NAIST-IS-DT9661031

Doctor Thesis

Hands-free Speech Recognition Using a Microphone Array

Takeshi Yamada

March 24, 1999

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

Doctor Thesis

submitted to Graduate School of Information Science,

Nara Institute of Science and Technology

in partial fulfillment of the requirements for the degree of

DOCTOR of ENGINEERING

Takeshi Yamada

Thesis committee: Kiyohiro Shikano, Professor

Kunihiro Chihara, Professor

Yoh'ichi Tohkura, Professor

Satoshi Nakamura, Associate Professor

Hands-free Speech Recognition Using a Microphone Array*

Takeshi Yamada

Abstract

This thesis focuses on microphone arrays to realize hands-free speech recognition in real environments. In hands-free situations, users can speak at arbitrary positions while moving. Therefore, it is very important for high quality speech acquisition using microphone arrays to localize a talker accurately. However, it is very difficult to localize a moving talker accurately in noisy and reverberant environments. The talker localization errors result in performance degradation of speech recognition. One way to solve this problem is to integrate the speech recognition process and the talker localization into a unified framework. This thesis proposes a new speech recognition algorithm based on a 3-dimensional Viterbi search. The 3-D Viterbi method extracts a direction-time sequence of parameter vectors by steering a beam to each direction in every frame, then finds the most likely path in a 3-dimensional trellis space composed of talker directions, input frames and HMM states. This means that speech recognition and talker localization are performed simultaneously within a statistical framework.

To perform initial evaluation of the 3-D Viterbi method, recognition experiments for simulated data were carried out. The results show that speech features

*Doctor Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT9661031, March 24, 1999.

extracted in low SNR conditions are insufficient to estimate a talker direction accurately. In this thesis, the following solutions are considered: (1) the use of a pitch harmonics weight function and (2) the use of adaptive beamforming. The effect of these solutions were evaluated through recognition experiments for real data. The results confirm that the 3-D Viterbi method achieves high recognition performance for a moving talker case as well as for a fixed-position talker case.

Keywords:

Speech recognition, real environments, hands-free situations, microphone arrays, talker localization

Acknowledgments

I would like to express my foremost gratitude to Professor Kiyohiro Shikano (Nara Institute of Science and Technology) for his constant guidance through the master course and the doctor course. His insight played a significant role in my study.

I would like to express my greatest appreciation to Professor Kunihiro Chihara (Nara Institute of Science and Technology) and Professor Yoh'ichi Tohkura (Nara Institute of Science and Technology and NTT Basic Research Laboratories) for their precious suggestions.

I would like to express my sincere gratefulness to Associate Professor Satoshi Nakamura (Nara Institute of Science and Technology) for his continuous guidance. The core of this work originated with his pioneering ideas in hands-free speech recognition, which led me to a new research direction. This work could not have been accomplished without his daily support, which often extended over the middle of the night. He also demonstrated to me an attitude toward not only research but also life. I am always proud of doing research with him.

I would like to thank Research Associate Dr. Shiro Ise (Nara Institute of Science and Technology) who is currently Associate Professor at Kyoto University, Research Associate Dr. Jinlin Lu (Nara Institute of Science and Technology), Dr. Makoto Shozakai (Asahi Chemical Industry Co. Ltd.) and Dr. Harald Singer (ATR Interpreting Telecommunications Research Laboratories) for their beneficial comments.

I would like to thank Dr. Satoru Hayamizu, Dr. Futoshi Asano (Electrotechnical Laboratory), Dr. Yutaka Kaneda and Mr. Masashi Tanaka (NTT Human Interface Laboratories) for their support when I was a student intern at their laboratories.

I would like to thank my colleagues: Mr. Mike Schuster (Nara Institute of Science and Technology and ATR Interpreting Telecommunications Research Laboratories) for his proofreading of this thesis; Mr. Tetsuya Takiguchi, Mr. Tadashi Yonezaki, Ms. Eli Yamamoto and all other members of our laboratory for their helpful discussions. I would also like to thank all members of the Acoustical Society of Japan (ASJ), who gave me useful comments concerning this work.

Finally, I would like to acknowledge my father, mother and sister. This thesis wouldn't have been possible without their help.

Dedication

To my father, mother and sister.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Hands-free Speech Recognition in Real Environments

During the last decade the performance of automatic speech recognition has been improved drastically by applying statistical approaches, namely hidden Markov models (HMMs) and neural networks [1, 2, 3, 4]. Continuous speech recognition is currently put to practical use for applications such as dictation systems, command-and-control systems, database inquiry and tourist information systems. Most available systems have to meet a number of requirements to achieve high recognition accuracy. Some of them are speaker-independence, huge vocabulary (more than a hundred thousand words), robustness for spontaneous speech and environment-independence. In particular, an urgent necessity is environment-independence because the performance of most recognizers seriously degrades if they encounter environmental mismatches between training and testing conditions. By solving this problem, the present speech recognition technology could be used in a wide variety of environments.

A simple way to achieve high recognition performance in a certain environment is to train a recognizer with speech collected in this environment. However, it is unrealistic to collect a large amount of speech in a large number of different environments. Therefore, most recognizers are trained with noise-free (clean) speech and force users to use a head-mounted or hand-held microphone to reduce environmental differences between training and testing conditions. However, it is troublesome for most users to be encumbered with a microphone. To make the best use of the speech modality, users should not be encumbered with microphone equipment and should be able to speak at a certain distance from a microphone while moving. Therefore, an issue to be addressed is hands-free speech recognition which is also referred to as distant-talking speech recognition. Hands-free speech recognition technology can contribute to the increase of flexibility of the present applications and can be used to create new applications.

Speech acquired by a microphone far from the talker is distorted by environmental interference. The environmental interference is divided into two factors: additive noise and convolutional distortion.

- The additive noise includes background noise such as mechanical noise of computer-fans and air-conditioners, other talkers and sound of television and music. The additive noise also includes electrical noise of a transducer which is composed of a microphone, an amplifier and a low/high-pass filter. Figure 1.1(a) shows a relationship between the desired speech and the additive noise.

- The convolutional distortion contains room acoustics which typically lead to reverberations and characteristics of the transducer. The convolutional distortion is represented as an impulse response from the talker to the transducer output. Figure 1.1(b) shows a relationship between the desired speech

and the convolutional distortion. In Figure 1.1(b), H_1 is the acoustic transfer function from the talker to the transducer input and H_2 is the acoustic transfer function from the transducer input to the output. H_1 is time-varying when the talker moves, while H_2 is approximately time-invarying.

To improve recognition performance for the distorted speech, research on compensation for these factors is necessary.

The final objective of this thesis is to realize hands-free speech recognition in small conference rooms, where users can speak at a distance up to $2-3$ meters from a microphone while moving. Since performance degradation in adverse environments is serious even for isolated-word recognizers with a limited vocabu-

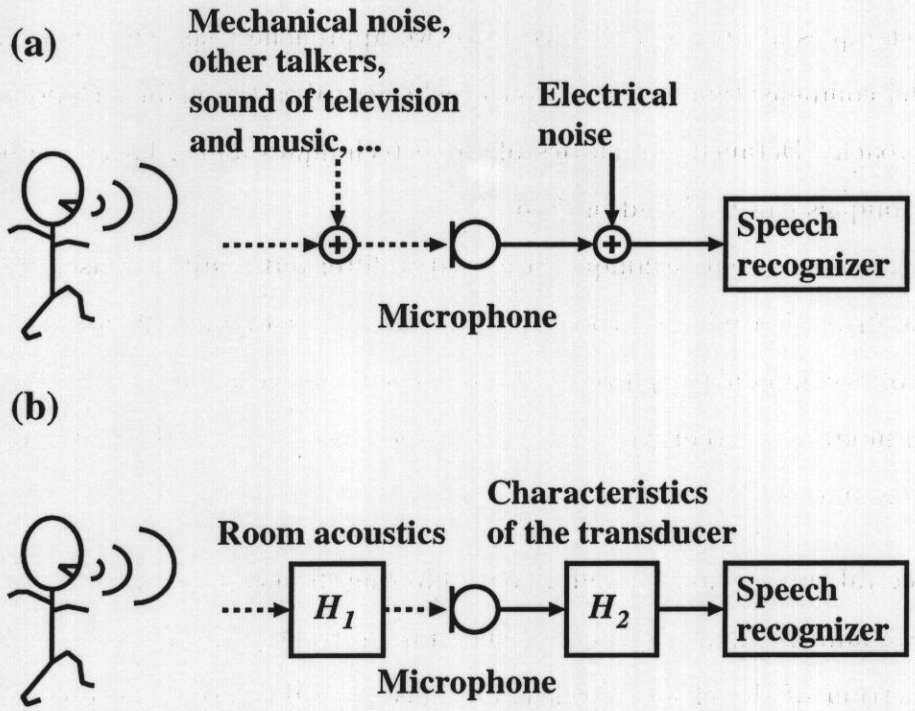

Figure 1.1. Relationships between desired speech and environmental interference. (a) Additive noise. (b) Convolutional distortion.

lary, the algorithms proposed in this thesis are tested using a simple isolated-word recognition system. To compensate for the environmental interference, this thesis introduces a microphone array-based approach and proposes a new speech recognition algorithm based on a 3-dimensional Viterbi search. The most attractive feature of this algorithm is to integrate the speech recognition process and the talker localization into a unified framework to deal with a moving talker.

1.2 Literature Review

The performance of most recognizers seriously degrades when training and testing environments don't match. Many techniques have been developed to cope with environmental interference such as additive noise and convolutional distortion. These techniques are broadly divided into two approaches: speech enhancement and model compensation. This section briefly describes the main techniques for each approach. Detailed summaries of these techniques and a large number of other techniques can be found in [5, 6, 7].

Speech enhancement techniques are used to filter out additive noise and convolutional distortion before recognition process starts. Figure 1.2 shows a block diagram of a common recognizer. The retrieved speech is fed into the pattern matching module together with the clean speech models through the feature extraction module.

- Spectral subtraction [8] and Wiener filtering [9] have been proposed to remove additive noise. Using the spectral subtraction technique, the power spectrum of the clean speech is obtained by subtracting the power spectrum of the additive noise from the power spectrum of the distorted speech. Using the Wiener filtering technique, the clean speech is estimated by filtering the distorted speech with a non-causal Wiener filter estimated from

the distorted speech and *a priori* knowledge of additive noise. Because the characteristics of the additive noise are needed to apply these techniques, they are estimated from non-speech frames if only a single microphone is used, and from a reference microphone located at the source of the noise if two microphones can be used. The performance of these techniques strongly depends on the accuracy of detecting the non-speech frames and the type of additive noise (stationary or non-stationary).

- Cepstrum mean normalization (CMN) [10] has been developed to remove the effect of convolutional distortion in a single microphone framework. Convolutional distortion in the time domain results in additive noise in the cepstral (log-spectral) domain. The cepstrum of the clean speech is obtained by subtracting the cepstrum of the convolutional distortion from the cepstrum of the distorted speech. The cepstrum of the convolutional distortion is estimated by averaging over the entire cepstrum sequence of the distorted speech. When using CMN, the implicit assumption is that the acoustic transfer function from the talker to the transducer output is time-invarying. However, the acoustic transfer function from the talker to the

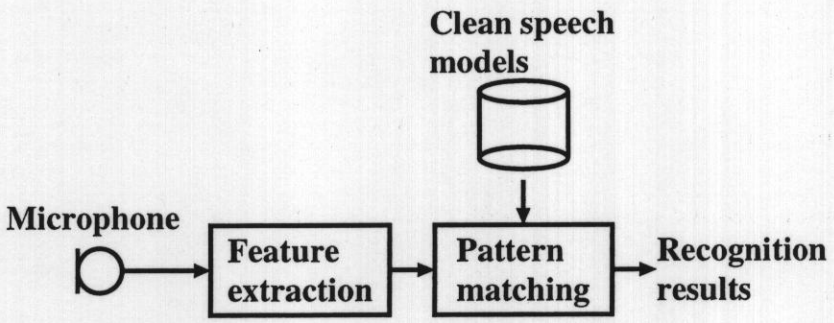

Figure 1.2. A block diagram of a common recognizer.

transducer input is time-varying when the talker moves (see Figure 1.1(b)).

Model compensation techniques adapt the clean speech models to the test conditions. The distorted speech is fed into the pattern matching module together with the adapted speech models through the feature extraction module.

- Model decomposition [11] and model composition [12, 13, 14] have been proposed to compensate for the effect of additive noise and convolutional distortion in a single microphone framework. Model decomposition decomposes the distorted speech by using the clean speech models and environmental interference models during the recognition process. Model composition is closely related to model decomposition. Distorted speech models are composed from the clean speech models and the environmental interference models before the recognition process. Additive noise models are trained with the non-speech frames. Also, convolutional distortion models are trained with the impulse response from the talker to the transducer output. To allow users to speak at arbitrary positions without measuring the impulse response every time, recently a few techniques have been developed [15]. These techniques estimate the convolutional distortion models from the distorted speech, the clean speech models and the additive noise models. However, these techniques still have a problem in the case when the talker is moving. It is necessary to establish a framework to deal with a moving talker.

As mentioned above, most techniques assume that:

- the additive noise is stationary,
- the convolutional distortion is time-invarying and
- *a priori* knowledge of environmental interference is available.

However, these assumptions do not hold for many practical environments. To realize hands-free speech recognition in a wide variety of environments, an ideal technique must have the following features:

- suppression of the effect of additive noise, whether stationary or non-stationary,
- removal of the effect of convolutional distortion, whether time-invarying or time-varying and
- adaptation to unknown environments and changing environments.

Recently, the use of microphone arrays has been investigated as a new paradigm, which offers a distinct performance advantage over single-channel algorithms [16]. A microphone array is composed of multiple microphones located at different positions. The output signal of each microphone includes not only acoustic information but also spatial information about sound sources. Therefore, signals originating from distinct directions can be separated by spatiotemporal filtering. In other words, a desired signal can be acquired selectively by forming a directive pattern sensitive to the direction. This means that the beamforming can effectively suppress other sound sources, whether stationary or non-stationary, including reverberations. Furthermore, a moving talker can be dealt with because the directive pattern can be electronically steered to arbitrary directions. Thus, this thesis focuses on the use of microphone arrays.

1.3 Thesis Overview

This thesis is organized as follows. Chapter 1 gives an introduction and reviews currently used techniques. Chapter 2 describes the basics of microphone array

processing: beamforming and acoustic source localization. This chapter also discusses their application to hands-free speech recognition. Chapter 3 describes a new database for microphone array research which includes speech data of a moving talker. This database contains generated data simulating ideal conditions which approximately correspond to those of an anechoic chamber. The database contains also real data recorded in a real room. Chapter 4 describes a new approach which integrates the speech recognition process and the talker localization into a unified framework to deal with a moving talker. In this chapter, a new speech recognition algorithm based on a 3-dimensional Viterbi search is proposed. The performance of the 3-D Viterbi method is evaluated through recognition experiments for simulated and real data. Chapter 5 describes the application of adaptive beamforming to the 3-D Viterbi method. The effect of adaptive beamforming is evaluated through recognition experiments for real data. Chapter 6 summarizes the contributions of this thesis and gives suggestions for future work.

Chapter 2

Microphone Arrays

A microphone array is composed of multiple microphones located at different positions. Signals received by a microphone array contain acoustical and spatial information about sound sources. By using spatial relationships among sound sources and microphones, signals originating from distinct directions can be separated. That is, a desired signal can be acquired by forming a directive pattern sensitive to the direction, whether other sound sources are stationary or non-stationary, including reverberations. Since the directive pattern can be steered to arbitrary directions, a moving talker can be dealt with.

This chapter describes the basics of microphone array processing and its application to hands-free speech recognition. Section 2.1 explains the principle of delay-and-sum beamforming [17, 18]. Section 2.2 also explains two acoustic source localization algorithms: spatiotemporal analysis [19, 20] and cross-power spectrum phase (CSP) analysis [21, 22, 23, 24]. Finally, Section 2.3 discusses problems of current speech recognizers using a microphone array.

