NAIST-IS-DT9861029

Doctor Thesis

Simultaneous Recognition of Distant-talking Speech of Multiple Sound Sources

Panikos Heracleous

February 5, 2002

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

Panikos Heracleous

Thesis committee: Kiyohiro Shikano, Professor
Tsukasa Ogasawara, Professor
Satoshi Nakamura, Dr.
Hiroshi Saruwatari, Associate Professor

Simultaneous Recognition of Distant-talking Speech of Multiple Sound Sources*

Panikos Heracleous

Abstract

This thesis deals with the recognition of distant-talking speech, and particularly with the simultaneous recognition of multiple talkers. The recognition of distant-talking speech plays an important role in any practical speech recognition system. Factors that should be considered include noisy and reverberant environments, the presence of multiple talkers, moving talkers, etc. Most of the hands-free speech recognition systems are microphone array-based, since the microphone can take advantage of the spatial and acoustical information of a sound source. More specifically, a microphone array can form multiple beams and therefore can be electronically steered simultaneously to multiple directions at the same time. In contrast, the use of a single microphone provides limited directional sensitivity and cannot be applied for the localization of multiple sound sources without physical steering. A serious problem that must me solved in the recognition of distant-talking speech is the talker localization. Some approaches localize the talker using the power information. However, in highly reverberant environments and under low SNR conditions, the talker localization appears to be difficult. The 3-D Viterbi search method integrates talker localization and speech

*Doctor Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT9861029, February 5, 2002.

recognition. Although this method performs efficiently even in the case of a moving talker, its applications are restricted to the presence of one talker. In order to deal with multiple sound sources we are proposing the 3-D N-best search method able to recognize simultaneously multiple sound sources. Our proposed method is one-pass search algorithm, which performs search in all directions and keeps N-best hypotheses for each word and direction hypothesis. Although our method integrates two existing technologies - 3-D Viterbi search and N-best search - into one complete system, the results obtained in the first evaluation were very poor. Implementing two additional techniques - a clustering and a likelihood normalization technique - into the baseline system drastically increased the performance of our system. The performance of our system was evaluated through experiments for the recognition of the distant-talking speech of two talkers and using several microphone array geometries. The obtained results are very promising, and especially in the case of 32-channels microphone array the achieved Simultaneous Word Accuracy in Top 1 hypothesis was 72.49 % and in Top 3 hypothesis was 86.25 %.

Keywords:

Speech recognition, distant-talking speech, microphone arrays, multiple sound sources, real environments

Acknowledgments

Professor Kiyohiro Shikano (Nara Institute of Science and Technology) is gratefully acknowledged for his guidance during my studies as research student and during the doctoral course.

I would like to express my greatest appreciation to Professor Tsukasa Ogasawara (Nara Institute of Science and Technology) and Associate Professor Hiroshi Saruwatari (Nara Institute of Science and Technology) for their precious suggestions.

I would like to express my sincere gratefulness to Associate Professor Satoshi Nakamura, who is currently Head of Department 1, Spoken Language Translation Research Labs at ATR. His ideas in the scientific part and his advises extended to the daily life in general, made this work to be accomplished.

I would like to thank Research Associate Dr. Shiro Ise who is currently Associate Professor at Kyoto University, and Research Associate Dr. Jinlin Lu who is currently a Researcher of Department 4 at ATR Spoken Language Translation Research Labs for their beneficial comments.

I would like to thank the members of Speech and Acoustics Laboratories in Nara Institute of Science and Technology and, especially my colleagues Dr. Takeshi Yamada, Dr. Tetsuya Takiguchi, Dr. Yamamoto Eli and Dr. Alexandre Girardi for their help at the beginning of the doctor course and for the useful comments and suggestions.

I would like to thank Dr. Seiichi Yamamoto, President of Spoken Language Translation Research Labs at ATR. My thanks go to Dr. Tomoko Matsui, Professor Kuldip Kumar Paliwal, and to all members of Department 1, Spoken Language Translation Research Labs at ATR, for their suggestions and helpful discussions during the meetings.

I would like to thank the members of Multimedia Interface Group, KDDI R&D Labs for their help and understanding during writing this thesis.

A great number of friends, especially Hakamata, Suyama, and Ishikawa families are acknowledged for their kindness and for giving me useful information about the life in Japan.

My greatest thanks to the Japanese Ministry of Education, Culture, Sports, Science and Technology for providing me the opportunity to study in Japan.

Finally, this work would not be possible without the supporting of my family.

Dedication

To my father and my mother

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Preface

Speech is the most friendly to human beings communication tool. By using speech we can easily communicate with each other and we can express our thinking and feelings in a compact and precise way. People realized the importance of the speech in very early times. The "civilized" human being tried to study and understand the process of the speech production and analyze the factors related to that from the beginning of his existence. As it was expected, the speech process stimulated the curiosity of the ancient Greek scientists, and first of all tried to analyze and understand the several voices. Moreover, very interesting and useful works related to speech and language were done. Romans followed the Greek scientists and offered us, also, very interesting studies.

At the end of the 18th century a great number of studies about speech were done. In the 1789 in the Academy of Science of St. Petersburg in Russia was a work whose objective was the recognition of five vowels, and in the 18th century the Hungarian scientist Kempelen Farkas built the first 'Talking Machine".

The number of the speech related studies was dramatically increased since it

had been realized the importance of the speech in the human-machine communication, too. In this kind of communication the speech offers a flexible solution and can replace in a very easy way other communication tools, such as the hands.

The rapidly increased information about speech and the increased expectations from that made the splitting of its studies into several fields necessary. Among others, speech coding, speech recognition, speech synthesis, speech translation and speech understanding are some of the fields, which a great number of researchers work on. In our days a main research topic is the speech recognition, which is a basic tool in the effective and reliable human-machine communication.

1.2 Distant-talking Speech Recognition Problem

The automatic speech recognition plays an important role in the human-machine communication, and a lot of applications, which can make our life easier and more comfortable, can be developed based on a speech recognizer. Among other uses, people with physical problems (blind people, handicapped, etc.) can easily communicate with a computer, a Automatic Cash Machine or other devices just by using speech. Moreover, a speech recognition based system can replace a secretary or an operator and can provide information automatically.

The first built speech recognition system recognized speech received by a close-talking microphone, and in relatively clean environments. However, for practical use the hands-free speech recognition should be also considered. A hands-free speech recognition system is not only user-friendly, but avoids also limitations that are required in a "close and manually" operated system. Most of the hands-free speech recognition systems are microphone array based. Figure 1.1 shows a microphone array based hands-free speech recognizer.

Considering a speech recognition system in the real world, the number of the

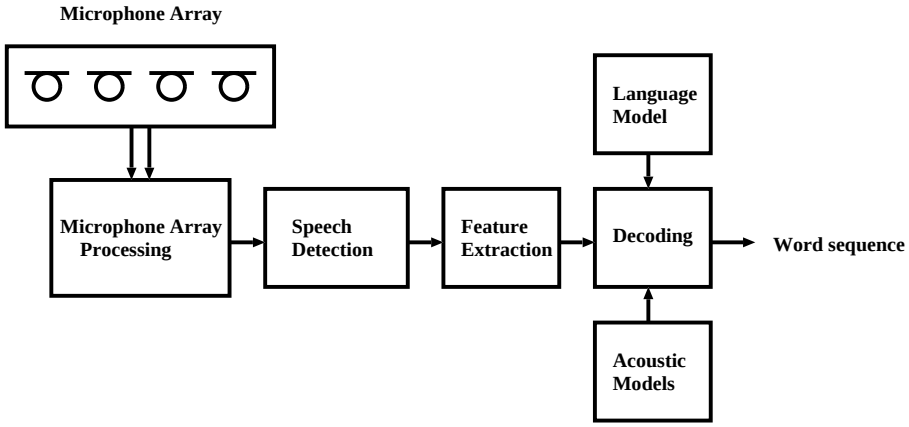

Figure 1.1. Microphone array-based hands-free speech recognition system

problems that must be solved is significant. The additional factors that should be considered include talker localization, noise, reverberation, echo, multiple moving talkers etc. In this thesis we deal with the simultaneous recognition of multiple sound sources in real environments. Our idea can be implemented in teleconferences with multiple and moving talkers, in Parliament's situation where the simultaneous talking and interruption is also allowed, in lectures, etc. Moreover, a system based on our idea can be integrated into a complete speech translation system able to recognize and translate simultaneously multiple talkers.

1.3 Current Status of Research Field

Although the investigation of distant-talking speech recognition is still at the beginning, very interesting works have been already reported [17, 18, 19, 20, 44, 21, 22, 16] and a great number of researchers are investigating the distant-talking speech recognition related fields. A critical issue in the distant-talking speech recognition is the choice of the device used for receiving the speech signals.

Although the use of a single channel was investigated, the microphone arrays [1, 4, 2] appear to have a significant role in this field and be an efficient tool in the distant-talking speech recognition [15, 6, 14, 32]. Several microphone array geometries exist, such as the linear uniform array, the planar array, etc. The number of the used channels is also critical issue. By increasing the number of microphones the performance of the speech recognizer is also increased [5]. However, a trade-off between the system complexity or cost and the performance is necessary. In most of the cases 16 to 32 channels are used.

The advantage, which makes the microphone array popular, is the directional sensitivity. More specifically, a microphone array can form a directive pattern and can acquire signals from the desired direction with high quality. Several beamforming techniques are used in hands-free speech recognition. The oldest and simplest one is the delay-and-sum beamforming [7]. The main disadvantage of the delay-and-sum beamforming is the limited speech enhancement. Therefore, more efficient beamforming techniques are used in some approaches [8], such as the adaptive beamforming [9, 10, 11].

The speech corpus collection is also important issue in the hands-free speech recognition. In some works data was collected in real environments, and in other cases the real data was simulated. However, instead of collecting real data, an effective and simple method is to collect impulse responses in the real environments using a microphone array [12, 13]. The real data are then obtained by the convolution of the speech data and the impulse responses.

1.4 Idea and Proposed Method

In the hands-free speech recognition an important factor is the talker localization. Most of the hands-free speech recognition systems are microphone array based,

since the microphone can take advantage of the spatial and temporal acoustic information. The conventional speech recognizers operate in two stages. First the talker is localized by using a localization method and then, speech information is extracted from the hypothesized direction. The extracted speech in parameterized format is the input to the speech recognizer.

This thesis introduces a new method for recognizing simultaneously speech of multiple talkers. The main idea of our proposed method is to integrate speech recognition and talker localization. Our method is based on the 3-D Viterbi search, extended to the 3-D N-best search. Our method is also microphone array-based, but for talker localization we use a probabilistic approach based on the acoustic information. The beauty of our approach is that avoids the problematic use of the power for the talker localization, and that operates without any a-priori knowledge of the sound direction. Moreover, our approach can be used for the recognition of the speech of multiple moving talkers, too. The proposed method will be described in details in later chapters.

1.5 Thesis Overview

This thesis focuses on the simultaneous distant-talking speech recognition of multiple talkers in real environments using microphone arrays. Our approach uses microphone array and it is based on the 3-D N-best search algorithm.

Chapter 2 covers the problem of the hands-free speech recognition of multiple sound sources and contains an overview of the existing approaches.

In Chapter 3 we address the speech recognition issues and we describe the several components of a complete speech recognition system.

Chapter 4 discusses the microphone arrays, the beamforming algorithms, and the sound source localization problem. Some beamforming and sound source

localization techniques are described in this chapter.

Chapter 5 contains an overview of the 3-D Viterbi search algorithm, which is the background of our research. The basic idea, the formulation, its advantages and disadvantages, and some results obtained in previous works are introduced in this chapter.

In Chapter 6 we introduce our proposed 3-D N-best Search method able to recognize speech of multiple talkers. Additional implemented techniques, such as a clustering technique and a likelihood normalization technique are also introduced.

The evaluation of the 3-D N-best search method based speech recognition system is covered in the Chapter 7. Experiments on simulated clean data were carried out for the speech recognition of two talkers located both at fixed position, and with one fixed and one moving talker. Experiments using simulated clean data were also carried out for the recognition of three talkers located at fixed positions. This chapter also describes the evaluation of the proposed 3-D N-best search based speech recognition system using reverberant data. The image method was used to simulate the reverberant environment. The real performance of our system was evaluated through experiments using data recorded in a noisy and reverberant environment.

Finally, Chapter 8 summarizes our work and discusses the remained problems and future work.

Chapter 2

Speech Recognition of Multiple Sound Sources

In this chapter, we discuss the speech recognition problem of multiple sound sources, and particularly the distant-talking speech recognition. First, we describe the problems that this task faces, and in the following based on the available literature, we will introduce recent approaches.

2.1 Problems

The recognition of distant-talking speech uttered by multiple talkers is a very difficult task and requires solution to a great number of problems. They should be considered and solved not only problems related to hands-free speech recognition, but also additional problems originated from the presence of the multiple talkers or multiple sound sources. Figure 2.1 illustrates the problem described above.

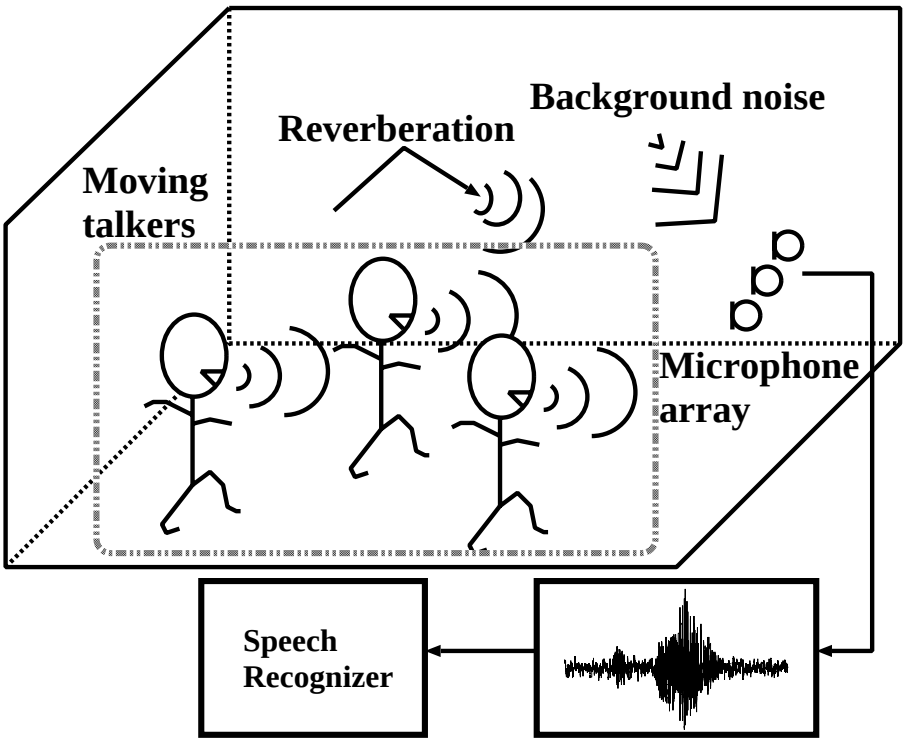

Figure 2.1. Distant-talking speech recognition problem

- **Sound source localization**

  The accurate hands-free speech recognition requires accurate sound source localization. The localization errors significantly degrade the speech recognition system performance and therefore, an important requirement in hands-free speech recognition is a robust sound source localization strategy [23, 24, 25, 26]. Although, efficient methods exist for the localization of sound source located at fixed position, the case of a moving talker's localization faces serious difficulties. Most of the sound source localization methods use short- or long-term power. However, in noisy and reverberant environments the sound source localization appears to be difficult.

- **Reverberation and Echo**

  The signal received by the hands-free speech recognizer is distorted by the echo and by the reverberant environmental conditions. Therefore, efficient de-reverberation techniques are necessary. Methods, such as de-reverberation by inverse processing or de-reverberation by composite inversion, provide efficient solution to the reverberation problem [27].

- **Additive Noise**

  The high quality input speech to a recognizer is a basic requirement for the high performance. However, in practice the input distant-talking speech is affected by noise components. A practical hands-free speech recognition system should consider the presence of several kinds of noise, too. The effect of the noise seriously degrades the performance of the system. A great number of works deals with the noise reduction problem and speech enhancement [28, 29, 30, 31, 33].

- **Mismatch between training and testing conditions**

  Most of the speech recognizers are based on acoustic models trained by clean speech received by a close-talking microphone. However, in the recognition stage the environmental conditions include reverberation and background noise. Since it is very difficult to collect training data that cover all possible environmental conditions, most of the hands-free speech recognizers try to solve this problem by adapting the clean models to the environmental conditions using an efficient adaptation technique.

- **Signals Estimation**

  In the recognition of the speech of multiple talkers the estimation of the signals from the input mixed signal is required. A problem that those

systems face is the correlation between the several talkers. More specifically, between closely located or crossing each other talkers, the correlation is very high and the signals separation is difficult. An effective tool to separate the speech signals originated from multiple talkers is the use of a microphone array. The microphone array forms the so-called directive pattern or beam, and can be steered to a specific direction. Therefore, a microphone can acquire signals from a desired direction with high quality and suppress the noise or other components arriving from other directions. In some approaches, a microphone array is used to extract speech signal with high quality by forming a beam to the hypothesized direction. The direction is estimated by a conventional localization method. In our approach, the microphone array is steered to all directions simultaneously without any a-priori information about the signal's direction, and the speech recognition is based on the signal's acoustic information. Finally, other approaches -such as the Blind Source Separation approaches- are signal processing oriented and can be integrated into a recognizer for multiple sound sources speech.

- **Voice activity detection**

  The voice activity detection is critical in the hands-free speech communication. In contrast with the earliest "push-to-talk" systems, new methods are reported to solve this problem [34, 35, 36].

2.2 Approaches

Although in the literature few works can be found which specifically deal with the simultaneous recognition of the speech of multiple sound sources, in this section we try to describe the possible solutions to this task and discuss the problems

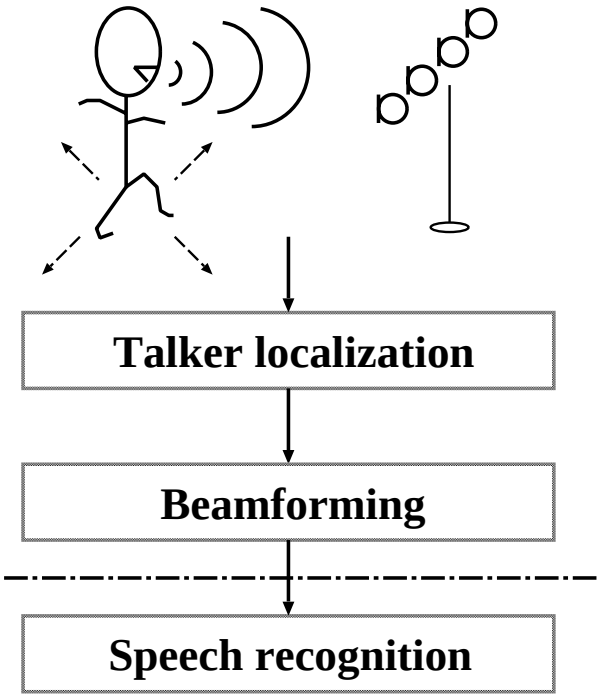

Figure 2.2. Conventional talker localization and speech recognition

that each of those approach faces.

