(\omega)|^2+|X_2(\omega)|^2-2\Re\left[X_1(\omega)X_2^*(\omega)e^{j\omega\tau}\right]\right)\right], \end{aligned} \tag{2.116}$$

we derive the optimal τ that gives the minimum J by the steepest descent method.

$$\tau_{i+1} = \tau_i - \mu\frac{\partial J_i}{\partial \tau_i} \tag{2.117}$$

$$\frac{\partial J_i}{\partial \tau_i} = E\left[\sum_{\omega}\left(2\omega\Im\left[X_1(\omega)X_2^*(\omega)e^{j\omega\tau_i}\right]\right)\right], \tag{2.118}$$

where μ is the stepsize and $E[\cdot]$ denotes the frame average. From τ_i, the direction estimate is given by

$$\overline{\theta_i} = \sin^{-1}\left(\frac{c}{d}\tau_i\right). \tag{2.119}$$

2.8 Summary of Chapter 2

This chapter summarises the overview of conventional technologies used in the speech signal processing to help readers to understand the works mentioned in the following part of this dissertation. In Sec. 2.2, we first explained the features of speech signals from the generating mechanism point of view, and referred to our viewpoint of their classification by the three signal domains. The following Sec. 2.3 mentioned some time-frequency analysis methods that can represent the

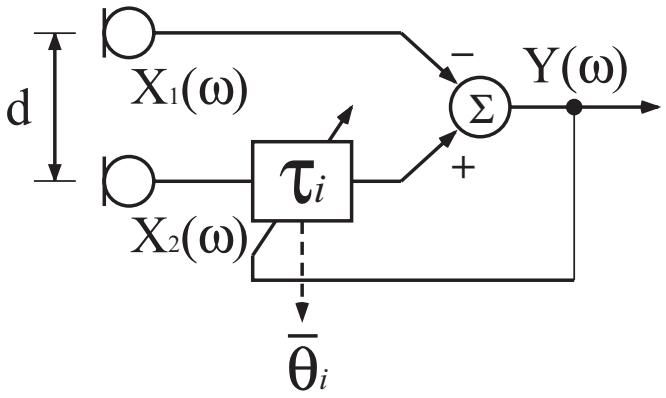

Figure 2.45: Speaker direction tracking by the minimization of null-beamformer output

speech signal features in temporal and spectral domains. Spending most of the latter part of this chapter, we explained the principles and some conventional studies on the use of the microphone array signal processing. In Sec. 2.4, the problems with, and fundamental principles of, microphone array signal processing are stated. Following in Sec. 2.5 and Sec. 2.6, the summaries of previous work relating to the two major functions of microphone array, namely, beamforming and direction estimation, are given respectively. Finally, some studies of speaker direction tracking are introduced in Sec. 2.7.

Chapter 3

Voice Activity Detection with Array Signal Processing in the Wavelet Domain [46][47][48]

3.1 Introduction

As mentioned in Chap. 1, detection of voice activity segments helps in many aspects of speech signal processing and works to improve their performance. In the area of Voice Activity Detection (VAD), many conventional methods have been proposed. They are classified into two groups by the number of microphones used in the system. The first group uses a single microphone. It utilizes the features of speech signal in the temporal or frequency domain, such as transition of signal power [2], zero crossing rate [2], and harmonic structure in its spectrum [49][50][51]. The second group uses the microphone array to extract spatial features, such as signal source directionality [52] and speaker's location [53][54]. The VAD method proposed in this chapter belongs to the latter group. This method exploits the speech signal features in all three signal domains as stated in Chap. 2.

An important VAD problem is to discriminate speech from various interference sounds. In indoor environments where the interface system is usually settled, the following types of interference sounds exist; (1)stationary interference without directionality, such as sensor noise, (2)stationary interference with directionality, such as noise generated by air conditioner etc., and (3)nonstationary interference

with directionality such as the sound of closing door, etc. Most of the conventional methods [2][49][50][51][52] consider the interferences of types (1) and (2).

The array-based VAD methods dealing with nonstationary interferences are studied by Kaneda [53] and Kiyohara *et al.* [54]. These VAD systems are implemented with the speech emphasizing microphone array system "AMNOR (Adaptive Microphone array for NOise Reduction)" [55]. The detector uses the direction difference between speech and interference. There are two aspects to be improved in these methods. One is the spatial resolution in AMNOR. Another is to make the detector capable to detect unvoiced sounds. In this research, a new VAD system for treating both voiced and unvoiced sounds is proposed, and we demonstrate that the method keeps its discriminability even if a nonstationary interference arrives from close direction to that of the target speech.

The first idea in the proposed method is the subband decomposition to capture the specific time-frequency features of speech signal. According to the spectral distribution difference between voiced and unvoiced sounds, particular band decomposition is used. The second idea is to use the directionality and direction of arrival (DOA). The directionality is defined when signals received at two spatially apart points are mutually correlated. The spatial information can be estimated through the eigenspace analysis in the wavelet domain.

The following Sec. 3.2 restates the speech features found in the temporal, spectral and spatial domains, and then explains the problem. The proposed method is described in Sec. 3.3, and simulation results are shown in Sec. 3.4. With some comments, we conclude this chapter in Sec. 3.5.

3.2 Speech signal features

Speech signal features can be categorized into three groups as summarised in Table 3.1. Generally, a speech signal is classified into voiced and unvoiced sounds according to the way of utterance, therefore, their temporal and spectral features are quite different. Now, we briefly review these features to explain our proposed method. For measuring speech quality, we use the SIR and SNR defined as a power ratio of target speech to that of the nonstationary interference and the sensor noise, respectively.

Table 3.1: Features of speech signal

	Voiced Sound	Unvoiced Sound
Temporal Features	Stationarity in a short period ($30-40$ms) Low power	High localization
Spectral Features	Harmonic Structure Power concentration in lower band	Power spread over wide band
Spatial Features	Existence of directionality Specific DOA	

1. **Temporal Features**

   One of the temporal features of speech signal is its power localization. That is, speech signal is conceived as a non-stationary signal. Nevertheless, the stationarity can be assumed for the voiced sound in a short period which is usually known about 30–40ms, while the unvoiced sound is highly localized and its power attenuates promptly. Furthermore, the SNR in unvoiced sound segment would be lower than that in voiced sound segment.

2. **Spectral Features**

   The power spectrum of a speech signal spreads over wide band and its distribution typically depends on such factors as speaker, phoneme, etc. The spectrum shape also largely depends on whether the sound is voiced or unvoiced. The voiced sound inherently has harmonic structure, that is, its power mostly concentrates on the specific harmonics in the lower frequency band. Each harmonic component is assumed to be a narrowband signal. Additionally, fundamental frequency is time-varying between the limited range of $80-400$Hz [2].

   In contrast, the spectrum of unvoiced sound spreads over wide band. The conventional VAD methods [49][50][51] discriminate into voiced and unvoiced by checking the existence of harmonic structure.

3. **Spatial Features**

   The spatial feature of speech is another important factor in this study. We

assume the speech signal emanates from a predetermined direction such as broadside. In addition, it is important to know whether the signal has directionality or not. In the method of [52], speech is characterized by the directionality, thus, we cannot distinguish directional interferences from speech.

While the VAD methods [2][49][50][51][52] use one of the above three features, Kaneda [53] combines the spatial and temporal features for VAD. In this research, we integrate the attributes in all three domains by considering the differences between voiced and unvoiced sounds. Figure 3.1 shows a diagram explaining our idea. Three axes represent signal features in the temporal, spectral and spatial domains respectively. The shadowed region in Fig. 3.1 indicates sub-domain by which we can determine the given signal to be speech.

At first, we set the following assumptions.

1. The DOA of desired speech signal is known *a priori*

2. There is no more than one nonstationary interference simultaneously

The first assumption is not too specific to lose generality. In adaptive beamformer design, the DOA of desired signal is usually set to be known and the null is adaptively steered to the unknown direction of interference.

The wavelet packet and eigenspace analysis are two significant tools in the proposed method. Wavelet packet decomposition in the time-frequency domain provides high frequency resolution for voiced sound and high temporal resolution for unvoiced sound. The frame (time interval) division and subband decomposition are discussed in the next chapter.

The eigenspace analysis is applied to the spatial covariance matrices for each subband signal. The first step is to check directionality. Then the direction angle of directive sound is estimated to determine whether it is the same as the supposed direction of the desired speech sound or not.

Eventually, VAD by integrating the speech features in the three signal domains is achieved by array signal processing on the wavelet coefficients. Figure 3.2 is the flow diagram of the proposed system. In the following section, we describe each procedure in detail.

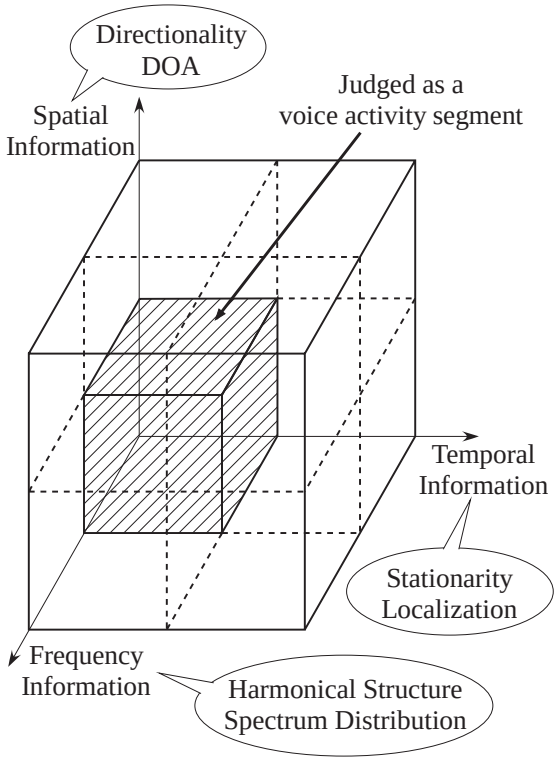

Figure 3.1: Concept of the proposed method

3.3 Proposed method

Figure 3.3 shows the proposed procedure and the data flow, each of which is described in this section.

3.3.1 Input signal modelling

From the assumptions stated in the Sec. 3.2, the signal received by the i-th sensor of the M sensors microphone array is modelled as

$$\hat{x}_i(n) = s(n - \tau_i^{(s)}) + d(n - \hat{\tau}_i^{(d)}) + \hat{n}_i(n), \tag{3.1}$$

where $s(n)$, $d(n)$ and $\hat{n}_i(n)$ are the speech whose active interval should be detected, interference and the sensor noise respectively, and $\tau_i^{(\mathrm{x})}$ indicates the delay at i-th sensor relating to each signal x. The length of the input signal $\hat{x}_i(n)$, to which the following VAD procedure is applied, is assumed to be nearly one word

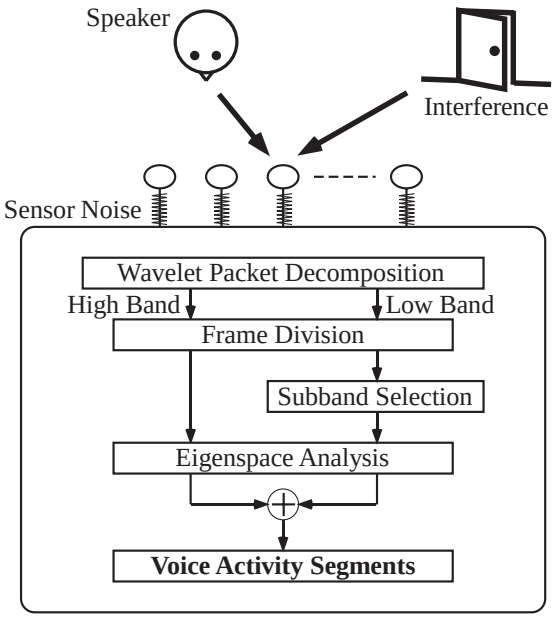

Figure 3.2: Flow diagram of the entire proposed method

length (approximately $1-2$ second). For each $\hat{x}_i(n)$, delay compensation with respect to $\tau_i^{(s)}$ yields

$$x_i(n) = s(n) + d(n - \tau_i^{(d)}) + n_i(n). \tag{3.2}$$

This means that the mean value of all $x_i(n)$, namely,

$$\begin{aligned} c(n) &= \frac{1}{M}\sum_{i=1}^{M} x_i(n) \\ &= s(n) + \frac{1}{M}\sum_{i=1}^{M}\{d(n - \tau_i^{(d)}) + n_i(n)\} \end{aligned} \tag{3.3}$$

performs as the delay-and-sum beamforming, and thus it enhances target speech.

3.3.2 Wavelet packet decomposition and frame division

At first, the wavelet packet decomposition is applied to $x_i(n)$ and $c(n)$. The delay compensated input signal $x_i(n)$ is transformed into subbands components $X_{i,\phi}(k_\phi)$ with the specific resolution explained in the following. The indices ϕ and k_ϕ mean the subband number and the sample point in the subband ϕ respectively.

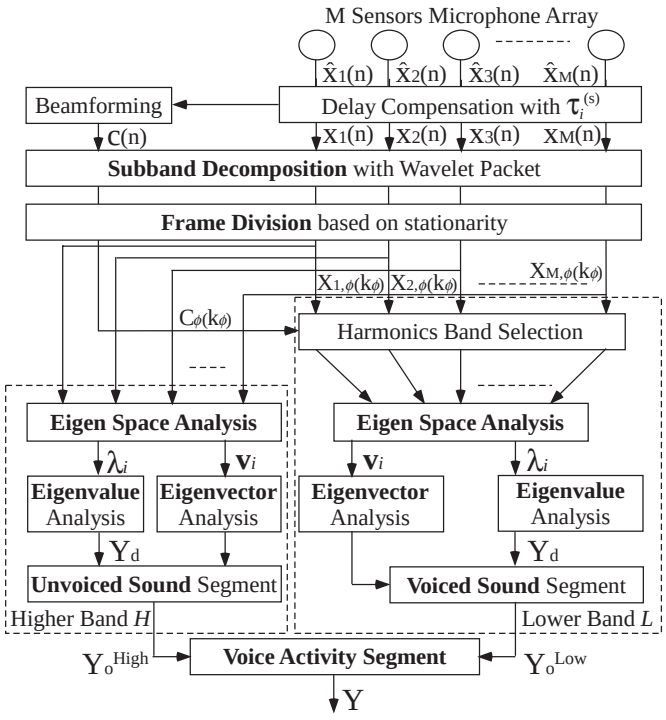

Figure 3.3: Data flow in the proposed method

Here the number of wavelet coefficients $X_{i,\phi}(k_\phi)$ in each subband is derived by calculating $K/2^{\beta_\phi}$ where K[sample] is the given input signal length and β_ϕ is the decomposition level.

For the tiling of time-frequency plane, a specific subband decomposition as illustrated in Fig. 3.4 is adopted. This decomposition is the reflection of the voiced and unvoiced speech characterization stated in Sec. 3.2. The first decomposition step is to separate the frequency band determined by the sampling frequency Fs into lower and higher bands by setting 2kHz as their boundary. The lower half band, we denote L, is used to extract voiced sounds according to its harmonic structure. Therefore, equi-bandwidth decomposition is performed by setting its bandwidth to be narrower than 80Hz. This decomposition is applied to prevent each band having more than one harmonic in it. In the higher half band, denoted as H, the octave band decomposition is adopted in order to detect unvoiced sound. The fast temporal fluctuation in unvoiced sounds will be detected in this case.

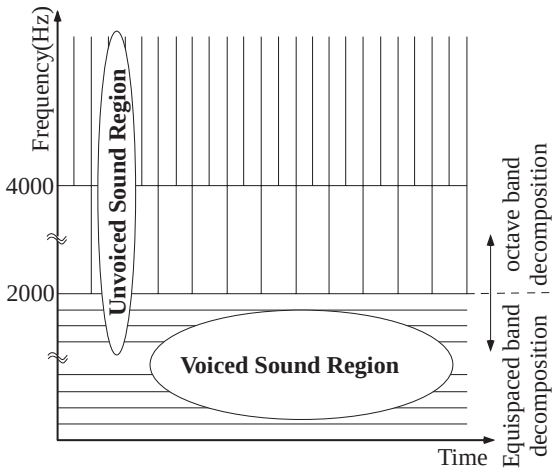

Figure 3.4: Time frequency resolution of proposed method

The obtained wavelet packet coefficients $X_{i,\phi}(k_\phi)$ are furthermore divided into frames. That is, we have $X^j_{i,\phi}(k_\phi)$ for j-th frame in subband ϕ. In the proposed VAD method, each voice activity decision is performed to every frame individually. The frame length is set to be sufficient to ensure the stationarity. In our case, 32ms and 8ms are taken for L and H, respectively. For simplicity, the frame number j is omitted in the following statement.

3.3.3 Determination of subbands containing major harmonic components in the lower band L

The procedure in this section is concerned with harmonic structure detection and the determination of subbands containing their harmonics in L. By the subband decomposition in L, harmonic components of voiced sound are supposed to be separately located in some subbands. The voiced signal stationarity within a frame is used to extract these subbands.

The extraction procedure is performed on the wavelet packet coefficients $C_\phi(k_\phi)$ of the delay-sum beamformer output $c(n)$ derived by Eq.(3.3). The wavelet domain power function on (ϕ, k_ϕ)-plane, as illustrated in Fig. 3.5(a), is utilized in subband determination. In Fig. 3.5(a), meshed portions illustrate dominated power component positions for speech and that of interference in the (ϕ, j)-plane. Each vertically divided section shows an equi-bandwidth subband in L, and the

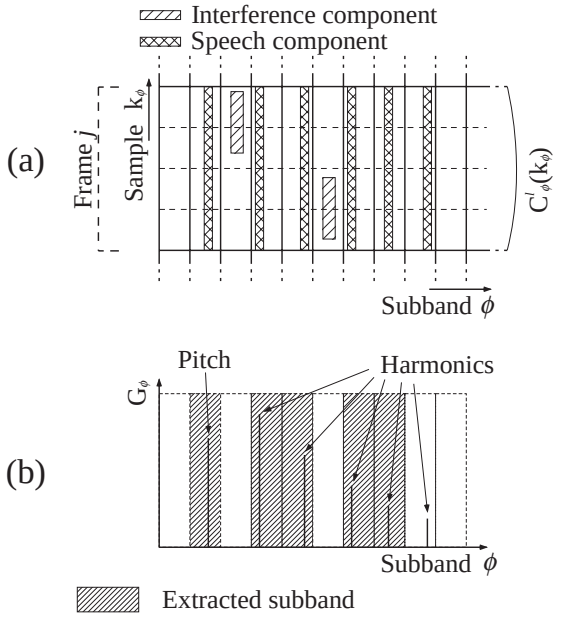

Figure 3.5: Harmonics band extraction with geometric mean

vertical axis shows sample point k_ϕ in the j-th frame. As Fig. 3.5(a) shows, time duration of dominant speech power would be larger than that of interference. We, therefore introduce

$$G_\phi = \sqrt{\prod_{k_\phi} |C_\phi(k_\phi)|^2} \tag{3.4}$$

which is the geometric mean of $|C_\phi|^2$ with respect to k_ϕ in each frame. Then, we extract the subband $\phi_p \in L$ which has p-th largest G_ϕ and define the set of the first N subbands by

$$\Phi \stackrel{\triangle}{=} \{\phi_p \mid p = 1, 2, \cdots, N\} \tag{3.5}$$

Assuming that the fundamental frequency should exist in the frequency band less than 400Hz, at least five harmonic frequencies must exist in L. Figure 3.5(b) illustrates an example of G_ϕ and extracted five subbands (meshed sections) according to Fig. 3.5(a). In the following sections, the method performs the array signal processing to the coefficients of these selected subbands $X_{i,\phi}(k_\phi)$ ($\phi \in \Phi$).

3.3.4 Eigenspace analysis for narrowband array signals

Before discussing the eigenspace analysis of subband signals, some mathematical preliminaries on narrowband array signal [8] are described and some results by which our VAD is obtained are derived. The features in the eigenspace of covariance matrix for narrowband array signal are summarised below.

In the case that M sensors array receives plane waves from m independent sources, consider the covariance matrix $\mathbf{R} \stackrel{\triangle}{=} E\left[\mathbf{x}\mathbf{x}^H\right]$ of received signal vector

$$\mathbf{x}(n) = \sum_{l=1}^{m} f_l(n)\mathbf{s}_l + \mathbf{n}(n), \tag{3.6}$$

where $\mathbf{s}_l$ is the direction vector of a signal $f_l(n)$.

$$\mathbf{s}_l = \left[\begin{array}{cccc} e^{-j\psi_1^l} & e^{-j\psi_2^l} & \cdots & e^{-j\psi_M^l} \end{array}\right]^T, \tag{3.7}$$

and ψ_i^l is the phase delay of $f_l(n)$ at i-th sensor. The $\mathbf{R}$ is decomposed into the following product of matrices by using eigenvalue λ_i of $\mathbf{R}$, and the corresponding eigenvector $\mathbf{v}_i$.

$$\begin{aligned} \mathbf{R} &= [\mathbf{v}_1 \quad \cdots \quad \mathbf{v}_M] \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_M \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^H \\ \mathbf{v}_2^H \\ \vdots \\ \mathbf{v}_M^H \end{bmatrix} \\ &\text{(subject to} \quad \lambda_1 > \lambda_2 > \cdots > \lambda_M) \end{aligned} \tag{3.8}$$

Theoretically, there is the following relation among the eigenvalues.

$$\lambda_1 > \lambda_2 > \cdots > \lambda_m > \lambda_{m+1} = \cdots = \lambda_M \tag{3.9}$$

Namely, each eigenvalue λ_i is classified into two groups, which correspond to signal subspace $\mathcal{S}$ and noise subspace $\mathcal{N}$ satisfying orthogonality $\mathcal{S} \perp \mathcal{N}$. Additionally, the corresponding set of eigenvectors $\mathbf{v}_i$ becomes a basis of each subspace.

$$\mathcal{S} = span\left\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m\right\} \tag{3.10}$$

$$\mathcal{N} = span\left\{\mathbf{v}_{m+1}, \mathbf{v}_{m+2}, \cdots, \mathbf{v}_M\right\} \tag{3.11}$$

The direction vector $\mathbf{s}_l (l = 1, 2, \cdots, m)$ satisfies

$$\mathbf{s}_l \in \mathcal{S} \tag{3.12}$$

The first subject what we should discuss is how to check the existence of directional signals. The second subject is the DOA of received directional signal. Since the delay compensated array signals $x_i(n)(i = 1, 2, \cdots, M)$ are in-phase with respect to the direction of target speech, the direction vector of desired speech sound is defined as

$$\mathbf{s}_0 \stackrel{\triangle}{=} \left[\begin{array}{cccc} 1 & 1 & \cdots & 1 \end{array} \right]^T \tag{3.13}$$

In addition, from the assumption that more than two directional signals do not exist simultaneously, there are two possible cases given in the following.

[Case 1: One directional signal]

In this case, it is required to check whether the DOA of the directional signal is a desired one or not. Let us substitute $m = 1$ into Eq.(3.6)–Eq.(3.11). We then describe the VAD condition as follow.

The direction vector $\mathbf{s}_1$ satisfies $\mathbf{s}_1 = \mathbf{s}_0$ if and only if the $M-1$ orthogonality conditions $\mathbf{s}_0 \perp \mathbf{v}_i \ \ (i = 2, 3, \cdots, M)$ are satisfied.

[Case 2: Two directional signals]

Let us define

$$P_l = E\left[|f_l(n)|^2\right] \tag{3.14}$$

$$N = E\left[|n(n)|^2\right], \tag{3.15}$$

and by substituting $m = 2$ in Eq.(3.6)–Eq.(3.11), we have

$$\mathbf{R} = P_1\mathbf{s}_1\mathbf{s}_1^H + P_2\mathbf{s}_2\mathbf{s}_2^H + N\mathbf{I}. \tag{3.16}$$

Now, we assume $\mathbf{s}_1 = \mathbf{s}_0$ without loss of generality. We define the VAD problem in this case is to check the condition $P_1 > P_2$. That is, voice activity segment (VAS) in the Case 2 is the segment in which desired speech power (P_1) is more dominant than interference power (P_2). We can obtain the following property through the theory of principle component analysis (PCA) for two sensors array case, and it is verified by computer simulation.

[Property]

In the Case 2 except for $\mathbf{s}_1 \approx \mathbf{s}_2$, the $M-1$ conditions $\mathbf{s}_0 \perp \mathbf{v}_i \ \ (i = 2, 3, \cdots, M)$ imply the condition $P_1 > P_2$.

