Title	Noise Reduction Based on Microphone Array and Post-filtering for Robust Hands-free Speech Recognition in Adverse Environments
Author(s)	李, 軍鋒
Citation	
Issue Date	2006-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/973
Rights	
Description	Supervisor:赤木　正人, 情報科学研究科, 博士

Noise Reduction Based on Microphone Array and Post-filtering for Robust Hands-free Speech Recognition in Adverse Environments

by

Junfeng Li

submitted to
Japan Advanced Institute of Science and Technology
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Supervisor: Professor Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology

March, 2006

Abstract

This research proposes a noise reduction system using microphone array and post-filtering with the goal of improving the recognition accuracy and robustness of hands-free speech recognition systems in adverse environments.

Acoustic interfering noise signals dramatically degrade the performance of many speech applications, such as automatic speech recognition system and speech communication system, in practical environments. For example, for automatic speech recognition system, noises result in the mismatch between the training and testing conditions, further degrading the performance of recognition system in real-world conditions. For speech communication system, acoustic noises degrade the quality and intelligibility of received speech signals. Therefore, noise reduction has been a fundamental enabling technology and an indispensable component for these applications that must recognize or transmit speech in noisy environments.

Though the problem of dealing with acoustic interfering noises has been researched for several decades and is still a challenging research topic due to the complex and time-varying characteristics of signals (speech and noise signals) and acoustic environments where the systems perform. In this research, interfering noise signals present in real conditions are considered to be of two components: localized noise coming from certain determinable directions and non-localized noise propagating in all directions. Note that localized noise might include stationary and non-stationary (e.g. sudden) noise components, white and colored noise components. Non-localized noise might include coherent and incoherent noise components as well. Noises with different characteristics from various kinds of sources make it difficult to construct an effective noise reduction system. Furthermore, the characteristics of noises do vary with time and environments, further increasing the difficulty of designing a noise reduction system. Moreover, only the system with small physical size is preferable because of the limited space, e.g., in car environments or hearing aid. Also, considering the practical implementation, real-time processing is generally a “must” for noise reduction systems in real conditions.

To suppress both localized and non-localized noises while keeping the desired speech signal distortionless, this research proposes a noise reduction system based on microphone array and post-filtering with the goal of improving the performance of speech recognition systems in adverse environments. This proposed noise reduction system follows the basic principle of the multi-channel Wiener filter, which is the optimal solution to the problem of minimizing the mean square error of the desired speech and its estimate and can further

be decomposed into a *minimum variance distortionless response* (MVDR) beamformer followed by a single-channel Wiener filter.

To deal with localized noise, Mizumachi *et al.* has reported a subtractive beamformer based algorithm which consists of three parts: noise direction estimation, noise spectral estimation and desired signal enhancement. However, this method fails to deal with localized noise in some frequencies and some directions because of the inherent spatial "NULLs" in its beam pattern. To solve this problem, we propose a hybrid noise estimation technique by combining the subtractive beamformer based multi-channel estimation approach and a soft-decision based single-channel estimation approach. The estimation accuracy of this hybrid technique is further improved by integrating a *robust and accurate speech absence probability* (RA-SAP) estimator. The experimental results show that this hybrid estimation technique provides much more accurate spectral estimates for localized noise than the multi-channel and single-channel estimation technique alone, respectively. The estimated spectrum of localized noise is then compensated and suppressed from that of noisy observation on each microphone. This algorithm is able to suppress various localized noise, especially sudden noise, using a small-size (3-channel) microphone array at a very low computational cost.

Moreover, note that the subtractive beamformer was derived based on paired microphones with the assumption of a perfectly coherent noise field. However, this assumption is seldom satisfied in practical environments. To solve this problem, we further develop a generalized subtractive beamformer by relaxing the assumption of a perfectly coherent noise field to the one of an arbitrary noise field. Following the ideas similar to those of the subtractive beamformer presented by Mizumachi *et al.*, the generalized subtractive beamformer with a *generalized sidelobe canceller* (GSC) like structure is derived. The theoretical analysis is also presented to show the linkage between these two beamformers and to show the theoretical noise reduction performance of the generalized algorithm in the theoretically well-defined noise fields. The comparison of two beamformers is also discussed based on the realistic experimental results.

To further deal with the residual non-localized noise (coherent and incoherent noise components), post-filtering is normally used at beamformer output. Many post-filters, such as, Zelinski post-filter and McCowan post-filter, have been published so far. However, their performance is degraded due to the unrealistic assumption of a perfectly incoherent noise field (Zelinski post-filter) and the assumed *a priori* coherence function of the noise field (McCowan post-filter). To solve these problems, we propose a hybrid post-filter for microphone arrays with an assumption of a diffuse noise field which was proven to be successful in modelling the noise conditions in many practical environments (e.g., car environments and reverberant rooms). In the proposed hybrid post-filter, a modified Zelinski post-filter, which is estimated using the signals on the microphone pairs on which noises are uncorrelated by considering the correlation characteristics of noise impinging

on different microphone pairs, is applied to the high frequencies to suppress the spatially uncorrelated noise; a single-channel Wiener filter is applied to the low frequencies for cancellation of spatially correlated noise. The proposed hybrid post-filter shows some advantages: in theory, it is a Wiener filter; in practice, it can deal with both high-correlated and low-correlated noise components in a diffuse noise field. Experimental results using various recordings confirm the superiority of this hybrid post-filter with regard to other comparative post-filters.

The performance of the proposed noise reduction system is finally investigated as a front-end processor for a speech recognition system. The speech recognition experiments are performed using multi-channel real-world noise recordings, and the performance of the proposed noise reduction system is further compared with other traditional noise reduction systems in terms of speech recognition rate. The speech recognition results show that the proposed noise reduction algorithm outperforms the other traditional algorithms in improving the speech recognition performance in the tested adverse environments.

Compared with other traditional noise reduction algorithms, this proposed algorithm demonstrates some advantages: (1) in theory, it provides the optimal solution to the problem of multi-channel noise reduction for broad-band inputs in *minimum mean square error* (MMSE) sense; (2) it is able to deal with various kinds of noise signals, including localized and non-localized noise, stationary and non-stationary (e.g., sudden) noise; (3) it avoids the problems of slow convergence rate and low stability in practical environments; (4) it can be implemented in real-time mode; (5) it is successful in improving the performance of hands-free speech recognition systems in adverse environments.

