NAIST-IS-DD0761003

# Doctoral Dissertation

# Noise Reduction Front-End for Robust Speech Recognition Using Multi-Channel Signals and Harmonic Structures

Osamu Ichikawa

September 30, 2008

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

A Doctoral Dissertation
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

Osamu Ichikawa

Thesis Committee:

| | |
|---|---|
| Professor Kiyohiro Shikano | (Supervisor) |
| Professor Masatsugu Kidode | (Co-supervisor) |
| Associate Professor Hiroshi Saruwatari | (Co-supervisor) |

# Noise Reduction Front-End for Robust Speech Recognition Using Multi-Channel Signals and Harmonic Structures[*]

Osamu Ichikawa

## Abstract

In general, automatic speech recognition (ASR) is sensitive to ambient noise. Therefore, the original commercial ASR products used close-talk microphones. Now many ASR products are equipped with far-field microphones, relying on noise-reducing front-ends and multi-style training in their acoustic models. Typical examples are car navigation systems and consumer electronic devices. However, most of them assume moderate and stationary noise sources and limited vocabularies of several hundreds words. Their noise robustness is still inadequate for many tasks.

In our daily life, we encounter a large variety of noises. For example, in automobiles, there are stationary noises such as cruising noises and fan noises, and non-stationary noises such as passengers' voices, radios or other audio devices, squeaking windshield wipers, or the sounds of passing traffic. Also, the mix of noises and signals will change, even with relatively stable noise sources such as cruising noise, resulting in fluctuations of the SNR that impact automatic speech recognition.

In this dissertation, three novel methods are proposed and evaluated to cope with variations in the noise. The first method is a new microphone array technology called Profile Fitting (PF) to cope with non-stationary noise using directional information. This method focuses on a profile of the shape of the power distribution according to the beamforming direction. An observed profile can be decomposed into known template profiles for directional sound sources and a non-directional background sound source. Evaluations confirmed this method significantly reduced the error rate in automatic

[*] Doctoral Dissertation, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0761003, September 30, 2008.

speech recognition.

PF can also be used for sound source localization. The sound source location (or direction) is essential information for beamformers unless Blind Signal Separation (BSS) is used for signal separation. Conventionally, localization methods such as MUSIC and CSP are used in addition to non-BSS beamformers. However, PF can integrate localization and signal separation into a single process. Furthermore, PF can extend the localization capability for a combination of sound reflectors, because the "profile" introduced by PF contains all of the localization cues such as reflections and diffusion effects as well as inter-channel time differences (ITD), and inter-channel sound intensity differences (IID). Experiments show this method combined with sound reflectors can provide a rough estimate of a vertical location even in a noisy environment, which was a difficult task for conventional microphone array technologies using two microphones.

The second method is a new echo cancellation technology named SSEC (Simultaneous adaptation of spectral Subtraction and Echo Cancellation) to cope with non-stationary noises such as music or human voices coming from a radio, car-navigation system, or other audio device. It uses reference signals from those devices to cancel echo components in the observed signals. Most of the conventional echo cancellers are based on time domain LMS, which requires heavy computations and suffers from performance degradation in high ambient noise environments. To avoid these difficulties, echo cancellation can be implemented using spectral subtraction. However, in automobiles, there is a practical problem of how to estimate the cruising noise while music is playing continuously. SSEC solves this problem by estimating the ambient noise component and the echo canceller's coefficients simultaneously under the assumption that ambient noises such as cruising noises and fan noises are relatively constant in automobiles. Experiments show SSEC significantly reduced the errors in automatic speech recognition compared with the conventional combination of an echo canceller and spectral subtraction.

The third method is a new speech enhancement method exploiting the harmonic structures observed in human voices. This is designed to improve the accuracy of automatic speech recognition in very low SNR situations such as high-speed cruising with an open window or a noisy fan. In such situations, speech signals are often buried in broadband noise and the accuracy of automatic speech recognition is greatly

degraded.

Microphone array technology can improve the output SNR. However, when adaptive beamformer is configured with small number of microphones and the noise source is non-directional (i.e. not from a single point), such as cruising noise, then the degree of improvement is very limited. Therefore, a different approach using harmonic structure was investigated to retrieve the speech information buried in the broadband noise. A new method called LPE (Local Peak Enhancement) was devised. Most of the conventional methods are based on comb filtering, which depends on accurate pitch frequency and reliable voiced/unvoiced detection. However, the detection is not accurate enough in very low SNR situations. LPE does not depend on this, because it designs a filter for speech enhancement directly from the observed spectrum. Experiments using automatic speech recognition show that LPE significantly improves the accuracy in very noisy conditions such as a noisy fan or an open window. They also confirmed that LPE can be combined with existing noise reduction algorithms such as SS and Wiener Filtering for further improvements.

**Keywords:**

automatic speech recognition, beamformer, acoustic echo canceller, harmonics, ITD, IID, microphone array, noise reduction, sound reflector, sound source localization, spectral subtraction, speech enhancement

# マルチチャンネル信号と調波構造を利用したロバスト音声認識のための雑音除去フロントエンド*

市川　治

## 内容概要

一般に音声認識は背景雑音の影響を受けやすい。そのため、音声認識が初めて実用化された当時は、音声の入力手段として接話マイクロフォンを利用するのが一般的であった。現在では、雑音除去フロントエンドと音響モデルのマルチスタイル学習の適用により、カーナビゲーションシステムや家庭用電気製品など遠隔マイクロフォンを利用した実用製品が数多く市販されるようになった。しかしながら、そのほとんどが中程度のレベルの定常雑音と数百単語の語彙サイズの単語認識を前提としており、十分な耐雑音性を備えているとは、言えないのが現状である。

日常生活において経験する雑音は、多種多様である。例えば、自動車内では、ほぼ定常と考えられる走行雑音や空調騒音の他に、助手席や後部座席の同乗者からの発声、ラジオなどオーディオ機器からの再生音、ワイパー動作音、他車通過音などの非定常雑音が存在する。また、走行雑音についても、低速・窓閉め走行などの比較的高い SN 比を確保できるケースと、高速・窓開け走行など、非常に低い SN 比のケースでは、音声認識に与える影響はかなり異なる。

本論文では、それら多様な雑音に対処するために、3つの手法を提案・検証する。1つ目の手法は、プロファイルフィッティング(PF)と名付けた新しいマイクロフォンアレイの技術である。音源方向性を利用することにより、非定常雑音に対処する。到来する音声の角度別パワー分布（観測プロファイル）に着目し、これを既知のテンプレートプロファイルに成分分解することにより、目的方向の信号成分を抽出するものである。実験によれば、この方式を音声認識のための雑音除去フロントエンドとして用いることにより、従来技術に比べ大幅に音声認識率を改善することができた。

* 奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 博士論文, NAIST-IS-DD0761003, 2008 年 9 月 30 日.

PFは、音源位置推定として用いることもできる。信号音源の位置を正しく推定することは、BSS (Blind Signal Separation)以外のビームフォーマにおいては、不可欠な要素である。従来は音源位置推定には、MUSICやCSPなど信号分離とは系統の異なる手法を併用することが多かったが、PFは、信号分離と音源位置推定の処理を統一することができる。さらに、PFが導入したプロファイルという概念は、音源位置推定の手がかりとなる、チャネル間の位相差と強度差、さらには反射や拡散性の情報を包含してので、これを用いて高度な音源位置推定が可能になった。実験では、マイクロフォンに装着した反射板との併用で、従来、2つのマイクロフォンでは困難であった正中面の音源仰角の推定精度を大幅に改善した。

2つ目の手法は、SSEC (Simultaneous adaptation of spectral Subtraction and Echo Cancellation) と名付けた新しいタイプのエコーキャンセラの技術である。自動車のオーディオ機器からの音楽やカーナビのガイダンス音声が雑音源である場合には、それらの機器から参照信号を得て、エコーキャンセラを構成することで、観測音声に含まれるそれら雑音成分を効果的に除去することができる。従来のエコーキャンセラは、時間領域の2乗誤差最小化の原理に基づくものが多い。しかし、それらは計算量負荷が大きく、また、背景雑音がある場合に性能が劣化する点が問題であった。これに対し、エコーキャンセラをスペクトルサブトラクションの形式に書き直し、処理を軽くした形式が期待されるが、自動車の場合、エコー成分（音楽など）とは無関係な走行雑音が存在し、この成分を容易には推定できないことが問題であった。SSECは、走行雑音が定常であるという仮定のもとに、スペクトルサブトラクション形式のエコーキャンセラの適応と定常雑音成分の推定とを同時に行う。これにより、走行中にオーディオ音声が再生され続けているという状況でも走行雑音成分とエコー成分（再生音）の両方を推定し除去することができる。実験では、従来技術であるエコーキャンセラとスペクトルサブトラクションの組み合わせよりも、大幅に音声認識率を向上させることができた。

3つ目の手法は、調波構造を利用した新しい音声強調の手法である。これは、自動車の高速・窓開け走行など、非常に低いSN比の状況での音声認識率を改善するためのものである。この領域では音声は広帯域に広がった雑音に埋もれかかっている。マイクロフォンアレイ技術によってもSN比は改善することができるが、マイクロフォン数が小規模の適応ビームフォーマで、かつ、雑音源が走行雑音のように拡散性（非点音源）である場合、そのSN比改善効果は限定的なものに留まることが知られている。そのため、ここでは、埋もれかかった音声を強調するために、音声の持つ調波構造を利用する。調波構造を利用する従来技術の多くは、くし型フィルタをベースとしており、正確なピッチ周波数の推定と有声音・無声音判定を前提としている。しかし、この推定・

判定は高騒音の環境下では不正確になるという問題があった。そこで、ここでは、観測パワースペクトルそのものから、直接フィルタを設計する手法 LPE (=Local Peak Enhancement)を提案する。くし型フィルタと異なり、ピッチ周波数の推定や有声音・無声音判定は必要ない。本編に示した音声認識実験では、自動車の高速・窓開け走行やファン最大のケースで特に大きな改善を示した。また、スペクトルサブトラクションやウィーナフィルターなど既存の雑音除去手法と組み合わせることで、さらに大きな改善が得られることを確認した。

**キーワード:**
ビームフォーマ、エコーキャンセラ、調波構造、チャネル間位相差、チャネル間強度差、マイクロフォンアレイ、雑音除去、音響反射板、音源位置推定、スペクトルサブトラクション、音声強調、音声認識

# Acknowledgements

The research described in this dissertation was accomplished as the results of several research projects in which I have been involved at the Tokyo Research Laboratory of IBM Japan and in the Graduate School of Nara Institute of Science. Since there are many people who were involved in these projects, I want to thank all of them for their time, efforts, and cooperation which eventually resulted in this work.

I would like to express my foremost gratitude to my supervisor, Professor Kiyohiro Shikano for his constant guidance and fruitful suggestions. He encouraged me to carry out my research for this dissertation through my current job. Without his support, I could not have made this dissertation a reality.

I would also like to express my sincere gratitude to a member of my thesis committee, Professor Masatsugu Kidode, for his acute and valuable comments on this dissertation.

I would especially like to express my greatest appreciation to Associate Professor Hiroshi Saruwatari for his enlightening guidance based on his wide knowledge and experiences. Actually, he kindly gave me many suggestions before my entrance into the doctoral course. Through him, I was strongly motivated to apply for this course.

I wish to thank all the members in Shikano Laboratory, who accepted me with a warm and friendly atmosphere. I especially thank Dr. Shigeki Miyabe (currently at the Georgia Institute of Technology) for giving me good hints and tips as I entered into this course.

I would like to express my sincere gratitude to Dr. Masafumi Nishimura (IBM Japan). He is my mentor and also my manager. In all aspects, he has been a great supporter of my research work. I also thank all of the team members (including former members) belonging to the Speech Technology team in the Tokyo Research Laboratory for their intelligent discussions. Especially, Dr. Takashi Fukuda and Dr. Tetsuya Takiguchi helped me a lot as the co-authors of many of my papers.

I would like to express my gratitude to my higher-level managers Ms. Mayumi Itakura and Dr. Kohichi Takeda and to the director of the Tokyo Research Laboratory, Dr. Hiroshi Maruyama. They clearly emphasized the importance of a doctoral degree in my Individual Development Plan, and encouraged me to tackle the challenge.

I would also like to express my gratitude to Mr. Shannon Jacobs. He reviewed my English in most parts of this dissertation and in many of my original papers. He is an intelligent person using his insights in different fields. His reviews went beyond simple refinements of my English expressions.

I owe a great deal to IBM Japan Ltd. for the scholarship support for this doctoral course. I thank Mr. Sakae Uno in TRL’s Human Resources section for his support and for the arrangement for my scholarship funding.

Finally, I would like to thank my wife Eiko Ichikawa, my mother Motoe Ichikawa and my late father Fujio Ichikawa from bottom of my heart. All my activities, including my research for this dissertation, are largely dependent on their support. Without their care and encouragement in my daily life, I could not even have begun this dissertation, let alone manage to complete it.

# Contents

# List of Figures

# List of Tables

# 1. Introduction

## 1.1. Background

For human beings, Automatic Speech Recognition (ASR) is a natural extension for the man-machine interface. Early ASR products required headsets with close-talk microphone to capture the speech without noise. Noise reduction technology and advances in acoustic modeling allowed far-field microphones to be used. This is called hands-free speech recognition and allows for a more natural interface without wires. Nowadays these interfaces are used in various situations, but users agree their noise robustness should be improved. Here are some restrictions of current products:

- **Car navigation system:** When the Talk-Button is pressed, the car audio and fan are automatically stopped or quieted in most systems.
- **Robot (Humanoid):** Many of them have a speech interface. However, it is rare to see a hands-free ASR demo in a noisy convention hall.
- **Consumer electronic devices:** In TVs or game machines, microphone is often equipped in a remote commander, so to get microphone closer to users..

Noise robustness is not easy to achieve, because we need to cope with various types of noises. For example, in robot applications, there are other people's voices, background noises, actuator noises of the robot itself, and the synthesized voice of the robot (Figure 1.1). In automobiles, there are cruising noises, audio noises, passenger's voices, and various environmental noises such as horns, the squeaks of windshield wipers, neighboring cars, train crossing signals, etc. Table 1.1 summarizes some features of these types of noises as observed in automobiles. Here, to simplify discussion in this dissertation, cruising noise is treated as a stationary noise, and non-stationary sounds such as potholes are classified as environmental noise.

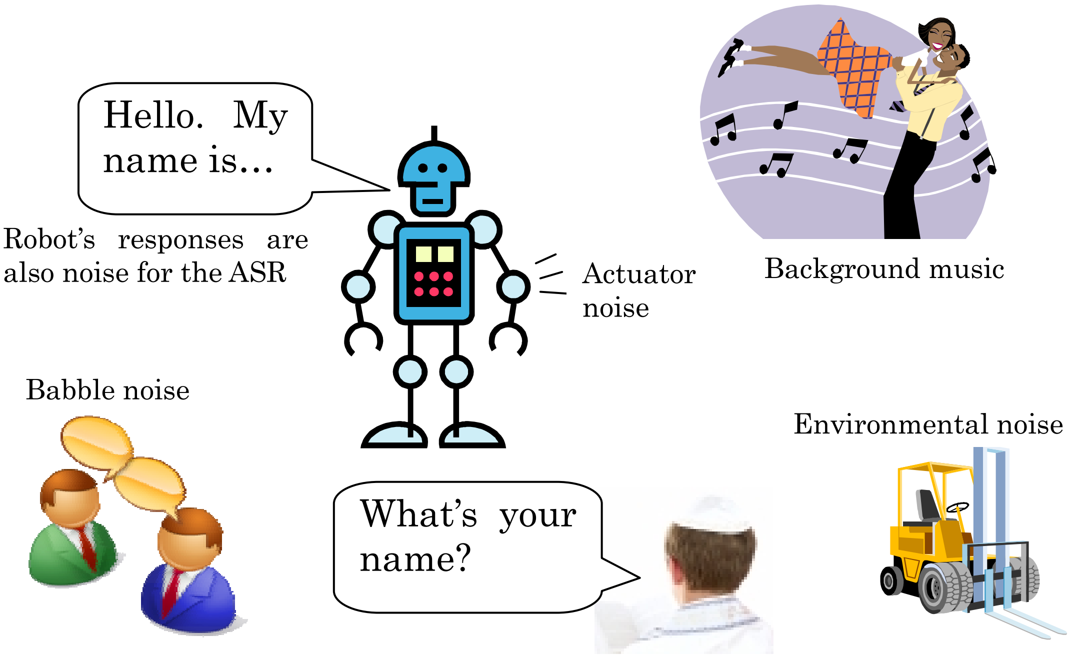


Figure 1.1. Various noises in a robot application.

For stationary noise, we can compensate in a straightforward manner by estimating the noise spectrum in the non-speech segments and subtracting it from the observed spectrum. For further improvements, we can use model-based compensation.

For non-stationary noise, we still have various methods as long as the noise does not have a harmonic structure while the target signal is speech. Comb filtering [TO98] is one of the approaches to enhance the harmonic spectrum of vowels. In this dissertation, we may refer to this type of compensation as speech enhancement rather than noise reduction.

If the noise is non-stationary and it has a harmonic structure, as is true with music or irrelevant human speech, then the compensation is not easy when using a single channel microphone. If two or more microphones are available, we can use beamformer focusing on the target by configuring the microphones as a microphone array. For automobiles, this may filter for only those sounds arriving from the driver's direction. Sounds from other directions such as passenger seats and audio loudspeakers can be filtered out. In current automobiles, standard position of the microphone is considered to be near a map-lamp on the ceiling. As shown in Figure 1.2, this position has also a good advantage to distinguish the driver's voice from passengers' voices by directionality of the sounds when multiple microphones are installed.

Table 1.1. Types of noises observed in automobiles

| | Directionality | Stability | Harmonic structure | Reference signal | Possible solution |
|---|---|---|---|---|---|
| Cruising noise | Non-directional | Stationary | No | Not available | Single channel noise reduction |
| Audio Sound (CD,TV,Radio, Car-Navigation) | Directional | Non-stationary | Yes | Possibly Available | Echo canceller, Beamformer |
| Passenger's voice | Directional | Non-Stationary | Yes | Not available | Beamformer |
| Environmental noise | Directional<br>or<br>Non-directional | Stationary<br>or<br>Non-stationary | Yes<br>or<br>No | Not available | Beamformer, Single channel noise reduction |

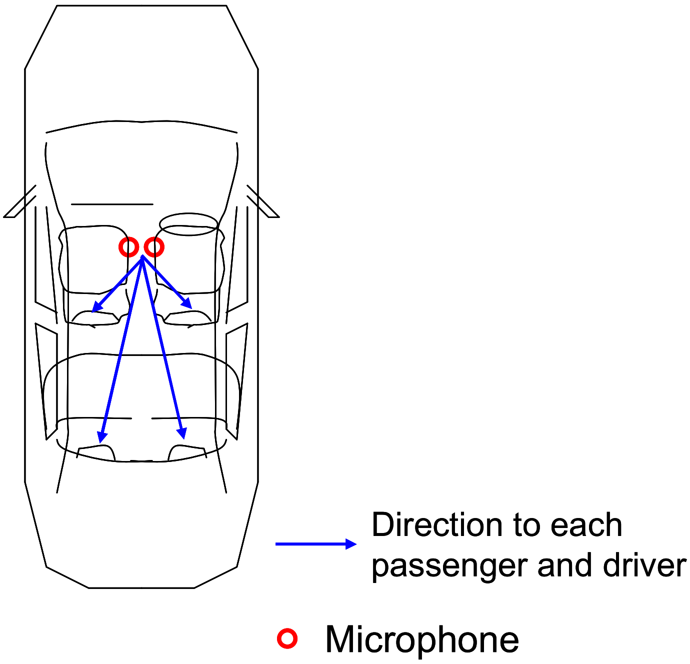


Figure 1.2. Standard position of microphone and directionality of driver's and passengers' voices in automobiles.

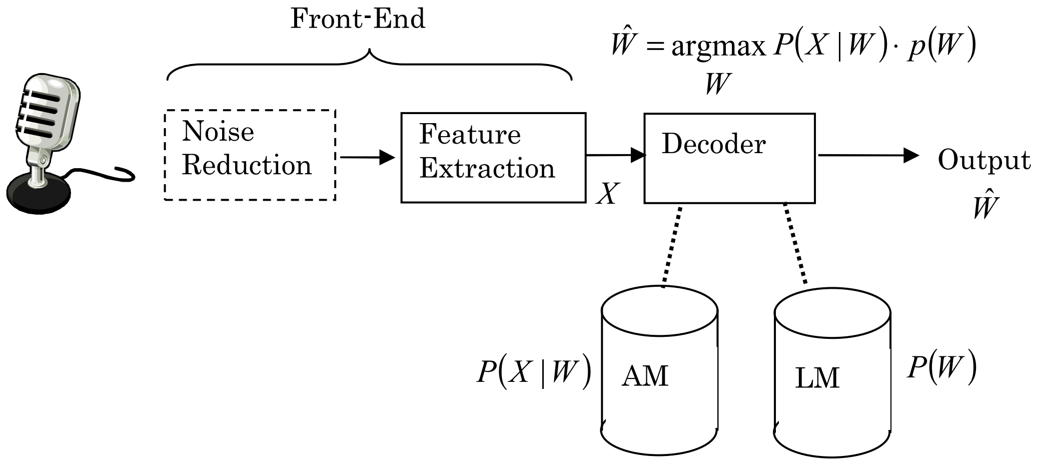


Figure 1.3. Automatic speech recognition system.

A beamformer depends on the direction (or location) of the sound sources. Therefore, a noise will not be reduced if it arrives from the same direction as the target. For example, guidance messages from the car navigation system, which are broadcasted from a loudspeaker on the driver's side, may escape filtering. In this case, we can use an echo canceller if a reference signal for the broadcast sound is available. An echo canceller can specifically reduce the noise component correlated with a reference signal.

However, most of current products support only stationary noise with single-channel noise reduction technology and an acoustic model trained with certain noises depending on their task..

# 1.2. Automatic Speech Recognition System

Figure 1.3 shows a diagram of a standard ASR system. The feature extraction part converts the input signal to feature vectors such as MFCC. It may also compensate for the multiplicative distortion and additive noise using CMS, CDCN [AS90], or other techniques. The Acoustic Model (AM) part contains statistical data trained from a large corpus of speech sounds with phonetic labels. The AM is often modeled as a HMM using EM algorithm. It is common to train the speech data along with noise to increase the noise robustness. This technique is called multi-style training. Adaptation methods such as MLLR [LW95] and HMM Composition (PMC) [GY96] may be done at this stage for speaker and environmental adaptations. The Language Model (LM) part contains statistical data about word sequences. For transcription, it is often modeled using an N-Gram approach. For command input, it is modeled using a constrained grammar. In both case, the LMs are often compiled into a Finite State Machine (FSM) to be used in decoding. The decoder part searches for the most likely word sequences using the AM and LM data. The Viterbi algorithm is often used for the search.

For noise robustness, a noise reduction part can be placed in the front-end part to pre-process the input data. This dissertation focuses on this noise reduction part. This includes multi-channel signal processing such as microphone arrays or echo cancellers.

Generally speaking, noise reduction may involve some side effects. The processed signal may have been altered in an undesirable way depending on the noise reduction method. Therefore, if necessary, the AM should be retrained with data that was processed with the noise reduction method.

# 1.3. Conventional Noise Reduction Technology

## 1.3.1. Single-channel noise reduction

For single-channel noise reduction, candidate methods include spectral subtraction (SS) [Bol79], Wiener Filters [ETS02], MMSE [EM84], CDCN [AS90], model-based compensation [STB+01][ATI06], noise and speech model reconstruction [KH03], and particle filtering [FN05]. Many noise reduction methods rely on the assumption that the noise is stationary. Particle filtering and model-based reconstruction appear to support filtering non-stationary noises. However, further research is still required for its use in realistic environments.

A comb filter [TO98] passes only the harmonic bins in the voiced segments. It works based on the estimated pitch (F0) information. Therefore, the performance is highly dependent on the accuracy of pitch detection [Boe93][NIZ03].

## 1.3.2. Multi-channel noise reduction

### 1.3.2.1. Microphone array

Microphone array technology includes beamformers, localization, Blind Signal Separation (BSS), and supporting technologies.

The most basic beamformer is called Delay and Sum (DS). This sums up the signals for all of the channels with their own delays so that all of the signals are synchronized to the target sound source. An Adaptive Beamformer (ABF) also sum up signals with individual delays and gains so that the residual noise may be minimized. When a noise source is directional and the room is non-reverberant, ABF will have a null-beamforming pattern for each noise source. ABF is also known as a Minimum Variance (MV) beamformer [AAM00]. Griffiths-Jim (GJ) [GJ82] consists of two beamformers. The added one is a null-beamformer used on the target signal source so as to enhance only the output noise component. The noise component is then subtracted from GJ's main beamformer output to further reduce the noise. This can be implemented using spectral subtraction as a Spatial Subtraction Array (SSA) [ONS+05]. This subtracts the sub-beamformer output from the main-beamformer output in the power spectrum domain. In this dissertation, the two-channel version of SSA is also referred to as 2-channel (Adaptive) SS [KAS+96].

A beamformer needs to know the location (or direction) of the target signal. Therefore, some localization methods such as MUSIC [JD] and CSP [OS96] are often used in conjunction. In contrast, BSS does not require that information. It can separate a mixed signal into separate signal components using the statistical independence between the signals based on the ICA algorithm [SKT+03]. For further improvements, it is extended to be combined with Binary Mask [MTS+06] and SSA [TTS+06].

### 1.3.2.2. Echo canceller

The most basic adaptation algorithm for an echo canceller is LMS. The normalized form is known as N-LMS. RLS, Sub-band LMS, and the ES algorithm [MK92] were developed for faster convergence. They are implemented in the time domain using long filter taps that should be sufficiently long relative to the the room's reverberation.

Therefore, they tend to require intensive computations. This drawback becomes much more severe when they are extended for stereo or a 5.1-channel surround system.

The echo canceller can be implemented in the spectrum domain or in the power spectrum domain for reduced computation. The implementation in the power spectrum domain works like spectral subtraction and is sometimes used as a post-processing step with a time domain echo canceller to suppress the remaining noise [DP97].

There are some extended versions of echo canceller. MCDCN [DDI04] that enhanced CDCN technique so to utilize reference signal. Semi-blind source separation (SBSS) [MTM+06] adopted independent component analysis (ICA) to achieve echo cancellation without double talk detector.

## 1.4. Contribution

In this dissertation, the following three novel methods for noise reduction or speech enhancement are proposed to improve the accuracy of automatic speech recognition in noisy environments.

- A new microphone array technology using a power distribution profile for beamforming and sound source localization.
- A new echo canceller that can simultaneously perform echo adaptation and ambient noise estimation.
- A new speech enhancement method using harmonic structures without relying on pitch and voiced/unvoiced detection.

As shown in Figure 1.4, the above three proposed methods should be placed in the noise reduction part of the front-end. They should be switched depending on the hardware availability or the variation of the noise to cope with.

## 1.5. Thesis Outline

This dissertation is organized as follows. In Chapter 2, we briefly review robust ASR to position the three novel methods proposed in this dissertation. In Chapter 3, a new microphone array technology that shows higher noise reduction capability with limited

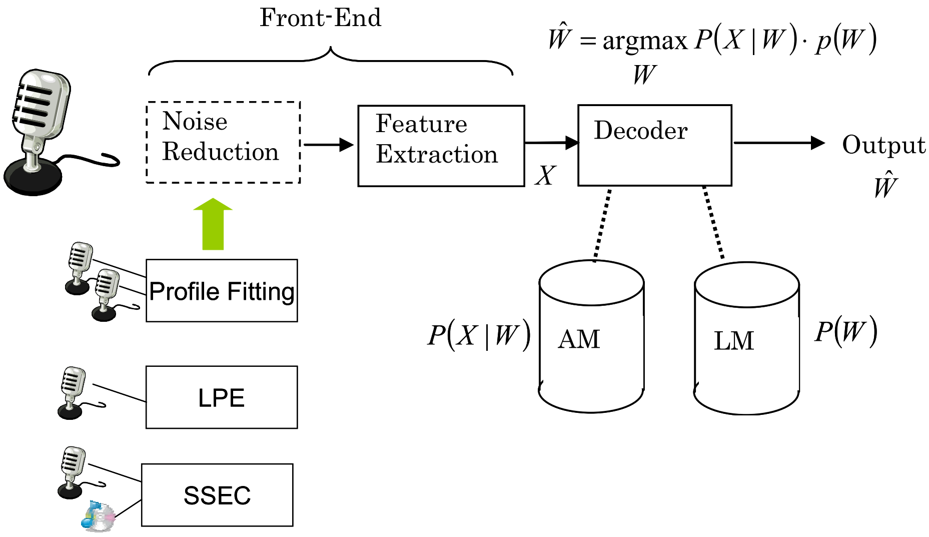


Figure 1.4. Integration of the proposed methods.

numbers of microphones is investigated, and the new method named Profile Fitting (PF) is proposed. In Chapter 4, PF is further discussed in an application of sound source localization. In Chapter 5, a new echo canceller named SSEC (Simultaneous adaptation of spectral Subtraction and Echo Cancellation) is proposed. In Chapter 6, a new speech enhancement method named Local Peak Enhancement (LPE) is proposed. Chapter 5 and Chapter 6 focus on the particular challenges in automobiles, but the proposed technologies are applicable for other applications. Finally, Chapter 7 summarizes this work and suggests future research directions.