The subspace analysis of $\mathbf{R}$ is perceived as PCA for the array received data

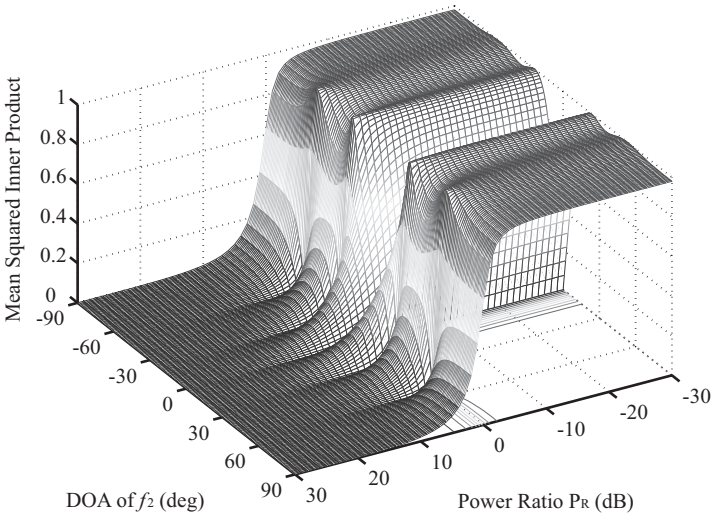

Figure 3.6: Variation of mean squared inner product

$x_i(n)$, that $\mathbf{v}_i$ and λ_i correspond to the i-th component vector and component score, respectively. While $f_1(n)$ is dominant, i.e. $P_1 \gg P_2$, the correlation distribution is dominated by the direction vector $\mathbf{s}_0$. Because the PCA gives the first component vector that maximizes the variance of the component scores, direction of $\mathbf{v}_1$ comes up to that of $\mathbf{s}_1$ for the case $P_1 \gg P_2$. An example is shown in Appendix A. Furthermore, Fig. 3.6 shows the simulation result for the mean squared inner product, namely $\frac{1}{M-1}\sum_{i=2}^{M}|\mathbf{s}_0^H \cdot \mathbf{v}_i|^2$, with respect to the power ratio $P_R = 10\log_{10}\frac{P_1}{P_2}$[dB] and the DOA ($[-90^\circ\ 90^\circ]$) of signal $f_2(n)$. The simulation result also ensures the Property except for the DOA of $f_2(n)$ being around 0 deg, which means both speech and interference arrive from almost the same direction. This is not a crucial problem because our proposed method still keeps its VAD ability by using the temporal and spectral features.

Now we can conclude this section by describing the VAD criterion for two cases as

$$\frac{1}{M-1}\sum_{i=2}^{M}|\mathbf{s}_0^H \cdot \mathbf{v}_i|^2 = 0 \tag{3.17}$$

3.3.5 Eigen decomposition of subband covariance matrix

The results obtained in the preceding section are applied to the signal $X_{i,\phi}(k_\phi)$. The eigen decomposition is performed to each subband covariance matrix $\mathbf{R}_\phi$ generated from signal vector $\mathbf{X}_\phi(k_\phi)$ given by Eq.(3.18) and Eq.(3.20).

$$\mathbf{R}_\phi = \sum_{k_\phi} \mathbf{X}_\phi(k_\phi)\mathbf{X}_\phi(k_\phi)^H \tag{3.18}$$

$$\mathbf{X}_\phi(k_\phi) = \left[\begin{array}{cccc} X_{1,\phi} & X_{2,\phi} & \cdots & X_{M,\phi} \end{array} \right]^T \tag{3.19}$$

$$\mathbf{R}_\phi = \mathbf{V}_\phi \mathbf{\Lambda}_\phi \mathbf{V}_\phi \tag{3.20}$$

$\mathbf{V}_\phi$ and $\mathbf{\Lambda}_\phi$ are the matrices consisting of normalized eigenvector $\mathbf{v}_{i,\phi}$ and corresponding eigenvalue $\lambda_{i,\phi}$ given by

$$\mathbf{V}_\phi = \left[\begin{array}{cccc} \mathbf{v}_{1,\phi} & \mathbf{v}_{2,\phi} & \cdots & \mathbf{v}_{M,\phi} \end{array} \right]^T \tag{3.21}$$

$$\mathbf{\Lambda}_\phi = \mathrm{diag} \left[\begin{array}{cccc} \lambda_{1,\phi} & \lambda_{2,\phi} & \cdots & \lambda_{M,\phi} \end{array} \right]. \tag{3.22}$$

3.3.6 Detection of directional signal segment

The directionality is used as an inherent character of speech in our VAD system. The detection of time interval in which directional signal exists (we call it directional signal segment (DSS)) is treated here. The directionality of a signal is measured by the eigenvalue distribution of covariance matrix $\mathbf{R}_\phi$ ($\phi \in \Phi$) as discussed in Sec. 3.3.4.

The entropy of eigenvalue distribution $\{\lambda_i\}$ is introduced as

$$E = -\sum_{\phi \in \Phi} \sum_{i=1}^{M} p_\phi(\lambda_i) \log_M p_\phi(\lambda_i), \tag{3.23}$$

where $p_\phi(\lambda_i)$ is the normalized eigenvalue defined as

$$p_\phi(\lambda_i) = \frac{|\lambda_{i,\phi}|}{\sum_{i=1}^{M} |\lambda_{i,\phi}|}. \tag{3.24}$$

Applying a threshold E_{Th}, which is determined from the input SNR and number of sensors M, we detect the DSS at which Y_d, defined by Eq.(3.25), takes 1.

$$Y_d^{Low} = \begin{cases} 1 & \text{if} \quad E \leq E_{Th}^{Low} \\ 0 & \text{otherwise} \end{cases} \tag{3.25}$$

Note that the following process mentioned in Sec. 3.3.7 is applied only to the segments whose Y_d is 1.

3.3.7 Detection of signal segments from specific direction

Because the extracted DSS possibly contains not only desired speech segment but also directive interference segment, we specify the DOA to discriminate the desired speech from interfering signals.

As we derive the condition $\frac{1}{M-1}\sum_{i=2}^{M}|\mathbf{s}_0^H \cdot \mathbf{v}_i|^2 = 0$ for VAD criterion in Sec. 3.3.4, same criterion for the signal $X_{i,\phi}(k_\phi)$ is adopted. To investigate the orthogonality, we define the sum of mean squared inner product over subband set Φ of Eq.(3.5).

$$V = \frac{1}{M-1}\sum_{\phi\in\Phi}\sum_{i=2}^{M}|\mathbf{s}_0^H \cdot \mathbf{v}_{i,\phi}|^2 \tag{3.26}$$

Applying a positive threshold V_{Th} which is nearly 0, the desired signal segment is determined as the interval Y_o^{Low}, defined by Eq.(3.27), takes 1.

$$Y_o^{Low} = \begin{cases} 1 & \text{if} \quad V \le V_{Th}^{Low} \\ 0 & \text{otherwise} \end{cases} \tag{3.27}$$

3.3.8 VAD in the higher band H

As we can find from Fig. 3.3, the procedures described in Sec. 3.3.3–Sec. 3.3.7, except for the subband selection in Sec. 3.3.3, are basically applied to the higher band H as well. The subband covariance matrix used in the following two subsections is computed for the wavelet coefficients $X_{i,\phi}(k_\phi)$ ($\phi \in H$). To cope with the difference in the spectral features of voiced and unvoiced sounds and the segmentation in the wavelet decomposition, the procedure in H is different from that in L at the following aspects.

Thresholds E_{Th} and V_{Th} determination

Since unvoiced sound has relatively low power, the desired unvoiced speech component tends to be contaminated by the sensor noise in the higher band. On the

other hand, the narrowband assumption which ensures the partiality of the eigenvalue distribution described by Eq.(3.9) is no longer distinctly satisfied for unvoiced sound. It is known that the directional signal having wide band spectrum does not have distinct partiality in eigenvalue distribution [56]. This increases the normalized entropy E. The eigenvector feature mentioned in Sec. 3.3.7 also loses its distinct orthogonality due to the wide spread spectrum. For these reasons, the threshold E_{Th} of Eq.(3.25) and V_{Th} of Eq.(3.27) in the higher band H are set to be larger than that in the lower band to make it more sensitive to the low power signal.

Union of detection results from different subbands in H

Due to the octave band decomposition, each subband in the higher band H has different temporal resolution. This prevents us taking simple sum of subbands adopted in Eq.(3.23) and Eq.(3.26). So we have the desired signal segment Y_o^{High} by taking the union of the VAD interval satisfying $Y_{o,\phi}^{High} = 1$ in each subband.

$$E'_\phi = -\sum_{i=1}^{M} p_\phi(\lambda_i) \log_M p_\phi(\lambda_i) \quad (\phi \in H) \tag{3.28}$$

$$Y_d^{High} = \begin{cases} 1 & \text{if} \quad E'_\phi \le E_{Th}^{High} \\ 0 & \text{otherwise} \end{cases} \tag{3.29}$$

$$V'_\phi = \frac{1}{M-1} \sum_{i=2}^{M} \left| \mathbf{s}_0 \cdot \mathbf{v}_{i,\phi} \right|^2 \quad (\phi \in H) \tag{3.30}$$

$$Y_{o,\phi}^{High} = \begin{cases} 1 & \text{if} \quad V'_\phi \le V_{Th}^{High} \\ 0 & \text{otherwise} \end{cases} \tag{3.31}$$

$$Y_o^{High} = \bigcup_\phi \{ Y_{o,\phi}^{High} = 1 \} \quad (\phi \in H) \tag{3.32}$$

3.3.9 Voice activity segment detection

Finally, the VAS is derived as follows.

$$Y = \left\{ Y_o^{Low} = 1 \right\} \cup \left\{ Y_o^{High} = 1 \right\} \tag{3.33}$$

Note that the values of two thresholds, E_{Th} and V_{Th}, appeared in the proposed method are determined by the rule of thumb.

Table 3.2: Parameters in the simulation

Band (Hz)	Low $(0-2000)$	High $(2000-8000)$
M	5	
d [cm]	2	
Fs [Hz]	16000	
SNR [dB]	30	
SIR [dB]	0	
θ [deg]*	10	
Wavelet	4B spline	
β_ϕ	7	1 and 2
E_{Th}	0.005	0.9
V_{Th}	0.002	0.26

(* difference angle between target and directional interference)

3.4 Simulation results

This section shows some simulation results to evaluate the proposed method. In order to obtain input data, a microphone receives a speech and nonstationary interference individually. Then the microphone array input signals are virtually generated by delaying the signal with appropriate samples for each sensor according to the DOA, and white noise is added. We use a male speaker uttering "start" in Japanese ($K = 13000$) as the target signal, and clap sound as the interference. Furthermore we show some results utilizing several kinds of speech and interferences to prove the effectiveness of our method in the general environment. We also confirm that the proposed method keeps its ability even for directional interference as long as the input SIR is relatively high.

As a conventional method for comparison, we adopt Kaneda's method [53], whose AMNOR is replaced with delay-sum beamformer (BF) to make it an impartial comparison, because AMNOR is an integrated beamformer running with sophisticated adaptive algorithm. We also compare with [53] by substituting subspace analysis (SS) for AMNOR to reveal the effects of VAD in the wavelet domain.

Table 3.3: Position of the interference signal

	Dominated period (samples)
Fig. 3.7 (Case I)	$8100 - 9600$
Fig. 3.8 (Case II)	$6300 - 7800$
Fig. 3.12	$7000 - 8500$

3.4.1 Performance evaluation

We consider the following two cases, where, the speech and interference do not exist at the same time (case I), and where they exist simultaneously (case II). The parameters in these simulations are shown in Table 3.2. Figure 3.7 and Fig. 3.8 show the results of case I and II respectively. The position of the interference signal domination in each case is shown in Table 3.3, and the correct VAS is designated by the dotted lines.

As we can find in Fig. 3.7, though the conventional method fails to discriminate the interference period due to close direction of it (10°) from the target, the proposed method succeeds in discriminating it correctly. On the other hand, Fig. 3.8 shows that the conventional method detects the overlapped segment (around 6400 sample point) as a speech period, meanwhile the proposed method precisely omitted the overlapped segment with low SIR. Such high discrimination ability of the proposed method is due to the integrated use of the features in three signal domains.

Additionally, in both cases of I and II, the proposed method is able to detect the low power unvoiced sound segments such as the phoneme /s/ in this simulation. This ability is not realised by the conventional method, because it detects the VAS based on the power transition without considering power difference between voiced and unvoiced sounds.

3.4.2 Quantitative evaluation

To evaluate the performance quantitatively, we introduce two criteria, the correct detection rate of speech segment R_S and that of non-speech segment R_N. To

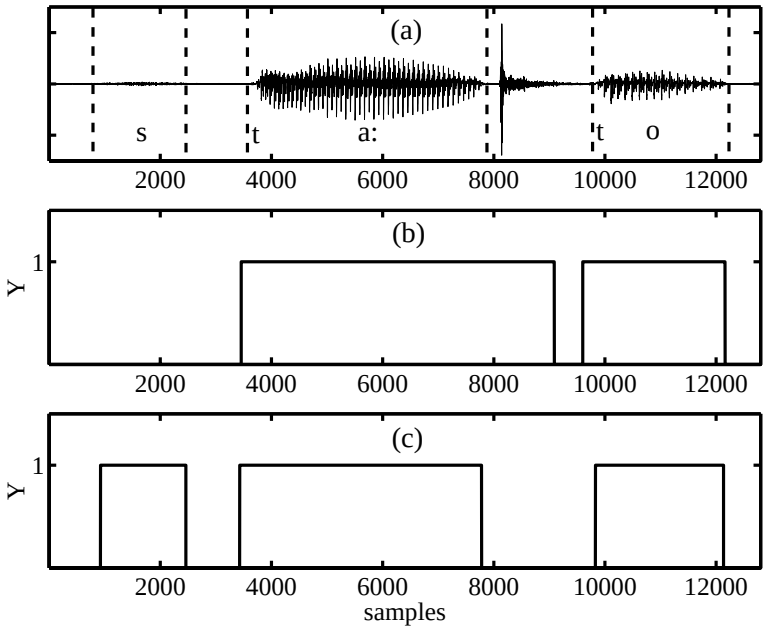

Figure 3.7: **Case I** : Voice and interference isolated exist : (a)Input signal (b)Result (Kaneda(SS)) (c)Result (Proposed)

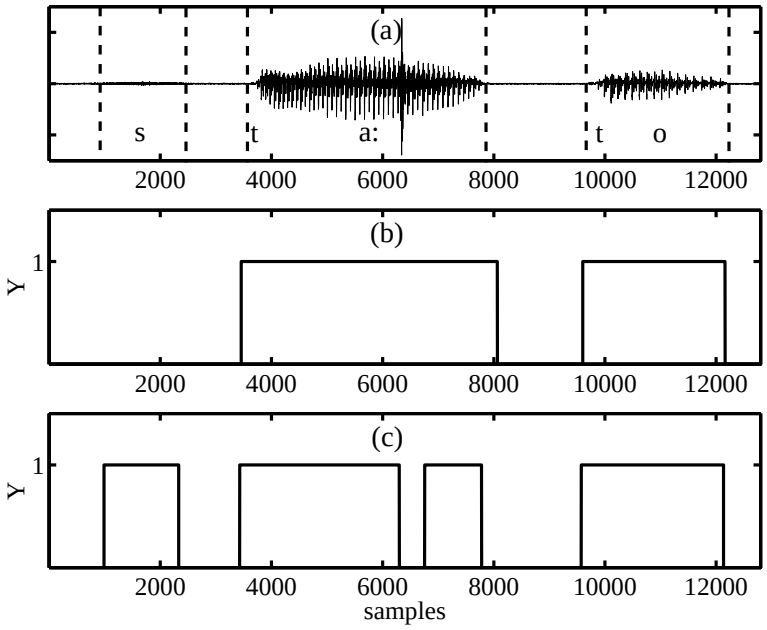

Figure 3.8: **Case II** : Voice and interference exist simultaneously

evaluate adequately, the quantitative evaluation is performed only for the case I. Each simulation is performed using the speech signal "start" with 9 people (4 male and 5 female) and the average is given in the results.(Fig. 3.9–Fig. 3.11)

As we can see in Fig. 3.9, the R_N of the conventional methods drop drastically when the interference source direction is close to the target speaker. This is caused by the misdetection of the interference segment. In contrast, the proposed method keeps its high discrimination ability even though the interference direction is not sufficiently far from the target signal direction. From this result, the robustness of the proposed method for the interference DOA is confirmed.

For various input SIRs, Fig. 3.10 shows that the proposed method keeps high detection rate even in a low SIR condition. Due to the utilization of power transition for detection makes the system to be sensitive to the SIR, the conventional methods lose their ability in the low SIR case. The proposed method uses the eigenspace feature, which is mostly independent on the SIR.

Finally the simulation results for various input signal SNRs are shown in Fig. 3.11. Referring to R_S, the proposed method decreases its performance as the SNR goes down. This is conceivable as a failure detection of unvoiced sound segments, because it may be covered with noise in low SNR condition. It is obviously impossible to detect such a signal without applying any noise reduction procedure, so we do not consider it as a fatal problem. On the other hand, even the proposed method as well as the conventional methods have a decreasing feature of R_N. The main subject for this deterioration is supposed to be the inappropriate value of the presettled thresholds.

3.4.3 Generality examination

In this section, we show some results applying different kinds of speech and interferences using the same parameters given in Table 3.2.

At first, we change the target speech. In the case shown in Fig. 3.12, the target signal is a female speech uttering "stop" in Japanese ($K = 12000$). It reveals that the detection rate seems to be independent of the speaker and sentence.

The result for a different interference is shown in Fig. 3.13(b). The interference is a sound given by placing one china plate over another [57]. Even though its spectrum is coloured relative to a clap sound, the proposed method succeeds in

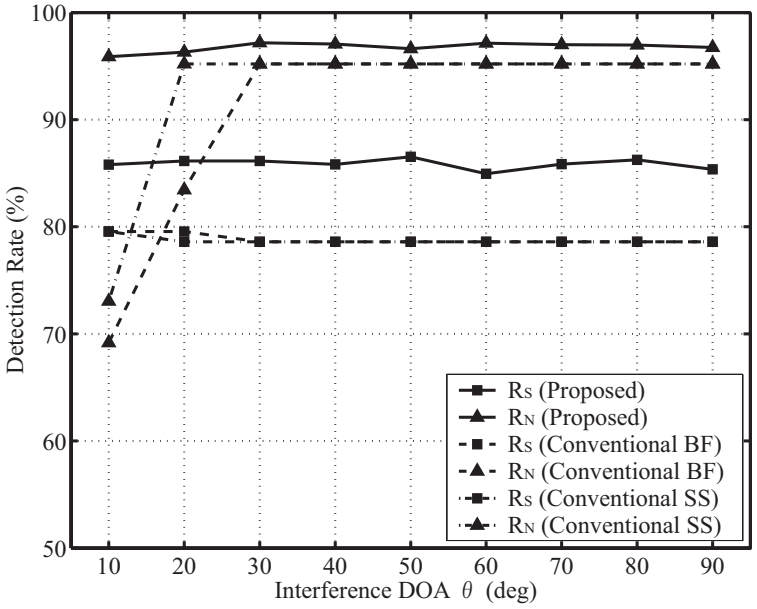

Figure 3.9: Detection rate for interference DOA changes (SNR:30dB, SIR:0dB)

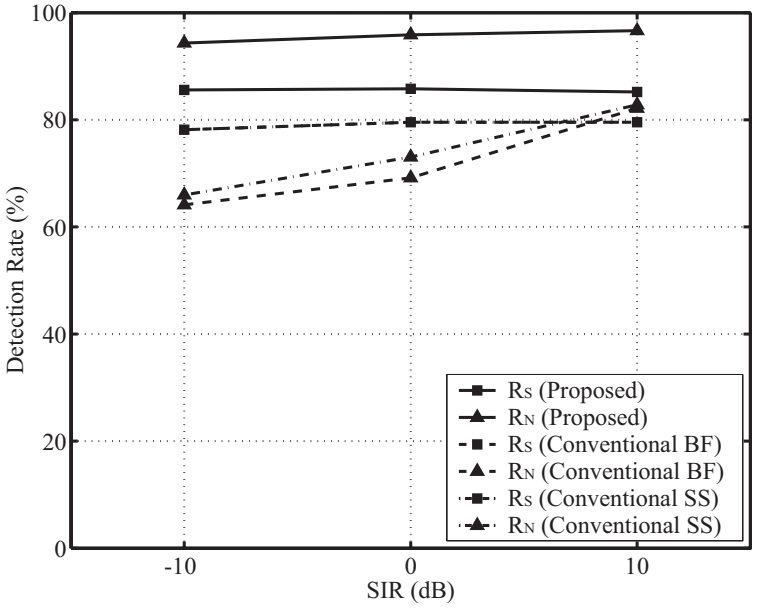

Figure 3.10: Detection rate for input SIR changes (DOA:10deg, SNR:30dB)

eliminating it.

Although the main purpose of the proposed method is to achieve high discriminability for nonstationary interference, it is desirable to keep its performance for stationary interference as well. To examine the discrimination capability for stationary interference, we performed a simulation using the sound of a hairdryer [57]. Figure 3.13(c) shows the result for an interference arriving from direction of 10° with 20dB in SIR. Proved from the result, our method keeps its discriminability even for stationary interference.

3.5 Conclusion of Chapter 3

We proposed a VAD method with array signal processing in the wavelet domain to utilize the temporal, spectral and spatial information in a desired speech signal. Applying the eigenspace analysis to the covariance matrix of array received signal, the proposed method keeps its high VAD precision even if the direction of the interference is close to that of the desired speech.

For future study, our method will be applied to a real room environment where the reflection and reverberation exist. They often degrade the system performance, because the eigenstructure of received signal is deteriorated. In addition, although we make some modifications in Sec. 3.3.8, the spatial resolution in the higher band H still goes down due to spread spectrum. Further improvement of the spatial information extraction in the higher band H is another future subject.

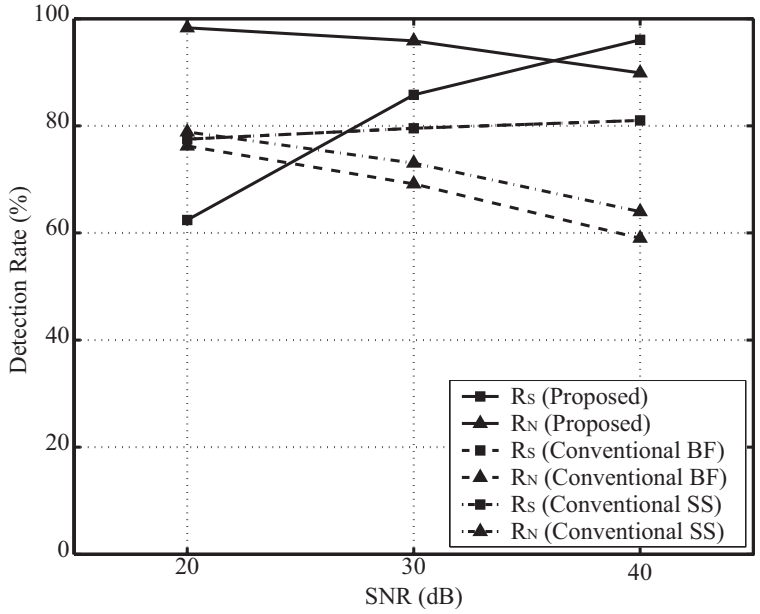

Figure 3.11: Detection rate for interference SNR changes (DOA:10deg, SIR:0dB)

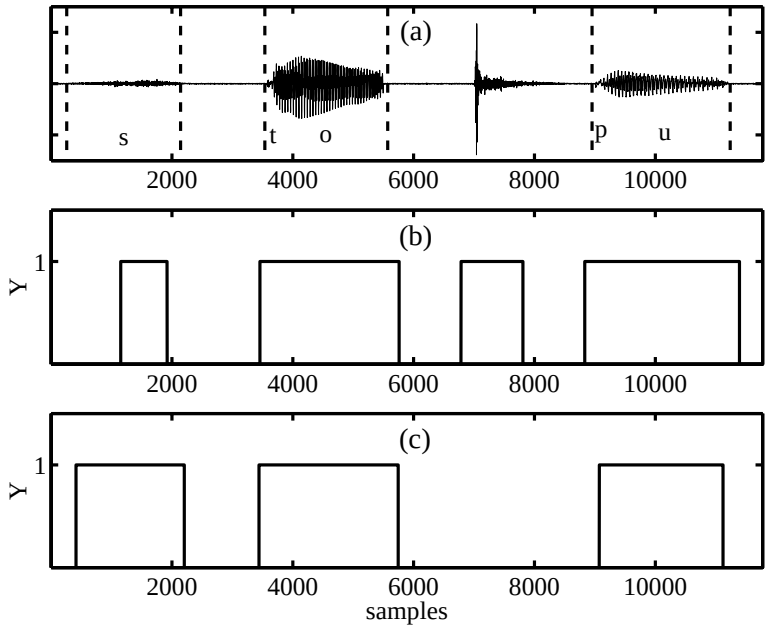

Figure 3.12: Result for a female speech : (a)Input Signal (b)Result (Kaneda(SS)) (c)Result (Proposed)

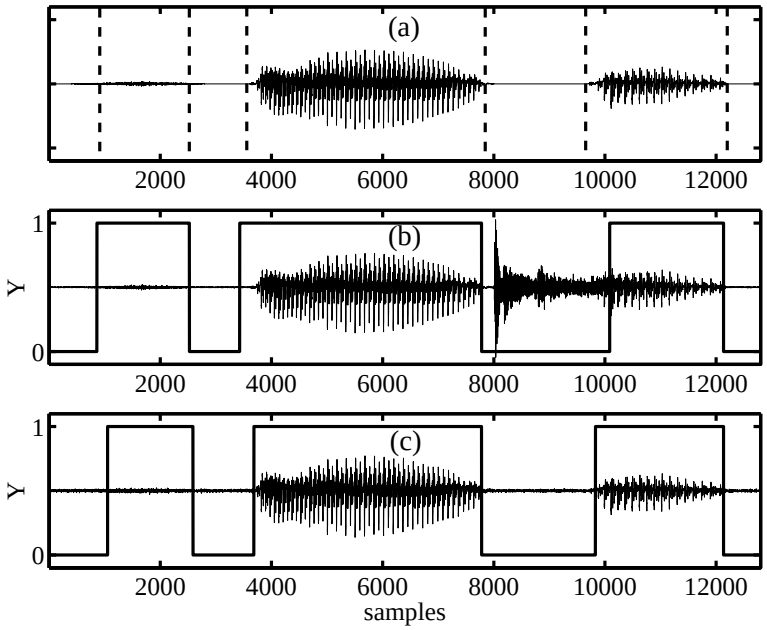

Figure 3.13: Results for different interferences : (a)Target Speech (b)Coloured Interference (China) (c)Stationary Interference (Hairdryer)

Chapter 4

Speaker Direction Estimation by the Integrated Use of Microphone Pairs

4.1 Introduction

Acquiring accurate speaker direction is an important element in speech signal processing. The following two chapters are dedicated to this topic. First of all, a new speaker direction estimation algorithm is proposed in this chapter.