In addition to hands-free speech recognition systems, the noise reduction system proposed in this thesis is also useful and preferable to many other applications. For example, for speech communication system, it is able to improve the quality and intelligibility of the received speech signals. For hearing aid, it is able to provide more clean and intelligible speech, enhancing the performance of hearing aid to hearing impaired with a small-size microphone array at a low computational complexity in adverse conditions.

Acknowledgments

I am very happy to write this page because I can be expected to finish my doctor course soon. At this key time, I would like to express my sincere gratitude to the following people.

Firstly, I would like to acknowledge my supervisor, Professor Masato Akagi, for his great help and directions. I would like to thank Professor Akagi for his welcome me into his Lab. when I came to Japan knowing absolutely nothing about signal processing. I would like to thank Professor Akagi for frequently finding time for discussion, mitigating my confusion and successfully introducing me to the fascinating worlds of microphone array signal processing. Without his directions and help, it is impossible for me to finish my doctor course. The kindness, knowledgeability and personality of Professor Akagi deeply impressed me and will affect my career and my life for ever.

I would like to acknowledge Professor Jianwu Dang and Associate Professor Masashi Unoki, as two members of my thesis committee, for their invaluable suggestions and comments in my research. The discussions with Professor Dang and Associate Professor Unoki make my research progress continuously. The assistances from them make my life joyful in JAIST.

I would like to acknowledge Professor Yôiti Suzuki of Tohoku University in Sendai, Japan, as one member of my thesis committee, for his interest in my research and his kind welcome me into his Lab. for further research on microphone array signal processing after my graduation. I also would like to thank Professor Suzuki for his invaluable comments to improve the quality of this thesis and for finding the time to participate in the defense of my thesis.

I would like to acknowledge Professor Joerg Bitzer of University of Applied Sciences in Oldenburg, Germany, for his fruitful discussions and constructive suggestions in this research. He was very generous with his time and always quick to respond to my questions.

I would like to thank Dr. Xugang Lu for his help and discussions in my research, especially in speech recognition experiments. I am very happy to spend the past three years with him and we also have been good friends.

I would like to thank Dr. Mitsunori Mizumachi of Kyushu Institute of Technology in Fukuoka, Japan. As one alumni of our Lab., he gave me helpful suggestions and continuous encouragements in this research.

I would like to thank all the people that I know through my numerous business trips for their helpful discussions and for the great moments together. Especially, I would like to thank Professor Yutaka Kaneda of Tokyo Denki University in Tokyo, Japan, Dr. Sharon

Gannot of Bar-Ilan University in Ramat-Gan, Israel, and Dr. Wolfgang Herbord of ATR in Kyoto, Japan, for their constructive comments and suggestions.

I would like to thank Dr. Kazuhito Ito for his so kind help in the past years. Now, I still clearly remember that he took me to the hospital in the first year. He told me the exact time when I would go to hospital and then took me there. I was really surprised at his so impressive kindness.

I would like to thank my tutor, Mr. Hironori Nishimoto, for his so kind help in my daily life. So far, I still remember the scenario that Mr. Nishimoto picked me up at Komatsu station when I came to Japan on April 2, 2003. As my tutor and my first Japanese friend, he gave me a lot of help in my daily life and my research.

Also, I would like to thank Dr. Yuichi Ishimoto, Mr. Takeshi Saitou, Mr. Atsushi Haniu and all other members of the acoustic information science Lab. in JAIST who have always been helpful in making my life easier and colorful. Also, I would like to thank all my Chinese and Japanese friends in JAIST with who I have spent the past unforgettable three years.

There is no way I would be where I am today without the immeasurable love, support and encouragement of my parents. I can not thanks them enough for all the opportunities they have given me in my life and for always supporting me. They are an incredible inspiration to me.

Finally, I would like to thank the “Graduate Research Program (GRP)” in JAIST. Since the beginning, this research was conducted as a program for the ”Fostering Talent in Emergent Research Fields” in Special Coordination Funds for promoting Science and Technology by Ministry of Education, Culture, Sports, Science and Technology. In the past years, I was supported by this program for my life in Japan and for my numerous business trips.

Contents

List of Figures

List of Tables

Glossary

Mathematical Notation

$s(t)$	desired speech signal
$x_m(t)$	signal on m-th microphone
X_m	STFT of $x_m(t)$
$\mathbf{X}$	stacked signal vector of X_m
$\mathbf{X}^T$	transpose of $\mathbf{X}$
$\mathbf{X}^H$	complex transpose of $\mathbf{X}$
$\mathbf{X}^\dagger$	conjugation transpose of $\mathbf{X}$
ϕ_{xx}	auto-power spectral density of $x(t)$
ϕ_{xy}	cross-power spectral density of $x(t)$ and $y(t)$
Φ_{xy}	cross-power spectral density vector of $\mathbf{X}$ and Y
$\boldsymbol{\Phi}_{xx}$	auto-power spectral density matrix of $\mathbf{X}$
$\boldsymbol{\Phi}_{\mathbf{xy}}$	cross-power spectral density matrix of $\mathbf{X}$ and $\mathbf{Y}$
$\Gamma_{\mu\nu}$	noise coherence function between μ-th and ν-th microphone signal
$\boldsymbol{\Gamma}$	noise coherence matrix
∂	differential operator
$\forall$	for all
arg	parameter operator
max	maximum operator
$\Re$	real part
$P(\boldsymbol{w}\|\mathcal{O})$	probability of $\boldsymbol{w}$ given $\mathcal{O}$
$P(\boldsymbol{w})$	probability of $\boldsymbol{w}$
$E\{\cdot\}$	expectation operator
$\|\boldsymbol{\Phi}\|$	cardinality of set $\boldsymbol{\Phi}$
$\{\mu, \nu\}$	microphone pair of μ-th and ν-th microphones
IFFT$[\cdot]$	inverse Fourier transform operator