- **Conventional microphone array-based methods**

  The conventional microphone array-based systems for hands-free speech recognition [37, 24, 38, 39, 40, 41, 42] operate in multiple stages. In the first stage, a source localization method - will be described in details in the next chapters - is used to determine the DOA (Direction Of Arrival) of the signals. Although most of those methods are applied for one sound source, we assume that can be also applied for the estimation of the DOA of multiple propagating signals. In the following, the microphone array is steered to the hypothesized directions by forming beams, and the signals from these directions are extracted with higher quality comparing to the

signals arriving from the undesired directions. The beamformed signals are the input to speech recognizer.

The majority of these methods are based on the use of short- or long-term power. Thus, a main disadvantage of these methods leads on this fact. However, in highly reverberant or under low SNR (signal-to-noise ratio) conditions the use of the power cannot be reliable. Moreover, in the case of a moving talker, his localization using power appears to be very difficult. A serious additional disadvantage of these methods is their dependency of the accurate talker localization. In such recognition systems, the talker localization is deterministic and the localization errors seriously decreases the performance. Figure 2.2 illustrates the operating stages of these methods.

- **Blind source separation-based methods**

  Blind Source Separation (BSS) methods [60] attempt to estimate the signals contained in the mixed signal and received by the array sensors without using any a-priori information about those signals. Independent Component Analysis (ICA) [60] is a set of techniques developed in the last few years to solve the BSS problem. The ICA techniques estimate a set of linear filters to separate the mixed signals under the assumption that the sources are statistically independent. Several ICA techniques exist, such as Instantaneous Mixing ICA, Convolutional Mixing ICA, etc.

  Assuming that M microphone signals $y_m[n]$, $\mathbf{Y}[n] = (y_1[n], y_2[n], ...y_M[n])$ are obtained by a linear combination of the M unobserved source signals $x_m[n]$, denoted by $\mathbf{X}[n] = (x_1[n], x_2[n], ..., x_M[n])$:

$$\mathbf{Y}[n] = \mathbf{V}\mathbf{X}[n] \tag{2.1}$$

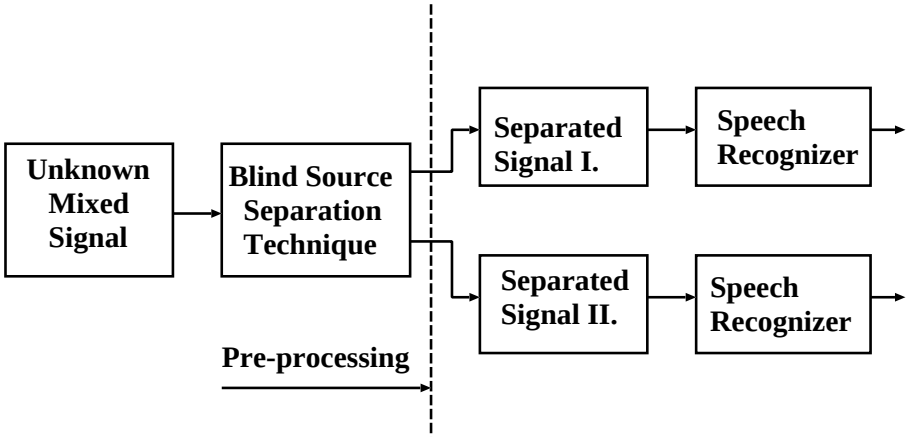

Figure 2.3. A BSS-based scenario for recognition of multiple sound sources

In the Eq. 2.1 $\mathbf{V}$ is the M×M mixing matrix. The blind source separation problem consists of estimating a separating matrix $\mathbf{W} = \mathbf{V}^{-1}$ from the observed signals. The source signals can then be recovered by:

$$\mathbf{X}[n] = \mathbf{W}\mathbf{Y}[n] \tag{2.2}$$

A BSS method can be a component of an integrated recognizer for the simultaneous recognition of multiple sound sources, providing the separated signals as the input to conventional recognizers operating in parallel. However, the BSS methods face a seriously problem. Namely, those methods are highly signal processing oriented and require significant amount of computation. Moreover, the knowledge in advance of the number of propagating sources is also a requirement. Figure 2.3 describes the scenario of the speech recognition of multiple sound sources based on BSS techniques.

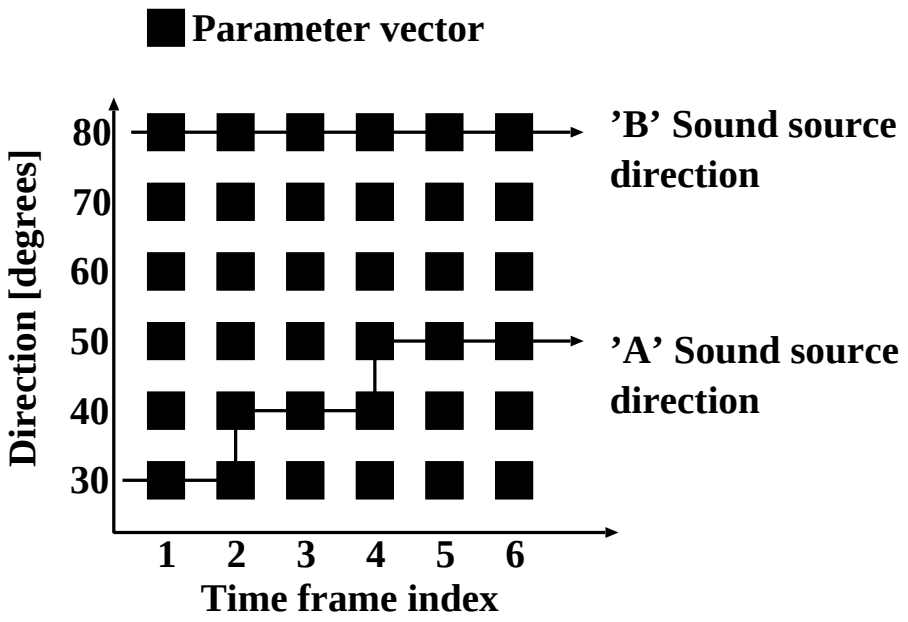

Figure 2.4. Integrated talker localization and speech recognition

- **Integration of talker localization and speech recognition**

  An additional approach for hands-free simultaneous recognition of the speech of multiple sound sources uses microphone array and it is based on the integration of the talker localization and speech recognition. This idea is originally implemented in the work done by Yamada et al. [43, 44, 8] for the recognition of one sound source. The authors proposed the 3-D Viterbi search algorithm, which performs simultaneous talker localization and speech recognition. In the evaluation system, a microphone is steered to every direction each time and speech is extracted. The speech extraction is followed by the matching between the input frames and the trained models. The advantage of this approach is that does not use any power information, it uses the speech information, and avoids the deterministic talker local-

ization. Moreover, it can be applied for the case of a moving talker. Our proposed algorithm evaluates also this idea, extended to deal with multiple sound sources. Figure 2.4 illustrates the idea of the integrated speech recognition and talker localization. By steering a beamformer to each direction, extracting the speech, and performing matching between the extracted speech and acoustics models, the most likely paths can be obtain. A path contains information about the uttered speech and the talker localization.

2.3 Summary

This chapter addresses the problems the recognition of distant-talking speech of multiple sources faces, and briefly describes several solutions. Factors that should be considered include the accurate sound source localization, the reverberation and the echo, the additive noise, the mismatch between training and target conditions, the signal estimation, and the voice activity detection.

Section 2.2 deals with the existing approaches for the recognition of distant talking speech of multiple sound sources. The conventional microphone array based- and the blind source separation-based methods are described. In this section is also described the integration of the talker localization and speech recognition which is evaluated in our proposed algorithm, too.

Chapter 3

Speech Recognition Issues

3.1 The Speech Recognition Task

The speech recognition is a very complex task and requires knowledge of different sciences. Signal processing, statistical analysis, information science and linguistics are some of the areas that are necessary for the investigation of speech recognition. The great number of problems that must be solved made the study of speech recognition more and more specialized and, therefore speech recognition can be classified into different branches, such as acoustic modeling, language modeling and decoding. However, the close cooperation of the three components is necessary in the building of a complete and efficient speech recognition system. The speech recognition process can be separated into two main stages. Namely, it consists of the training and the decoding stage.

- **Training**

  In the training stage, statistical approaches are used and the parameters of specific models are estimated using the data chosen for the training purpose. More specifically, recent speech recognition systems are based on the Hidden

Markov Models (HMM)[45, 46, 47, 48, 49, 50] or on the Neural-networks [51, 52]. Usually, the training is based on the Maximum Likelihood Estimation (MLE) [53, 54] or on the Maximum Mutual Information Estimation (MMIE) [55, 56]. In the training stage, the choice of the unit model is critical. Most of the HMM-based speech recognizers use sub-words as unit models. Earlier speech recognizers used context-independent monophone models. Recent speech recognizers consider also the context-dependency of each phoneme (biphones, triphones). However, the use of the context-dependent models faces with a serious problem. Namely, due to the large number of variations, it may happen that some context-dependent models do not appear in the training data. In order to solve this problem, efficient clustering algorithms had been reported [57] with the Decision Tree Clustering [58] to be an efficient solution to this problem. In some other approaches, parameters are tied for the more robust training.

The accurate estimation of the model parameters requires a large amount of training data. Moreover, the training should consider a large number of varieties, such as speakers, gender, etc. Since it is very difficult to collect data, which cover all the possible conditions, standard models are usually trained using sufficient amount of data and then the obtained models are adapted to specific conditions.

- **Decoding**

  In the decoding stage, a search strategy is applied which performs the matching between the trained models and the input unknown speech [59]. The search strategy incorporates acoustic and language knowledge in order to find to most likely word sequence. The search strategy can be Time Synchronous, or Time Asynchronous. The Viterbi search is a widely used

search algorithm.

In a speech recognition system for real use, the decoding speed is an important issue. The decoder has to deal with a huge number of candidates in order to find to most likely word sequence. In the literature, the considered candidates are known as search space. Although the more powerful and faster computers decrease significantly the time required for the decoding, most of the developers try to speed-up the speech recognizers by implementing efficient algorithms for fast decoding. A widely used approach is the multi-pass decoding [61, 62]. More specifically, in these cases the recognizer operates in multiple stages. In a first stage the so-called fast-match is performed [63], and resources that not require large amount of computation are used (monophone acoustic models, bigram language model). The fast-match eliminates the candidates to a small number, and those are in the following re-scored by using computationally more expensive resources (for example triphones, trigram language model).

In both training and decoding stage the speech is represented by specific parameters, with the Mel Frequency Cepstral Coefficients (MFCC) parameters to be the most popular.

A serious problem that usually the speech recognizers face is the mismatch between training and decoding conditions. More specifically, in most of the cases the statistical models are trained using clean speech of particular speakers. However, the input unknown speech that must be recognized is usually distorted due to the noise or reverberation. An effective method that used in order to solve this problem is the adaptation of the clean models to the environmental and speaker variations. Effective adaptation techniques, such as Maximum Likelihood Linear Regression (MLLR), or Maximum A

Posteriori (MAP) [64] were introduced.

Given an input acoustic observation sequence $\mathbf{O} = o_1, o_2, ..., o_n$, the mathematical formulation of the speech recognition problem can be given by the following equation:

$$\hat{\mathbf{W}} = \underset{W}{argmax}\, P_r(\mathbf{W}|\mathbf{O}) \tag{3.1}$$

Thus, the speech recognizer attempts to find that $\mathbf{W} = w_1, w_2, ..., w_m$ word sequence, which maximizes the probability. By using the Bayes' formula Eq. (3.1) can be written as follows:

$$P_r(\mathbf{W}|\mathbf{O}) = \frac{P(\mathbf{O}|\mathbf{W})P(\mathbf{W})}{P(\mathbf{O})} \tag{3.2}$$

The acoustic models give the $P(\mathbf{O}|\mathbf{W})$ probability and the language model gives the $P(\mathbf{W})$. The task of the search strategy is to find that word sequence which maximizes the $P(\mathbf{O}|\mathbf{W})P(\mathbf{W})$. Figure 3.1 shows the block diagram of a complete speech recognition system.

3.2 Hidden Markov Models(HMM)

A main problem in the speech recognition is that there are many uncertainties. Stochastic modeling is a flexible method for modeling such problems. Hidden Markov modeling is such a stochastic technique, which permits modeling with many of the classical probability distributions and is well suited to the incorporation of temporal information. The Hidden Markov model is a statistical model that uses a number of states and the associated state transitions to jointly model the temporal and spectral variations of speech. Since the HMM can characterize both the temporal and spectral varying nature of the speech, it has been used to model fundamental speech units. More specifically, in speech recognition

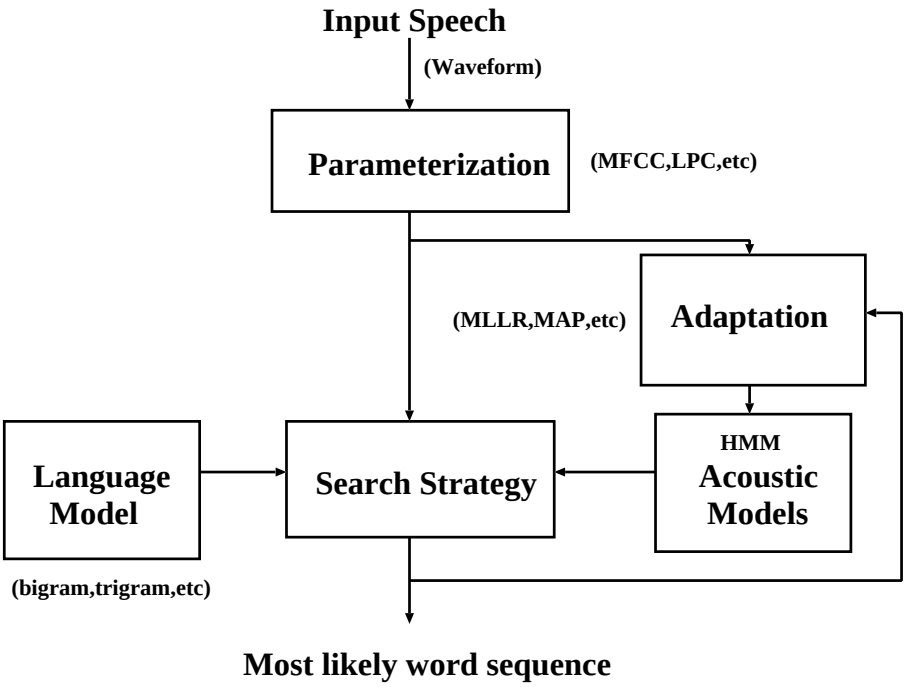

Figure 3.1. An HMM-based speech recognition system

the HMM model is a finite automata that changes state once every time frame, and each time t that a j state entered, a $\mathbf{x}_t$ speech vector is generated from the $b_j(\mathbf{x}_t)$ probability density. The transition from the state i to the state j is also probabilistic and it is given by the transition probability.

Recent HMM-based recognizers use many variants for the sub-word modeling. However, the most commonly used is a left-to-right without skip HMM. Figure 3.2 shows a HMM model with 3 emitting states. In this model only self and forward transitions are allowed.

In practice the state sequence is hidden and only the observation sequence $\mathbf{X}$ is known. The required likelihood is computed over all possible state sequences

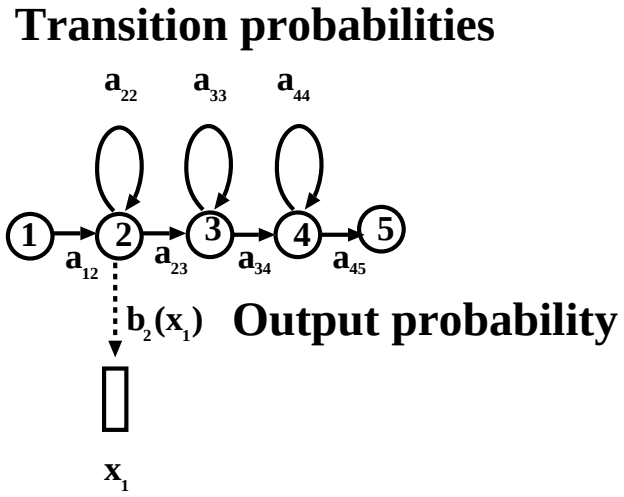

Figure 3.2. A 3-states Hidden Markov Model

$S = s(1), s(2), ..., s(T)$ according to the following equation:

$$P(\mathbf{X}|\mathbf{M}) = \sum_{s} \alpha_{s(0)s(1)} \prod_{t=1}^{T} b_{s(t)}(\mathbf{x}_t)\alpha_{s(t)s(t+1)} \tag{3.3}$$

where $s(0)$ is the model entry state and $s(T+1)$ the model exit state. The following equation gives an approximated way to the computation of the likelihood, when only the most likely state sequence is considered.

$$\hat{P}(\mathbf{X}|\mathbf{M}) = \underset{S}{max}\{\alpha_{s(0)s(1)} \prod_{t=1}^{T} b_{s(t)}(\mathbf{x}_t)\alpha_{s(t)s(t+1)}\} \tag{3.4}$$

3.3 The Viterbi Algorithm

The Viterbi algorithm is used for the decoding and it is based on the maximum likelihood. Namely, the Viterbi search attempts to find the most likely state sequence by searching a trellis space composed of states and input frames. Figure 3.3 describes the Viterbi algorithm.

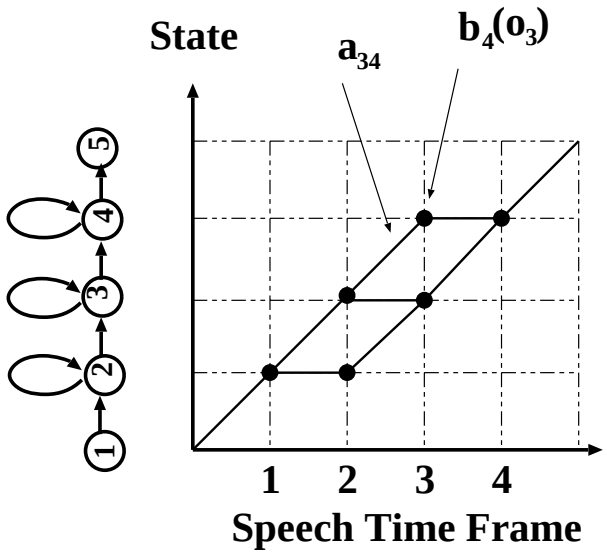

Figure 3.3. 2-D Trellis space and the Viterbi algorithm

The maximum likelihood can be given by the following recursive equation:

$$\Phi_j(t) = \max_i\{\Phi_i(t-1)a_{ij}b_j(\mathbf{x}_t)\} \tag{3.5}$$

In the equation (3.5) $\Phi_j(t)$ is the maximum likelihood of observing speech vectors $\mathbf{x}_1$ to $\mathbf{x}_t$ and being in state j at time t.