2.1 Beamforming

Beamforming is the name given to microphone array processing which acquires a high quality signal by forming a directive pattern sensitive to the propagating direction. In the last several decades, a number of beamforming algorithms have been developed. Delay-and-sum beamforming [17, 18] is the oldest and simplest algorithm and is still a powerful algorithm today because of its simplicity, computational efficiency and reliability. However, the delay-and-sum beamforming algorithm has the problem that many microphones are necessary to narrow the width of the main beam of the directive pattern. To achieve a so-called high resolution beam without increasing the number of microphones, a number of different beamforming algorithms have been proposed such as adaptive beamforming [25, 26, 27], subspace-based beamforming [28], matched filter array processing and multiple beamforming [29, 30]. These algorithms achieve superior signal acquisition performance by adapting to specific conditions. In this thesis, the delay-and-sum beamforming algorithm is used for initial evaluation of the speech recognition algorithm proposed in Chapter 4. The remainder of this section explains the principle of delay-and-sum beamforming.

Figure 2.1 shows a block diagram of the delay-and-sum beamforming algorithm. A plane wave from direction θ is received by a microphone array which is composed of M microphones set up linearly and separated at distance d. The signal $x_m(t)$ received by the mth microphone is represented as follows:

$$x_m(t) = x_1\left(t - (m-1)\tau\right), \tag{2.1}$$

where $\tau = d\cos\theta/c$ and c is the sound propagation speed. The output signal $y(t)$ of delay-and-sum beamforming is given by

$$y(t) = \sum_{m=1}^{M} x_m\left(t + (m-1)\tau\right). \tag{2.2}$$

The amplitude of the signal from direction θ becomes M times as large as the original signal by the co-phasing operation. On the other hand, signals from other directions will not be co-phased. This means that a directive pattern sensitive to direction θ is formed. Figure 2.2(a)$\sim$(c) show the effect of noise reduction by delay-and-sum beamforming. An original speech waveform of a Japanese utterance /ikioi/ and its noisy version are shown in Figure 2.2(a) and (b), where the speech signal and a white Gaussian noise signal come from 30 degrees and 120 degrees, respectively. Figure 2.2(c) shows the enhanced version using the delay-and-sum beamforming algorithm with 16 microphones. The results show that the white Gaussian noise is successfully suppressed.

Equation (2.1) can be written in the frequency domain as follows:

$$X_m(f) = X_1(f)\, e^{-j2\pi f(m-1)d\cos\theta/c}, \tag{2.3}$$

where $X_m(f)$ is the spectrum of the signal received by the mth microphone and f is the frequency. The spectrum $Y(f)$ of the output signal of delay-and-sum

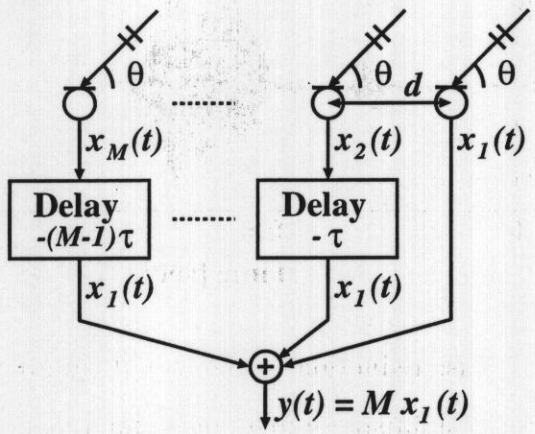

Figure 2.1. A block diagram of the delay-and-sum beamforming algorithm.

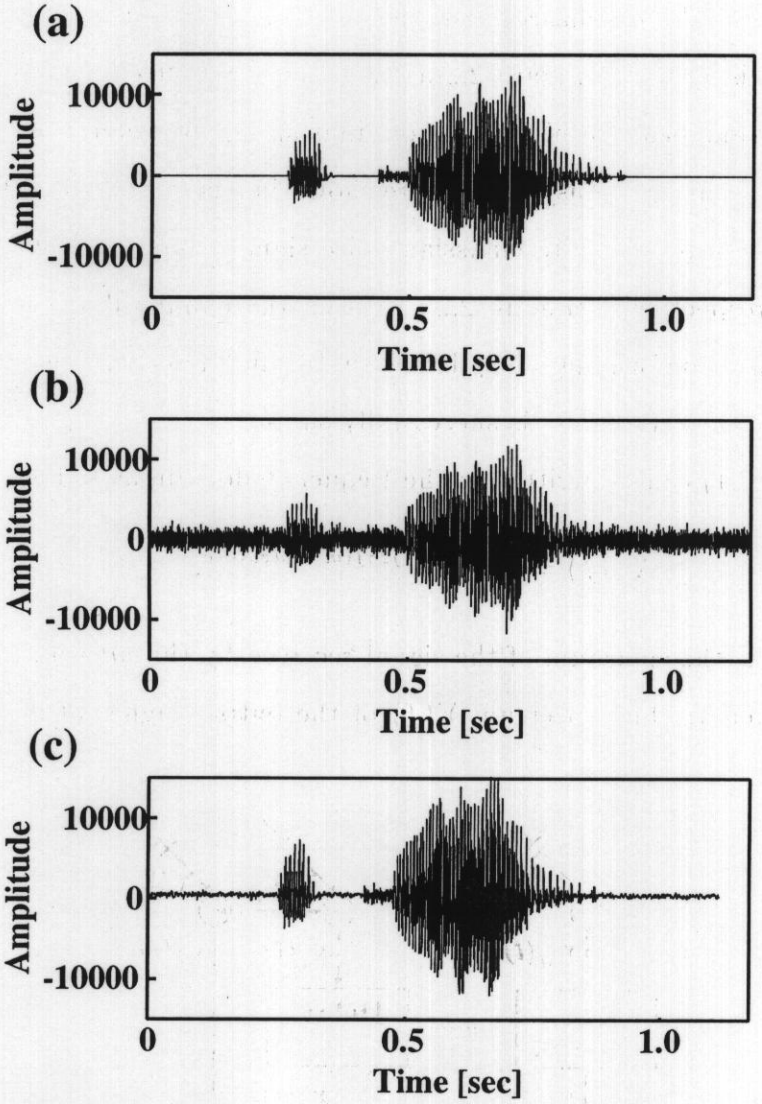

Figure 2.2. The effect of noise reduction by delay-and-sum beamforming. (a) An original speech waveform of a Japanese utterance /ikioi/. (b)(c) A noisy version and its enhanced version of the same utterance.

beamforming is represented as follows:

$$Y(f) = \sum_{m=1}^{M} X_m(f)\, e^{j2\pi f(m-1)d\cos\theta/c}. \tag{2.4}$$

The output gain $G(\phi, f)$ of delay-and-sum beamforming in direction ϕ is derived from Equation (2.3) and (2.4) as follows:

$$G(\phi, f) = \frac{|Y(f)|}{|X_m(f)|} = \left| \frac{\sin(M\pi f d(\cos\phi - \cos\theta)/c)}{\sin(\pi f d(\cos\phi - \cos\theta)/c)} \right|. \tag{2.5}$$

Figure 2.3 shows directive patterns calculated from Equation (2.5), where the wavefront arrival direction is 90 degrees ($\theta = 90$). In Figure 2.3(a), the number of microphones is 8 ($M = 8$), the distance between adjacent microphones is 17 cm ($d = 17$) and the frequency is 500 Hz ($f = 500$). The width W of the main lobe can be given by

$$W = 2\sin^{-1}\frac{c}{M\ d\ f}. \tag{2.6}$$

Equation (2.6) shows that the width of the main lobe is narrowed as the frequency is raised. In Figure 2.3(b), the frequency is twice as large as that in Figure 2.3(a). Equation (2.6) also shows that the width of the main lobe can be narrowed by increasing the number of microphones or spreading the distance between adjacent microphones. In Figure 2.3(c) and (d), the number of microphones and the distance between adjacent microphones are twice as large as those in Figure 2.3(a), respectively. Spatial aliasing occurs when the distance between adjacent microphones is too wide. In Figure 2.3(e) and (f), the secondary lobes become as large as the main lobe. To avoid spatial aliasing, the following condition must be satisfied.

$$d \leq \frac{c}{2f_{\max}}, \tag{2.7}$$

where $f_{\max}$ is the Nyquist frequency.

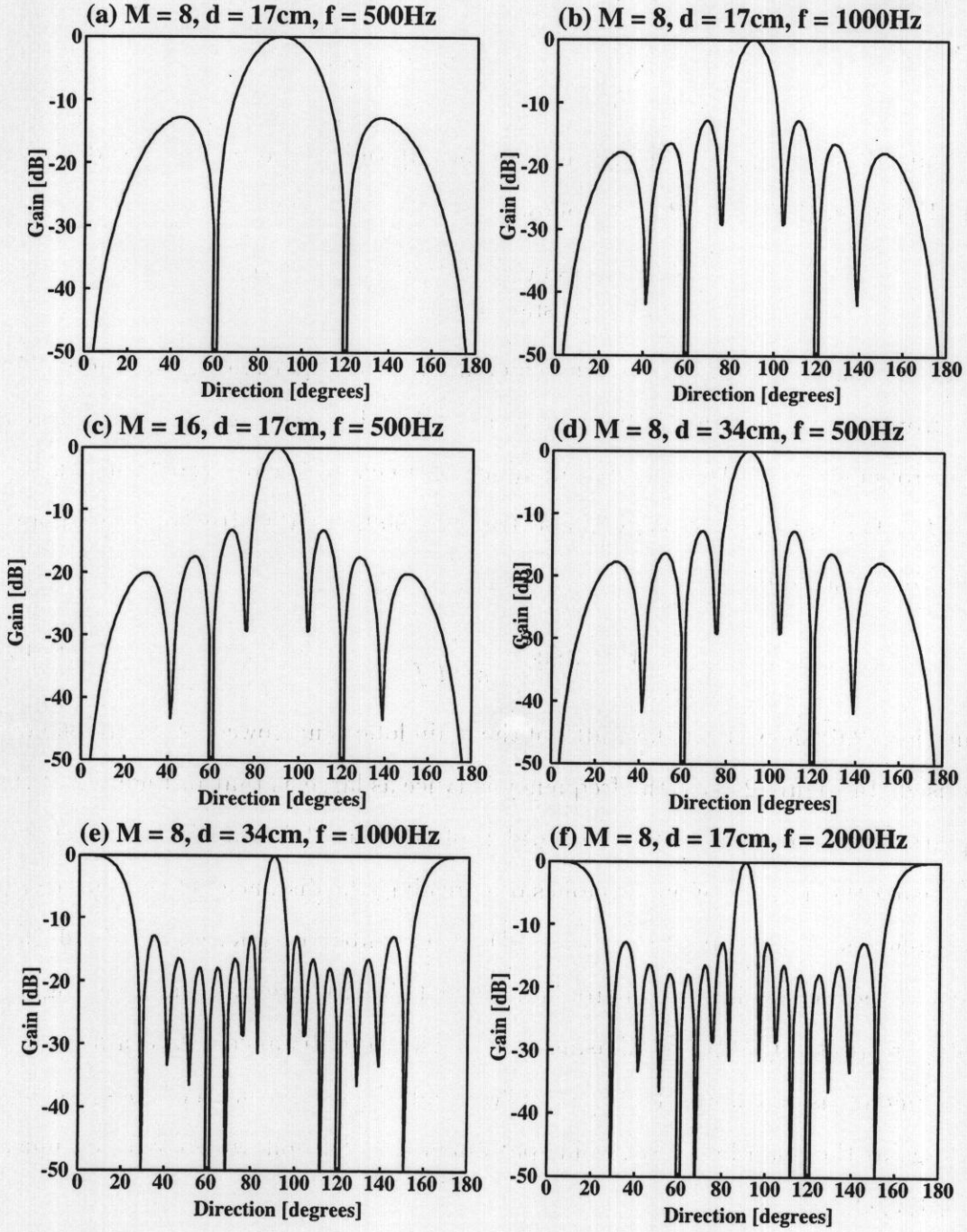

Figure 2.3. Directive patterns calculated from Equation (2.5), where the wavefront arrival direction is 90 degrees ($\theta = 90$).

2.2 Acoustic Source Localization

When a wavefront arrival direction of a desired signal is known, beamforming algorithms can effectively reduce noise. However, in hands-free situations where users can speak at arbitrary positions while moving, the wavefront arrival direction is generally unknown. If a steering direction is different from the wavefront arrival direction, the desired signal is distorted due to the frequency dependency of directive patterns (see Section 2.1) and other sound sources. Therefore, it is very important for high quality signal acquisition by beamforming algorithms to localize a target sound source accurately. During the last several decades acoustic source localization algorithms such as spatiotemporal analysis [19, 20], cross-power spectrum phase (CSP) analysis [21, 22, 23, 24] and eigenvector-based analysis [31] have been proposed. In particular, the spatiotemporal analysis and the CSP analysis are widely used for speech recognizers using a microphone array because of their computational efficiency and stability. This section describes the spatiotemporal analysis and the CSP analysis in detail.

In the spatiotemporal analysis, a spatial power spectrum is calculated by steering a beam to each direction. To simplify the problem, the delay-and-sum beamforming algorithm is used for this purpose. The spatial power spectrum $P(\phi, f)$ is derived from Equation (2.3) and (2.4) as follows:

$$P(\phi, f) = \left| \sum_{m=1}^{M} X_m(f)\, e^{j2\pi f(m-1)d\cos\phi/c} \right|^2 \tag{2.8}$$

and

$$X_m(f) = X_1(f)\, e^{-j2\pi f(m-1)d\cos\theta/c}, \tag{2.9}$$

where ϕ is the steering direction and θ is the wavefront arrival direction. If the steering direction ϕ is equal to the wavefront arrival direction θ, the output

power of delay-and-sum beamforming is maximized. Therefore, the wavefront arrival direction θ can be estimated by finding the maximum value of the spatial power spectrum. When there are multiple sound sources, the wavefront arrival directions can also be estimated by finding the peak values of the spatial power spectrum. Figure 2.4 shows an example of the spatial power spectrum, where two white Gaussian noise signals come from 30 degrees and 120 degrees. In Figure 2.4, the spatial power spectrum is averaged over all frequency bands. The spatial power spectrum has peak values at 30 degrees and at 120 degrees, showing that the wavefront arrival directions are estimated correctly.

In the CSP analysis, the delay of arrival (DOA) between signals received by two microphones, as shown in Figure 2.5, is estimated to obtain the wavefront

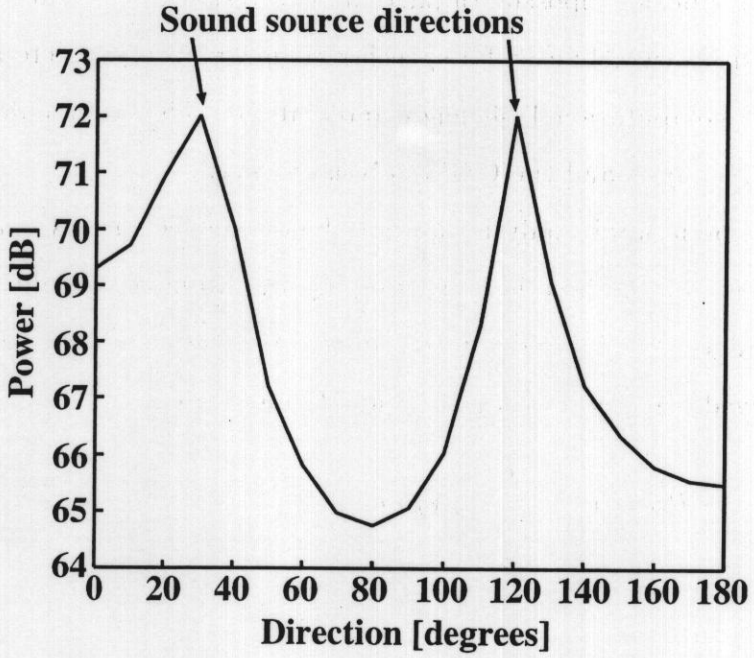

Figure 2.4. An example of the spatial power spectrum, where two white Gaussian noise signals come from 30 degrees and 120 degrees.

arrival direction θ. The CSP function is defined by

$$CSP(k) = \mathrm{DFT}^{-1}\left[\frac{\mathrm{DFT}\left[x_1(n)\right]\mathrm{DFT}\left[x_2(n)\right]^*}{\left|\mathrm{DFT}\left[x_1(n)\right]\right|\left|\mathrm{DFT}\left[x_2(n)\right]\right|}\right], \tag{2.10}$$

where n and k is the time index, DFT $[\cdot]$ (or $\mathrm{DFT}^{-1}[\cdot]$) is the discrete Fourier transform (or the inverse discrete Fourier transform) and * is the complex conjugate. The CSP function gives stable estimation performance by whitening the magnitude spectra of $x_1(n)$ and $x_2(n)$ compared to the normal cross-correlation function. The DOA can be estimated by finding the maximum value of the CSP function. When there are multiple sound sources, the DOAs can also be estimated by finding the peak values of the CSP function. Figure 2.6 shows an example of the CSP function, where two white Gaussian noise signals come from 30 degrees and 120 degrees. The CSP function has peak values at -0.43 msec and at 0.25 msec which correspond to 30 degrees and 120 degrees.