Fixed Symbols

a_m	impulse response between speech source and m-th microphone
b	normalized window for frequency smoothing

c	velocity of sound
d	distance between adjacent microphones (uniform linear array)
d_{mfcc}	Mel-frequency cepstral coefficient distance
$d_{\mu,\nu}$	distance between μ-th microphone and ν-th microphone
e	error signal
f_s	sampling frequency
f_t	transient frequency
f_t^m	transient frequency of m-th microphone pair
h	window function (hamming window)
i	sub-band index
j	imaginary unit: $\sqrt{-1}$
k	frequency index
$\tilde{k}$	frequency index in sub-band
l	point index in a frame
m	microphone index
$mfcc_i$	i-th mel-frequency cepstral coefficient
$n_m(t)$	additive noise signal on m-th microphone
$n_m^c(t)$	localized noise component on m-th microphone
$n_m^{uc}(t)$	non-localized noise component on m-th microphone
n_o	noise signal at beamformer output
p	index for localized noise
q	speech absence probability
$q^{'}$	the *a priori* speech absence probability
s_o	desired signal at beamformer output
$x_m(t)$	observed noisy signal on m-th microphone
y	system output

A_m	transfer function between speech source and m-th microphone
B_m	m-th sub-band
D	length of the normalized window b
D_i	noise components after localized noise suppression
G_{mz}	gain function of modified Zelinski post-filter
G_s	gain function of single-channel Wiener post-filter
I	number of sub-band

K	length of *short-time Fourier transform*
L	frame length, window length
N_m	STFT of n_m
$\hat{N}^c$	localized noise spectral estimate
$\hat{N}^c_{mul,i}$	localized noise spectral estimate in i-th sub-band using the multi-channel technique
$\hat{N}^c_{sig,i}$	localized noise spectral estimate in i-th sub-band using the single-channel technique
P	number of localized noise sources
$P_{local}, P_{global}, P_{frame}$	speech presence probability in a local frequency window, a larger frequency window, and neighboring frames
Q	state number for a word
R	frame shift
S	STFT of s
SNR_{priori}	the *a priori* SNR for post-filter Wiener post-filter
SNR_{post}	the *a posteriori* SNR for Wiener post-filter
$U_{\mu\nu}$	STFT of $u_{\mu\nu}$
W_m	gain function on m-th channel
W_{MVDR}	gain function of the superdirective beamformer
Y	STFT of y
Y_{FBF}	fixed beamformer output
Y_{NC}	noise canceller output
Y_o	output of the generalized subtractive beamformer
Z_i	output of localized noise suppression

$\mathbf{A}$	transfer function vector of A_m
$\mathbf{B}$	blocking matrix
$\mathbf{B}_1, \mathbf{B}_2$	matrixes for blocking matrix
$\mathbf{H}$	noise canceller filter
$\mathbf{H}_{opt}$	optimal solution of $\mathbf{H}$
$\hat{\mathbf{H}}_{opt}$	estimate of $\hat{\mathbf{H}}_{opt}$
$\mathbf{N}$	noise signal vector of N_m
$\mathcal{S}$	set of all possible state sequences
$\boldsymbol{s}$	state sequence
$\mathbf{W}$	gain function vector
$\mathbf{W}_{opt}$	optimal gain function vector
$\boldsymbol{w}$	set of all possible word sequence that can be hypothesized by the recognition system

α	overestimation factor
β	spectral floor factor
$\alpha_n, \alpha_s, \beta_s$	forgetting factors
ℓ	frame index
ζ	time delay of desired speech between adjacent microphones
δ_p	time delay of p-th localized noise between adjacent microphones
μ, ν	index of microphone
$u_{\mu\nu}$	subtractive beamformer using signals x_μ and x_ν
ε	band-width of sub-band
ε_1, ε_2	thresholds
τ	any value expect to zero
ξ	the *a priori* SNR
γ	the *a posteriori* SNR
$\tilde{\xi}, \tilde{\gamma}$	frequency-smoothed ξ, γ
$\bar{\xi}, \bar{\gamma}$	time-frequency-smoothed ξ, γ
$\delta_{\mu\nu}$	time delay of localized noise between μ-th and ν-th microphone
Ω_m	m-th microphone pair set
ϕ_{xx}	auto-power spectral density of x
$\mu^w_{\imath\rho}$	mean vector associated with the k-th Gaussian in the mixture density of state i of the HMM of word w
$\mathbf{\Phi}$	set of frames with speech present
$\mathcal{O}$	sequence of feature vectors
$\mathcal{N}$	Gaussian distribution
$\sum^w_{\imath\rho}$	covariance matrix associated with the k-th Gaussian in the mixture density of state $\imath$ of the HMM of word w

Acronyms and Abbreviations

AR	auto-regressive
ASR	automatic speech recognition
BM	blocking matrix
BSS	blind source separation
DCT	discrete cosin transform
DFT	discrete Fourier transform
DI	directivity index
DOA	direction of arrival
DSBF	delay-and-sum beamformer

DSWF	delay-and-sum beamformer with Wiener post-filter
AR	auto-regressive
ASR	automatic speech recognition
BM	blocking matrix
BSS	blind source separation
DCT	discrete cosin transform
DFT	discrete Fourier transform
DI	directivity index
DOA	direction of arrival
DSBF	delay-and-sum beamformer
DSWF	delay-and-sum beamformer with Wiener post-filter
FBF	fixed beamformer
GCC	generalized cross-correlation
GSC	generalized sidelobe canceller
GSVD	generalized singular value decomposition
HMM	hidden markov model
ISTFT	inverse short-time Fourier transform
ITD	interaural time difference
KLT	Karhunen-Loeve transform
LCMV	linearly constrained minimum variance
LMS	least mean square
MAP	maximum a posterior
MA-LSA	microphone arrays with OM-LSA based post-filtering
MFCC	mel-frequency cepstral coefficient
MMSE	minminum mean square error
MSC	magnitude-squared coherence
MVDR	minimum variance distortionless response
MWF	multi-channel Wiener filter
NC	noise canceller
NEE	normalized estimation error
NR	noise reduction performance
OLA	overlap-and-add
OM-LSA	optimally-modified log-spectral amplitude

ORG-GSC	original generalized sidelobe canceller
ORG-SBF	original subtractive beamformer
PATH	phase transform
PRO-GSBF	proposed generalized subtractive beamformer
PRO-MAPF	proposed noise reduction algorithm with microphone and post-filtering
PSD	power spectral density
RA-SAP	robust and accurate speech absence probability
SAP	speech absence probability
SEGSNR	segmental SNR
SNR	signal-to-noise ratio
STFT	short-time Fourier transform
SVD	singular value decomposition
TDC	time delay compensation
TDE	time delay estimation
TF-GSC	transfer function generalized sidelobe canceller
VAD	voice activity detection

Chapter 1

Introduction

Speech is the most natural and most important means of communication between human beings. Hence, research on speech sciences and technologies has been going on for centuries to understand the mechanism and process of the production, communication and perception of speech.