# 2. Robust Automatic Speech Recognition

## 2.1. Introduction

In this chapter, we briefly review robust ASR to have a clear image of what robustness is, what is involved in robustness, and how the three novel methods proposed in this dissertation can contribute to the robustness of ASR systems.

## 2.2. Robustness of Automatic Speech Recognition System

As shown in Figure 2.1, ASR systems have been evolving to push against the following limitations:

- Environmental robustness
- Usability
- Speaking style
- Task complexity
- Speaker dependency

Speaking in the broad sense, advances in any of these areas affects the robustness of the ASR. The various kinds of research described in this dissertation are related to primarily to environmental robustness and usability. These topics will be further addressed in the following chapters.

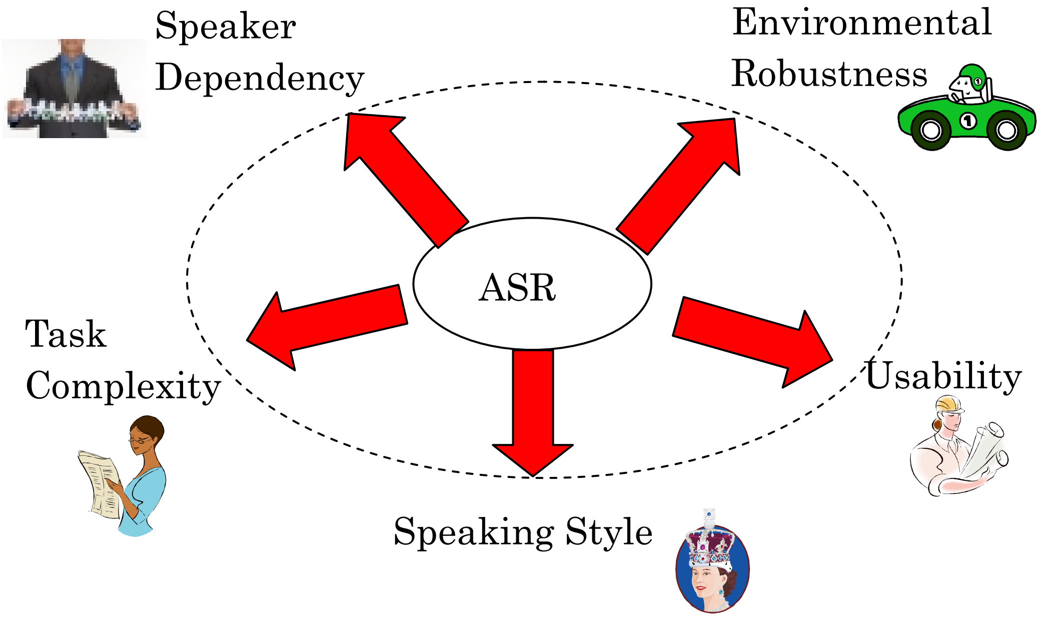


Figure 2.1. Evolution of automatic speech recognition.

### 2.2.1. Environmental robustness

Modern automatic speech recognition is based on statistical modeling from a large corpus. Therefore, the first steps involve training the acoustic model with realistic speech data. If the situation permits, the data should be recorded in the same environment as will be used and with the same conditions for types of noise, SNR, microphones, and room reverberation. This approach is known as multi-style training and is widely used in commercial products. However, the environments of the actual usage are highly unpredictable and usually there is some acoustic mismatch. Environmental robustness in speech recognition represents the ability to minimize performance degradation that occurs as a result of mismatches between system training and test conditions [Ros04].

Acoustic mismatches are classified into two types, additive noise and multiplicative distortion. Figure 2.2 illustrates how a speech signal is corrupted by these two factors. The variables *X*, *S*, N, and *H* denote the observed noisy speech signal, the clean speech signal, the observed noise signal, and the acoustic transmission function. The observed signal can be represented by Equation (2.1):

$$X = H \otimes S + N, \tag{2.1}$$

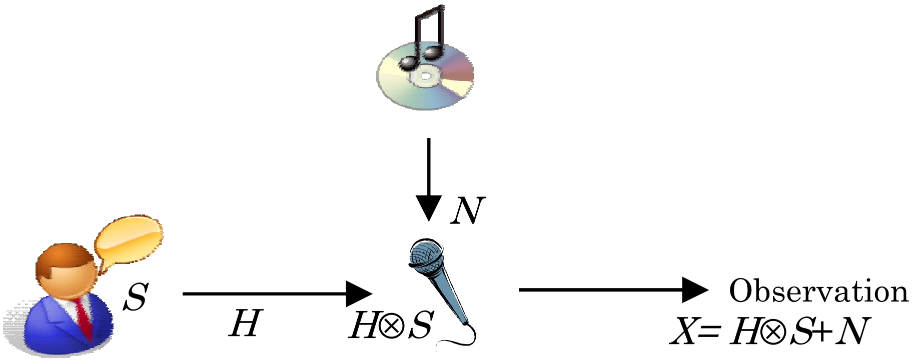


Figure 2.2. Transmission of a speech signal.

where ⊗ denotes the convolution operation. Distortion in the transmission channel is ignored here. As shown in Equation (2.1), $N$ is additive and $H$ is multiplicative. In automobiles, cruising noise, fan noise, passengers' voices, and sounds from radios or other audio devices are possible additive noises. Multiplicative distortion becomes conspicuous when the distance from the subject speaker to the microphone is large or when the transmission path involves reflection or diffusion. Also, a reverberant room significantly increases multiplicative distortion.

In order to minimize the acoustic mismatch that might be involved in realistic environments, these $N$ and $H$ components should be compensated in both the training and decoding. As long as the subject speaker is not moving relative to the microphone, Cepstrum Mean Subtraction (CMS) is a simple but effective method to compensate multiplicative distortion. CDCN can also compensate for it. However, these methods work within frames, and they cannot handle the later parts of the reverberations beyond the frame size. Therefore, some de-reverberation algorithms [SC00][NM03] can be used in very reverberant environments.

For additive noise, we should be aware that there are two type of additive noise, stationary noise (which is almost time-constant) and non-stationary noise (which is time-varying). Unless the SNR is very low, it is not difficult to compensate for stationary noise. We can subtract the estimated value of $N$ in Equation (2.1) in the power spectrum domain or in the log-power spectrum or cepstrum domain. When the SNR is very low, the compensation may be excessive and result in losing speech information or it may insufficient and leave too much residual noise. Therefore, other methods that not only reduce the noise but that also enhance the speech signal should be combined. We follow up on this idea in Chapter 6 of this dissertation.

For non-stationary noise, it is difficult to estimate time-varying $N$ in a reliable manner in a single channel system. Therefore, multiple-channel signal processing technologies such as microphone arrays and echo cancellation are pursued in Chapter 3 and Chapter 5 in this dissertation.

Instead of using front-end processing, the minimizing of acoustic mismatch can be done in the acoustic model. MLLR [LW95] and HMM Composition (PMC) [GY96] can be used for environmental adaptation. HMM Composition can be extended to support moving speakers to compensate for location dependent multiplicative distortions [Tak99].

Another approach for minimizing acoustic mismatch is to use robust features instead of MFCC. RASTA-PLP [HM94], multi-band spectral features [NSI+04], distinctive phonetic features [Fuk05], spectral peak-weighted liftering [KL00], SBCOR [KI95], and MVDR [DR01] are some of the candidates.

### 2.2.2. Usability

Usability involves various features of the ASR systems that make them more friendly for users.

Voice Activity Detection (VAD) has a large impact on usability. Most of the current car-navigation systems use a Push-To-Activate (PTA) mode, for which a user need to push a button when starting to speak. The end of the utterance is detected automatically by the car-navigation system. Using a Push-To-Talk (PTT) mode, the user would need to continue pressing the talk button while speaking. As the driver may need to turn the steering wheel, PTT is considered as an unacceptable scenario for driving. VAD typically utilizes information about the likelihood of speech, the speech power and also the sound directivity when a microphone array is used. The noise reduction methods described in Chapters 3, 5, and 6 of this dissertation can contribute to VAD by enhancing evaluation of the likelihood of speech and the speech power. Also, the sound source localization from Chapter 4 can contribute to the use of sound directivity information.

In near future automobiles, PTA mode will evolve to Always Listening (AL) mode, allowing the user to start speaking at any time without pushing a talk button. AL mode is already a common interface for humanoid robots. For cars, this will require higher VAD accuracy, as well as some language understanding technologies. Therefore, VAD improvements are important.

Another usability issue is the restriction that the car audio be stopped or the volume minimized upon pressing a talk button. Many current car-navigation systems have this restriction. The enhancement of echo cancellation technology discussed in Chapter 5 of this dissertation can contribute to addressing this problem.

In the near future, car-navigation systems will support a Barge-In interface that allows a user to start speaking even while the car-navigation system is broadcasting informative messages. This will require higher capabilities for the echo cancellation. Therefore, improvements in echo cancellation are important.

In the use of microphone arrays, reducing the number of microphones is also an improvement of the usability in terms of lower failure rates and lower costs for the equipment. In Chapter 3 in this dissertation, we pursue small-scale microphone arrays that outperform conventional approaches.

### 2.2.3. Speaking style

Speaking style affects pronunciation and thus the recognition. In general, discrete utterances such as commands and digits have less ambiguity. Continuous speech such as dictation has a lot of variation caused by co-articulation, stress, and so forth. Spontaneous speech has even more, because of the increased co-articulation, intonational phrasing, disfluencies, and speech repairs.

In order to support spontaneous speech in automatic speech recognition, many corpus-based approaches have been investigated. In Japan, a large corpus called the Corpus of Spontaneous Japanese (CSJ) was developed and many research projects have used it [Fru05].

### 2.2.4. Task complexity

Task complexity has a direct impact on ASR accuracy. If the task is a grammar-based task with a small vocabulary, we can probably expect higher ASR accuracy, but less flexibility in the acceptable expressions. If the task involves transcription with a large vocabulary, Large Vocabulary Continuous Speech Recognition (LVCSR), it is not easy to achieve higher ASR accuracies, because there are vastly more variations in what can be said.

In the LM context, task complexity can be measured by perplexity. An LM with lower perplexity is considered to be a good LM. To build such an LM, it should be trained with a text corpus well-matched to the actual usage. In that sense, an LM with lower perplexity tends to be task-specific, resulting in a narrow scope. Therefore, some topic detection algorithm or topic adaptation algorithm are sometimes used to switch or adapt the prepared LMs according to the detected current task.

### 2.2.5. Speaker dependency

Historically, original ASR systems were speaker dependent. The AM had to be trained with each speaker’s voice. Now, it is common to train the AM with a large speech corpus containing various subject speakers’ voices. It is called the speaker-independent model.

However, there are still some speakers who have poor results with ASR systems that use the speaker-independent model. Model adaptation technologies such as MAP, VTLN [LR96], and MLLR [LW95] are effective to boost the accuracy. Also, some canonical modeling such as SAT [AMS+96] is known to increase the benefit of model adaptation.

Though it is not a speaker adaptation, it is worth noting that discriminative training such as MPE [PW02] generally boosts the accuracy in ASR including such problematic speakers.

# 3. Noise Reduction by Profile Fitting Method

## 3.1. Introduction

Previous research estimated that more than 50 microphones are required to achieve high performance for automatic speech recognition using microphone arrays at distances of 1 m [Elk01]. This also requires a special interface to enable simultaneous multi-channel audio input. However this requirement would not be acceptable for many consumer products like mobile PCs, PDAs, etc.

On the other hand, the directional pattern formed by a small-scale microphone array such as a 2-channel (left and right) Delay and Sum beamformer is not sufficiently focused on the target. This means the output of the beamformer will contain too much noise arriving from other directions, and therefore additional logic to estimate and subtract the noise signal mixed in the output of the beamformer is essential.

The basic concept was provided by the Griffiths-Jim-type adaptive beamformer [GJ82]. It can be implemented in spectral subtraction as Spatial Subtraction Array (SSA) [ONS+05]. It subtracts sub-beamformer output from main-beamformer output in power spectrum domain. In this dissertation, two-channel version of SSA is also referred as 2-channel (Adaptive) SS. This logic is shown in Figure 3.1 or Figure 3.2. The main-beamformer forms a directivity pattern focused on the target direction and the sub-beamformer forms a directional null on the target. In Figure 3.1, the output of the sub-beamformer is assumed to be the noise power and it is simply subtracted from the output of the main-beamformer [INK97]. In this chapter, we call this method "Two-Channel Spectral Subtraction (2-channel SS)."

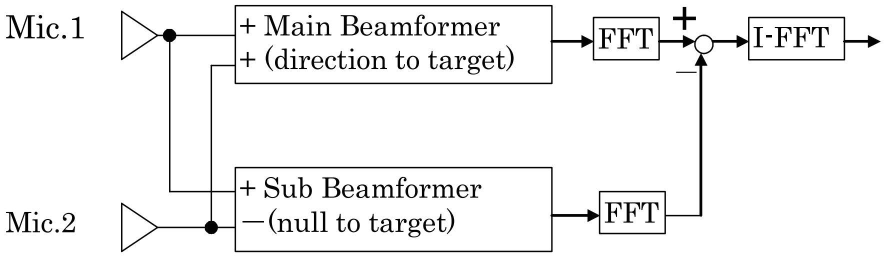


Figure 3.1. Two-channel Spectral Subtraction.

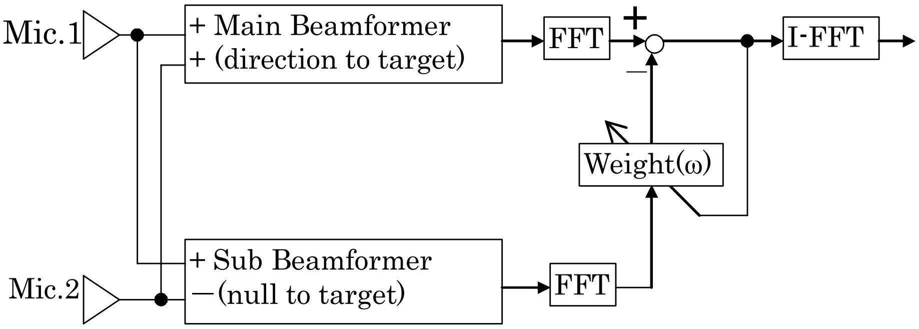


Figure 3.2. Two-channel Adaptive Spectral Subtraction.

In order to improve the performance, the level of the noise power to be subtracted should be estimated more accurately. Figure 3.2 shows one of the solutions, which estimates the subtraction weight at each frequency adaptively using an LMS algorithm so as to minimize the output when the target sound is absent (i.e. when only the noise is active). In this chapter, we call this method "Two-Channel Adaptive Spectral Subtraction (2-channel Adaptive SS)." This method is equivalent to the method of Kim et al. [KAS+96] except for the online adaptation capability. Saruwatari et al. described the weight as complementary weight vectors at the microphone elements to provide twice the directional nulls compared to the conventional adaptive beamformer [SKT+00]. Mizumachi et al. added a third microphone to detect the arrival direction of the noise so the weight can be estimated analytically without relying on the adaptation [MA98].

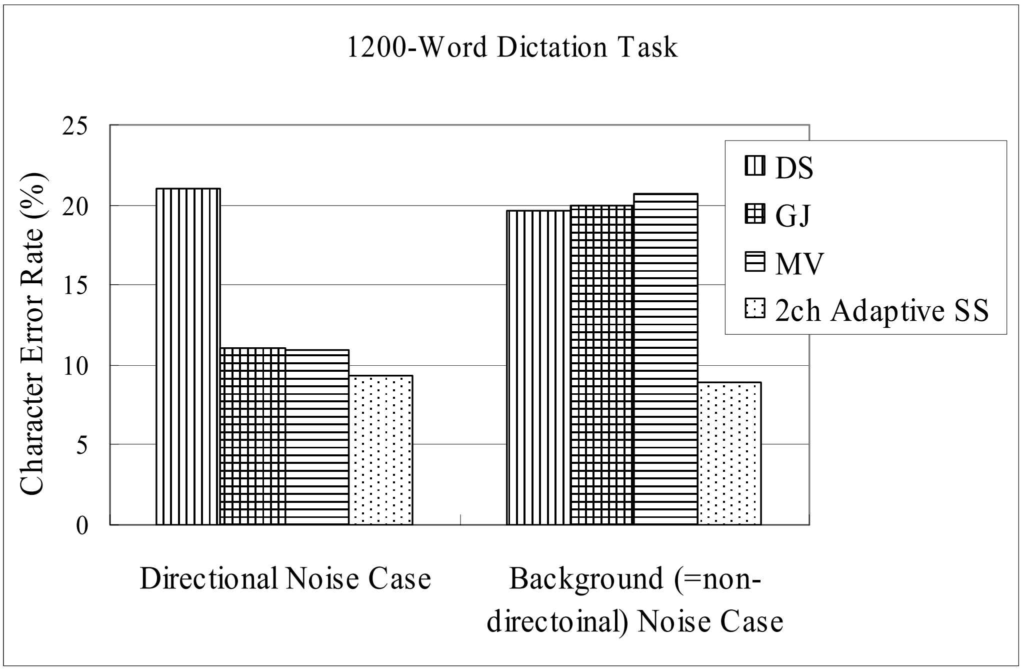


Figure 3.3. Preliminary experimental results of automatic speech recognition on data from two microphone processed by various conventional methods. The experimental data was prepared from simulations with impulse responses in RWCP and recordings of 125 utterances by a male speaker. Noise data for the Exhibition Hall in the Denshi-Kyo DB was overlapped at an SNR of 20 dB. The distance between the microphones was 11.3 cm. The distance to the subject speaker was 2 m. The Acoustic Model was trained with clean speech.

These methods are much more effective compared with the conventional beamformers such as Delay and Sum (DS) and Minimum Variance (MV) as shown in Figure 3.3 of our preliminary investigation. However, the performance is not still optimized, because they rely on the information from only the 2 points that are the focal directions of the main beamformer and the sub-beamformer. In this chapter, an optimized method that utilizes the information from all spatial directions is proposed. This approach makes the noise estimation more accurate and it provides a reasonable solution for the case of multiple noise sources, for which 2-channel Adaptive SS is not well adapted.

## 3.2. Two-Channel Adaptive Spectral Subtraction Method and Problem

In Figure 3.2, the beamformer output ($Z_{\omega,T}$) before the I-FFT can be written as Equation (3.1):

$$\begin{aligned} Z_{\omega,T} &= M1_{\omega,T} - W_{\omega} \cdot M2_{\omega,T} \\ &= S_{\omega,T} + \{N_{\omega,T} - W_{\omega} \cdot M2_{\omega,T}\}, \end{aligned} \tag{3.1}$$

where the index $\omega$ is a frequency for each sub-band, and the index $T$ is a time frame number. The variable $M1_{\omega,T}$ represents the power spectrum of the main beamformer output. The variable $M2_{\omega,T}$ represents the power spectrum of the sub beamformer output. The variable $S_{\omega,T}$ represents the signal power and the variable $N_{\omega,T}$ represents the noise power in the main beamformer output. The variable $W_{\omega}$ is the subtraction weight parameters that minimize the following Equation (3.2):

$$V_{\omega} = E[\{N_{\omega,T} - W_{\omega} \cdot M2_{\omega,T}\}^2], \tag{3.2}$$

where the expectation operation denoted by $E$[ ] should be performed only when the signal is absent. Using this adaptation, we can minimize the noise power in the beamformer output of Equation (3.1).

Figure 3.4 shows an example of the weight parameters $W_{\omega}$ adapted for a single noise source in a non-reverberant environment. We see the weight value is very large in certain frequency ranges. In those ranges, the variance of the remaining noise power defined by the Equation (3.2) is large. As a result, the beamformer output will have more remaining noise. That banded distortion of the output power spectrum causes an adverse affect on the mel-cepstrum coefficients (MFCC) that are used in automatic speech recognition.

For multiple noise sources, another issue arises with the 2-channel Adaptive SS. Let’s assume that there are two noise sources around the signal source. Adaptation of the weight parameters will be performed over long intervals in an averaging manner. However the noise sources are not necessarily stationary. In the frame-by-frame view, one of the noise powers may sometimes be zero or very small. Therefore, the weight parameters are not always consistent for all the time frames.

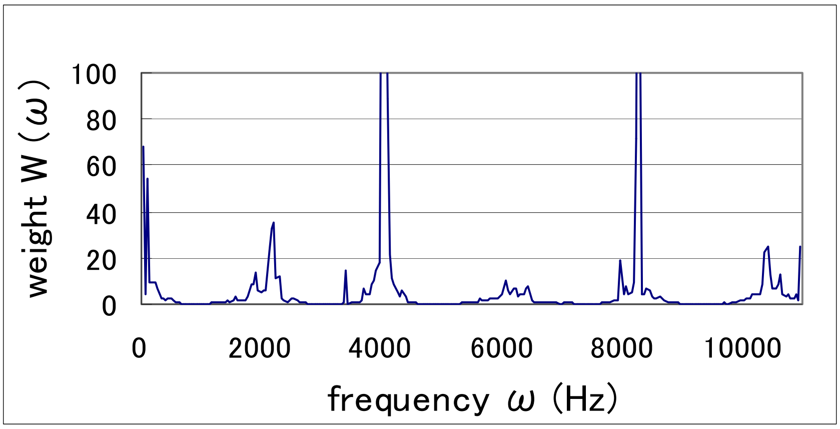


Figure 3.4. An example of the weight parameters (single noise case). The distance between the two microphones is 30 cm.

Experiments in the later section show that the resulting SNR (Signal Noise Ratio) of the 2-channel Adaptive SS is acceptable, but the output has considerable distortion as remarked above.

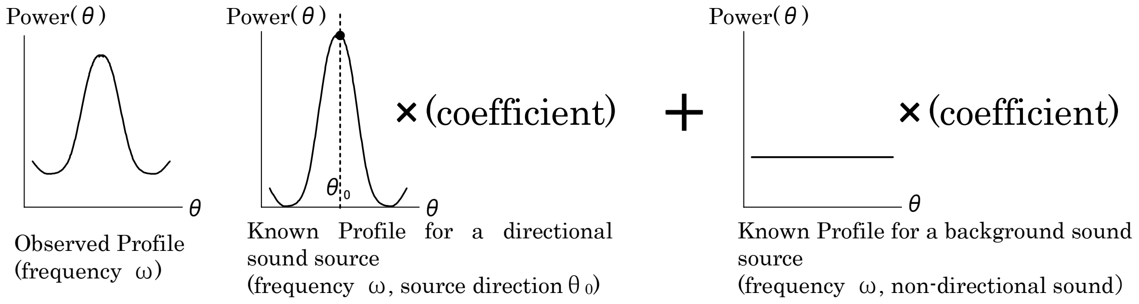
## 3.3. Proposed Method (Profile Fitting)

In order to support multiple non-stationary noise sources, a new method that does not rely on adaptation is proposed. Instead, we introduce the critical assumption that the locations of the noise sources are known.

Our proposed method is based on the information in a profile consisting of a series of points. Figure 3.8 and Figure 3.9 show some examples of profiles, which are power distribution patterns observed at varying look directions $\theta$ for a Delay and Sum beamformer.

Profiles are measured at each frequency. In general, they have specific peaks corresponding to a sound source direction. If the sound source is non-directional or it involves distinct reflections in the transmission paths, the profile does not have steep peaks and valleys. If the sound source is in near field, the two microphones often involve gain imbalance so to make the profile have a bias value.

MV beamformer sometimes uses similar measures associated with each candidate location of sound source, known as steering vector. But, it only contains time delay



Observed Profile
(frequency ω)

Known Profile for a directional sound source
(frequency ω, source direction θ0)

Known Profile for a background sound source
(frequency ω, non-directional sound)

Figure 3.5. Decomposition of power distribution pattern.

information and distant information, based on a simple sound source assumption. On the other hand, the profile can include the effect of reflection and diffusion.

Since the locations of the sound sources are known, we can prepare the profiles a priori. Observed sound signals can be decomposed into linear combinations of the profiles on a frame-by-frame basis. This approach is still valid even if one (or more) of the noise sources is intermittently inactive.

For a single noise source, the proposed method will also have banded distortion due to the aliasing we experienced with the 2-channel Adaptive SS. However the distortion is more moderate in favor of the decomposition process that utilizes the information from all spatial directions.

The first step is to prepare "known profiles." We need to imagine placing a sound source in a possible direction and measuring its power distribution profiles ($P_{\omega}(\theta_0,\theta)$, $Q_{\omega}(\theta)$) for this microphone array at each frequency $\omega$ by using white noise or any standard signal. $P_{\omega}(\theta_0,\theta)$ represents a profile for a directional sound source in the direction $\theta_0$, while $Q_{\omega}(\theta)$ represents a profile for a non-directional background sound source. After making these measurements, they should be normalized so that the area of the pattern at each frequency is equal to 1, because the shape is the only essential information. These shapes are considered as the characteristics of the microphone array. They do not represent any acoustic features of the target signal or the noise signals. They are referred to as the known profiles.

The next step is to work with an "observed profile" for each time frame $T$. When an observed sound signal can be assumed to consist of a directional target sound and a non-directional background noise as in Figure 3.5, the observed profile $X_{\omega,T}(\theta)$ can be approximately represented as the weighted sum of the two known profiles as Equation (3.3):

$$X_{\omega,T}(\theta) \cong \alpha_{\omega,T} \cdot P_{\omega}(\theta_0,\theta) + \beta_{\omega,T} \cdot Q_{\omega}(\theta), \tag{3.3}$$

where we assumed there is no correlation between the target sound and the noise. The variable $\alpha_{\omega,T}$ is a weight coefficient for a profile $P_\omega$ and $\beta_{\omega,T}$ is a weight coefficient for a profile $Q_\omega$ . These coefficients can be determined so as to minimize the following evaluation function $\Phi_{\omega,T}$:

$$\Phi_{\omega,T} = \int_{\min\_\theta}^{\max\_\theta} \left\{X_{\omega,T}(\theta) - \alpha_{\omega,T} \cdot P_{\omega}(\theta_0,\theta) - \beta_{\omega,T} \cdot Q_{\omega}(\theta)\right\}^2 d\theta. \tag{3.4}$$

The values of $\alpha_{\omega,T}$ and $\beta_{\omega,T}$ can be determined from $\partial\Phi_{\omega,T} / \partial\alpha_{\omega,T} = 0$ and $\partial\Phi_{\omega,T} / \partial\beta_{\omega,T} = 0$ under the following constraints:

1) $\alpha_{\omega,T} \geq 0$.

2) $\beta_{\omega,T} \geq 0$.

3) $\alpha_{\omega,T} \leq X_{\omega,T}(\theta_0) / P_{\omega}(\theta_0,\theta_0)$.