Considering log likelihoods the Eq.(3.5) can be written as

$$\alpha_j(t) = \max_i\{\alpha_i(t-1) + loga_{ij}\} + logb_j(\mathbf{x}_t) \tag{3.6}$$

Equation (3.6) is the so-called Viterbi formula and a modified version will be used in this thesis.

3.4 Summary

In this section the speech recognition issues are addressed. The speech recognition task can be classified into the training and the decoding stage. In the training

stage, a method is chosen for estimating the parameters of statistical models. Most of the speech recognizers are HMM- or Neural Network-based. The second stage of the recognition process is the decoding, where matching is performed between the input unknown speech and the trained acoustic models. Since our proposed system is HMM-based, in section 3.2 we describe the Hidden Markov Models and in section 3.3 we describe the Viterbi algorithm.

Chapter 4

Microphone Arrays

4.1 Beamforming

Sources, which produce propagating signals, can be expressed by the information contained in those signals. The signal waveform expresses the source nature and by using its temporal and spatial characteristics the source location can be determined. In the real world we should consider not only the one of interest source, but the presence of several sources, too. Therefore, the signal processing methods must focuses on selected signals, and moreover signals must be separated according to their directions and their frequency content, by applying spatiotemporal filtering. A flexible approach is the use of array, which acts as spatial filter attenuating all signals and saving those, propagated from a certain direction.

Our method is based on this approach and uses microphone array for separating the signals of the different speakers. A microphone array is composed of several microphones located linearly or non-linearly. A signal from a desired direction can be acquired by forming a directive pattern sensitive to the direction. The directive pattern can be steered electronically to a particular direction and the microphone array can suppress signals from the other directions saving those

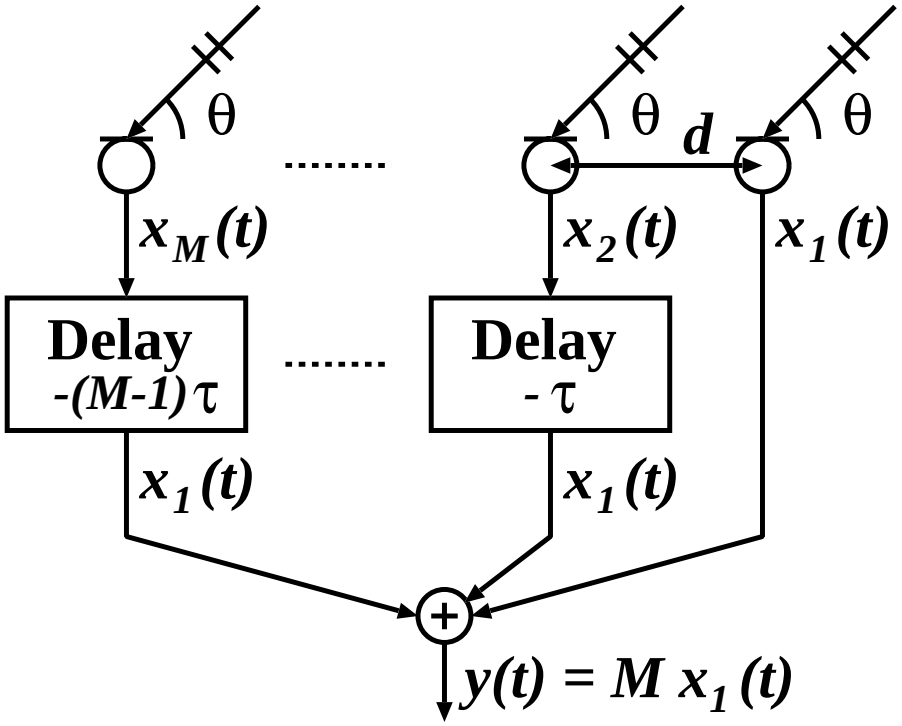

Figure 4.1. Delay-and-sum beamformer

ones of the interest.

Beamforming is the name given to array signal processing algorithms, which focuses the array-capture abilities in a particular direction. The oldest and simplest array processing algorithm is called Delay-and-Sum beamforming and still remains a powerful and widely used method. Figure 4.1 explains the idea the delay-and-sum beamforming is based on. The output signal of the beamformer is the summation of the outputs of each microphone after a delay has been applied to each one.

Let's assume a source that is located in the far-field and which is propagating a plane wave. The propagated signal is received by a microphone array, composed of M microphones, linearly located in d distance from each other. The signal

received by the i-th microphone is:

$$x_i(t) = x_1(t - (i-1)\tau) \tag{4.1}$$

where $\tau = dcos\theta/c$, θ is the direction of the arrival of the speech signal, and c is the sound propagation speed. The output $y(t)$ signal of the beamforming is given by the equation:

$$y(t) = \sum_{i=0}^{M-1} x_i(t + (i-1)\tau) \tag{4.2}$$

By adjusting the τ, the array can be steered to a desired direction. In the frequency domain the Eq. (4.2) can be written as follows:

$$x_i(f) = x_1(f)e^{j2\pi f(i-1)dcos\theta/c} \tag{4.3}$$

The spectrum of the signal received by the i-th microphone is represented by the Eq. (4.3). The spectrum of the output beamformed signal is given by the Eq. (4.4).

$$Y(f) = \sum_{i=1}^{M-1} x_m(f)e^{j2\pi f(i-1)dcos\theta/c} \tag{4.4}$$

Considering a signal propagated from direction θ, Eq. (4.5) represents the array gain in ϕ direction .

$$H(\phi, f) = \frac{|sin(M\pi fd(cos\phi - cos\theta)/c)|}{|sin(\pi fd(cos\phi - cos\theta)/c)|} \tag{4.5}$$

The main disadvantage of the delay-and-sum beamformer is the limited enhancement ability. However, a sharp beam requires a great number of microphones that increases the system complexity. In some other approaches an adaptive beamformer is used.

Several microphone array geometries are used. In the literature we can meet with Uniform Linear Arrays (ULA), Linear Random Arrays, Planar Arrays, etc. [3]. In this thesis, two uniform linear arrays composed of 16 and 32 microphones are used.

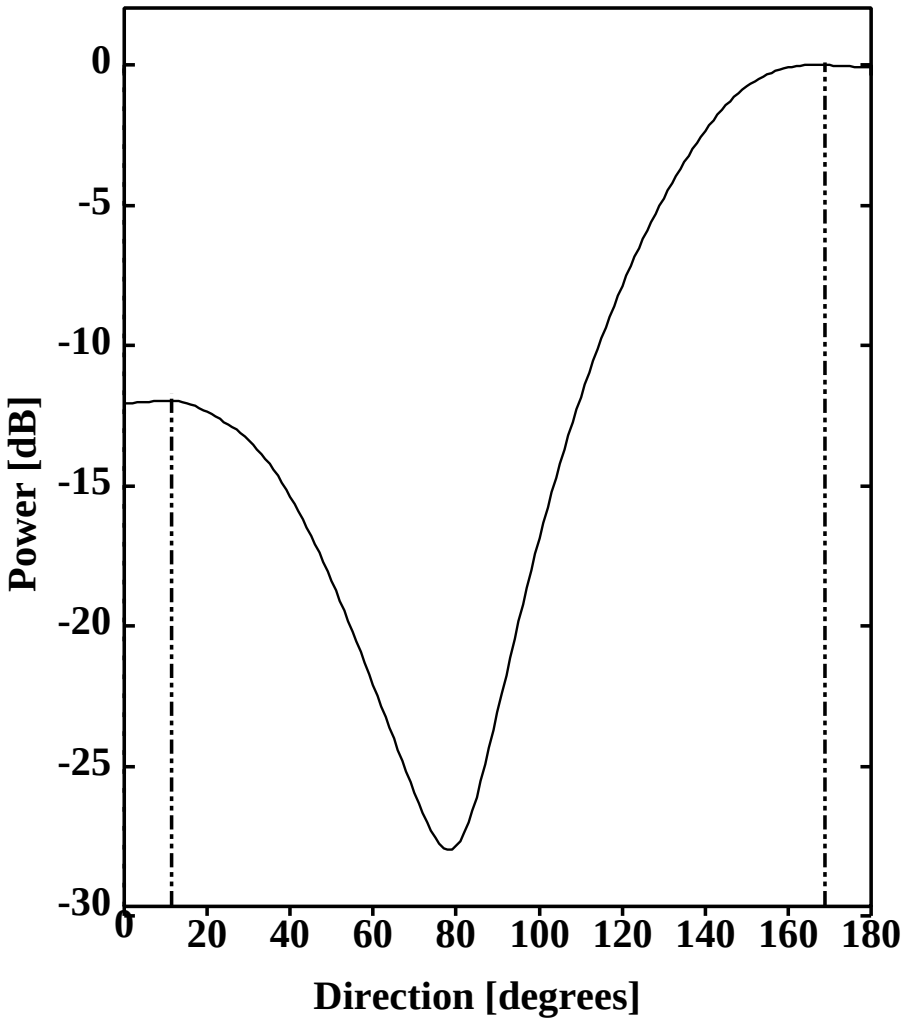

Figure 4.2. Talker localization using spatiotemporal analysis

4.2 Sound Source Localization

In this section, we briefly discuss two conventional localization methods based on the microphone array. Perhaps, the easiest method is the spatiotemporal analysis. In this method, the microphone array is steered to each direction and a spatial power spectrum is calculated. The DOA is found by searching for the

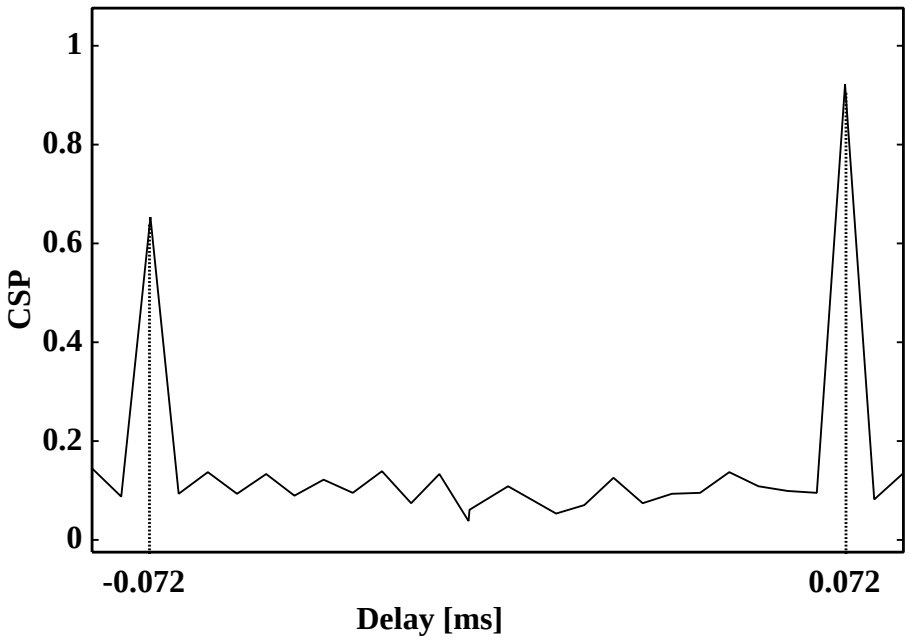

Figure 4.3. Talker localization using CSP method

peaks in the power spectrum. Figure 4.2 shows an example of a spatial power spectrum. In this experiments two signals are propagating from 10 and 170 degrees directions. In the figure, the two peaks at the two directions can be seen. The main disadvantage of this method is the difficulty to localize moving talker. Moreover, under low SNR conditions the use of power appears to be problematic.

The second method that we are addressing in this section, is the Cross-power Spectrum Phase Analysis (CSP) [23]. In this method, the direction of arrival of the speech signal is obtained by estimating the time delay. Formula 4.6 shows the so-called CSP function. The maximum of the CSP function gives the estimation of the DOA.

$$CSP(k) = \mathrm{DFT}^{-1}\left[\frac{\mathrm{DFT}\left[x_1(n)\right]\mathrm{DFT}\left[x_2(n)\right]^*}{\left|\mathrm{DFT}\left[x_1(n)\right]\right|\left|\mathrm{DFT}\left[x_2(n)\right]\right|}\right], \tag{4.6}$$

In the CSP function the signals received by two channels are used ($x_1(n)$ *and* $x_2(n)$). In the Formula 4.6, k and n is time index, DFT[.] is the Discrete Fourier Transform, and $*$ is the complex conjugate. Figure 4.3 shows the CSP function in the case of two sources located at 30 and 150 degrees, respectively. In this experiment a linear microphone array composed of 16 channels was used.

Although the CSP is an efficient localization method for fixed sound sources, it faces serious difficulties in localizing moving talkers. In this thesis, we carried out experiments to compare the CSP method with our proposed method. The results show higher performance obtained by our method.

4.3 Summary

The microphone array has an important role in the recognition of distant-talking speech. The microphone array can form a directive pattern sensitive to the direction, and a signal from a desired direction can be acquired with high quality. Moreover, the microphone array can be steered electronically to a particular direction. The simplest and oldest array processing algorithm is the delay-and-sum beamforming. In this chapter, we also describe two sound source localization methods, namely the spatiotemporal analysis and the Cross-power Spectrum Phase Analysis.

Chapter 5

Thesis Background - 3-D Viterbi Search Method

The 3-D Viterbi search method proposed by Yamada et al. [43, 8, 44], implements the idea of the integration of talker localization and speech recognition. More specifically, instead of the deterministic talker localization, that is followed by the speech recognition, the 3-D Viterbi search method based system uses a simultaneous talker localization and speech recognition. In the evaluation system, at each time frame a beamformer is steered to all directions and acoustic feature vectors are extracted from each direction. Then matching is performed between the extracted features vectors and the acoustic models.

In this approach, Viterbi search is performed in a 3-D trellis space (Fig. 5.1) composed of input frames, direction and HMM models, and the path with the highest likelihood is obtained. The path is a (q, d) $(state, direction)$, which corresponds to the uttered speech and talker locus. Therefore, the talker localization and speech recognition is performed simultaneously. The (q, d) with the highest

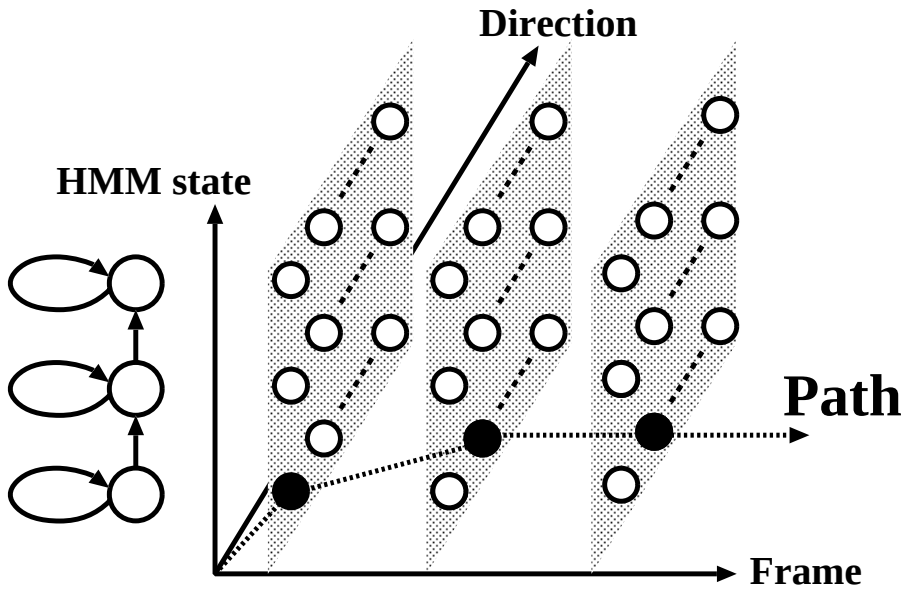

Figure 5.1. 3-D trellis space

likelihood can be obtained using the following Equation:

$$(q,d) = \underset{q',d'}{argmax}\, P(\mathbf{X}(d)|q',d',M) \tag{5.1}$$

In Eq. (5.1) M are the HMM models and $\mathbf{X}$ is the observation vector sequence. The likelihoods of each (q,d) at the t time frame can be computed using the following formula :

$$\alpha(q,d,t) = \underset{q',d'}{max}\{\alpha(q',d',t-1) + loga_1(q,q') + loga_2(d,d')\} + logb(\mathbf{X}(d,t)) \tag{5.2}$$

In Formula (5.2), a_1 is the state transition probability and a_2 is the direction transition probability. The state transition is provided by the acoustic models. However, the training of the direction transition probability, that indicates the movement of the talker, is very difficult and therefore a heuristic approach for its computation is used.

Table 5.1. Results obtained by using simulated data for talker located at fixed position.

	Clean	SNR 20 dB	SNR 10 dB
3-D Viterbi method - Initial evaluation	96.2	72.6	28.2
3-D Viterbi method with the weight function	96.2	94.9	88.4

Table 5.2. Results obtained by using simulated data for moving talker.

	Clean	SNR 20 dB	SNR 10 dB
3-D Viterbi method - Initial evaluation	96.2	74.5	26.8
3-D Viterbi method with the weight function	96.7	93.9	84.7

The performance of the 3-D Viterbi search method based speech recognition system was evaluated through experiments on simulated and real data. The obtained results showed that the method provides efficiently high performance, even in the case of a moving talker talker. Tables 5.1, 5.2, 5.3 and 5.4 show the obtained results as described in [65]. In the initial evaluation a delay-and-sum beamformer was used and the experiments were carried out both on simulated and real data. The introduction of a weight function, which increases the likelihood of the speech-likely directions resulted improvement in the performance. Finally, experiments were carried out by using adaptive beamforming. In that case, significant improvement could be obtained. The results are obtained from

Table 5.3. Results obtained by using real data for talker located at fixed position.

	21 dB	SNR 18 dB	SNR 10 dB
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

Table 5.4. Results obtained by using real data for moving talker.

	SNR 21 dB	SNR 18 dB	SNR 10 dB
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

experiments for speaker-dependent isolated word recognition.

The described results shows that the 3-D Viterbi search method and using adaptive beamforming provides high recognition rates, even in the case of the moving talker, too. However, a main disadvantage of this method is that it cannot deal with multiple sources. The reason is that because it considers only the most likely path in the 3-D trellis space. An additional problem of the method is the huge computation amount occurred by the steering of the beamformer to each direction.

In the following chapters we will explain our idea to solve the problem of the presence of multiple sound sources. Our proposed method is an extension of the

described 3-D Viterbi search method to the 3-D N-best search method.

5.1 Summary

This chapter describes the idea, the formulation, and some obtained results by using the 3-D Viterbi search method. This method integrates sound source localization and speech recognition process. The obtained results show that the 3-D Viterbi search-based recognition system performs efficiently, even in the case of a moving talker, too. The main disadvantage of this method is that can deal only with single sound source. Our proposed method attempts to solve this problem, by extending the 3-D Viterbi search to the 3-D N-best search.