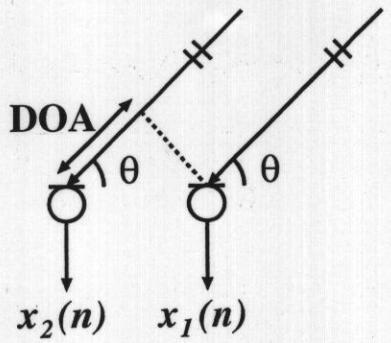

Figure 2.5. The delay of arrival (DOA) between signals received by two microphones.

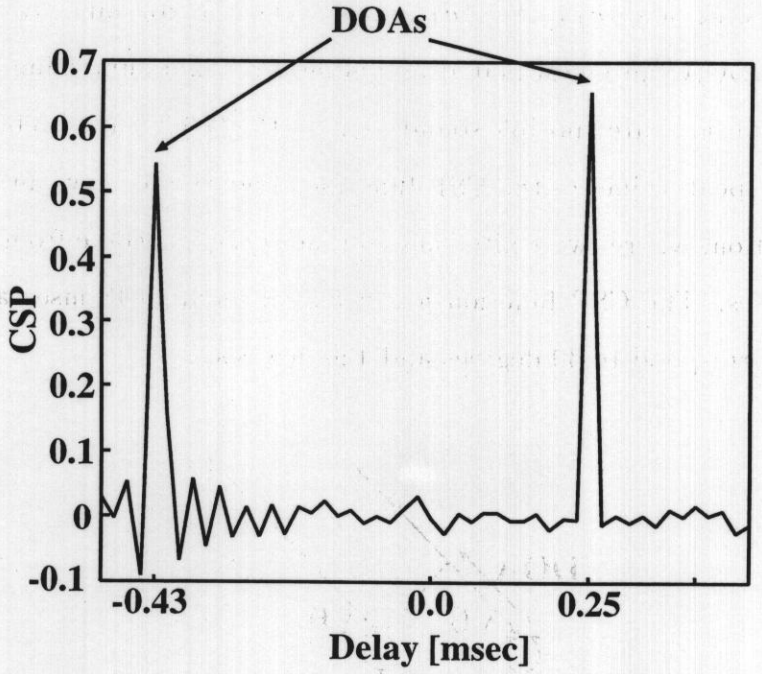

Figure 2.6. An example of the CSP function, where two white Gaussian noise signals come from 30 degrees and 120 degrees.

2.3 Problems of Current Speech Recognizers Using a Microphone Array

Applying microphone array processing to hands-free speech recognition, issues to be addressed are:

- the microphone array configuration;
- the beamforming algorithm;
- the talker localization and tracking algorithm.

The recognition performance is affected by the microphone array configuration (the number of microphones and the arrangement of microphones) and the beamforming algorithm. One way to decide these factors is a trial-and-error approach [32]. However, this approach is usually impractical because recognition experiments have to be run for a lot of the possible combinations, which takes a lot of time. To decide these factors automatically, it is necessary to develop a measure which reflects the recognition performance. One recently proposed measure to design microphone arrays is the normalized distortion to signal ratio (NDSR) [33, 34, 35, 36]. The NDSR gives good prediction of the recognition performance and is therefore a useful measure to evaluate beamforming algorithms. However, to maximize the effect of noise reduction, a reliable talker localization and tracking algorithm is necessary even if an optimal combination of the microphone array configuration and the beamforming algorithm is available. In particular, the talker localization and tracking algorithm becomes more important as the directive pattern is sharpened. Therefore, this thesis focuses on the last issue of the talker localization and tracking algorithm.

In the last several years, a number of speech recognition systems using a microphone array have been developed [37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48,

49]. Most systems are composed of a talker localization module, a beamforming module and a speech recognition module as shown in Figure 2.7. These systems estimate a talker direction in every frame, then steer a beam to the direction. High quality speech acquisition by beamforming algorithms is achieved only in situations where the talker direction is known. However, the talker direction is generally unknown in hands-free situations because users can talk at arbitrary positions while moving. If the steering direction differs from the talker direction, the output speech is distorted due to frequency dependency of directive patterns and other sound sources. The localization of multiple sound sources is generally possible by using the acoustic source localization algorithms described in Section 2.2. However, it is very difficult to distinguish a talker from the other sound sources. Several talker localization algorithms have been proposed [50, 51, 52, 45, 46, 47, 48, 49]. These algorithms estimate a talker direction by only using information about speech such as short-term power, long-term power and pitch harmonics. Therefore, the performance of these algorithms strongly depends on the characteristics of the other sound sources. Figure 2.8 shows an example of talker localization results and signal to noise ratio (SNR) when a talker is located at an angle of 40 degrees and a white Gaussian noise source at an angle of 90 degrees. The direction with the maximum short-term power is selected in every

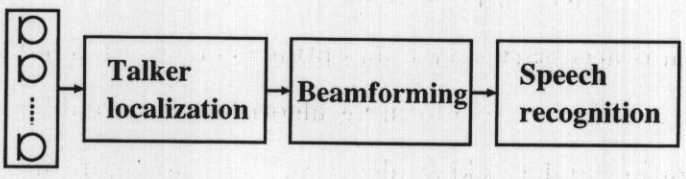

Figure 2.7. A block diagram of conventional speech recognition systems using a microphone array.

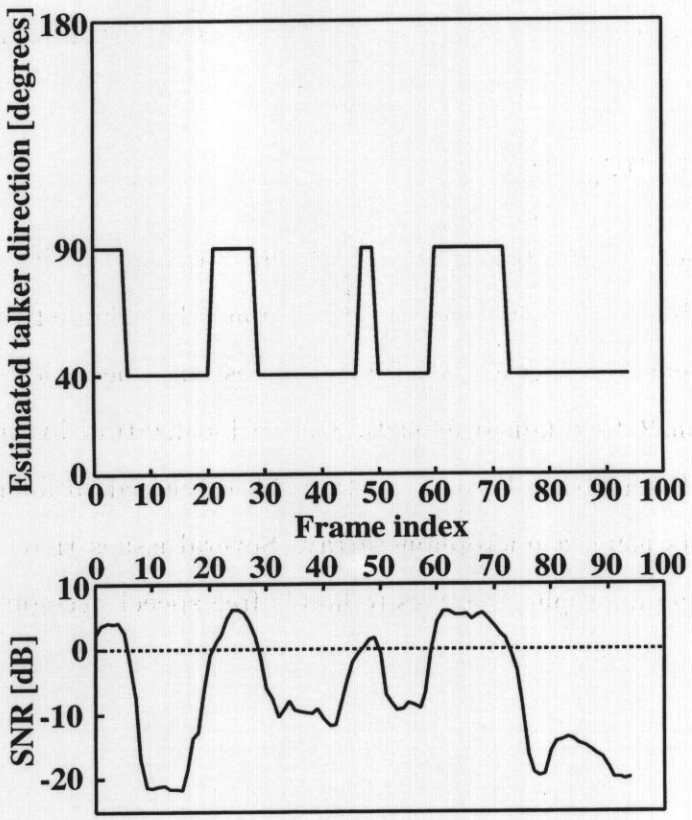

Figure 2.8. An example of talker localization results and signal to noise ratio (SNR) when a talker is located at an angle of 40 degrees and a white Gaussian noise source at an angle of 90 degrees.

frame and is taken as the talker direction. It can be seen that the noise source is mistaken as the talker for the frames with a low SNR. Since in conventional systems the microphone array processing is a pre-processor to speech recognition, it is very difficult to recover the beamforming error during the recognition process. To solve this problem, the integration of the speech recognition process and the talker localization into a unified framework is described in Chapter 4.

2.4 Summary

This chapter described the basics of microphone array processing and their application to hands-free speech recognition. Section 2.1 explained the principle of delay-and-sum beamforming, and its frequency response behavior was described in detail. Section 2.2 explained acoustic source localization by spatiotemporal analysis and CSP analysis. Finally, Section 2.3 discussed problems of current speech recognizers using a microphone array. Several issues that should be addressed in applying microphone arrays to hands-free speech recognition were also described.

Chapter 3

A Database for Microphone Array Research

It is very important for the development and evaluation of microphone array processing to collect a large amount of signals recorded through microphone arrays in various environments. Recently, several databases have been collected to evaluate the performance of beamforming, acoustic source localization and speech recognition algorithms [53, 54]. However, the use of these databases is very difficult because the microphone array configuration (e.g. the number of microphones and the arrangement of microphones) is specific to the database.

This chapter describes a new database of microphone array data in detail. This database contains simulated data and real data. The simulated data is used for development and initial evaluation of the speech recognition algorithm proposed in Chapter 4. The real data is used for evaluation of this algorithm in a real room. The most attractive feature of this database is that speech data of a moving talker is included. The remainder of this chapter is organized as follows. Section 3.1 describes the data collection procedure for the real data. Section 3.2 also describes the data generation procedure under ideal conditions which

approximately correspond to the acoustic conditions in an anechoic chamber. This database is partially open to the public through the Real World Computing Partnership (RWCP) [55].

3.1 Real Data

The real data recorded through a microphone array is collected in a real experiment room. The room is surrounded by sound absorption walls on the four sides and the floor. A reverberant time (T_{60}) of the room is approximately 0.18 sec. There is background noise such as computer-fans and air-conditioners in the room.

A list of equipment for data collection is shown in Table 3.1. The microphone array used in this thesis is composed of 14 omni-directive microphones as shown in Figure 3.1. The 14 microphones are linearly located at distances of 2.83 cm which

Table 3.1. A list of equipment for data collection.

Microphone array	ONSOKU
16-channel AD/DA conversion	SDS DASBOX
16-channel microphone amplifier with low pass filters	ONSOKU OAF-411
Loudspeaker	JBL Control 5 Plus
Loudspeaker amplifier	YAMAHA P4050

is the maximum distance to avoid spatial aliasing (see Equation (2.6)), where the sampling frequency is 12 kHz. The 16-channel AD/DA conversion system is used for the data collection, where 14 channels are for the microphone array and the other two channels are reserved for future needs (e.g. a head-mounted microphone and a reference microphone).

Microphone array data is collected for the following set-ups:

(a) The positions of the talker and the white Gaussian noise source are fixed.

(b) The talker moves, while the position of the white Gaussian noise source is fixed.

Figure 3.2(a) and (b) illustrate the two set-ups. Two loudspeakers facing the microphone array are used instead of the talker and the white Gaussian noise source. The distance between the microphone array and each loudspeaker is

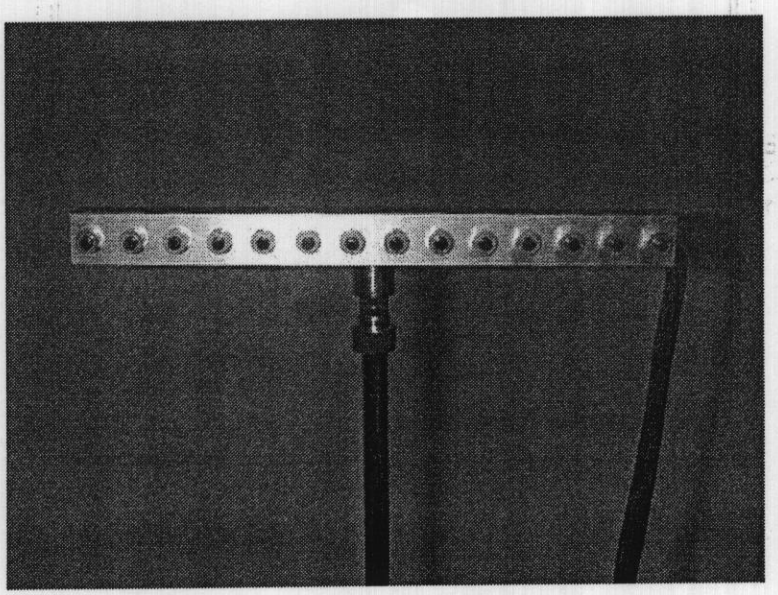

Figure 3.1. A microphone array used for data collection.

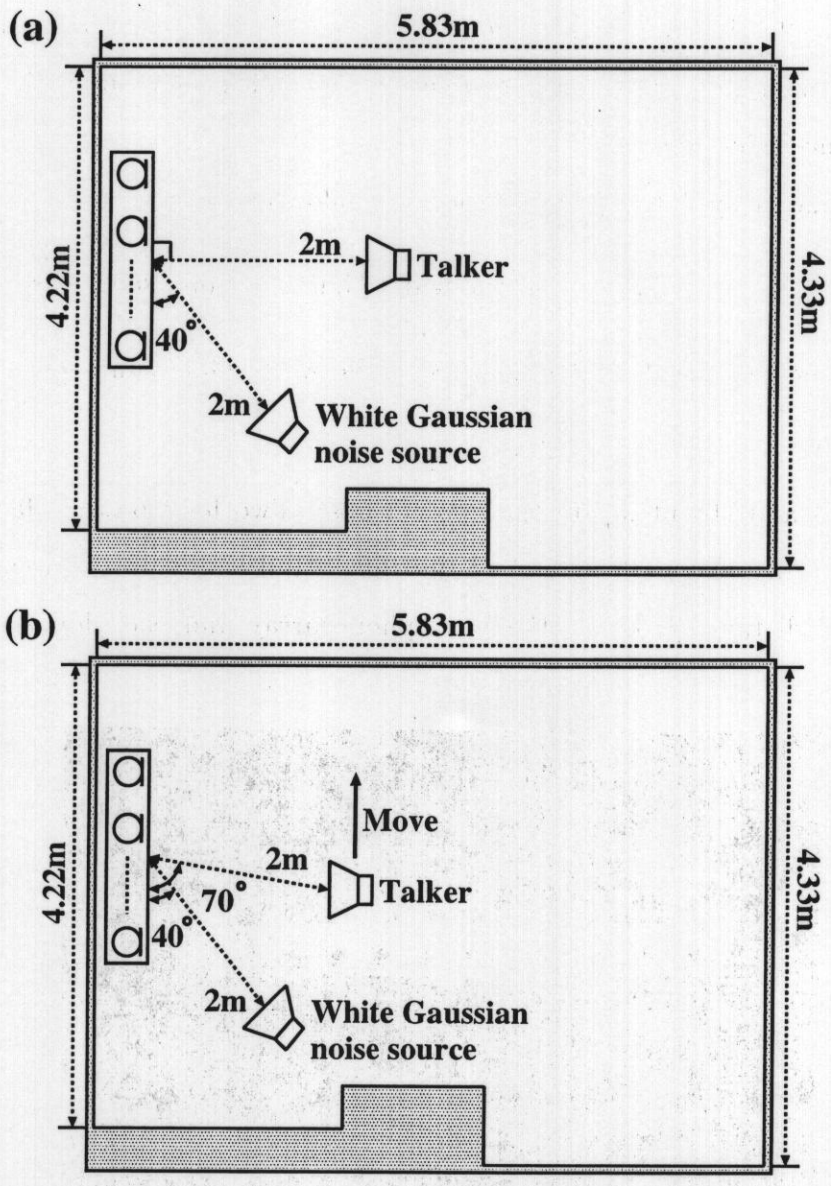

Figure 3.2. Two set-ups for the collection of the real data through a microphone array.

about 2 m. 216 phonetically-balanced words of a male speaker (MHT) from the ATR Japanese speech database Set-A [56] are used as speech data. In Figure 3.2(a), the talker is located at an angle of 90 degrees (in front of the microphone array) and the white Gaussian noise source at an angle of 40 degrees. In Figure 3.2(b), the talker moves from 70 degrees to 140 degrees while uttering each word, walking at a speed of about 80 cm per second. The white Gaussian noise source is located at an angle of 40 degrees.