The conditions 1) and 2) mean the power should not be negative. Condition 3) means the output should be less than the observed power as the noise is reduced.

$$\frac{\partial\Phi_{\omega,T}}{\partial\alpha_{\omega,T}} = -2 \cdot \int_{\min\_\theta}^{\max\_\theta} P_{\omega}(\theta_0,\theta) \cdot \left\{X_{\omega,T}(\theta) - \alpha_{\omega,T} \cdot P_{\omega}(\theta_0,\theta) - \beta_{\omega,T} \cdot Q_{\omega}(\theta)\right\} d\theta = 0 . \tag{3.5}$$

$$\frac{\partial\Phi_{\omega,T}}{\partial\beta_{\omega,T}} = -2 \cdot \int_{\min\_\theta}^{\max\_\theta} Q_{\omega}(\theta) \cdot \left\{X_{\omega,T}(\theta) - \alpha_{\omega,T} \cdot P_{\omega}(\theta_0,\theta) - \beta_{\omega,T} \cdot Q_{\omega}(\theta)\right\} d\theta = 0 . \tag{3.6}$$

Equations (3.5) and (3.6) can be expressed in a matrix and vectors as Equation (3.7).

$$\mathbf{C}_{\omega,T} = \mathbf{A}_{\omega} \cdot \mathbf{B}_{\omega,T}, \tag{3.7}$$

where $\mathbf{A}_\omega$, $\mathbf{B}_{\omega,T}$, and $\mathbf{C}_{\omega,T}$ are defined as follows:

$$\mathbf{A}_{\omega} = \begin{bmatrix} \int_{\min\_\theta}^{\max\_\theta} P_{\omega}(\theta_0,\theta) \cdot P_{\omega}(\theta_0,\theta) d\theta & \int_{\min\_\theta}^{\max\_\theta} Q_{\omega}(\theta) \cdot P_{\omega}(\theta_0,\theta) d\theta \\ \int_{\min\_\theta}^{\max\_\theta} P_{\omega}(\theta_0,\theta) \cdot Q_{\omega}(\theta) d\theta & \int_{\min\_\theta}^{\max\_\theta} Q_{\omega}(\theta) \cdot Q_{\omega}(\theta) d\theta \end{bmatrix} . \tag{3.8}$$

$$\mathbf{B}_{\omega,T} = \begin{bmatrix} \alpha_{\omega,T} \\ \beta_{\omega,T} \end{bmatrix} . \tag{3.9}$$

$$\mathbf{C}_{\omega,T} = \begin{bmatrix} \int_{\min\_\theta}^{\max\_\theta} P_\omega(\theta_0,\theta)\cdot X_{\omega,T}(\theta)d\theta \\ \int_{\min\_\theta}^{\max\_\theta} Q_\omega(\theta)\cdot X_{\omega,T}(\theta)d\theta \end{bmatrix}. \quad (3.10)$$

The values of $\alpha_{\omega,T}$ and $\beta_{\omega,T}$ can be determined from Equation (3.11):

$$\mathbf{B}_{\omega,T} = \mathbf{A}_\omega^{-1}\cdot \mathbf{C}_{\omega,T}. \quad (3.11)$$

If $\beta_{\omega,T}$ is less than 0, the variable $\beta_{\omega,T}$ should be set to 0. In this case, $\mathbf{A}_\omega$ and $\mathbf{C}_{\omega,T}$ should be modified per Equations (3.12) and (3.13), and $\mathbf{B}_{\omega,T}$ should be re-calculated using Equation (3.11).

$$\mathbf{A}_\omega = \begin{bmatrix} \int_{\min\_\theta}^{\max\_\theta} P_\omega(\theta_0,\theta)\cdot P_\omega(\theta_0,\theta)d\theta & 0 \\ 0 & 1 \end{bmatrix}. \quad (3.12)$$

$$\mathbf{C}_{\omega,T} = \begin{bmatrix} \int_{\min\_\theta}^{\max\_\theta} P_\omega(\theta_0,\theta)\cdot X_{\omega,T}(\theta)d\theta \\ 0 \end{bmatrix}. \quad (3.13)$$

Finally, the variable $\alpha_{\omega,T}$ should be adjusted as follows:

$\alpha_{\omega,T} = 0$ if $\alpha_{\omega,T} < 0$ .

$\alpha_{\omega,T} = X_{\omega,T}(\theta_0) / P_\omega(\theta_0,\theta_0)$ if $\alpha_{\omega,T} > X_{\omega,T}(\theta_0) / P_\omega(\theta_0,\theta_0)$ .

Now we can determine the power of the enhanced speech signal $Z_{\omega,T}$ at the frequency $\omega$ for that time frame $T$ as Equation (3.14):

$$Z_{\omega,T} = \alpha_{\omega,T}\cdot P_\omega(\theta_0,\theta_0). \quad (3.14)$$

In the process above, the noise power estimated as $\beta_{\omega,T}\cdot Q_\omega(\theta_0)$ is actually subtracted from the observed power $X_{\omega,T}(\theta_0)$, which is the output of the conventional Delay and Sum beamformer.

The observed profiles $X_{\omega,T}(\theta)$ are obtained for each time frame (every 10–20 ms). The above decomposition should be done for every time frame $T$ and at every frequency $\omega$.

If there is not only a background noise but also a directional noise arriving from the direction $\theta_1$, we can add the profile for the directional sound source as $R_\omega(\theta_1,\theta)$ with coefficient $\gamma_{\omega,T}$ to the right hand side of the Equation (3.3), producing Equation (3.15).

The additional term accounts for the power distribution from the additional noise source. Additional known profiles can be added if there are more known noise sources.

$$X_{\omega,T}(\theta) \cong \alpha_{\omega,T} \cdot P_{\omega}(\theta_0,\theta) + \beta_{\omega,T} \cdot Q_{\omega}(\theta) + \gamma_{\omega,T} \cdot R_{\omega}(\theta_1,\theta). \tag{3.15}$$

Similar to the two-profile case, the negative value check should be done for each coefficient. First, the coefficient $\beta_{\omega,T}$ or $\gamma_{\omega,T}$ should be checked. If it is negative, it should be set to zero and all other coefficients should be recalculated per Equations (3.12) and (3.13). The coefficient $\alpha_{\omega,T}$ should be checked last.

## 3.4. Spectral Smoothing and Inverse Smoothing

When the SNR of the observed sound is small, the correlation term omitted in Equation (3.3) cannot be ignored. Because no profiles are available for the correlation term, the decomposition becomes inaccurate.

Kitaoka et al. proposed SMT for spectral smoothing over the time dimension for single channel spectral subtraction in order to minimize the correlation term [KAN01]. We applied this technique to Profile Fitting using Equation (3.16):

$$\overline{X}_{\omega,T}(\theta) = \sum_{t=0}^{L-1} c_t \cdot X_{\omega,T-t}(\theta), \tag{3.16}$$

where $\overline{X}_{\omega,T}(\theta)$ is the smoothed observed profile, the $c_t$ are the smoothing coefficients, and $L$ is the smoothing width. When SMT is applied, $\overline{X}_{\omega,T}(\theta)$ should be used instead of $X_{\omega,T}(\theta)$ in Equations (3.3), (3.4), (3.5), (3.6) and the follow-on equations.

As a side effect of SMT, the output of the enhanced speech signal is also smoothed**.** This means the dynamic features detected by automatic speech recognition will be affected by SMT. In order to compensate for this, we used I-SMT (Inverse SMT) with a limiter for stability. When I-SMT is used, the enhanced speech signal $Z_{\omega,T}$ can be obtained as follows:

$$\overline{Y}_{\omega,T} = a_{\omega,0,T} \cdot P_{\omega,0}(\theta_0). \tag{3.17}$$

$$Y_{\omega,T} = \max\left[0, \quad \frac{1}{c_0}\left\{\overline{Y}_{\omega,T} - \sum_{t=1}^{L-1} c_t \cdot Y_{\omega,T-t}\right\}\right]. \tag{3.18}$$

$$Z_{\omega,T} = \min\left[X_{\omega,T}(\theta_0), \quad Y_{\omega,T}\right]. \tag{3.19}$$

Above SMT and I-SMT operation is an option for Profile Fitting. In our experiments at moderate SNRs higher than 15 dB described in this chapter, we found the accuracy of the automatic speech recognition was not degraded without SMT and I-SMT operation, although the resulting SNR was sometimes degraded. Therefore, SMT and I-SMT operation was not used in our experiments in this chapter.

In this chapter, the SNR was measured simply by the power histogram from Equation (3.20):

$$SNR = 10 \cdot \log_{10}(S_{max} / N_{mod}). \qquad (3.20)$$

The variable $N_{mod}$ is the mode value in the noise power histogram, and $S_{max}$ is the 90th percentile value above the mode value in the signal power histogram.

## 3.5. Preliminary Experiment

Before evaluating the proposed method in automatic speech recognition, we briefly checked the distortion of the beamformer output from the original sound. We defined the MFCC distance (MCEP) in Equation (3.21) as the measurement of the distortion.

$$MCEP = \left[ \frac{1}{N_{cep}} \sum_{i=1}^{N_{cep}} \left\{ C_{out}(i) - C_{original}(i) \right\}^2 \right]^{\frac{1}{2}}, \qquad (3.21)$$

where *C(i)* is the i-th mel-cepstrum, and $N_{cep}$ is the number of the mel-cepstrum except *C(0)*. We set $N_{cep}$ = 23 for the sampling rate 22.05 kHz so that it would be consistent with the decoder of the automatic speech recognition program. The subscript "*out*" means the output of the beamformer and "*original*" means the original sound without adding noise.

### 3.5.1. Preliminary experiment stationary noise case

We placed two microphones at a distance of 30 cm in the soundproof chamber. The arrival angle of the target signal was 0° (directly in front), and the distance was 15 cm. As a directional noise, white noise was played back at an arrival angle of +40° (right side) at a distance of 1 m. The SNR was 18.3 dB.

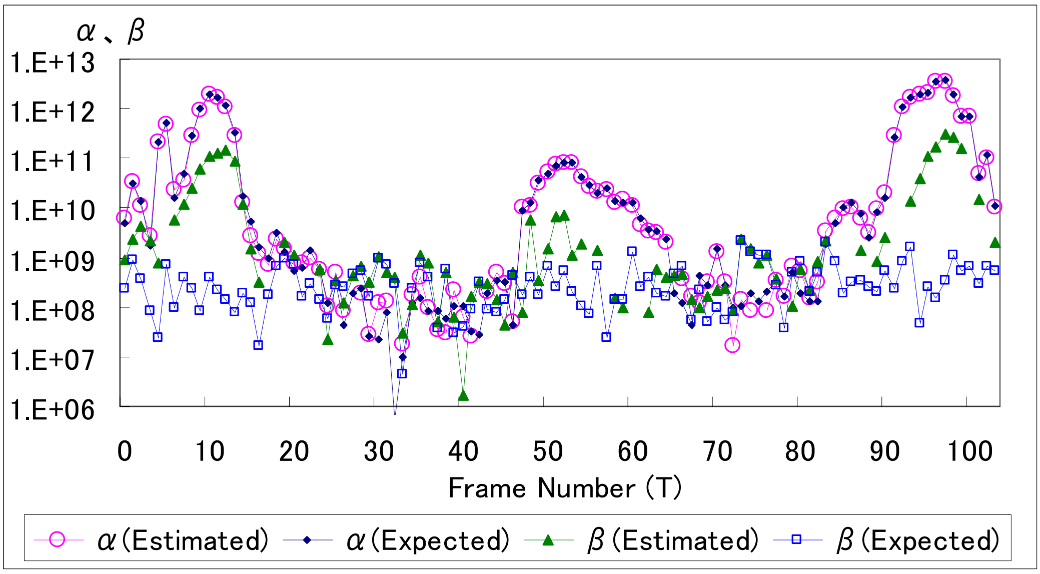


Figure 3.6. Estimated and expected coefficients at **ω**=600 Hz.

For Profile Fitting, two profiles were used, each associated with one directional sound source. Figure 3.6 shows the estimated and expected coefficients at $\omega$ = 600 Hz. The expected values were calculated using the separated signal and noise. When the target signal is active (i.e. when the expected value of $\alpha_{\omega,T}$ is large), the estimated $\alpha_{\omega,T}$ matches well to the expected $\alpha_{\omega,T}$. The estimated $\beta_{\omega,T}$ seems to be affected by the large $\alpha_{\omega,T}$, but the absolute value of the estimated $\beta_{\omega,T}$ is still very small compared with the estimated $\alpha_{\omega,T}$, since Figure 3.6 is plotted with a logarithmic scale. When the target signal is not active, the estimated value of $\beta_{\omega,T}$ matches well to the expected $\beta_{\omega,T}$.

Table 3.1 shows the averaged MFCC distance. It shows the 2-channel Adaptive SS has a larger MFCC distance than Profile Fitting.

### 3.5.2. Non-stationary noise case

In addition to the configuration in the previous paragraph, we added a directional noise source playing back white noise at an arrival angle of -50° (left side) at a distance of 1 m. In order to simulate the worst case of multiple non-stationary noise sources, the new noise source was stopped in the entire speech period, and both noise sources were active during the adaptation of the 2-channel Adaptive SS. Profile Fitting does not require any

Table 3.1. MFCC distance between clean speech and output of beamformer. Sample utterance is /oi henjishiro/ in Japanese (stationary case)

| | MCEP |
|---|---|
| 2-channel Adaptive SS | 16.8 |
| Profile Fitting | 11.6 |

Table 3.2. MFCC distance between clean speech and output of beamformer. Sample utterance is /oi henjishiro/ in Japanese (non-stationary case)

| | MCEP |
|---|---|
| 2-channel Adaptive SS | 22.6 |
| Profile Fitting | 9.6 |

adaptation. Instead, three profiles were prepared, each associated with one directional sound source. The SNR without the additional noise source was 20.5 dB.

Table 3.2 shows the averaged MFCC distance. The advantage of Profile Fitting over the 2-channel Adaptive SS is more evident than for the single stationary noise case.

# 3.6. Experiment in Automatic Speech Recognition

## 3.6.1. Non-reverberant environment

Figure 3.7 shows the configuration for this test. The distance between the two microphones was 30 cm. The speech recognition task was a transcription of a robot conversation (size of vocabulary = 1,200, test set perplexity = 9.2). Two sets of 125

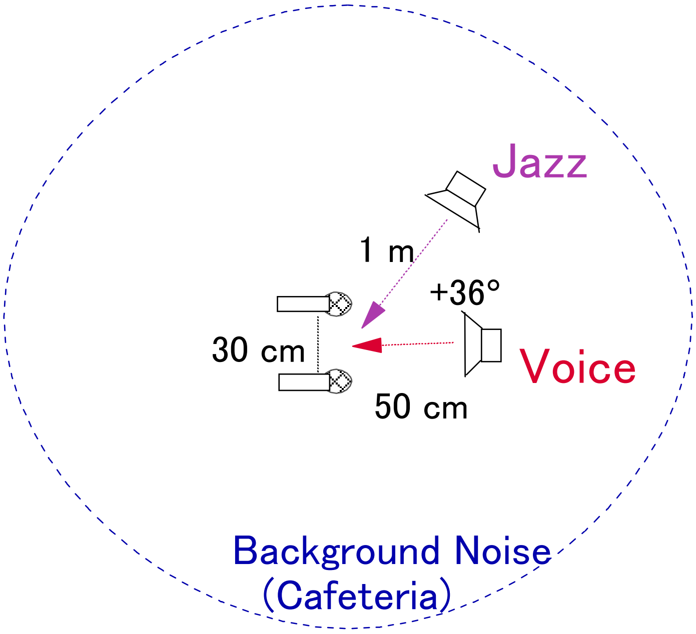


Figure 3.7. Testing configuration (in soundproof chamber).

sentences, each spoken by a male speaker and a female speaker in our soundproof chamber, were used for the evaluation as the target signal. The arrival angle of the target signal was 0° (directly in front), and the distance was 50 cm. Jazz music as a directional noise was recorded in the soundproof chamber. The arrival angle was 36° (right side), and the distance was 1 m. The background noise was recorded separately at lunchtime in our cafeteria. Those recorded noises were mixed with the target signal data manually so that the SNR could be controlled.

The sampling frequency of the audio stream was 22.05 kHz. The frame shift was 10 ms. The windowing function was a Hamming Window. The FFT width was 512 samples. Profiles were measured by the Delay and Sum beamformer in the time domain. The horizontal axis of the profiles represents the time delay measured as the number of delayed samples. Here, this value corresponds to the look direction of beamformer. This was varied from the $-\max$ to the $+\max$ value at every specified step value. For lower frequency profiles (< 1 kHz), we used a 5 times larger maximum value and a bigger step value to acquire the whole pattern, since the shapes have gentle slopes at those frequencies. For higher frequency profiles, we used the original maximum value and the minimum step value (= 1 sample) to be more accurate.

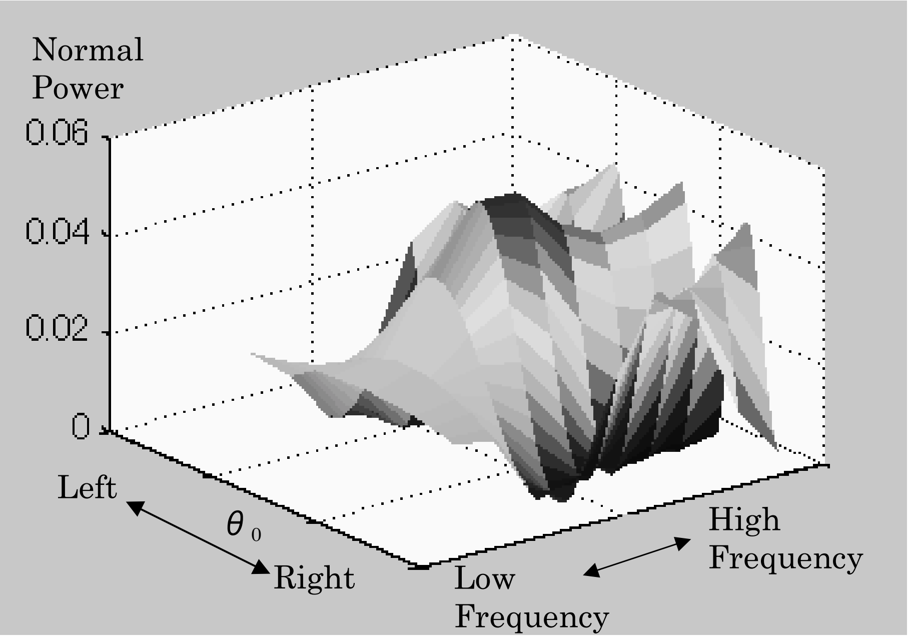


Figure 3.8. An example of a profile for a directional sound source.

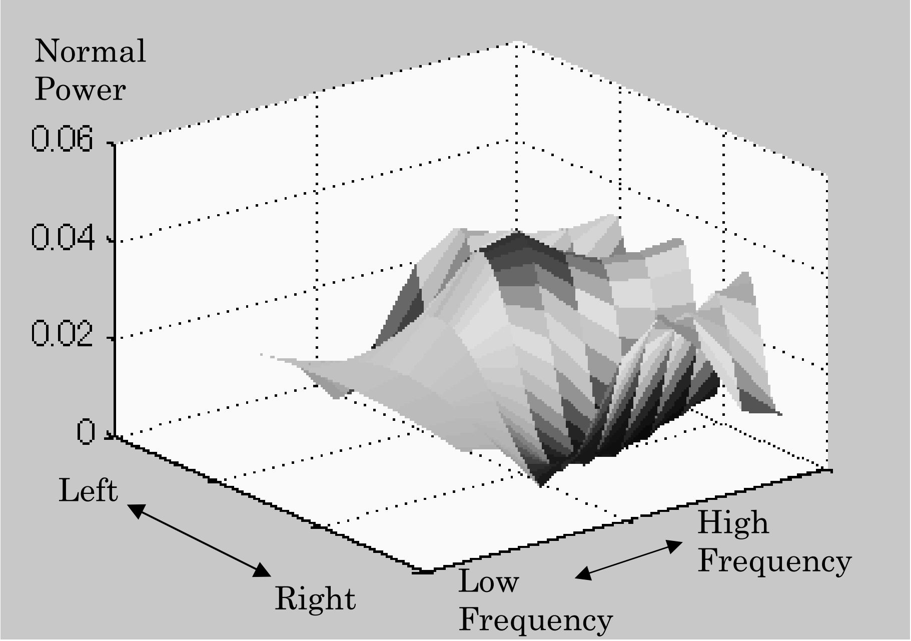


Figure 3.9. An example of a profile for a background sound source.

Figure 3.8 shows the actual profile for the directional sound source at 0°/50 cm. Figure 3.9 shows one for a non-directional background sound source. Although it is not shown here, the profile for the directional sound source at 36°/1 m was also used in the decomposition process. In total, three profiles were used in this experiment, two for the directional sound sources, and one for the background noise source. The decomposition is defined by Equation (3.15). As the directivity of the background noise is not rigorously determined, the profile for the background noise is only valid in terms of the average. In other words, it was introduced for an approximate solution. In general, if there is a distinct noise source in the background, it should be defined separately as a directional noise source.

These profiles were measured by using white noise in our soundproof chamber before the experiment began. As shown in Figure 3.9, the profiles are not completely flat even for a non-directional sound, because of the directivity pattern of the unit microphone.

The test cases are as follows:

1) Only a background noise was added
2) Only a directional noise was added
3) Both a background noise (reduced to 83%) and a directional noise (reduced to 40%) were added

All the three profiles were used for all of the test cases.

The SNR for each case was almost constant around 20 dB. The SNR of the original signal without adding noise was 31.4 dB for the male speaker and 36.0 dB for the female speaker.

We measure the error rate with CER (Character Error Rate), because the evaluation task is transcription and there is some ambiguity in word segmentation in Japanese as an agglutinating language. The definition of CER is in Equation (3.22).

$$CER = \frac{(\text{number of substituted characters}) + (\text{number of inserted characters}) + (\text{number of deleted characters})}{(\text{number of all expected characters})} \ , \tag{3.22}$$

The CER measured using only the left channel of the original signal without adding noise was 3.3% for the male speaker and 5.4% for the female speaker.

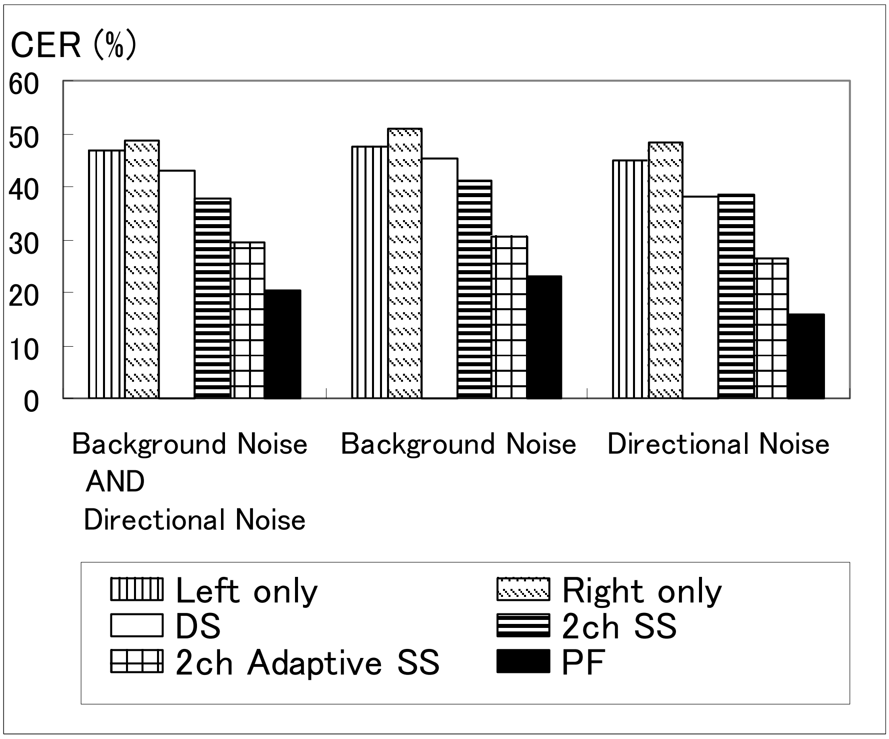


Figure 3.10. Resulting character error rate (in soundproof chamber).

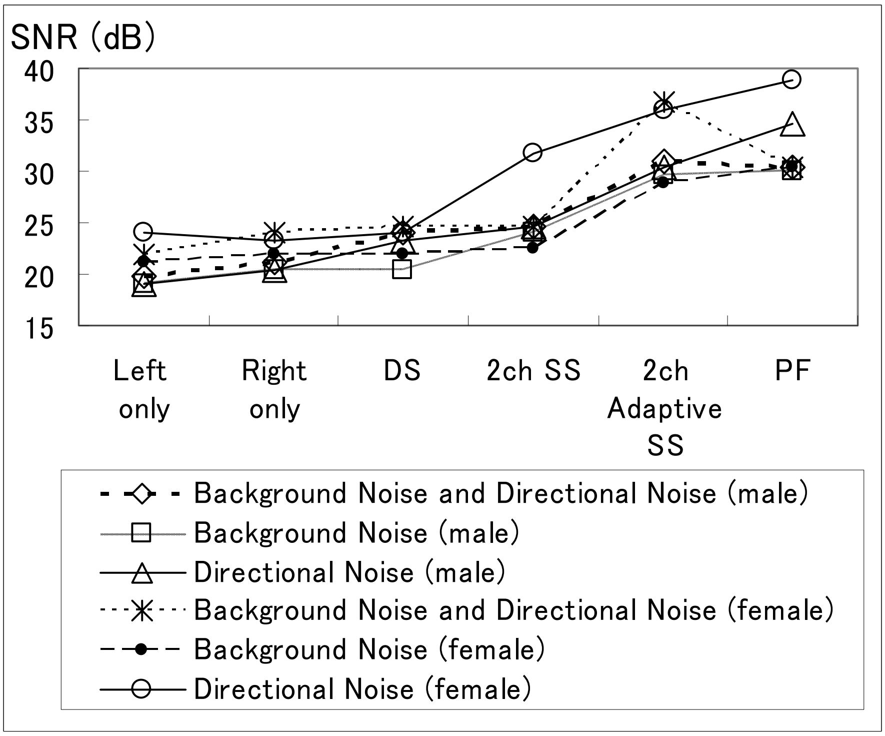


Figure 3.11. Resulting signal-to-noise ratios (in soundproof chamber).

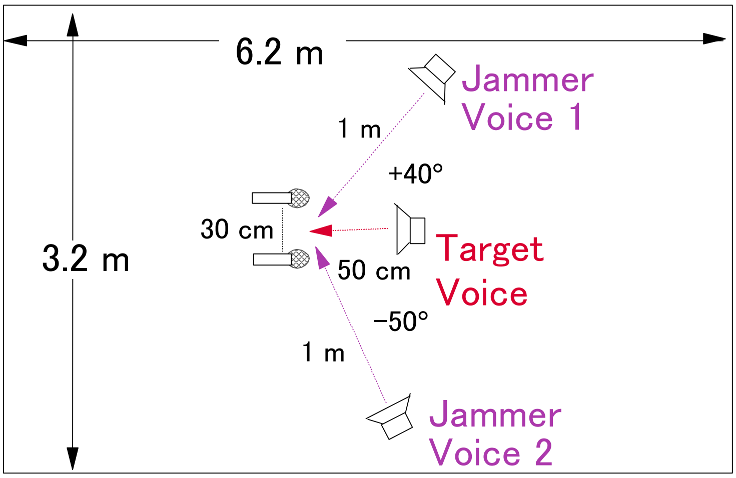


Figure 3.12. Testing configuration (in meeting room).

The conventional methods to be compared were chosen as follows:

1) No beamformer (Left or Right only)
2) Delay and Sum (DS)
3) Two-Channel Spectral Subtraction (2-ch SS)
4) Two-Channel Adaptive Spectral Subtraction (2-ch Adaptive SS)

Figure 3.10 and Figure 3.11 show the resulting CERs and SNRs, respectively, for this experiment. The speech recognition was done only when the target signal was active. Compared with the conventional methods, Profile Fitting (PF) shows superior performance for CER. Generally speaking, CER was reduced by more than 20% from the best result of the conventional beamformers (2-ch Adaptive SS). The SNR was almost the same as for 2-channel Adaptive SS.

## 3.6.2. Realistic environment

We also evaluated the performance in a more realistic environment. Figure 3.12 shows the testing configuration in our meeting room with a reverberant time of 0.22 seconds. The geometry of the microphone array, the tested recognition task and the signal processing parameters were the same as in the previous test.

Two sets of 125 sentences, each spoken by a male speaker and a female speaker were played back using a loudspeaker located at the angle of 0° (directly in front) and at the distance of 50 cm. The jammer voices were two human speeches by a male and a female speaker respectively, and they were played back continuously using 2 loudspeakers that were located at the angles of +40° and –50°, both at a distance of 1 m. In the decomposition process, we used three profiles for the directional sound sources of +40° at 1 m, 0° at 50 cm, and -50° at 1 m. These profiles were measured in the meeting room before the experiment began.

Figure 3.13 and Figure 3.14 show the resulting CERs and SNRs, respectively, for this experiment. The speech recognition was done only when the target signal was active. Profile Fitting (PF) reduced the CER by approximately 11% from the best result of the conventional beamformers (2-ch SS). The extent of the improvement was relatively smaller than in the previous experiment. The SNR for Profile Fitting was almost the same as for the 2-channel Adaptive SS.

## 3.7. Concluding Remarks

The proposed method focuses on the power distribution profile of a microphone array to decompose an observed profile into some known profiles so as to extract the target signal only.

Experiments in a non-reverberant environment with a dictation system configured with 2 microphones showed the proposed method (Profile Fitting) reduced CER by more than 20% from the best results of the conventional beamformers (2-ch Adaptive SS). However, in a realistic environment, the extent of the improvement was reduced to 11%. One of the factors of this degradation could be the reverberation in the room.

The application of Profile Fitting is not limited to the 2-microphone system. It can be easily extended to systems with small numbers of microphones like 3 or 4 microphones configured in 2-dimensional or 3-dimensional geometries, where the associated profiles should have multiple directional axis, each associated directly or indirectly with the spatial dimensions

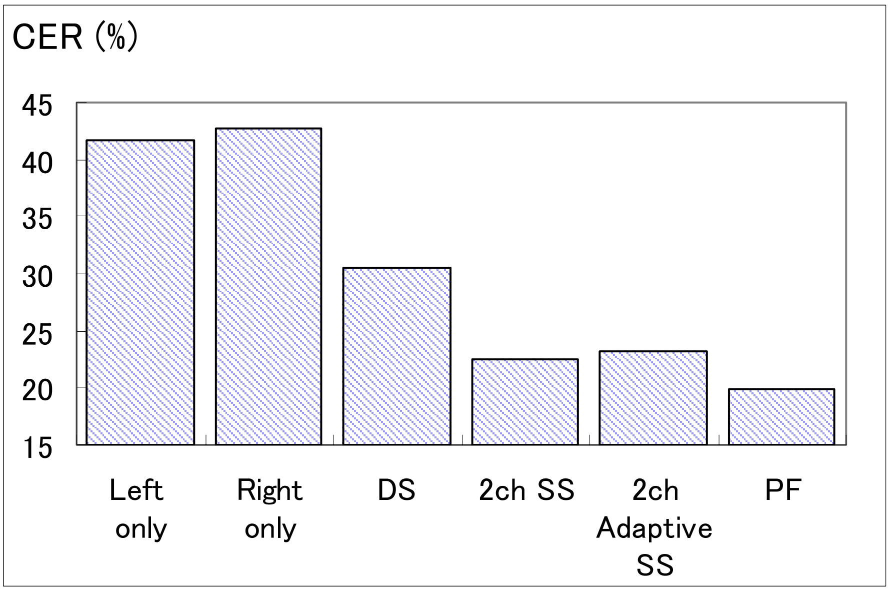


Figure 3.13. Resulting character error rates (in meeting room).