The production process begins with formulating a message that is to be transmitted from the talker to the listener via speech. The message is subsequently converted into a language code and a sequence of neuromuscular commands are executed, resulting in the vibration of a series of structures in the human vocal system and thereby producing an acoustic signal at the final output. The machine counterpart to the process of speech production is the speech synthesizer [127].

Once the speech signal is transmitted to the listener via communication channel, the perception process begins. The incoming acoustic signal is first analyzed along the basilar membrane in the inner ear. The output signal at the output of the basilar membrane is subsequently converted into activity signal. Finally the neural activity signal along the auditory nerve is converted into a language code, which is further understood and comprehended within the brain. The machine counterpart to the process of speech perception is the speech recognizer [127].

From the point of view of signal processing, the field of speech signal processing is essentially an application of signal processing techniques to speech signals. The explosive advances in recent years in the field of digital signal processing have provided a tremendous boost to the field of speech signal processing. The rapid development of speech signal processing techniques has stimulated the emergence and application of many speech techniques and products, such as, speech synthesizer, mobile phone and *automatic speech recognition* (ASR) system.

1.1 Speech recognition applications

Among the speech applications which emerged in recent years, speech recognition systems are becoming increasingly important in many aspects in modern society. Some applications of speech recognition systems are of high interest to be mentioned, which provide the interest and motivation to further research on the speech recognition technology.

Speech recognition systems have influenced on the writing process. The powerful and intricate connection between thought and speech has recently been recognized. Often, it is believed that dictating a document allows a writer to produce much more fluid, natural and expressive writing than if he/she had typed it manually. One extremely promising area in which speech recognition has already yielded significant benefits is enabling or facilitating the writing process for disabled writers [166].

Speech recognition systems have influenced on communication. At its core, speech recognition is a technology centered around the human voice and still our most fundamental means of communicating, connecting, and collaborating with others. Speech recognition could unlock an altogether new form of human-computer communication: the *dialogue-based interface*. Dialogue enhances the richness of the interaction and allows more complex information to be conveyed than is possible in a single utterance. Moreover, such a means of interaction would provide substantially more flexibility to the user and offer a more intuitive interface than do conventional systems. Speech recognition could also have a profound impact on the way humans communicate with each other. Current forms of interaction, such as blogging or instant messaging, might be forced to adapt (or become obsolete), as speech systems become more prevalent [166].

Speech recognition systems have influenced on the human-computer interface. Speech recognition systems hold the potential to unlock the treasure trove of data, creating a searchable index of information and placing it at users' fingertips just as conventional search engines have done with the World Wide Web. Speech recognition systems have already been proven useful in a number of specific domains of knowledge. Gaming is another realm of human-computer interaction in which speech recognition could play a significant role. It has been recognized that the opportunity to add functionality and enhance the user experience using this technology, making games more lifelike [166].

Specifically, more and more recognition systems are put into use in our daily lives to switch lights on and off, to control electronic equipments (e.g., TV, keyboards and buttons), etc. in a easy and user-friendly interface [123, 127]. Another promising application is in vehicle environments where recognition systems can be used to retrieve information from navigation system or perform simple control tasks [26, 61, 108, 119, 164]. As a fundamental human activity, meetings also provide an important and potential application domain for ASR technology [120].

1.2 Hands-free speech recognition challenges

The past several decades have witnessed the significant advances on ASR technology. As a result, state-of-the-art recognition systems have demonstrated high recognition accuracy for the situations where there is a good match between testing and training conditions. However, their performance drastically degrades when they are applied to real-world environments. Obstacles to robust recognition systems include degradations produced by acoustical disturbances, the effects of linear filtering, nonlinearities in transduction or transmission, as well as impulsive interfering sources, and diminished accuracy caused by changes in articulation produced by the high-intensity noise sources (i.e. Lombard effect) [123, 127]. Additionally, when the language/dialogue model becomes more complex, the variability in talking style may increase and one can expect that the talker will often speak in spontaneous mode, which further deteriorates the performance of speech recognition systems [127]. As speech recognition and spoken language technologies are being transferred to real applications, therefore, robustness in recognition technology is increasingly called for.

This research is particularly interested in those environments in which either safety or convenience precludes the use of close-talking microphones. For example, while operating a vehicle, the act of wearing microphones is distracting and dangerous. In a meeting room, microphones restrict the movement of the participants. In these situations and others, the users suffer a frustrating experience caused by the close-talking interaction. Hence, flexibility in the recognition technology is substantially called for to extend its use in a wide variety of real applications [51, 61, 63, 123].

One of the most attractive feature that improves the flexibility of recognition systems is hands-free interaction, where the user is not encumbered anymore by hand-held or head-mounted microphones and can talk up to a distance of some meters from the microphones. Therefore, hands-free speech recognition offers a remarkable flexibility and represents a very ambitious task, especially when considered for the applications of moderate and high complexity [40, 42, 49, 50, 123, 131].

In hands-free technology, as the distances between the user and the microphones grow, the speech signals become increasingly corrupted by the effects of acoustical interfering signals (e.g., environmental noise, reverberation and acoustical echo) [42, 123]. Sources of ambient noises are abundant. For example, in a room, noise sources might include personal computer, typewriter (from some certain directions) and background conversation of other people (from all directions, or, from some undeterminable directions). In a vehicle, noises mainly come from all directions, e.g., generated by wind, especially when the car is running at high speeds; other noises might come from radio or other passengers with certain directions. Moreover, environments in which hands-free recognition systems perform are generally reverberant conditions to a certain degree, which is caused by the reflections of signals by the walls and the furniture existing in the room [40]. In addition, acoustic echo is another type of disturbance for the signals picked up by distant microphones [145]. These acoustic interfering signals substantially degrade the speech recognition accuracy in noisy environments. Practically, to apply hands-free recognition system in real-world applications, it would be necessary to account for other various factors related to the means of hands-free interaction. For example, the talker's position may be unknown and time-varying in an unpredictable fashion; head movements, even subtle movements, may influence the quality of the input signal, due to the sound attenuation and talker radiation effects [40, 119, 145, 164]. Especially, background noise and acoustic characteristics (e.g., reverberation and acoustic echo) of the environment play an important role for hands-free speech recognition systems.