3.2 Simulated Data

Microphone array data is generated to simulate ideal conditions which are approximately equal to those of an anechoic chamber. This means that there is no background noise and reverberations. This data is very useful for development and initial evaluation of microphone array-based algorithms.

A signal coming from a certain direction is generated using Equation (2.1). If the delay τ in Equation (2.1) is represented as nT (n is an integer and T is the sampling period), the delay operation is equal to a shift of n samples. If not, the delay operation can be approximated by an FIR filter as follows:

$$h(l) = \frac{1}{\pi(l - (\tau + \tau_0)/T)} \sin(\pi(l - (\tau + \tau_0)/T)), \tag{3.1}$$

where l is the sample index and τ_0 is the delay for causality.

The two set-ups described in Section 3.1 are simulated as shown in Figure 3.2(a) and (b). The microphone array configuration, the sampling frequency and the speech data are the same as those used in Section 3.1. In Figure 3.3(a), the talker is located at an angle of 90 degrees and the white Gaussian noise source at an angle of 40 degrees. In Figure 3.3(b), the talker moves from 0 degrees to 180 degrees while uttering each word. The white Gaussian noise source is located

at an angle of 40 degrees. The speech data of the moving talker is generated as follows:

$$x_m(n) = \sum_{l=0}^{L-1} h_{\theta(n-l),m}(l)x(n-l), \tag{3.2}$$

where $h_{\theta(n-l),m}(l)$ is the impulse response from the direction $\theta(n-l)$ to the mth microphone input, $\theta(n-l)$ is the talker direction at the sample index $n-l$, L is the filter length and $x(n)$ is the source signal.

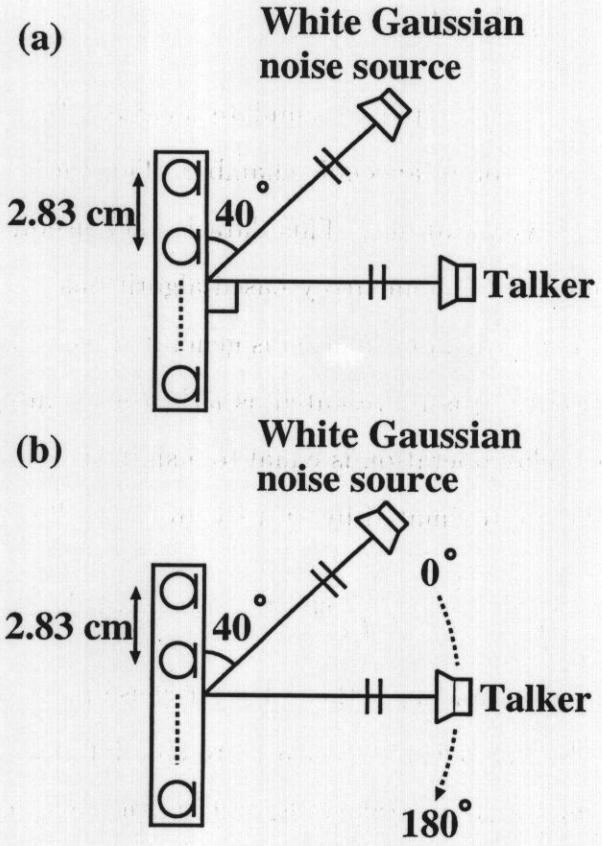

Figure 3.3. Two set-ups for the generation of the simulated data.

3.3 Summary

This chapter described a new database of microphone array data in detail. Section 3.1 described the collection of the real data. Section 3.2 also described data generation under ideal conditions which approximately correspond to those of an anechoic chamber.

Chapter 4

Hands-free Speech Recognition Based on 3-D Viterbi Search

The talker localization and tracking algorithm is critical for high quality speech acquisition by beamforming algorithms because the talker direction is unknown in hands-free situations. Generally, acoustic source localization is possible by using the algorithms described in Section 2.2 even if there are multiple sound sources. However, it is very difficult to distinguish a talker from other sound sources. Conventional algorithms estimate the talker direction by only using information about speech such as short-term power, long-term power and pitch harmonics [50, 51, 52, 45, 46, 47, 48, 49]. The problem with these algorithms is that the performance strongly depends on the characteristics of the other sound sources. Since most recognizers using a microphone array [37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49] use the microphone array processing as a pre-processor to speech recognition, it is very difficult to recover errors caused by the beamforming procedure during the recognition process.

To cope with this problem, this thesis introduces a new approach which integrates the speech recognition process and the talker localization into a unified

framework. Section 4.1 describes this approach in detail and proposes a new speech recognition algorithm based on a 3-dimensional Viterbi search [57, 58, 59, 60, 61, 62, 63]. Section 4.2 gives an overview over the recognition experiments run for this thesis. In Section 4.3, initial evaluation of the 3-D Viterbi method through recognition experiments for simulated data is discussed. Section 4.4 describes a pitch harmonics weight function and evaluates its effect. Section 4.5 evaluates the performance of the 3-D Viterbi method through recognition experiments for real data.

4.1 Integration of Speech Recognition Process and Talker Localization

A direction-time sequence of parameter vectors (e.g. mel-frequency cepstral coefficients) can be obtained by steering a beam to each direction for every frame as shown in Figure 4.1. Each box stands for a parameter vector and the solid line is the talker direction. The parameter vectors on the talker direction are extracted from the speech acquired by a beamforming algorithm. Therefore, the talker direction may be estimated by matching the direction-time sequence of the parameter vectors with the speech models. This means that the talker localization and speech recognition are performed simultaneously. This approach is formulated as follows:

$$(\hat{\mathbf{q}}, \hat{\mathbf{d}}) = \underset{(\mathbf{q},\mathbf{d})}{\operatorname{argmax}} P(\mathbf{x}, \mathbf{q}, \mathbf{d} \mid \mathbf{M}), \tag{4.1}$$

where $(\hat{\mathbf{q}}, \hat{\mathbf{d}})$ is the most likely combination of the phone sequence $\mathbf{q}$ of the speech and the direction sequence $\mathbf{d}$ of the talker, $\mathbf{M}$ is the speech model and $\mathbf{x}$ is the direction-time sequence of parameter vectors. Given the speech models $\mathbf{M}$, $(\hat{\mathbf{q}}, \hat{\mathbf{d}})$ can be obtained by the Viterbi search algorithm which finds the most likely path

in a 3-dimensional trellis space composed of talker directions, input frames and HMM states shown in Figure 4.2. The likelihood is give by

$$\begin{aligned}\alpha(q,d,n) &= \max_{q',d'}\{\alpha(q',d',n-1) \\ &\qquad + \log a_1(q',q) + \log a_2(d',d)\} \\ &\quad + \log b(q,\mathbf{x}(d,n)),\end{aligned} \tag{4.2}$$

where q is the HMM state index, d is the direction and n is the frame index. Furthermore, $a_1(q',q)$ is the transition probability from the HMM state q' to q, $a_2(d',d)$ is the transition probability from the direction d' to d and b is the output probability. The transition probability $a_2(d',d)$ represents how likely the talker moves. However, it is very difficult to train this probability automatically. In this thesis, according to the fact that the talker moves to neighboring directions at most during about 10 msec corresponding to the frame period, the range of

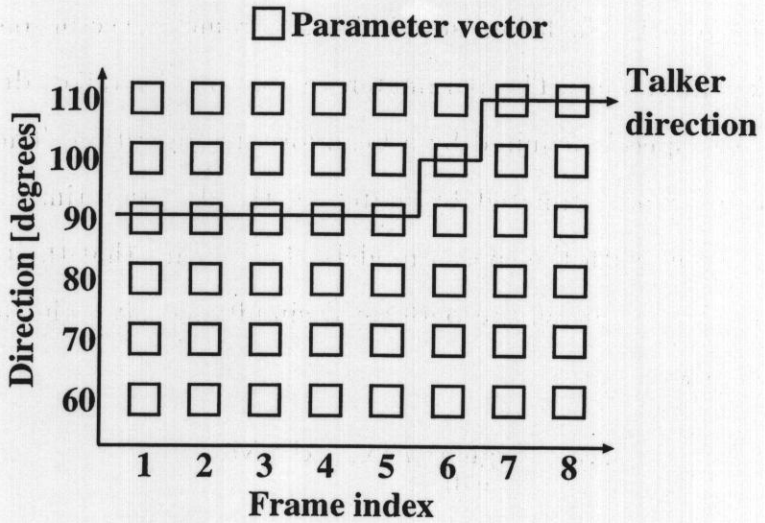

Figure 4.1. A direction-time sequence of parameter vectors.

the talker movements is restricted as follows:

$$a_2(d', d) = \begin{cases} \dfrac{1}{2\Delta d} & , \quad |d - d'| \leq \Delta d \\ 0 & , \quad |d - d'| > \Delta d \end{cases} , \tag{4.3}$$

where Δd is the range of the talker movements.

4.2 Experiment Conditions

To perform initial evaluation of the 3-D Viterbi method, isolated-word recognition experiments are conducted. In this chapter, a speaker-dependent case is dealt with to remove the effect of speaker variability. Recognition experiments in a speaker-independent case are shown in Chapter 5.

An overview of experiment conditions is shown in Table 4.1. The speech recognizer is based on a tied-mixture HMM with 256 distributions [64]. The HMMs

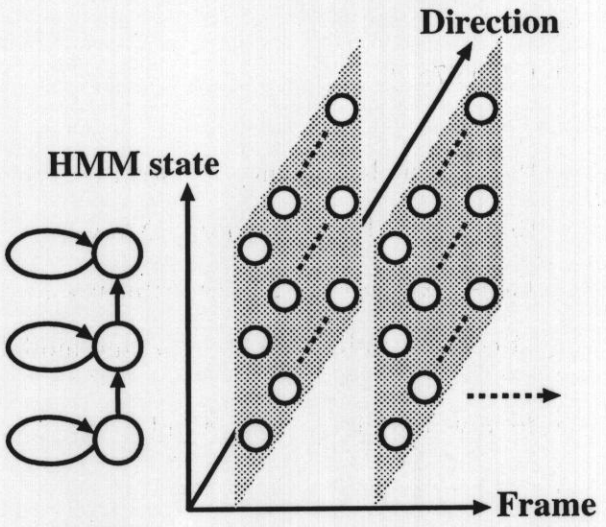

Figure 4.2. A 3-dimensional trellis space composed of talker directions, input frames and HMM states.

are 54 context-independent phoneme models, which are trained with 2620 words of a male speaker (MHT) of the ATR Japanese speech database Set-A. The test data are another 216 phonetically-balanced words. The speech signal is sampled at 12 kHz and windowed by a 32 msec Hamming window every 8 msec. Then the parameter vectors of 16th-order mel-frequency cepstral coefficients (MFCCs), 16th-order ΔMFCCs and 1st-order Δpower are calculated.

In this chapter, the delay-and-sum beamforming algorithm is used. Figure

Table 4.1. An overview of experiment conditions.

Sampling frequency	12 kHz
Frame length	32 msec
Frame period	8 msec
Pre-emphasis	$1 - 0.97z^{-1}$
Parameter vectors	16-order mel-frequency cepstral coefficients (MFCCs), 16-order ΔMFCCs, 1-order Δpower
HMM	Tied-mixture with 256 distributions, 54 context-independent phoneme models
Training data	2620 words of the male MHT
Test data	216 phonetically-balanced words of the male MHT

4.3 shows a directive pattern for the microphone array configuration described in Section 3.1, which is calculated from Equation (2.5) for each frequency. It can be seen that the width of the main lobe becomes narrower as the frequency raises and that spatial aliasing does not occur. The direction-time sequence of the parameter vectors is sampled every 10 degrees $(0, 10, \cdots, 180$ degrees) because preliminary results showed that recognition performance does not degrade when a difference between the steering direction and the talker direction is within ± 5 degrees.

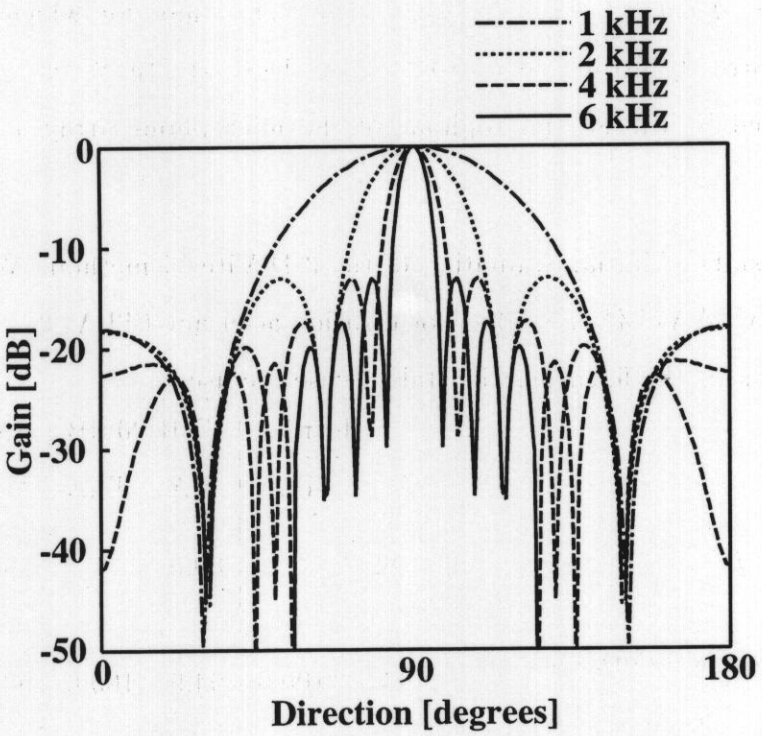

Figure 4.3. A directive pattern for the microphone array configuration described in Section 3.1, which is calculated from Equation (2.5) for each frequency.

4.3 Initial Evaluation of the 3-D Viterbi Method for Simulated Data

To perform initial evaluation of the 3-D Viterbi method, recognition experiments are conducted for the simulated data described in Section 3.2. Word recognition accuracy and talker localization accuracy for the fixed-position talker case are shown in Table 4.2. The talker localization accuracy is defined as follows:

$$\text{TLA} = \frac{number\ of\ correct\ frames}{number\ of\ total\ speech\ frames} \times 100[\%], \tag{4.4}$$

where *number of correct frames* is the number of frames for which the talker direction is correctly estimated $(0, 10, \cdots, 180$ degrees). In *Single microphone*, speech recorded by the 8th microphone of the microphone array is fed to the

Table 4.2. Results of initial evaluation of the 3-D Viterbi method. Word recognition accuracy (WA) [%] and talker localization accuracy (TLA) [%] are shown. Simulated data for the fixed-position talker case are used.

	Clean		SNR 20 dB		SNR 10 dB	
	WA	TLA	WA	TLA	WA	TLA
Single microphone	96.2	—	80.0	—	25.9	—
Delay-and-sum beamforming to the correct talker direction	96.2	100.0	94.9	100.0	90.7	100.0
Delay-and-sum beamforming with conventional talker localization	96.2	100.0	83.3	62.0	34.2	32.4
3-D Viterbi method	96.2	99.0	72.6	40.9	28.2	17.9

recognizer. In *Delay-and-sum beamforming to the correct talker direction*, the correct talker direction is known. The time sequence of the parameter vectors is obtained by steering the beam to the correct talker direction. In *Delay-and-sum beamforming with conventional talker localization* and *3-D Viterbi method*, the correct talker direction is unknown. In *Delay-and-sum beamforming with conventional talker localization*, a direction with the maximum short-term power is selected as the talker direction for every frame, then the time sequence of the parameter vectors is obtained by steering the beam to the estimated direction. In *3-D Viterbi method*, the Viterbi search in the 3-dimensional trellis space is performed, where Δd in Equation (4.3) is 10. The results are summarized as follows:

- The word recognition accuracy of *Delay-and-sum beamforming to the correct talker direction* is improved by 14.9 % for a SNR of 20 dB and by 64.8 % for a SNR of 10 dB compared to that of *Single microphone*. It is confirmed that the delay-and-sum beamforming almost suppresses the effect of the white Gaussian noise when the correct talker direction is known.