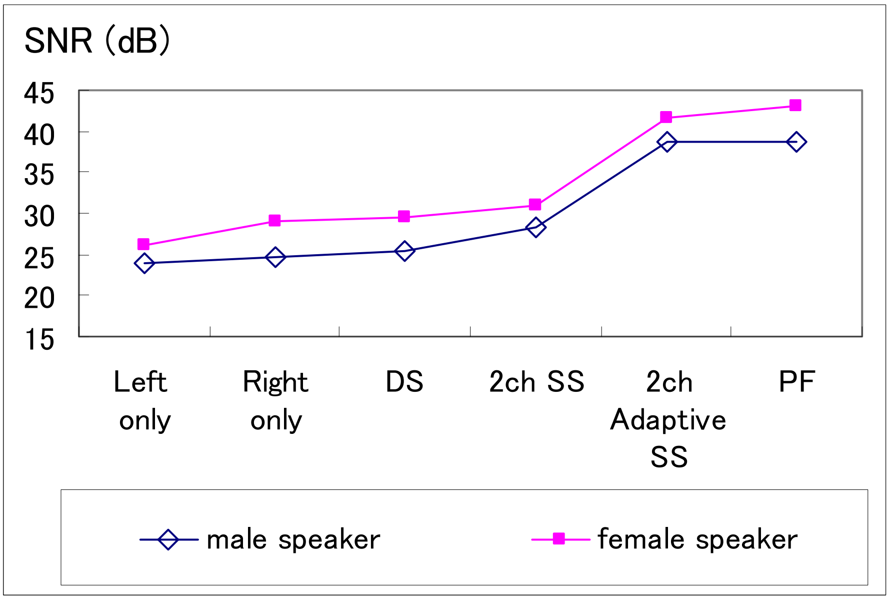


Figure 3.14. Resulting signal-to-noise ratios (in meeting room).

.

# 4. Sound Source Localization by Profile Fitting Method

## 4.1. Introduction

Profile Fitting can be used also for sound source localization. The sound source location (or direction) is essential information for beamformers so as to focus on the target. Therefore conventional beamformers combine some external logic to detect the target source location.

In a two-microphone array system, the interchannel cues (ITD and IID) are often referred to for horizontal localization. There have also been several attempts to apply ITD and IID for vertical localization outside of the median plane [Mar95]. In the median plane, ITD and IID do not contribute to vertical localization [MN82] since they are minimized. To achieve vertical localization in the median plane, it was suggested that a spectral cue model [ZC93][HO97] be integrated. However, since the spectral cues depend on the spectrum of the signal source, they are not robust enough against signal variations and environmental noise. Also, it may require special considerations to consolidate the interchannel cues (ITD and IID) and the spectral cues in one localization system [MII02].

In this chapter, we enhance the localization cues for a specific reflection by using reflectors correlated with the location of the sound source. We call this a reflection cue. It can be detected by CSP analysis directly, or it can be observed as a modification of the ITD, IID, or the profile. By using this reflection cue, we believe equi-distant vertical localization in the median plane becomes possible without relying on the spectral cues.

For noise robustness, we introduce Profile Fitting (PF) method for sound source localization. It was originally proposed for speech enhancement in Chapter 3, but we show it is also effective for localization in a noisy field because of its noise reduction feature. For the conventional method using ITD and IID, several techniques have been proposed to improve the performance in noisy fields [Mar95][NH01][NOK02]. One of them is to use the onsets to get a locally high signal-to-noise ratio (SNR). Another technique is to train the probability density function of the sound location in the actual noise field. However those methods do not have a function to subtract noise, so they depend on the SNR where ITD and IID are trained.

## 4.2. Reflector Design

### 4.2.1. Reflector design for vertical localization

In the HRTF approach, the pinna shape is just a given parameter. In our approach, we deliberately designed the shape of a pinna-like reflector so that the following process can retrieve the localization cues provided by the reflector.

Figure 4.1 shows the concept of the design. The ellipses are plotted where the two foci for each ellipse are at the microphone location and one of the candidate locations of the sound source. The reflector shape is given by the envelope curve for these ellipses. At the upper part of the reflector, sound waves from a high elevation are reflected to focus on the microphone. At the lower part of the reflector, sound waves from a low elevation are reflected so as to focus on it. Sound waves from unmatched elevations should be diffused by the reflection. Therefore the microphone receives both a direct wave and a reflected wave whose delay time is correlated with the sound source elevation. It should be noted that the actual reflector has a 3D-shape designed as an envelope of the revolutions of the ellipses (spheroids).

### 4.2.2. Verifying prototype reflector using CSP analysis

For our experiment, the reflector was made of gypsum molded from a handmade clay model. We verified the working accuracy by Cross-power Spectrum Phase (CSP)

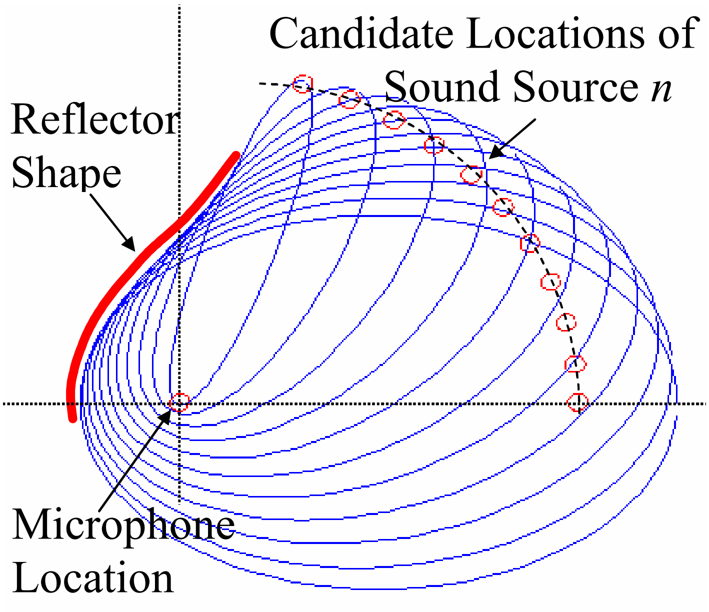


Figure 4.1. Concept of reflector design.

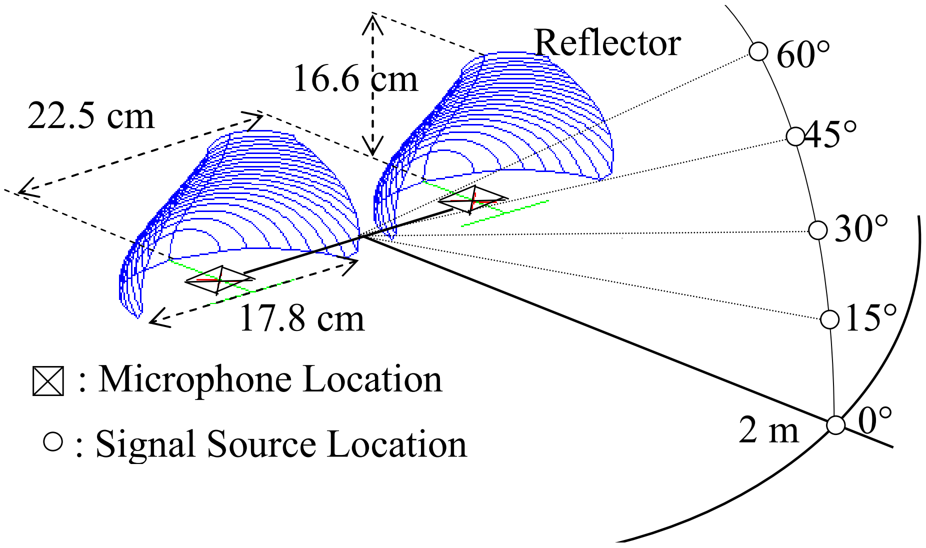


Figure 4.2. Testing configuration for the verification of the prototype reflector.

analysis [OS94] to check that the reflector generated the desired main reflected wave according to the sound source location.

Figure 4.2 shows the configuration for this test. Human speech in calls for attention (“oh-i”, “moshi-moshi”, etc. in Japanese) of about 5 seconds in length were played back

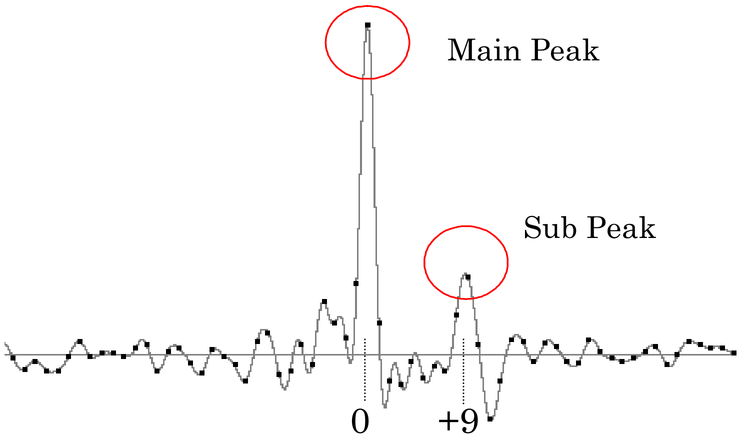


Figure 4.3. Output of CSP analysis with reflector for a signal source at an elevation angle of 30°.

Table 4.1. Peak locations detected by CSP analysis

| Elevation angle of sound source | 0° | 15° | 30° | 45° | 60° |
|---|---|---|---|---|---|
| Peak in 1st place | **0** | **0** | **0** | **0** | **0** |
| Peak in 2nd place | N/A | **10** | **9** | **6** | **2** |
| Peak in 3rd place | N/A | N/A | N/A | **-6** | -10 |
| Design point | ±14 | ±12 | ±9 | ±5.5 | ±2.5 |

in a soundproof chamber using a loudspeaker located directly in front at a distance of 2 m with elevation angles of 0°, 15°, 30°, 45°, and 60°. Two microphones with reflectors recorded the sound signal at a 48 kHz sampling frequency.

As shown in Figure 4.3, the output of CSP analysis shows many sub-peaks, so the criteria of the intensity for the acceptable sub-peaks are arbitrary. Here we took the top 3 peaks whose intensities were greater than a tenth of the main peak as valid peaks. Table 4.1 shows the result of the analysis. The peak in first place is the main peak representing a direct wave. It was observed at position 0. This means the signal source was directly in front. In second and third places, two sub-peaks caused by correlations between the direct wave and the reflected wave should be detected at the designated positions. In these experiments, we observed at least one sub-peak at the designated positions except for 0°, where the area of the designed surface for the reflection (at the root of the reflector) was zero. The absence of an intense reflection can also be treated as a localization cue.

## 4.3. Sound Source Localization

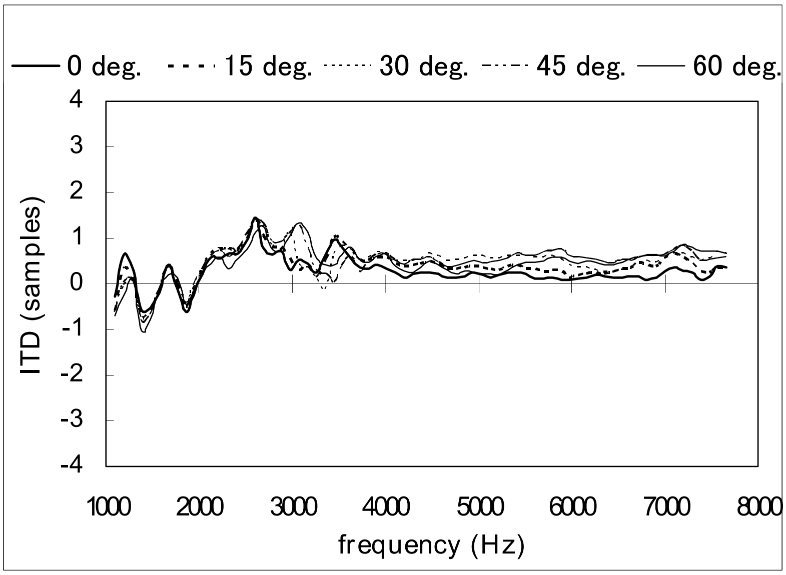
CSP analysis can be used for sound source localization. However, this depends on the assumption that the specific reflected wave is distinct. In a noisy environment, it is difficult for CSP analysis to detect the specific reflected wave, because the sub-peaks associated with the noise sources become dominant. Also, the specific reflected wave can be distinct only when a signal source is located exactly on the designated positions and the working accuracy of the reflectors is precise. Therefore, the conventional method using ITD and IID, and Profile Fitting using a profile are investigated in this section. They do not directly utilize the specific reflected wave, but we expect the design method discussed in Section 4.2 will work to make the large modification in the ITDs, IIDs, and profiles, so that the localization methods can utilize these reinforced localization cues.

### 4.3.1. Conventional method using ITD and IID

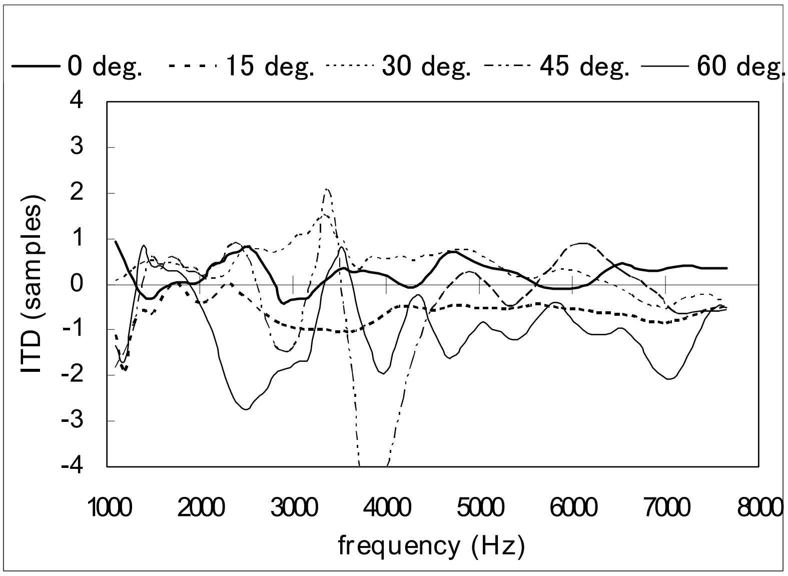
The probability density function, the likelihood that a source is located at a particular position, can be approximated by the product of the marginal distribution of the ITD and IID at each sub-band frequency [Mar95][NH01]. We applied the Gaussian distribution for the likelihood as Equation (4.1):

$$\Psi_n = K \cdot \exp\left[ -\frac{1}{2} \sum_{\omega} \sum_{T} \left\{ \frac{\left( ITD_{\omega,T} - \overline{ITD}_{n,\omega} \right)^2}{\sigma^2_{ITD,\omega}} + \frac{\left( IID_{\omega,T} - \overline{IID}_{n,\omega} \right)^2}{\sigma^2_{IID,\omega}} \right\} \right], \qquad (4.1)$$

where $\Psi_n$ is the likelihood expected for a signal source at $n$, $\omega$ is the sub-band frequency, $T$ is the time frame number, $\sigma^2_{ITD,\omega}$ and $\sigma^2_{IID,\omega}$ are the variances of the interchannel differences under consideration, and $K$ is a normalizing constant.



(a) ITD plots without reflectors.



(b) ITD plots with reflectors.

Figure 4.4. ITD plots with and without reflectors for a signal source in the median plane at various elevation angles. The plots are smoothed over 8 sub-bands (=375 Hz).

We defined the interchannel differences and the variances in Equations (4.2) to (4.7):

$$ITD_{\omega,T} = \angle\left(R_{\omega,T} \cdot L^{*}_{\omega,T}\right) \cdot \frac{f}{2\pi\omega}, \tag{4.2}$$

$$IID_{\omega,T} = 10\log\left(\frac{\left|R_{\omega,T}\right|}{\left|L_{\omega,T}\right|}\right), \tag{4.3}$$

$$\overline{ITD}_{n,\omega} = \frac{1}{N_T}\sum_{T} ITD_{\omega,T}\Big|_{source=n}, \tag{4.4}$$

$$\overline{IID}_{n,\omega} = \frac{1}{N_T}\sum_{T} IID_{\omega,T}\Big|_{source=n}, \tag{4.5}$$

$$\sigma^{2}_{ITD,\omega} = \frac{1}{N_n}\frac{1}{N_T}\sum_{n}\sum_{T}\left(ITD_{\omega,T}\Big|_{source=n} - \overline{ITD}_{n,\omega}\right)^{2}, \tag{4.6}$$

$$\sigma^{2}_{IID,\omega} = \frac{1}{N_n}\frac{1}{N_T}\sum_{n}\sum_{T}\left(IID_{\omega,T}\Big|_{source=n} - \overline{IID}_{n,\omega}\right)^{2}, \tag{4.7}$$

where $R_{\omega,T}$ and $L_{\omega,T}$ are the short-time Fourier transforms of the observations for each of the right and left channels, $N_T$ is the total number of frames to be examined, and $N_n$ is the total number of candidate locations. IID is measured in dB and ITD is measured in units of the sampling count. We selected time frames of 0.2 sec around the onset for the each utterance to be examined.

Before the experiment, $\overline{ITD}_{n,\omega}$, $\overline{IID}_{n,\omega}$, $\sigma^{2}_{ITD,\omega}$, and $\sigma^{2}_{IID,\omega}$ should be trained using a signal from each candidate location $n$ with or without noise at a specific SNR.

## 4.3.2. Reflector effect on ITD and IID

If the left and right reflectors are configured completely symmetrically, ITD and IID still take near-zero values. However, as shown in the CSP output of our prototype (Figure 4.3), the desired reflected waves generated by the actual left and right reflectors are not necessarily at the same level. In that case, the ITD and IID values are significantly modified by the reflected waves. Also, it is difficult to predict the actual modification before measurement, because there are many reflected waves and their

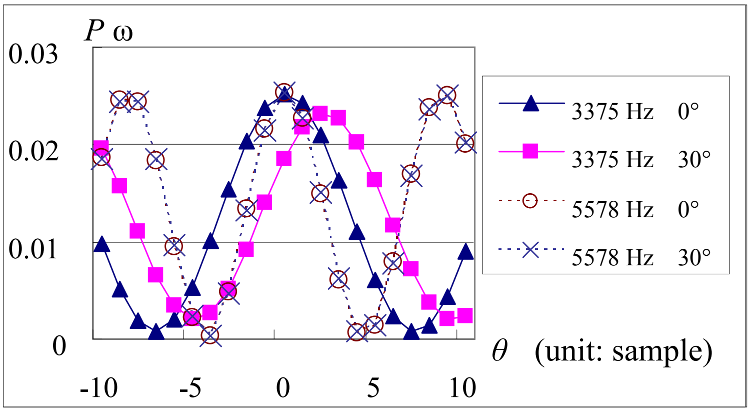


Figure 4.5. Template profiles for a signal source at elevation angles of 0° and 30° measured with reflector.

levels are not balanced. The expectation here is that the reflectors should just cause large modifications at the characteristic positions. For an example, Figure 4.4 shows the ITDs with and without our reflectors. Without reflectors, ITD plots are similar against variations of signal source elevation. This implies it is difficult to determine the signal source elevation by ITD without reflectors. With reflectors, we can observe the shape of ITD plots varies a lot against signal source elevation. As the localization process checks the shapes as a whole, it should not be a problem, even if they are partially similar, under the assumption that the signal is broadband.

## 4.3.3. Profile Fitting

For robustness against noise, we introduce a Profile Fitting for sound source localization utilizing the residual of the approximate decomposition of signal and noise. It is based on the concept that the power distribution observed at varying look direction can be approximated by the linear combinations of the template distributions, each associated with a signal source and a noise source. When the assumed location $n$ is correct, Equation (4.8) is justified.

$$X_{\omega}(\theta) \cong \alpha_{n,\omega} \cdot P_{n,\omega}(\theta) + \beta_{n,\omega} \cdot Q_{\omega}(\theta), \qquad (4.8)$$

where $X_{\omega}(\theta)$ is the power distribution of the sub-band frequency $\omega$ observed at the particular look direction $\theta$ for a delay and sum beamformer. This is called an “observed profile.” *$P_{n,\omega}(\theta)$* is a “template profile” measured by white noise coming from the candidate location *n* for the signal source. *$Q_{\omega}(\theta)$* is a “template profile” measured for the noise source. The template profile for the noise source can be measured using a white noise originating from the noise source before the experiment if the location of the noise source is known a priori. Otherwise it should be measured from the actual noise by averaging over noise segments during the experiment.

Profile Fitting determines each of the weight coefficients $\alpha_{n,\omega}$ and $\beta_{n,\omega}$ for the template profiles of a signal source and a noise source, so as to minimize the evaluation function $\Phi_{n,\omega}$ defined by Equation (4.9):

$$\Phi_{n,\omega} = \int_{\min\_\theta}^{\max\_\theta} \left\{ X_{\omega}(\theta) - \alpha_{n,\omega} \cdot P_{n,\omega}(\theta) - \beta_{n,\omega} \cdot Q_{\omega}(\theta) \right\}^2 d\theta \quad . \tag{4.9}$$

We configure the delay and sum beamformer in the time domain, using Equation (4.10), and the observed profile *$X_{\omega}(\theta)$* is derived by using Equations (4.11) and (4.12):

$$s(t,\theta) = l(t) + r(t+\theta), \tag{4.10}$$

$$S_{\omega,T}(\theta) = DFT[s(t,\theta)], \tag{4.11}$$

$$X_{\omega}(\theta) = \frac{1}{N_T} \sum_T S_{\omega,T}(\theta) \cdot S_{\omega,T}(\theta)^* \quad , \tag{4.12}$$

where *l*(*t*) and *r*(*t*) are the time domain observations of the left and right channels at the *t*-th sample, and the look direction *θ* is measured by the delay in the samples. *T* is the time frame number and $N_T$ is the total number of frames. Since the template profile should contain only the directivity information, it is normalized by the power at each sub-band as Equation (4.13):

$$P_{n,\omega}(\theta) = \frac{X_{\omega}(\theta)\big|_{source=n}}{\int_{\min\_\theta}^{\max\_\theta} X_{\omega}(\theta)\big|_{source=n} d\theta} \quad . \tag{4.13}$$

For speech enhancement, the decomposition using Equation (4.9) should be done in each time frame, but for sound source localization, it should be done only once. Therefore, *$X_{\omega}(\theta)$* is an averaged observation over a few seconds. As Profile Fitting does not rely on onsets, test data can include non-speech frames before and after the utterances.

The coefficients $\alpha_{n,\omega}$ and $\beta_{n,\omega}$ can be determined by variation method with non-negative conditions.

Once the coefficients are determined, then the residual $\Phi_{n,\omega}$ can be determined. With Equation (4.14), we calculate the normalized residual $\overline{\Phi}_n$ as a function of $n$ by dividing the sub-band power and averaging over the $\Omega$ sub-bands. Using Equation (4.15), the location of the signal source is estimated as $\hat{n}$ so as to minimize the normalized residual.

$$\overline{\Phi}_n = \frac{1}{\Omega}\sum_{\omega}\frac{\Phi_{n,\omega}}{\int_{\min\_\theta}^{\max\_\theta}\{X_\omega(\theta)\}^2 d\theta} \quad . \tag{4.14}$$

$$\hat{n} = \arg\min_{n}\left(\overline{\Phi}_n\right) \quad . \tag{4.15}$$

## 4.3.4. Reflector effect on profile

A profile contains ITD information as peak-shifts and IID information as a bias. Also, diffusion or reflection of the target signal increases the bias of the profile. Therefore, it should be noted that even though the desired reflected waves generated by the left and right reflectors are completely identical, the bias of the profile still retains the reflection cue, while the peak-shift might be zero in that case.

Figure 4.5 compares the template profile for an elevation angle of 30° with the one for 0°. At the frequency of 3,375 Hz, the peak-shift and bias are observed in the profile for 30°. They are caused by the reflected waves arriving with their own delays.

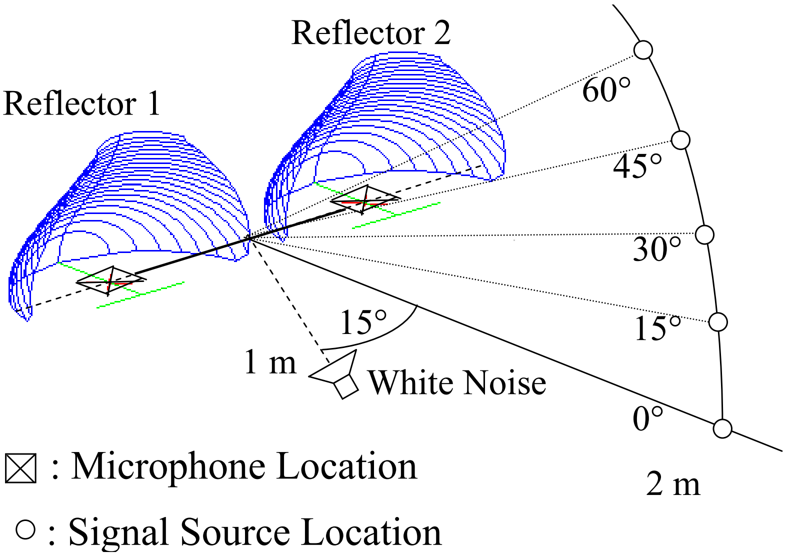


Figure 4.6. Testing configuration for the preliminary experiment.

# 4.4. Experiments and Results

## 4.4.1. Preliminary experiment

In order to verify Profile Fitting with the designed reflectors works correctly, we performed a preliminary experiment using a limited amount of data for vertical localization in a sound proof chamber.

The recording parameters and the geometry are the same as in Section 4.2.2 for the CSP analysis. In a soundproof chamber, four utterances about 5 seconds in length were played back from each candidate location for a signal source. As a noise source, white noise was played from a loudspeaker at an azimuth angle of 15°, a distance of 1 m, and an elevation angle of 0° (Figure 4.6). The recorded noise was manually mixed with the recorded signal, so that the SNR could be controlled.

Before the experiment, the template profiles for the signal sources and the noise source were individually measured using white noise coming from each sound source location.

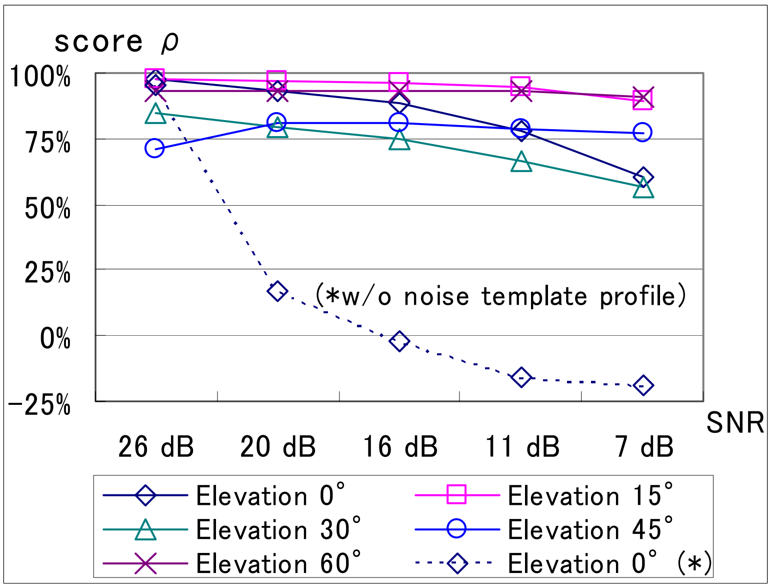


Figure 4.7. Resulting score for sound source localization by Profile Fitting. (*) denotes a reference trial without using the template profile for the noise source.