Among the conventional direction estimation methods [58][59], the MUSIC (MUltiple SIgnal Classification) [60] is well-known for providing high spatial resolution. The MUSIC, in its original form, is applicable for narrowband input signal. For broadband signals, in which speech signal is included, many DOA estimation methods have been reported as well [27][34][37][61][62][63][64][65]. Wang *et al.* [34] proposed the Coherent Signal-Subspace (CSS) that enables the application of MUSIC to broadband signals. In the method, the components in several frequency bands are gathered together (this process is called *focusing*). The focusing requires rough DOA estimation in advance, and the pre-estimation error highly affects the final estimation result in practice.

The physical scale of array is another subject to be considered from a practical point of view. Generally, the performance of direction estimation, as well as that of interference rejection, is improved by increasing both the number of

sensors and the array aperture size. However, they are often restricted due to the limited physical size of the apparatus on which the array sensors are installed. Some studies of direction estimation using a few microphones have been reported [63][64][66][67][68]. The propagation delay and the sound pressure difference between sensors are used in [63] and in [64], respectively.

In the papers [66][67][68], we proposed a speaker direction estimation method using two microphones. This study exploits the harmonic structure of voiced speech signal to refine the method. It uses the virtually generated multichannel array data, called *frequency array data*, given by a pair of microphones [65]. However, there are two drawbacks in this method. One is that the estimation resolution degrades as the propagating direction apart from the array broadside. Another one is that the method cannot discriminate whether the speaker is in front or behind. These properties are inherent in linear array, and the two-sensor array is the simplest linear array.

In this chapter, a new method to estimate omni-directional DOA with realising spatially uniform resolution by the integrated use of microphone pairs is proposed.

The main proposals in the research are summarised as follows.

- Array of three microphones located at vertices of equilateral triangle (we call it "equilateral-triangular microphone array" in the below).
- A new direction estimation scheme by integrating the frequency array data for three pairs of microphones

The former aims to realise uniform resolution with respect to direction. In the equilateral-triangular microphone arrangement, there are three different pairs of microphones and each of them is a linear 2-sensors array. Since the broadside of each microphone pair faces at a different angle individually, the resolution is expected to be uniform by integrating the frequency array data of these three pairs. The second idea is the integrated use of the three frequency array data. At the estimation stage of the method, we apply the subspace analysis to the *integrated frequency array data*. The process does not require any *a priori* knowledge of speaker direction that is necessary for MUSIC with CSS. The additional advantage of the method is its robustness to reverberation. That is because of a spatial averaging effect by the array data integration.

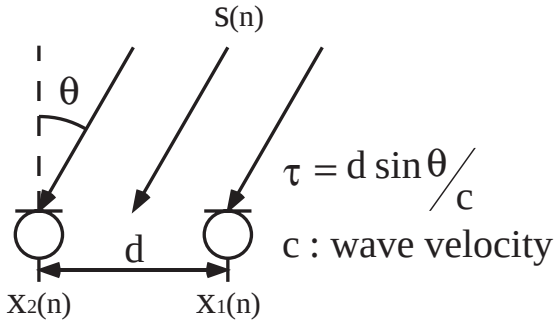

Figure 4.1: Microphone pair model to derive a frequency array data

Next, we extend this method to estimate not only the azimuth but also the elevation angle of speaker direction. In this extension, we use four microphones to form a tetrahedral microphone array to get the discriminability for elevation angle.

This chapter is organized as follows. In the following Sec. 4.2, we have a brief review of our previous method [66][67][68], and then the details of the direction estimation method [69][70][71] including the results of simulation and experiment are described in Sec. 4.3. The extended scheme for estimating both azimuth and elevation angles [72][73][74] is introduced in Sec. 4.4, and some concluding remarks are stated in Sec. 4.5.

4.2 Speaker direction estimation using a pair of microphones [66][67][68]

We briefly review the speaker direction estimation method using a pair of microphones proposed in [66][67][68]. Let us consider the two channel signals $\{x_1(n), x_2(n)\}$ in Fig. 4.1, obtained by a pair of microphones, represented by

$$x_1(n) = s(n) + n_1(n) \tag{4.1}$$

$$x_2(n) = s(n-\tau) + n_2(n) \tag{4.2}$$

where $s(n)$ is a voiced speech signal, τ is the time delay between two microphones which is a function of the sound source direction θ, and $n_1(n)$ and $n_2(n)$ are

mutually uncorrelated white noise signals. Thus, the Fourier transforms of $x_1(n)$ and $x_2(n)$, and their cross spectrum, are represented by

$$X_1(\omega) = S(\omega) + N_1(\omega) \tag{4.3}$$

$$X_2(\omega) = S(\omega)e^{-j\omega\tau} + N_2(\omega) \tag{4.4}$$

and

$$G_{12}(\omega) = E\left[X_1^*(\omega)X_2(\omega)\right] = P_S(\omega)e^{-j\omega\tau} \tag{4.5}$$

respectively, where $P_S(\omega)$ and the expectation $E[\cdot]$ denote the power spectral density of $s(n)$ and the average of DFT at several frames respectively, and $*$ means the complex conjugate. When we set $\omega = \omega_m$, where ω_m is the m-th higher harmonics of the fundamental frequency ω_0 of $s(n)$, i.e.

$$\omega_m = m\omega_0, \tag{4.6}$$

the phase term in $G_{12}(\omega_m)$ is replaced by $e^{-j\omega_0 m\tau}$. This phase term is interpreted as a time delay, m times τ, of a narrow-band signal whose central frequency is ω_0. This interpretation leads us to the idea that the $G_{12}(\omega_m)$ might be the virtual multichannel array signals, which are narrow-band signals acquired by an equally spaced linear multiple microphones. For determining the frequency ω_0 and its harmonics, the harmonic structure of voiced sound $s(n)$ is employed. That is, we set ω_0 as the fundamental frequency of voiced sound in speech. Because the power of a voiced sound is localized in its harmonic frequencies, the SNRs at these frequencies are rather high, and as a result, harmonic elements contribute to improving the estimation accuracy. Thus, we define the following frequency array data $\mathbf{G}(\omega_0)$ for a pair of microphone signals.

$$\mathbf{G}(\omega_0) = \left[\begin{array}{ccc} \frac{G_{12}(a\omega_0)}{|G_{12}(a\omega_0)|} & \frac{G_{12}(b\omega_0)}{|G_{12}(b\omega_0)|} & \cdots \end{array} \right]^T \tag{4.7}$$
$$(a, b, \cdots \in \mathbf{m})$$

Since power spectrum distribution depends on speaker and phoneme, here we select the $\hat{M}$ harmonics that contains the voiced speech components in higher SNR condition. In Eq.(4.7), $\mathbf{m}$ is a set of the $\hat{M}$ harmonics order selected by thresholding the magnitude-squared coherence function [75], as given by Eq.(4.8)–Eq.(4.10).

$$\mathbf{m} = (m \mid |\eta_{xy}(m\omega_0)| \geq T, \quad (m = 1, 2, \cdots, M)) \tag{4.8}$$

$$\eta_{xy}(m\omega_0) = 10\log_{10}\frac{|\gamma_{xy}(m\omega_0)|^2}{1-|\gamma_{xy}(m\omega_0)|^2} \tag{4.9}$$

$$|\gamma_{xy}(m\omega_0)|^2 = \frac{|E[X^*(m\omega_0)Y(m\omega_0)]|^2}{E[|X(m\omega_0)|^2]E[|Y(m\omega_0)|^2]} \tag{4.10}$$

The M is the highest order of the candidate harmonics determined by the criterion stated in [66]. The fundamental frequency ω_0 is estimated by evaluating logarithmic harmonic product spectrum [76]. Here we note that the magnitude of each component in $\mathbf{G}(\omega_0)$ are normalized as shown in Eq.(4.7). Finally, the direction estimation is performed by applying the MUSIC [60] to the frequency array data $\mathbf{G}(\omega_0)$.

4.3 Speaker direction estimation using equilateral-triangular microphone array [69][70][71]

4.3.1 Problem settings

An equilateral-triangular microphone array as shown in Fig. 4.2 is used. The three microphones are located at the vertices of an equilateral triangle, and a speaker in the direction θ utters a voiced speech signal $s(n)$. The microphones receive the signal $x(n)$, $y(n)$ and $z(n)$ respectively, with additive sensor noise signals $n_x(n), n_y(n)$ and $n_z(n)$ that can be modelled as spatially uncorrelated. From such configuration, we can take three pairs of microphones that have equal distance D between microphones and each pair faces to different direction of every $\frac{2\pi}{3}$[rad]. For a linear array, including a microphone pair as the simplest case, it has the highest spatial accuracy to its facing (broadside) directions. The aim of the proposed method is to realise uniform accuracy by integrating these three microphone pairs.

Here we set the following assumptions for the input signal without loss of generality.

a) One voiced speech signal is received.

A speech signal mainly contains voiced sound localizing at some time segments [2] and usually it is possible to extract a single speaker time segments even for a double speak case.

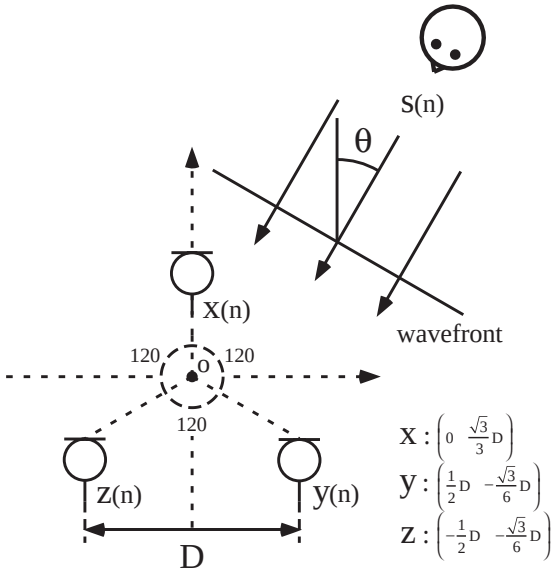

Figure 4.2: Model of input signal to the equilateral-triangular microphone array

b) The location of the speaker is restricted on the array plane.
Later, the effects of circumstances violating this assumption are considered in Sec. 4.3.3.

4.3.2 Proposed method

Model of input signal

The short-time Fourier transforms of each microphone input signals $x(n)$, $y(n)$ and $z(n)$ in Fig. 4.2 are given by

$$\begin{cases} X(\omega) = S(\omega)e^{-j\omega\tau_x} + N_x(\omega) \\ Y(\omega) = S(\omega)e^{-j\omega\tau_y} + N_y(\omega) \\ Z(\omega) = S(\omega)e^{-j\omega\tau_z} + N_z(\omega) \end{cases} \tag{4.11}$$

where $\tau_{\mathtt{x}}(\mathtt{x} = x, y, z)$ denotes the signal arrival delay at microphone $\mathtt{x}$ with respect to the reference point located at the array origin o. Here we can define the cross spectra of three microphone pairs as shown in Eq.(4.12).

$$\begin{cases} G_{xy}(\omega) = E\left[X^*(\omega)Y(\omega)\right] = P_S(\omega)e^{-j\omega\tau_{xy}} \\ G_{yz}(\omega) = E\left[Y^*(\omega)Z(\omega)\right] = P_S(\omega)e^{-j\omega\tau_{yz}} \\ G_{zx}(\omega) = E\left[Z^*(\omega)X(\omega)\right] = P_S(\omega)e^{-j\omega\tau_{zx}} \end{cases} \tag{4.12}$$

The delay variables in Eq.(4.12) are the function of speaker direction θ given by

$$\tau_{xy}(\theta) = D\sin(\theta + \tfrac{2}{3}\pi)/c \tag{4.13}$$

$$\tau_{yz}(\theta) = D\sin\theta/c \tag{4.14}$$

$$\tau_{zx}(\theta) = D\sin(\theta - \tfrac{2}{3}\pi)/c \tag{4.15}$$

where c denotes the sound velocity. Then, we form the frequency array data $\mathbf{G}_{xy}(\omega_0, \theta)$, $\mathbf{G}_{yz}(\omega_0, \theta)$ and $\mathbf{G}_{zx}(\omega_0, \theta)$ by extracting the $\hat{M}$ harmonic components as shown in Sec. 4.2. For simplicity, we omit the ω_0 in the following part of this paper.

Integration of three frequency array data

Now let us consider the difference of delay term (which determines the phase value) between two frequency array data for a signal propagating from direction ϕ.

$$\tau_{x2y}(\phi) \equiv \tau_{yz}(\phi) - \tau_{xy}(\phi) = \sqrt{3}D\sin(\phi - \tfrac{\pi}{6})/c \tag{4.16}$$

$$\tau_{z2y}(\phi) \equiv \tau_{yz}(\phi) - \tau_{zx}(\phi) = \sqrt{3}D\sin(\phi + \tfrac{\pi}{6})/c \tag{4.17}$$

Then, we define the following $\hat{M} \times \hat{M}$ diagonal matrices called *rotation matrices* that consist of the phase compensating components with respect to the signal from direction ϕ.

$$\mathbf{G}_{x2y}(\phi) \equiv \mathrm{diag}\left[\, e^{-ja\omega_0\tau_{x2y}(\phi)} \;\; e^{-jb\omega_0\tau_{x2y}(\phi)} \;\; \dots \right] \tag{4.18}$$

$$\mathbf{G}_{z2y}(\phi) \equiv \mathrm{diag}\left[\, e^{-ja\omega_0\tau_{z2y}(\phi)} \;\; e^{-jb\omega_0\tau_{z2y}(\phi)} \;\; \dots \right] \tag{4.19}$$

Using these rotation matrices, we define the following data called integrated frequency array data.

$$\mathbf{G}_m(\phi, \theta) = \{\mathbf{G}_{x2y}(\phi)\mathbf{G}_{xy}(\theta) + \mathbf{G}_{yz}(\theta) + \mathbf{G}_{z2y}(\phi)\mathbf{G}_{zx}(\theta)\}/3 \tag{4.20}$$

It is noted that the phases of each terms in the right side of Eq.(4.20) are equal if and only if $\phi = \theta$.

Subspace analysis of integrated array data matrix

Here let us note the delay term of each rotated frequency array data in Eq.(4.20).

$$\begin{aligned} \mathbf{G}_{x2y}(\phi)\mathbf{G}_{xy}(\theta) &\rightarrow \tau_{x2y}(\phi) + \tau_{xy}(\theta) \equiv \breve{\tau}_{xy}(\phi,\theta) \\ \mathbf{G}_{yz}(\theta) &\rightarrow \tau_{yz}(\theta) \equiv \breve{\tau}_{yz}(\theta) \\ \mathbf{G}_{z2y}(\phi)\mathbf{G}_{zx}(\theta) &\rightarrow \tau_{z2y}(\phi) + \tau_{zx}(\theta) \equiv \breve{\tau}_{zx}(\phi,\theta) \end{aligned} \tag{4.21}$$

The sign " $\rightarrow$ " above denotes to extract the delay term of a frequency array data. Now the following lemma is satisfied[1].

[Lemma]

The equation $\breve{\tau}_{xy}(\phi,\theta) = \breve{\tau}_{yz}(\theta) = \breve{\tau}_{zx}(\phi,\theta)$ is satisfied if and only if $\phi = \theta$.

From this lemma, we can replace our direction estimation problem by searching rotation matrices, which equalize the delay terms of all three frequency array data. To solve this subject, we can find the following theorem.

[Theorem]

The integrated frequency array data $\mathbf{G}_m(\phi,\theta)$ is equal to a steering vector $\mathbf{s}(\phi)$ defined by

$$\mathbf{s}(\phi) = \left[\begin{array}{ccc} e^{-ja\omega_0\tau_{yz}(\phi)} & e^{-jb\omega_0\tau_{yz}(\phi)} & \cdots \end{array} \right]^T \tag{4.22}$$

for ϕ that satisfies $\breve{\tau}_{xy}(\phi,\theta) = \breve{\tau}_{yz}(\theta) = \breve{\tau}_{zx}(\phi,\theta)$, and vice versa. That is

$$\breve{\tau}_{xy}(\phi,\theta) = \breve{\tau}_{yz}(\theta) = \breve{\tau}_{zx}(\phi,\theta) \Longleftrightarrow \mathbf{G}_m(\phi,\theta) = \mathbf{s}(\phi) \tag{4.23}$$

[Proof of Theorem]

The magnitude of k-th element in $\mathbf{G}_m$ is less than 1 not as far as all the interpolated delay terms are equal.

$$\begin{aligned} |[\mathbf{G}_m]_k| &= \left| e^{-jk\omega_0\breve{\tau}_{xy}(\phi,\theta)} + e^{-jk\omega_0\breve{\tau}_{yz}(\theta)} + e^{-jk\omega_0\breve{\tau}_{zx}(\phi,\theta)} \right| /3 \\ &\leq \left\{ |e^{-jk\omega_0\breve{\tau}_{xy}(\phi,\theta)}| + |e^{-jk\omega_0\breve{\tau}_{yz}(\theta)}| + |e^{-jk\omega_0\breve{\tau}_{zx}(\phi,\theta)}| \right\} /3 \\ &= 1 \end{aligned}$$

The equality is satisfied only if the three complex values are equal.■

This theorem leads the DOA estimation problem to search the parameter ϕ satisfying the equality $\mathbf{G}_m(\phi,\theta) = \mathbf{s}(\phi)$. In order to determine ϕ, we use the

[1]The proof is described in the Appendix B

subspace structure of the following covariance matrix $\mathbf{R}_m(\phi)$ for $\mathbf{G}_m(\phi, \theta)$.

$$\mathbf{R}_m(\phi) = \mathbf{G}_m \mathbf{G}_m^H. \tag{4.24}$$

Because $\mathbf{R}_m(\phi)$ is an Hermitian matrix, each eigenvector $\mathbf{v}_i$ of $\mathbf{R}_m(\phi)$ is mutually orthogonal. Namely,

$$\mathbf{v}_i^H \mathbf{v}_j = \delta_{ij}, \tag{4.25}$$

where δ_{ij} is the Kronecker delta given by

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}. \tag{4.26}$$

We also have

$$\mathbf{R}_m(\phi) = \sum_{i=1}^{\hat{M}} \lambda_i \mathbf{v}_i(\phi) \mathbf{v}_i^H(\phi). \tag{4.27}$$

From the well-known theorem on array covariance matrix [8], the eigenvector $\mathbf{v}_1$ corresponding the largest eigenvalue λ_1 is equal to the vector $\mathbf{G}_m$ in the case of rank-1 model. The estimated value $\bar{\theta}$ is given by the following null search strategy.

$$\bar{\theta} = \arg\max_{\phi} |P(\phi)|, \tag{4.28}$$

where,

$$P(\phi) = \frac{1}{\sum_{i=2}^{\hat{M}} \mathbf{s}^H(\phi) \mathbf{v}_i(\phi) \mathbf{v}_i^H(\phi) \mathbf{s}(\phi)}. \tag{4.29}$$

Figure 4.3 shows the flow diagram of the proposed method.

Uniform spatial resolution

Let us recall the array of two microphones in Fig. 4.1. The received signal defined by Eq.(4.3) and Eq.(4.4) are rewritten in matrix form as

$$\begin{aligned} \mathbf{X}(\omega) &= S(\omega) \begin{bmatrix} 1 \\ e^{-j\omega\tau} \end{bmatrix} + \begin{bmatrix} N_1(\omega) \\ N_2(\omega) \end{bmatrix} \\ &\equiv S(\omega)\mathbf{s} + \mathbf{N}(\omega). \end{aligned} \tag{4.30}$$

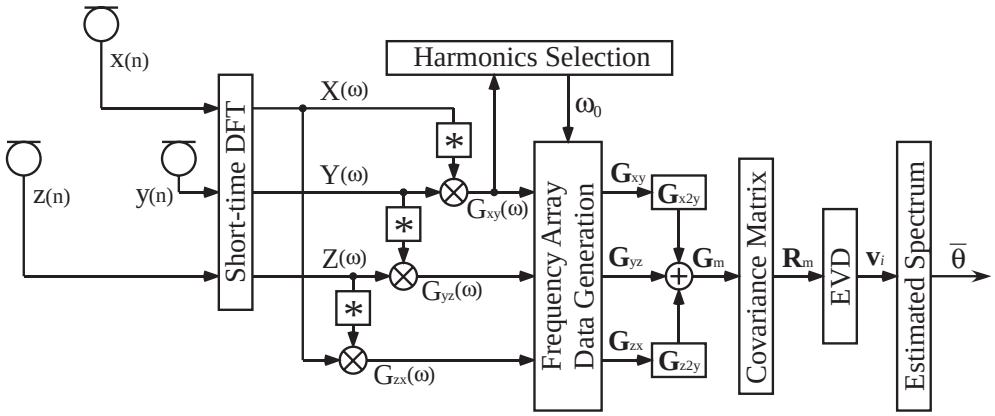

Figure 4.3: Flow diagram of the proposed method

The DOA information of the signal $S(\omega)$ is contained in the steering vector $\mathbf{s}$ that involves the phase difference $\omega\tau$ between the two microphones. While the frequency ω is determined, the phase term is a function of θ, therefore, the following relation is derived.

$$\begin{aligned} \theta &= \arcsin\left(\frac{c}{d}\tau\right) \\ &\equiv \arcsin(\hat{\tau}) \end{aligned} \tag{4.31}$$

Due to the noise in practice, the phase value of each sensor signal deviates randomly. Eventually, $\hat{\tau}$ in Eq.(4.31) is deviated and this causes the direction estimation error. For measuring the sensitivity of DOA estimation with respect to noise, we introduce *noise robustness factor* (NRF) $I(\theta)$, which is defined by

$$I(\theta) \equiv \left|\frac{d\theta}{d\hat{\tau}}\right|^{-1} \tag{4.32}$$

$$\begin{aligned} &= \left|\frac{\pm 1}{\sqrt{1-\hat{\tau}^2}}\right|^{-1} \\ &= \left|\sqrt{1-\hat{\tau}^2}\right| \\ &= \left|\sqrt{1-\sin^2\theta}\right| \\ &= |\cos\theta|. \end{aligned} \tag{4.33}$$

We can find from $I(\theta)$ that the estimation accuracy degrades as θ being apart from the broadside direction.

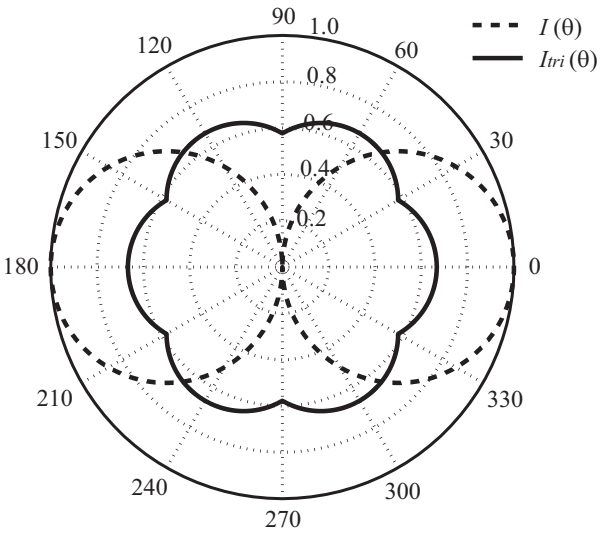

Figure 4.4: Noise robustness factor

On the other hand, the three pairs of microphones in the proposed method face at different directions of every $\frac{2\pi}{3}$[rad] and they are integrated by summation. So averaging NRFs of three frequency array data as follows derives the NRF of the equilateral-triangular microphone array.

$$I_{tri}(\theta) = \frac{1}{3}\left\{ I\left(\theta - \frac{2\pi}{3}\right) + I\left(\theta\right) + I\left(\theta + \frac{2\pi}{3}\right)\right\}. \tag{4.34}$$

In Fig. 4.4, we show the NRF $I(\theta)$ and $I_{tri}(\theta)$. The $I_{tri}(\theta)$ can keep its value nearly constant to omni-direction.

Discriminability for opposite-direction

Another advantage of the proposed method is the discriminability for a direction and its opposite. Now, let us consider a signal arriving from the back of the linear array, i.e. $\pi > |\theta| > \frac{\pi}{2}$. From $\sin\theta = \sin(\pi - \theta)$, the linear array cannot discriminate whether the signal arrives from its front or the back. In contrast, the proposed method can discriminate the signal from omni-direction because the preceding theorem and lemma in Sec. 4.3.2 hold for $\theta = [-\pi, \pi]$.

Table 4.1: Parameters for simulation

Input SNR	20dB
Sampling Frequency	16000Hz
Wave Velocity c	340m/s
D	0.08m
Threshold T	15dB
Window	Hamming
FFT point	4096
Frame Length	600
Frame Overlap	300
Data Length	625ms

Suppressing effects to the reverberation

The integrated use of plural frequency array data in our method is expected to be effective for suppressing the influence of reverberation. Usually, reverberation is known to be spatially diffuse due to the multiple reflection paths. On the other hand, each frequency array data is generated by the data acquired at different spatial position. From these facts, the reverberation components in each term of Eq.(4.20) are mutually uncorrelated. Thus, the integrated frequency array data can be expected to possess an anti-reverberation effect.