For these reasons and others, there are many challenging and as yet unsolved problems in this field. As environmental noise has become one main obstacle to commercial use of speech recognition techniques in a hands-free interaction. This thesis is mainly focusing on combating environmental undesirable acoustic noises and enhancing the desired speech signal, with the objective of reducing the mismatch between training and testing conditions and further improving the performance and robustness of hands-free recognition systems in real-world adverse environments.

1.3 Noise reduction for hands-free speech recognition

In real conditions, speech recognition systems are often exposed to various kinds of noises, which might arise from audio equipments, traffic and other speakers present in the environments (i.e., cocktail party noise). Noises degrade the quality of speech, resulting in the mismatch between training and testing conditions and further degrading the recognition rate of speech recognition systems.

To combat the background acoustical noises and improve the performance of speech recognition systems in the presence of disturbances, two basic ways are possibly adopted: (1) training the acoustic speech models of the recognizer engine using the speech database corrupted by the corresponding noises, which is referred to as model adaptation; (2) applying a front-end noise reduction system to suppress the background noises and improve the quality of the speech signals which are to be recognized [119, 123, 127, 153]. The first option may yield robust and high recognition performance if sufficient noise scenarios are included in the training procedure, but the drastic recognition performance decrease is expected if only limited noise conditions are considered in the training phase (i.e., the mismatch between training and testing conditions can not be reduced) and/or high time-varying non-stationary acoustic noise signals are present. Therefore, although the model adaptation technique has shown acceptable recognition performance in some controlled conditions, only limited performance improvement can be achieved by using the model adaptation technique in real noisy conditions [123, 127, 153]. Considering the complex and time-varying characteristics of real noisy environments, the second option (i.e., a front-end processor) provides a promising solution to the problem of suppressing the undesired noise signals, and has been widely researched and used as a front-end processor for speech recognition systems due to its effectiveness and flexibility. This kind of algorithm is based on the fact that the increased speech quality will also improve the speech recognition performance, which was proved to be effective although they are not correlated directly [119, 123, 127, 153]. This research is focusing on developing a practically effective and computationally efficient noise reduction system as a front-end processor to improve the recognition performance and robustness of a speech recognition system in adverse environments.

1.4 Noise reduction challenges

As a very effective way of increasing the speech quality and improving the performance of speech recognition systems, noise reduction has been studied for several decades and is currently still a challenging research topic. So far, a wide variety of noise reduction algorithms have been published [1, 3, 6, 11, 13, 22, 28, 43, 44, 52, 54, 62, 79, 80, 100, 114, 152], however, few of them can be applied to and can achieve acceptable noise reduction performance in practical environments.

The challenges are mainly caused by the complex and time-varying characteristics of the signals (speech and noise signals) and practical acoustic environments where recognition systems perform. Desired speech signals have a broad-band and high time-varying spectral components [40]. In practical environments, interfering noise signals are of very complex and time-varying properties. Take the noise condition in a car environment as an example. Noises generated by winds around the car come from all directions and have slow time-varying spectral components including coherent and incoherent noise components as well, which are generally modelled as diffuse noises [40, 88, 108]. Noises generated by engine come from certain directions and have slow time-varying spectral components. While, undesired interfering noises, such as passenger's voice and radio, have some determinable directions and highly non-stationary speech-like spectral components. Noises with different characteristics from various kinds of sources make it difficult to construct an effective noise reduction system. Furthermore, the characteristics of noises do vary with time and environments in a unpredictable fashion, further increasing the difficulty of designing a noise reduction system [26, 61, 119, 164]. Additionally, only the system with small physical size is preferable because of the limited space, e.g., in car environments and for hearing aids. Also, considering the practical implementation, real-time processing is generally a "must" for noise reduction systems in real conditions [1, 2, 3, 40, 145].

1.5 State-of-the-art noise reduction techniques

To suppress various background noises, a variety of noise reduction algorithms have been published in the literature [1, 2, 3, 6, 11, 13, 22, 28, 43, 44, 52, 54, 62, 79, 80, 100, 114, 142, 144, 152]. The different noise reduction algorithms can be classified into two categories, single-channel technique and multi-channel technique, according to the number of microphones needed in the implementation. In this section, we will summarize the different state-of-the-art noise reduction algorithms presented in the past several decades.

1.5.1 Single-channel noise reduction

A variety of single-channel noise reduction techniques, which exploits spectral and temporal differences between the speech and noise signals to suppress acoustical noises, have been proposed for speech recognition purposes [46, 67, 102]. Basically, these single-channel techniques compute estimates of the short-term spectral characteristics of the speech and the noise. These estimates are then combined according to a certain optimization criterion to produce an enhanced speech signal.

Single-channel noise reduction algorithms can be broadly classified into *parametric* and *non-parametric* approaches. *Parametric techniques* model the speech and sometimes also the noise as a stochastic auto-regressive (AR) model [53, 124]. Based on the estimates of AR-parameters, a Kalman filter is computed which is then applied to the noisy speech signal. *Non-parametric techniques* do not estimate the speech parameters, but rather exploit an estimate of the noise statistics to produce an enhanced speech signal [6, 13, 43, 44, 45, 152]. In recent years, *non-parametric techniques* have been paid more attention and been dominant techniques in the single-channel scenarios. Single-channel noise reduction and speech enhancement techniques normally operate in the transform domain: the frequency domain by the *discrete Fourier transform* (DFT) [6, 13, 43, 44], the wavelet domain by the wavelet transform [70], the *discrete cosin transform* (DCT) domain [142, 144] and in the domain using the *Karhunen-Loeve Transform* (KLT) [45, 128]. In *non-parametric techniques*, several typical noise reduction algorithms with their variants are of interest to be mentioned. Spectral subtraction first calculates the short-time spectral estimates of noise signals and then reduce the noise estimates from those of noisy observations [6]. Some improvements on spectral subtraction were performed by non-linear techniques [13], in other transform domains (e.g., wavelet domain, cepstral domain and DCT domain) and by combining some other signal modelling or estimation techniques [64, 72, 121, 140, 162]. Wiener filter, in principle, is closely related to spectral subtraction and yield the optimal solution in *minimum mean square error* (MMSE) sense [146]. Single-channel subspace-based techniques decompose the space of noisy signals into the perpendicular noise-only subspace and speech-plus-noise subspace by the means of a *generalized singular value decomposition* (GSVD) (or the KLT). The desired speech signal is then enhanced by extracting the speech components from the components in speech-plus-noise subspace based on some optimization criterion with certain constraints [45, 128]. Another class of single-channel noise reduction techniques, referred to as stochastic modelling based algorithms, has been paid more attention in recent years [22, 58, 59, 107, 134]. In these techniques, speech and noise are assumed to follow a certain priori distribution (e.g., Gaussian distribution, Laplacian distribution and Gamma distribution) in some transformed domain (e.g., spectral domain, power spectral domain). Model parameters of speech signal are then estimated according to a certain optimization criterion (e.g., MMSE or *maximum a posterior* (MAP)) and speech signal is finally recovered based on the estimated parameters of speech model.