- The word recognition accuracy of *Delay-and-sum beamforming with conventional talker localization* decreases by 11.6 % for a SNR of 20 dB and by 56.5 % for a SNR of 10 dB compared to that of *Delay-and-sum beamforming to the correct talker direction*. This is caused by the low talker localization accuracy.

- The word recognition accuracy of *3-D Viterbi method* decreases by 10.7 % for a SNR of 20 dB and by 6.0 % for a SNR of 10 dB compared to that of *Delay-and-sum beamforming with conventional talker localization*.

Talker directions estimated by *3-D Viterbi method* for a Japanese word /ikioi/ for a clean condition and for a SNR 20 dB are shown in Figure 4.4. Since the

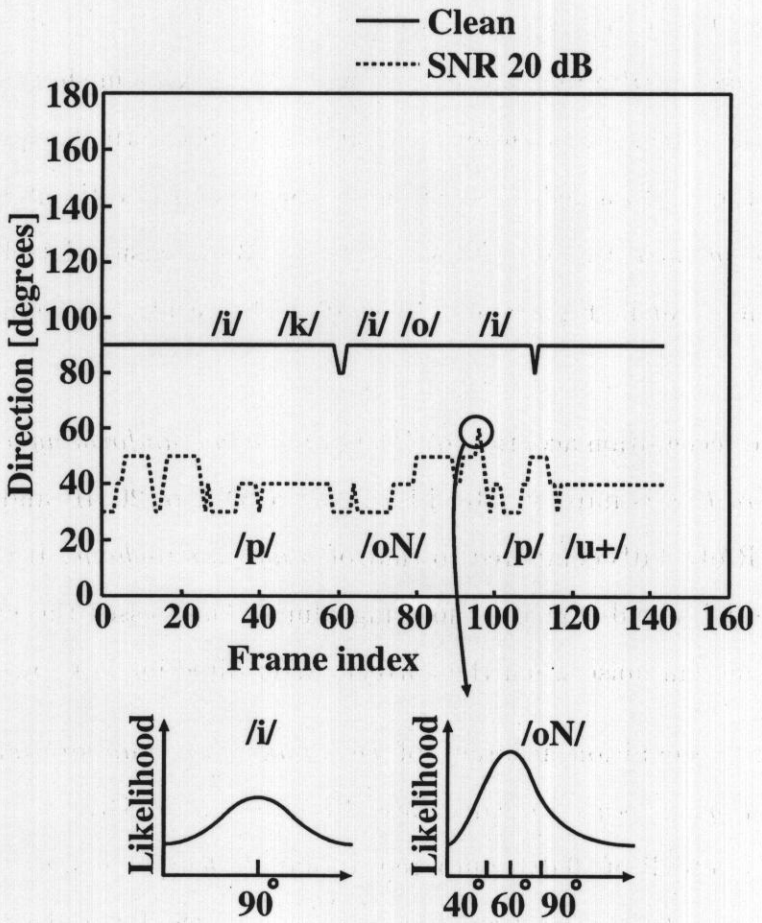

Figure 4.4. Talker directions estimated by *3-D Viterbi method* for a Japanese word /ikioi/ for a clean condition and for a SNR of 20 dB.

talker direction estimated by *3-D Viterbi method* for a clean condition is almost correct, the correct recognition result is obtained. However, the talker direction estimated by *3-D Viterbi method* for a SNR of 20 dB includes many errors and the word /ikioi/ is incorrectly recognized as the word /poNpu+/. For example, at frame index 97 the likelihood for the correct phoneme /i/ at 90 degrees is lower than that for the wrong phoneme /oN/ at 60 degrees. The results show that the speech features extracted in the low SNR conditions are insufficient to estimate the talker direction accurately. In this thesis, the following solutions are considered:

- The use of a weight function which raise the likelihood in the correct talker direction by using pitch harmonics information about speech.
- The use of beamforming algorithms with a high resolution.

The next section describes a pitch harmonics weight function and evaluates its effect. The second solution is addressed in Chapter 5.

4.4 A Pitch Harmonics Weight Function

4.4.1 Formulation

As mentioned above, speech features extracted in low SNR conditions are insufficient to estimate the talker direction accurately. To solve that problem, it will be effective to raise the likelihood in directions with speech-like characteristics. Since pitch harmonics information about speech can be used as a measure of speech-like

characteristics, a pitch harmonics weight function is introduced as follows:

$$w(d,n) = \log \frac{\sum_{n'=n-(\nu-1)}^{n} \{c(d,n')\}^{\mu}}{\sum_{d'=0}^{180} \sum_{n'=n-(\nu-1)}^{n} \{c(d',n')\}^{\mu}}, \tag{4.5}$$

where $c(d,n)$ is the maximum value of the cepstral coefficients in the high quefrency region, which is obtained by cepstral analysis for the beamformed signal in the direction d at the frame index n. Furthermore, μ is the parameter to control the weight effect and ν is the parameter to adjust the continuation of the pitch harmonics. $c(d,n)$ and $w(d,n)$ become larger as more of the pitch harmonics are included. $w(d,n)$ is added to the right part in Equation (4.2).

Figure 4.5 shows an example of $c(d,n)$ and $w(d,n)$ obtained for the phoneme /i/ at a frame for a SNR of 20 dB. In Figure 4.5, μ and ν are set to 40 and 10, the talker direction is located at an angle of 90 degrees and the white Gaussian noise source at an angle of 40 degrees. It can be seen that $c(d,n)$ and $w(d,n)$ in the talker direction is larger than those in the other directions, since the phoneme /i/ includes the pitch harmonics.

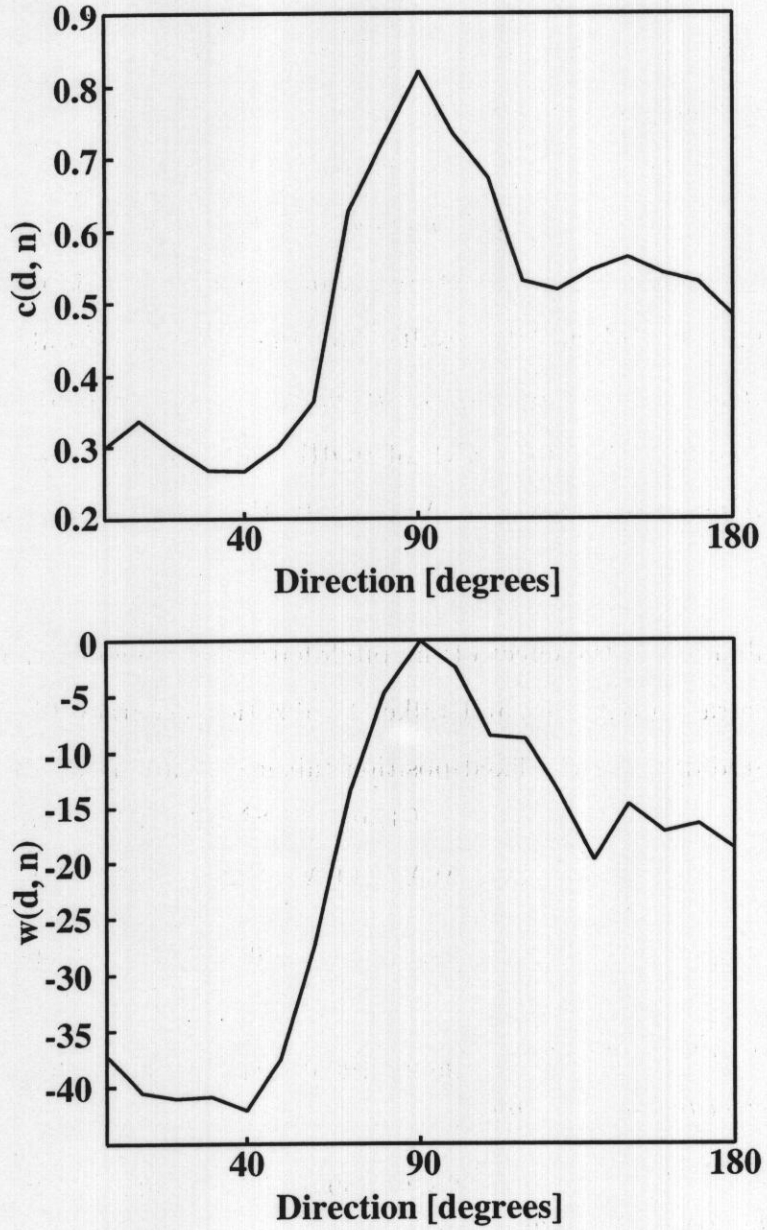

Figure 4.5. An example of $c(d,n)$ and $w(d,n)$ obtained for the phoneme /i/ at a frame for a SNR of 20 dB.

4.4.2 Results of Experiment

To evaluate the effect of the pitch harmonics weight function, recognition experiments are conducted for simulated data. The word recognition accuracy and the talker localization accuracy for the fixed-position talker case are shown in Table 4.3. *3-D Viterbi method without the weight function* is the same as *3-D Viterbi method* in Table 4.2. In *3-D Viterbi method with the weight function*, the pitch harmonics weight function is used, where Δd in Equation (4.3) is 10, μ and ν in Equation (4.5) are 40 and 20. The results show that the talker localization accuracy of *3-D Viterbi method with the weight function* is improved by 36.7 % for a SNR of 20 dB and by 53.3 % for a SNR of 10 dB compared to that of *3-D Viterbi method without the weight function.* As a result, the word recognition accuracy

Table 4.3. Evaluation of the effect of the pitch harmonics weight function. Word recognition accuracy (WA) [%] and talker localization accuracy (TLA) [%] are shown. Simulated data for the fixed-position talker case are used.

	Clean		SNR 20 dB		SNR 10 dB	
	WA	TLA	WA	TLA	WA	TLA
Single microphone	96.2	—	80.0	—	25.9	—
Delay-and-sum beamforming to the correct talker direction	96.2	100.0	94.9	100.0	90.7	100.0
3-D Viterbi method without the weight function	96.2	99.0	72.6	40.9	28.2	17.9
3-D Viterbi method with the weight function	96.2	99.1	94.9	77.6	88.4	71.2

of *3-D Viterbi method with the weight function* is almost equal to that of *Delay-and-sum beamforming to the correct talker direction.* Talker directions estimated by *3-D Viterbi method with the weight function* and *3-D Viterbi method without the weight function* for a Japanese word /ikioi/ for a SNR of 20 dB are shown in Figure 4.6. It can be seen that the talker can be localized more accurately by using the pitch harmonics weight function in the speech period.

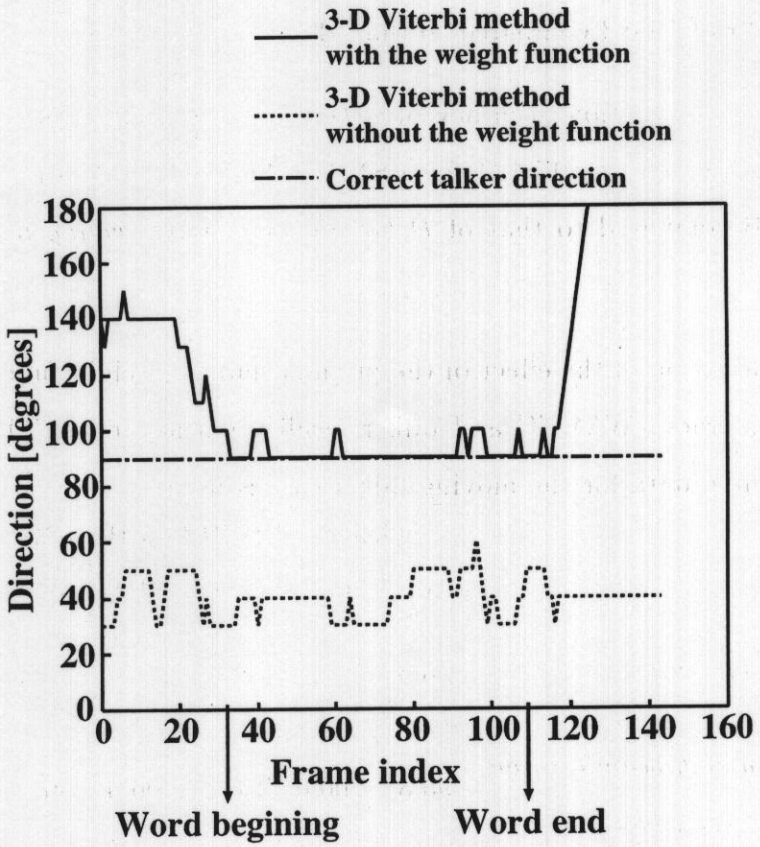

Figure 4.6. Talker directions estimated by *3-D Viterbi method with the weight function* and *3-D Viterbi method without the weight function* for a Japanese word /ikioi/ for a SNR of 20 dB (for the fixed-position talker case).

The word recognition accuracy and the talker localization accuracy for the moving talker case are also shown in Table 4.4. In *3-D Viterbi method without the weight function*, Δd in Equation (4.3) is 10. In *3-D Viterbi method with the weight function*, Δd in Equation (4.3) is 10, μ and ν in Equation (4.5) are 40 and 10. The results are summarized as follows:

- The word recognition accuracy of *Delay-and-sum beamforming to the correct talker direction* is improved by 12.0 % for a SNR of 20 dB and by 64.3 % for a SNR of 10 dB compared to that of *Single microphone.*
- The word recognition accuracy of *3-D Viterbi method without the weight function* decreases by 17.1 % for a SNR of 20 dB and by 59.7 % for a SNR of 10 dB compared to that of *Delay-and-sum beamforming to the correct*

Table 4.4. Evaluation of the effect of the pitch harmonics weight function. Word recognition accuracy (WA) [%] and talker localization accuracy (TLA) [%] are shown. Simulated data for the moving talker case is used.

	Clean		SNR 20 dB		SNR 10 dB	
	WA	TLA	WA	TLA	WA	TLA
Single microphone	95.8	—	79.6	—	22.2	—
Delay-and-sum beamforming to the correct talker direction	95.8	100.0	91.6	100.0	86.5	100.0
3-D Viterbi method without the weight function	96.2	86.7	74.5	44.7	26.8	21.0
3-D Viterbi method with the weight function	96.7	84.0	93.9	63.1	84.7	51.9

talker direction.

- The talker localization accuracy of *3-D Viterbi method with the weight function* is improved by 18.4 % for a SNR of 20 dB and by 30.9 % for a SNR of 10 dB compared to that of *3-D Viterbi method without the weight function.* As a result, the word recognition accuracy of *3-D Viterbi method with the weight function* is almost equal to that of *Delay-and-sum beamforming to the correct talker direction.*

Talker directions estimated by *3-D Viterbi method with the weight function* and *3-D Viterbi method without the weight function* for a Japanese word /ikioi/ for a SNR of 20 dB are shown in Figure 4.7. It can be seen that *3-D Viterbi method with the weight function* tracks the moving talker in the speech period.

Finally, the effect of μ and ν in the pitch harmonics weight function is investigated, where μ is the parameter to control the weight effect and ν is the parameter to adjust the continuation of the pitch harmonics. Figure 4.8 and 4.9 show the effect of μ and ν for the fixed-position talker case and the moving talker case for a SNR of 20 dB, respectively. In Figure 4.8 and 4.9, the vertical axis is the word recognition accuracy and the horizontal axis is the combination of μ and ν. When ν is set to 1, the word recognition accuracy becomes higher as μ is increased. However, when μ is set to ∞ which corresponds to the case that only one direction is selected as the talker direction according to the maximum value of the cepstral coefficients in the high quefrency region, the word recognition accuracy decreases. This means that the integration of the speech recognition process and the talker localization is more effective than the conventional approach. In Figure 4.9, when μ is set to 40, the word recognition accuracy becomes lower as ν is increased. This is caused by movements of the talker.