4.3.3 Simulation and experiment results

Evaluation with computer simulation

For the computer simulation, we use the real 5 phoneme data (/a/,/e/,/i/,/o/,/u/) uttered by 10 subjects (5 each for male and female) as a source signal and take 5 trials for every data. As the conventional methods for comparison, we adopt our previous method [66] with linearly located 2, 3 and 4 microphones. In the case of 3 and 4 microphones, we use the average of multiple frequency array data before the covariance matrix derivation. Furthermore, we also compare with the MUSIC [60] with CSS [34] on the harmonics [37]. In the method, the pre-estimation is obtained by the beamformer method [77]. The parameters shown in Table 4.1

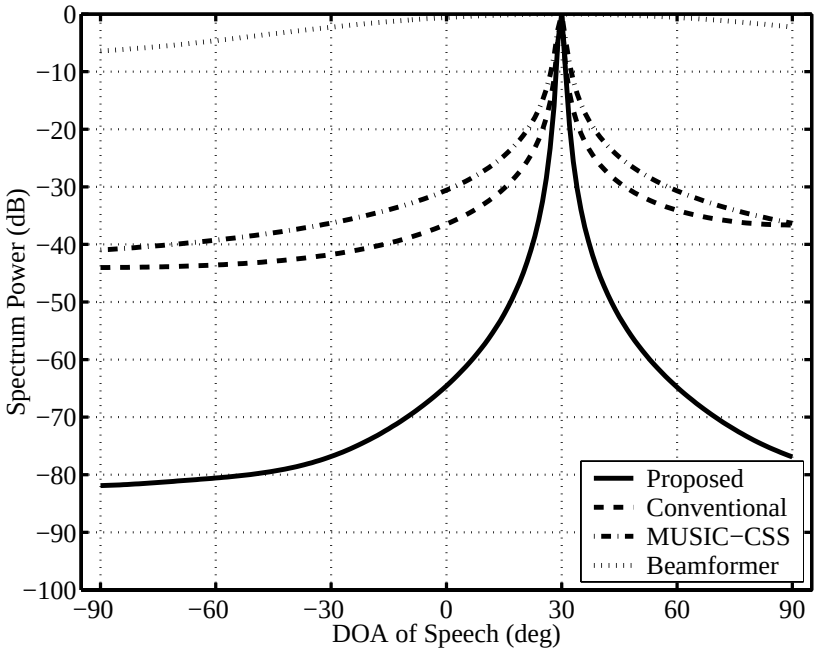

Figure 4.5: Estimated spectra (female /a/, $\theta = 30°$)

are adopted to every method. For the conventional methods, we use the same harmonics selected in the proposed method.

The spatial resolution is evaluated by the deviation of estimation error (DEE), which is given by

$$DEE = \sqrt{\overline{|\bar{\theta}_i - \theta_T|^2}}, \tag{4.35}$$

where $\bar{\theta}_i$ and θ_T are the estimated and true direction respectively, and $\overline{\cdot}$ means average for i.

Evaluation for the anechoic case

We perform a numerical simulation using ideally generated microphone array input signals without reverberation. The microphone array input signal is virtually generated by delaying the signal with an appropriate samples according to θ, and sum up with additive white noise as a sensor noise.

Figure 4.5 shows the $P(\phi)$ (called "spectrum") of the proposed method and the spectra given by the conventional methods. The spectrum given by the pro-

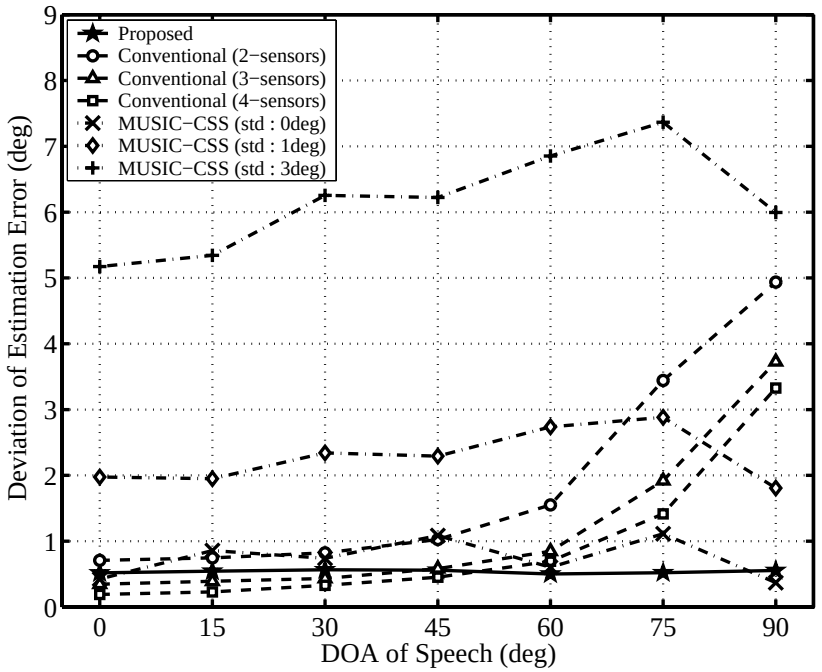

Figure 4.6: Deviation of estimation error at ideally anechoic case (solid line : Proposed, broken line : Conventional, dash dotted line : MUSIC-CSS)

posed method shows the most prominent peak[2] at the estimated angle. Figure 4.6 shows the DEEs for each method. In this simulation, we also compare with the MUSIC-CSS whose pre-estimated DOA involves random errors having Gaussian probability density function. From this result, we can recognise that the proposed method keeps its nearly constant spatial resolution in every direction, and its accuracy is better than that of the MUSIC-CSS with precise pre-estimated angle. Later in Fig. 4.14 (see the case $\psi = 0$), a nearly constant spatial resolution for omni-direction is shown for the proposed method.

Evaluation for a simulated reverberant condition

For evaluating reverberation suppressing effect in the proposed method, we perform computer simulation using the room impulse responses simulated by the image method [78]. The room model for the simulated reverberant condition is summarised in Fig. 4.7, and for the reflection coefficients β in [78], we use the

[2]The prominent peak of the spectrum results in the definiteness of the estimated DOA

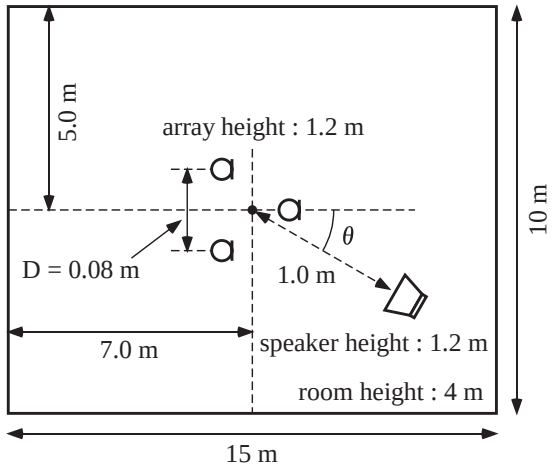

Figure 4.7: Room model for reverberant room simulation

Table 4.2: Conditions for simulated reverberant room

Case	β	T_R[sec]
I	0.5	0.27
II	0.6	0.31
III	0.7	0.38

values as shown in Table 4.2 except for the values relating to ceiling and floor which are fixed at 0.5 in every case. We also denote the approximate reverberation time T_R calculated using the given impulse response [79] in Table 4.2. The other parameters are the same as given in Table 4.1 except that the threshold T is 10dB. In Fig. 4.8, the DEEs for each method are shown. From these results, the proposed method keeps its accuracy and uniformity even when reverberation exists.

Experiments at real acoustic environment

To verify that the proposed method is effective even at real acoustic environment, we performed some experiments in a conference room whose physical sizes are shown in Fig. 4.9. The equilateral-triangular microphone array system used for sound acquisition is shown in Fig. 4.10 and Fig. 4.11. The speech data and

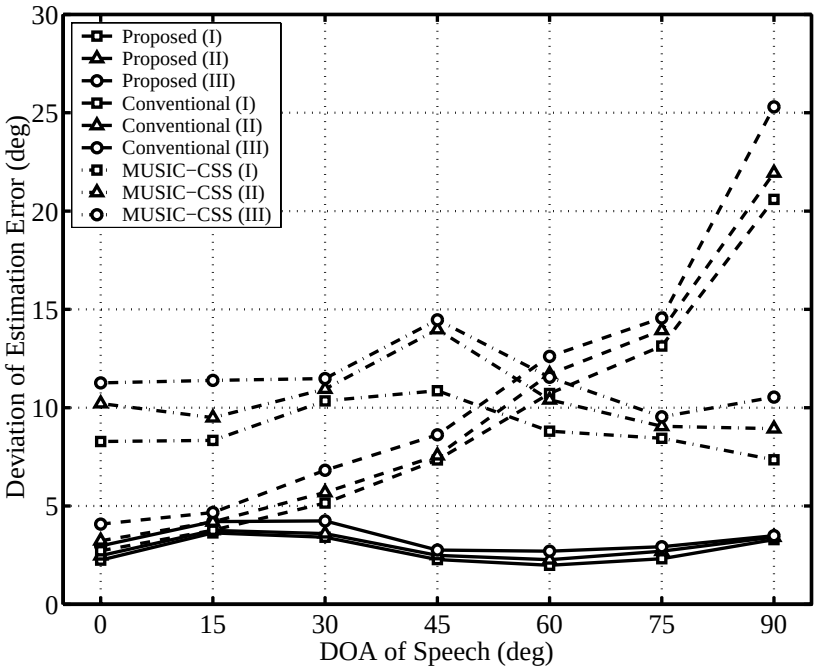

Figure 4.8: Estimation results in the simulated reverberant condition (solid line : Proposed, broken line : Conventional, dash dotted line : MUSIC-CSS)

parameters are the same as in the preceding computer simulation except for the input SNR being around 18dB and the threshold T is 12dB, and here we also undertook 5 trials for each data. We regard the mean value of the estimation results $\bar{\theta}_{MEAN}$ shown in Table 4.3 as the true direction θ_T and evaluate the DEE around this value. Figure 4.12 shows the results of the experiment. This result shows that the proposed method provides the best resolution. In contrast, the MUSIC-CSS is crucially degraded by the existing pre-estimation error. It is noted that the resolution of the proposed method is nearly uniform for omni-direction. (see the case $\psi = 0$[deg] in Fig. 4.15).

Influence by the elevation angle deviation

Although we assume that the speaker is located on the array plane, he/she may deviate from this plane in practice. Here we measure the influence occurred by the deviation of the elevation angle ψ between the speaker's location and array plane. Fig. 4.13 and Fig. 4.14 show the influence of the elevation angle to the

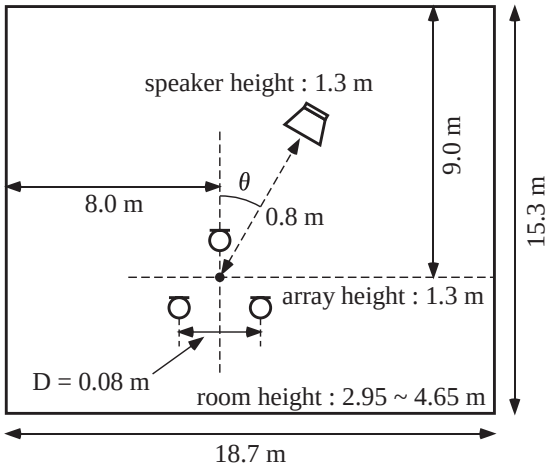

Figure 4.9: Acoustic environment for the experiment

Table 4.3: Mean value of the estimation results $\bar{\theta}_{MEAN}$

DOA[deg]	0	15	30	45	60	75	90
Proposed	−0.90	14.66	29.15	45.04	60.21	75.13	90.60
Conventional	−0.26	14.36	28.86	42.94	57.16	72.32	76.31
MUSIC-CSS	−0.05	14.72	24.80	40.75	58.56	73.66	92.07
Beamformer	−1.09	14.56	29.39	44.17	60.68	76.31	91.90

spectra and DEE, respectively. From these results, the elevation angle error below 10 degree degrades the estimation merely by a few degrees. Furthermore, this fact is also confirmed through the experimental results as shown in Fig. 4.15.

4.4 Estimation of azimuth and elevation direction using microphones located at apices of regular tetrahedron [72][73][74]

As we referred in preceding Sec. 4.3.3, the assumption that restricting the speaker movement to the elevation direction is quite unrealistic in practice. In some situations, we need to know the amount of the speaker deviation from the array

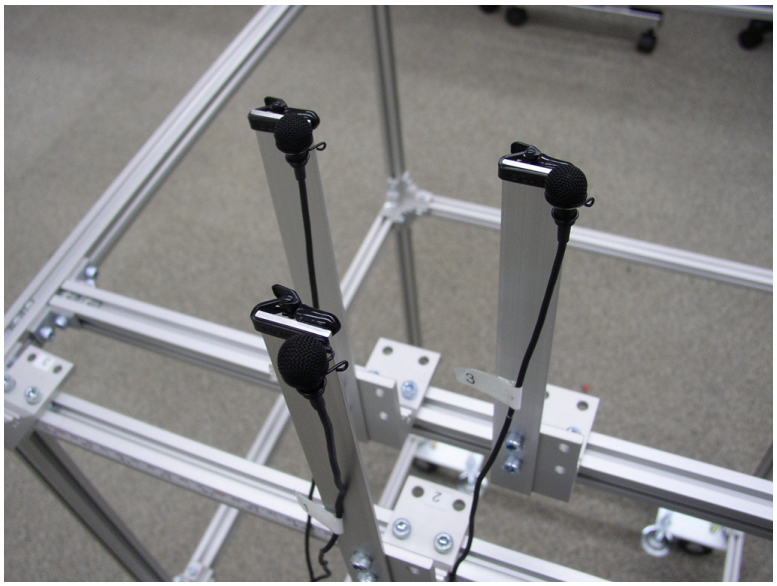

Figure 4.10: Equilateral-triangular microphone array used in the experiments

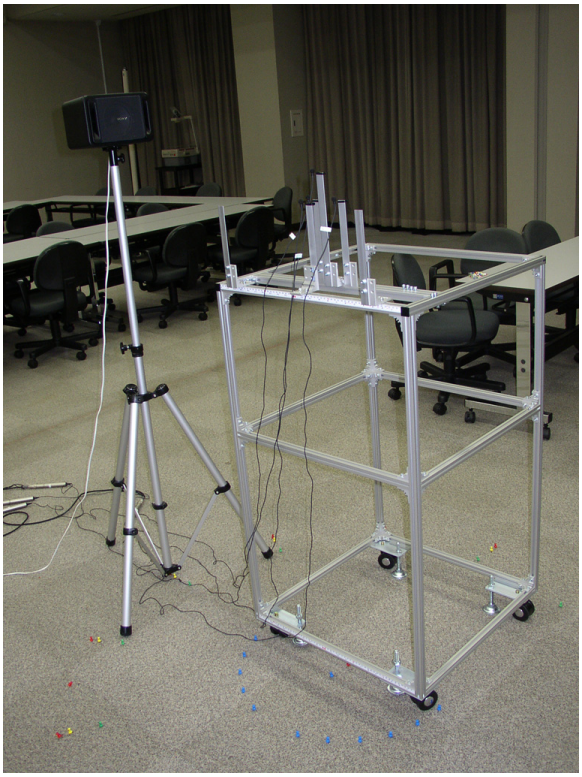

Figure 4.11: Microphone array system and loudspeaker

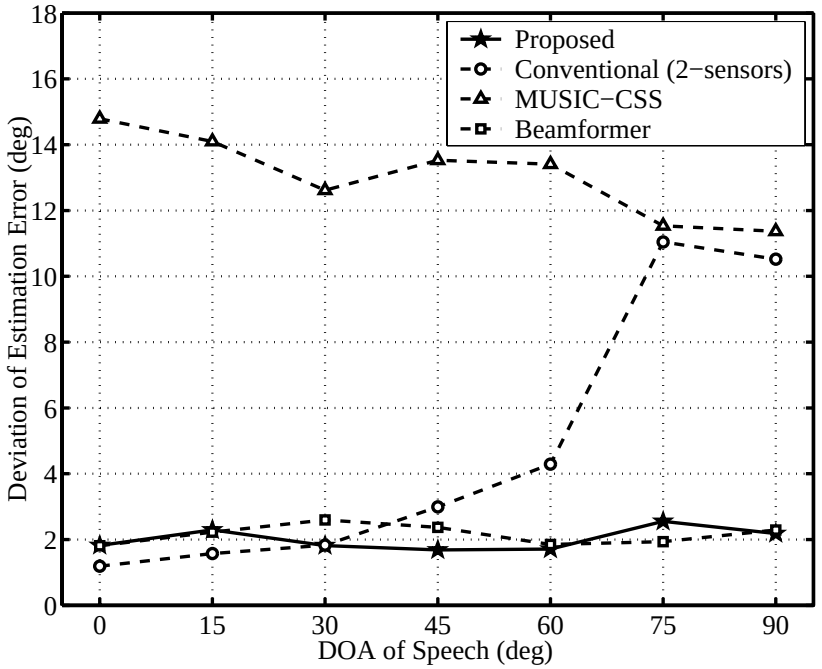

Figure 4.12: Results of experiment at real acoustic environment

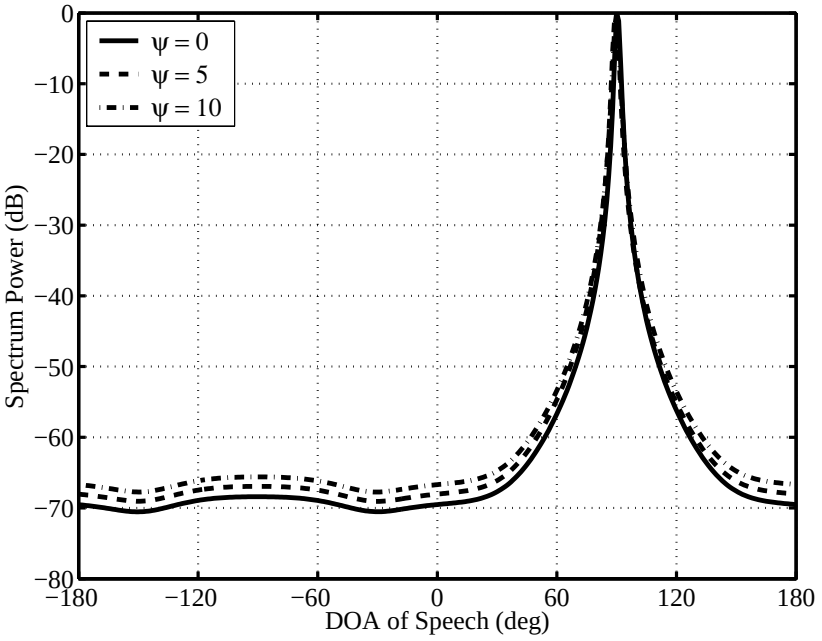

Figure 4.13: The influence of the elevation angle error to the spectra (same speech used in Fig. 4.5 arriving from $\theta = 90^\circ$)

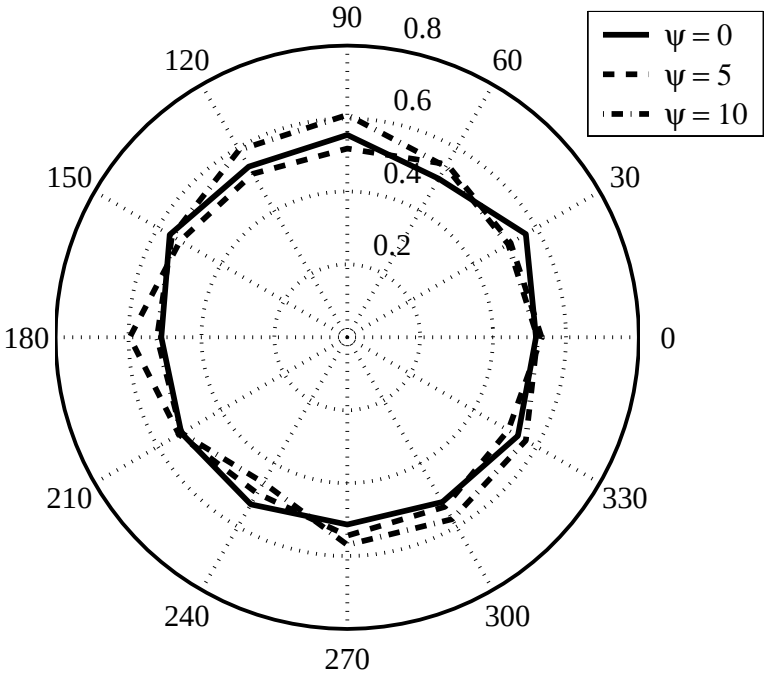

Figure 4.14: The influence of the elevation angle error to the DEE

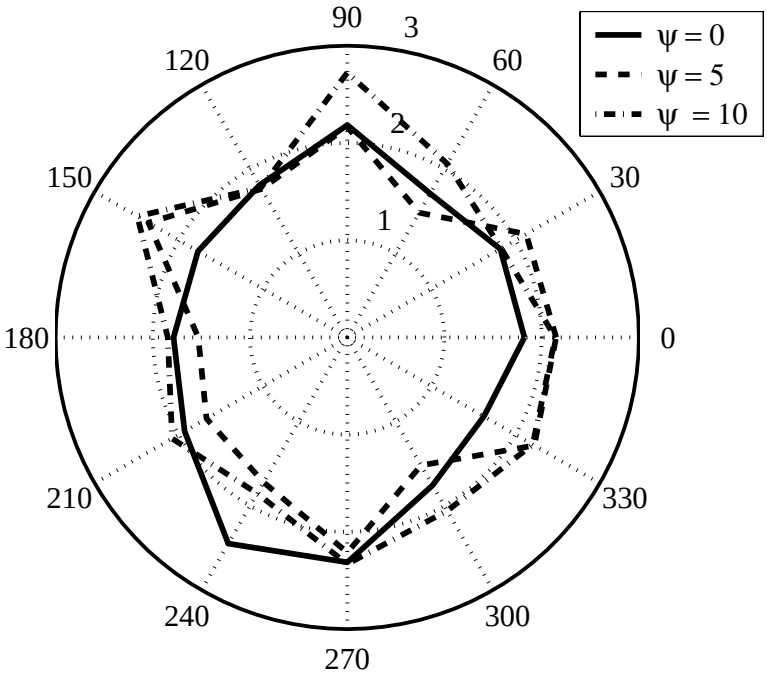

Figure 4.15: DEEs for different elevation angles in a real acoustic environment

plane accurately. Conscious of this fact, we propose an extension of the method to be able to estimate the elevation angle as well as the azimuth direction. The main ideas of the extension are summarised in the following two proposals.

- Use of four microphones located at the apices of a regular tetrahedron
- Azimuth and elevation angles are separately estimated

The former aims to provide data on the elevation angle as well as the azimuth angle. In addition to the three pairs of microphones making up the planar equilateral triangular array, we can have another three pairs of microphones located vertically to the planar triangular array. Using these latter pairs, we can measure the elevation angle around the plane of the triangular array. The idea of separate estimation is effectively used to reduce the calculation load.

4.4.1 Problem settings

In the proposed system, we receive the target signal from 4 microphones located at the apices of regular tetrahedron as shown in Fig. 4.16. A speaker in the direction {azimuth,elevation} = $\{\theta, \psi\}$ utters a voiced speech signal $s(n)$, and the microphones receive the signals $a(n)$, $b(n)$, $c(n)$ and $h(n)$ respectively, with additive sensor noise signals $n_a(n)$, $n_b(n)$, $n_c(n)$ and $n_h(n)$ that can be modelled as spatially uncorrelated. From such configuration, we have six pairs of microphones whose distances between microphones are equal and we can derive the frequency array data for each of them. Here we assume for the input signal that only one voiced speech signal is received as in the previous sections.