The key point for single-channel non-parametric noise reduction algorithms is to calculate the noise spectral estimates with a high accuracy. Generally, single-channel speech enhancement techniques assume that the noise statistics are more stationary than the statistics of speech so that they can be estimated during noise-only periods. Hence, traditionally, the noise signal estimate is commonly adapted from the most recent recording, i.e., a few seconds before the speech is present, or *voice activity detection* (VAD) algorithms are used to distinguish the each frame and/or each frequency bin to noise-only or speech-plus-noise period and the noise signal is then estimated in the detected noise-only periods [15, 27, 58, 69, 150]. Recently, a minimum statistics approach has been proposed by Martin [105, 106]. In this approach, the power spectral densities of the observed noisy signals in the past several frames are stored and the noise power spectrum is then estimated by tracking the minimum value of the stored spectra. This noise spectral estimate is finally compensated by the fixed [105] or adaptive [106] bias compensator. The minimum statistic noise estimation technique is able to update the noise spectrum even in speech present periods [105, 106]. In addition, note that the VAD-based noise estimation technique is exactly a hard-decision mechanism since each frame and frequency band are judged as speech-present or speech-absent state absolutely. The performance of this hard-decision technique can be further improved. Therefore, recently, a soft-decision based noise estimation approach has been proposed and widely used as well [5, 28, 29, 31, 32, 34, 55, 101, 141, 143]. The soft-decision estimation approach considers, from the stochastic point of view, the probability of one frame and one frequency band which include desired speech components. Therefore, it also can update the noise spectral estimates even in speech active periods in a soft-decision mode by integrating the speech presence/absence probability.

Remark

Although an increase in global SNR has been reported in many cases, single-channel noise reduction algorithms have so far produced no or limited benefit for improving the local SNR in each frequency band since they can only differentiate between signals that have different temporal and spectral characteristics. Further, they only showed very limited capability in improving the performance of speech recognition systems [6, 13, 43, 44, 46, 67, 152]. This fact indicates that an increase in SNR does not automatically yield an increase in recognition rate. In real conditions, the speech and noise signals are considerably overlapped in the time-frequency domain, which makes it extremely difficult for single-channel techniques to substantially eliminate most of noise components without introducing speech distortion and artifacts (e.g., musical noises). Especially in low SNRs and spectrally highly non-stationary noise (such as babble noise) which are typical ingredients of a cocktail-party situation, the single-channel noise reduction techniques suffer from a low noise reduction performance [6, 44, 46, 67]. As a result, single-microphone techniques can achieve very limited improvements in suppressing noise and enhancing the speech recognition performance.

The limited benefit of single-microphone techniques for speech recognition is manifested by the growing tendency in the development of recognition systems towards the use of directional microphone(s) and/or multi-microphone techniques in recent years [12, 17, 111, 120, 127]. In addition to the temporal and spectral characteristics, the multi-microphone techniques also allow to exploit the spatial diversity of the speech and noise signals, resulting in the highly improved noise reduction performance and speech recognition accuracy [12, 17, 110, 111].

1.5.2 Multi-channel noise reduction

To overcome the performance limitations of single-channel noise reduction techniques which use the temporal/spectral characteristics of speech and noise signals, multi-channel techniques have attracted more research interests and showed great potential ability in reducing noise by exploiting the additional spatial information of signals and environments [1, 2, 3, 8, 9, 12, 17, 18, 23, 24, 26, 32, 33, 34, 36, 39, 40, 42, 48, 49, 51, 52, 54, 61, 62, 73, 74, 79, 81, 82, 83, 95, 100, 103, 113, 114, 115, 120, 123, 135, 136, 145, 148, 154, 155, 157, 163, 164]. In most scenarios, the desired speech source and interfering noise source are physically located at different positions in space. Exploiting the spatial diversity of the signals, multi-channel techniques can steer a main beam towards the desired speech source and/or nulls towards the interfering noise sources. The use of spatial diversity further provides more noise reduction ability to multi-channel techniques. Generally, multi-channel techniques can be classified into beamforming techniques and blind source separation techniques.

Beamforming techniques

The first class of beamforming techniques is *fixed beamforming.* In fixed beamforming techniques, the filter coefficients are normally optimized so that a beam is steered to the direction of the desired signal while suppressing the background noise coming from other directions as much as possible. These optimized filters are fixed, independent of the input signals, and then applied to the multi-channel microphone inputs. Typical fixed beamforming techniques include delay-and-sum beamforming [17, 83], differential microphone arrays [42, 83] and superdirective beamforming [8, 39]. Fixed beamforming techniques are widely used in the conditions where the acoustical characteristics do not change with time. However, using the fixed beamforming techniques, it is generally not possible to design arbitrary spatial directivity patterns for arbitrary microphone array configurations and design spatial directivity patterns which can be optimized to the time-varying acoustical environments [83].

The second class of beamforming techniques is *adaptive beamforming.* In contrast to fixed beamforming techniques, adaptive beamforming techniques make use of data-dependent filter coefficients that are adapted to respond to time-varying environments, yielding a better noise reduction performance than fixed beamforming techniques, particularly if the number of interferences is small (i.e., smaller than the number of microphones) and in the acoustic environments with less reverberation [7, 12, 23, 24, 40, 49, 52, 62, 79, 83, 103, 108, 145].