Chapter 6

The proposed 3-D N-best Search Method

6.1 Idea and Formulation

The idea to solve the problem of the recognition of distant-talking speech of multiple sound sources is to introduce the N-best paradigm into the 3-D Viterbi search method. More specifically, instead of keeping only the hypothesis with the highest likelihood, we keep the N-best hypotheses, and in this way we can consider speech signals arriving from multiple directions. Our N-best approach [20, 22, 66, 67, 68, 26, 69] is different from the conventional ones in the sense that we keep N-best hypotheses for each direction. Our algorithm is one-pass search strategy, which performs full search in all directions and keeps N-best for word and direction hypothesis.

In a similar way to 3-D Viterbi search method based system, our system is also microphone array based. At each time frame a beamformer is steered to each direction and feature vectors are extracted. The N-best hypotheses are found by

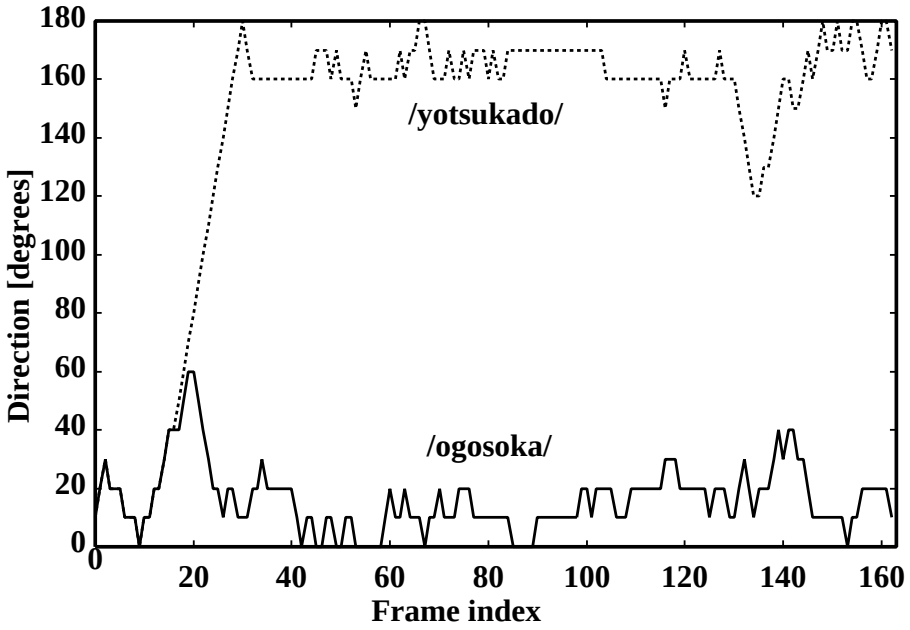

Figure 6.1. Direction sequences of the hypotheses /ogosoka/ and /yotsukado/

matching between the feature vectors and the HMM models, and by keeping the hypotheses with the highest likelihoods. The obtained N-best list includes hypotheses from different directions, and therefore multiple sound sources can be recognized simultaneously. The phoneme sequences of a hypothesis corresponds to the uttered speech and the direction sequence to the talker locus. Tables 6.1 and 6.2 show the obtained N-best lists. Figure 6.1 and Figure 6.2 show the obtained direction sequences. In this experiment two sound sources located at fixed position at 10 and 170 degrees, pronounces in the first case the Japanese words */ogosoka/* and */yotsukado/* and in the other case the Japanese words */yuumoa/* and */omowazu/*. As can be seen both words are included in the N-best list and the direction sequences follow the correct talkers locations. Table 6.3 shows the N-best list in the case of the pronounces words */ikioi/* and */kakurepyuritan/*

Table 6.1. Top5 results. Both sound sources are included in the list

Input	MHT /ogosoka/	FTK /yotsukado/

Best	Word	Likelihood
1	/ogosoka/	-77.0445
2	/yotsukado/	-77.3181
3	/monosugoi/	-77.3501
4	/naosara/	-77.4224

and Fig. 6.3 the obtained direction sequences of the two hypotheses. In this case only the word with the longer duration is included in the N-best lists and the other one doesn't appear. Although we haven't investigated this problem in details, a possible reason for this is that the algorithm attempts to search the directions where the signal lasts longer than the other directions. The direction sequences justify this observation. A possible solution to this problem is to implement additional techniques, similar to those used in word-spotting in order to terminate the search in one direction if for a number of frames silence is observed, and extract the hypotheses of that direction.

The N-best hypotheses are chosen based on the Formula (6.1). Considering a (q, d) $(state, direction)$ node at t time frame, the $\alpha^N(q, d, t)$ N-best hypotheses can be obtained by adding the a_1 state and a_2 direction transition probability, as well, to the $\alpha^N(q, d, t-1)$ arriving from all connected nodes hypotheses, then sorting the overall hypotheses and, finally by adding the $b(\mathbf{X}(d, t))$ output probability of the appropriate d direction.

$$\alpha^N(q, d, t) = \underset{q', d'}{sort}\{\alpha^N(q', d', t-1) + loga_1(q, q') + loga_2(d, d')\} + logb(\mathbf{X}(d, t)) \quad (6.1)$$

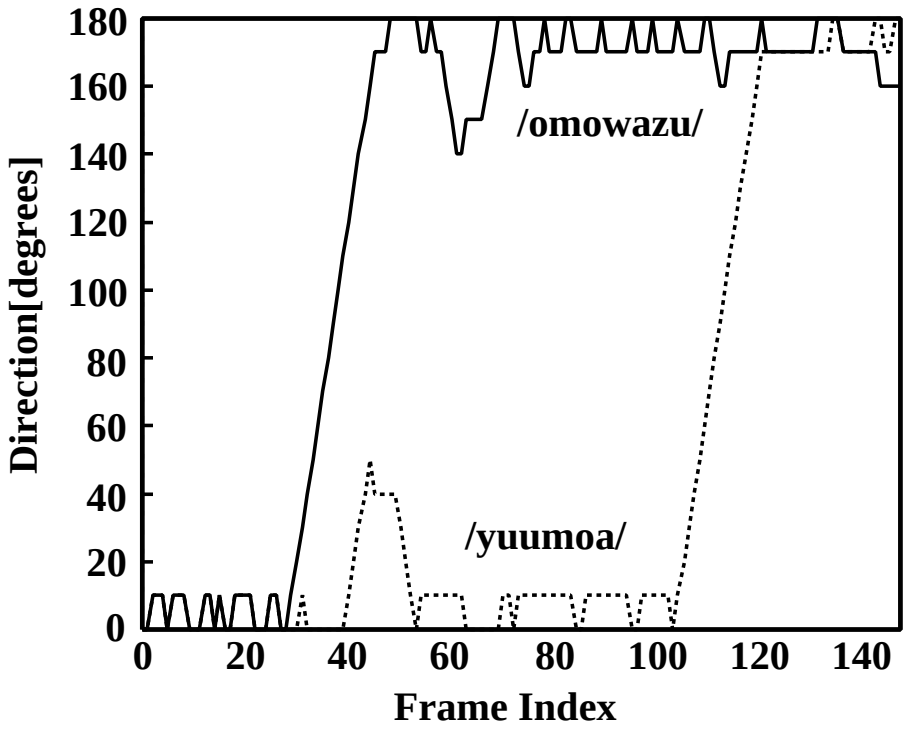

Figure 6.2. Direction sequences of the hypotheses /yuumoa/ and /omowazu/

The a_1 state transition probability is provided by the transition matrix of the acoustic models. However, for the a_2 direction transition probability a heuristic approach is used. More specifically the a_2 can be computed as follows:

$$a_2(d', d) = \begin{cases} \dfrac{1}{2\Delta d} & , \quad |d - d'| \leq \Delta d \\ 0 & , \quad |d - d'| > \Delta d \end{cases} , \tag{6.2}$$

where Δd is the range of the talker movements. In this thesis the Δd is 10 degrees.

Table 6.2. Top5 results. Both sound sources are included in the list

Input	MHT /yuumoa/	FTK /omowazu/

Best	Word	Likelihood
1	omowazu	-75.5039
2	imagoro	-75.5795
3	yuumoa	-75.8902
4	uyamau	-75.9920
5	ikioi	-75.9998

6.2 The Proposed Path Distance-based Clustering Technique

Table 6.4 shows the N-best list obtained in an experiment for the recognition of two talkers located at fixed position, at 10 and 170 degrees, respectively. The 'Speaker A' pronounces the Japanese word */omoshiroi/* and the 'Speaker B' the Japanese word */wagamama/*. Figure 6.4 shows the direction sequences of the Top N hypotheses. The first observation is that only one talker appears in the higher orders of the N-best list, and the other one appears in low order. The second observation is that the hypotheses appeared in the higher order originated from the 170 degrees direction. This is exactly a serious problem that our baseline system faced. More specifically, if in one direction the likelihood is much higher than the other directions then the N-best list is occupied by the hypotheses of that direction.

In order to solve this problem, we introduce our proposed path distance-

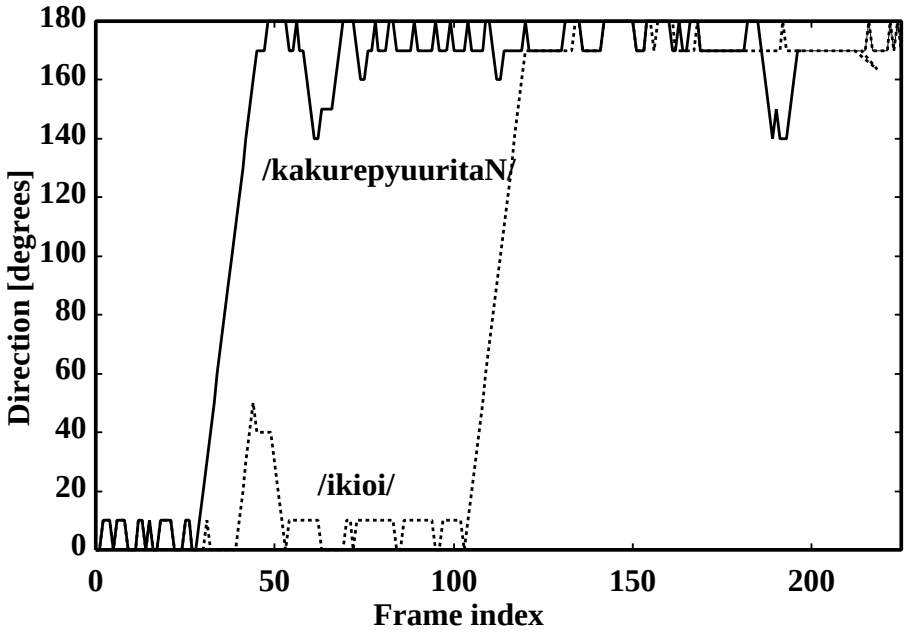

Figure 6.3. Direction sequences of the hypotheses /ikioi/ and /kakurepyuuritan/

based clustering technique [66, 67]. The main idea is to separate the hypotheses according to the direction. The clustering technique provides multiple N-best lists, which correspond to the multiple talkers. The results are obtained by picking up the Top N from each cluster. The direction of the multiple talkers can be obtained by examining the direction sequence of the Top 1 of each cluster.

Figure 6.5 shows the direction and power sequences of two hypotheses. Based on this information, we calculate an Euclidean distance weighted with the power. We name this distance **Path Distance**, and this is the measure our clustering technique is based on. At the last frame of the input speech signal, for each hypothesis-pair the path distance is computed and all the hypotheses are clustered into a pre-defined number of clusters. In this thesis, we assume that the number of sound sources is known and that a cluster corresponds to each sound source.

Table 6.3. Top5 Results. Only one source is included in the list.

Input	MHT /ikioi/	FTK /kakurepyuuritaN/

Best	Word	Likelihood
1	kakurepyuuritaN	-79.2763
2	hiQkurikaesu	-80.3220
3	oiharau	-80.3852
4	akegata	-80.4967
5	atarimae	-80.5042

Due to its simplicity, a bottom-up clustering technique was chosen. The Path Distance is computed using the following formula:

$$D(k,k') = \sum_{t=0}^{T-1} \{(d_k(t) - d_{k'}(t))^2 (p(d_k(t),t) + p(d_{k'}(t),t))\} \tag{6.3}$$

where k and k' are the ending directions, t is the time frame, T is the total frame number, $d_k(t)$ is the direction value of the hypothesis ending in k at time t, and $p(d_k(t))$ is the power sequence of this hypothesis. The reason that we introduce the power as weight in the path distance's computation is because our algorithm cannot guarantee the correct direction in the silence region. Therefore, by introducing the power we reduce the importance of the silence region in the computation of path distance. In the Figure 6.5 we can observe that in the speech region the two hypotheses can be separated based on their direction sequences. However, in the silence region the two hypotheses appear to be originated from the same direction.

Table 6.5 shows results described in Table 6.4 after the clustering is performed. As can be seen, the two pronounced words are included in different

Table 6.4. Top 4 results. Sorting according to the likelihood. Only one sound source was included in the N-best list

Input	Speaker A /omoshiroi/	Speaker B /wagamama/

Top	Word	Likelihood
1	**/wagamama/**	-78.5579
2	/hanahada/	-78.9105
3	/hanabanashii/	-78.9776
4	/wazawaza/	-79.2003
..	..	..
7	**/omoshiroi/**	-79.5485

cluster and in the first order.

6.3 The Proposed Likelihood Normalization Techniques

The N-best hypotheses of a (q, d) $(state, direction)$ are found by sorting the overall arriving hypotheses and choosing the top N. However, hypotheses arriving from different directions correspond to different sound sources with different likelihood dynamic ranges. Therefore, the comparison of the hypotheses according to their likelihoods can not be accurate. In order to avoid this problem we introduce a technique for likelihood normalization [68].

The technique used for likelihood normalization is similar to the method proposed by Matsui T. et al. [70]. That method was used for speaker recognition,

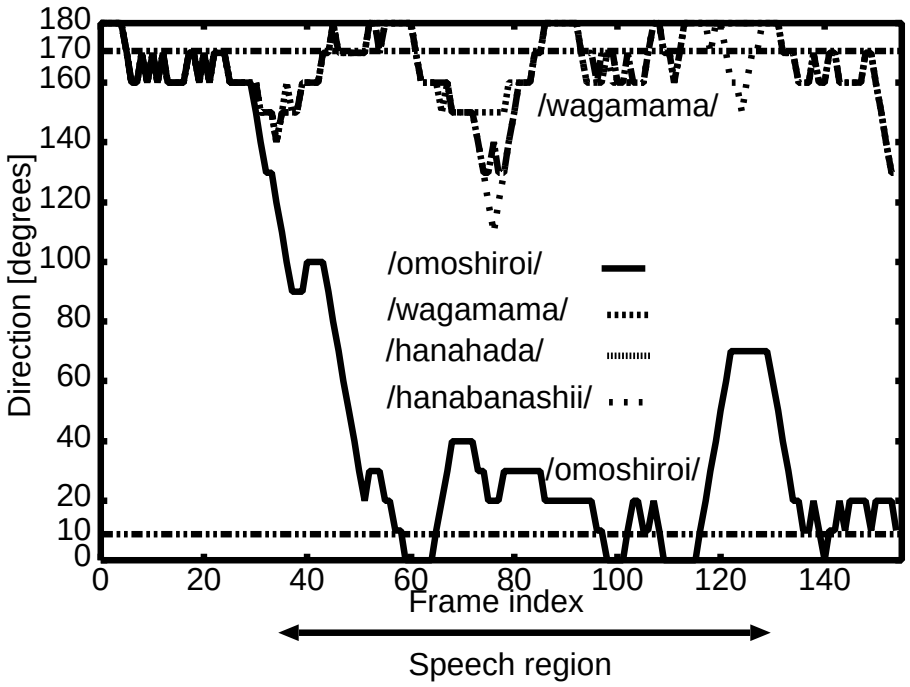

Figure 6.4. Direction sequences of Top N hypotheses

but it was found that it can also be efficiently applied for our task. Our one-state Gaussian mixture (GM) (1 state, 64 mixtures) model is close to that proposed by Matsui T. et al., but its objective is different. More specifically, this model runs in parallel with the other models and its accumulated likelihood is used to normalize the likelihoods of the hypotheses involved. Two different techniques are implemented and compared. The two likelihood normalization techniques, $L1$ and $L2$ are the following :

- **L1 Likelihood Normalization Technique**

  In the first approach we normalize the likelihoods only at the last frame. The actual likelihoods $\alpha(q, d, T)$ of every state q and direction d are normalized at the last frame T by dividing with the likelihood $\alpha_G(d, T)$ of the

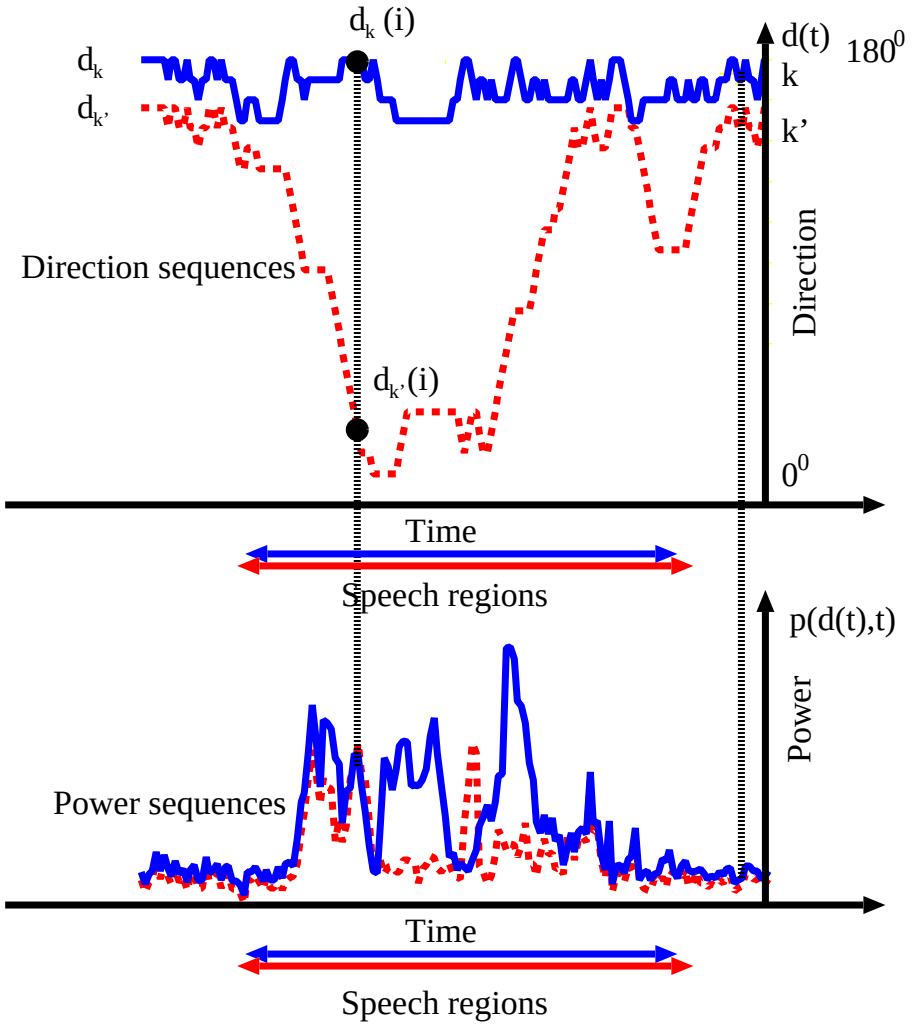

Figure 6.5. Direction and power sequences of the hypothesis-pair

one-state model. Considering logarithmic likelihoods, Eq. (6.4) gives the normalized likelihood $\Lambda(d, q)$.

$$\Lambda(q, d) = \alpha(q, d, T) - \alpha_G(d, T) \tag{6.4}$$

Table 6.5. Top 4 results. The hypotheses are classified using the path distance.