Using Equations (4.16) and (4.17), a score $\rho$ is introduced to define the relative degrees of superiority using the second best (smallest) normalized residual as the base value. Here, $n^{\circ}$ denotes the correct location. When the correct location has the minimum value, it should be selected by Equation (4.15) and the score will have a positive value. If the normalized residual is zero, the score becomes 100%. A positive large score means it is estimated with high confidence. If the score decreases close to zero, it means the chances increase that the second best candidate might be incorrectly taken as a result of noise or some other influence. If the correct location does not have the minimum value, then Equation (4.15) will fail to select the correct location, and the score will have a negative value.

$$\rho = \frac{\overline{\Phi}_{\bar{n}} - \overline{\Phi}_{n^{\circ}}}{\overline{\Phi}_{\bar{n}}} \quad . \tag{4.16}$$

$$\bar{n} = \underset{n \neq n^{\circ}}{\operatorname{argmin}} \left( \overline{\Phi}_n \right) \quad . \tag{4.17}$$

On calculating the normalized residual in Equation (4.14), an averaging operation was performed over the sub-band frequencies from 938 Hz to 7,453 Hz where the reflector effect is most apparent.

Figure 4.7 shows the experimental results. All elevations maintain large positive scores in spite of SNR degradation. This means the correct signal location was selected from the five candidates without being affected by noise, showing the superiority of the approximate decomposition by Profile Fitting. On the other hand, the reference experiment (marked * in Figure 4.7) without using the template profile for the noise source failed in the noisy environment.

### 4.4.2. Experiments in a realistic environment

In order to evaluate the capability in more realistic conditions, we performed an experiment using more utterances from more locations in a slightly reverberant meeting room with realistic noise.

As shown in Figure 4.8, 21 locations were defined as a signal source location. They are also candidate locations for the localization. They have 5 horizontal steps from -30° to +30°, and 5 vertical steps from 0° to 60°. As a noise source, cafeteria noise in stereo was played from two loudspeakers at azimuth angles of 30° and -30°, a distance of 2 m, and an elevation angle of 0°. The recorded noise was manually mixed with the recorded signal, so that the SNR could be controlled. The recording was done in our meeting room whose reverberant time is about 0.22 sec.

Per location, a total of 108 utterances of personal names spoken by 6 male and 6 female speakers were played back. In order to evaluate the robustness, we projected an imaginary grid around each candidate location as shown in Figure 4.9, and played back almost same numbers of utterances from each grid point. Here, we categorize the utterances by the offset error from the candidate location. Category A is for the utterances from the exact candidate location. Category B is for the utterances whose azimuth angle and elevation angle are correct but whose distance contains about ±10% error. Category C is for the utterances whose azimuth angle is correct but whose elevation angle contains about ±4° error or whose distance contains about ±10% error. Category D is for all the utterances that contain at least one of the errors in azimuth of about ±4°, elevation of about ±4°, or distance of about ±10%. It should be noted Category B, Category C and Category D do not include Category A, and therefore the Categories other than A involve offset errors in one or multiple dimensions.

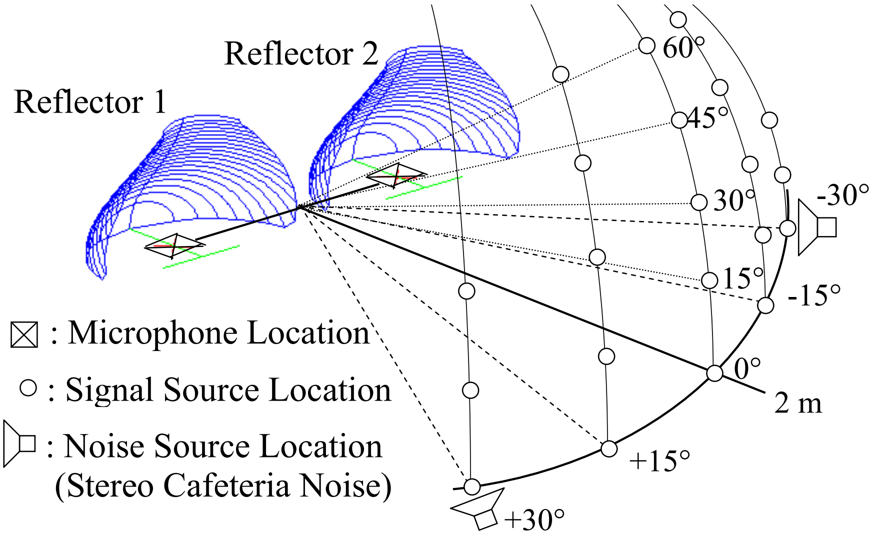


Figure 4.8. Testing configuration for the experiment in a realistic environment.

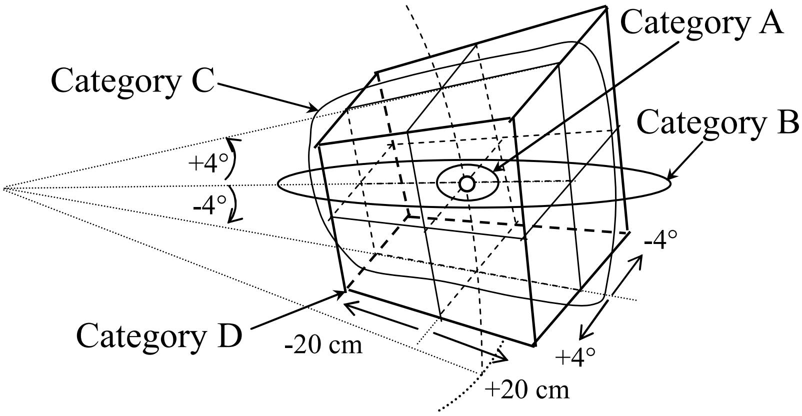


Figure 4.9. Category by the offset from the location.

The sizes of the offset errors should not be too large with reference to the design points and the neighboring candidate locations. Here, the offset errors in azimuth and elevation are about a quarter of the angles between the candidate locations. The offset error in distance is chosen as a simple fraction of the distance between the microphone and the candidate locations, so that it will be near to the actual length of the offset errors in azimuth and elevation.

For Profile Fitting, the template profile for the noise source was measured from the actual noise for 1 sec just before each utterance. It should be noted that the template does not contain any spectral information, but just records the directivity information as it is normalized by a power at each sub-band.

Both for Profile Fitting and the conventional method, the sub-bands to be examined were selected from 938 Hz to 7,453 Hz where the reflector effect is most apparent.

Figure 4.10 shows the success rates for the localization of 5 signal source locations in the median plane out of 21 candidate locations. The SNR was 11 dB. Both Profile Fitting and the conventional method (trained by the utterances in Category A) showed high success rates for the utterances in Category A that have very little offset error from the candidate locations. On the other hand, the success rates are significantly decreased for the utterances in Category D that have much larger offset errors. Figure 4.10 also shows the result of the conventional method trained using the utterances in Category D. This improved the success rate for the utterances in Category D. In that case, the probability density functions have broad distributions, as they are trained with large offset errors associated with Category D. Therefore, that causes a significant loss of accuracy for the utterances in Category A.

In order to evaluate the dependency on SNR, we also tried this localization without adding noise. The SNR was 28 dB. Figure 4.11 shows the resulting success rate. It also shows the result of the conventional method that was trained in a noisy environment (11 dB). In that case, the SNR was unmatched between the training and the localization. The success rate of this unmatched case was worse than the matched cases shown in Figure 4.10 (at 11 dB) and Figure 4.11 (at 28 dB). We conclude the conventional method is dependent on the SNR when it is trained. Also, there is concern that the conventional method is dependent on the noise color as well as the SNR, because the probability density functions are trained for each sub-band. On the other hand, Profile Fitting is less dependent on them, because it does not require any training in advance.

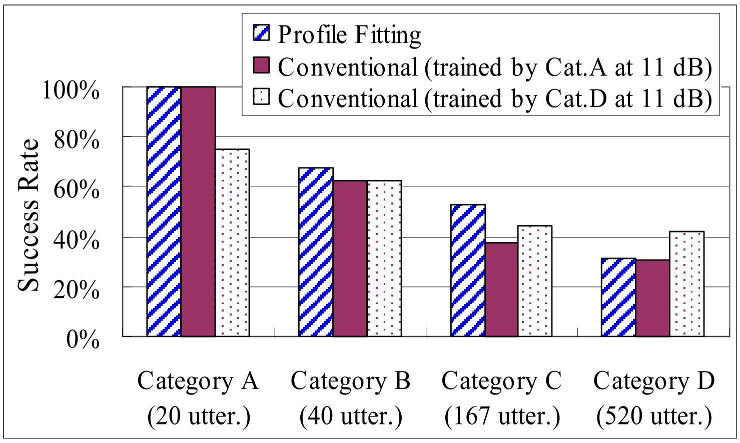


Figure 4.10. Success rates for the localization of 5 signal source locations in the median plane out of 21 candidate locations at the SNR of 11 dB.

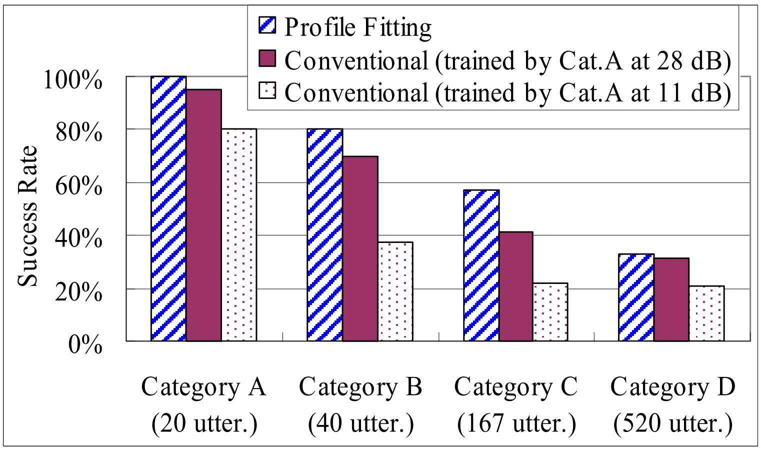


Figure 4.11. Success rates for the localization of 5 signal source locations in the median plane out of 21 candidate locations at the SNR of 28 dB.

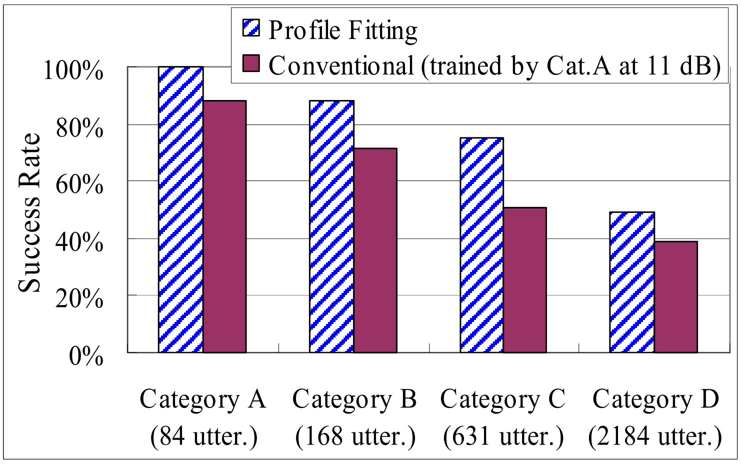
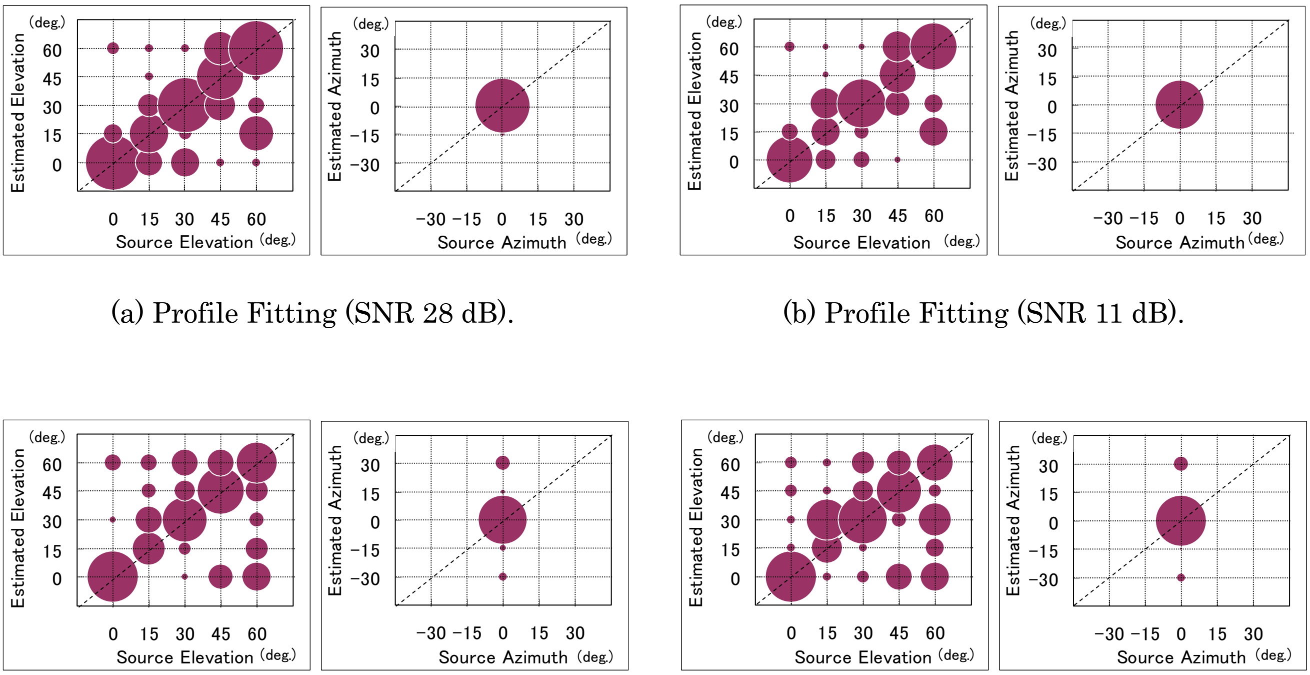


Figure 4.12. Success rates for the localization of 21 signal source locations out of 21 candidate locations at the SNR of 11 dB.



(a) Profile Fitting (SNR 28 dB).

(b) Profile Fitting (SNR 11 dB).

(c) Conventional method (SNR 28 dB).

(d) Conventional method (SNR 11 dB).

Figure 4.13. Maps of the signal source locations and the estimated locations for the utterances included in Categories A, B, and C (187 utterances), localizing 5 signal source locations in the median plane out of 21 candidate locations at the SNRs of 28 dB and 11 dB.
The conventional method was trained using the utterances in Category A at the matched SNR. The area of the each bubble is proportional to the number of estimations.

Using not only the 5 signal source locations in the median plane, but also using all of the 21 signal source locations, Figure 4.12 shows the success rates resulting for the localization out of 21 candidate locations. We see Profile Fitting outperformed the conventional method in all categories. It should be noted that the conventional method checked the utterances only around onsets where the SNR was locally high, both for training and localization. Profile Fitting did not use this technique and still had an advantage in the experimental results.

Figure 4.13 shows maps of the signal source locations and the estimated locations for the utterances from the 5 signal source locations in the median plane in Categories A, B, and C. In the error cases, the locations estimated by Profile Fitting were closer to the correct locations than the ones using the conventional method. This trend was still observed when the SNR was reduced to 11 dB. In both methods, the azimuth estimation was very accurate.

## 4.5. Concluding Remarks

We have proposed a framework for sound source localization using Profile Fitting. This can reduce the effect of noise by exploiting the approximate decomposition of signal and noise. In Profile Fitting combined with reflectors, the process for horizontal localization and the process for vertical localization can be consolidated into a single process. Experiments showed this method can correctly provide a rough estimate of the vertical location in the median plane even in a noisy environment. Profile Fitting showed more robustness against SNR variations than the conventional method using ITD and IID.

# 5. Echo-Cancellation and Noise Reduction by SSEC Method

## 5.1. Introduction

Automatic speech recognition is widely used in cars to input commands for car navigation and hands-free telephone dialers. However, the current systems are not sufficiently robust against noise. As most of the current systems are based on the techniques of multi-conditional training and spectral subtraction [Bol79][BSM79], they rely on the assumption that there is only a stationary cruising noise. Therefore, the recognition rate is degraded when there are non-stationary noises such as those created by road bumps or oncoming cars. The degradation is much more severe when there is music or news coming from a radio or a CD player in the car.

Music and news are actually non-stationary noises. However, if they are coming from a radio or a CD player in a car, we may have a chance to use an echo canceller, because it is not technically difficult to rout the reference signals from such devices to the recognition system.

Previous research reported that an echo canceller works well in a quiet environment. However its performance is poor for low signal to noise ratios [BSN00]. There has been a lot of research on ways to improve the performance of echo cancellers along with noise reduction [MV96][AFB96][DP97][SNH+03]. However, many of the target uses were for teleconference and hands-free telephones, where auditory intelligibility has the highest priority. Our objective is to find a solution for automatic speech recognition with high performance echo cancellation and noise reduction. Our second objective is to retain practical compatibility with the current acoustic model trained with stationary cruising noise and spectral subtraction. In this chapter, we assume the cruising noise can be treated as stationary.

## 5.2. Conventional Methods

In order to improve the performance of an echo canceller in a noisy environment, the background noise should be reduced before echo cancellation. If many microphones are available, a beamformer can be used to reduce the noise before or at the same step as the echo cancellation [DCN97][KFK04].

Since we assume a single microphone, we need to consider one-channel noise reduction instead of using a beamformer approach. A Wiener Filter [LO79], MMSE [EM84], and spectral subtraction are candidates for the noise reduction. For automatic speech recognition, spectral subtraction is often used because of the computational cost and the performance. As the output is not for humans, the annoying side effect known as musical noise is acceptable. However, the problem for this application is that we cannot place the noise reduction stage before the echo canceller because of the nonlinearity in the echo path [BSN00]. Therefore, the conventional combination is echo cancellation first and noise reduction second, as shown in Figure 5.1.

The conventional time-domain echo-cancellers based on LMS rely on phase correlation as well as magnitude correlation between the observed signal and the reference signal. However the phase information is susceptible to noise. That’s one of the reasons why it takes longer time in adaptation in noisy environment. On the other hand, the echo canceller based on spectral subtraction does not rely on phase information. Therefore, the adaptation will quickly converge with some trade off in accuracy due to the lack of phase information. However, it should be noted that the remaining echo can be further reduced by introducing an over-subtraction technique with the echo canceller based on spectral subtraction.

If the echo canceller is implemented using spectral subtraction, the noise reduction stage can be placed before or at the same step as the echo canceller, and we can expect better performance. However, the question is how to estimate the stationary noise power for the noise reduction under the influence of the echo. If the application is a telephone, we can expect noise-only periods in which no one is speaking [Tak97]. However, we cannot expect such a period in our application, because a car-radio or a car-CD

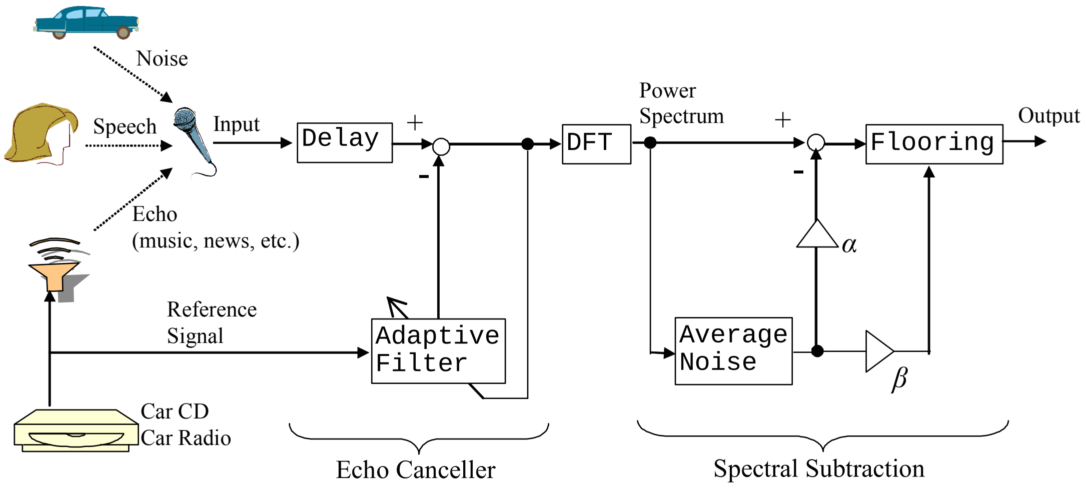


Figure 5.1. Conventional combination of echo canceller and spectral subtraction.

produces sound continuously. Therefore, we propose a new method that estimates the stationary noise power during the adaptation of the echo canceller using spectral subtraction.

Dreiseitel et al. placed a time-domain echo canceller before the combination of noise reduction and echo canceller in spectral subtraction form [DP97]. By preprocessing the input using the echo canceller, the stationary noise is estimated more reliably at the noise reduction stage. Our proposed method can also work with this type of preprocessing for further improvement.

Since the reverberation in a car is longer than the processing frame, it degrades the performance of frame-based echo cancellation using spectral subtraction. In order to solve this problem, Sakauchi et al. introduced a second term, a scaled echo component estimated in the previous frame [SHN+03]. However, their scaling factor should be preset depending on the reverberation in the room. In contrast, our system does not require any a priori knowledge about the room reverberation, because we extended the echo cancellation to refer to the last several frames, and the factors can be determined through the adaptation. The structure is similar to the taps of an adaptive filter in the time domain. In this way, our echo canceller can be adapted to cancel the echoes including the reverberations from past frames.

## 5.3. Proposed Method (SSEC)

We propose a method named SSEC (Simultaneous adaptation of spectral Subtraction and Echo Cancellation). A stationary noise component for spectral subtraction is estimated through the adaptation of an echo canceller. Figure 5.2 shows a block diagram of our proposed Method 1 (without preprocessing), and Figure 5.3 shows our proposed Method 2 (with preprocessing). As the preprocessing stage is a standard N-LMS echo canceller in the time domain, we describe our method after the preprocessor.

The echo canceller stage and the spectral subtraction stage are integrated into the same stage. This estimates both the stationary noise power $\overline{N_\omega}$ and the echo power $Q_\omega(T)$. They are subtracted from the observed noise power $X_\omega(T)$ with the subtraction weights $\alpha_1$ and $\alpha_2$, respectively. The compensated output $Y_\omega(T)$ is written as Equation (5.1).

$$Y_\omega(T) = X_\omega(T) - \alpha_2 \cdot Q_\omega(T) - \alpha_1 \cdot \overline{N_\omega}, \tag{5.1}$$

where $T$ is a frame number. The index $\omega$ is a bin number of the DFT corresponding to the sub-band frequency, and the process described in this section should be performed for each $\omega$.

In general, flooring is an essential technique for spectral subtraction. The floored output $Z_\omega(T)$ is given by Equations (5.2a) and (5.2b).

$$Z_\omega(T) = Y_\omega(T) \qquad \text{if } Y_\omega(T) \geq \beta \cdot \overline{N_\omega} \ , \tag{5.2a}$$

$$Z_\omega(T) = \beta \cdot \overline{N_\omega} \qquad \text{if } Y_\omega(T) < \beta \cdot \overline{N_\omega} \ , \tag{5.2b}$$

where $\beta$ is a flooring coefficient. $\alpha_1$ and $\beta$ should be set to the same value with which the acoustic model was trained. The value of $\alpha_2$ can be larger than $\alpha_1$ for over-subtraction in order to cancel more of the echo component, which has a large effect on the performance of automatic speech recognition. We introduce an over-subtraction factor $\gamma$ as Equation (5.3).

$$\alpha_2 = \gamma \cdot \alpha_1 . \tag{5.3}$$

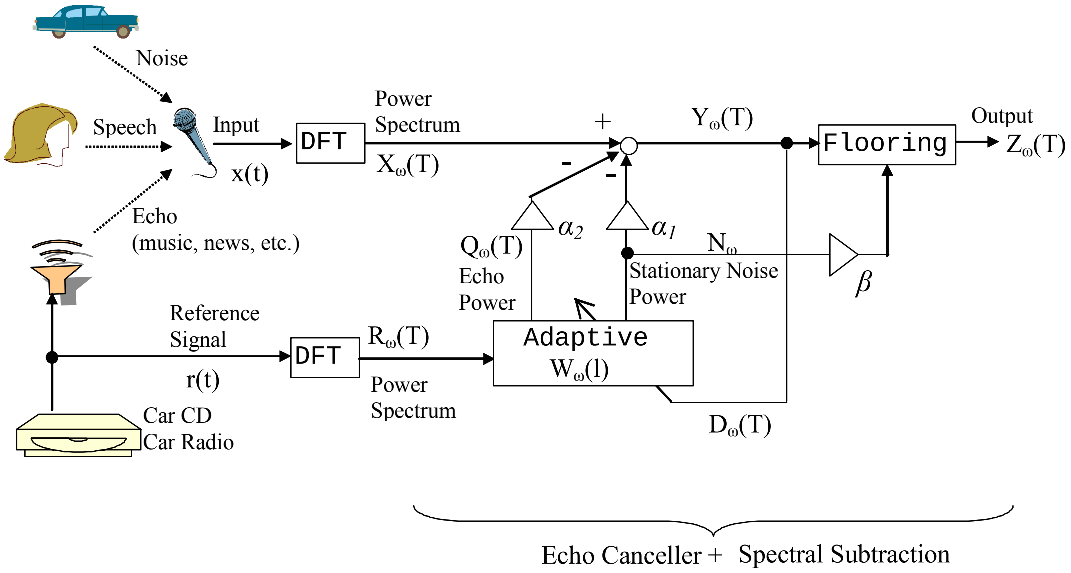


Figure 5.2. Proposed method 1 (without preprocessor).

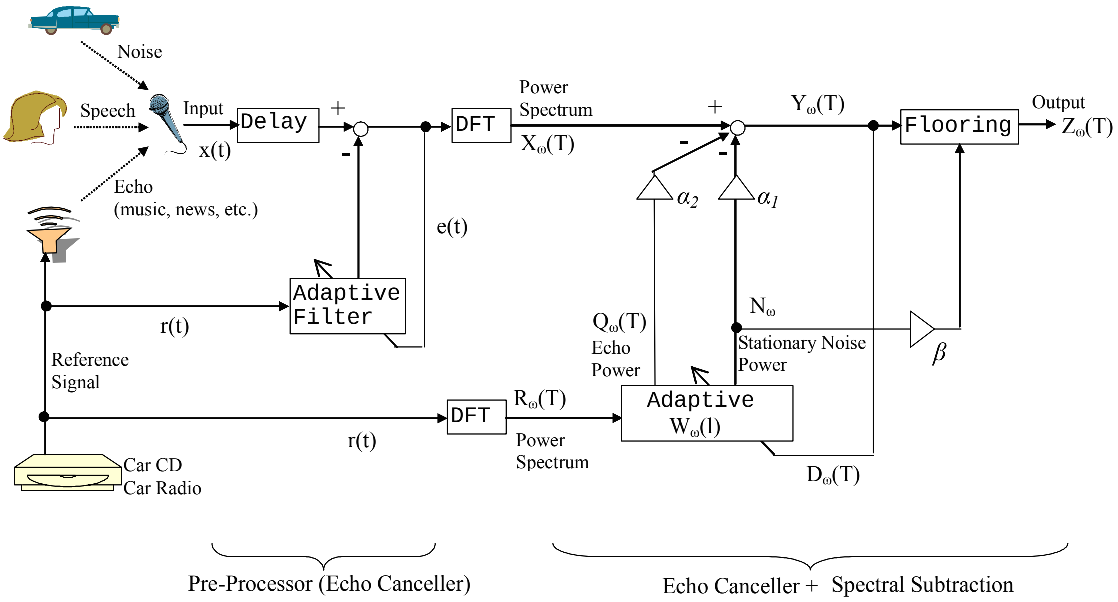


Figure 5.3. Proposed method 2 (with preprocessor).

Next we describe how to estimate $\overline{N_\omega}$ and $Q_\omega(T)$. The value of $Q_\omega(T)$ is estimated as the weighted sum of the reference signal power $R_\omega(T)$ for the present and the most recent $L$ frames so as to cope with reverberation that lasts longer than the processing frame.

$$Q_\omega(T) = \sum_{l=0}^{L-1} W_\omega(l) \cdot R_\omega(T-l). \tag{5.4}$$

For convenience, $\overline{N_\omega}$ is formulated as a product of an arbitrary constant $C$ and its weight.

$$\overline{N_\omega} = W_\omega(L) \cdot C. \tag{5.5}$$

Although we only consider the stationary cruising noise of a car, the stationary noise power may fluctuate around the average in the frame-wise observation, so $\overline{N_\omega}$ can be estimated as an averaged value. Figure 5.4 shows the concept of the estimation.