Adaptive beamforming techniques (e.g., the Frost beamformer [52, 148]) typically solve a *linearly constrained minimum variance* (LCMV) optimization problem, keeping the signals arriving from the desired look-direction (i.e., ideally the direction of the desired speech source) distortionless while suppressing the signals from other directions by minimizing the output power or output noise power. A *generalized sidelobe canceller* (GSC) beamformer [62], first presented by Griffiths and Jim as an alternative implementation structure of the LCMV beamformer, has also been widely researched. The GSC beamformer transforms the constrained optimization problem as an unconstrained optimization problem. The GSC beamformer consists of a fixed beamformer, creating a so-called speech reference signal; a blocking matrix, creating the so-called noise reference signals; and a multi-channel adaptive filter, eliminating the (noise) components in the speech reference signal which are correlated with the noise reference signals. In addition, a wide variety of noise reduction algorithms based on the GSC beamformer have so far been suggested, which are of interest to be mentioned [8, 17, 34, 34, 49, 54, 69, 122, 145]. Bitzer *et al.* presented an alternative implementation algorithm with a GSC structure for the superdirective beamformer and its performance was also analyzed in a diffuse noise field [8]. Fischer *et al.* proposed to apply a Wiener filter in the upper path of the GSC beamformer to suppress the uncorrelated noise components and then the correlated noise components are then reduced by the adaptive noise canceller in the lower path [49]. Recently, the

GSC beamformer was extended to a *transfer function generalized sidelobe canceller* (TF-GSC) beamformer by considering the transfer functions which relate the speech source and the microphones [54], which was shown to yield high noise reduction performance in practical environments. Moreover, the theoretical performance of the GSC beamformer and TF-GSC beamformer was widely examined in the diffuse noise field [7, 122, 145]. However, in all variants of the LCMV and GSC beamformers, adaptive signal processing techniques (e.g., *least mean square* (LMS)) were normally used to avoid cancellation of the desired speech signal, which introduces low convergence rate in practical environments and low ability in reducing non-stationary noise (e.g., sudden noise) [40, 52, 62, 54, 145]. To accelerate the convergence rate of the adaptive beamformers, the frequency-domain implementation of the GSC beamformer and the two-dimensional LMS implementation were introduced and further applied to the GSC beamformer [4, 23]. However, adaptive processing systems still do not show a high enough convergence rate and a high stability in real conditions.

Another class of multi-channel noise reduction techniques is *multi-channel Wiener filtering* (MWF) [17, 40, 145]. These techniques provide a *minimum mean square error* (MMSE) estimate of the (reverberant) speech signal in one of the microphone signals. In contrast to adaptive beamformer techniques, MWF techniques exploit both spectral and spatial differences between the speech and the noise sources, resulting in a higher noise reduction performance and inevitably introduces some speech distortion. Different MWF techniques include the GSC with single-channel post-filter [17], the MWF using calibration sequences [57] and the MWF with unknown reference [40, 145]. Traditionally, the MWF with unknown reference does not need the priori information about the signals, therefore, it provides much robust noise reduction performance. However, the MWF with unknown reference techniques introduces very high computational complexity, making it unreasonable and unfeasible for the practical real-time applications [40, 145].

Blind source separation

Blind source separation (BSS) is another class of multi-channel noise reduction techniques, which has also been researched in recent years [73, 100, 125, 153]. BSS recovers independent source signals by using only the information of mixed signals observed at all input channels. In this technique, neither the sources nor any information about the way these sources are mixed is known to the user. The basic assumption of BSS is that the source signals are statistically independent, which is however not always true in practical environments. For BBS technique, there are two basic kinds of implementation approaches, i.e., the time-domain BSS and the frequency-domain BSS. The time-domain BSS demonstrates a slow convergence rate due to its high computational cost for long FIR filters in the convolutive mixture scenarios. To accelerate the time-domain BSS, the frequency-domain BSS is widely considered by separately considering the instantaneous mixtures at each frequency. However, some other problems, e.g., permutation problem, underdetermination and circularity, are involved in the frequency-domain BSS [73, 100]. These problems are considered to be inevitable even if the time-domain BSS and the frequency-domain BSS are combined, where some benefits from the frequency-domain BSS are expected. Therefore, although BSS seems to be promising approach to estimate speech signal (in another sense, reducing background noise), their performance will be degraded due to a large number of problems in the implementation procedure in practical environments.

Remark

Target speech and interfering noise generally originate from different spatial positions, though they might have similar spectral properties in the time-frequency domain. Therefore, in comparison of single-channel noise reduction algorithms, multi-channel noise reduction algorithms have shown high noise reduction performance with minimum speech distortion due to the use of the spatial information of the signals in addition to the temporal and spectral information of the signals. However, a large number of microphones (for the delay-and-sum beamformer) and adaptive signal processing techniques (for e.g., the Frost and GSC beamformes) are involved for the beamformering based algorithms, and the independent assumption between different sources are needed for the BSS algorithms. Those associated problems degrade the noise reduction performance of the traditional multi-channel noise reduction algorithms, resulting in the limited improvement of the recognition accuracy of speech recognition systems in adverse environments. Hence, high-performance and small-size computationally efficient noise reduction system is preferred in the development of multi-microphone noise reduction algorithms for speech recognition systems in real-world conditions. This is also the research objective of this thesis.

1.6 Outline of the thesis and main contributions

In this section, we begin with describing the objectives of the research that is done in this thesis. Then, we provide a chapter by chapter overview of this thesis and summarize the main contributions that this research achieved.

1.6.1 Research objectives

In this research, we propose a noise reduction system which is constructed using microphone array and post-filtering for robust speech recognition in adverse environments.

As already mentioned, acoustic interfering noise signals degrade the performance of many applications in noisy conditions. For example, for speech recognition systems, background noises result in the mismatch between the training and testing conditions, further drastically degrading their recognition performance. To improve the performance of recognition systems in noisy environments, noise reduction techniques are called for. The objective of this research is to suppress undesired noises with the goal of improving the recognition performance of hands-free speech recognition systems in adverse environments.

Noise reduction in real conditions is a challenging research topic due to the complexity and time-variation of the signals and acoustic environments in real conditions. In practical conditions, interfering noises might include coherent and incoherent noises, stationary and non-stationary noises, white and colored noises. Therefore, the developed system should be effective in suppressing various kinds of noise signals. In addition, since noise signals are also time-varying, the developed system should be adaptive to deal with time-changing acoustic environments. Moreover, because of some practical factors, e.g. economy and space limitations, the noise reduction algorithm with small physical size is acceptable for practical applications, e.g. hearing aid and in car environments. In addition, considering the practical implementation, real-time processing is generally a “must” for noise reduction algorithms in real conditions.