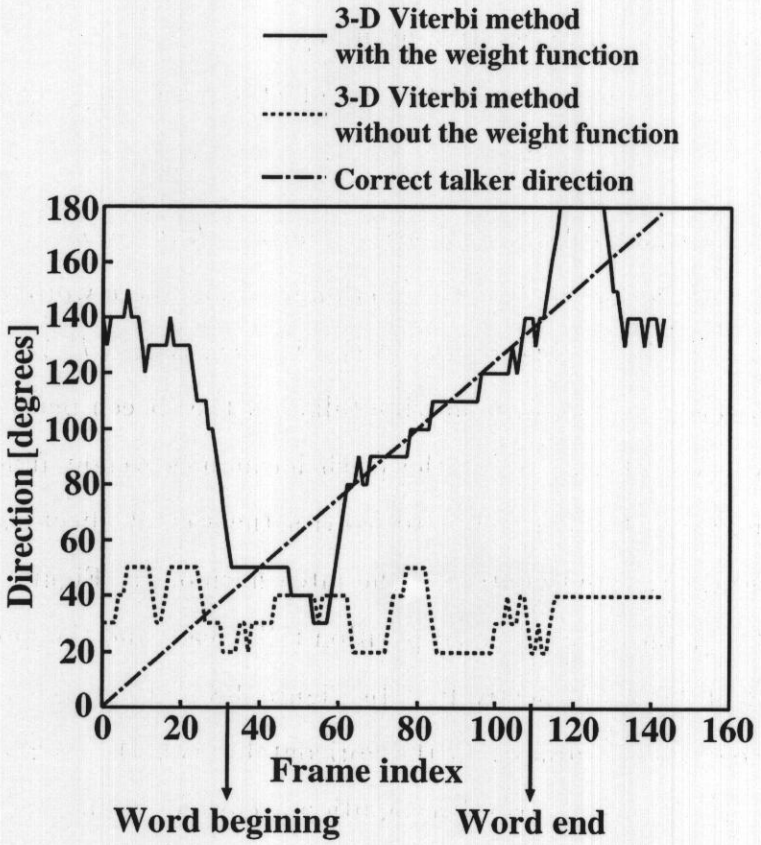

Figure 4.7. Talker directions estimated by *3-D Viterbi method with the weight function* and *3-D Viterbi method without the weight function* for a Japanese word /ikioi/ for a SNR of 20 dB (for the moving talker case).

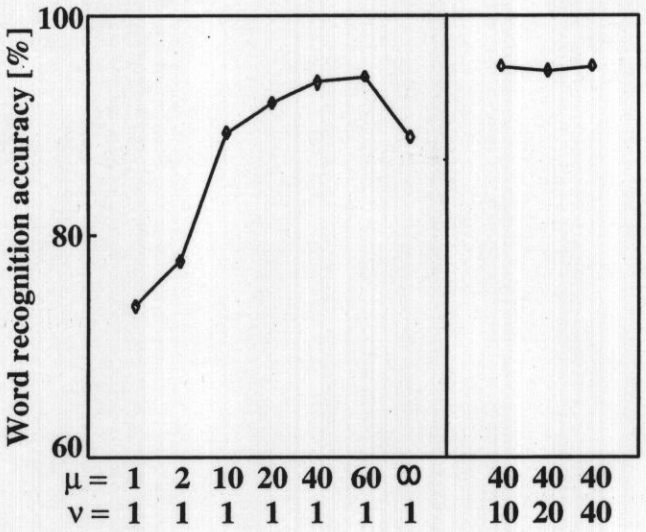

Figure 4.8. The effect of the pitch harmonics weight function for the fixed-position talker case for a SNR of 20 dB.

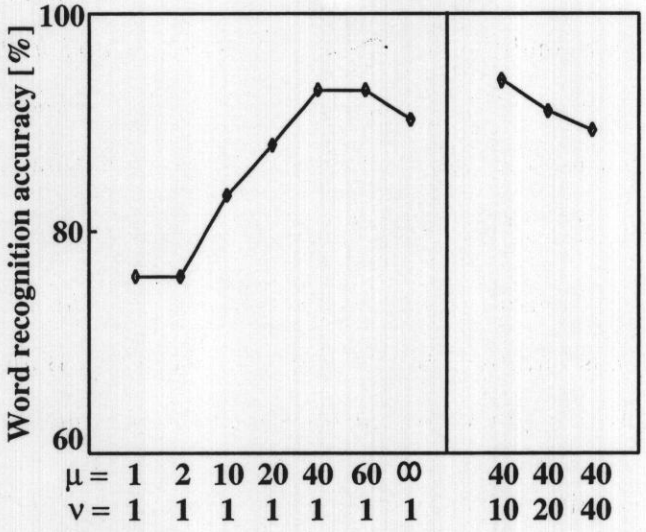

Figure 4.9. The effect of the pitch harmonics weight function for the moving talker case for a SNR of 20 dB.

4.5 Evaluation of the 3-D Viterbi Method for Real Data

The simulated data used for initial evaluation is generated for ideal conditions which approximately correspond those of an anechoic chamber. However, in practical situations, there are background noise and reverberations. In this section, to investigate the performance of the 3-D Viterbi method in a real room, recognition experiments are conducted for the real data described in Section 3.1.

Several techniques have been developed to compensate for the characteristics of a transducer and the distortion due to a difference between the steering direction and the talker direction [65, 66, 67]. In the following experiments, cepstrum mean normalization [10] is applied to the pre-processed speech.

The word recognition accuracy for the fixed-position talker case is shown in Table 4.5. In Table 4.5, a SNR of 21 dB corresponds to the case where there are no white Gaussian noise sources. However, there is background noise from computer-fans and air-conditioners. In *3-D Viterbi method without the weight function*, Δd in Equation (4.3) is 10. In *3-D Viterbi method with the weight function*, Δd in Equation (4.3) is 10, and μ and ν in Equation (4.5) are 80 and 10. The results are summarized as follows.

- The word recognition accuracy of *Delay-and-sum beamforming to the correct talker direction* is improved by 9.7 % for a SNR of 18 dB and by 38.0 % for a SNR of 10 dB compared to that of *Single microphone.*
- The word recognition accuracy of *3-D Viterbi method without the weight function* decreases by 33.8 % for a SNR of 18 dB and by 61.2 % for a SNR of 10 dB compared to that of *Delay-and-sum beamforming to the correct talker direction.*

- The word recognition accuracy of *3-D Viterbi method with the weight function* is improved by 26.4 % for a SNR of 18 dB and by 39.4 % for a SNR of 10 dB compared to that of *3-D Viterbi method without the weight function.*

The word recognition accuracy for the moving talker case is shown in Table 4.6. In *3-D Viterbi method without the weight function*, Δd in Equation (4.3) is 10. In *3-D Viterbi method with the weight function*, Δd in Equation (4.3) is 10, and μ and ν in Equation (4.5) are 80 and 10. Since the correct talker direction couldn't be measured, experiments of *Delay-and-sum beamforming to the correct talker direction* couldn't be carried out. The results show that the word recognition accuracy of *3-D Viterbi method with the weight function* is improved by 21.3 % for a SNR of 18 dB and by 29.2 % for a SNR of 10 dB compared to that of *3-D Viterbi method without the weight function.*

Table 4.5. Evaluation of the 3-D Viterbi method in a real room. Word recognition accuracy [%] is shown. Real Data for the fixed-position talker case are used.

	SNR 21 dB	SNR 18 dB	SNR 10 dB
Single microphone	89.8	76.8	37.0
Delay-and-sum beamforming to the correct talker direction	92.1	86.5	75.0
3-D Viterbi method without the weight function	89.3	52.7	13.8
3-D Viterbi method with the weight function	92.5	79.1	53.2

The results show that the pitch harmonics weight function is effective in a real room for the moving talker case as well as for the fixed-position talker case. Also, it can be seen that the performance improvement by *3-D Viterbi method with the weight function* compared to *Single microphone* is small. Since the performance of *Delay-and-sum beamforming to the correct talker direction* is insufficient, a high resolution beamforming algorithm is necessary.

Table 4.6. Evaluation of the 3-D Viterbi method in a real room. Word recognition accuracy [%] is shown. Real Data for the moving talker case are used.

	SNR 21 dB	SNR 18 dB	SNR 10 dB
Single microphone	92.5	77.7	38.4
3-D Viterbi method without the weight function	89.8	60.6	23.1
3-D Viterbi method with the weight function	89.3	81.9	52.3

4.6 Summary

This chapter focused on a talker localization and tracking algorithm. Section 4.1 introduced a new approach which integrates the speech recognition process and the talker localization into a unified framework and proposed a new speech recognition algorithm based on a 3-dimensional Viterbi search. Section 4.2 gave an overview of the recognition experiments. In Section 4.3 the initial evaluation of the 3-D Viterbi method through recognition experiments for simulated data was described. The results showed that speech features extracted in low SNR conditions are insufficient to estimate the talker direction accurately. To solve this problem, Section 4.4 introduced a pitch harmonics weight function and evaluated its effect. The results showed that the performance of the 3-D Viterbi method drastically improved for the moving talker case as well as for the fixed-position talker case. Section 4.5 investigated the performance of the 3-D Viterbi method in a real room. The results showed that the pitch harmonics weight function is effective in a real room. Also, it was shown that a high resolution beamforming algorithm is necessary to achieve high recognition performance for practical use.

Chapter 5

Hands-free Speech Recognition Using Adaptive Beamforming

In Chapter 4, results of experiments showed that the pitch harmonics weight function is very effective when the SNR improvement by a beamforming algorithm is insufficient. Also, it was confirmed that the recognition accuracy in a real room is still insufficient for practical use. This chapter focuses on adaptive beamforming as a high resolution beamforming algorithm [25, 26, 27]. Figure 5.1 shows conceptual directive patterns of delay-and-sum beamforming and adaptive beamforming. The delay-and-sum beamforming algorithm forms the so-called super-directive pattern which is independent of noise source directions. The width of the main lobe can be narrowed by increasing the number of microphones. However, it is undesirable to increase the number of microphones because it increases the cost of the equipment. On the other hand, the output gain of the adaptive beamforming algorithm is attenuated especially in noise source directions. Therefore, adaptive beamforming can achieve a high SNR improvement compared to delay-and-sum beamforming without increasing the number of microphones.

This chapter describes the application of adaptive beamforming to the 3-D

Viterbi method [68, 69, 70, 71]. Section 5.1 explains an adaptive beamforming algorithm. Section 5.2 evaluates the effect of the adaptive beamforming algorithm through recognition experiments for real data. Finally, Section 5.3 evaluates the performance of the 3-D Viterbi method through speaker-independent recognition experiments for a real moving talker.

5.1 Adaptive Beamforming

Figure 5.2 shows a block diagram of adaptive beamforming. In Figure 5.2, $S(\omega)$ is the spectrum of the desired signal and $Y(\omega)$ is the spectrum of the output signal. $G_m(\omega)$ is the acoustic transfer function from the desired sound source to the mth microphone output and $H_m(\omega)$ is the frequency response of the mth filter. The frequency response $F(\omega)$ of adaptive beamforming to the desired signal is

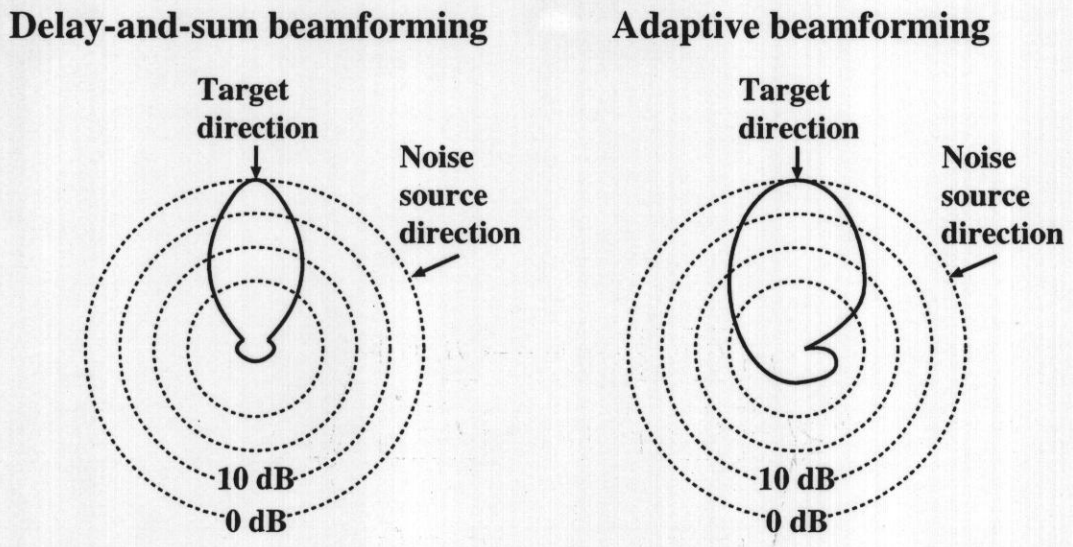

Figure 5.1. Conceptual directive patterns of delay-and-sum beamforming and adaptive beamforming.

represented as follows:

$$F(\omega) = \sum_{m=1}^{M} G_m(\omega) H_m(\omega), \tag{5.1}$$

where M is the number of microphones. The concept of adaptive beamforming is to minimize the output power while constraining $F(\omega)$ to a frequency response. In this thesis, an AMNOR constraint [27] as shown in Equation (5.2) is used.

$$D = \int |1 - F(\omega)|^2 d\omega. \tag{5.2}$$

The AMNOR constraint achieves maximum noise reduction, while allowing a small distortion D in the frequency response to the desired signal.

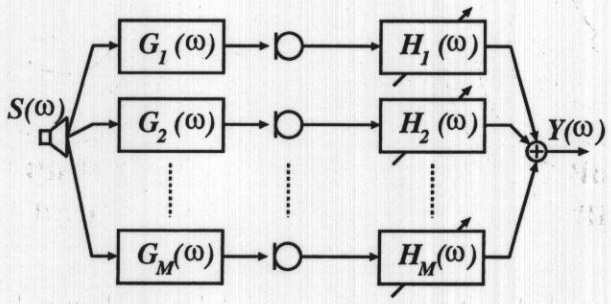

Figure 5.2. A block diagram of adaptive beamforming.

5.2 Evaluation of the Effect of Adaptive Beamforming

To evaluate the effect of adaptive beamforming on the performance of the 3-D Viterbi method, recognition experiments are conducted for real data.

The conditions of the experiments are the same as those in Section 4.2. In the following experiments, cepstrum mean normalization is applied to the pre-processed speech. The filter coefficients for adaptive beamforming are calculated using pre-recorded noise signals in the following two set-ups:

(1) There is only background noise.

(2) There is the white Gaussian noise source as well as background noise.

Directive patterns of adaptive beamforming and delay-and-sum beamforming are shown in Figure 5.3. In Figure 5.3, the directive pattern of adaptive beamforming is calculated for the set-up (2). The delay-and-sum beamforming algorithm forms the super-directive pattern. It can be seen that the output gain of adaptive beamforming in the white Gaussian noise source direction (40 degrees) is lower than that of delay-and-sum beamforming.

The word recognition accuracy for the fixed-position talker case is shown in Table 5.1. In Table 5.1, a SNR of 21 dB corresponds to the case that there are no white Gaussian noise sources. However, there is background noise from computer-fans and air-conditioners. In *Single microphone*, speech recorded by the 8th microphone of the microphone array is fed to the recognizer. In *Delay-and-sum beamforming to the correct talker direction* and *Adaptive beamforming to the correct talker direction*, the correct talker direction is known. The time sequence of the parameter vectors is obtained by steering each beam to the correct talker direction. In *3-D Viterbi method with delay-and-sum beamforming* and

3-D Viterbi method with adaptive beamforming, the correct talker direction is unknown. The direction-time sequence of the parameter vectors is obtained by each beamforming, then the Viterbi search in the 3-dimensional trellis space is performed, where Δd in Equation (4.3) is 10, and μ and ν in Equation (4.5) are 80 and 10. The results can be summarized as follows:

- The word recognition accuracy of *Adaptive beamforming to the correct talker direction* is improved by 2.3 % for a SNR of 21 dB, by 4.7 % for a SNR of 18 dB, by 14.3 % for a SNR of 10 dB compared to that of *Delay-and-sum beamforming to the correct talker direction.*

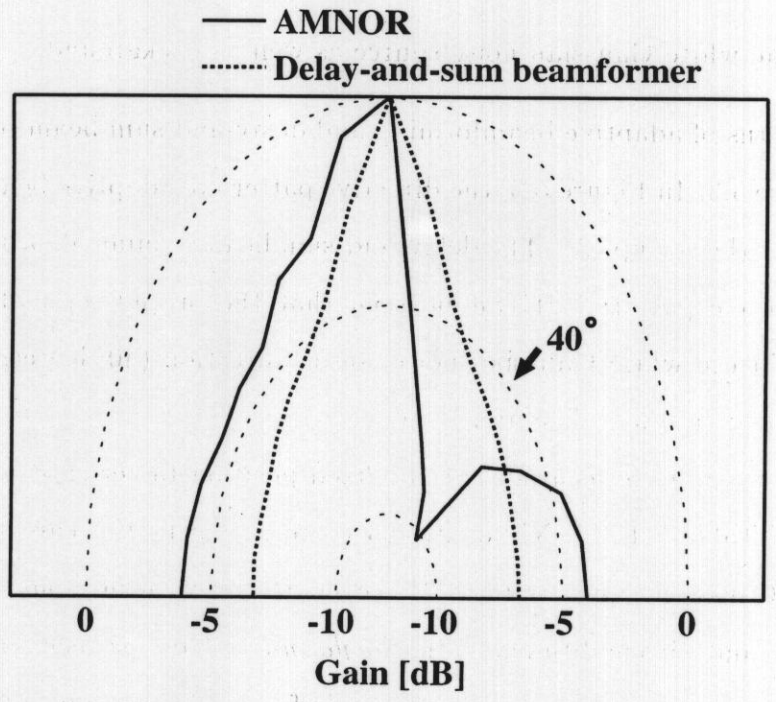

Figure 5.3. Directive patterns of adaptive beamforming and delay-and-sum beamforming.