Therefore, our goal is to estimate the non-negative adaptive weights $W_\omega(l)$ where $l$ ranges from 0 *to L*. They should be set so as to minimize Equation (5.6) during non-speech periods with the subtraction weights $\alpha_1$ and $\alpha_2$ set to 1.

$$\Phi_\omega = E\left[\{D_\omega(T)\}^2\right], \tag{5.6}$$

where $D_\omega(T)$ is the error signal as defined in Equation (5.7). $E[\ ]$ denotes the expectation operator and we calculate it as the frame-wise average during non-speech periods.

$$\begin{aligned} D_\omega(T) &= X_\omega(T) - Q_\omega(T) - \overline{N_\omega} \\ &= X_\omega(T) - \left[R_\omega(T), \ \cdots \ R_\omega(T-(L-1)), \ C\right] \cdot \begin{bmatrix} W_\omega(0) \\ \vdots \\ W_\omega(L-1) \\ W_\omega(L) \end{bmatrix}. \end{aligned} \tag{5.7}$$

The values of $W_\omega(l)$ can be determined from $\partial\Phi_\omega / \partial W_\omega(l) = 0$. This can be expressed in a matrix and vectors as Equations (5.8) to (5.11).

$$\mathbf{C}_\omega = \mathbf{A}_\omega \cdot \mathbf{B}_\omega. \tag{5.8}$$

$$\mathbf{A}_{\omega}=\begin{bmatrix} \sum_{T} R_{\omega}(T)\cdot R_{\omega}(T) & \cdots & \sum_{T} R_{\omega}(T-(L-1))\cdot R_{\omega}(T) & \sum_{T} C\cdot R_{\omega}(T) \\ \vdots & \ddots & \vdots & \vdots \\ \sum_{T} R_{\omega}(T)\cdot R_{\omega}(T-(L-1)) & \cdots & \sum_{T} R_{\omega}(T-(L-1))\cdot R_{\omega}(T-(L-1)) & \sum_{T} C\cdot R_{\omega}(T-(L-1)) \\ \sum_{T} R_{\omega}(T)\cdot C & \cdots & \sum_{T} R_{\omega}(T-(L-1))\cdot C & \sum_{T} C\cdot C \end{bmatrix}. \tag{5.9}$$

$$\mathbf{B}_{\omega}=\begin{bmatrix} W_{\omega}(0) \\ \vdots \\ W_{\omega}(L-1) \\ W_{\omega}(L) \end{bmatrix}. \tag{5.10}$$

$$\mathbf{C}_{\omega}=\begin{bmatrix} \sum_{T} R_{\omega}(T)\cdot X_{\omega}(T) \\ \vdots \\ \sum_{T} R_{\omega}(T-(L-1))\cdot X_{\omega}(T) \\ \sum_{T} C\cdot X_{\omega}(T) \end{bmatrix}. \tag{5.11}$$

The values of $W_{\omega}(l)$ can be determined from Equation (5.12).

$$\mathbf{B}_{\omega}=\mathbf{A}_{\omega}^{-1}\cdot\mathbf{C}_{\omega}. \tag{5.12}$$

Since this off-line form requires the inverse matrix, it has considerable computational cost. By introducing the diagonal approximation for the matrix $\mathbf{A}_{\omega}$, we can formulate the adaptive weights $W_{\omega}(l)$ so as to be successively updated in each non-speech frame using Equations (5.13a), (5.13b) and (5.14). The parameter $\theta$ is an updating factor and $\varepsilon$ is a constant for stability.

$$\Delta W_{\omega}(l)=\theta\cdot\frac{R_{\omega}(T-l)\cdot D_{\omega}(T)}{\sum_{T} R_{\omega}(T-l)\cdot R_{\omega}(T-l)+\varepsilon}, \qquad \text{if} \quad l<L. \tag{5.13a}$$

$$\Delta W_{\omega}(l)=\theta\cdot\frac{C\cdot D_{\omega}(T)}{\sum_{T} C\cdot C+\varepsilon}, \qquad \text{if} \quad l=L. \tag{5.13b}$$

$$W_{\omega}(l)^{T}=W_{\omega}(l)^{T-1}+\Delta W_{\omega}(l). \tag{5.14}$$

This on-line form has a weak dependency on the constant $C$.

Table 5.1. The signal-to-noise ratio of the data in the experiment. They are the averaged values for all 24 subject speakers. $N_{cruise}$, $N_{music}$ and $N_{all}$ denote the cruising noise component, the music noise component and the total noise respectively

| (dB) | Stationary | City Drv. | Highway |
|---|---|---|---|
| $S/N_{cruise}$ | 10.5 | 4.5 | 2.6 |
| $S/N_{music}$ | 10.1 | 6.5 | 9.8 |
| $S/N_{all}$ | 6.4 | 1.1 | 1.2 |

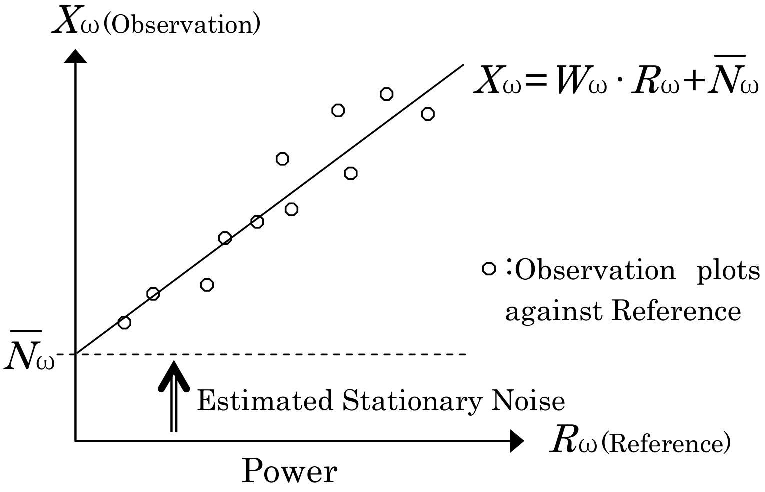

$X_\omega$ (Observation)
$X_\omega = W_\omega \cdot R_\omega + \overline{N_\omega}$
○ :Observation plots against Reference
$\overline{N_\omega}$
Estimated Stationary Noise
$R_\omega$ (Reference)
Power

Figure 5.4. The concept how to estimate the averaged stationary noise power $\overline{N_\omega}$ through the adaptation of $W_\omega$. ($L$ is set to 1 for simplicity.)

## 5.4. Preliminary Experiment

Before evaluating the proposed method in automatic speech recognition, we first checked that it can properly estimate the stationary noise power through the adaptation of the echo canceller. We selected one male utterance from the test set for this trial. It was recorded in an actual car driving on a highway, and manually mixed with the music sound recorded separately. As shown in Figure 5.5, our method was able to estimate the stationary noise power in an acceptable way. Here, the actual stationary noise was measured as the average of the cruising noise. If we estimated the stationary noise power by the simple average of the observation regardless of the existence of the echo, it would be very different from the actual values, as plotted with the "by Direct Average" line in Figure 5.5. For the proposed methods, we see some deviation in the rage of 1,600-2400 Hz, 3,500-4,000 Hz and 4,500-5,000 Hz. Our interpretation is that the estimation error increased because the echo component is much larger than the stationary noise in those range. In other words, the deviation is relatively small compared to the echo power, which is why they are plotted with a logarithmic scale.

## 5.5. Experiment in Automatic Speech Recognition

A microphone was installed on the visor in a car. The subject speakers were 12 females and 12 males. Each speaker read 13 Japanese sentences for the digit recognition task and for the command recognition task in a car at each of three speeds (stationary, city driving, or highway speed). The total number of utterances was 936 for each test subject over all of the tasks. They were recorded with a sampling frequency of 22 kHz. The cruising noise in the recorded data was almost constant.

The music playing from the in-vehicle loud speakers was recorded separately by a microphone, along with a reference signal. The music was up-tempo popular music with a female vocalist. The in-vehicle loud speakers are stereo, but the music source was monaural in this experiment. The recorded music was mixed with the recorded utterances to generate the test data. The averaged SNRs are shown in Table 5.1. The noise power and the signal power were measured by the average in the non-speech and speech periods respectively in the recorded data.

The digit recognition task involves connected digits with no grammar constraints on the length. Therefore, it is sensitive to insertion errors, mostly occurring in the non-speech periods, and this allows measuring the amount of residual echo.

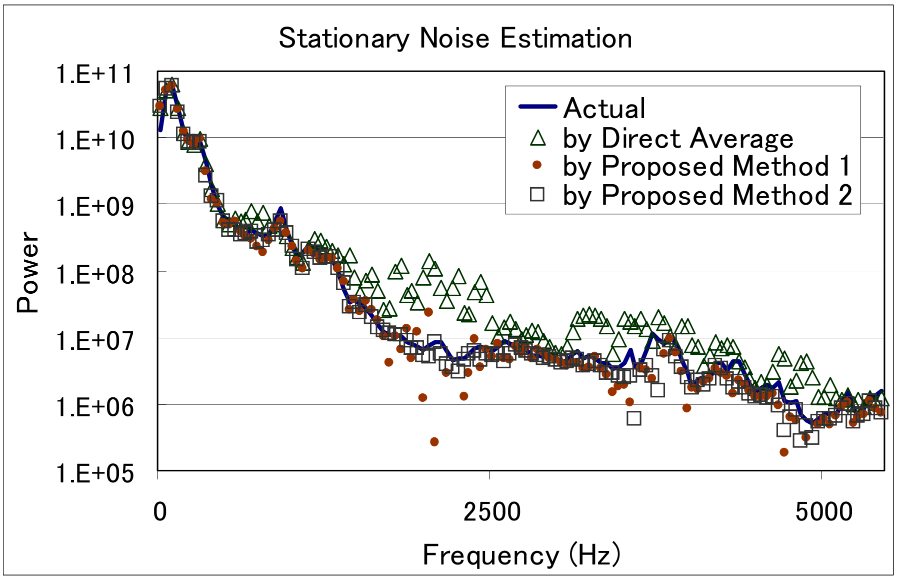


Figure 5.5. Power plots for the actual stationary noise and the estimated stationary noise under the influence of the echo. The proposed method estimated the stationary noise using the off-line formula.

The command recognition task is a set of commands used in a car, such as "North Up", "Input Address", etc. As the grammar only allows 1 command per utterance, we do not have to worry about insertion errors. Therefore, this allows measuring the amount of distortion of the speech (possibly caused by the echo canceller).

The acoustic model used for this automatic speech recognition was a speaker independent model trained with various cruising noises including idling, city driving and highway driving. The acoustic model was trained using spectral subtraction with the subtraction weight set to 1.0. Since the training data was sampled at 11 kHz, the test data was down-sampled before recognition. In this experiment, we did not use a speech-silence detector for the automatic speech recognition and the complete utterances were decoded in order to measure the front-end performance accurately.

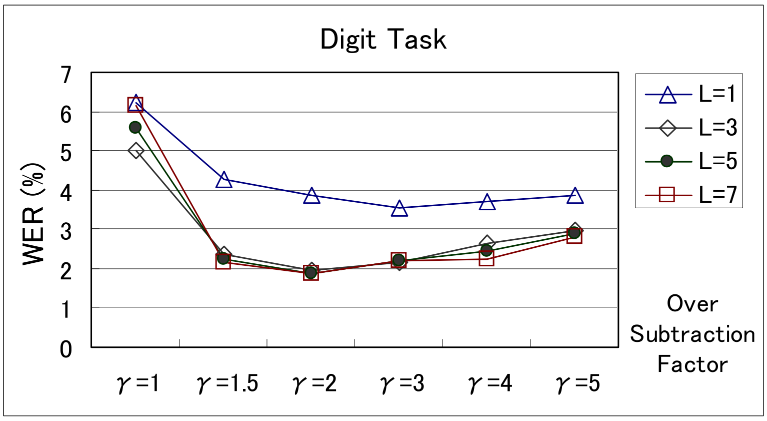


Figure 5.6. Word error rate using the proposed method 1 for various values of the over-subtraction factor *γ* and the length of the adaptive weights *L*, for the digit recognition task.

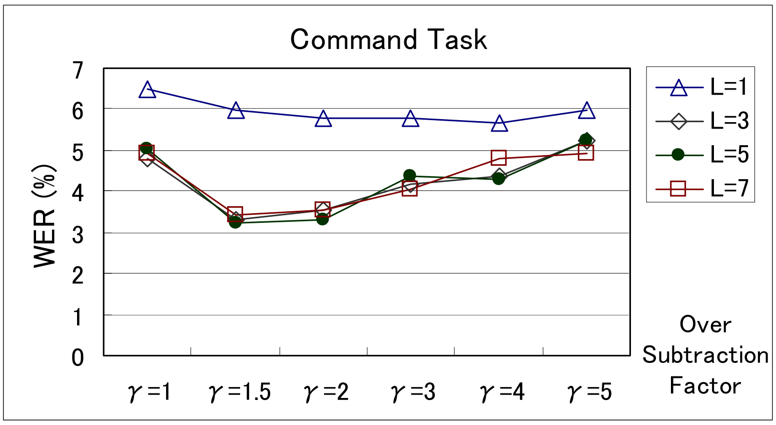


Figure 5.7. Word error rate using the proposed method 1 for various values of the over-subtraction factor *γ* and the length of the adaptive weights *L*, for the command recognition task.

On the other hand, the performance of speech-silence detector is critically important for the front-end processing including echo cancellation, spectral subtraction and the proposed method. In this experiment, we used the oracle speech-silence information for the front-end processing. This was prepared using the data without adding the music. In order to get the most reliable speech-silence information, we installed two additional microphones to do the speech-silence detection based on the coherence between the microphone outputs [AM97].

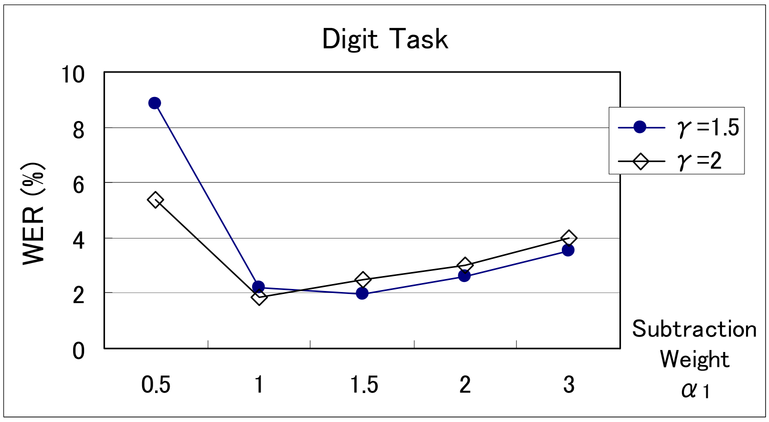


Figure 5.8. Word error rate using the proposed method 1 for various values of the base subtraction weight $\alpha_1$ and the over-subtraction factor $\gamma$ for the echo component, for the digit recognition task.

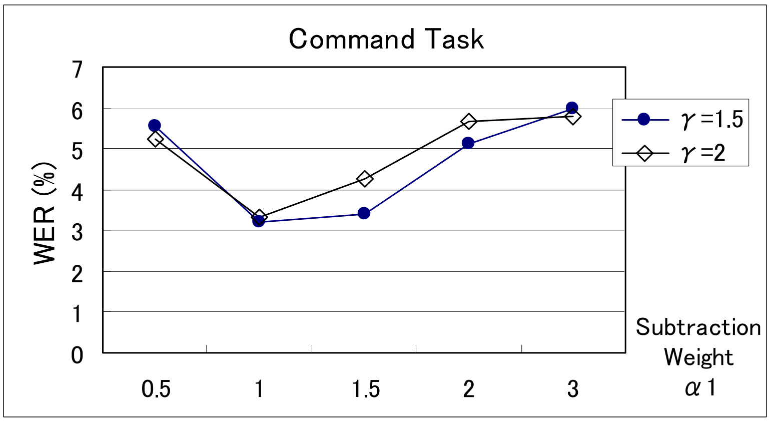


Figure 5.9. Word error rate using the proposed method 1 for various values of the base subtraction weight $\alpha_1$ and the over-subtraction factor $\gamma$ for the echo component, for the command recognition task.

We measure the error rate with WER (Word Error Rate) defined in Equation (5.15), because the evaluation involves lots of insertion words and deletion words with the digit task.

$$WER = \frac{(\text{number of substituted words}) + (\text{number of inserted words}) + (\text{number of deleted words})}{(\text{number of all expected words})}, \quad (5.15)$$

Table 5.2. Detailed word error rates for the conventional methods and the proposed methods

Digit Task WER (%)

| | Stationary | City Drv. | Highway | Average |
|---|---|---|---|---|
| Case 1: SS only (reference without music) | 0.5 | 0.6 | 1.1 | 0.8 |
| Case 2: SS only | 3.1 | 14.1 | 12.1 | 9.8 |
| Case 3: Echo Canceller + SS | 1.4 | 2.2 | 3.6 | 2.4 |
| Case 4: Proposed Method 1 (L=5, $\alpha_1$=1.0, $\alpha_2$=2.0) | 1.0 | 2.0 | 2.6 | 1.9 |
| Case 5: Proposed Method 2 with Preprocessor (L=5, $\alpha_1$=1.0, $\alpha_2$=2.0) | 1.0 | 1.2 | 1.5 | 1.2 |

Command Task WER (%)

| | Stationary | City Drv. | Highway | Average |
|---|---|---|---|---|
| Case 1: SS only (reference without music) | 2.6 | 1.0 | 3.5 | 2.4 |
| Case 2: SS only | 3.5 | 11.9 | 12.5 | 9.3 |
| Case 3: Echo Canceller + SS | 4.2 | 1.9 | 4.8 | 3.6 |
| Case 4: Proposed Method 1 (L=5, $\alpha_1$=1.0, $\alpha_2$=2.0) | 3.2 | 2.6 | 4.2 | 3.3 |
| Case 5: Proposed Method 2 with Preprocessor (L=5, $\alpha_1$=1.0, $\alpha_2$=2.0) | 2.9 | 1.0 | 3.2 | 2.4 |

Figure 5.6 and Figure 5.7 show the resulting WERs depending on the various over-subtraction factors $\gamma$ and the lengths of the adaptive weights $L$, for the proposed Method 1. This used the on-line formula with the parameters $Const$=$10^3$, $\theta$=0.1, and $\varepsilon$=$10^4$. The WERs are averaged values for the three speeds and the 24 subject speakers. Based on the results, the over-subtraction of the echo improved the recognition accuracy. The optimum factor was around 1.5 to 2.0. Also, introducing a sufficient length of adaptive weights improved the recognition accuracy. In the following experiment, we select $\gamma$=2.0 and $L$=5 as the default setting.

Figure 5.8 and Figure 5.9 show the resulting WERs depending on the various over-subtraction factors $\gamma$ and the base subtraction weight $\alpha_1$, for the proposed Method 1. Based on the results, the optimum weight was around 1.0 to 1.5, which is close to the

Table 5.3. Word error rates for component reduction only cases

Digit Task WER (%)

| | Stationary | City Drv. | Highway | Average |
|---|---|---|---|---|
| Case 4: Proposed Method 1 ($L$=5, $\alpha_1$=1.0, $\alpha_2$=2.0) | 1.0 | 2.0 | 2.6 | 1.9 |
| Case 6: Proposed Method 1 --- stationary noise reduction only ($L$=5, $\alpha_1$=1.0, $\alpha_2$=0.0) | 7.6 | 20.8 | 19.6 | 16.0 |
| Case 7: Proposed Method 1 --- echo reduction only ($L$=5, $\alpha_1$ =0.0 , $\alpha_2$ =2.0) | 1.3 | 2.5 | 3.5 | 2.5 |

Command Task WER (%)

| | Stationary | City Drv. | Highway | Average |
|---|---|---|---|---|
| Case 4: Proposed Method 1 ($L$=5, $\alpha_1$=1.0, $\alpha_2$=2.0) | 3.2 | 2.6 | 4.2 | 3.3 |
| Case 6: Proposed Method 1 --- stationary noise reduction only ($L$=5, $\alpha_1$=1.0, $\alpha_2$=0.0) | 3.5 | 3.2 | 6.1 | 4.3 |
| Case 7: Proposed Method 1 --- echo reduction only ($L$=5, $\alpha_1$=0.0, $\alpha_2$=2.0) | 4.2 | 4.5 | 4.8 | 4.5 |

value used for the acoustic model training. In the following experiment, we select $\alpha_1$=1.0 as the default setting.

Table 5.2 shows performance comparisons with the conventional methods. Case 1 is for reference. Music was NOT mixed into the test data. It was processed by conventional spectral subtraction and decoded. Automatic speech recognition performs very well for the stationary cruising noise. Case 2 and the following cases have music mixed into the test data. Case 2 processed the test data only with conventional spectral subtraction. Since there is no echo cancellation, the recognition performance was severely degraded. Case 3 processed the test data by using the conventional combination of echo cancellation and spectral subtraction as shown in Figure 5.1. The echo canceller was N-LMS in the time domain with a tap length of 2,048. The recognition performance was much improved from Case 2 as a result of the echo canceller. Case 4 processed the test data using the proposed Method 1 with the parameters $\gamma$=2 and $L$=5. $L$ was selected so to be comparable with the tap length in Case 3. This shows performance superior to Case 3. Case 5 processed the test data by the proposed Method 2 with the parameters $\gamma$=2 and $L$=5. The tap length of the preprocessing echo canceller was 512. The performance is improved in favor of the preprocessing.

Table 5.3 shows the performance of the two additional cases in order to measure the contributions of the proposed Method 1 to the stationary noise reduction and the echo reduction separately. Case 6 reduces only the stationary noise component, and Case 7 reduces only the echo component, while the adaptation processes were the same as in Case 4. Based on the results, the echo component reduction of the proposed method was very effective in the digit task. Also, the stationary noise reduction of the proposed method was effective in the command task.

## 5.6. Concluding Remarks

In order to reduce both background noise and echo effectively for automatic speech recognition in a car, we proposed a new method that adapts echo cancellation and spectral subtraction simultaneously. The stationary noise component is estimated through the adaptation of an echo canceller. As the echo canceller is also implemented using spectral subtraction, the echo component can be further reduced by introducing over-subtraction. We can still use the existing acoustic model trained only with the background noises and spectral subtraction, since we kept the subtraction weight the same as for the stationary noise and introduced over-subtraction only for the echo. The performance can be improved even more by introducing a shot-tap echo cancellation as a preprocessor. In our experiment, this method showed superior recognition accuracy compared to the conventional combination of echo cancellation and spectral subtraction.

# 6. Local Peak Enhancement

## 6.1. Introduction

The performance of automatic speech recognition in automobiles is affected by various noises. Beamformer [SSL+03] technology reduces directional noise such as voices from passengers and sounds coming from a car radio, TV, or CD player. However, it does not have sufficient signal recovery in very low SNR situations with ambient noise (such as “Fan high” or “Window open”) unless the size of the beamformer is very large. For single channel signal processing, existing noise reduction algorithms such as a Wiener Filter [ETS02] or Spectral Subtraction (SS) [Bol79] are known to improve the accuracy, but improvements are still needed in those situations. Therefore, different approaches beyond reducing noise should be combined with existing noise reduction algorithms.

One of the candidate approaches involves enhancements of the harmonic structures in human voices. Comb filtering [TO98] and its variants [GR01] were proposed and showed good performance, especially in mixed speech cases. However, they are rarely integrated into commercial ASR products, and especially not for automobiles. This is because designing a comb filter relies on the accurate estimation of F0 (the fundamental frequency or pitch) and the accurate discrimination between voiced and unvoiced speech. It was reported that errors at this stage have detrimental effects on the performance [NIZ03]. Szymanski et al. proposed Comb Filter Decomposition [SB05] that does not require F0 estimation, but their experiment was limited to white Gaussian noise.

Another candidate would use a matching algorithm to put larger weights on frequencies having larger spectral powers as the decoder calculates likelihoods [SS80][NSI+04]. This is based on the assumption that frequencies having more spectral power are noise robust and most likely to be the formant frequencies in voiced speech frames. Huang et al. enhanced the logic for the MFCC domain [HHS+06], but this involved adding autocorrelation into their decoding process.

In this chapter, we propose a novel approach for the speech enhancement. It uses a filter designed to enhance the harmonic structure which is observed as local peaks at regular distances in the spectrum domain. It does not depend on F0 or voiced/unvoiced detection. Since it works as a front-end for both training and decoding, it does not require any changes in existing decoders. This new method will be referred to as LPE (Local Peak Enhancement) in the following sections.

## 6.2. Proposed Method (LPE)

Figure 6.1 shows the whole process of LPE and sample outputs at each step for both a voiced frame and a noise frame. The process is the same for entire frames, but the generated filter looks very different depending on whether or not the frame is voiced speech, as shown in the figure.

In the first step, an observed spectrum $y_T(j)$ is converted to a log power spectrum $Y_T(j)$.

$$Y_T(j) = \log(y_T(j)), \tag{6.1}$$

where, the index *T* is a frame number and *j* is the bin number of the DFT corresponding to the subband frequency. The process described in this section should be performed for each *T*.

Then the log power spectrum is converted to a cepstrum $C_T(i)$ by using $D(i,j)$, a DCT (Discrete Cosine Transformation) matrix.

$$C_T(i) = \sum_j D(i,j) \cdot Y_T(j). \tag{6.2}$$

The cepstra represent the curvatures of the log power spectra. The lower cepstra correspond to long oscillations, and the upper cepstra correspond to short oscillations. We need only the medium oscillations. The range of the cepstra is chosen to cover possible harmonic structures in the human voice. Therefore the lower and the upper cepstra should be filtered out.

$$\hat{C}_T(i) = \begin{cases} \varepsilon \cdot C_T(i) & \text{if } i < \mathrm{I}_{\text{lower}} \text{ or } i > \mathrm{I}_{\text{upper}} \\ C_T(i) & \text{otherwise} \end{cases}, \tag{6.3}$$

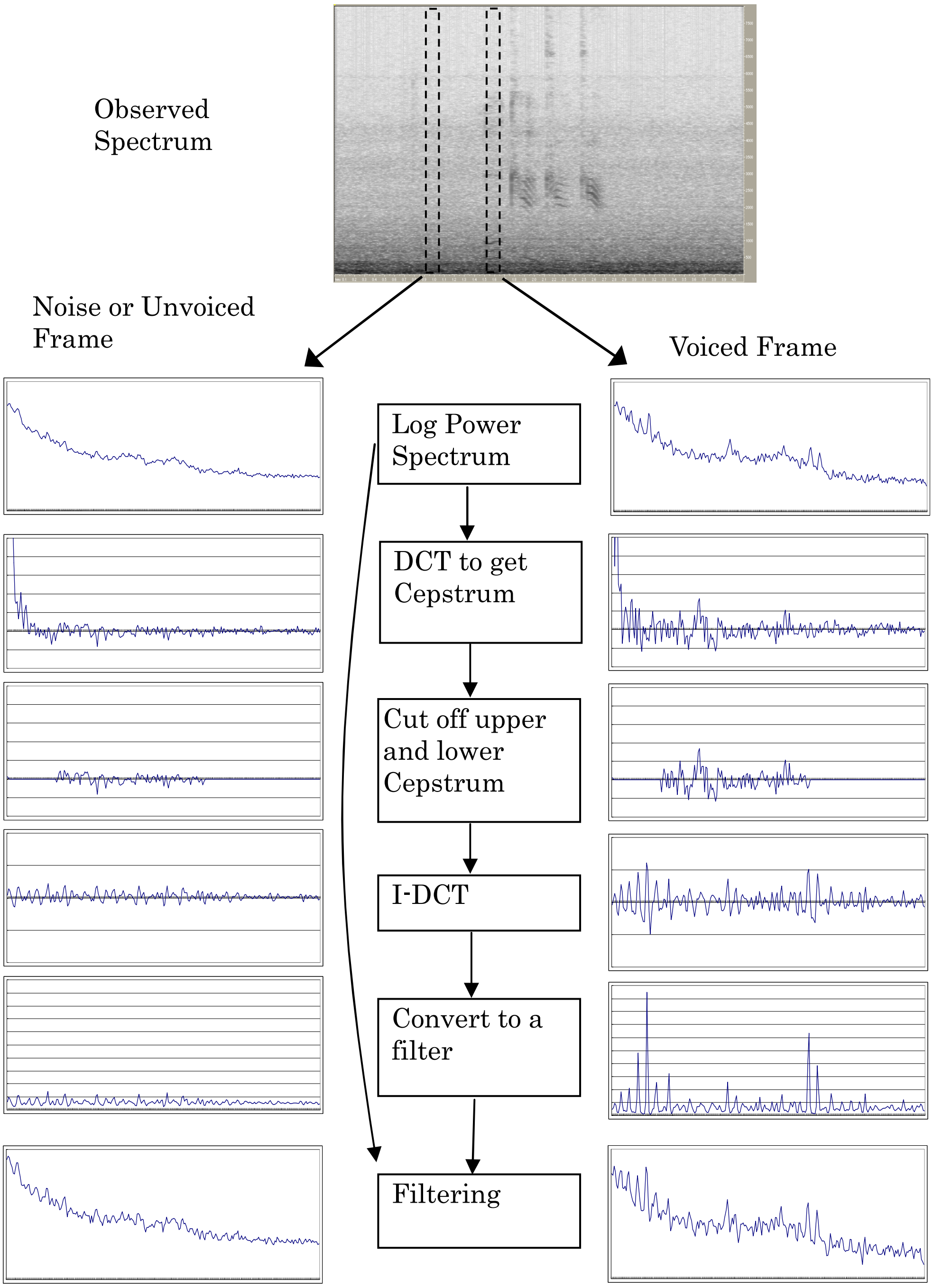


Figure 6.1. Process of LPE.