Input	Speaker A /omoshiroi/	Speaker B /wagamama/

Top	1st Cluster	2nd Cluster
1	**/omoshiroi/**	**/wagamama/**
2	-	/hanahada/
3	-	/hanabanashii
4	-	/wazawaza/

- **L2 Likelihood Normalization Technique**

  In this approach, the actual accumulated likelihoods $\alpha(q, d, t)$ of every state q and direction d are normalized at each time frame t by dividing them with the accumulated likelihood $\alpha_G(d, t)$ of the one-state model. Considering logarithmic likelihoods, Eq. (6.5) gives the normalized likelihood $\Lambda(d, q, t_f)$ at time t_f.

$$\Lambda(d, q, t_f) = \alpha(q, d, t_f) - \alpha_G(d, t_f) \tag{6.5}$$

Figure 6.6 shows the results of the comparison of the two techniques. In this experiments two talkers are used located at fixed position at 10 and 170 degrees. A linear microphone array composed of 16 microphones is used. The distance between the microphones is 2.83 cm. Results show that the likelihood normalization in every frame provides higher performance. Therefore, we finally choose this method for likelihood normalization.

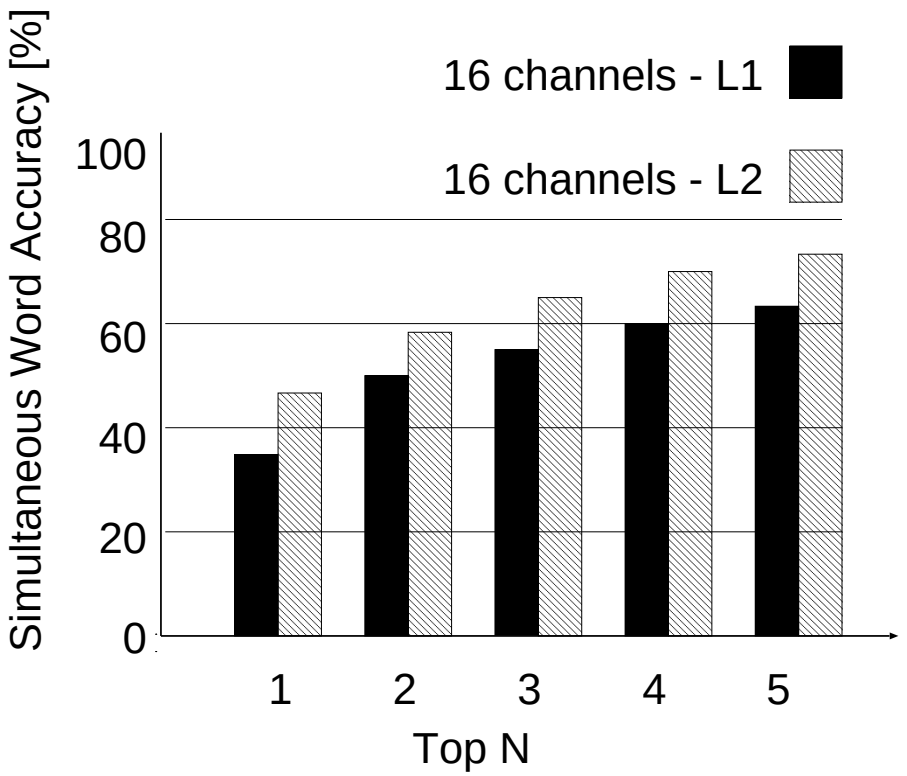

Figure 6.6. Comparison of the two implemented likelihood normalization techniques

6.4 Summary

This chapter describes our proposed 3-D N-best search algorithm able to recognize distant-talking speech of multiple sound sources. Our method is one-pass search strategy, which considers multiple hypotheses for each direction and word hypothesis. In our system, a microphone array is steered to each direction in each time frame, feature vectors are extracted, and matching is performed between the acoustic models and the input feature vectors. The baseline system integrates the 3-D Viterbi search method and the N-best paradigm into a com-

plete system. However, the simple integration is not sufficient for our purpose and, therefore we further improved the system by developing and implementing a clustering and a likelihood normalization technique.

Chapter 7

Evaluation of the 3-D N-best Search Based Speech Recognition System

The performance of our speech recognition system based on the 3-D N-best search was evaluated through experiments carried out for the recognition of the speech of two and three talkers.

7.1 Experiments Using Data With Time Delay

In these experiments simulated clean data (only time delay) are used for the simultaneous recognition of the speech of two sound sources.

7.1.1 Experimental Conditions

The two sound sources are located at fixed position at 10 and 170 degrees and pronounce a different word. Several speaker- and word-pairs are used. Namely,

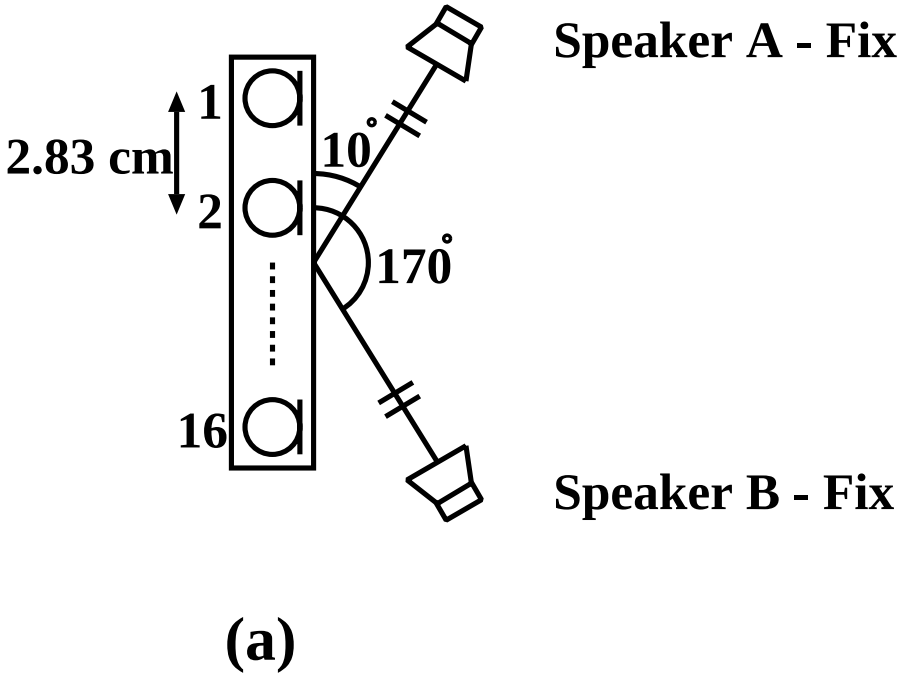

Figure 7.1. Position of sound sources

we use as test data 216 phoneme words from four talkers and we form a large number of variations. In total we use 2150 test word-pairs. Two linear microphone array composed of 16 and 32 microphones are used. The distance between the microphones is 2.83 cm. Figure 7.1 shows the experimental arrangement and Table 7.1 the specifications of the system.

7.1.2 Results Using a 16-channels Microphone Array

Three kind of results are given by our experiments. More specifically, the Word Accuracy (WA) of each speaker separately and the Simultaneous Word Accuracy is also, described. The results are obtained by examining the Top N of the two provided clusters. The two accuracies are defined as follows:

Table 7.1. System specification

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	ASJ continuous speech database (64 Speakers)
Test data	216 phonetically-balanced words of ATR Set A database

- Word Accuracy (WA)

$$WA = \frac{Word\ Correct\ \&\ Cluster\ Correct}{Total\ Test\ Words} \times 100\ [\%] \tag{7.1}$$

- Simultaneous Word Accuracy (SWA)

$$SWA = \frac{Both\ Words\ Correct\ \&\ Both\ Clusters\ Correct}{Total\ Test\ Words} \times 100\ [\%] \tag{7.2}$$

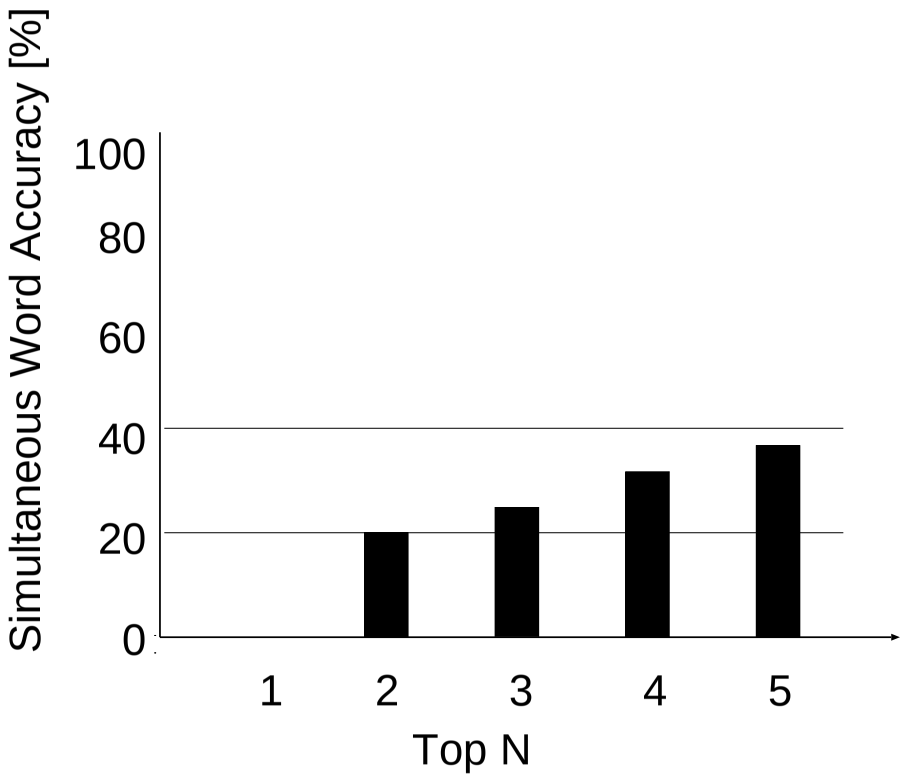

Figure 7.2. Results of the initial experiments

Figure 7.2 shows the Simultaneous Word Accuracy obtained by the initial experiments. In these experiments, neither the clustering and likelihood normalization techniques are not implemented. As can be seen, the achieved results are poor.

The results are further improved by implementing the clustering technique. Figures 7.3, 7.4 and 7.5 show the achieved WA and SWA. Figure 7.6 illustrates the comparison of the cases when there is no clustering and when clustering is implemented. As can be seen, by implementing clustering technique improvements in word accuracies can be obtained. However, the results are not sufficiently high.

The performance of our speech recognition system was significantly increased by implementing the likelihood normalization technique. In these experiments, the

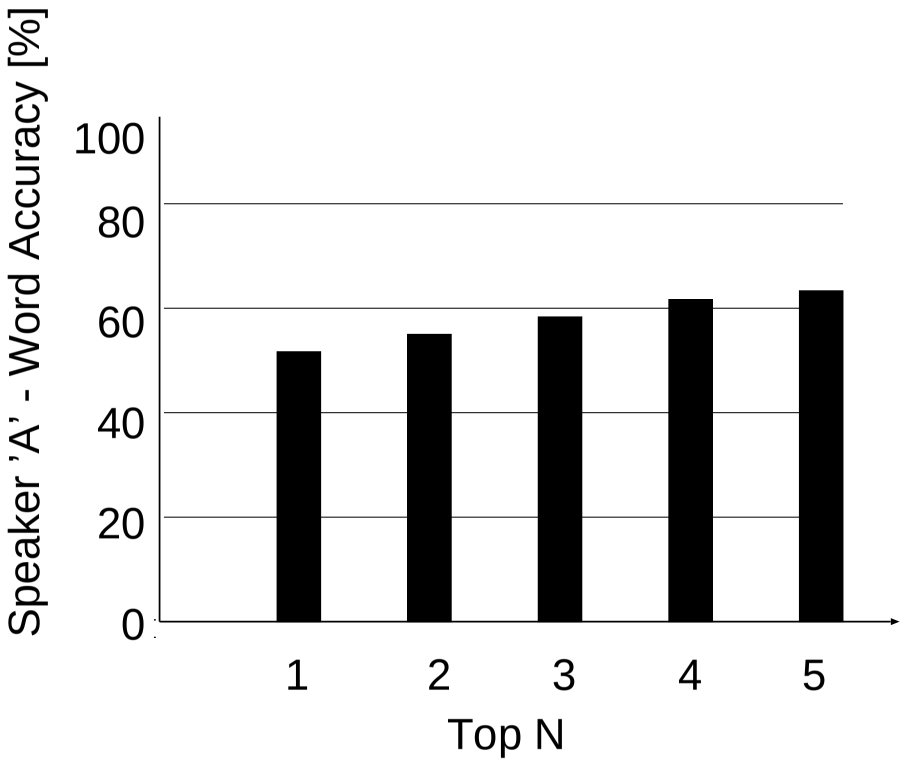

Figure 7.3. Speaker 'A' Word Accuracy - Clustering technique is implemented

likelihoods are normalized each time frame, and the N-best are chosen using those normalized likelihoods. Figure 7.7, 7.8 and 7.9 show the achieved results when both likelihood and clustering techniques are implemented. As results show, the improvements in recognition rates drastically increased.

Considering the initial evaluation, where only the 3-D Viterbi search and the N-best paradigm are integrated into a complete system, easily we can realize the effectiveness of the two additional developed techniques. The evaluation of the baseline system provides Simultaneous Word Accuracy between 25 % to 38 %. After developing and implementing the Path Distance-based clustering technique, our system shows 27 % relative improvement in the Simultaneous Word Accuracy.

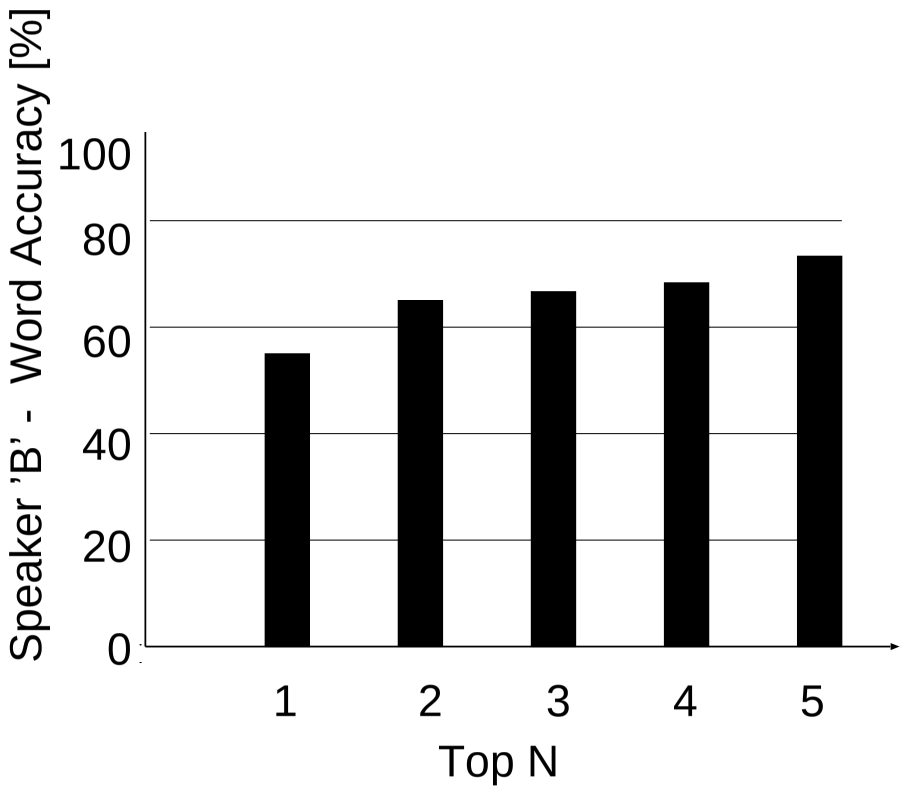

Figure 7.4. Speaker 'B' Word Accuracy - Clustering technique is implemented

The performance of our system is further increased by developing and implementing the likelihood normalization technique. More specifically, comparing to the initial system the relative improvement in the Simultaneous Word Accuracy is 62 %. The obtained results justify the necessarily of implementing additional ideas and techniques into the baseline system. The results obtained when both techniques are implemented are very promising.

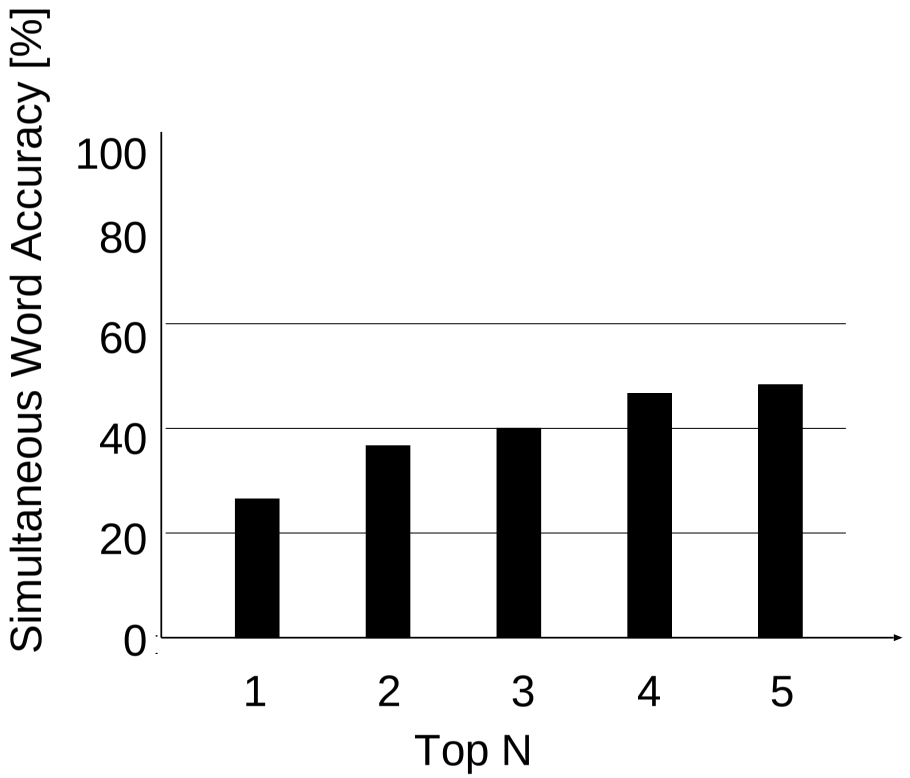

Figure 7.5. Simultaneous Word Accuracy - Clustering technique is implemented

7.1.3 Comparison between 3-D N-best search and a CSP-based method

In these experiments our proposed 3-D N-best method based system is compared with a conventional talker localization-based method. As conventional method the CSP method is chosen. The CSP-based speech recognition system operates in two stages. In the first stage the talker is localized using the CSP method. Then, a beamformer is steered to each hypothesized direction, featured vectors are extracted and 2-D Viterbi search is performed. In the case of our 3-D N-best search simultaneous talker localization and speech recognition is performed.