- The word recognition accuracy of *3-D Viterbi method with adaptive beamforming* is improved by 1.4 % for a SNR of 21 dB, by 10.7 % for a SNR of 18 dB, by 30.1 % for a SNR of 10 dB compared to that of *3-D Viterbi method with delay-and-sum beamforming.*

The word recognition accuracy for the moving talker case is also shown in Table 5.2. In *3-D Viterbi method with delay-and-sum beamforming* and *3-D Viterbi method with adaptive beamforming*, Δd in Equation (4.3) is 10, and μ and ν in Equation (4.5) are 80 and 10. Since the correct talker direction couldn't be measured, experiments of *Delay-and-sum beamforming to the correct talker direction*

Table 5.1. Evaluation of the effect of the adaptive beamforming. Word recognition accuracy [%] is shown. Real Data for the fixed-position talker case are used.

	SNR 21 dB	SNR 18 dB	SNR 10 dB
Single microphone	89.8	76.8	37.0
Delay-and-sum beamforming to the correct talker direction	92.1	86.5	75.0
Adaptive beamforming to the correct talker direction	94.4	91.2	89.3
3-D Viterbi method with delay-and-sum beamforming	92.5	79.1	53.2
3-D Viterbi method with adaptive beamforming	93.9	89.8	83.3

and *Adaptive beamforming to the correct talker direction* couldn't be carried out. The results are summarized as follows.

- The word recognition accuracy of the 3-D Viterbi method for the moving talker case is almost equal to that for the fixed-position talker case.
- The word recognition accuracy of *3-D Viterbi method with adaptive beamforming* is improved by 3.2 % for a SNR of 21 dB, by 6.9 % for a SNR of 18 dB, by 28.7 % for a SNR of 10 dB compared to that of *3-D Viterbi method with delay-and-sum beamforming.*

The results confirmed that adaptive beamforming drastically improves the performance of the 3-D Viterbi method compared to delay-and-sum beamforming for the moving-talker case as well as for the fixed-position talker case.

Table 5.2. Evaluation of the effect of the adaptive beamforming. Word recognition accuracy [%] is shown. Real Data for the moving talker case is used.

	SNR 21 dB	SNR 18 dB	SNR 10 dB
Single microphone	92.5	77.7	38.4
3-D Viterbi method with delay-and-sum beamforming	89.3	81.9	52.3
3-D Viterbi method with adaptive beamforming	92.5	88.8	81.0

5.3 Evaluation of the 3-D Viterbi method for Speaker-independent HMMs and a Real Moving Talker

In the previous sections, a loudspeaker is used instead of a moving talker and speaker-dependent recognition experiments are performed. In this section, to evaluate the performance of the 3-D Viterbi method in a more practical situation, speaker-independent recognition experiments are performed for a real moving talker (not a loudspeaker).

The HMMs are 54 context-independent phoneme models, which are trained with sentences of 64 speakers of the ASJ Japanese speech database. Cepstrum mean normalization is applied to the pre-processed speech. Speech data of a real moving talker is collected using a set-up shown in Figure 5.4, which is almost same as that used to collect the real data described in Section 3.1. The talker moves from 70 degrees to 140 degrees while uttering each word. The talker faces the microphone array and walks at a speed of about 80 cm per second. The distance between the talker and the microphone array is about 2 m. 216 phonetically-balanced words of the ATR Japanese speech database Set-A are recorded by the microphone array and a head-mounted microphone (SENNHEISER HMD410).

The word recognition accuracy is shown in Table 5.3. In *Head-mounted microphone*, speech recorded by the head-mounted microphone is fed to the recognizer. In *Single microphone*, speech recorded by the 8th microphone of the microphone array is fed to the recognizer. In *3-D Viterbi method with adaptive beamforming*, Δd in Equation (4.3) is 10, and μ and ν in Equation (4.5) are 80 and 10. The results showed that the word recognition of *3-D Viterbi method with adaptive beamforming* decreases by 8.8 % compared to *Head-mounted microphone* and is

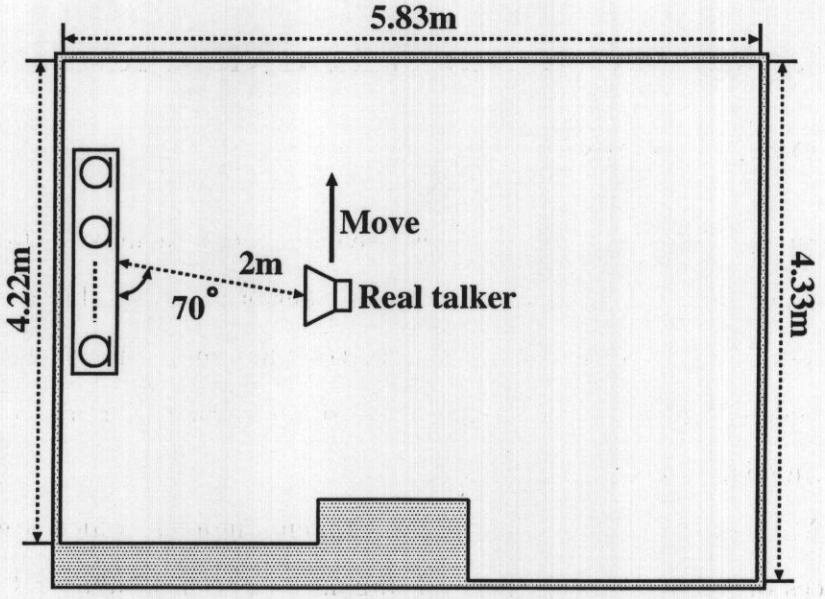

Figure 5.4. Speech data collection of a real moving talker.

Table 5.3. Evaluation of the 3-D Viterbi method for speaker-independent HMMs. Word recognition accuracy [%] is shown. Speech data of a real moving talker is used.

	SNR 21 dB
Head-mounted microphone	89.3
Single microphone	78.2
3-D Viterbi method with adaptive beamforming	80.5

slightly higher than that of *Single microphone.*

5.4 Summary

In this chapter, the application of adaptive beamforming to the 3-D Viterbi method was described. Section 5.1 explained the algorithm of adaptive beamforming. Section 5.2 evaluated the effect of adaptive beamforming through recognition experiments for real data. The results confirmed that adaptive beamforming drastically improves the performance of the 3-D Viterbi method compared to delay-and-sum beamforming for the moving-talker case as well as for the fixed-position talker case. Finally, Section 5.3 evaluated the performance of the 3-D Viterbi method through speaker-independent recognition experiments for a real moving talker. The results showed that the word recognition of *3-D Viterbi method with adaptive beamforming* decreases by 8.8 % compared to *Head-mounted microphone* and is slightly higher than that of *Single microphone.*

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This thesis has focused on the use of a microphone array to realize hands-free speech recognition in small conference rooms, where users can speak at a distance up to 2 or 3 meters from a microphone while moving.

Chapter 2 explained the principle of delay-and-sum beamforming and acoustic source localization by spatiotemporal analysis and CSP analysis. This chapter also discussed problems of current speech recognizers using a microphone array and several issues to be addressed in applying microphone arrays to hands-free speech recognition.

Chapter 3 described a new database of microphone array data in detail. This database includes speech data of a moving talker. The data collection procedure in a real room and the data generation procedure to simulate ideal conditions which approximately correspond to those of an anechoic chamber were explained.

Chapter 4 addressed the problem of a talker localization and tracking algorithm. This chapter introduced a new approach which integrates the speech recognition process and the talker localization into a unified framework and pro-

posed a new speech recognition algorithm based on a 3-dimensional Viterbi search. To perform initial evaluation of the 3-D Viterbi method, recognition experiments for simulated data were conducted. The results showed that speech features extracted in low SNR conditions are insufficient to estimate the talker direction accurately. To solve this problem, a pitch harmonics weight function was introduced and its effect was evaluated. The results showed that the performance of the 3-D Viterbi method drastically improved for the moving talker case as well as for the fixed-position talker case. Furthermore, the performance of the 3-D Viterbi method in a real room was investigated. The results showed that the pitch harmonics weight function is effective in a real room. Also, it was confirmed that a high resolution beamforming algorithm is necessary to achieve high recognition performance for practical use.

Chapter 5 described the application of adaptive beamforming to the 3-D Viterbi method. To evaluate the effect of adaptive beamforming, recognition experiments for real data were carried out. The results confirmed that adaptive beamforming drastically improves the performance of the 3-D Viterbi method compared to delay-and-sum beamforming for the moving-talker case as well as for the fixed-position talker case. Furthermore, the performance of the 3-D Viterbi method was evaluated through speaker-independent recognition experiments for a real moving talker. It was shown that the use of the 3-D Viterbi method in combination with adaptive beamforming yields slightly better results than using only a single microphone, for which adaptive beamforming and the 3-D Viterbi method cannot be used.

6.2 Future Work

Although the 3-D Viterbi method is a promising way to realize hands-free speech recognition in real environments, the following problems still remain for practical use. These are:

- the cost of the microphone array equipment,
- the computational efficiency of the beamforming algorithms,
- the computational efficiency of searching in the 3-D trellis space and
- the effect of multiple talkers.

A high SNR improvement can be achieved by increasing the number of microphones irrespective of the beamforming algorithms. However, this is impractical because of the increased cost of the equipment and of the computation time. Therefore, research on achieving a high SNR improvement by using a small number of microphones is necessary. Real-time processing is very important for speech recognition applications. Another issue to be addressed is the computational efficiency of the 3-D Viterbi method. This might be achieved by using a stack decoding approach [72]. The development of the real-time signal processing system is currently proceeded by members of our research group. The performance of the 3-D Viterbi method might degrade if multiple talkers have to be dealt with. In this case, it might be useful to apply an N-best algorithm for searching in the 3-dimensional trellis space [73]. Also, this paradigm might be applied to recognition of sound scenes, namely recognition of multiple sound sources which include talkers and noise sources.

References

[1] S. Furui and M. M. Sondhi, editors. *Advances in speech signal processing.* Marcel Dekker. Inc., 1991.

[2] K. F. Lee, editor. *Automatic speech recognition: the development of the SPHINX system.* Kluwer Academic, Boston, 1989.

[3] X. D. Huang, Y. Ariki, and M. A. Jack, editors. *Hidden Markov Models for Speech Recognition.* Edinburgh University Press, 1990.

[4] Jr. J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete-time processing of speech signals.* Macmillan Publishing Company, 1993.

[5] J. C. Junqua and J. P. Haton. *Robustness in automatic speech recognition.* Kluwer Adademic Publishers, 1996.

[6] C. H. Lee. On stochastic feature and model compensation approaches robust speech recognition. *Speech Communication*, 25:29–47, August 1998.

[7] M. J. F. Gales. Predictive model-based compensation schemes for robust speech recognition. *Speech Communication*, 25:49–74, August 1998.

[8] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27:113–120, April 1979.

[9] J. S. Lim and A. V. Oppenheim. All-pole modeling of degraded speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:197–210, June 1978.

[10] B. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55:1304–1312, June 1974.

[11] A. P. Varga and R. K. Moore. Hidden Markov model decomposition of speech and noise. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 845–848, 1990.

[12] M. J. F. Gales and S. J. Young. An improved approach to the hidden Markov model decomposition of speech and noise. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 233–236, 1992.

[13] F. Martin, K. Shikano, and Y. Minami. Recognition of noisy speech by composition of speech and noise. In *Proceedings of European Conference on Speech Communication and Technology*, pages 1031–1034, 1993.

[14] S. Nakamura, T. Takiguchi, and K. Shikano. Noise and room acoustics distorted speech recognition by HMM composition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 69–72, 1996.

[15] T. Takiguchi, S. Nakamura, Q. Huo, and K. Shikano. Adaptation of model parameters by HMM decomposition in noisy reverberant environments. In *Proceedings of ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 155–158, 1997.

[16] M. Omologo, P. Svaizer, and M. Matassoni. Environmental conditions and acoustic transduction in hands-free speech recognition. *Speech Communication*, 25:75–95, August 1998.

[17] S. U. Pillai. *Array signal processing.* Springer-Verlag, 1989.

[18] D. H. Johnson and D. E. Dudgeon. *Array signal processing — concepts and techniques —.* P T R Prentice Hall. Inc., 1993.

[19] J. Capon. High resolution frequency-wavenumber spectrum analysis. In *Proceedings of IEEE*, 57:1408–1418, 1969.

[20] S. W. Lang and J. H. McClellan. Frequency estimation with maximum entropy spectral estimators. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28:716–724, December 1980.

[21] D. Giuliani, M. Omologo, and P. Svaizer. Acoustic event localization using a crosspower-spectrum phase based technique. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 273–236, 1994.

[22] M. Omologo and P. Svaizer. Acoustic source location in noisy and reverberant environment using CSP analysis. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 921–924, 1996.

[23] P. Svaizer, M. Matassoni, and M. Omologo. Acoustic source location in a three-dimensional space using cross-power spectrum phase. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 231–234, 1997.

[24] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24:320–327, 1976.

[25] O. L. Frost. An algorithm for linearly constrained adaptive array processing. In *Proceedings of IEEE*, 60(8):926–935, 1972.

[26] L. J. Griffiths and C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas Propagation*, 30(1):27–34, January 1982.

[27] Y. Kaneda and J. Ohga. Adaptive microphone array system for noise reduction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(6):1391–1400, December 1986.

[28] F. Asano and S. Hayamizu. Speech enhancement using array signal processing based on the coherent-subspace method. *IEICE Transactions on Fundamentals*, E80-A(11):2276–2285, November 1997.

[29] J. L. Flanagan, A. C. Surendran, and E. E. Jan. Spatially selective sound capture for speech and audio processing. *Speech Communication*, 13:207–222, 1993.

[30] E. E. Jan. *Parallel processing of large scale microphone arrays for sound capture.* PhD thesis, The State University of New Jersey, 1995.

[31] R. O. Schmidt. Multiple emitter location and signal parameter estimation. In *Proceedings of RADC Spectral Estimation Workshop*, pages 243–258, 1979.

[32] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer. Use of different microphone array configurations for hands-free speech recognition in noisy

and reverberant environment. In *Proceedings of European Conference on Speech Communication and Technology*, pages 347–350, 1997.

[33] M. Inoue, T. Yamada, S. Nakamura, and K. Shikano. Comparative experiments of microphone arrays for speech recognition. In *Technical Report of the Institute of Electronics, Information and Communication Engineers, SP96-89*, pages 13–20, 1997 (in Japanese).

[34] M. Inoue, T. Yamada, S. Nakamura, and K. Shikano. Measure of microphone array designs for speech recognition. In *Proceedings of Acoustical Society of Japan, 2-3-6*, pages 515–516, March 1997 (in Japanese).

[35] M. Inoue, S. Nakamura, T. Yamada, and K. Shikano. Microphone array design measures for hands-free speech recognition. In *Proceedings of European Conference on Speech Communication and Technology*, pages 331–334, 1997.