4.4.2 Proposed method

Separate DOA estimation using two groups of microphone pairs

A linear array, including a microphone pair as its simplest case, has spatial discriminability in the direction along the aperture line, and the highest resolution of it is obtained on its facing (broadside) direction. In the case of regular tetrahedral arrangement, its six microphone pairs are separated into two groups as shown in Fig. 4.17 from the spatial resolution point of view. The first group,

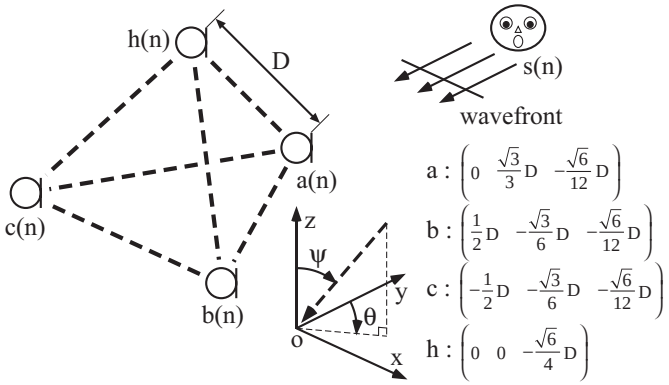

Figure 4.16: Model of input signal

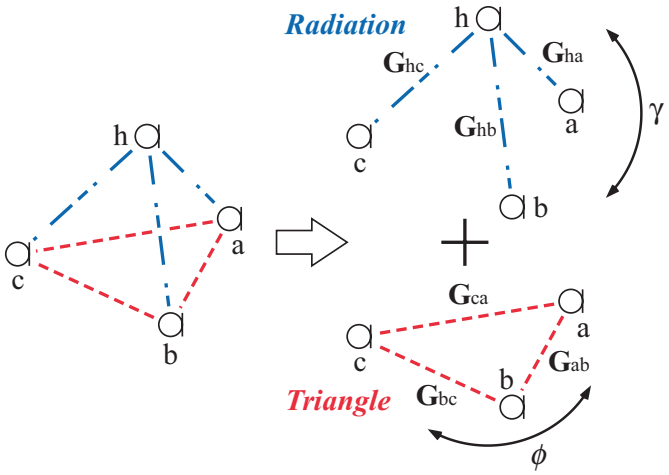

Figure 4.17: Separate DOA estimation

we call *Triangle*, consists of three pairs parallel to the x-y plane, i.e. $[ab]$, $[bc]$ and $[ca]$, where $[ij]$ means a pair of microphones i and j. The pairs in *Triangle* form the equilateral-triangular array as treated in Sec. 4.3. The *Triangle* has high discriminability to the azimuth angle but not to the elevation angle. The second group, we call *Radiation*, comprises the rest of the pairs, i.e. $[ha]$, $[hb]$ and $[hc]$. Because their apertures are nearly vertical to the x-y plane, they could be used for elevation angle estimation. Thus we propose to perform the respective estimations separately. Fig. 4.18 shows three steps algorithm as described in the followings.

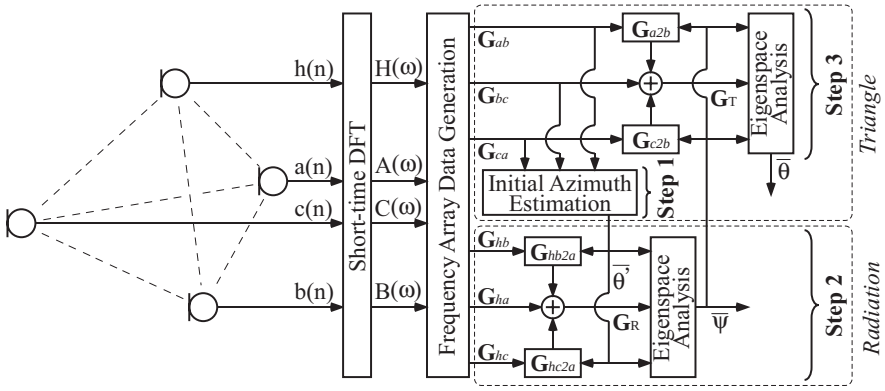

Figure 4.18: Data flow diagram of the proposed method

Step 1 : Initial azimuth estimation using *Triangle*

We first estimate an initial azimuth $\overline{\theta'}$, using *Triangle* by assuming $\psi = \frac{\pi}{2}$. In the paper [69][70][71], we confirmed through simulations and experiments that the deviation effect of ψ from $\frac{\pi}{2}$ to the azimuth estimation result is small as far as the deviation is within 10 degrees, and by the further simulations, we proved that the permissible range of the deviation is extended to nearly 50 degrees.

For a signal propagating from direction (ϕ, γ), we consider the following difference of delay terms (which determine the phase values) between two frequency array data associated with *Triangle*,

$$\tau_{a2b}(\phi,\gamma) \equiv \tau_{bc} - \tau_{ab} = \sqrt{3}D\sin(\phi - \tfrac{\pi}{6})\sin\gamma/c \tag{4.36}$$

$$\tau_{c2b}(\phi,\gamma) \equiv \tau_{bc} - \tau_{ca} = \sqrt{3}D\sin(\phi + \tfrac{\pi}{6})\sin\gamma/c, \tag{4.37}$$

where τ_{ij} means the time delay difference between microphones i and j. Then we define the following rotation matrices,

$$\mathbf{G}_{a2b}(\phi,\gamma) \equiv \mathrm{diag}\left[e^{-j\alpha\omega_0\tau_{a2b}} \quad e^{-j\beta\omega_0\tau_{a2b}} \quad \dots\right] \tag{4.38}$$

$$\mathbf{G}_{c2b}(\phi,\gamma) \equiv \mathrm{diag}\left[e^{-j\alpha\omega_0\tau_{c2b}} \quad e^{-j\beta\omega_0\tau_{c2b}} \quad \dots\right], \tag{4.39}$$

to generate the following integrated frequency array data.

$$\begin{aligned}\mathbf{G}_T(\theta,\psi,\phi,\gamma) \equiv \{&\mathbf{G}_{a2b}(\phi,\gamma)\mathbf{G}_{ab}(\theta,\psi) + \mathbf{G}_{bc}(\theta,\psi) \\ &+\mathbf{G}_{c2b}(\phi,\gamma)\mathbf{G}_{ca}(\theta,\psi)\}/3\end{aligned} \tag{4.40}$$

It is noted that the phases of three terms in the right side of Eq.(4.40) are equal if and only if $\phi = \theta$ and $\gamma = \psi$.[3] From this fact, our problem results in searching ϕ and γ that satisfies $\mathbf{G}_T = \mathbf{s}_{bc}$, where $\mathbf{s}_{bc}$ is the steering vector that contains the time delay between microphones b and c, defined by

$$\mathbf{s}_{bc}(\phi, \gamma) = \left[e^{-j\alpha\omega_0\tau_{bc}(\phi,\gamma)} \;\; e^{-j\beta\omega_0\tau_{bc}(\phi,\gamma)} \;\; \dots \right]^T. \tag{4.41}$$

To solve this problem, we analyse the subspace structure of $\mathbf{G}_T$, namely we perform eigenvalue decomposition to the covariance matrix $\mathbf{R}_T = \mathbf{G}_T\mathbf{G}_T^H$, and find the angles by the following maximum search.

$$\{\overline{\theta}, \overline{\psi}\} = \arg\max_{\{\phi,\gamma\}} |P(\phi, \gamma)| \,|_{\phi\in\Theta,\gamma\in\Psi}, \tag{4.42}$$

where

$$P(\phi, \gamma) = \frac{1}{\sum_{i=2}^{\hat{M}} \mathbf{s}_{bc}^H \mathbf{v}_i \mathbf{v}_i^H \mathbf{s}_{bc}}, \tag{4.43}$$

and $\mathbf{v}_i$ is the eigenvector corresponding to the i-th largest eigenvalue of $\mathbf{R}_T$. The search regions of ϕ and γ are given by the set of degrees Θ and Ψ, respectively. As stated in the first part of this section, we solve above problem by fixing $\Psi = \frac{\pi}{2}$.

Step 2 : Elevation estimation using *Radiation*

For the elevation estimation, we use the three frequency array data generated from *Radiation*. The estimation scheme is same as that given in the previous section, except for the rotation matrices, the integrated frequency array data and the steering vector given by the followings.

$$\mathbf{G}_{hb2a}(\phi, \gamma) \equiv \mathrm{diag}\left[e^{-j\alpha\omega_0\tau_{hb2a}} \;\; e^{-j\beta\omega_0\tau_{hb2a}} \;\; \dots \right] \tag{4.44}$$

$$\mathbf{G}_{hc2a}(\phi, \gamma) \equiv \mathrm{diag}\left[e^{-j\alpha\omega_0\tau_{hc2a}} \;\; e^{-j\beta\omega_0\tau_{hc2a}} \;\; \dots \right] \tag{4.45}$$

$$\begin{aligned}\mathbf{G}_R(\theta, \psi, \phi, \gamma) \equiv \{&\mathbf{G}_{ha}(\theta, \psi) + \mathbf{G}_{hb2a}(\phi, \gamma)\mathbf{G}_{hb}(\theta, \psi) \\ &+\mathbf{G}_{hc2a}(\phi, \gamma)\mathbf{G}_{hc}(\theta, \psi)\}/3\end{aligned} \tag{4.46}$$

$$\mathbf{s}_{ha}(\phi, \gamma) = \left[e^{-j\alpha\omega_0\tau_{ha}(\phi,\gamma)} \;\; e^{-j\beta\omega_0\tau_{ha}(\phi,\gamma)} \;\; \dots \right]^T, \tag{4.47}$$

[3]The proof is derived by the same way as stated in Sec. 4.3.

where

$$\tau_{hb2a}(\phi,\gamma) \equiv \tau_{ha} - \tau_{hb} = D\sin(\phi - \tfrac{\pi}{3})\sin\gamma/c \quad (4.48)$$

$$\tau_{hc2a}(\phi,\gamma) \equiv \tau_{ha} - \tau_{hc} = D\sin(\phi - \tfrac{2\pi}{3})\sin\gamma/c. \quad (4.49)$$

In this step, the azimuth is restricted to the initial estimate as $\Theta = \overline{\theta'}$, thus only the elevation $\overline{\psi}$ is estimated.

Step 3 : Azimuth determination using *Triangle*

As the final step, we renew the azimuth using *Triangle* again. Except for setting $\Psi = \overline{\psi}$, the estimation method is same as that in Step 1.

For further improvement of accuracy, we can repeat from Step.2 to Step.3 as far as the increase of computation cost is allowed.

4.4.3 Simulation and experimental results

Evaluation with computer simulation

For the computer simulation, we use the real 5 phoneme data (/a/,/e/,/i/,/o/,/u/) uttered by 10 subjects (5 each for male and female) as the source signal and had 5 trials for every data. The microphone array input signal is virtually generated by delaying the signal with appropriate samples according to θ and white noise is added as the sensor noise. MUSIC with CSS is the conventional method used for comparison. For the pre-estimated DOA information, we add estimation error factor following Gaussian probability density distribution due to reflect the pre-estimation inaccuracy. All the same parameters shown in Table 4.4 are adopted for every method, and for the conventional method, we use the same harmonics selected in the proposed method.

Fig. 4.19 and Fig. 4.20 show the deviation of final estimation error (DEE) defined in Sec. 4.3.3. From these results, we can recognise that the proposed method keeps its high accuracy, which is almost the same level as that of the MUSIC-CSS with accurately pre-estimated DOA, at every azimuth direction and the elevation angle nearly ranging 30 to 150 degrees. In practical use, a speaker's location to the elevation direction is restricted within the area around the plane, so that the proposed method would be enough in practice.

Table 4.4: Parameters for simulation

Input SNR	20dB
Sampling Frequency	16000Hz
D	8cm
Sound Velocity v	340m/s
Threshold T [66]	15dB
Window	Hamming
Frame Length	600
Frame Overlap	300
Data Length	625ms

Evaluation in real acoustic environment

To verify that the proposed method is still effective even in real acoustic environment, we performed some experiments in a conference room as shown in Fig. 4.21. The setup of utilized regular tetrahedral microphone array is shown in Fig. 4.22. The speech data and parameters are the same as in the computer simulation except for the threshold T settled at 10dB, and here we also made 5 trials for each data. The results of the experiment as shown in Fig. 4.23 and Fig. 4.24 show that the proposed method gives better results than that of the MUSIC-CSS.

4.5 Conclusion of Chapter 4

In this chapter, new algorithms of speaker direction estimation that can discriminate the omni-direction with uniform accuracy have been proposed. In the method, the frequency array data calculated from a pair of microphones is introduced at first in Sec. 4.2, and the proposals of direction estimation using the equilateral-triangular microphone array is stated in Sec. 4.3. Added to these, in Sec. 4.4, we also proposed an extension of the method to make it applicable for more realistic situations. A future research topic is the tracking of moving speaker direction. The following Chap. 5 talks about the idea to solve this problem.

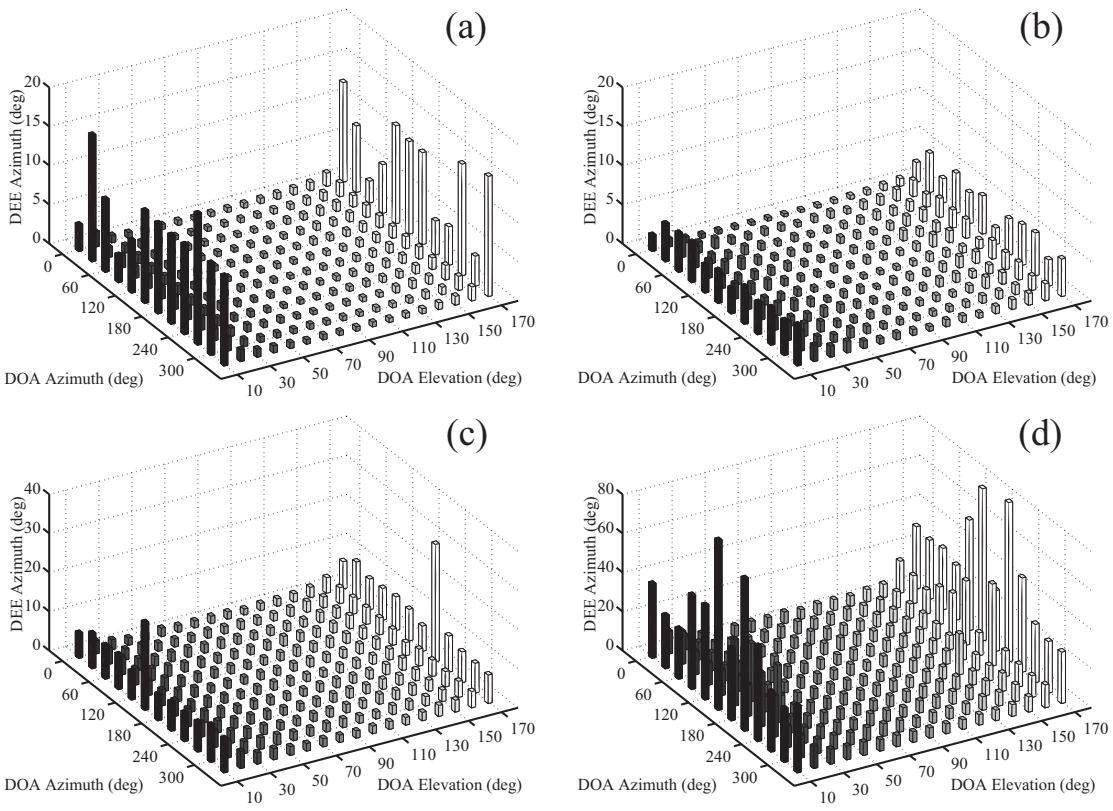

Figure 4.19: DEE of azimuth at ideally anechoic case : (a)Proposed (b)–(d)MUSIC-CSS (standard deviation of pre-estimation error : (b)0[deg] (c)1[deg] (d)3[deg])

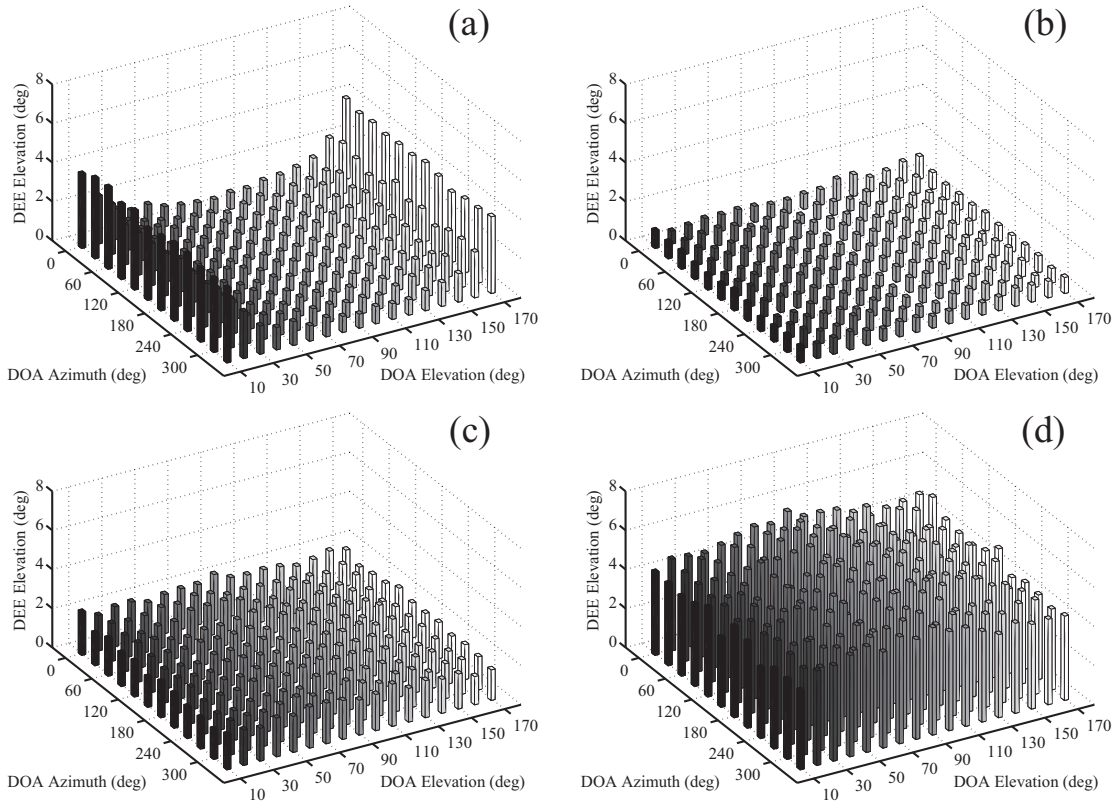

Figure 4.20: DEE of elevation at ideally anechoic case : (a)Proposed (b)–(d)MUSIC-CSS (standard deviation of pre-estimation error : (b)0[deg] (c)1[deg] (d)3[deg])

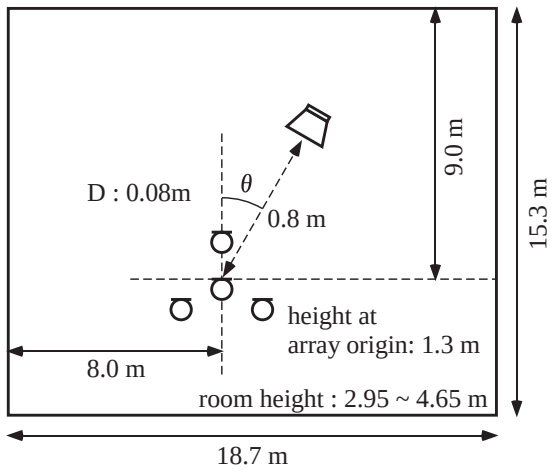

Figure 4.21: Room configuration for the experiment

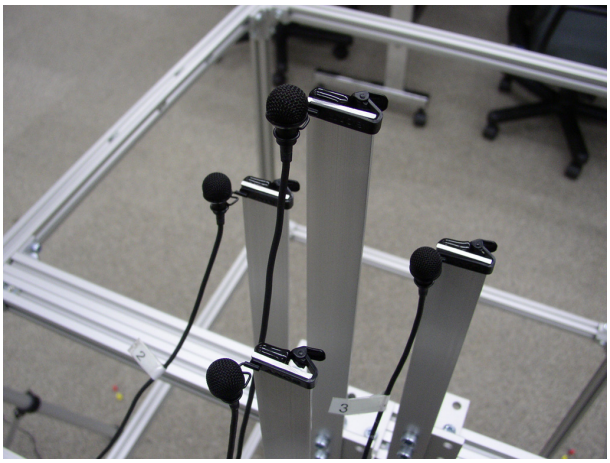

Figure 4.22: Regular tetrahedral microphone array used in the experiments

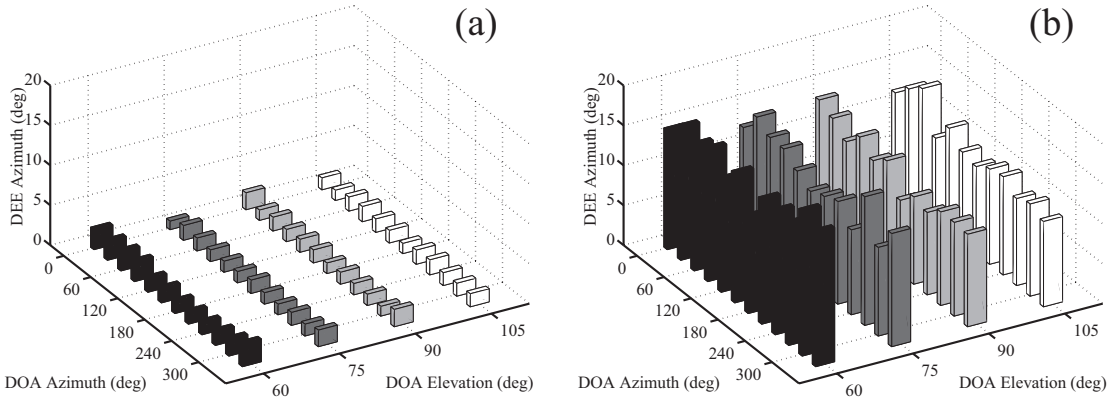

Figure 4.23: DEE of azimuth at experiment under real acoustic environment : (a)Proposed (b)MUSIC-CSS

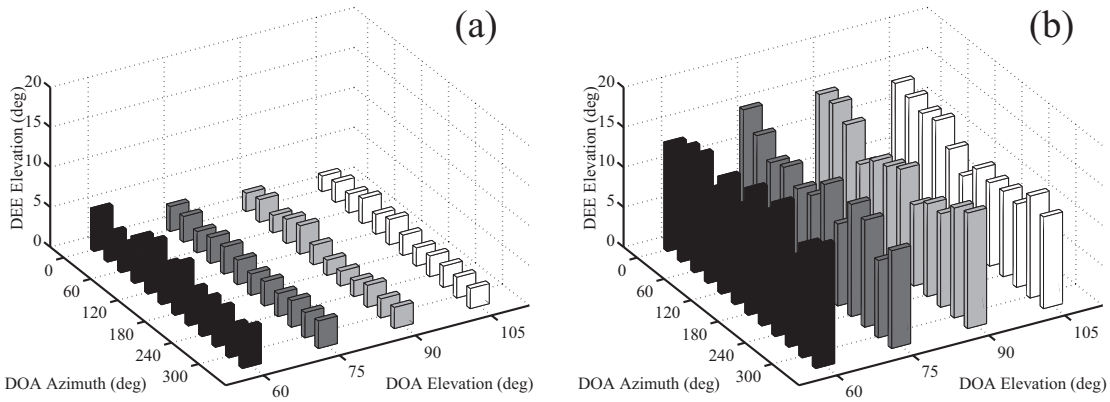

Figure 4.24: DEE of elevation at experiment under real acoustic environment : (a)Proposed (c)MUSIC-CSS

Chapter 5

Tracking of Speaker Direction by Equilateral-Triangular Microphone Array [80][81][82]

5.1 Introduction

Tracking of speaker direction is the issue discussed in this chapter. Teleconference and remote learning systems require the ability to track the target speaker direction rapidly. Several conventional methods for speaker direction tracking have been reported [42][43][44][45][83][84]. Some of them rely on adaptive beamforming, such as LCMV [43] and GSC [44] that capture the desired speech while suppressing the directional interferences adaptively, so that the speaker direction is determined from the beampattern of the given array weights. However, the method requires the beampattern calculation at every weight update, therefore, it heavily increases the computational cost and the accuracy is dependent upon the beampattern resolution. On the other hand, some methods without beampattern calculation have been proposed. Kawakami *et al.* [45] proposed a method that achieves speaker tracking by minimizing the output power of null steering fixed beamformer, and Suyama *et al.* [83] extended this method to double talk situation by introducing the data classification in the time-frequency domain. Zhang *et al.* [84] proposed another tracking method by combining CSS [34] and TDE.

In the case of teleconference and remote learning systems, there are numerous delegates speaking in turn, therefore, abrupt alternations in the current speaker direction often occur as well as gradual speaker movement. However, the application of the methods [45][83][84] is restricted to the single speaker with smooth movement. The tracking performance should be spatially uniform for omni-direction, and it is preferable that the number of microphones and array aperture are small from the practical point of view.

In this chapter, a new tracking algorithm of speaker direction using the equilateral-triangular microphone array is proposed. In the method, we aim at overcoming the huge computational cost problem caused by the eigenanalysis adopted in our DOA estimation method. The main proposals in the new method are summarised as follows.

- A novel tracking mechanism realised by the integrated use of three cross spectra obtained from the equilateral-triangular microphone array.
- Alternate the performance index during the adaptation to achieve fast and accurate global convergence.

The former is to realise the tracking system for omni-direction with uniform resolution, which is the modification of the method in Chap. 4. Since each microphone pair in the equilateral-triangular microphone array has different directional resolution [69], the method localizes the speaker direction by minimizing the performance index that consists of the cross spectra at three different microphone pairs. The second idea aims at the enhancement of the tracking accuracy and convergence speed. Because the power of the speech signal concentrates at specific harmonic frequencies, the SNRs at these harmonics are relatively high and they provide high accuracy of estimation. In addition, we select the harmonic components utilized for the performance index according as the convergence state. This contributes to enhance the convergence speed and for the assurance of global convergence.

This chapter is organized as follows. The following Sec. 5.2 summarises the problem formulation with the equilateral-triangular microphone array, and the proposed method is described in Sec. 5.3. Some discussions about the parameter settings and system performance based on the simulation results are stated in

Sec. 5.4, and experimental results in a real acoustic environment are shown in Sec. 5.5. Finally in Sec. 5.6, we conclude with some comments.