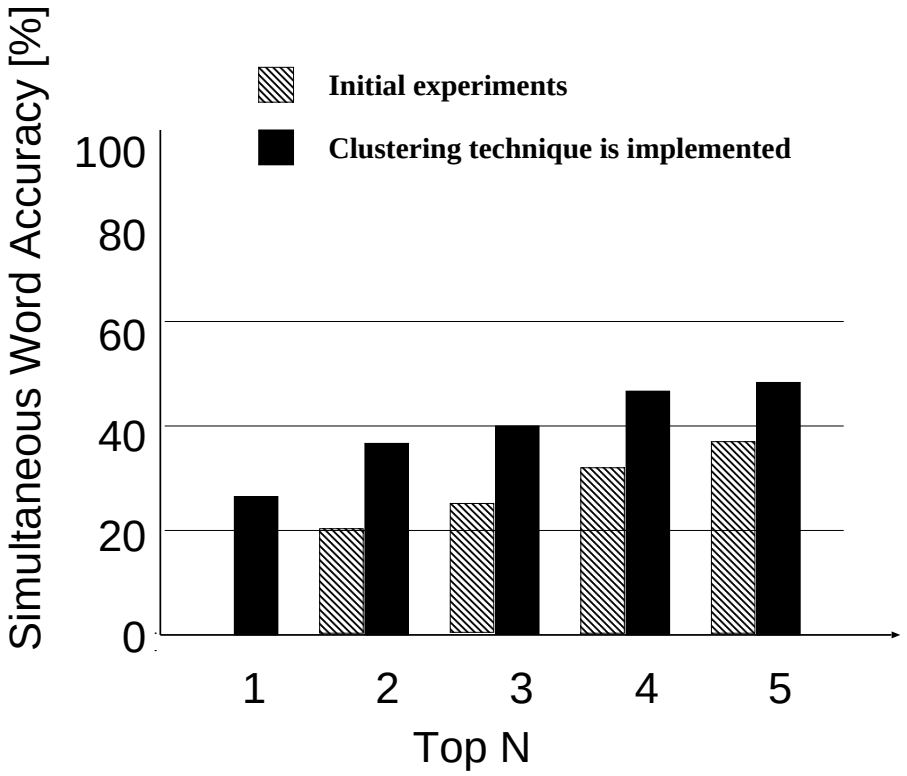

Figure 7.6. Improvement by implementing clustering technique

Table 7.2 shows the obtained results. As can be seen, in the case of the 'Speaker A' the CSP-based method operates slightly better. However, in the case of the 'Speaker B' the performance of our system is significantly higher. Therefore, the SWA provided by our system is also higher.

7.1.4 Results Including the Recognition of a Moving Talker

In order to evaluate the performance of our system in the case of a moving talker, we carry out experiment for the simultaneous speech recognition of a moving and a fixed talker. The fixed talker is located at 10 degrees and the other one moves

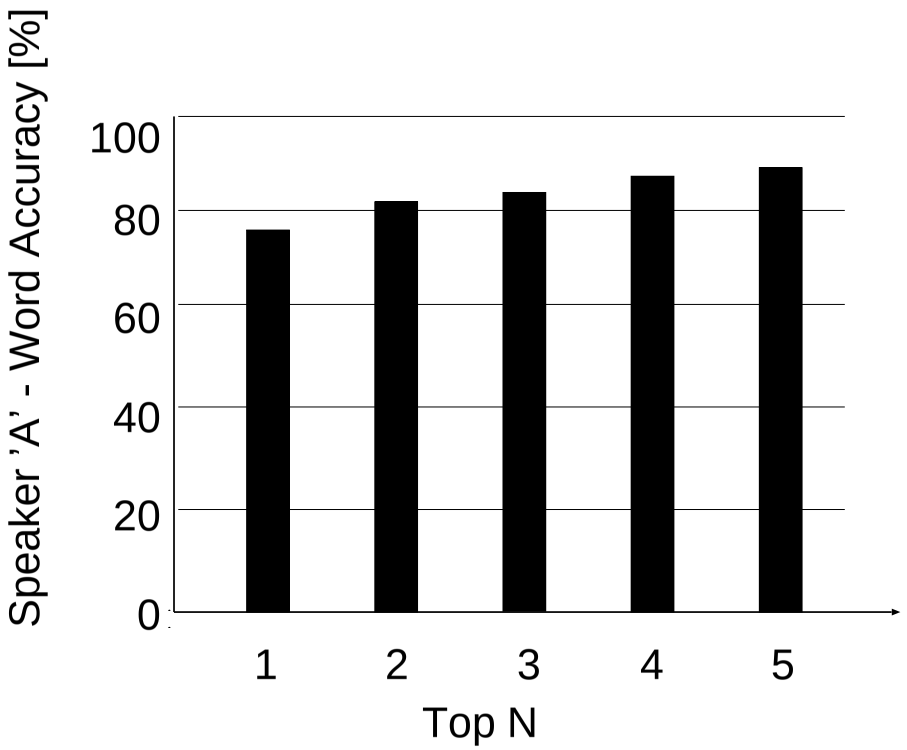

Figure 7.7. Speaker 'A' Word Accuracy - Both Clustering and likelihood normalization technique is implemented

from 0 to 180 degrees while uttering a word (Fig. 7.10). The situation we choose for this experiment is a very general case, when the two talkers are crossing each other. This fact makes our task very difficult. However, results shows that our method can be applied for the recognition of distant-talking speech of a moving talker,too. Namely, in the case of the moving talker for Top 5 the Word Accuracy is 72.01 %.

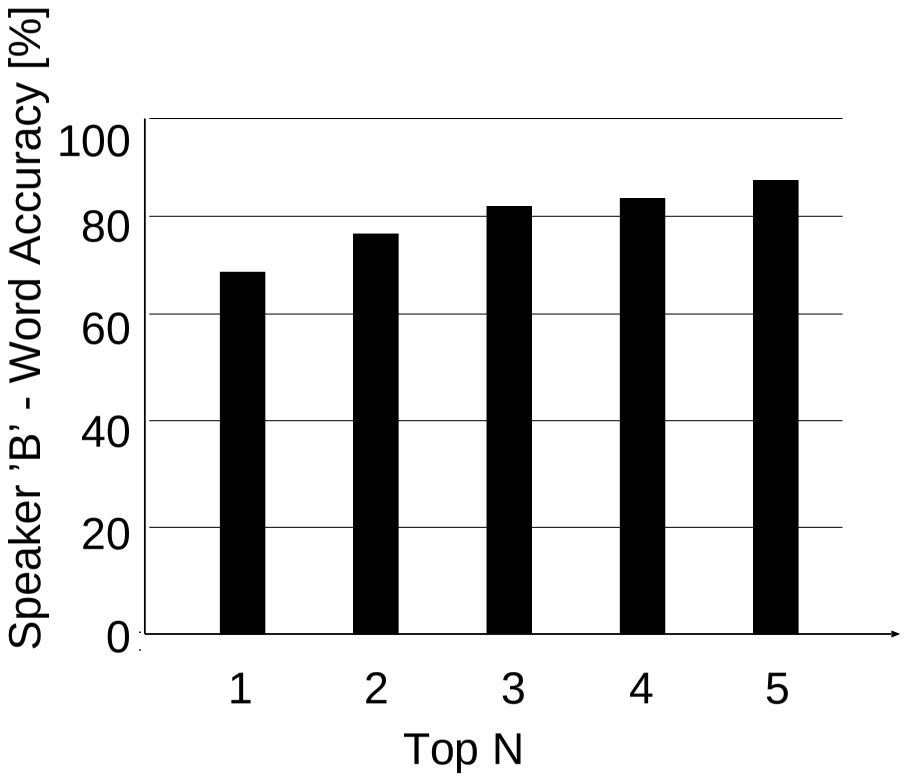

Figure 7.8. Speaker 'B' Word Accuracy - Both clustering and likelihood normalization technique is implemented

7.1.5 Results Using a 32-channels Microphone Array

In the previous experiments, a microphone array composed of 16 microphones is used. In these experiments a 32-channels linear microphone array is used. Figures 7.11, 7.12 and 7.13 show the obtained results. The obtained results show significant improvement by increasing the number of channels. More specifically in the case of the 32-channels microphone array the Top 1 results for Simultaneous Word Accuracy is higher than 72 %, which is very promising result.

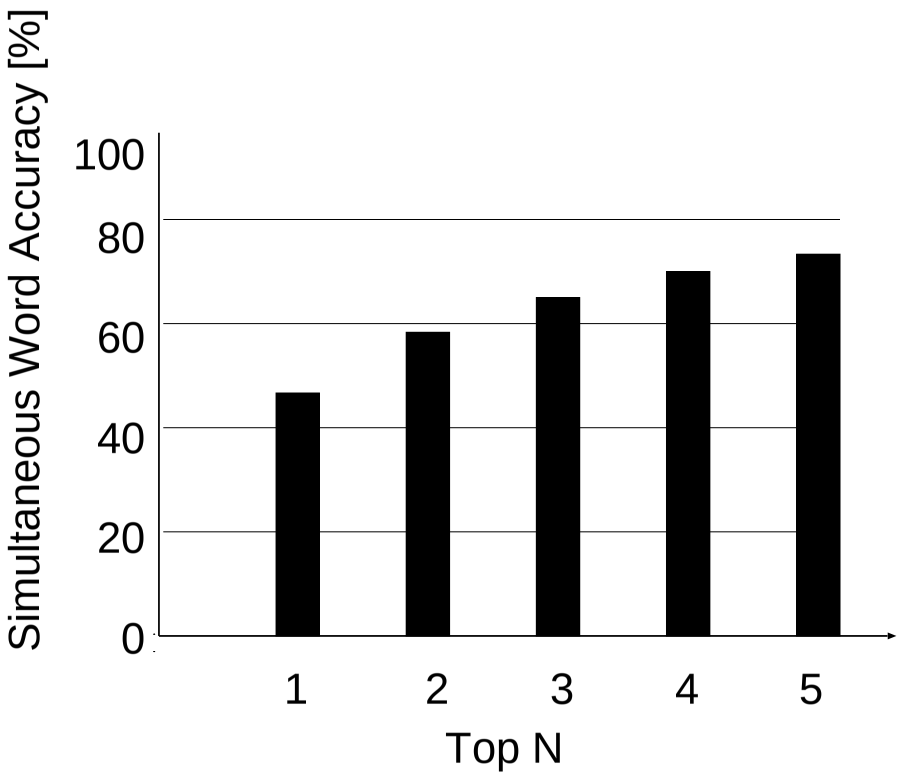

Figure 7.9. Simultaneous Word Accuracy - Both Clustering and likelihood normalization technique is implemented

7.1.6 Localization Error

In order to explain the difference in the performance of our system using 16-channels and 32-channels microphone array we introduce the so-called Localization error defined as follows:

$$\Delta(t) = |d(t) - d^{'}(t)|\{p(d(t), t) + p(d^{'}(t), t)\} \tag{7.3}$$

where t is the time frame, $\Delta(t)$ is the localization error, $d(t)$ is the correct direction, $d^{'}(t)$ is the hypothesized direction, and $p(d, t)$ is the corresponding power. We introduce the power in order to reduce the importance of the localization er-

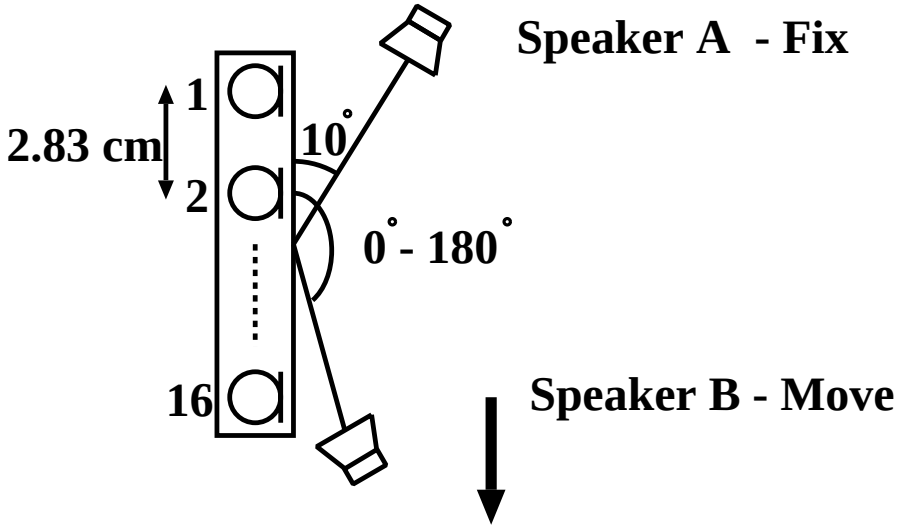

Figure 7.10. Position of sound sources including a moving talker

ror in the silence region, where the hypothesized direction is not accurate. Figure 7.14 shows the histogram of the localization errors in both cases of microphone arrays. As can be seen, in the case of the 32-channels microphone array the talker is localized with much lower errors. Therefore, the word accuracy is higher. The explanation of this fact can be given by examining the *Directive Patterns* (Fig. 7.15). In the case of the 32 microphones, the delay-and-sum beamformer forms sharper beam and therefore the localization accuracy is higher. As conclusion, in our experiments the 16 microphones can not form efficiently sharp beam.

Table 7.2. Comparison between the 3-D N-best search method-based and a CSP-based system

Localization Method	Search Method	#Source	Directions		WA [%]		SWA [%]
			A	B	A	B	
3-D N-best Search		2	Fix-30	Fix-150	90.09	84.19	73.68
CSP	2-D Viterbi	2	Fix-30	Fix-150	91.63	69.77	63.72

7.2 Experiments Using Simulated Reverberated Data

In these experiments we use the *Image Method* [71] to simulate reverberated data, and evaluate the performance of our 3-D N-best search based system in real environments. The impulse responses provided by the image method are convoluted with the clean speech in order to obtain the reverberated speech.

7.2.1 Experimental Conditions

Figure 7.16 describes the arrangement of the experiments. Two sound sources are used located at fixed position, at 10 and 170 degrees. A linear microphone array composed of 32 microphones is used. The distance between the microphones is 2.83 cm. The distance between the sound sources and the microphone array is 2 m. The (x,y,z) dimension of the room is (5.8 m, 4.3 m, 2.7 m). The microphone array is located in the center of the room in 1 m distance from the wall.

The reverberation time is measured by using square integration. Based on impulse response, the reverberation attenuation curve can be obtained. The

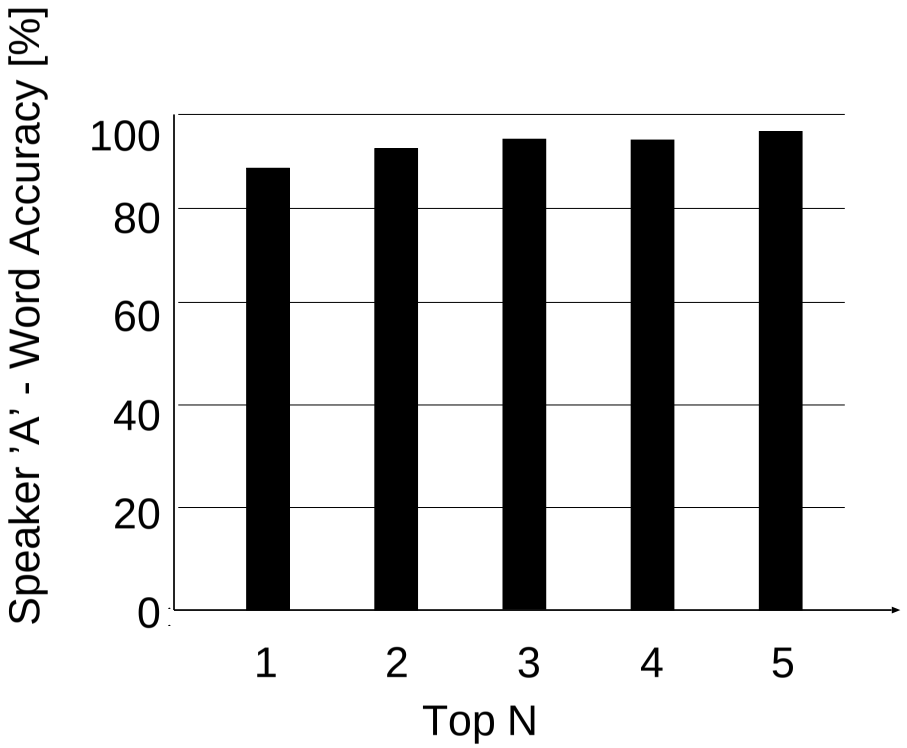

Figure 7.11. Speaker 'A' Word Accuracy - Microphone array composed of 32 channels

reverberation time is considered the time when the attenuation becomes 60 dB. Figure 7.17 illustrates the method used for the measurement of the reverberation time based of the impulse response.

7.2.2 Results

We carry out experiments on different reverberation times. Namely, we carry out speaker-independent isolated word recognition experiments for the cases of 162, 200 and 240 ms reverberation time. Figures 7.18, 7.19 and 7.20 show the achieved results.