[36] M. Inoue, T. Yamada, S. Nakamura, and K. Shikano. Microphone array design measures for hands-free speech recognition. *Transactions of the Institute of Electronics, Information and Communication Engineers*, J81–D-II(11):2511–2518, 1998 (in Japanese).

[37] Q. Lin, E. Jan, C. Che, and B. Vries. System of microphone arrays and neural networks for robust speech recognition in multimedia environment. In *Proceedings of International Conference on Spoken Language Processing*, pages 1247–1250, 1994.

[38] D. Giuliani, M. Omologo, and P. Svaizer. Talker localization and speech recognition using a microphone array and a cross-powerspectrum phase analysis. In *Proceedings of International Conference on Spoken Language Processing*, pages 1243–1246, 1994.

[39] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer. Continuous speech recognition in noisy environment using a four microphone array. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 860–863, 1995.

[40] K. C. Yen and Y. Zhao. Robust automatic speech recognition using a multi-channel signal separation front-end. In *Proceedings of International Conference on Spoken Language Processing*, pages 1337–1340, 1996.

[41] K. Kiyohara, Y. Kaneda, S. Takahashi, H. Nomura, and J. Kojima. A microphone array system for speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 215–218, 1997.

[42] E. Lleida, J. Fernandez, and E. Masgrau. Robust continuous speech recognition system based on a microphone array. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 241–244, 1998.

[43] T. Hughes, H. Kim, J. DiBiase, and H. Silverman. Using a real time, tracking microphone array as input to an HMM speech recognizer. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 249–252, 1998.

[44] S. Nakamura, T. Yamada, and K. Shikano. Speech recognition with source localization by microphone array. In *Proceedings of Acoustical Society of Japan, 1-5-8*, pages 15–16, March 1995 (in Japanese).

[45] T. Yamada, S. Nakamura, and K. Shikano. Speech recognition with speaker localization by microphone array. In *Proceedings of Acoustical Society of Japan, 1-2-4*, pages 7–8, Sept. 1995 (in Japanese).

[46] T. Yamada, S. Nakamura, and K. Shikano. Speech recognition with speaker localization by microphone array. In *Technical Report of the Institute of Electronics, Information and Communication Engineers, SP95-65*, pages 27–34, 1995 (in Japanese).

[47] T. Yamada, S. Nakamura, and K. Shikano. Robust speech recognition with speaker localization by a microphone array — performance evaluation in real environments —. In *Proceedings of Acoustical Society of Japan, 1-5-19*, pages 45–46, March 1996 (in Japanese).

[48] T. Yamada, S. Nakamura, and K. Shikano. Robust speech recognition with speaker localization by a microphone array. In *Proceedings of International Conference on Spoken Language Processing*, pages 1317–1320, 1996.

[49] T. Yamada, S. Nakamura, and K. Shikano. Hands-free speech recognition with talker localization by a microphone array. *Transactions of Information Processing Society of Japan*, 39(5):1275–1284, 1998 (in Japanese).

[50] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko. Computer-steered microphone arrays for sound transduction in large rooms. *Journal of the Acoustical Society of America*, 78(5):1508–1518, November 1985.

[51] H. F. Silverman and S. E. Kirtman. A two-stage algorithm for determining talker location from linear microphone array data. *Computer, Speech and Language*, 6(5):129–152, June 1992.

[52] M. Omologo and P. Svaizer. Talker localization and speech enhancement in a noisy environment using a microphone array based acquisition system. In *Proceedings of European Conference on Speech Communication and Technology*, pages –, 1993.

[53] M. Crawford, G. J. Brown, M. Cooke, and P. Green. Design, collection and analysis of a multi-simultaneous-speaker corpus. In *Proceedings of the Institute of Acoustics*, 16:183–190, 1994.

[54] E. E. Jan, P. Svaizer, and J. L. Flanagan. A database for microphone array experimentation. In *Proceedings of European Conference on Speech Communication and Technology*, pages 813–816, 1995.

[55] S. Nakamura, K. Hiyane, F. Asano, and T.Endo. Sound scene database in real acoustical environments. In *Proceedings of International Workshop on East-Asian Language Resource and Evaluation (EALREW), — Oriental COCOSDA Workshop'98 —*, pages 17–20, 1998.

[56] K. Takeda, Y. Sagisaka, and S. Katagiri. Acoustic-phonetic labels in a japanese speech database. In *Proceedings of European Conference on Speech Communication and Technology*, pages 13–16, 1987.

[57] T. Yamada, S. Nakamura, and K. Shikano. Hands-free speech recognition based on 3-d trellis search using a microphone array. In *Technical Report of Information Processing Society of Japan, 97-SLP-15-6*, pages 35–40, 1997 (in Japanese).

[58] T. Yamada, S. Nakamura, and K. Shikano. Hands-free speech recognition based on 3-d trellis search using a microphone array. In *Proceedings of Acoustical Society of Japan, 3-6-21*, pages 129–130, March 1997 (in Japanese).

[59] T. Yamada, S. Nakamura, and K. Shikano. Hands-free speech recognition based on 3-d viterbi search using a microphone array. In *Technical Report of the Institute of Electronics, Information and Communication Engineers, SP97-22*, pages 31–38, 1997 (in Japanese).

[60] T. Yamada, S. Nakamura, and K. Shikano. Speech recognition of a moving talker based on 3-D Viterbi search using a microphone array. In *Proceedings of International Joint Conference on Artificial Intelligence Workshop on Computational Auditory Scene Analysis*, pages 113–116, 1997.

[61] T. Yamada, S. Nakamura, and K. Shikano. Hands-free speech recognition based on 3-d viterbi search using a microphone array — performance evaluation in real environments —. In *Proceedings of Acoustical Society of Japan, 2-Q-23*, pages 161–162, Sept. 1997 (in Japanese).

[62] T. Yamada, S. Nakamura, and K. Shikano. Effect of cmn on 3-d trellis search using a microphone array. In *Proceedings of Acoustical Society of Japan, 3-6-10*, pages 101–102, March 1998 (in Japanese).

[63] T. Yamada, S. Nakamura, and K. Shikano. Hands-free speech recognition based on 3-D Viterbi search using a microphone array. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 245–248, 1998.

[64] J. R. Bellegarda and D. Nahamoo. Tied mixture continuous parameter models for large vocabulary isolated speech recognition. In *Proceedings of European Conference on Speech Communication and Technology*, pages 13–16, 1989.

[65] D. Giuliani, M. Omologo, and P. Svaizer. Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation. In *Proceedings of International Conference on Spoken Language Processing*, pages 1329–1332, 1996.

[66] M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani. Microphone array based speech recognition with different talker-array positions. In *Proceedings*

of IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 227–230, 1997.

[67] J. E. Adcock, Y. Gotoh, D. J. Mashao, and H. F. Silverman. Microphone-array speech recognition via incremental MAP training. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 897–900, 1996.

[68] T. Yamada, S. Nakamura, and K. Shikano. An effect of adaptive beamforming on hands-free speech recognition based on 3-d viterbi search. In *Technical Report of the Institute of Electronics, Information and Communication Engineers, EA98-3*, pages 13–20, 1998 (in Japanese).

[69] T. Yamada, S. Nakamura, and K. Shikano. Hands-free speech recognition based on 3-d viterbi search and amnor. In *Proceedings of Acoustical Society of Japan, 2-1-12*, pages 61–62, Sept. 1998 (in Japanese).

[70] T. Yamada, S. Nakamura, and K. Shikano. An effect of adaptive beamforming on hands-free speech recognition based on 3-D Viterbi search. In *Proceedings of International Conference on Spoken Language Processing*, pages 381–384, 1998.

[71] T. Yamada, S. Nakamura, and K. Shikano. Hands-free speech recognition based on 3-d viterbi search using adaptive beamforming. *Transactions of Information Processing Society of Japan*, 40(2):460–468, 1998 (in Japanese).

[72] L. R. Bahl, R. Bakis, P. S. Cohen, A. G. Cole, F. Jelinek, B. L. Lewis, and R. L. Mercer. Speech recognition of a natural text read as isolated words. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1168–1171, 1983.

[73] P. Heracleous, T. Yamada, S. Nakamura, and K. Shikano. Simultaneous recognition of multiple sound sources based on 3-d n-best search. In *Proceedings of Acoustical Society of Japan*, March 1998 (to be appeared).

List of Publications

Journal Paper

1. T. Yamada, S. Nakamura, and K. Shikano. Hands-free speech recognition with talker localization by a microphone array. *Transactions of Information Processing Society of Japan*, 39(5):1275–1284, May 1998 (in Japanese).

2. T. Yamada, S. Nakamura, and K. Shikano. Hands-free speech recognition based on 3-D Viterbi search using adaptive beamforming. *Transactions of Information Processing Society of Japan*, 40(2):460–468, Feb. 1999 (in Japanese).

Journal Paper (co-author)

3. M. Inoue, T. Yamada, S. Nakamura, and K. Shikano. Microphone array design measures for hands-free speech recognition. *Transactions of the Institute of Electronics Information and Communication Engineers*, J81–D-II(11):2511–2518, November 1998 (in Japanese).

International Conference

4. T. Yamada, S. Nakamura, and K. Shikano. Robust speech recognition with speaker localization by a microphone array. In *Proceedings of International Conference on Spoken Language Processing*, pages 1317–1320, October 1996.

5. T. Yamada, S. Nakamura, and K. Shikano. Speech recognition of a moving talker based on 3-D Viterbi search using a microphone array. In *Proceedings of International Joint Conference on Artificial Intelligence Workshop on Computational Auditory Scene Analysis*, pages 113–116, August 1997.

6. T. Yamada, S. Nakamura, and K. Shikano. Hands-free speech recognition based on 3-D Viterbi search using a microphone array. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 245–248, May 1998.

7. T. Yamada, S. Nakamura, and K. Shikano. An effect of adaptive beamforming on hands-free speech recognition based on 3-D Viterbi search. In *Proceedings of International Conference on Spoken Language Processing*, pages 381–384, December 1998.

International Conference (co-author)

8. S. Nakamura, T. Yamada, T. Takiguchi, and K. Shikano. Hands free speech recognition by a microphone array and HMM composition. In *Proceedings of International Workshop on Human Interface Technology*, pages 33–38, October 1995.

9. S. Nakamura, T. Yamada, T. Takiguchi, and K. Shikano. Hands-free speech recognition by a microphone array and HMM composition. In *Proceedings of ASA and ASJ Third Joint Meeting*, pages 1149–1154, December 1996.

10. K. Shikano, S. Nakamura, T. Yamada, T. Takiguchi, E. Yamamoto, M. Inoue, R. Nagai, and T. Aoki. Hands-free speech recognition and lip-reading/synthesis. In *Proceedings of International Workshop on Human Interface Technology*, pages 47–54, March 1997.

11. M. Inoue, S. Nakamura, T. Yamada, and K. Shikano. Microphone array design measures for hands-free speech recognition. In *Proceedings of European Conference on Speech Communication and Technology.* pages 331–334, September 1997.

Domestic Conference

12. T. Yamada, S. Nakamura, and K. Shikano. Speech recognition with speaker localization by microphone array. In *Technical Report of the Institute of Electronics, Information and Communication Engineers, SP95-65*, pages 27–34, October 1995 (in Japanese).

13. T. Yamada, S. Nakamura, and K. Shikano. Hands-free speech recognition based on 3-D trellis search using a microphone array. In *Technical Report of Information Processing Society of Japan, 97-SLP-15-6*, pages 35–40, February 1997 (in Japanese).

14. T. Yamada, S. Nakamura, and K. Shikano. Hands-free speech recognition based on 3-D Viterbi search using a microphone array. In *Technical Report of the Institute of Electronics, Information and Communication Engineers, SP97-22*, pages 31–38, June 1997 (in Japanese).

15. T. Yamada, S. Nakamura, and K. Shikano. An effect of adaptive beamforming on hands-free speech recognition based on 3-D Viterbi search. In *Technical Report of the Institute of Electronics, Information and Communication Engineers, EA98-3*, pages 13–20, April 1998 (in Japanese).

16. T. Yamada, S. Nakamura, and K. Shikano. Speech recognition with speaker localization by microphone array. In *Proceedings of Acoustical Society of Japan, 1-2-4*, pages 7–8, September 1995 (in Japanese).

17. T. Yamada, S. Nakamura, and K. Shikano. Robust speech recognition with speaker localization by a microphone array — performance evaluation in real environments —. In *Proceedings of Acoustical Society of Japan, 1-5-19*, pages 45–46, March 1996 (in Japanese).

18. T. Yamada, S. Nakamura, and K. Shikano, Hands-free speech recognition based on 3-D trellis search using a microphone array. In *Proceedings of Acoustical Society of Japan, 3-6-21*, pages 129–130, March 1997 (in Japanese).

19. T. Yamada, S. Nakamura, and K. Shikano. Hands-free speech recognition based on 3-D Viterbi search using a microphone array — performance evaluation in real environments —. In *Proceedings of Acoustical Society of Japan, 2-Q-23*, pages 161–162, September 1997 (in Japanese).

20. T. Yamada, S. Nakamura, and K. Shikano. Effect of CMN on 3-D trellis search using a microphone array. In *Proceedings of Acoustical Society of Japan, 3-6-10*, pages 101–102, March 1998 (in Japanese).

21. T. Yamada, S. Nakamura, and K. Shikano. Hands-free speech recognition based on 3-D Viterbi search and AMNOR. In *Proceedings of Acoustical Society of Japan, 2-1-12*, pages 61–62, September 1998 (in Japanese).

Domestic Conference (co-author)

22. M. Inoue, T. Yamada, S. Nakamura, and K. Shikano, Comparative experiments of microphone arrays for speech recognition, In *Technical Report of the Institute of Electronics, Information and Communication Engineers, SP96-89*, pages 13–20, January 1997 (in Japanese).

23. F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura. Application of subspace-based speech enhancement to speech recognition. In *Technical Report of the Institute of Electronics, Information and Communication Engineers, EA97-17*, pages 17–24, June 1997 (in Japanese).

24. R. Nishioka, T. Yamada, S. Nakamura, and K. Shikano. Multiple beamforming microphone array with autodirective source localization. In *Technical Report of the Institute of Electronics, Information and Communication Engineers, EA97-98*, pages 49–56, January 1998 (in Japanese).

25. S. Nakamura, T. Yamada, and K. Shikano. Speech recognition with source localization by microphone array. In *Proceedings of Acoustical Society of Japan, 1-5-8*, pages 15–16, March 1995 (in Japanese).

26. T. Aoki, T. Yamada, T. Takiguchi, S. Nakamura, and K. Shikano. Speech recognition experiments in real environments using a microphone array and HMM composition. In *Proceedings of Acoustical Society of Japan, 2-Q-2*, pages 133–134, September 1996 (in Japanese).

27. M. Inoue, T. Yamada, S. Nakamura, and K. Shikano. Measure of microphone array designs for speech recognition. In *Proceedings of Acoustical Society of Japan, 2-3-6*, pages 515–516, March 1997 (in Japanese).

28. F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura. Evaluation of subspace-based speech enhancement by speech recognition. In *Proceedings of Acoustical Society of Japan, 2-4-6*, pages 539–540, September 1997 (in Japanese).

29. R. Nishioka, T. Yamada, S. Nakamura, and K. Shikano. Multiple beamforming microphone array with autodirective source localization. In *Proceedings of Acoustical Society of Japan, 2-5-19*, pages 527–528, March 1998 (in Japanese).

30. T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano. Localization of multiple sound sources based on maximum likelihood estimation using a microphone array. In *Proceedings of Acoustical Society of Japan, 2-9-15*, pages 543–544, September 1998 (in Japanese).

31. T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano. Localization of multiple sound sources using a synchronous addition of crosspower spectrum phase coefficients. In *Proceedings of Acoustical Society of Japan*, March 1998 (in Japanese) (to be appeared).

32. P. Heracleous, T. Yamada, S. Nakamura, and K. Shikano. Simultaneous recognition of multiple sound sources based on 3-D N-best search. In *Proceedings of Acoustical Society of Japan*, March 1999 (to be appeared).

Master thesis

33. T. Yamada. Robust speech recognition with speaker localization by a microphone array. Master thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, March 1996 (in Japanese).