5.2 Problem formulation

We use the equilateral-triangular microphone array as shown in Fig. 5.1. A speaker in the direction θ utters a speech signal $s(n)$. The microphones receive the signal $\mathtt{x}(n)(\mathtt{x} = x, y, z)$ given by

$$\mathtt{x}(n) = s(n) \otimes a_{\mathtt{x}}(n) + w_{\mathtt{x}}(n), \tag{5.1}$$

where $a_{\mathtt{x}}$ and $w_{\mathtt{x}}(n)(\mathtt{x} = x, y, z)$ are the impulse response between the speaker and microphone $\mathtt{x}$ and additive sensor noise signals that can be modelled as spatially uncorrelated, respectively. The symbol $\otimes$ denotes the convolution. Under the assumption that we receive a plane wave at anechoic environment, the input signals are simplified as Eq.(5.2).

$$\mathtt{x}(n) = s(n - \tau_{\mathtt{x}}) + w_{\mathtt{x}}(n) \tag{5.2}$$

Here, $\tau_{\mathtt{x}}$ is signal delay at microphone $\mathtt{x}$ with respect to the reference point located at the array origin o, and n is the sampling index.

In the triangular configuration, we can take three pairs of microphones that have equal distance D between microphones and each pair faces to different direction of every $\frac{\pi}{3}$[rad]. As we also stated in [69], the cross spectrum of each microphone pair contains the speaker direction information in its phase term. So the proposed method aims to achieve the tracking of direction θ by using the cross spectra derived from three different microphone pairs.

Here we assume next a) and b) for the input signal.

a) Only one speech signal is received.
   In a situation such as a teleconference, it is usual to assume that more than one speaker do not speak simultaneously.

b) The location of the speaker is restricted on the array plane.
   In the real environment, this assumption may not be satisfied. Thus, we will describe the influence of it through experiments later.

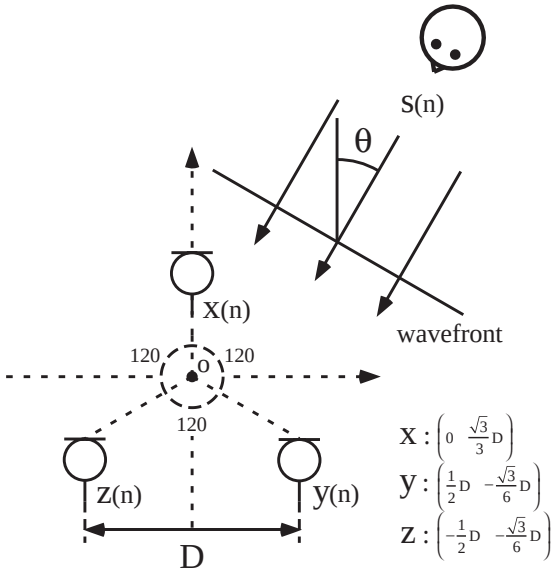

Figure 5.1: Model of input signal to the equilateral-triangular microphone array

5.3 Proposed method

5.3.1 Model of input signal

The short-time Fourier transforms of each microphone input signals $x(n)$, $y(n)$ and $z(n)$ in Fig. 5.1 are given by

$$\begin{cases} X(\omega) = S(\omega)e^{-j\omega\tau_x} + W_x(\omega) \\ Y(\omega) = S(\omega)e^{-j\omega\tau_y} + W_y(\omega) \\ Z(\omega) = S(\omega)e^{-j\omega\tau_z} + W_z(\omega) \end{cases}, \tag{5.3}$$

where $S(\omega)$ and $W_{\mathtt{x}}(\omega)$ are the Fourier transform of the speech $s(n)$ and noise $w_{\mathtt{x}}(n)(\mathtt{x} = x, y, z)$, respectively. Here we can define the cross spectra of three microphone pairs given by

$$\begin{cases} \hat{G}_{xy}^{(\omega)}(\theta) = E\left[X^*(\omega)Y(\omega)\right] = P_S(\omega)e^{-j\omega\tau_{xy}(\theta)} \\ \hat{G}_{yz}^{(\omega)}(\theta) = E\left[Y^*(\omega)Z(\omega)\right] = P_S(\omega)e^{-j\omega\tau_{yz}(\theta)} \\ \hat{G}_{zx}^{(\omega)}(\theta) = E\left[Z^*(\omega)X(\omega)\right] = P_S(\omega)e^{-j\omega\tau_{zx}(\theta)} \end{cases}, \tag{5.4}$$

where $P_s(\omega)$ and the expectation $E[\cdot]$ denote the power spectral density of $s(n)$ and the average of DFT at several frames respectively, and $*$ means the complex

conjugate. The delay constants in Eq.(5.4) are the function of θ given by

$$\begin{cases} \tau_{xy}(\theta) = D\sin(\theta + \frac{2}{3}\pi)/c \\ \tau_{yz}(\theta) = D\sin\theta/c \\ \tau_{zx}(\theta) = D\sin(\theta - \frac{2}{3}\pi)/c \end{cases}, \tag{5.5}$$

where c denotes the sound velocity.

Since major power of speech signal is localized in its harmonic frequencies, the SNRs at these frequencies are rather high, and as a result, harmonic elements contribute to improving the estimation accuracy. Thus in the following process, we utilize the cross spectra at the harmonic frequencies ω_m (m is the order of harmonics) selected by the SNR η_m higher than a threshold T[1], i.e. $m \in \mathcal{M} \equiv \{m|\eta_m \geq T\}$. Here the selected harmonic frequencies ω_m should be smaller than $\omega_{max} = \frac{\pi c}{D}$ to follow the spatial sampling theory [8].

5.3.2 Direction estimation by the cross spectra integration

Now let us consider the difference between delay terms of two cross spectra for a signal propagating from direction ϕ.

$$\tau_{x2y}(\phi) \equiv \tau_{yz}(\phi) - \tau_{xy}(\phi) = \sqrt{3}D\sin(\phi - \tfrac{\pi}{6})/c \tag{5.6}$$

$$\tau_{z2y}(\phi) \equiv \tau_{yz}(\phi) - \tau_{zx}(\phi) = \sqrt{3}D\sin(\phi + \tfrac{\pi}{6})/c \tag{5.7}$$

Then, we define the following *phase rotation factors* composed of the above phase compensating components with respect to the signal that arrives from a direction ϕ.

$$G^{(\omega)}_{x2y}(\phi) \equiv e^{-j\omega\tau_{x2y}(\phi)} \tag{5.8}$$

$$G^{(\omega)}_{z2y}(\phi) \equiv e^{-j\omega\tau_{z2y}(\phi)} \tag{5.9}$$

Using these *phase rotation factors*, we define the following *integrated cross spectrum*.

$$\begin{aligned} G^{(\omega)}_{\phi,\theta} &= G^{(\omega)}_{x2y}(\phi)G^{(\omega)}_{xy}(\theta) \\ &\quad + G^{(\omega)}_{yz}(\theta) + G^{(\omega)}_{z2y}(\phi)G^{(\omega)}_{zx}(\theta), \end{aligned} \tag{5.10}$$

[1]The derivation of η_m and T is explained in [66] and [69].

where

$$G_{\mathtt{Y}}^{(\omega)}(\theta) = \frac{\hat{G}_{\mathtt{Y}}^{(\omega)}(\theta)}{|\hat{G}_{\mathtt{Y}}^{(\omega)}(\theta)|} \quad (\mathtt{Y} = xy, yz, zx) \tag{5.11}$$

are the filtered cross spectra by the whitening prefilter which is generally called Phase Transform (PHAT) [62].

Now for the $G_{\phi,\theta}^{(\omega)}$, following theorem is derived.

[Theorem]

In general,

$$|G_{\phi,\theta}^{(\omega)}| \leq 3 \tag{5.12}$$

and the equality is satisfied if and only if $\phi = \theta$.

[Proof]

Because the magnitudes of both *phase rotation factors* and normalized (pre-filtered) cross spectra are unity, the following relation holds.

$$\begin{aligned} |G_{\phi,\theta}^{(\omega)}| &= |G_{x2y}^{(\omega)}(\phi)G_{xy}^{(\omega)}(\theta) + G_{yz}^{(\omega)}(\theta) + G_{z2y}^{(\omega)}(\phi)G_{zx}^{(\omega)}(\theta)| & (5.13)\\ &\leq |G_{x2y}^{(\omega)}(\phi)G_{xy}^{(\omega)}(\theta)| + |G_{yz}^{(\omega)}(\theta)| + |G_{z2y}^{(\omega)}(\phi)G_{zx}^{(\omega)}(\theta)| & (5.14)\\ &= 3 & (5.15) \end{aligned}$$

The equality between Eq.(5.13) and Eq.(5.14) is satisfied if and only if the three complex terms are equal. By the proof of lemma as given in Appendix B, the arguments of three terms take the same value only at $\phi = \theta$. Therefore, the theorem holds.■

From this theorem, we define the following non-negative performance index which takes its global minimum at $\phi = \theta$.

$$Q_{\phi,\theta}^{(\omega)} = 9 - \left|G_{\phi,\theta}^{(\omega)}\right|^2 \geq 0 \tag{5.16}$$

Thus, the tracking problem results in searching ϕ that satisfies $Q_{\phi,\theta}^{(\omega)} = 0$.

5.3.3 Steepest descent method using harmonics

The speaker direction tracking is achieved by minimizing a performance index that consists of the combination of $Q_{\phi,\theta}^{(\omega)}$ for several selected harmonic frequencies

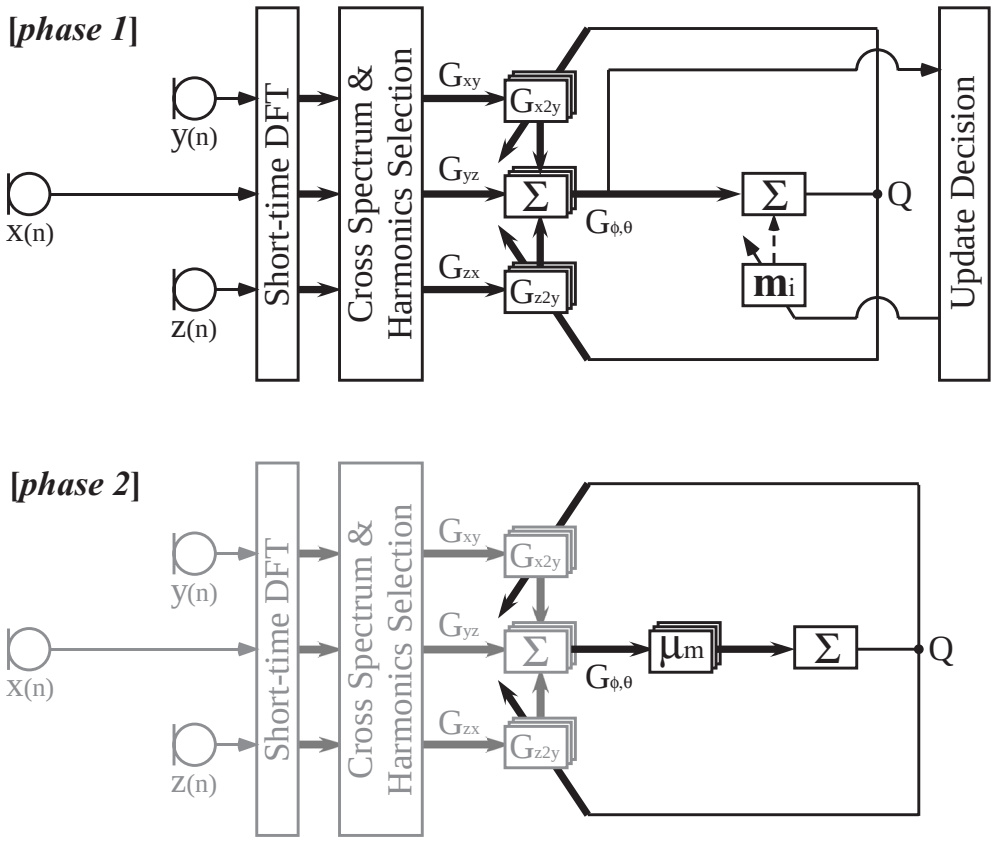

Figure 5.2: Flow diagram of the proposed method

$\omega = \omega_m (m \in \mathcal{M})$. The minimization is performed by steepest descent method. As stated in Sec. 5.1, we aim at tracking abruptly moving speaker direction while avoiding the local minimum convergence in the adaptation. To realise it, the adaptation process takes two steps. The first step aims at achieving fast and global convergence, and the second step contributes to improve the estimation accuracy. These two adaptation steps are combined according to the state of convergence. In *Phase 1*, we select $\omega \in \{\omega_m | m \in \mathcal{M}\}$, at which $Q^{(\omega)}_{\phi,\theta}$ provides global convergence in the steepest descent algorithm. The performance index in this phase is the sum of the selected $Q^{(\omega)}_{\phi,\theta}$s. In contrast, in *Phase 2*, the weighted sum of all $Q^{(\omega)}_{\phi,\theta}$ for $\omega = \omega_m (m \in \mathcal{M})$ is used as its performance index. We show the flow diagram of the proposed method in Fig. 5.2.

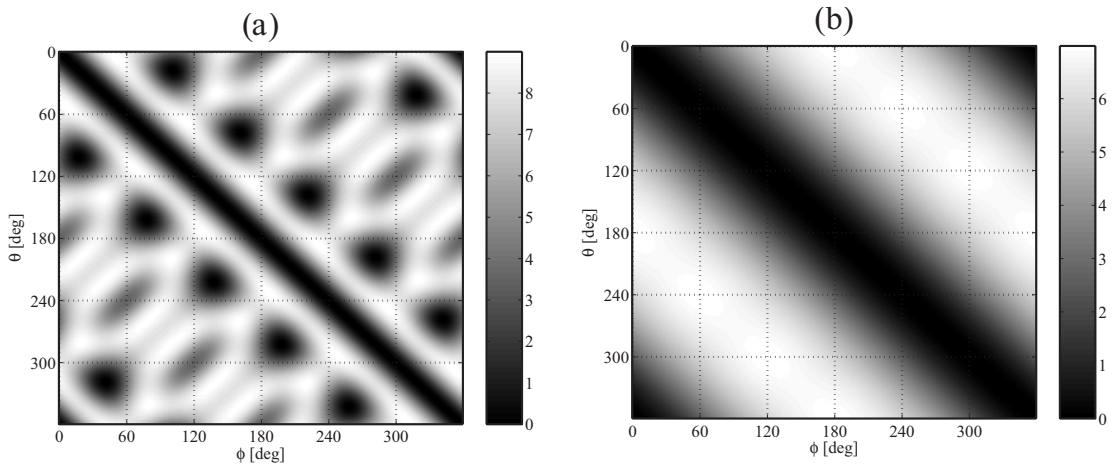

Figure 5.3: Performance index $Q_{\phi,\theta}^{(\omega)}$ at (a)$\omega = \omega_{max}$ (b)$\omega = 0.25\omega_{max}$

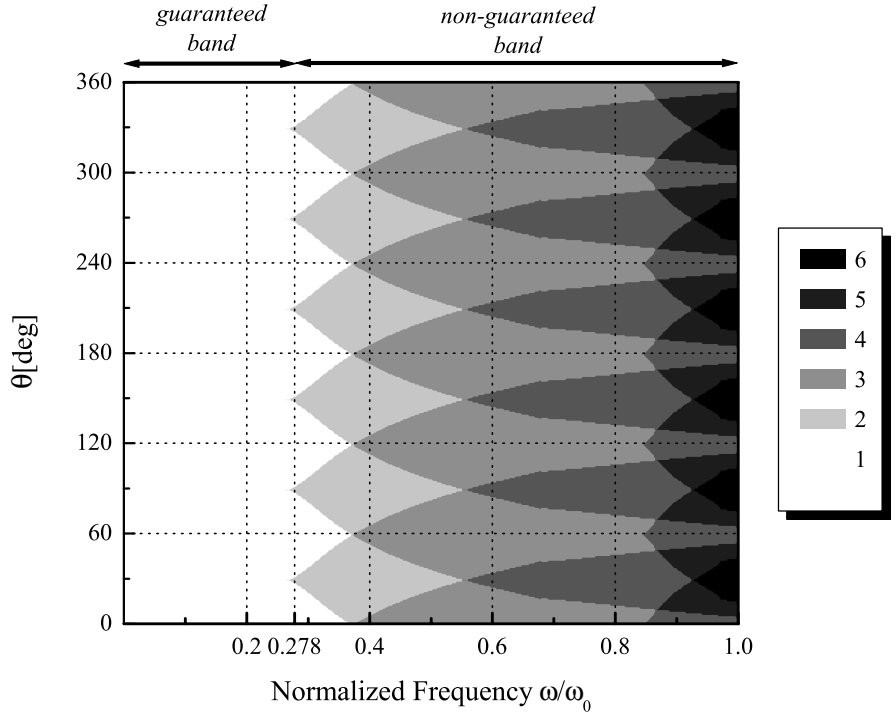

Figure 5.4: Number of local minima in the performance index $Q_{\phi,\theta}^{(\omega)}$

[*Phase 1*] Steepest descent method with performance index selection

Fig. 5.3 shows the feature of $Q_{\phi,\theta}^{(\omega)}$ with respect to ϕ and θ at different frequencies. Fig. 5.4 shows the number of local minima in $Q_{\phi,\theta}^{(\omega)}$ for all ω and θ. We find unique local minimum at the lower band, and so we have verified that the global convergence is guaranteed at the lower band $\omega \leq 0.278\omega_{max}$ (We call this band as "*guaranteed band*"). In contrast, $Q_{\phi,\theta}^{(\omega)}$ around $\phi \approx \theta$ steeply decreases as ω goes to higher (even though it is "*non-guaranteed band*"), so that the convergence speed would be faster.

Motivated by these $Q_{\phi,\theta}^{(\omega)}$'s features, we propose the following recursive method to obtain optimal ϕ. We start to update ϕ using the performance index within the *guaranteed band*, then, use that in the *non-guaranteed band* by switching the set of selected harmonics' indexes $\mathbf{m}_i \subset \mathcal{M}$ according to the convergence rate. Now we update the ϕ by

$$\phi_{i+1} = \phi_i - \frac{\rho}{[\mathbf{m}_i]} \sum_{m \in \mathbf{m}_i} \nu(\omega_m) \frac{\partial Q_{\phi,\theta}^{(\omega_m)}}{\partial \phi}, \tag{5.17}$$

where i and ρ are the iteration index and the stepsize parameter, respectively. The $[\mathbf{m}_i]$ denotes the number of elements in the set $\mathbf{m}_i$. Since $\partial Q/\partial\phi$ is nearly proportional to ω^2, we normalize it by $\nu(\omega) = \omega_{max}/\omega$. $\mathbf{m}_i$ is utilized for ϕ_i update, and it is modified by the following decision rules as shown in Fig. 5.5.

[Initial setting]

For the initial $\mathbf{m}_0$, we use a set of harmonic frequencies in the *guaranteed band*

$$\mathbf{m}_0 = \{m | \omega_m < \alpha\omega_{max}\}. \tag{5.18}$$

α should be $\alpha_{min} < \alpha < 0.278$, where α_{min} is determined to assure that the fundamental frequency ω_1 is contained in $\mathbf{m}_0$. In later simulation and experiment, we set $\alpha = 0.25$.

[Update of $\mathbf{m}_i$]

We update $\mathbf{m}_i$ as

$$\mathbf{m}_i = \mathbf{m}_{i-1} \cup \min\langle\overline{\mathbf{m}_{i-1}}\rangle \tag{5.19}$$

[2]See the Appendix C for the derivation.

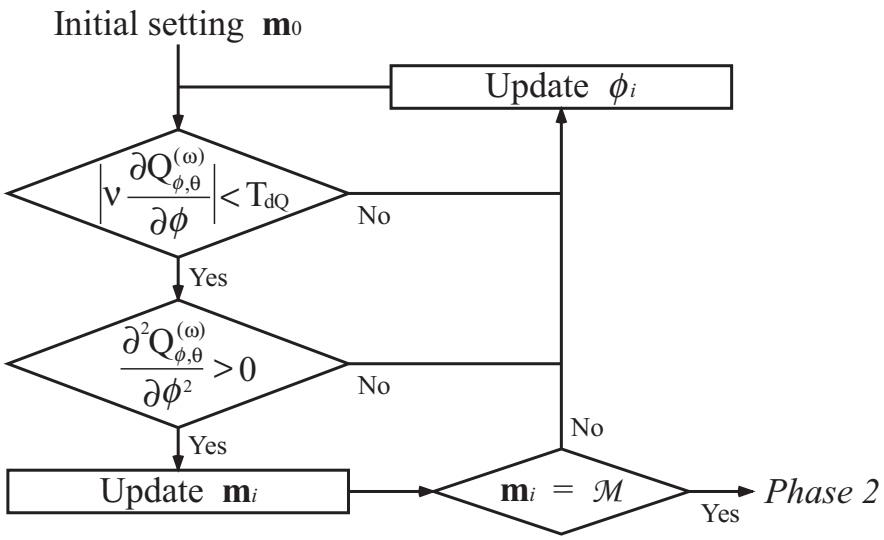

Figure 5.5: Decision rule of the $\mathbf{m}_i$ update in the *Phase 1*

if the following conditions in Eq.(5.20) and Eq.(5.21) are simultaneously satisfied at $\phi = \phi_i$.

$$\left| \nu(\omega_{\max\langle \mathbf{m}_{i-1}\rangle}) \frac{\partial Q^{(\omega_{\max\langle \mathbf{m}_{i-1}\rangle})}_{\phi_i,\theta}}{\partial \phi} \right| < T_{dQ} \tag{5.20}$$

$$\frac{\partial^2 Q^{(\omega_{\max\langle \mathbf{m}_{i-1}\rangle})}_{\phi_i,\theta}}{\partial \phi^2} > 0, \tag{5.21}$$

where $\overline{\mathbf{m}_{i-1}}$ is the complement set of $\mathbf{m}_{i-1}$ and $\max / \min\langle \mathbf{m}_{i-1}\rangle$ means the maximum/minimum element in the real integer set $\mathbf{m}_{i-1}$. T_{dQ} is the threshold about the steepness of the performance index to decide the update of $\mathbf{m}_i$.

[Termination of *Phase 1*]

The *Phase 1* is terminated if the following condition is satisfied.

$$\mathbf{m}_i = \mathcal{M} \tag{5.22}$$

[*Phase 2*] Weighted steepest descent method based on SNR

In *Phase 2*, we adopt the following weighted steepest descent method given by

$$\phi_{i+1} = \phi_i - \frac{\rho}{[\mathcal{M}]} \sum_{m \in \mathcal{M}} \mu_m \nu(\omega_m) \frac{\partial Q_{\phi,\theta}^{(\omega_m)}}{\partial \phi}. \tag{5.23}$$

The weight μ_m is given by

$$\mu_m = \frac{\eta_m}{\sum_m \eta_m}, \tag{5.24}$$

where η_m is the SNR at ω_m. This weighting aims to improve the accuracy of final estimation result.

5.3.4 Determination of threshold T_{dQ}

The value of T_{dQ} is a critical factor to ensure the global convergence. As we can find in Fig. 5.3 and Fig. 5.4, the range of ϕ guaranteeing global convergence is narrowest at $\omega = \omega_{max}$. Therefore, the determination of T_{dQ} should be performed as follows. Fig. 5.6 illustrates an example to explain how the T_{dQ} is determined.

$$0 < T_{dQ} < \min_{m \in \mathcal{M}} |\Gamma(\omega_m)|, \tag{5.25}$$

where

$$\Gamma(\omega) \equiv \nu(\omega) \left. \frac{\partial Q_{\phi,\theta}^{(\omega)}}{\partial \phi} \right|_{\phi=\phi_{local}}. \tag{5.26}$$

ϕ_{local} is the nearest ϕ to θ among the ϕ's at which $Q_{\phi,\theta}^{(\omega_{max})}$ takes local maximum. Fig. 5.7 shows the analytically calculated values of $|\Gamma(\omega)|(\omega \in$ *guaranteed band*$)$. Since $|\Gamma(\omega)|$ monotonically decreases as ω decreases, $\min_{m \in \mathcal{M}} |\Gamma(\omega_m)|$ is determined by the lowest harmonic component $\omega_m (m = \min\langle \mathcal{M} \rangle)$ in the input speech signal.