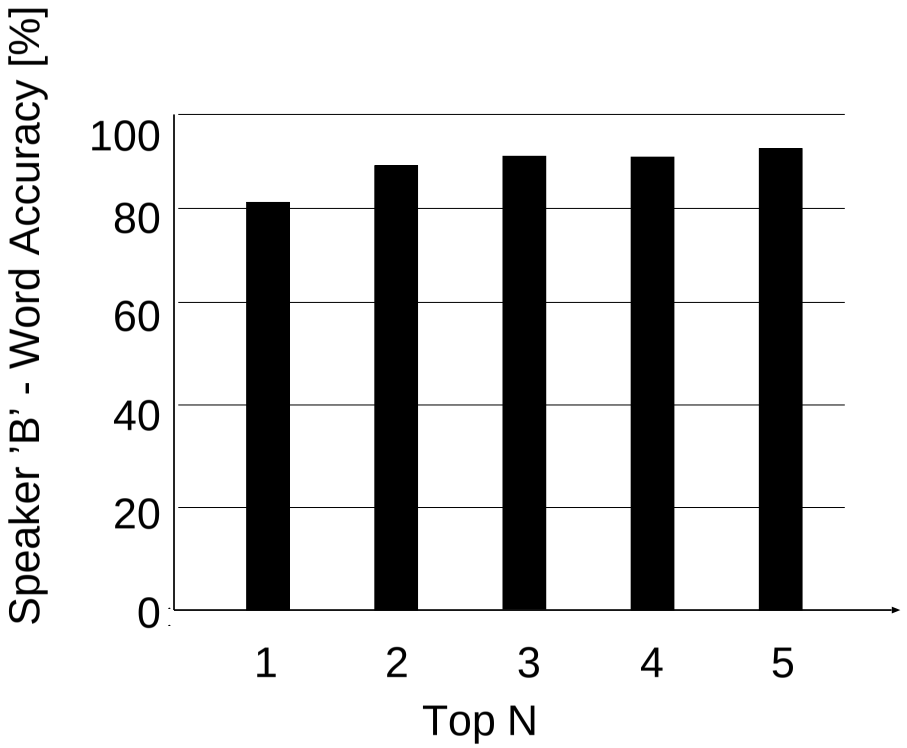

Figure 7.12. Speaker 'B' Word Accuracy - Microphone array composed of 32 channels

The obtained results show the impact effect of the reverberation in the speech recognition. The performance of our system under reverberant environment is decreased. The relative decrease in Simultaneous Word Accuracy is 27 %, 37.5 %, and 52 % for the case of 162, 200 and 240 ms reverberation time.

The reverberation belongs to the difficult problems that the hands-free speech recognition systems face. A great number of researchers work on this field developing methods for the de-reverberation. In this thesis we do not implement any additional technique for de-reverberation, but only the delay-and-sum beamforming. However, the use of delay-and-sum beamforming is not efficient for this

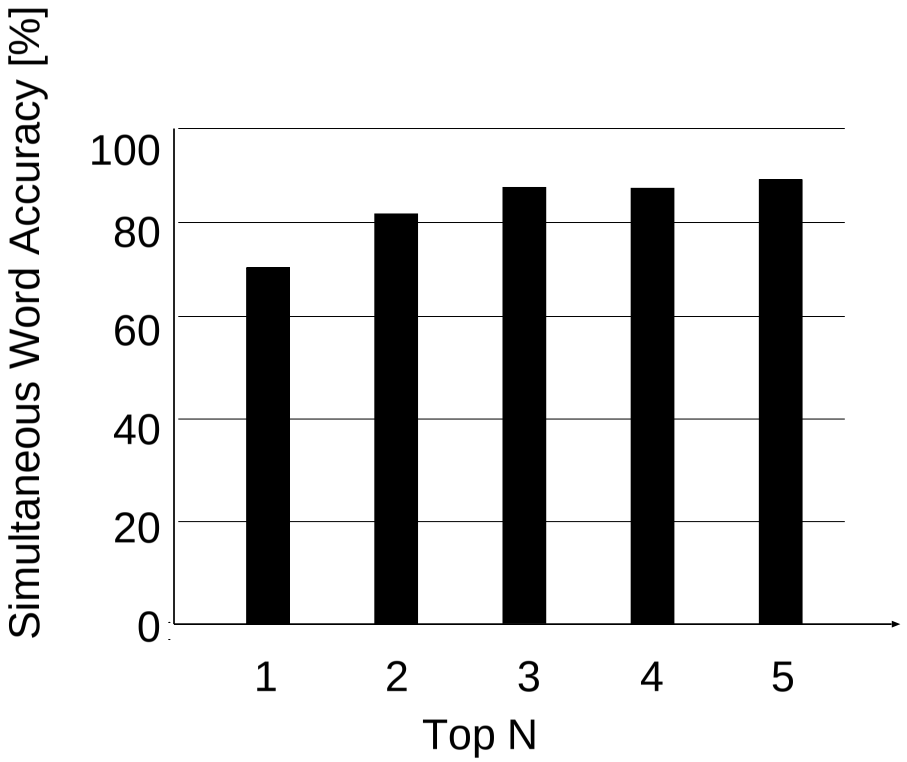

Figure 7.13. Simultaneous Word Accuracy - Microphone array composed of 32 channels

purpose.

7.3 Experiments Using Real Data

In this section, we describe the experiments carried out in real environment for the simultaneous recognition of distant-talking speech in real environment. Figure 7.21 shows the experiments conditions. The two sound sources are located at fixed positions at 10 and 170 degrees respectively. The speech data are played back through loudspeakers. A linear microphone array composed of 32 channels is

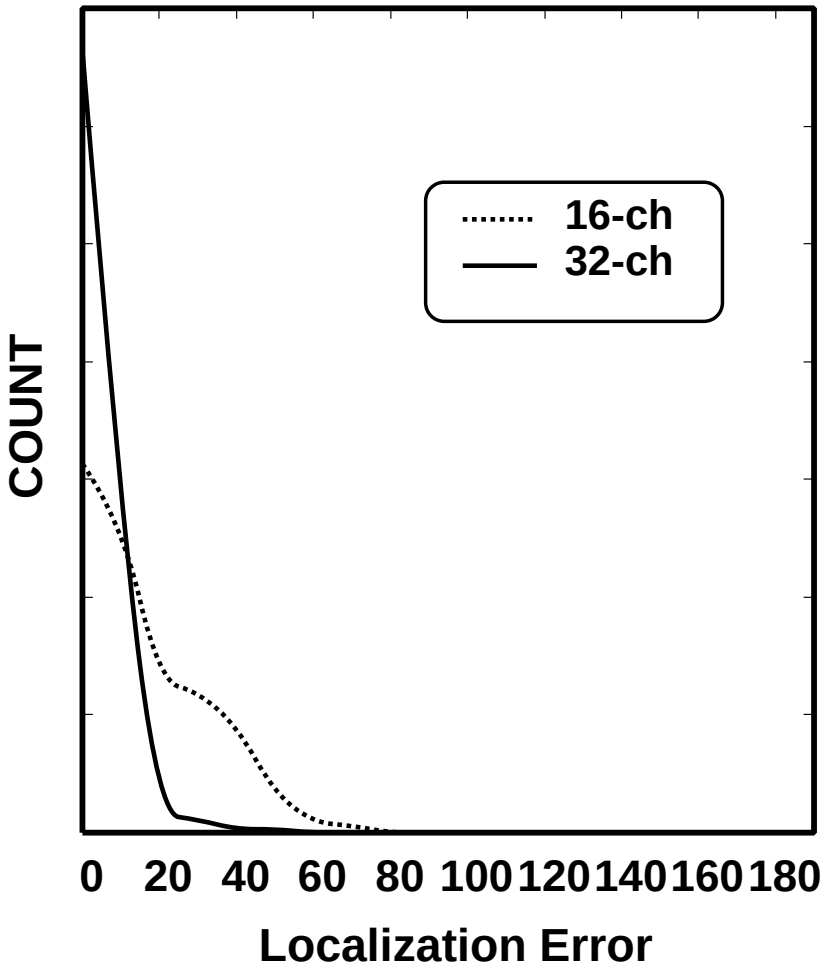

Figure 7.14. Histogram of the localization Error

used. The distance between the microphones is 2.80 cm. The distance between the loudspeakers and the microphone array is 2 m. The reverberation time ($T_{[60]}$) in the experimental room is 280 ms.

Figures 7.22, 7.23, and 7.24 show the obtained results in comparison with the achieved results by using the image method. Although the performance of our system in real environments is decreased, the achieved results are comparable

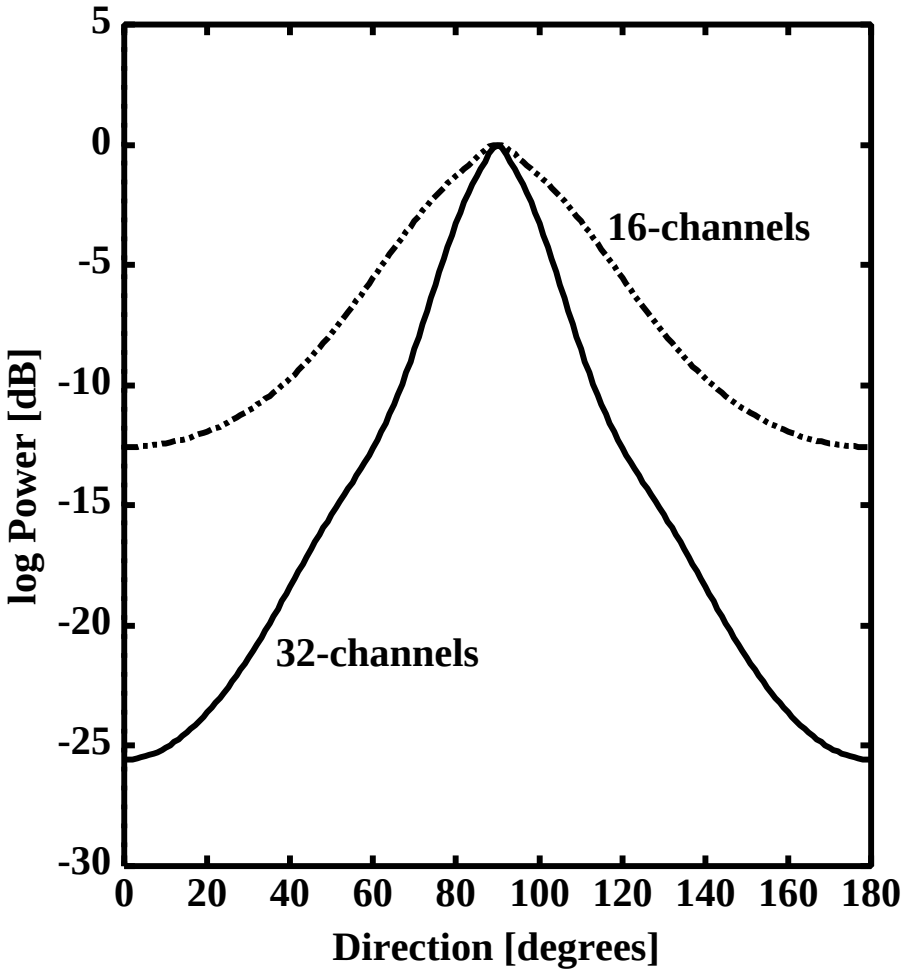

Figure 7.15. Directive patterns

with those when simulated reverberated data were used. However, in the case of using real data we should consider also the presence of ambient noise and the longer reverberation time. Therefore, we can conclude that the obtained results using real data are expected lower than the case when we use simulated data, and the comparison between the two cases is reasonable.

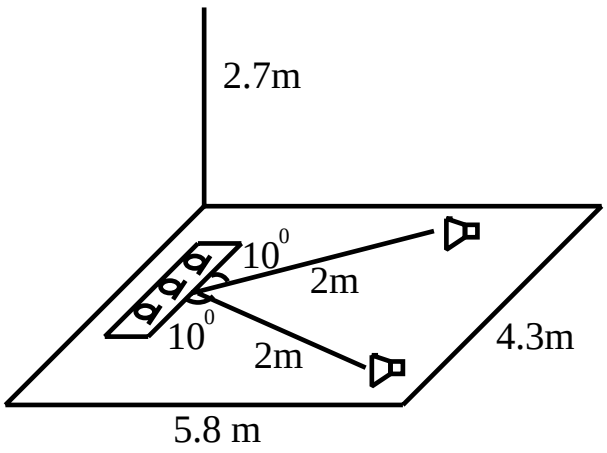

Figure 7.16. Experiment arrangement for reverberated environment

7.4 Experiments for the Recognition of Three Talkers

In this section we describe the experiments carried out for the simultaneous recognition of speech of three talkers. The three talkers are located at fixed positions at 10, 90, and 170 degrees, as Fig. 7.25 shows. In total we use 645 test word-pairs. The microphone array is linear and it is composed of 32 channels. The distance between two microphones is 2.83 *cm*. Figure 7.26 shows the Top 5 results. The A, B, and C indicate the Word Accuracy of the three talkers. $A+B$, $B+C$, and $A+C$ indicate the Simultaneous Accuracy for two talkers. Finally, the $A+B+C$ shows the Simultaneous Word Accuracy of all the three talkers. Comparing with the two talkers case, the performance is degraded, since the used delay-and-sum cannot separate efficiently the speech signals of the three talkers. However, results show that our system performs well even in the case of three talkers, too.

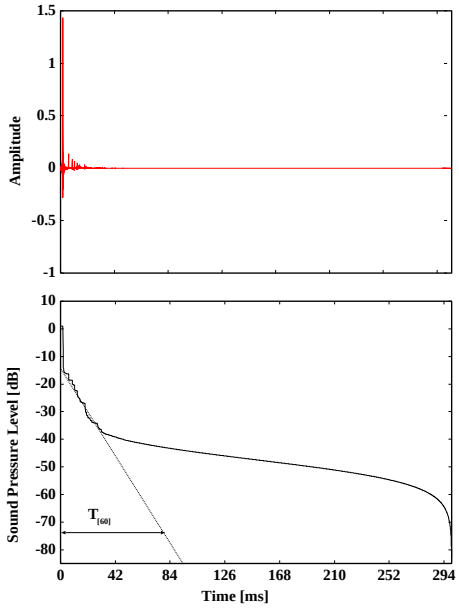

Figure 7.17. Measurement of reverberation time

7.5 Summary

This chapter describes the experiments carried out in order to evaluate the performance of our system. In the initial experiments the 3-D N-best search method was used without implementing the two additional techniques. However, in that case the performance of the our system was low. The performance was further improved by implementing the two additional techniques described in the previous sections. We carried out experiments using two microphone array geometries, namely we used a 16-channels and a 32-channels microphone array. By using a 32-channels microphone array much higher recognition rates could be obtained.

This section describes also the comparison between our method and a conventional talker localization method-based system. Finally, in this chapter we introduced the results of the experiments using reverberated data, and results

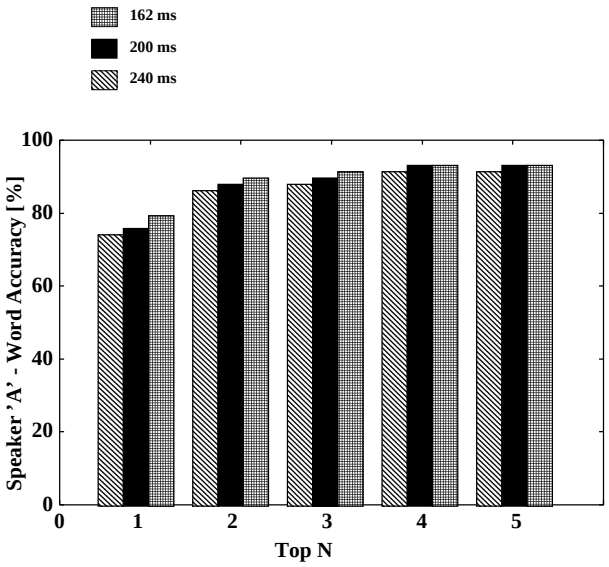

Figure 7.18. Speaker 'A' Word Accuracy

obtained for the recognition of three talkers.