In this experiments, $I_{lower}$=40 and $I_{upper}$=160 for a 16 kHz sampling frequency with an FFT length of 512 samples. This corresponds to an F0 range from 100 Hz to 400 Hz for the human voice, with $\varepsilon$ being close to zero. We set it to $10^{-3}$.

The filtered cepstrum $\hat{C}_T(i)$ is converted back to a log power spectrum by using an I-DCT.

$$W_T(j) = \sum_i D^{-1}(j,i) \cdot \hat{C}_T(i). \tag{6.4}$$

Then it is converted back to a linear power spectrum, and it is normalized so that the average is 1.0.

$$w_T(j) = \exp(W_T(j)). \tag{6.5}$$

$$\overline{w}_T(j) = w_T(j) \cdot \frac{N_{bin}}{\sum_k^{N_{bin}} w_T(k)}, \tag{6.6}$$

where $N_{bin}$ is the number of bins used in the FFT. The filter is obtained as $\overline{w}_T(j)$. Finally, the enhanced output $z_T(j)$ is obtained as

$$z_T(j) = \overline{w}_T(j) \cdot y_T(j). \tag{6.7}$$

In order to reduce the amount of computation, the steps of the Equations (6.2), (6.3), and (6.4) can be combined into a single step using the pre-calculated matrix $A$ as follows.

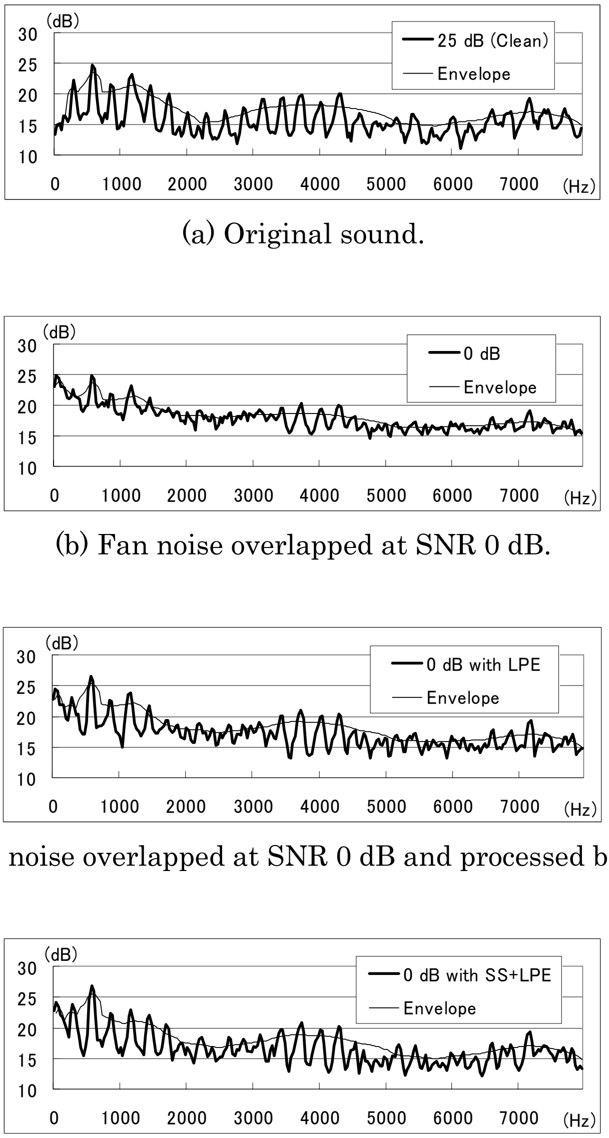
$$\Lambda(i,j) = \begin{cases} 0 & \text{if } i \neq j \\ \varepsilon & \text{if } i = j \text{ and } (i < \mathrm{I}_{\text{lower}} \text{ or } i > \mathrm{I}_{\text{upper}}) \\ 1 & \text{otherwise} \end{cases}, \tag{6.8}$$

$$A = D^{-1} \Lambda D. \tag{6.9}$$

$$W_T = A Y_T. \tag{6.10}$$

As shown in Figure 6.1, the filter for LPE is derived directly from the observed spectrum. Therefore, F0 estimation is not required. For a noise frame or an unvoiced speech frame, it will be designed to be almost flat. This means LPE does almost nothing to such frames, and therefore, LPE does not require voiced/unvoiced detection.

For voiced speech frames, the LPE filter is designed to enhance the harmonic structures in the observed spectrum. Unlike a comb filter, the LPE filter is not uniform over all frequencies. It is more focused on the frequencies where harmonic structures are observed in the input spectrum. Therefore the acoustic model should be retrained with LPE for automatic speech recognition.



(a) Original sound.

(b) Fan noise overlapped at SNR 0 dB.

(c) Fan noise overlapped at SNR 0 dB and processed by LPE.

(d) Fan noise overlapped at SNR 0 dB and processed by LPE after SS.

Figure 6.2. Spectrums of vowel /u/ recorded in a stationary car with and without fan noise overlapping at the specified SNR. The spectrum envelope is plotted with Mel-Filtering.

Figure 6.2 shows how a spectrum is degraded by a noise. In Figure 6.2(a), the original clean spectrum shows three formants around 600 Hz, 1200 Hz, and 3500 Hz. However, in Figure 6.2(b), they are less conspicuous, and the spectrum contour is close to flat. In contrast, LPE retains more of the characteristics of the formants, as shown in Figure 6.2(c). The combination of SS and LPE retains even more, as shown in Figure

6.2(d). An advantage of LPE is that voiced speech immersed in heavy noise should be more distinct and distinguishable for decoding.

Harmonic structures are conspicuous around frequencies having larger spectral powers in the voiced speech frames, and they are most likely to be formant frequencies. Therefore, this approach inherently involves formant enhancement as well as harmonic enhancement, under the assumption that the noise has a broad spectrum and the harmonic structure is not locally destroyed by the noise.

# 6.3. Experiments

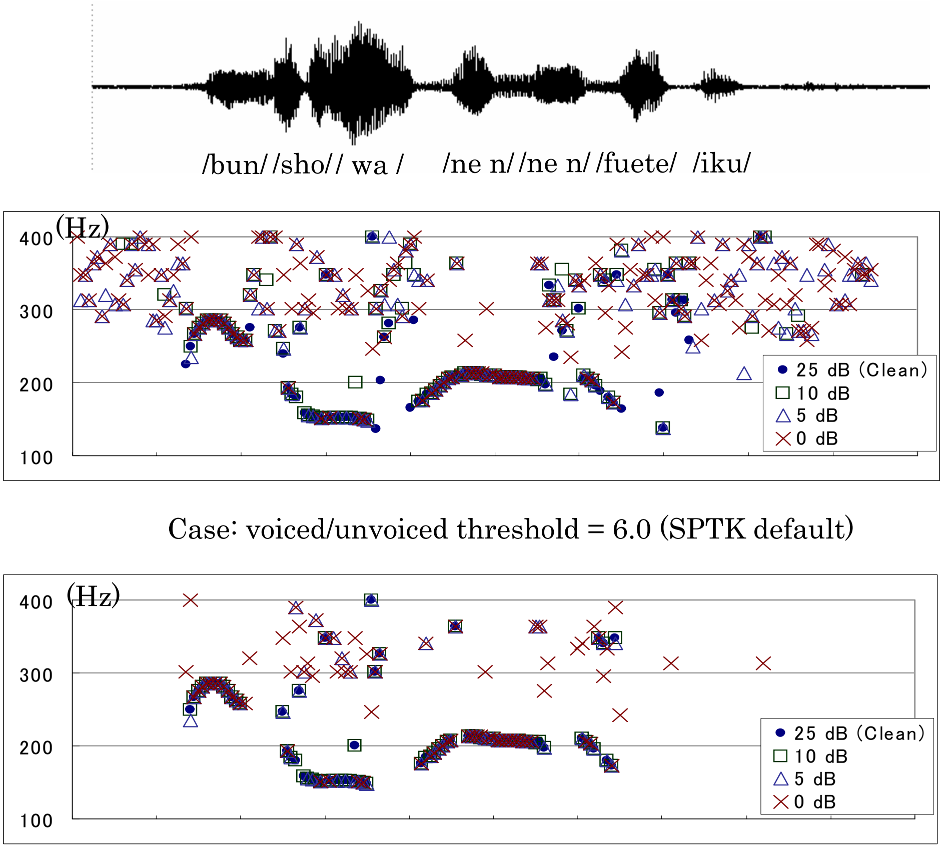
## 6.3.1. Testing data

We used CENSREC-3, an evaluation framework for isolated Japanese word recognition in actual moving-automobile environments. This data was collected by IPSJ, and is widely used to evaluate noise reduction algorithms [FNT+05]. It has speech data both for training and testing for automatic speech recognition using matched acoustic models.

The test data in the database was recorded under 16 environmental conditions using combinations of three vehicle speeds and six kinds of in-car environments as shown in Table 6.1. A total of 14,216 utterances spoken by 18 speakers (8 males and 10 females) were recorded at a 16 kHz sampling frequency. The performance is measured with word accuracy as CENSREC-3 defines.

For training, each driver’s speech saying phonetically balanced sentences was recorded under two conditions: while idling and while driving on a city street in a normal in-car environment. A total of 14,050 utterances spoken by 293 drivers (202 males and 91 females) were recorded with a close-talking microphone and a hands-free microphone.

In this experiment, we used only hands-free microphone data for both training and testing. The acoustic models were trained with both idling data and driving data for the front-end processing being tested. This corresponds to Condition 3 as defined in CENSREC-3. The evaluation category is zero, which means no changes at the backend.



Case: voiced/unvoiced threshold = 6.0 (SPTK default)

Case: voiced/unvoiced threshold = 7.0

Figure 6.3. F0 output by Pitch command in SPTK. For unvoiced frames, SPTK outputs zero. The test data was prepared by overlapping noise at different SNRs.

## 6.3.2. Conventional methods

Comb-filtering needs F0 estimation and voiced/unvoiced detection. We used the “Pitch command” in SPTK-3.0 [SPTK] to obtain this information. We used a low-end frequency of 100 Hz and an upper frequency limit of 400 Hz, so to be compatible with LPE experiment. The voiced/unvoiced threshold was empirically set to 7.0, because it gave us a better result than the SPTK default value. Figure 6.3 shows an example of F0 information by SPTK. We see many outliers in the low SNR conditions. Also, the vowels in the last part of the sentence were not recognized as voiced sounds. Based on the F0 and voiced/unvoiced information, the comb filter was designed in the spectrum domain for each frame as in Equation (6.11), and the comb-filtering output was obtained using Equation (6.12).

$$Wcomb_T(j)=\begin{cases}1.0 & \text{if } T \text{ is unvoiced frame}\\ 1.0 & \text{if } T \text{ is voiced frame and } j \text{ is harmonic bin}\\ 0.01 & \text{otherwise}\end{cases}, \tag{6.11}$$

$$z_T(j)=y_T(j)\cdot Wcomb_T(j). \tag{6.12}$$

For the combination of LPE and existing noise reduction algorithms, SS and ETSI Advanced Front-End (ES202-050) [ETS02] were introduced in the evaluations. For SS processing, the first 0.1 second of each utterance was assumed to be a non-speech segment where the noise spectrum $N(j)$ could be estimated. The SS output was obtained as Equation (6.13).

$$z_T(j)=\begin{cases}y_T(j)-\alpha\cdot N(j) & \text{if } y_T(j)-\alpha\cdot N(j)\geq\beta\cdot N(j)\\ \beta\cdot N(j) & \text{otherwise}\end{cases}, \tag{6.13}$$

In this experiment, the subtraction weight $\alpha$ was set to 1.0, and the flooring coefficient $\beta$ was set to 0.1.

### 6.3.3. Results of standalone test

Table 6.1 shows the resulting word accuracies for various environmental conditions. The baseline is the evaluation without using any speech enhancement or noise reduction algorithms. Table 6.1 also shows the estimated SNRs of the test data using the VAD (Voice Activity Detection) information came from the ETSI ES202-050. Note that the accuracy of SNR depends on the VAD information. Table 6.2 shows the estimated SNRs of the training data. We see CENSREC-3 trains an acoustic model at relatively better SNRs than for the test data. Therefore, speech enhancement and noise reduction are expected to help the test performance.

Table 6.1. Word accuracy and estimated SNRs according to the environmental conditions. SNR was calculated for the baseline data after a 250 Hz high-pass filtering

| CENSREC-3 | | | | Word Accuracy (%) | | |
|---|---|---|---|---|---|---|
| (Condition 3) | | | SNR (dB) | Base Line | Comb Filter | LPE |
| Idling | Audio off | Normal | 16.2 | 99.7 | 98.8 | 99.7 |
| | | Hazard on | 15.3 | 98.7 | 95.3 | 96.8 |
| | | Fan low | 11.3 | 94.6 | 87.7 | 94.8 |
| | | Fan high | 6.2 | 53.4 | 55.0 | 60.3 |
| | | Window open | 10.5 | 90.0 | 85.4 | 92.7 |
| | Audio on | | 9.9 | 81.4 | 73.2 | 56.4 |
| Low speed | Audio off | Normal | 10.9 | 99.3 | 96.6 | 98.7 |
| | | Fan low | 9.7 | 95.1 | 91.8 | 94.7 |
| | | Fan high | 6.7 | 62.7 | 66.2 | 69.1 |
| | | Window open | 9.3 | 66.2 | 70.6 | 74.3 |
| | Audio on | | 6.7 | 79.0 | 74.7 | 61.6 |
| High speed | Audio off | Normal | 7.5 | 95.0 | 94.3 | 96.2 |
| | | Fan low | 7.1 | 89.0 | 86.7 | 89.7 |
| | | Fan high | 6.1 | 58.2 | 62.1 | 63.6 |
| | | Window open | 7.2 | 22.2 | 35.8 | 40.4 |
| | Audio on | | 3.9 | 79.3 | 69.0 | 66.6 |
| Average (ALL) | | | | 78.9 | 77.6 | 78.4 |
| Average (Audio off) | | | | **78.8** | **78.9** | **82.4** |
| Average (Audio on) | | | | 79.9 | 72.3 | 61.5 |
| Average (Fan high) | | | | 58.1 | 61.1 | 64.3 |
| Average (Window open) | | | | 59.5 | 63.9 | 69.1 |

LPE enhances the local peaks considered to be harmonic structures. Therefore, a drawback is expected with LPE when the background noise contains music or speech from audio devices such as a radio, TV, or CD player, because the filter is designed to enhance that audio, too. This is a known restriction of LPE. Comb filtering shares this problem, and a multi-pitch tracker was proposed to address it [WWB02]. In this chapter, we accept this restriction and we focus only on the results of the “Audio off” cases. The restriction should not matter with current car navigation systems, because most of them are designed to disable audio on pushing a talk button. Also, we can expect an echo canceller to eliminate audio components before processing by LPE.

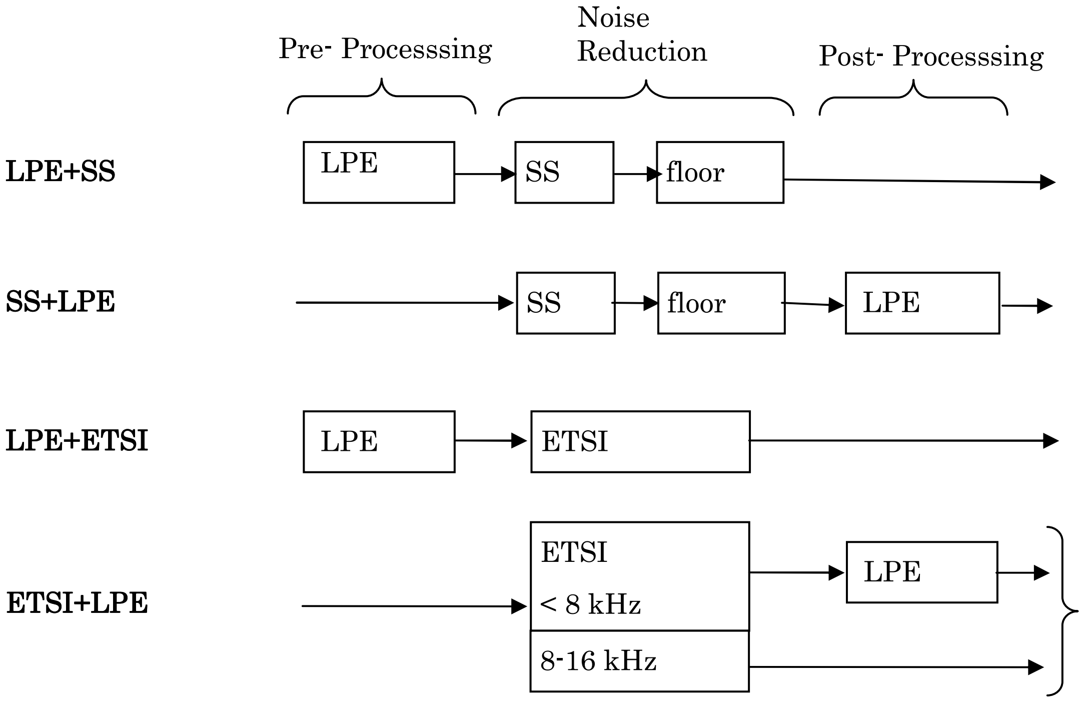


Figure 6.4. Combinations of LPE and noise reduction algorithms.

For the average "Audio off" case, LPE outperformed the baseline by 17.0% in error reduction. Most of the improvement was gained in very noisy conditions of "Fan high" and "Window open" conditions with error reductions of 14.8% and 23.7%, respectively. An advantage of LPE is that voiced speech immersed in heavy noise should be more distinct and distinguishable for decoding. Comb-filtering also improved the accuracy in these conditions. However, the improvement was smaller than LPE.

In relatively clean conditions such as "Normal" or "Fan low" at "Idling" or "Low speed", the accuracy of LPE was almost the same or slightly degraded from the baseline. However, the degree of loss was small enough for practical use. In contrast, comb-filtering shows noticeable degradation in these conditions, possibly caused by inaccurate F0 estimation and errors in the voiced/unvoiced detection.

Table 6.2. Estimated SNRs of CENSREC-3 training data. SNR was calculated for the baseline data after a 250 Hz high-pass filter

| Training Data | SNR (dB) |
|---|---|
| Idling | 21.1 |
| Driving | 18.7 |

Table 6.3. Word accuracy with existing noise reduction methods and the combinations of LPE

| CENSREC-3<br>(Condition 3) | | | Word Accuracy (%) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | SS | LPE + SS | SS + LPE | ETSI | LPE + ETSI | ETSI + LPE |
| Idling | Audio off | Normal | 99.8 | 99.6 | 99.0 | 100.0 | 99.8 | 100.0 |
| | | Hazard on | 96.8 | 96.9 | 96.7 | 98.1 | 98.1 | 98.6 |
| | | Fan low | 95.2 | 95.7 | 95.3 | 99.2 | 99.6 | 99.7 |
| | | Fan high | 58.1 | 65.7 | 67.6 | 85.3 | 89.9 | 88.9 |
| | | Window open | 90.4 | 94.1 | 93.8 | 97.2 | 98.2 | 98.0 |
| | Audio on | | 74.8 | 57.0 | 61.4 | 89.5 | 77.7 | 82.6 |
| Low speed | Audio off | Normal | 98.4 | 97.8 | 97.5 | 99.7 | 98.6 | 99.7 |
| | | Fan low | 94.6 | 94.4 | 94.2 | 97.8 | 97.5 | 98.7 |
| | | Fan high | 66.9 | 71.1 | 74.3 | 87.9 | 89.5 | 91.5 |
| | | Window open | 72.4 | 76.7 | 78.5 | 87.0 | 89.6 | 88.7 |
| | Audio on | | 79.5 | 62.1 | 62.8 | 90.8 | 81.3 | 87.6 |
| High speed | Audio off | Normal | 97.8 | 95.3 | 95.9 | 98.1 | 97.2 | 98.8 |
| | | Fan low | 91.7 | 91.9 | 91.6 | 96.7 | 94.8 | 97.6 |
| | | Fan high | 61.3 | 68.3 | 69.6 | 88.4 | 89.1 | 88.1 |
| | | Window open | 40.1 | 44.2 | 45.4 | 65.0 | 69.4 | 66.7 |
| | Audio on | | 84.3 | 67.4 | 69.1 | 92.8 | 84.0 | 89.7 |
| Average (ALL) | | | 81.3 | 79.8 | 80.7 | 92.1 | 90.9 | 92.1 |
| Average (Audio off) | | | **81.8** | **84.0** | **84.6** | **92.3** | **93.2** | **93.5** |
| Average (Audio on) | | | 79.5 | 62.2 | 64.4 | 91.0 | 81.0 | 86.6 |
| Average (Fan high) | | | 62.1 | 68.4 | 70.5 | 87.2 | 89.5 | 89.5 |
| Average (Window open) | | | 67.6 | 71.7 | 72.6 | 83.1 | 85.7 | 84.5 |

### 6.3.4. Results of combination test

LPE can be used in combination with existing noise reduction algorithms. In Table 6.3, SS and ETSI ES202-050 were introduced in the evaluations. Figure 6.5 shows the average word accuracies in combined "Audio off" cases.

As shown in Figure 6.4, "LPE+SS" means LPE pre-processes the input of SS, and "SS+LPE" means LPE post-processes the output of SS. Since ETSI ES202-050 splits the 16 kHz input into a less-than-8-kHz part and an upper-8-kHz part, "ETSI+LPE" applied LPE only to the less-than-8-kHz part of the ETSI ES202-050 output.

The "SS+LPE" combination outperformed SS or LPE alone, as well as the baseline. It reduced the average error rate for the "Audio off" case by 27.3% from the baseline. Likewise, the "ETSI+LPE" combination showed the best performance, reducing the error rate by 69.2%.

## 6.4. Concluding Remarks

We are proposing a new approach to speech enhancement to improve automatic speech recognition in very noisy conditions. It generates a filter to enhance the harmonic structure observed in the input spectrum, without relying on F0 estimation and voiced/unvoiced detection. Experiments using automatic speech recognition showed this method significantly improved the accuracy in very noisy conditions such as "Fan high" or "Window open." However, it showed some drawbacks in "Audio on" cases. This method can be combined with existing noise reduction algorithms such as SS and ETSI ES202-050 for further improvements.

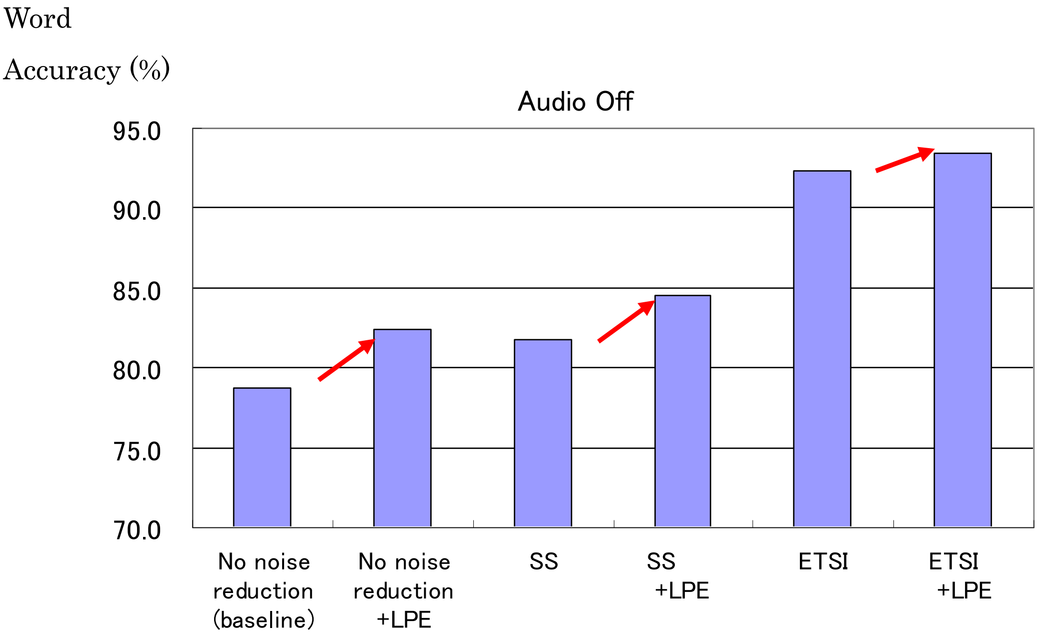


Figure 6.5. Averaged word accuracy of “Audio off” cases for the combinations of noise reduction method and LPE.

# 7. Conclusion

## 7.1. Thesis Summary

To increase the applications of ASR in the real world, improved robustness against noise is a key. For better noise reduction, we may need to consider such questions as “What is the definition of noise?” Someone may say “Any undesired signal is noise.” Yet that begs the question of how can we determine whether a signal is desired or undesired? In this dissertation, features such as the direction of a sound, correlations to already known reference signals, constancy, or harmonic structure are used as cues to determine whether some sound is noise or signal, depending on the assumed noise characteristics. Using these traits, three novel approaches are proposed to perform noise reduction or speech enhancement.

In Chapter 3, a new microphone array technology named Profile Fitting (PF) is proposed. It focuses on the directivity of arriving sounds. The directivity is measured as a distribution profile. PF decomposes an observed profile into certain known profiles so as to extract only the target signal. Experiments in a non-reverberant environment with a dictation system configured with 2 microphones showed PF reduced error rate by more than 20% from the best results of the conventional beamformers (2-ch Adaptive SS). In a realistic environment, the extent of the improvement was 11%.

In Chapter 4, PF is further discussed in an application of sound source localization. It is shown that PF is noise robust and the concept of profiles allows extended sound source localization in combination with sound reflectors.

In Chapter 5, a new echo canceller named SSEC (Simultaneous adaptation of spectral Subtraction and Echo Cancellation) is proposed. In automobiles, sound from audio devices may be overlapping with the speech signal. In the echo canceller

Table 7.1. Coverage of noise variations in automobiles and the current achievement levels in subjective views

| | Cruising noise | Fan | Radio, Navi, CD | Passenger voice | Door slam, Wiper | Road bump | Outside events | Current achievement levels |
|---|---|---|---|---|---|---|---|---|
| PF | √ | √ | √ | √ | | | (√) | • Acceptable accuracy<br>• Need some improvement to work in real-time |
| SSEC | √ | √ | √ | | | | | • Satisfactory accuracy<br>• Can work in real-time |
| LPE | √ | √ | | | | | | • Acceptable accuracy<br>• Can work in real-time |
| Acoustic model | √ | √ | | | (√) | (√) | | |

framework, such sources are treated as echos to be cancelled. However, conventional echo cancellers do not perform well in noisy environments such as moving cars. SSEC solves this difficulty by simultaneous adaptation of echo cancellation and spectral subtraction. This assumes that cruising noise can be treated as stationary. In the experiment, SSEC showed superior recognition accuracy compared to the conventional combination of echo cancellation and spectral subtraction.

In Chapter 6, a new speech enhancement method named Local Peak Enhancement (LPE) is proposed. The objective of LPE is to retrieve a voiced speech signal immersed in broadband noise with a very low SNR, such as occurs in a "window open" or "fan high" situation in a moving car. It uses the harmonic structure in the human voice and assumes that the noise does not contain the same structure. Unlike a comb filter, LPE does not require pitch estimation or voiced/unvoiced detection. In the "Audio off" case, LPE outperformed the baseline by 17.0% in error reduction, and it showed further improvements in combination with existing noise reduction methods.

Table 7.1 summarizes the coverage of noise variation in automobiles with the above three methods. It also indicates over-all achievement measures at the current technology levels, from my subjective view. The coverage is almost satisfactory, but they still require some improvement in accuracy or speed. As the current PF formulation requires intensive computation, it is somewhat heavy to run it in real-time on many of the current embedded devices. It may require some improvement in the implementation to speed up the whole process, or more powerful processors that possibly appear in the near future. LPE improved ASR accuracy in “window open” or “fan high” situation in a moving car, but the accuracy needs to be further improved up to around 90% to be acceptable for many of the users.

## 7.2. Future Research

This dissertation proposed three novel approaches for noise reduction and speech enhancement to improve the accuracy in automatic speech recognition. They are designed to work in specific configurations and with specific types of noise. In other words, they have their own limitations and they are not universal solutions for every situation as shown in Table 7.2.

PF can reduce both directional and ambient (non-directional) noise. It supports non-stationary noise including music and human speech. However, the major drawbacks of this method are the requirement for multiple microphones and the availability of pre-measured template profiles. Also, the location of noise source must be different from the signal source.

SSEC can reduce any kind of noise whose reference signal is available. That can be non-stationary noise including music and human speech. The location of the noise source does not matter. However, the availability of reference signals is the critical requirement, which is sometimes not satisfied in actual situations.