5.4 Evaluation with simulation results

To evaluate the performance of the proposed method and to investigate parameter setting, we performed computer simulation using real 5 phoneme data

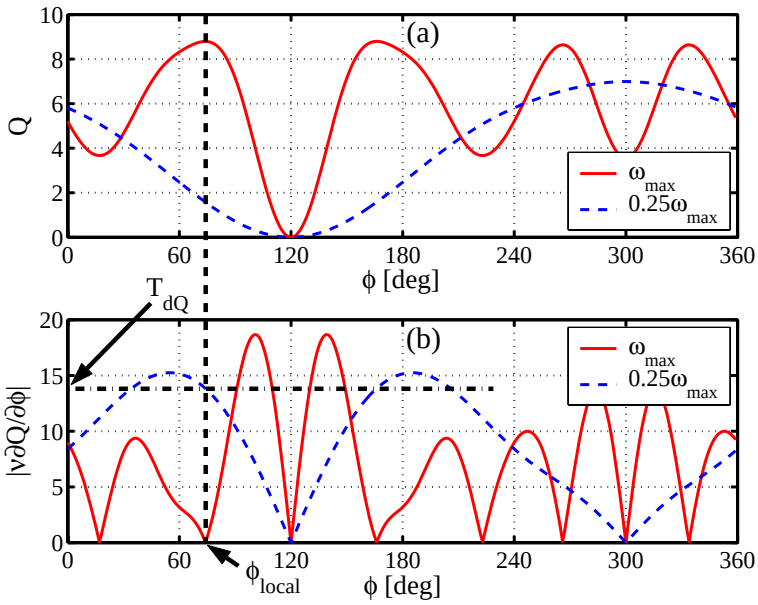

Figure 5.6: Example of T_{dQ} determination : (a)Q (b)$\left|\nu\frac{\partial Q}{\partial \phi}\right|$ ($\theta = 120°$)

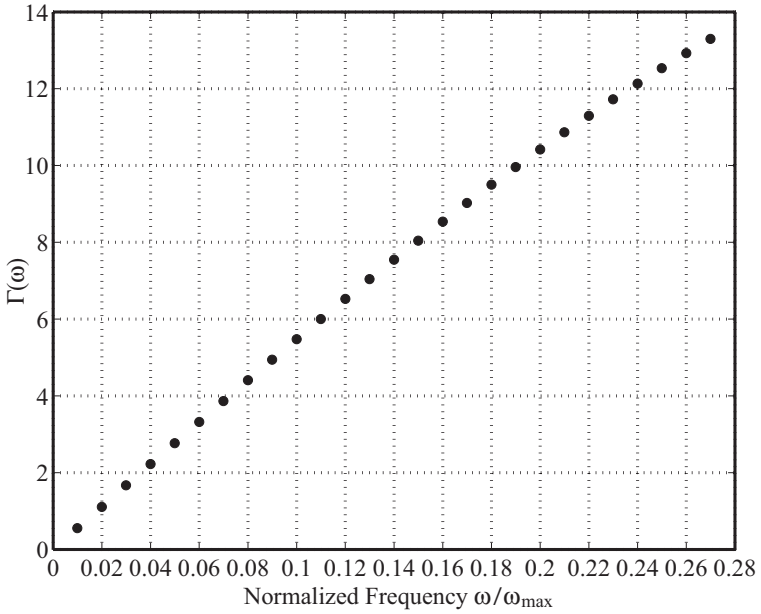

Figure 5.7: Critical value of T_{dQ} to assure the global convergence

Table 5.1: Parameters for simulation

Input SNR	20dB
Sampling Frequency	16000Hz
Wave Velocity c	340m/s
Microphone Distance D	0.08m
T	15dB
T_{dQ}	2
α	0.25
Window	Hamming
FFT point	1024
Frame Length	512 samples
Frame Overlap	384 samples
Data Length	128ms

(/a/,/e/,/i/,/o/,/u/) uttered by 10 subjects (5 each for male and female) as the source signal. The array input signal is generated by delaying the source signal with an appropriate delay samples according to θ [4] and white noise is added as the sensor noise. For quantitative evaluation criterion, we use the deviation of estimation error (DEE) defined by

$$DEE = \sqrt{\overline{|\phi_{\text{final}} - \theta_T|^2}}, \tag{5.27}$$

where ϕ_{final} and θ_T are the final estimated and true DOAs respectively, and $\overline{\cdot}$ denotes average calculation. Table 5.1 shows the parameters used in the simulation.

5.4.1 Stepsize determination

The determination of appropriate stepsize ρ is an important subject because there is a typical trade off between stability and speed of convergence. Fig. 5.8 shows the simulation results for various stepsizes, numbers of iterations and the initial direction in adaptation. From these results, we can find that the estimation process converges faster at large stepsize, but the accuracy of final estimated value

Table 5.2: Conditions for simulated reverberant room

Case	β	T_R[sec]
I	0.5	0.25
II	0.6	0.33
III	0.7	0.43
IV	0.8	0.73

is degraded due to the fluctuation around the optimal value. Taking a compromise about the tradeoff, we set the stepsize $\rho = 2$ in the following simulation.

5.4.2 Directional uniformity in accuracy

To confirm the estimation accuracy and its uniformity for omni-direction, we evaluate the results at several different acoustic environments simulated using the image method [78]. The room model for the simulated reverberant condition is summarised in Fig. 5.9, and for the reflection coefficients β in [78], we take the values as shown in Table 5.2 except for the values relating to ceiling and floor which are fixed at 0.5 in every case. We also denote the approximate reverberation time T_R calculated using the given impulse response in Table 5.2. The other parameters are the same as given in Table 5.1 except that the threshold T is 10dB. Fig. 5.10 shows the estimation accuracy after 500 iterations starting from the furthest position, i.e. $\phi_0 = \theta + 180°$. At the anechoic case, we can confirm that the proposed method keeps the spatially uniform estimation accuracy at every speaker direction and such performance can be found even in reverberated conditions.

5.5 Experiments at real acoustic environment

In this section, we will show some experimental results and discussions to verify the effectiveness of the proposed method in real acoustic environments. The experiments were performed in a reverberant conference room whose size is given in Fig. 5.11. The speech data and parameters are the same as in the preceding com-

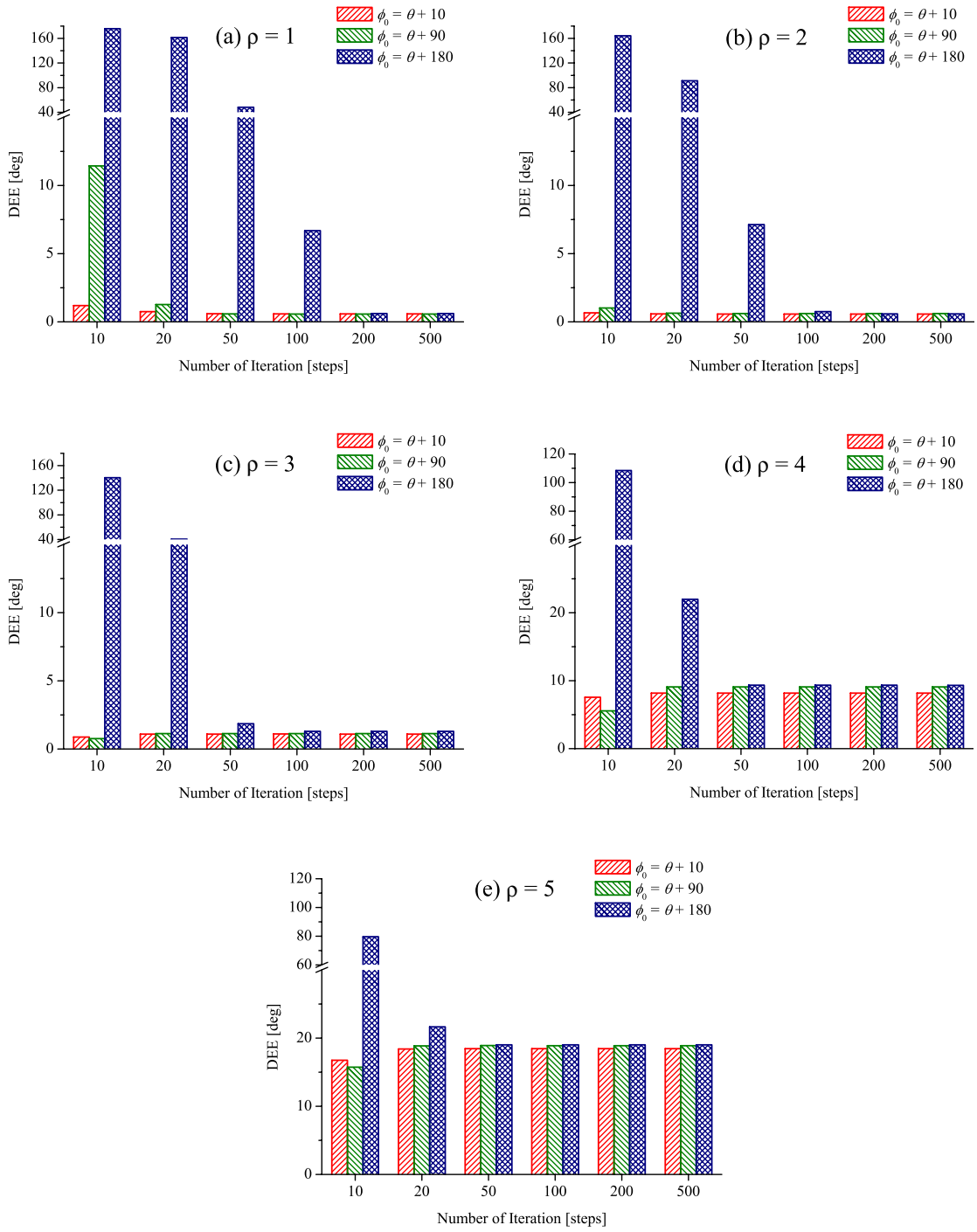

Figure 5.8: Estimation accuracy at different stepsize and number of iteration

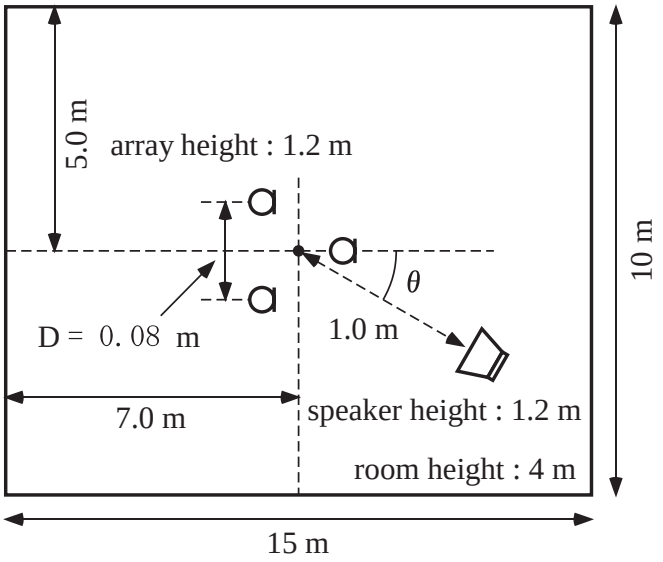

Figure 5.9: Model for reverberant room simulation

puter simulation except for the input SNR lying around 18dB and the threshold T is 10dB.

5.5.1 Evaluation of estimation accuracy

Experiments were undertaken to evaluate the estimation accuracy with the same conditions that applied in the computer simulation with 5 trials for each data. We also measured the influence of the deviation of the elevation angle ψ between the speaker's location and array plane due to the fact that speakers often deviate from the array plane in real situations. We regard the mean value of estimation results as the true direction θ_T and evaluate the DEE around this value.

The DEEs shown in Fig. 5.12 reveal that the proposed method is sufficiently effective even at real acoustic environment. Furthermore, the non zero elevation angle deviation from the array plane gives little influence to the estimation accuracy.

5.5.2 Comparison to the conventional method

Now we confirm the global convergence of the proposed method with a comparison to the conventional method. Among many conventional methods, we take

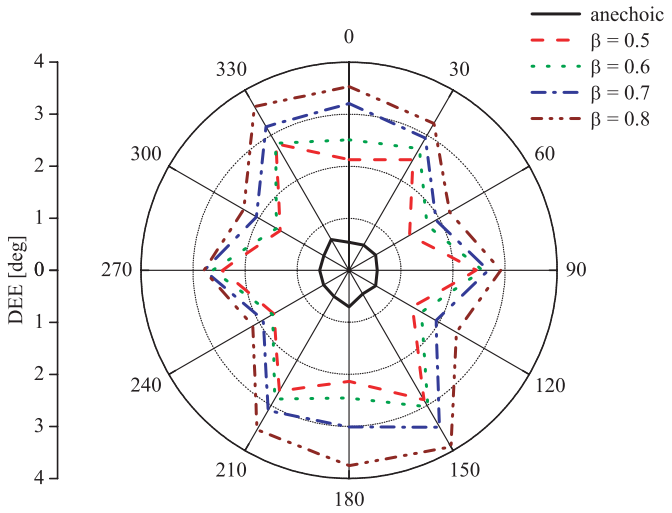

Figure 5.10: DEEs of computer simulation

the method [45] for impartial comparison. One reason is that the method [45] does not calculate beamformer's gain pattern in the same way as our proposed method. Another reason is that it directly uses direction angle as the parameter of adaptive process. In the conventional method, we used the data received at the microphones Y and Z in the Fig. 5.11. The convergence state of both the conventional and the proposed methods are shown in Fig. 5.13 and Fig. 5.14, respectively. Here the speaker is located at $\theta = 0$[deg] and the initial parameter ϕ_0 is settled at different directions. From these results, the proposed method succeeds in converging to the global optimum wherever the initial parameter is settled. In contrast, the conventional method tends to converge at the nearest local optimum around the initial value. This is because the method [45] assumes gradual speaker's movement. Therefore, it does not take any countermeasure to cope with local optimum convergence problem.

5.5.3 Examples of tracking to abruptly alternating speaker directions

In this subsection, we show some experimental results in practical cases. Fig. 5.15 shows the experimental result of tracking 5 speakers whose locations and pronunciation of vowels are given in Table 5.3. For the received data, we performed

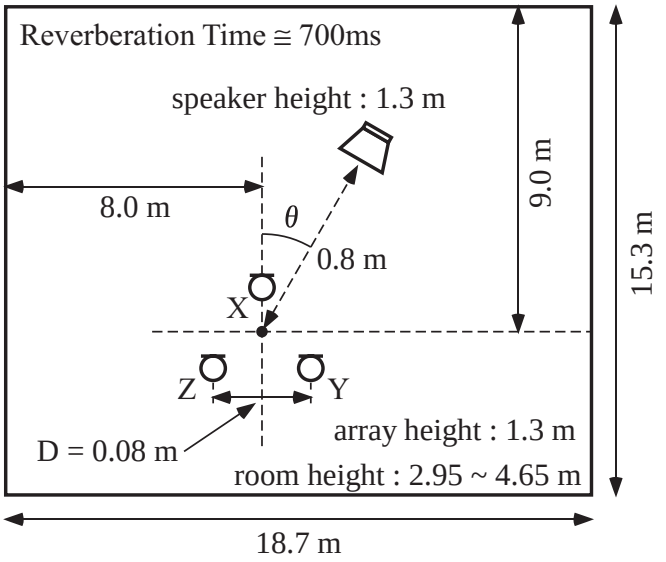

Figure 5.11: Acoustic environment for experiment

Table 5.3: Position of speakers

	Direction	Period (ms / frame*)	Phoneme
Speaker 1	0	0 – 512 / 1 – 16	Female /a/
Speaker 2	180	513 – 768 / 17 – 24	Male /i/
Speaker 3	90	769 – 1152 / 25 – 36	Female /u/
Speaker 4	−120	1153 – 1408 / 37 – 44	Male /e/
Speaker 5	−60	1409 – 1920 / 45 – 59	Female /o/

* Number of frame set

the proposed method with 50 iteration on every set of 4 frames, and the initial parameter ϕ_0 is settled at 60[deg]. From the result, the method promptly tracks the speakers' directions even they alternate abruptly.

Finally in Fig. 5.16 and Fig. 5.17, we show examples of tracking results for moderately and abruptly moving speaker directions respectively. Here the speakers utter general Japanese sentences whose contents are given in Table 5.4, and the data length for each frame set is 1sec. These satisfactory results confirm the ability of the method in practical use.

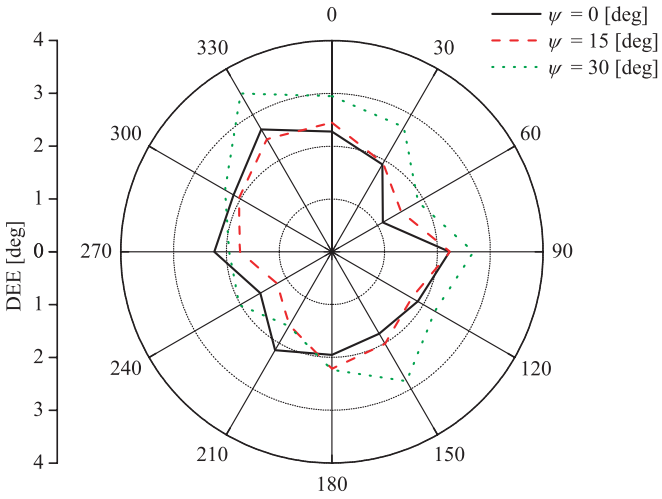

Figure 5.12: DEEs of experiment

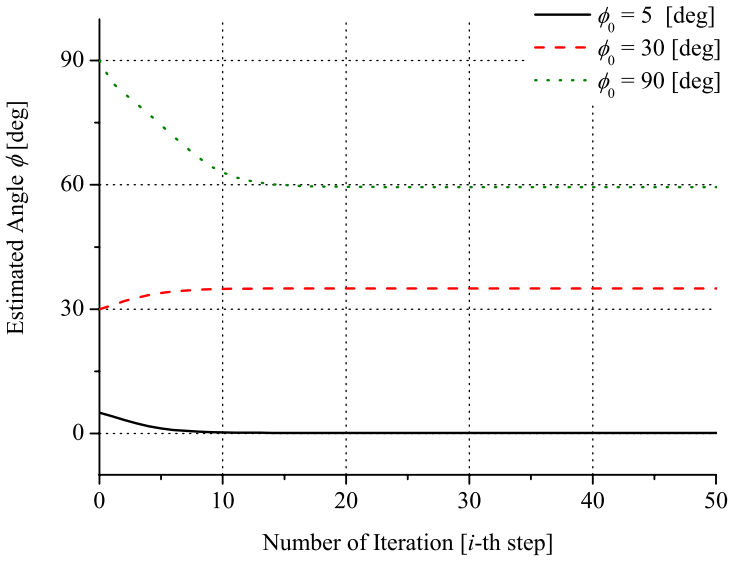

Figure 5.13: Adaptation profile of the conventional method

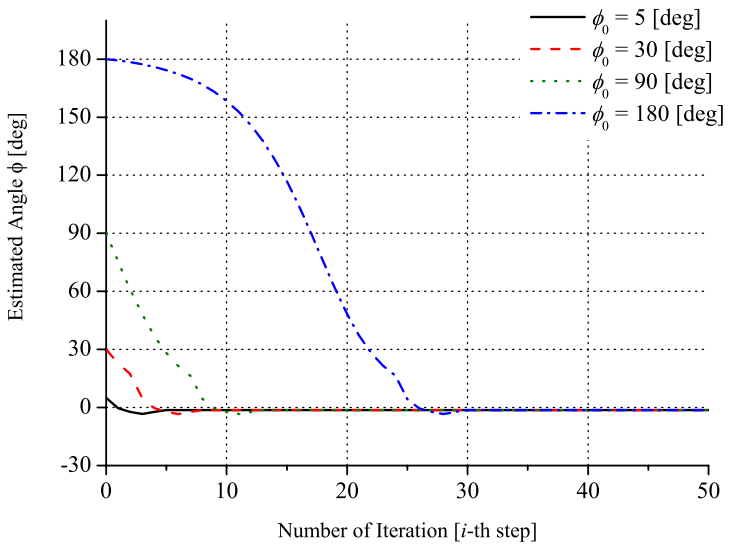

Figure 5.14: Adaptation profile of the proposed method

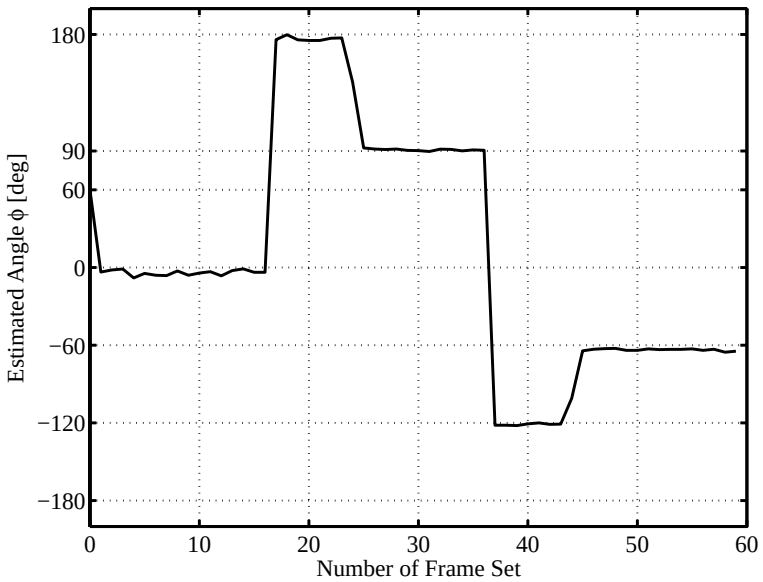

Figure 5.15: Tracking result of abruptly alternating speakers

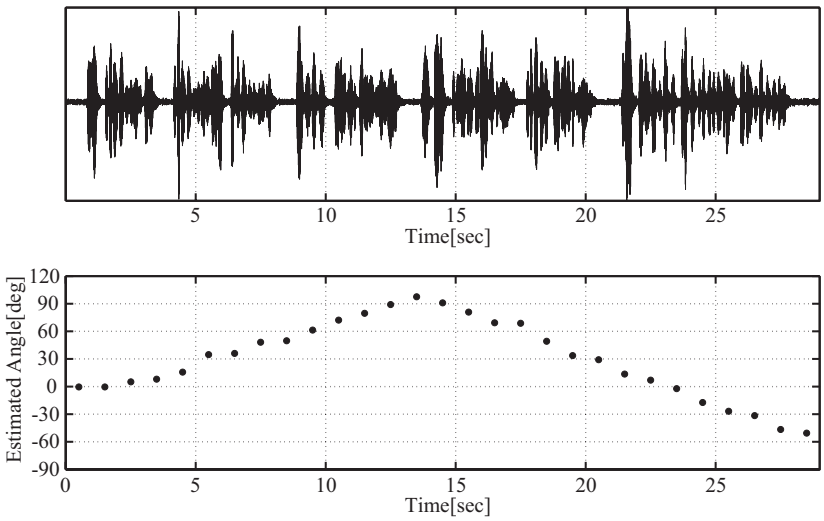

Figure 5.16: Example of tracking a generally moving speaker direction (A male speaker moves from 0° to 90° and goes backward up to −50°)

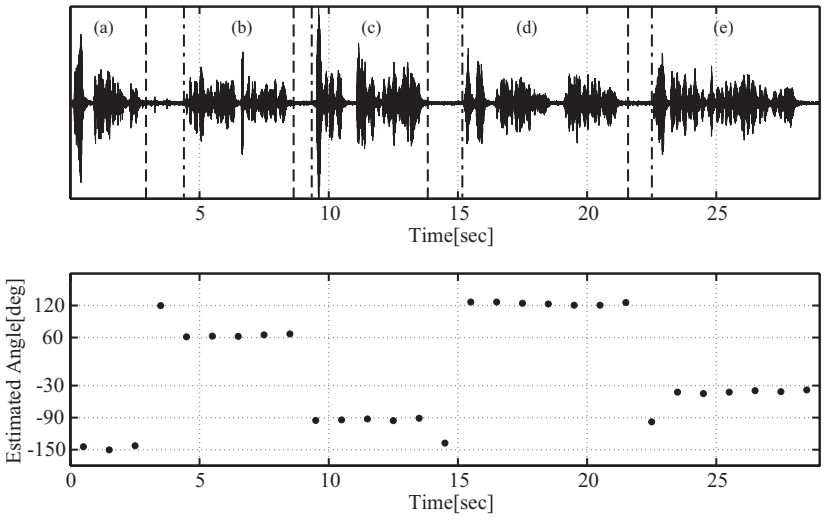

Figure 5.17: Example of tracking abruptly alternating speakers' direction (5 speakers surrounding the microphone array speak alternately : (a)−150° (b)60° (c)−90° (d)120° (e)−30°, Dash-dot line : Beginning of sentence, Dash line : End of sentence)

Figure 5.18: A scene of experiment in Fig. 5.16

Figure 5.19: A scene of experiment in Fig. 5.17

5.6 Conclusion of Chapter 5

In this chapter, we have proposed a new algorithm for speaker direction tracking. This is an extension of the method proposed in Chap. 4 and is designed to solve the tracking problem of both gradually and abruptly moving speaker directions. As stated in Sec. 5.3, the speaker direction is estimated by minimizing the performance index derived by the three pairs of microphones in the equilateral-triangular microphone array, using the steepest descent algorithm. In the adaptation procedure, we introduced an idea to avoid the local minimum convergence by exploiting the harmonic structure of speech signal. In Sec. 5.4 and Sec. 5.5, some computer simulations and experimental results show that the method keeps uniform and high accuracy for omni-direction, and it can track the speaker direction even if it moves abruptly. An extension of the studies would be to use the detection of voiced speech segments to our method for more practical use. The tracking problem for multiple and simultaneous speakers should also be considered.

Table 5.4: Sentences used in the experiment (English translation is given in the bracket)

- **Speaker A**
  Korewa wasya houkou tsuiseki no jikken desu.
  (This is an experiment of speaker direction estimation.)
- **Speaker B**
  Wasya houkou tsuiseki no mondai wa hutatsu no ke-su ni bunrui dekimasu.
  (The problem of speaker direction tracking is classified into two cases.)
- **Speaker C**
  Dai-ichi no ke-su wa hitori no wasya ga yukkuri to idou suru youna ba-ai desu.
  (Tracking a speaker moving moderately is the first case.)
- **Speaker D**
  Mou ippou wa hukusuu no wasya ga koutai de hanasutoki no youni, wasya houkou ga kyuugeki ni hennka suru ba-ai desu.
  (Another case is the tracking of abruptly moving speaker directions that occurs in the case of multiple speakers utter alternately.)
- **Speaker E**
  Kokode wa kousya no ke-su ni tsuite tadashi ku wasya tsuiseki wo okonau syuhou ni tsuite kenkyuu shite imasu.
  (Here we deal with the subject to track the speaker in the latter case.)