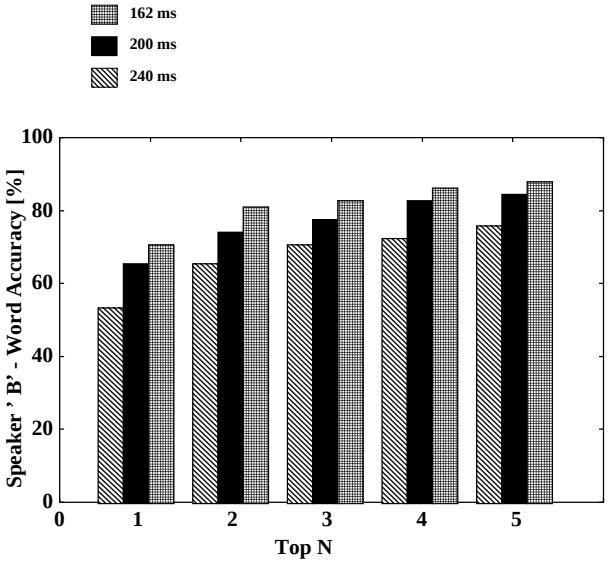

Figure 7.19. Speaker 'B' Word Accuracy

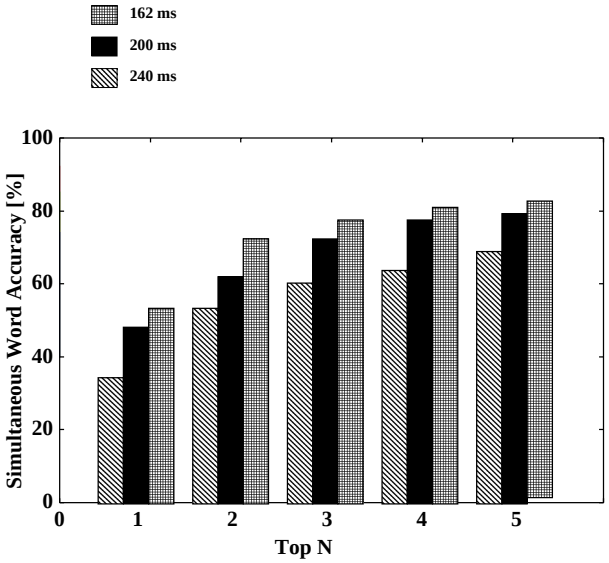

Figure 7.20. Simultaneous Word Accuracy

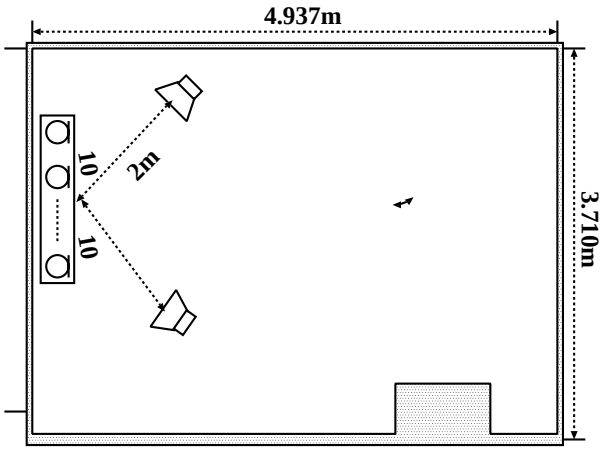

Figure 7.21. Experimental room

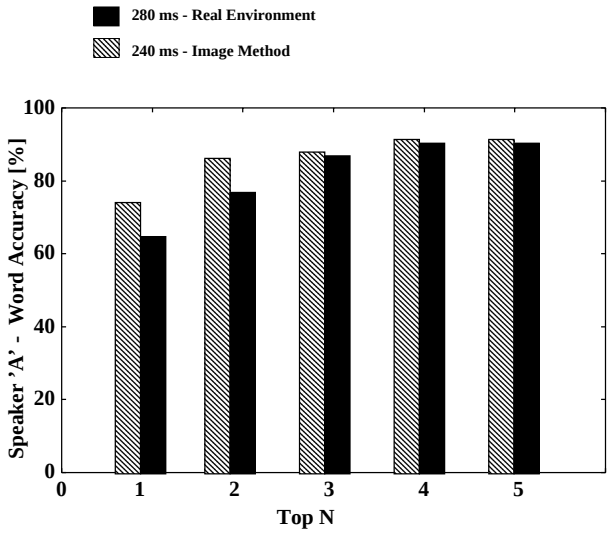

Figure 7.22. Speaker 'A' Word Accuracy

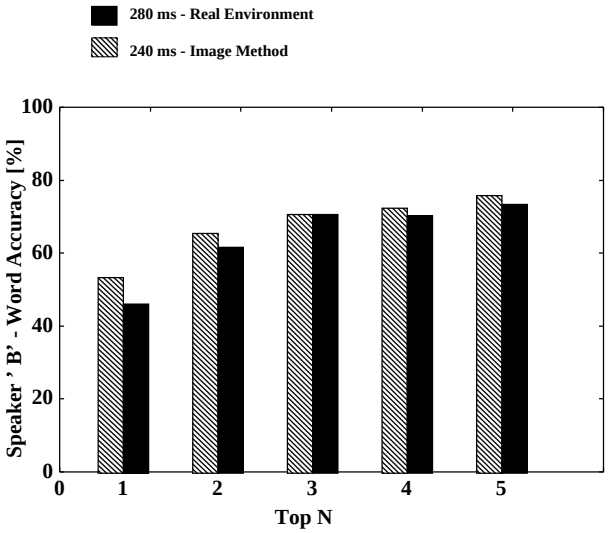

Figure 7.23. Speaker 'B' Word Accuracy

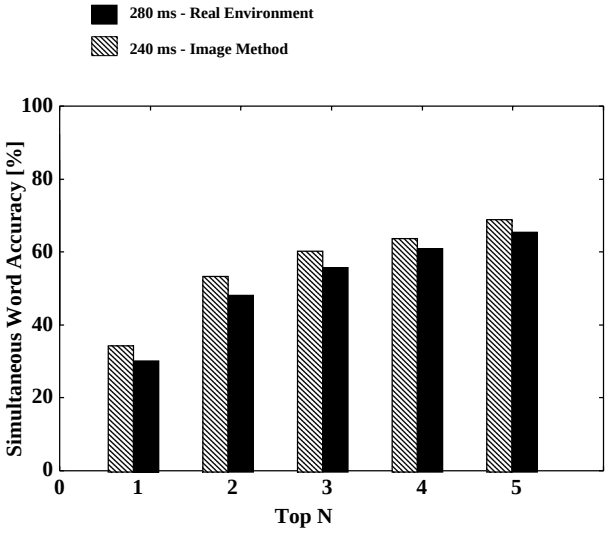

Figure 7.24. Simultaneous Word Accuracy

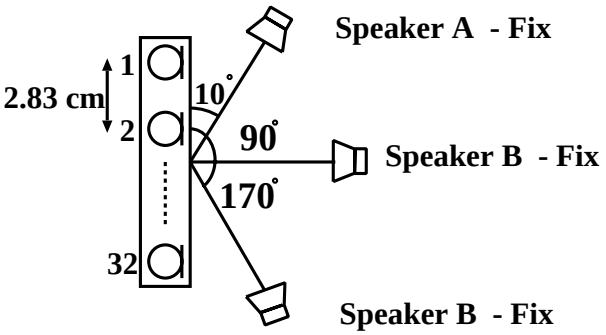

Figure 7.25. Experiment arrangement for three talkers

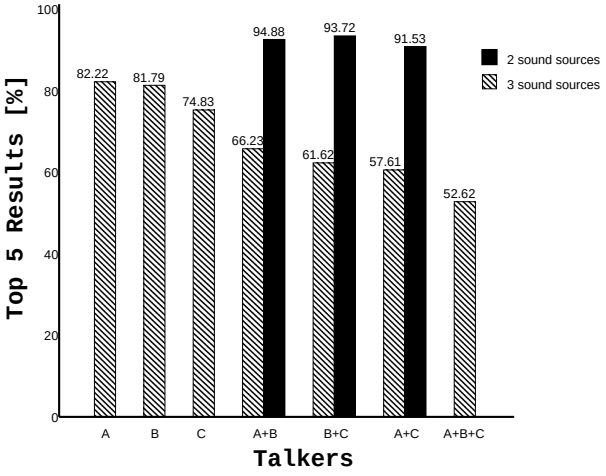

Figure 7.26. Word Accuracy for three talkers

Chapter 8

Conclusions and Future Work

8.1 Conclusions

This thesis deals with the simultaneous distant-talking speech recognition of multiple sound sources.

Chapter 1 described the distant-talking speech recognition problem and the current status of the research field. In this chapter we introduced our idea and proposed method to solve the problem of the simultaneous recognition of distant-talking speech.

Chapter 2 addressed the speech recognition of multiple sound sources problem and deals with possible solutions to this problem.

Chapter 3 dealt with speech recognition issues and explained about the speech recognition task, the Hidden Markov Models and the Viterbi algorithm.

Chapter 4 briefly addressed the microphone arrays, the delay-and-sum beamforming and some sound source localization techniques.

Chapter 5 described the 3-D Viterbi search method, which is the base of this thesis. We explained the idea, the formulation, the advantages and the disadvantages of this method.

Chapter 6 introduced our proposed method to solve the problem of the simultaneous distant-talking speech recognition of multiple sound sources. In this chapter we explained in details our proposed 3-D N-best search. Moreover we introduced two additional techniques, which drastically increased the performance of our system. Namely, a likelihood normalization technique and a path distance-based clustering were introduced.

Chapter 7 described the experiments were carried out on simulated data in order to evaluate the performance of our system. The experiments cover the cases of two fix talkers, one fix and one moving talker and two microphone array geometries. The obtained results are very promising and particularly in the case of 32-channels microphone array we could achieve higher than 72 % Simultaneous Word Accuracy. The results showed that by increasing the number of the microphones the performance of the system was significantly increased. The improvements in word accuracy by implementing the two additional techniques are also described. In this chapter are also given the results of the comparison of our proposed method and a conventional talker localization method based system. More specifically, we compared our proposed 3-D N-best search method based system with a CSP-based system. Results showed that higher Simultaneous Word Accuracy could be achieved by our system.

In the Chapter 7 we described also the experiments carried out on reverberated data, which was obtained by using the image method and using data recorded in real environments. The performance of our system under reverberant conditions was decreased.

8.2 Future Work

In this thesis we introduced our proposed 3-D N-best search algorithm able to recognize distant-talking speech of multiple sound sources. The obtained results are very promising and they justify the existence of our idea. However, problems are still remaining and further improvement is possible. Among others the following problems are still remaining:

- **Computational Cost**

  The search in the 3-D trellis space requires a large amount of computation. Moreover, the introduction of the N-best paradigm further increases the search space. Therefore, efficient methods are necessary in order to reduce the computational amount that is required. A possible solution is the use of the power as measure to eliminate the directions on which search is performed.

- **Multiple Talkers and Sound Sources**

  In this thesis we dealt only with the presence of two sound sources without considering noise components. However, for practical use the presence of noise sources should be also considered.

- **Speech Enhancement**

  The delay-and-sum beamforming algorithm provides limited speech enhancement. For efficiently sharp beams a large number of microphones is required. However, this increases the system's complexity and cost. A possible solution is the use of the adaptive microphone array.

- **Clustering Techniques**

  In this thesis we use a simple bottom-up method with the assumption that the number of the clusters is pre-defined and it is the same as the number of sources. The clustering could be more precise by avoiding this assumption and by using a more sophisticated clustering method. Stopping rules can be also used and in this way we can deal with unknown number of sound sources.

References

[1] G. W. Elko *Superdirectional microphone arrays.* Acoustic signal processing for telecommunications, Kluwer Academic Publishers, 2000

[2] D. H. Johnson, D. .E. Dudgeon Array Signal Procecing. Concepts and techniques P T R Prentice-Hall, Inc., 1993

[3] P. S. Naidu Sensor array signal processing CRC Press, 2000

[4] J. L. Flanagan, D. A.Berkley, G. W. Elko, J. E. West, and M. M.Sondhi. Autodirective microphone systems. Acoustica (75), 1991.

[5] G. W. Elko. Microphone arrays. In *Proceedings of International Workshop on Hands-free Speech Communication*, pages 11–14, 2001.

[6] M. Omologo. Hands-free speech recognition: Current activities and future trends. In *Proceedings of International Workshop on Hands-free Speech Communication*, pages 23–26, 2001.

[7] W. Kellerman. A self steering digital microphone array. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3581–3584, 1991.

[8] T. Yamada, S. Nakamura, and K. Shikano. An effect of adaptive beamforming on hands-free speech recognition based on 3-D Viterbi search. In *Pro-*

ceedings of International Conference on Spoken Language Processing, pages 381–384, 1998.

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

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

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

[12] 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 CO-COSDA Workshop'98* , pages 17–20, 1998.

[13] S. Nakamura. Acoustic sound database collected for hands-free speech recognition and sound scene understanding. In *Proceedings of International Workshop on Hands-free Speech Communication*, pages 43–46, 2001.

[14] 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.

[15] 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.

[16] 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, 1995.

[17] T. Takiguchi, S. Nakamura, and K. Shikano. Speech recognition for a distant moving speaker based on HMM composition and separation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1403–1406, 2000

[18] 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.

[19] S. Nakamura, T. Yamada, P. Heracleous, and K. Shikano. Recognition of distant-talking speech based on 3-D trellis search using a microphone array and adaptive beamforming. In *Proceedings of Workshop on Robust Methods for Speech Recognition*, pages 219–222, 1999.

[20] P. Heracleous, T. Yamada, S. Nakamura, and K. Shikano. Simultaneous recognition of multiple sound sources based on 3-D N-best search using microphone array. In *Proceedings of European Conference on Speech Communication and Technology*, pages 69–72, 1999.

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

[22] 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*, pages 91–92, 1999.

[23] 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.

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

[25] 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.

[26] P. Heracleous, S. Nakamura, and K. Shikano. Integrated talker localization and speech recognition based on the 3-D N-best search In *Proceedings of Acoustical Society of Japan*, pages ???-???,2001.

[27] B. H.Juang and F. K.Soong. Hands-free telecommunications. In *Proceedings of International Workshop on Hands-free Speech Communication*, pages 5–10, 2001.

[28] Y. Zhao. Frequency-domain maximum likelihood estimation for automatic speech recognition in additive and convolutive noises. In *IEEE Transactions on SAP, vol. 8, no.3*, pages 255-266, 2000

[29] A. Acero and R. Stern. Environment robustness in automatic speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 849–852, 1990.

[30] Y. Zhao. Speech enhancement - Issues and recent advances. In *Proceedings of International Workshop on Hands-free Speech Communication*, pages 19–22, 2001.

[31] S. K.Gupta,F. Soong, and R. Haimi-Cohen High Accuracy connected digit recognition for mobile applications. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 57–60, 1996.

[32] T. Nishiura, S. Nakamura, and K. Shikano. Robust speech recognition by multiple beamforming with reflection signal equalization. In *Proceedings of International Workshop on Hands-free Speech Communication*, pages 119–22, 2001.

[33] T. Nishiura, K. Miki, S. Nakamura, and K. Shikano. Complimentary combination of microphone array and HMM composition for noisy speech recognition. In *Proceedings of International Workshop on Hands-free Speech Communication*, pages 167–170, 2001.

[34] M. S.Brandstein and Darren B. Ward,editors Microphone arrays: Signal processing techniques and applications. Springer-Verlag, 2001

[35] M. Omologo and P. Svaizer. Use of the cross-power spectrum phase in acoustic event location. In *IEEE Transactions on SAP,* **5** *3*, pages 288–292, 1997

[36] D.Van Compernolle. Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings. In *Proceedings of*

IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 833–836, 1990.

[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. 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

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

[46] L. R. Rabiner. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. In *Proceeding of IEEE, Vol. 77*, pages 257–286, 1989.

[47] L. R. Rabiner and B. H.Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1990.

[48] F. Jelinek. *Statistical Methods for Speech Recognition* . MIT Press, 1999.

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

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

[51] H. Bourland and N. Morgan. Links between Markov models and multi-layer perceptron. In *IEEE Transactions on Pattern Analysis, Machine Intelligence, Vol. 12*, pages 1167–1178, 1992

[52] H. Bourlard and N. Morgan. *Connectionist speech recognition - A hybrid approach* . Kluwer Academic Publishers, 1994.

[53] B. H.Juang. Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains. In *AT&T Technical Journal, Vol. 64*,1985

[54] L. R.Liporase. Maximum likelihood estimation for multivariate observations of Markov sources. In *IEEE Transactions on Information Theory, Vol. 28*,pages 729-734,1982

[55] L. R.Bahl, P. F.Brown,P. V. de Souza, and R. L.Mercer Maximum mutual information estimation of hidden Markov models parameters for speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 49–52, 1986.

[56] Y. Normandin and D. Morgera An improved MMIE training algorithm for speaker-independent small vocabulary, continuous speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 537–540, 1991.

[57] J. A.Hartigan. *Clustering Algorithms* . 1975.

[58] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J.Stone. *Classification and regression trees* . Wadsworth &Brooks, 1984

[59] L. R.Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. In *IEEE Transactions on Pattern Analysis, Machine Intelligence, Vol. 5*, pages 179–190, 1983.

[60] A. Acero, S. Altschuler, and L. Wu Speech/noise separation using two microphones and a VQ model of speech signals. In *Proceedings of International Conference on Spoken Language Processing*, pages 532–535, 2000.

[61] R. Schwartz and Y. L.Chow. The N-best algorithm: An efficient and exact procedure for finding the N most likely sentence hypotheses. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 81–84, 1990.

[62] F. K.Soong and E. F.Huang. A tree-trellis based fast search for finding the N-best sentence in continuous speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 703–706, 1981.

[63] C. H. Lee, F. K.Soong, and K. K.Paliwal,editors. *Automatic Speech and Speaker Recognition - Advanced Topics : Chapter 17* Kluwer Academic Publishers, 1996.

[64] J. L.Gauvain and C. H.Lee. Maximum A Posteriori estimation of multivariate Gaussian mixture observations of Markov chains. In *IEEE Transanctions on Speech and Audio Processing, Vol 2*, pages 291–298, 1994.

[65] T. Yamada. *Hands-free Speech Recognition Using a Microphone Array.* Doctor Thesis, 1999.

[66] P. Heracleous, S. Nakamura, and K. Shikano. An improvement to 3-D N-best search using path distance-based clustering. In *Technical Report of the Institute of Electronics, Information and Communication Engineers*,pages 85–89, Dec. 1999.

[67] P. Heracleous, S. Nakamura, and K. Shikano. Evaluation of 3-D N-best search using path distance-based clustering for recognizing multiple sound sources. In *Proceedings of Acoustical Society of Japan*, pages 157–158, 2000.

[68] P. Heracleous, S. Nakamura, and K. Shikano. A technique for likelihood normalization in the 3-D N-best search for simultaneous recognition of multiple sound sources. In *Proceedings of Acoustical Society of Japan*, pages 117–118, 2000.

[69] P. Heracleous, S. Nakamura, and K. Shikano. A microphone array-based 3-D N-best algorithm for the simultaneous recognition of multiple sound sources in real environments. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages ?–?, May 2001.

[70] T. Matsui and S. Furui. Likelihood normalization for speaker verification using a phoneme- and speaker-independent model. In *Speech Communication 17*,pages 109–116, 1995.

[71] J. B.Allen and D. A.Berkley. Image method for efficiently simulating small-room acoustics. Journal of Acoustical Society of America, vol. 65, No 4,pages 943–950, 1979.

List of Publications

Journal Paper

1. P. Heracleous, S. Nakamura, T. Yamada, and K. Shikano. A Microphone array-based 3-D N-best search for simultaneous recognition of multiple sound sources. *Transactions of the Institute of Electronics Information and Communication Engineers*, (ACCEPTED)

2. P. Heracleous, S. Nakamura, and K. Shikano. Simultaneous recognition of distant-talking speech of multiple talkers based on the 3-D N-best search method. *Journal of VLSI Signal Processing*, (SUBMITTED)

International Conference

3. P.Heracleous, T. Yamada, S. Nakamura, and K. Shikano. Simultaneous recognition of multiple sound sources based on 3-D N-best search using microphone array. In *Proceedings of European Conference on Speech Communication and Technology*, pages 69–72, Sep. 1999.

4. P.Heracleous, S. Nakamura, and K. Shikano. Multiple sound sources recognition by a microphone array-based 3-D N-best search with likelihood normalization. In *Proceedings of International Workshop on Hands-free Speech*

Communication, pages 103–106, April 2001.

5. P.Heracleous, S. Nakamura, and K. Shikano. A microphone array-based 3-D N-best algorithm for the simultaneous recognition of multiple sound sources in real environments. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 193–196, May M2001.

6. P.Heracleous, S. Nakamura, and K. Shikano. Simultaneous recognition of distant-talking speech of multiple sound sources based on 3-D N-best search algorithm. In *Proceedings of ASRU2001 Workshop*, pages –, December 2001.

7. S.Nakamura, T.Yamada, P.Heracleous, and K.Shikano. Recognition of distant-talking speech based on 3-D trellis search using a microphone array and adaptive beamforming. In *Proceedings of Workshop on Robust Methods for Speech Recognition*, pages 219–222, May 1999.

Domestic Conference

8. 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*, pages 91–92, March 1999.

9. P. Heracleous, S. Nakamura, and K. Shikano. An improvement to 3-D N-best search using path distance-based clustering. In *Technical Report of the Institute of Electronics, Information and Communication Engineers*, pages 85–89, Dec. 1999.

10. P.Heracleous, S.Nakamura, and K. Shikano. Evaluation of 3-D N-best search using path distance-based clustering for recognizing multiple sound

sources. In *Proceedings of Acoustical Society of Japan*, pages 157–158, March 2000.

11. P.Heracleous, S.Nakamura, and K. Shikano. A technique for likelihood normalization in the 3-D N-best search for simultaneous recognition of multiple sound sources. In *Proceedings of Acoustical Society of Japan*, pages 117–118, Sep. 2000.

12. P.Heracleous, S.Nakamura, and K. Shikano. Integrated talker localization and speech recognition based on the 3-D N-best search. In *Proceedings of Acoustical Society of Japan*, pages 185–186, March 2001.

Master thesis

13. P. Heracleous. Comparison of Speech Coding Algorithms. Master Thesis. Department of Communication Engineering, Technical University of Budapest, October 1992.