LPE can enhance speech signals degraded to very low SNRs. It does not require multiple microphones or reference signals. However, LPE does not allow harmonic structure in noise and the spectrum of the noise need to be broad. This means the noise cannot be music or human speech.

Table 7.2. Noise reduction capabilities and requirements of the proposed methods

| | Noise reduction capabilities | | Requirements | |
|---|---|---|---|---|
| | Work with non-stationary noise? | Work with harmonic noise? | Multiple microphones required? | Reference signal required? |
| PF | Yes | Yes | Yes | No |
| SSEC | Yes<br>(No for ambient noise) | Yes | No | Yes |
| LPE | Yes | No | No | No |

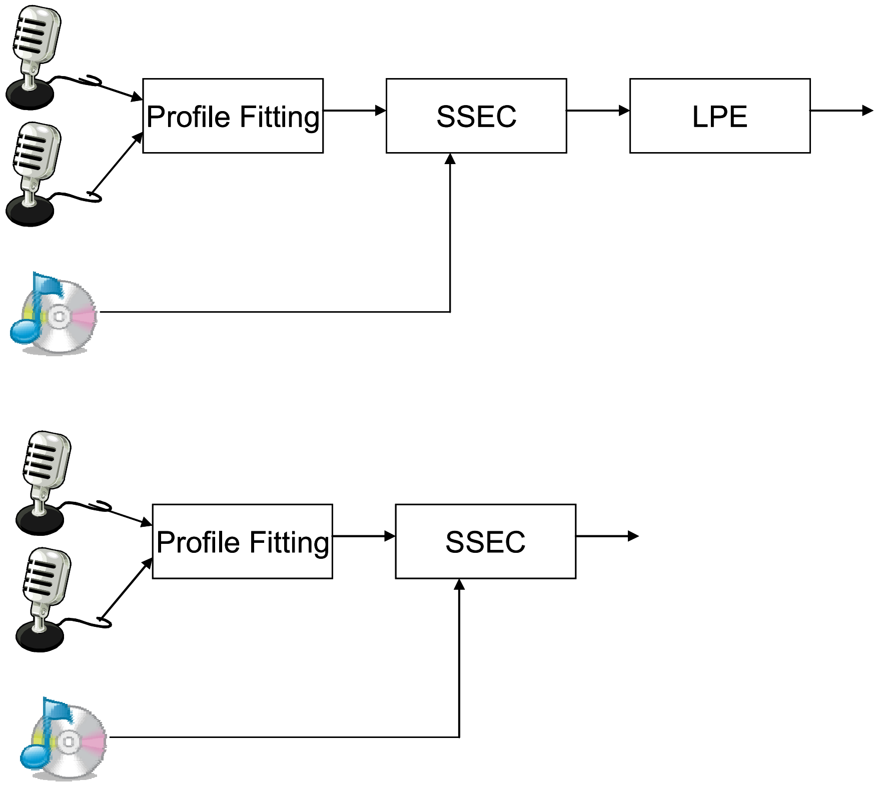


Figure 7.1. Possible combinations of the proposed methods.

Therefore, some new methods were desired, which can reduce any kind of noise including music and human speech without using multiple microphones or reference signals. There are already several research projects with these goals [AS01][KH03][FN05]. However, they still more improvements for practical applications in computational cost and accuracy.

Another approach would be a combination of the three proposed methods. Figure 7.1 shows some possible combinations of the methods. In automobiles, PF is unable to reduce guidance messages from the car navigation system, which are broadcasted from a loudspeaker on the driver's side, then SSEC successively processes the output to reduce it. LPE may be introduced to enhance the speech output, under the assumption that remaining noise does not contain harmonic structure.

This dissertation only discussed the noise robustness of automatic speech recognition. However, from the viewpoint of human interface systems, we may also need to consider two major capabilities for the near future, "Always Listening" and "Barge-In." "Always Listening" would allow us to talk to a system without pushing a talk button. A "Barge-In" system would allow us to initiate utterances before the completion of system messages. Both of these are essential capabilities for natural man-machine interactions, especially in robot applications. They require a critical level of noise reduction technology, as well as speech command detection technology. I will continue my research on noise reduction for automatic speech recognition, focusing on ways to solve these problems.

## 7.3. Future Applications

Improved levels of noise reduction will make various advanced ASR applications a reality.

In automobiles, speech will become the main interface to input complicated information. Even in a very noisy situation such as an open car cruising at high speed, the driver's natural phrases will be correctly transcribed and interpreted for the desired actions. Since the supported vocabulary will be very large, the recognized words can even be used for Internet searches as with a personal computer. The system will be able to search for music titles and facilities information and the returned information will be sent to the car’s audio and navigation systems. The passengers and car audio do not have to be silent when the driver initiates his/her utterance. Of course such systems will not need any talk button, so that signal for silence will be gone. The driver can change settings of auxiliary machines or retrieve information interactively as the dialog system asks for the missing information needed to complete actions. In such interactions, the driver will not have to wait for the completion of each of the dialog system’s messages, because the system will support a Barge-In mode.

Robots will have similar capabilities. However, they will need more robustness than automobiles. For example, people will be able to talk at robots from any direction from up to several meters distance even in very noisy rooms such as convention halls, factories, or living rooms with noisy TVs.

Advances in noise robustness will also support military uses of ASR systems. There are already needs for translators supporting local languages.

Currently, manufacturing industries are threatened by losing the skills of older and experienced workers before those skills are transmitted to the next generation. Noise-robust ASR will help address this situation by storing the skills within manufacturing machines so that new comers can retrieve them via voice.

There is even an ambitious idea of a personal life recorder, a portable device that will record every sound the wearer hears for 24 hours a day. Since people cannot use 24 hours to check the recorded content of each recorded 24 hours, noise robust ASR will be indispensable in the future to analyze and search such voluminous recordings.

# Appendix

## Abbreviation List

| | |
|---|---|
| ABF | Adaptive Beam Former |
| AL | Always Listening |
| AM | Acoustic Model |
| ASR | Automatic Speech Recognition |
| BSS | Blind Signal Separation |
| CDCN | Codeword Dependent Cepstral Normalization |
| CER | Character Error Rate |
| CMS | Cepstrum Mean Subtraction |
| CSJ | Corpus of Spontaneous Japanese |
| CSP | Cross-power Spectrum Phase |
| DS | Delay and Sum |
| EM | Expectation Maximization |
| ETSI | European Telecommunications Standards Institute |
| GJ | Griffiths-Jim |
| GMM | Gaussian Mixture Model |
| HMM | Hidden Markov Model |
| ICA | Independent Component Analysis |
| IID | Inter-channel sound Intensity Differences |
| ITD | Inter-channel Time Differences |
| LM | Language Model |
| LMS | Least Mean Square |
| LPE | Local Peak Enhancement |
| MFCC | Mel Frequency Cepstrum Coefficient |

| | |
|---|---|
| MLLR | Maximum Likelihood Liner Regression |
| MMSE | Minimum Mean Square Error |
| MUSIC | Multiple Signal Classification method |
| MV | Minimum Variance |
| PF | Profile Fitting |
| PMC | Parallel Model Combination |
| PTA | Push To Activate |
| PTT | Push To Talk |
| RLS | Recursive Least Squares |
| SAT | Speaker Adaptive Training |
| SMT | Smoothing Method of Time direction |
| SNR | Signal to Noise Ratio |
| SS | Spectral Subtraction |
| SSA | Spatial Subtraction Array |
| SSEC | Simultaneous adaptation of spectral Subtraction and Echo Cancellation |
| VAD | Voice Activity Detection |
| VTLN | Vocal Tract Length Normalization |
| WER | Word Error Rate |

# References

[AAM00] F. Asano, H. Asoh and T. Matsui, "Sound source localization and separation in near field," *IEICE Trans.*,Vol.E83-A,No.11,pp.2286-2294, 2000.

[AFB96] B. Ayad, G. Faucon and R.L. Bouquin-Jeannes, “Optimization of a noise reduction preprocessing in an acoustic echo and noise controller,” *Proc. of ICASSP ‘96*, Vol.2, pp. 953-956, 1996.

[AM97] H. Agaiby and T.J. Moir, “Knowing the wheat from the weeds in noisy speech,” *Proc. EUROSPEECH’97*, Vol.3, pp. 1119-1122, 1997.

[AMS+96] T. Anastasakos, J. McDonough, R. Schwartz and J. Makhoul, “A compact model for speaker-adaptive training,” *Proc. of ICSLP96*, Vol.2, pp. 1137-1140, 1996.

[AS01] M. Afify and O. Siohan, “Sequential noise estimation with optimal forgetting for robust speech recognition,” *Proc. of ICASSP 2001*, Vol. I, pp. 229-232, May 2001.

[AS90] A. Acero and R.M. Stern, “Environmental robustness in automatic speech recognition,” *Proc. of ICASSP ‘90*, pp. 849-852, 1990.

[ATI06] T. Arakawa, M. Tsujikawa and R. Isotani, "Model-based wiener filter for noise robust speech recognition," *Proc. of ICASSP 2006*, Vol. I, pp.537-540, 2006.

[Boe93] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proc. of Institute of Phonetic Sciences 17*, pp. 97-110, 1993.

[Bol79] S.F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech & Signal Process.*, Vol. ASSP-27, No. 2, pp. 113-120, Apr. 1979.

[BSM79] M. Berouti, R. Schwartz and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” *Proc. ICASSP ‘79,* pp. 208-211, Apr. 1979.

[BSN00] F. Basbug, K. Swaminathan and S. Nandkumar, “Integrated noise reduction and echo Cancellation for IS-136 Systems," *Proc. of ICASSP 2000*, Vol. 3, pp. 1863-1866, Jul. 2000.

[DCN97] M. Dahl, I. Claesson and S. Nordebo, "Simultaneous echo cancellation and car noise suppression employing a microphone array," *Proc. of ICASSP '97*, Vol. 1, pp. 239-242, 1997.

[DDI04] S. Deligne, S. Dharanipragada and O. Ichikawa, "Robust speech recognition with Multi-Channel Codebook Dependent Cepstral Normalization (MCDCN)", *Proc. of International Congress on Acoustics 2004*, pp.IV2599-IV2602, 2004.

[DP97] P. Dreiseitel and H. Puder, "A combination of noise reduction and improved echo cancellation," *Proc. of IWAENC '97*, pp.180-183, 1997.

[DR01] S. Dharanipragada and B.D.Rao, "MVDR based feature extraction for robust speech recognition," *Proc. of ICASSP01*, Vol.1, pp.309-312, 2001.

[Elk01] G.W. Elko, "Microphone arrays", *Proc. of International Workshop on Hands-Free Speech Communication*, pp. 11-14, Apr. 2001.

[EM84] Y. Ephraim, and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol.32, pp11009-1121, 1984.

[ETS02] ETSI ES 202 050 v1.1.1, "Distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," 2002.

[FN05] M. Fujimoto and Satoshi Nakamura, "Particle filtering and polyak averaging-based non-stationary noise tracking for ASR in noise," *Proc. ASRU Workshop*, pp. 337- 342, 2005.

[FNT+05] M. Fujimoto, S. Nakamura, K. Takeda, S. Kuroiwa, T. Yamada, N. Kitaoka, K. Yamamoto, M. Mizumachi, T. Nishiura, A. Sasou, C. Miyajima and T. Endo, "CENSREC-3: Data collection for in-car speech recognition and its common evaluation framework," *Proc. of International Workshop on Real-world Multimedia Corpora in Mobile Environments*, RWCinME2005, pp. 53-60, 2005.

[Fru05] S. Furui, "Recent progress in corpus-based spontaneous speech recognition," *IEICE Trans.*, Vol.E88-D, No.3, pp.366-375, 2005.

[Fuk05] T. Fukuda, “A Study on Feature Extraction and Canonicalization for Robust Speech Recognition,” *Ph.D thesis of Toyohashi University of Technology*, 2005.

[GJ82] L.J. Griffiths and C.W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Trans. Antennas & Propag.*, Vol. AP-30, No. 1, pp. 27-34, Jan. 1982.

[GR01] L. Gu and K. Rose, “Perceptual harmonic cepstral coefficients for speech recognition in noisy environment,” *Proc. of ICASSP*, Vol. 1, pp. 125-128, 2001.

[GY96] M.J.F.Gales and S.J.Young, ”Robust speech recognition using parallel model combination,” *IEEE Trans. on Speech and Audio Processing*, Vol.4, pp.352-359, 1996.

[HHS+06] C. Huang, Y. Huang, F. Soong and J. Zhou, “Weighted likelihood ratio (WLR) hidden Markov model for noisy speech recognition,” *Proc. of ICASSP*, Vol. 1, 2006.

[HM94] H. Hermansky, N. Morgan, “RASTA processing of speech,” *IEEE Trans. On Speech and Audio Processing*, Vol.2, pp.578-589, 1994.

[HO97] P. Hofman and J.V. Opstal, “Identification of spectral features as sound localization cues in the external ear acoustics,” *Proc. 1997 International Work-Conference on Artificial and Natural Neural Networks*, pp. 1126-1135, 1997.

[INK97] T. Isaka, Y. Nagata and H. Kanazawa, ”A study of voice detection using microphone array”, *Proc. of The Acoustical Society of Japan Autumn Meeting*, 2-Q-25, pp. 165-166, Sep. 1997 (in Japanese).

[JD] D. Johnson and D. Dudgeon, "Array signal processing", *Prentice Hall*

[KAN01] N. Kitaoka, I. Akahori and S. Nakagawa, “Speech recognition under noisy environments using spectral subtraction with smoothing of time direction and real-time cepstral mean normalization”, *Proc. of International Workshop on Hands-Free Speech Communication*, pp. 159-162, Apr. 2001.

[KAS+96] H.Y. Kim, F. Asano, Y. Suzuki and T. Sone, “Speech enhancement based on short-time spectral amplitude estimation with two-channel beamformer", *IEICE Trans. Fundamentals*, Vol. E79-A, No. 12, pp. 2151-2158, Dec. 1996.

[KFK04] K. Kobayashi, K. Furuya and A. Kataoka, "A microphone array system with echo canceller," *IEICE Trans.*, Vol. J87-A, No. 2, pp. 143-152, 2004 (in Japanese).

[KH03] T. Kristjansson and J. Hershey, "High resolution signal reconstruction," *Proc. ASRU workshop*, pp. 291- 296, 2003.

[KI95] S. Kajita and F. Itakura, "Robust speech feature extraction using SBCOR analysis," *Proc. of ICASSP'95*, Vol.1, pp.421-424, 1995.

[KL00] H.K. Kim and H.S. Lee, "Spectral peak-weighted liftering of cepstral coefficients for speech recognition," *IEICE*, *Trans.*, Vol.E83-D, No.7, pp.1540-1549, 2000.

[LO79] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. of IEEE*, Vol.67, No. 12, pp.1586-1604, 1979.

[LR96] Li Lee and R.C. Rose, "Speaker normalization using efficient frequency warping procedures," *Proc. of ICASSP'96*, Vol.1, pp.353-356, 1996.

[LW95] C. J. Leggetter and P. C.Woodland, "Maximum likelihood linear regression for speaker adaption of continuous density hidden Markov models," *Computer Speech and Language*, Vol. 9, pp.171-185, 1995.

[MA98] M. Mizumachi and M. Akagi "Noise reduction by paired microphones using spectral subtraction", *Proc. of ICASSP'98*, Vol.2, pp. 1001-1004, 1998.

[Mar95] K.D. Martin, "Estimating azimuth and elevation from interaural differences," *IEEE Mohonk Workshop on Applications of Signal Processing to Acoustics and Audio*, Oct. 1995.

[MII02] M. Morimoto, M. Itoh and K. Iida, "3-D sound image localization by interaural differences and the median plane HRTF," *Proc. 2002 International Conference on Auditory Display Kyoto*, Jul. 2002.

[MK92] S. Makino and Y. Kaneda, "Exponentially Weighted Step-Size Projection Algorithm for Acoustic Echo Cancellers," *IEICE Trans.*, Vol.E75-A, No.11, pp.1500-1508, 1992.

[MN82] M. Morimoto and K. Nomachi, “Binaural disparity cues in median-plane localization,” *J. Acoust. Soc. Jpn.*, (E) 3, 2, 99-103, 1982.

[MTM+06] S. Miyabe, T. Takatani, Y. Mori, H. Saruwatari, K. Shikano and Y. Tatekura, “Double-talk free spoken dialogue interface combining sound field control with semi-blind source separation,” *Proc. of ICASSP 2006*, Vol.I, pp.809--812, May 2006.

[MTS+06] Y. Mori, T. Takatani, H. Saruwatari, K. Shikano, T. Hiekata and T. Morita, "Blind Source Separation Combining SIMO-ICA and SIMO-Model-Based Binary Masking," *Proc. of ICASSP*, pp. V-81-84, 2006.

[MV96] R. Martin and P. Vary, “Combined acoustic echo control and noise reduction for hands-free telephony-state of the art and perspectives,” *Proc. EUSIPCO ’96*, pp. 1107-1110, 1996.

[NH01] J. Nix and V. Hohmann, “Enhancing sound sources by use of binaural spatial cues,” *Consistent & Reliable Acoustic Cues for sound analysis One-day workshop*, Sep. 2001.

[NIZ03] T. Nakatani, T. Irino and P. Zolfaghari, “Dominance spectrum based V/UV classification and F0 estimation,” *Proc. of EuroSpeech*, pp. 2313-2316, 2003.

[NM03] T. Nakatani and M. Miyoshi, “Blind dereverberation of single channel speech signal based on harmonic structure,” *Proc. of ICASSP 2003*, Vol.1, pp. 92-95, Apr., 2003.

[NOK02] K. Nakadai, H.G. Okuno and H. Kitano, “Real-time sound source localization and separation for robot audition,” *Proc. of ICSLP-2002*, pp. 193-196, Sep. 2002.

[NSI+04] Y. Nishimura, T. Shinozaki, K. Iwano, and S. Furui, “Noise-robust speech recognition using multi-band spectral features,” *Acoustical Society of America Journal*, Vol.116, Issue 4, pp. 2480-2480, 2004.

[ONS+05] Y. Ohashi, T. Nishikawa, H. Saruwatari, A. Lee and K. Shikano, "Noise-robust hands-free speech recognition based on spatial subtraction array and known noise superimposition," *Proc. of Intelligent Robots and Systems*, pp.2328 -

2332, 2005

[OS94] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," *Proc. of ICASSP'94*, pp. 273-276, 1994.

[OS96] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using CSP analysis," *Proc. of ICASSP'96*. Vol.2, pp.921-924, 1996.

[PW02] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminativetraining," *Proc. of ICASSP 2002*, Vol.1, pp.105-108, 2002.

[Ros04] R. C. Rose. "Environmental robustness in automatic speech recognition," *ISCA Workshop on Robustness in Conversational Interaction (Robust2004)*, August 2004.

[SB05] L. Szymanski and M. Bouchard, "Comb filter decomposition for robust ASR," *Proc. of InterSpeech 2005*, pp. 2645-2648, 2005.

[SC00] M. L. Shire and B. Y. Chen, "Data-driven RASTA filters in reverberation," *Proc. of ICASSP 2000*, Vol.3, pp. 1627-1630, 2000.

[SKT+00] H. Saruwatari, S. Kajita, K. Takeda and F. Itakura, "Speech enhancement using nonlinear microphone array based on noise adaptive complementary beamforming", *IEICE Trans. Fundamentals*, Vol.E83-A, No. 5, pp. 866-876, May 2000.

[SKT+03] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*, Vol.2003, No.11, pp.1135-1146, 2003.

[SNH+03] S. Sakauchi, A. Nakagawa, Y. Haneda and A. Kataoka, "Implementing and evaluating an audio teleconferencing terminal with noise and echo reduction," *Proc. of IWAENC 2003*, pp. 191-194, 2003.

[SPTK] http://www.sp.nitech.ac.jp/~tokuda/SPTK/

[SS80] M. Sugiyama and K. Shikano, "LPC peak weighted spectral matching measures," *ASJ Trans. of the Com. on Speech Res.*, S80-13, pp.101-108, 1980.

[SSL+03] H. Saruwatari, K. Sawai, A. Lee, K. Shikano, A. Kaminuma and M. Sakata, "Speech enhancement and recognition in car environment using blind source separation and subband elimination processing," *Proc. of 4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pp.367-372, 2003.

[STB+01] J.C. Segura, A. de la Torre, M.C. Benitez and A.M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. Experiments using the Aurora -II database and tasks," *Proc. EuroSpeech*, pp.221-224, 2001.

[Tak97] K. Takagi, "Japan Tokkai Hei 9-252268," *Patent Journal published by Japan Patent Office*, 1997 (in Japanese).

[Tak99] T. Takiguchi, "Statistical acoustic model adaptation for robust speech recognition in noisy reverberant environments," *Ph.D theis of Nara Institute of Science and Technology*, 1999.

[TO98] H. Tolba and D. O'Shaughnessy, "Robust automatic continuous-speech recognition based on a voiced-unvoiced decision," *Proc. of ICSLP*, paper 0342, 1998.

[TTS+06] Y. Takahashi, T. Takatani, H. Saruwatari and K. Shikano, "Blind spatial subtraction array with independent component analysys for hands-free speech recognition," *Proc. of International Work Shop on Acoustic Echo and Noise Control*, 2006.

[WWB02] M. Wu, D. Wang and G.J. Brown, "A multi-pitch tracking algorithm for noisy speech," *Proc. of ICASSP*, Vol.1, pp.369-372, 2002.

[ZC93] P. Zakarauskas and M.S. Cynader, "A computational theory of spectral cue localization", *Journal Acoustical Society of America*, Vol.94, pp.1323-1331, 1993.

# List of Publications

## Journal papers

1. Osamu Ichikawa, Tetsuya Takiguchi and Masafumi Nishimura, "Speech enhancement by Profile Fitting method," *IEICE Transaction, Special Issue on Speech Information Processing*, Vol.E86D No.3, pp.514-521, 2003.

2. Osamu Ichikawa, Tetsuya Takiguchi and Masafumi Nishimura, "Sound source localization using a pinna-based Profile Fitting method," *IEICE Transaction, Special Issue on Speech Dynamic by Ear, Eye, Mouth and Machine*, Vol.E87-D No.5, pp.1138-1145, 2004.

3. Osamu Ichikawa and Masafumi Nishimura, "Simultaneous adaptation of echo cancellation and spectral subtraction for in-car speech recognition," *IEICE Transaction, Special Section on Multi-channel Acoustic Signal Processing*, Vol.E88A No.7, pp.1732-1738, 2005.

4. Osamu Ichikawa and Kozo Fujii, "Computation of the flow field around arbitrary three-dimensional body geometry using Cartesian grid," *Transactions of the Japan Society of Mechanical Engineers. B*, Vol.68 No.669, pp.1329-1336, 2002 (in Japanese).

## Journal letter

1. Osamu Ichikawa, Takashi Fukuda and Masafumi Nishimura, "Local peak enhancement for in-car speech recognition in noisy environment," *IEICE Transaction, Special Section on Robust Speech Processing in Realistic Environments*, Vol.E91D No.3, pp.635-639, 2008.

# International conference papers (peer-reviewed)

1. Osamu Ichikawa, Tetsuya Takiguchi and Masafumi Nishimura, "Sound source localization using a pinna-based Profile Fitting method," *Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC) 2003*, pp.263-266, 2003.

2. Sabine Deligne, Satya Dharanipragada and Osamu Ichikawa, "Robust speech recognition with Multi-channel Codebook Dependent Cepstral Normalization (MCDCN)," *Proc. of International Congress on Acoustics (ICA) 2004*, pp.IV2599-IV2602, 2004.

3. Osamu Ichikawa and Masafumi Nishimura, "Simultaneous adaptation of echo cancellation and spectral subtraction for in-car speech recognition," *Proc. of European Conference on Speech Communication and Technology (EuroSpeech / InterSpeech) 2005*, pp.2293-2296, 2005.

4. Osamu Ichikawa, Takashi Fukuda and Masafumi Nishimura, "Local peak enhancement combined with noise reduction algorithms for robust automatic speech recognition in automobiles," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2008*, pp.4869-4872, 2008.

5. Takashi Fukuda, Osamu Ichikawa and Masafumi Nishimura, "Short- and long-term dynamic features for robust speech recognition," *Proc. of International Conference on Spoken Language Processing (ICSLP / InterSpeech) 2008*, pp. (to appear), 2008

6. Takashi Fukuda, Osamu Ichikawa and Masafumi Nishimura, "Phone-duration-dependent long-term dynamic features for stochastic model-based voice activity detection," *Proc. of International Conference on Spoken Language Processing (ICSLP / InterSpeech) 2008*, pp. (to appear), 2008

# Technical report

1. Osamu Ichikawa, T. Takiguchi and M. Nishimura, “Noise reduction by profile fitting method,” *IEICE Technical Report*, SP2002-21, pp.19-23, May 2002 (in Japanese).

# Meetings

1. Osamu Ichikawa, Tetsuya Takiguchi and Masafumi Nishimura, “Speech emphasis by Profile Fitting,” *Spring Meeting of Acoustic Society Japan*, Vol.I 2-2-7, pp.69-70, March 2002 (in Japanese).
2. Osamu Ichikawa and Masafumi Nishimura, “Real time implementation of 2 channel beamformer on a personal computer,” *Fall Meeting of Acoustic Society Japan*, Vol.I 3-Q-6, pp.165-166, September 2002 (in Japanese).
3. Osamu Ichikawa, Tetsuya Takiguchi and Masafumi Nishimura, “Sound source localization by Profile Fitting method,” *Spring Meeting of Acoustic Society Japan*, Vol.I 2-8-18, pp.687-688, March 2003 (in Japanese).
4. Osamu Ichikawa, Tetsuya Takiguchi and Masafumi Nishimura, “Sound source localization using a pinna-based Profile Fitting method in a reverberant environment,” *Fall Meeting of Acoustic Society Japan*, Vol.I 1-5-14, pp.505-506, September 2003 (in Japanese).
5. Osamu Ichikawa, Tetsuya Takiguchi and Masafumi Nishimura, “Rank-based spectral subtraction method for musical noise reduction,” *Spring Meeting of Acoustic Society Japan*, Vol.I 2-8-17, pp.93-94, 2004 (in Japanese).
6. Osamu Ichikawa and Masafumi Nishimura, “Simultaneous adaptation of echo cancellation and spectral subtraction for in-car speech recognition,” *Spring Meeting of Acoustic Society Japan*, Vol.I 2-Q-10, pp.117-118, 2005.
7. Takashi Fukuda, Osamu Ichikawa and Masafumi Nishimura, “Designing filter for reverberant speech based on reverberant segment detection at end of utterance,” *Fall Meeting of Acoustic Society Japan*, 2-P-1, pp.95-96, September 2006 (in Japanese).

8. Takashi Fukuda, Osamu Ichikawa and Masafumi Nishimura, “A study on dynamic features over long time periods,” *Spring Meeting of Acoustic Society Japan*, 1-P-1, pp.125-126, March 2007 (in Japanese).
9. Osamu Ichikawa, Takashi Fukuda and Masafumi Nishimura, “Local peak enhancement for automatic speech recognition,” *Fall Meeting of Acoustic Society Japan*, 1-P-24, pp.185-186, September 2007 (in Japanese).
10. Takashi Fukuda, Osamu Ichikawa and Masafumi Nishimura, “Long-term speech information for voice activity detection in low S/N environment,” *Spring Meeting of Acoustic Society Japan*, 1-10-6, pp.19-20, March 2008 (in Japanese).
11. Takashi Fukuda, Osamu Ichikawa and Masafumi Nishimura, “Harmonic structure-based feature extraction for noise-robust voice activity detection,” *Fall Meeting of Acoustic Society Japan*, 1-1-11, pp.25-26, September 2008 (in Japanese).
12. Osamu Ichikawa, Takashi Fukuda and Masafumi Nishimura, “Sound source localization with local peak weighted CSP,” *Fall Meeting of Acoustic Society Japan*, 3-P-26, pp.821-822, September 2008 (in Japanese).

# Patents (issued)

1. Osamu Ichikawa, Takashi Fukuda and Masafumi Nishimura, "Low-cost method for determining filter coefficient in dereverberation," *issued as patent 4107613 in US*, 2008
2. Gakuto Kurata, Masafumi Nishimura and Osamu Ichikawa, "Apparatus, method, and program for supporting speech interface design," *issued as patent 4156639 in US*, 2008
3. Masafumi Nishimura, Osamu Ichikawa and Tetsuya Takiguchi, "Speech recording method for court report," *issued as patent 4082611 in US*, 2008
4. Osamu Ichikawa, "Rank-based spectral subtraction method for musical noise reduction," *issued as patent 3909709 in US*, 2007
5. Osamu Ichikawa and Masafumi Nishimura, "A method to design the shape of outer-ear suitable for sound source localization," *issued as patent 3999689 in US*, 2007
6. Osamu Ichikawa, "Business model of web page facsimile service for an internet

appliance disconnected to a printer," *issued as patent 7079291 in US*, 2006

7. Osamu Ichikawa, "Computer system having a data buffering system which includes a main ring buffer comprised of a plurality of sub-ring buffers connected in a ring," *issued as patent 5948082 in US*, 1999