Chapter 6

Concluding Remarks

This dissertation is a summary of the research into the acquisition of voice activity interval, speaker direction and its tracking using microphone array. As the applications of digital signal processing to multimedia signals are becoming a reality, the speech signal is expected to be the main interface between humans and machines. Among several features in speech signal, the temporal and spatial information, i.e. **"When the speaker utters?"** and **"Where the speaker is?"**, are often required by various applications. In this research, new strategies for acquiring this information about the speaker are proposed.

The contents in this dissertation are summarised as followings. Chap. 1 states the background and the purpose of the research. Chap. 2 is the summary of the technologies relating to speech signal processing, especially focusing on the array signal processing which is the core tool of our research. At the beginning of Chap. 2, the features of speech signal, which are based on its generating mechanism, are explained. Then some time-frequency analysing techniques are referred to. The latter part of the chapter is spent on an explanation of the principle of microphone array signal processing and some conventional techniques including beamforming, direction estimation and tracking.

Chap. 3 to Chap. 5 focused on the following three topics.

- Chap. 3 $\cdots$ Voice activity detection
- Chap. 4 $\cdots$ Spaeker direction estimation
- Chap. 5 $\cdots$ Speaker direction tracking

Chapter 3 : Voice Activity Detection by the combined use of speech signal features in multiple signal domains

The proposed method applied the array signal processing to the input signal represented in the wavelet domain for achieving the combination of speech signal features in temporal, spectral, and spatial signal domains. As a result of this new strategy, we succeeded in discriminating the segments dominated by nonstationary interferences which most of the conventional VAD methods were not able to cope with. From the simulation results, the method keeps its discriminability even if the desired speech and undesired nonstationary interference arrive from close direction, i.e. around 10°. Added to this, we confirmed that the method could detect, not only, the voiced sound segments but also the very short time intervals dominated by unvoiced components. This is because the method takes account of the feature difference between voiced and unvoiced sounds.

Chapter 4 : Speaker direction estimation by the equilateral-triangular microphone array

A new method for speaker direction estimation is proposed. The purpose of this study is to achieve the omni-directionality of estimation accuracy and discriminability. To achieve this, we introduced the equilateral-triangular microphone array, and integrated the data in each of three different microphone pairs in the array. The proposed method makes use of the harmonic structure in the speech signal spectrum to improve the estimation accuracy. Through some simulations and experiments, we confirmed that the method is able to specify the speaker direction with uniform accuracy for omni-direction at the DEEs of a few degrees.

In addition, the extended DOA estimation method using four microphones in the tetrahedral arrangement was proposed. This method achieved the estimation of both azimuth and elevation angles without the increase of calculation load and the loss of accuracy.

Chapter 5 : An expanded scheme of direction estimation to realise tracking of abruptly changing speaker directions

We developed the tracking algorithm of speaker direction as a refinement of the method in the above study. In this research, the system has been improved to make it possible to track both abruptly and moderately moving speaker directions. The method searches for the speaker direction using the steepest descent method. During the adaptation, this strategy alters the performance index in order to avoid the convergence to the local optima, which occurs because the shape of the object function varies depending on the frequency band.

In the computer simulations and experiments, the method has shown the proper convergence property to the correct angle with sufficient accuracy wherever the initial parameter of the adaptation is set. This is the critical advantage of the strategy that was impossible to be achieved by existing methods. Furthermore, we have confirmed the effectiveness for tracking both abruptly and moderately moving speaker directions.

In conclusion, we proposed and discussed new strategies for voice activity detection and speaker direction estimation. In this research, we considered various speech signal features, and tried to integrate them into our strategy. For future study, a method to combine both the VAD and speaker direction tracking technologies should be considered to devise an online system for speaker information collection. Finally, the proposed methods will be beneficial for various applications such as speech enhancement, videoconferences, and so on.

Appendix A

Example of the subspace analysis

Here we demonstrate the features of the eigenvector with a simple example using two sensor array model. In this case, the covariance matrix is given by

$$\mathbf{R} = P_1 \begin{bmatrix} 1 & e^{j\phi_1} \\ e^{-j\phi_1} & 1 \end{bmatrix} + P_2 \begin{bmatrix} 1 & e^{j\phi_2} \\ e^{-j\phi_2} & 1 \end{bmatrix}, \tag{A.1}$$

and the eigenvalues and eigenvectors are derived as

$$\begin{aligned} \lambda &= (P_1 + P_2) \pm \sqrt{P_1^2 + P_2^2 + 2P_1P_2\cos\psi}, \\ \psi &= \phi_1 - \phi_2. \end{aligned} \tag{A.2}$$

Then the eigenvector corresponding to the larger eigenvalue $\lambda_1 = (P_1 + P_2) + \sqrt{P_1^2 + P_2^2 + 2P_1P_2\cos\psi}$ is denoted as

$$\mathbf{v}_1 = \begin{bmatrix} 1 & e^{-j\arg(P_1e^{-j\phi_1} + P_2e^{-j\phi_2})} \end{bmatrix}^T. \tag{A.3}$$

Thus, the direction of the $\mathbf{v}_1$

$$\arg(P_1e^{-j\phi_1} + P_2e^{-j\phi_2}) \tag{A.4}$$

is determined by the dominating component.

Appendix B

Proof of Lemma in Sec. 4.3.2

By the following three relations, the compensated delays $\breve{\tau}_{xy}(\phi,\theta)$, $\breve{\tau}_{yz}(\theta)$ and $\breve{\tau}_{zx}(\phi,\theta)$ are equal if and only if $\phi=\theta$.

[Relation 1]

The equality $\breve{\tau}_{xy}(\phi,\theta)=\breve{\tau}_{yz}(\theta)$ holds at $\phi=\theta$ and $\phi=-\frac{2}{3}\pi-\theta$. Substituting $\phi=-\frac{2}{3}\pi-\theta$ for the delay compensation, the another equality does not hold as

$$\breve{\tau}_{zx}(\phi,\theta)|_{\phi=-\frac{2}{3}\pi-\theta} \neq \breve{\tau}_{xy}(\phi,\theta)|_{\phi=-\frac{2}{3}\pi-\theta} = \breve{\tau}_{yz}(\theta). \tag{B.1}$$

[Relation 2]

The equality $\breve{\tau}_{yz}(\theta)=\breve{\tau}_{zx}(\phi,\theta)$ holds at $\phi=\theta$ and $\phi=\frac{2}{3}\pi-\theta$. Substituting $\phi=\frac{2}{3}\pi-\theta$ for the delay compensation, the another equality does not hold as

$$\breve{\tau}_{xy}(\phi,\theta)|_{\phi=\frac{2}{3}\pi-\theta} \neq \breve{\tau}_{yz}(\theta) = \breve{\tau}_{zx}(\phi,\theta)|_{\phi=\frac{2}{3}\pi-\theta}\,. \tag{B.2}$$

[Relation 3]

The equality $\breve{\tau}_{zx}(\phi,\theta)=\breve{\tau}_{xy}(\phi,\theta)$ holds at $\phi=\theta$ and $\phi=-\theta$. Substituting $\phi=-\theta$ for the delay compensation, the another equality does not hold as

$$\breve{\tau}_{yz}(\theta) \neq \breve{\tau}_{zx}(\phi,\theta)|_{\phi=-\theta} = \breve{\tau}_{xy}(\phi,\theta)|_{\phi=-\theta}\,. \tag{B.3}$$

Appendix C

The derivative of the performance index

The performance index and its derivative are derived as followings.

$$\begin{aligned} Q_{\phi,\theta}^{(\omega)} &= 9 - \left| G_{\phi,\theta}^{(\omega)} \right|^2 \\ &= 6 - 2\cos\left[\sqrt{3}\tfrac{D}{c}\omega\left\{\sin(\theta - \tfrac{\pi}{6}) - \sin(\phi - \tfrac{\pi}{6})\right\}\right] \\ &\quad -2\cos\left[\sqrt{3}\tfrac{D}{c}\omega\left\{\sin(\theta + \tfrac{\pi}{6}) - \sin(\phi + \tfrac{\pi}{6})\right\}\right] \\ &\quad -2\cos\left[\sqrt{3}\tfrac{D}{c}\omega\left\{\cos\theta - \cos\phi\right\}\right] \end{aligned} \tag{C.1}$$

$$\begin{aligned} \frac{\partial Q_{\phi,\theta}^{(\omega)}}{\partial\phi} &= -2\sqrt{3}\frac{D}{c}\omega\Big[\cos(\phi - \tfrac{\pi}{6})\sin\left\{\sqrt{3}\tfrac{D}{c}\omega\left[\sin(\theta - \tfrac{\pi}{6}) - \sin(\phi - \tfrac{\pi}{6})\right]\right\} \\ &\quad + \cos(\phi + \tfrac{\pi}{6})\sin\left\{\sqrt{3}\tfrac{D}{c}\omega\left[\sin(\theta + \tfrac{\pi}{6}) - \sin(\phi + \tfrac{\pi}{6})\right]\right\} \\ &\quad - \sin\phi\sin\left\{\sqrt{3}\tfrac{D}{c}\omega\left[\cos\theta - \cos\phi\right]\right\}\Big] \end{aligned} \tag{C.2}$$

In the Eq.(C.2), we find that the derivative is a function of frequency ω. Although it is not completely proportional to ω, the major influence of ω is caused by the multiplication positioned at outside of the bracket $[\cdot]$. Thus we regard that the derivative of the performance index is nearly proportional to the frequency ω.

Bibliography

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

[2] S. Furui, Digital Speech Processing, Synthesis and Recognition -Second Edition-, Marcel Dekker Inc., 2001.

[3] http://www.ibm.com/

[4] J. Ohga, Y. Yamazaki, and Y. Kaneda, Acoustic Systems and Digital Processing for Them, IEICE, Tokyo 1995. (in Japanese)

[5] M. Brandstein and D. Ward, Microphone Arrays, Springer-Verlag, 2001.

[6] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol.27, No.2, pp.113–120, Apr. 1979.

[7] T. Ishida and A. Taguchi, "A Noise Suppression Method by Data Dependent Wiener Filtering for Noisy Speech Signal," Proc. 19th SIP Symposium, No.A3-3, Nov. 2004. (in Japanese)

[8] D. H. Johnson and D. E. Dudgeon, "Array Signal Processing -Concepts and Techniques-," PTR Prentice Hall, 1993.

[9] O.L. Frost, "An Algorithm for Linearly Constrained Adaptive Array Processing," Proc. IEEE, Vol.60, No.8, pp.926–935, Aug. 1972.

[10] O. Hoshuyama and A. Sugiyama, "Robust Adaptive Beamforming" in Microphone Arrays, ed. M. Brandstein and D. Ward, pp.87–109, Springer-Verlag, Berlin, 2001.

[11] O. Hoshuyama and A. Sugiyama, "A Robust Generalized Sidelobe Canceller with a Blocking Matrix Using Leaky Adaptive Filters," IEICE Trans. Fundamentals, Vol.J79-A, No.9, pp.1516–1524, Sep. 1996. (in Japanese)

[12] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A Robust Adaptive Beamformer with a Blocking Matrix Using Coefficient-Constrained Adaptive Filters," IEICE Trans. Fundamentals, Vol.E82-A, No.4, pp.640–647, Apr. 1999.

[13] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A Robust Adaptive Beamformer for Microphone Arrays with a Blocking Matrix Using Constrained Adaptive Filters," IEEE Trans. Signal Processing, Vol.47, No.10, pp.2677–2684, Oct. 1999.

[14] O. Hoshuyama, B. Begasse, and A. Sugiyama, "A new adaptation-mode control based on cross correlation for a robust adaptive microphone array," IEICE Trans. Fundamentals, Vol.E84-A, No.2, pp.406–413, Feb. 2001.

[15] T. Kikuchi, T. Yamaoka, and N. Hamada, "A robust adaptive beamforming being superior to processing burst signals," The 6th IEEE Int. Workshop on Intelligent Signal Processing and Communication Systems, pp.772–776, 1998.

[16] D. V. Compernolle, W. Ma, F. Xie, and M. V. Diest, "Speech recognition in noisy environments with the aid of microphone array," Speech Communication, Vol.9, pp.433–442, 1990.

[17] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech Enhancement Based on the Subspace Method," IEEE Trans. Speech and Audio Processing, Vol.8, No.5, pp.497–507, Sep. 2000.

[18] A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.

[19] T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA," IEICE Trans. Fundamentals, Vol.E86-A, No.4, pp.846–858, Apr. 2003.

[20] H. Saruwatari, T. Kawamura, T. Nishikawa, and K. Shikano, "Fast-Convergence Algorithm for Blind Source Separation Based on Array Signal Processing," IEICE Trans. Fundamentals, Vol.E86-A, No.3, pp.286–291, Mar. 2003.

[21] M. Furukawa, Y. Hioka, T. Ema, and N. Hamada, "Introducing New Mechanism in the Learning Process of FDICA-based Speech Separation," Proc. IWAENC2003, pp.291–294, Sep. 2003.

[22] K. Kobayashi, K. Furuya, and A. Kataoka, "A Microphone Array System with Echo Canceller," IEICE Trans. Fundamentals, Vol.J87-A, No.2, pp.143–152, Feb. 2004. (in Japanese)

[23] http://www.itu.int/

[24] "Continuous Speech Corpus for Research Vol.1–3," Japan Information Processing Development Center, ©Shuichi Itahashi (Edited by the Acoustical Society of Japan) 1991.

[25] S. Sakakibara, Wavelet Beginners Guide, TDU Pub., Tokyo, 1995. (in Japanese)

[26] M. Vetterli and J. Kovacevic, Wavelets and Subband Coding, Prentice-Hall, 1995.

[27] F. Asano, H. Asoh, and T. Matsui, "Sound Source Localization and Separation in Near Field," IEICE Trans. Fundamentals, Vol.E83-A, No.11, pp.2286–2294, Nov. 2000.

[28] J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," J. Acoust. Soc. Am., Vol.78, No.5, pp.1508–1518, Nov. 1985.

[29] L. J. Griffiths and C. W. Jim, "An alternative approach to linear constrained adaptive beamforming," IEEE Trans. on Antennas and Propagation, Vol.30, No.1, pp.27–34, 1982.

[30] S. Fischer and K.U. Simmer, "Beamforming microphone arrays for speech acquisition in noisy environments," Speech Communication, Vol.20, No.3–4, pp.215–227, Dec. 1996.

[31] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein, "Robust Localization in Reverberant Rooms," in Microphone Arrays, ed. M. Brandstein and D. Ward, pp.157–180, Springer-Verlag, Berlin, 2001.

[32] C.H. Knapp and G.C. Carter, "The Generalized Correlation Method for Estimation of Time Delays," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol.24, No.4, pp.320–327, Aug. 1976.

[33] M.S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," J. Acous. Soc. America, Vol.105, No.5, pp.2914–2919, May. 1999.

[34] 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 Acoustics, Speech, and Signal Processing, Vol.33, No.4, pp.823–831, 1985.

[35] F. Asano and S. Hayamizu, "Speech Enhancement Using Array Signal Processing Based on the Coherent-Subspace Method," IEICE Trans. Fundamentals, Vol.E80-A, No.11, pp.2276–2285, Nov. 1997.

[36] F. Asano, H. Asoh, and T. Matsui, "Sound Source Localization and Separation in Near Field," IEICE Trans. in Fundamentals, Vol.E83-A, No.11, pp.2286–2294, Nov. 2000.

[37] T. Kikuchi, T. Yamaoka, and N. Hamada, "Microphone Array System with DOA Estimation by using Harmonic Structure of Speech Signals," IEICE Technical Report, DSP98-164, pp.23–28, Jan. 1999. (in Japanese)

[38] H. Hung and M. Kaveh, "Focussing Matrices for Coherent Signal-Subspace Processing," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol.36, No.8, pp.1272–1282, Aug. 1988.

[39] M.A. Doron and A.J. Weiss, "On Focusing Matrices for Wide-Band Array Processing," IEEE Trans. on Signal Processing, Vol.40, No.6, pp.1295–1302, Jun. 1992.

[40] W. Hong and A.H. Tewfik, "Focusing Matrices for Wideband Array Processing with No a Priori Angle Estimates," Proc. IEEE ICASSP-92, Vol.2, pp.493–496, Mar. 1992.

[41] H. Hung and M. Kaveh, "On the statistical sufficiency of the coherently averaged covariance matrix for the estimation of the parameters of wideband sources," Proc. IEEE ICASSP-87, Vol.12, pp.33–36, Apr. 1987.

[42] V.C. Raykar, R. Duraiswami, B. Yegnanarayana, and S.R. Mahadeva Prasanna, "Tracking a moving speaker using excitation source information," Proc. Eurospeech2003, pp.69–72, Sep. 2003.

[43] G. Nokas and E. Dermatas, "Speaker Tracking for Hands-Free Continuous Speech Recognition in Noise Based on a Spectrum-Entropy Beamforming Method," IEICE Trans. on Inf. & Syst., Vol.E86-D, No.4, pp.755–758, Apr. 2003.

[44] Y. Nagata and M. Abe, "Two-Channel Adaptive Microphone Array with Target Tracking," IEICE Trans. on Fundamentals, Vol.J82-A, No.6, pp.860–866, Jun. 1999. (in Japanese)

[45] H. Kawakami, M. Abe, and M. Kawamata, "A Two-Channel Microphone Array with Adaptive Target Tracking Using Frequency Domain Generalized Sidelobe Cancellers," IEEE Int. Symp. on Intelligent Sign. Process. & Commun. Systems, pp.291–296, Nov. 2002.

[46] Y. Hioka and N. Hamada, "Voice Activity Detection with Array Signal Processing in the Wavelet Domain," IEICE Trans. on Fundamentals, Vol.E86-A, No.11, pp.2802–2811, Nov. 2003.

[47] Y. Hioka and N. Hamada, "Voice Activity Detection with Array Signal Processing in the Wavelet Domain," The 2002 European Signal Processing Conference, pp.255–258, Sep. 2002.

[48] Y. Hioka and N. Hamada, "Voice Activity Detection Using Microphone Array Combining with Wavelet Analysis," Technical Report of IEICE, SP2001-126, pp.9–16, Jan. 2002. (in Japanese)

[49] I. Abdallah, S. Montresor, and M. Baudry, "Robust speech/non speech detection in adverse condition using an entropy based estimator," Proc. 13th International Conference on DSP Processing, Vol.2, pp.757–760, 1997.

[50] J. I. Agbinya, "Discrete wavelet transform techniques in speech processing," IEEE TENCON, vol.2, pp.514–519, 1996.

[51] S. Kadambe and G. F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," IEEE Trans. Information Theory, Vol.38, No.2, pp.917–924, 1992.

[52] J-F. Chen and W. Ser, "Speech detection using microphone array," Electronic Letters, Vol.36, No.2, pp.181–182, 2000.

[53] Y. Kaneda, "Speech period detection using a microphone array under noisy environments," Trans. IEICE Vol.J73-A, No.8, pp.1391–1398, 1990. (in Japanese)

[54] K. Kiyohara, Y.Kaneda, S. Takahashi, H. Nomura, and J. Kijima,"A microphone array system for speech recognition," Proc. IEEE ICASSP 97, Vol.1, pp.215–218, 1997.

[55] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol.34, No.6, pp.1391–1400, 1986.

[56] G. Su and M. Morf, "The signal subspace approach for multiple wide-band emitter location," IEEE Trans., Acoustics, Speech, and Signal Processing, Vol.31, No.6, pp.1502–1522, 1983.

[57] "RWCP Sound Scene Database in Real Acoustical Environments," Real World Computing Partnership, ©1998–2001.

[58] S.U. Pillai, Array Signal Processing, Springer-Verlag, 1989.

[59] N. Kikuma, Adaptive Signal Processing with Array Antenna, Science and Technology Publishing Company, Inc., 1999. (in Japanese)

[60] R.O. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," IEEE Trans. Antennas and Propagation, Vol.34, No.3, pp.276–280, 1986.

[61] G. Su and M. Morf, "Signal subspace approach for multiple wide-band emitter location," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol.31, No.12, pp.1502–1522, 1983.

[62] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," IEEE Trans. on Speech and Audio Processing, Vol.5, No.3, pp.288–292, 1997.

[63] S. Okada, F. Sato, and T. Morita, "3-Dimentional sound localization and voice separation by 3-microphone system –Theoretical Study and Simulation," Journal of the Institute of Systems, Control and Information Engineers, Vol.6, No.3, pp.149–155, 1993. (in Japanese)

[64] T. Okada, K. Mitomi, and H. Sato, "Direction measurement of a sound source in 2D space utilizing difference of sound levels given by two sets of condenser microphones," Journal of the Robotics Society of Japan, Vol.18, No.4, pp.569–575, 2000. (in Japanese)

[65] S. Tanigawa and N. Hamada, "Direction-of-Arrival Estimation of Speech Using Virtually Generated Multichannel Data from Two-Channel Microphone Array," IEICE Trans. Fundamentals, Vol.J85-A, No.2, pp.153–161, Feb. 2002.

[66] Y. Hioka, Y. Koizumi, and N. Hamada, "Improvement of DOA Estimation Using Virtually Generated Multichannel Data from Two-Channel Microphone Array," Journal of Signal Processing, Vol.7, No.1, pp.105–109, Jan. 2003.

[67] Y. Hioka, Y. Koizumi, and N. Hamada, "Improvement of DOA Estimation Method Using Virtually Generated Multichannel Data from Two-channel Microphone Array," Proc. of the ISITA2002, pp.735–738, Oct. 2002.

[68] Y. Koizumi, Y. Hioka, S. Tanigawa, and N. Hamada, "DOA Estimation Using Virtually Generated Multichannel Data from Two-channel Microphone array -Improvement by selecting harmonic elements-," Proc. The 2002 IEICE General Conference, No.A-4-69, p.179, Mar. 2002. (in Japanese)

[69] Y. Hioka and N. Hamada, "DOA Estimation of Speech Signal Using Microphones Located at Vertices of Equilateral Triangle," IEICE Trans. on Fundamentals., Vol.E87-A, No.3, pp.559–566, 2004.

[70] Y. Hioka and N. Hamada, "DOA Estimation of Speech Signal with Equilateral-Triangular Microphone Array," Proc. of Eurospeech2003, pp.1717–1720, Sep. 2003.

[71] Y. Hioka and N. Hamada, "DOA Estimation of Speech Signal using Microphones located at Vertices of triangle," Technical Report of IEICE, EA2003-44, pp.9–16, Jun. 2003.

[72] Y. Hioka and N. Hamada, "Estimation of azimuth and elevation DOA using microphones located at apices of regular tetrahedron," IEICE Trans. Fundamentals, Vol.E87-A, No.8, pp.2058–2062, Aug. 2004.

[73] Y. Hioka and N. Hamada, "Separate Estimation of Azimuth and Elevation DOA Using Microphones Located at Apices of Regular Tetrahedron," Proc. ICASSP2004, Vol.2, pp137–140, May. 2004.

[74] Y. Hioka and N. Hamada, "Estimation of speech DOA using microphones located at apices of regular tetrahedron," Proc. of 2004 Spring Meeting of ASJ, Vol.I, No.3-10-2, pp.585–586, Mar. 2004. (in Japanese)

[75] H. Kanai, Spectrum Analysis of Sound and Vibration, CORONA Pub., 1999. (in Japanese)

[76] M.R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," J. Acoust. Soc. Am. Vol.43, pp.829–834, 1968.

[77] N. Kikuma, "Beamformer Method," in Adaptive Signal Processing with Array Antenna, pp.178–181, Science and Technology Publishing Company, Inc., 1999. (in Japanese)

[78] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. America, Vol.65, No.4, pp.943–950, Apr. 1979.

[79] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The Fundamental Limitation of Frequency Domain Blind Source Separation for Convolutive Mixtures of Speech," IEEE Trans. on Speech and Audio Processing, Vol.11, No.2, pp.109–116, 2003.

[80] Y. Hioka and N. Hamada, "Tracking of speaker direction by integrated use of microphone pairs in equilateral-triangle," IEICE Trans. on Fundamentals, Vol.E88-A, No.3, Mar. 2005. (in press)

[81] Y. Hioka and N. Hamada, "A Tracking Algorithm of Speaker Direction using Microphones Located at Vertices of Equilateral-Triangle," Proc. 12th European Signal Processing Conference EUSIPCO2004, pp.1983–1986, Sep. 2004.

[82] Y. Hioka and N. Hamada, "Tracking of speaker direction by the integrated use of microphone pairs in equilateral-triangle," Technical Report of IEICE, EA2004-17, Vol.104, No.143, pp.1–6, Jun. 2004.

[83] K. Suyama and T. Tasaki, "A Study on Target Talker Tracking via Two Microphones," Proc. 18th IEICE DSP Symposium, A5-6, Nov. 2003. (in Japanese)

[84] M. Zhang and M.H. Er, "An alternative Algorithm for Estimating and Tracking Talker Location by Microphone Arrays," J. Audio Eng. Soc., Vol.44, No.9, pp.729–736, Sep. 1996.

[85] Y. Hioka and N. Hamada, "DOA Estimation of Speech Signal with a Small Number of Microphone Array at Real Acoustical Environment," Proc. IWAENC2003, pp.227–230, Sep. 2003.

[86] Y. Hioka and N. Hamada, "DOA Estimation of Speech Signal at a Reverberant Condition," Technical Report of IEICE, EA2002-111, pp.35–40, Jan. 2003. (in Japanese)