In this thesis, we concentrate on dealing with the challenging problem of designing a low-computation, high-performance system with a small physical size to suppress various kinds of noise signals for further improving the performance of speech recognition systems in adverse environments. To do this, we propose a multi-channel noise reduction system. This is motivated by the fact that more (temporal, spectral and spatial) characteristics of desired signals and interfering signals can be exploited in multi-channel techniques. Consequently, compared to single-channel techniques, multi-channel techniques provide substantial superiority in reducing noise and enhancing speech due to their spatial filtering capability in suppressing the interfering signals coming from directions other than the specified look-direction. The high noise reduction capability of multi-channel techniques make them preferable to improving the performance and robustness of hands-free speech recognition systems in noisy conditions.

Specifically, in this research, the undesired noise signals are considered to be composed of localized noise components coming from certain directions and non-localized noise components coming from all (or, undeterminable) directions. Subsequently, we propose a multi-channel noise reduction algorithm which applies a (3-channel) microphone array system based on the beamforming technique to eliminate the localized noise components due to the high ability of microphone arrays to suppress the localized noises, and applies a hybrid post-filter which is designed with the assumption of a diffuse noise field to eliminate the non-localized noise components which might be coherent or incoherent. To suppress the localized noises, we propose a hybrid noise estimation technique which combines a multi-channel noise estimation approach and a single-channel noise estimation approach. This combination is further enhanced by integrating a *robust and accurate speech absence probability* (RA-SAP) which considers the strong correlation of speech absence uncertainty between adjacent frequency bins and consecutive frames, significantly improving the spectral estimation accuracy for localized noises. The more accurate localized noise spectral estimate is then subtracted from that of the noisy observation by non-linear spectral subtraction. To suppress the non-localized noise, we propose a hybrid Wiener post-filter under the assumption of a diffuse noise field. In this hybrid post-filter, a modified Zelinski post-filter, which fully considers and utilizes the spatial correlations of noise signals on different microphone pairs, is applied to the high frequencies to suppress the spatially uncorrelated noise; a single-channel Wiener filter is applied to the low frequencies for cancellation of spatially correlated noise. As a result, the proposed noise reduction system based on microphone array and post-filtering is expected to be able to suppress both localized noise and non-localized noise with minimum speech distortion, further improving the performance and robustness of hands-free speech recognition systems in adverse environments.

In comparison of the traditional noise reduction systems, the proposed noise reduction system has the following advantages: in theory, it follows the principle of the multi-channel Wiener filter (a MVDR beamformer followed by a single-channel Wiener post-filter) which provides the optimal solution to the problem of minimizing the mean square error of the desired speech signal and its estimate; in practice, it is effective to deal with various kinds of undesired noise signals, it is effective to deal with time-varying highly non-stationary (e.g., sudden) noise signals, it is a small physical size system with only three microphones, it is able to be implemented in a real-time mode.

1.6.2 Chapter by chapter overview and contributions

This thesis consists of six chapters. A schematic overview of this thesis is presented in Fig. 1.1.

Chapter 2 starts with the description of the characteristics of acoustic (speech and noise) signals and acoustic environments. Special attention is paid to the complex (e.g., localized and non-localized, stationary and non-stationary) and time-varying properties of noise signals present in practical noisy conditions. According to the source types where noises are generated, a noise signal model is explicitly suggested, consisting of localized noises coming from certain determinable directions and non-localized noises propagating in all directions. Based on this signal model, we develop a novel noise reduction system using microphone array and post-filtering. In theory, the proposed noise reduction system follows the basic principle of the multi-channel Wiener filter (a MVDR beamformer followed by a single-channel Wiener filter); in practically, it is able to suppress various kinds of noise signals (e.g., localized and non-localized, stationary and non-stationary noises) with small-size (3-channel) microphone array. This noise reduction system based on microphone array and post-filtering is the main contribution of this research. The basic principle and implementation procedure of this system are explained briefly as well.

In **Chapter 3**, the problem of localized noise suppression is dealt with. The basic idea is to first estimate spectra of localized noises which are then subtracted from those of noisy observations. In this chapter, we first give a brief review of a subtractive beamformer based noise reduction algorithm on which the proposed algorithm is based and its drawbacks are also discussed. To overcome these drawbacks, we propose a novel hybrid noise estimation technique which combines the subtractive beamformer based multi-channel noise estimation technique and a soft-decision based single-channel noise estimation technique to improve the spectral estimation accuracy of localized noises. The estimation accuracy is further enhanced by the a RA-SAP estimator which exploits the strong correlations of speech presence/absence uncertainty between adjacent frequency bins and consecutive frames, yielding the highly accurate localized noise spectral estimates. The more accurate spectral estimates are then reduced from those of noisy observations by non-linear spectral subtraction, improving the localized noises suppression performance.

Moreover, a generalized expression for the subtractive beamformer is developed by relaxing the assumption of a perfectly coherent noise field to the one of an arbitrary noise field. Following the ideas similar to those of the subtractive beamformer, the generalized algorithm with a GSC-like structure is derived. The theoretical analysis is also presented to show the linkage between these two beamformers and to show the theoretical noise reduction performance of the generalized algorithm in well-defined noise fields. The comparison of two beamformers is also discussed and performed based on the realistic experimental results.

Publications related to this chapter are [85, 86, 87, 88, 89, 92].

In **Chapter 4**, the problem of non-localized noise suppression is dealt with. To suppress non-localized noises, we propose a hybrid post-filter for microphone arrays with an assumption of a diffuse noise field which was proven to be a reasonable model for many noise conditions. In the proposed hybrid post-filter, a modified Zelinski post-filter, which is estimated using the signals on the microphone pairs on which noises are uncorrelated by considering the correlation characteristics of noise impinging on different microphone pairs, is applied to the high frequencies to suppress the spatially uncorrelated noise; a single-channel Wiener filter is applied to the low frequencies for cancellation of spatially correlated noise. The proposed hybrid post-filter shows some advantages: in theory, it is a Wiener filter; in practice, it can deal with both high-correlated and low-correlated noise components in a diffuse noise field. Experimental results using multi-channel real-world noise recordings confirm the superiority of this hybrid post-filter with regard to other comparative post-filters.

Publications related to this chapter are [90, 93].

In **Chapter 5**, we begin with the introduction of the principle of speech recognition system which consists of feature extraction and decoding. A speech recognition system is then constructed to evaluate the performance of the proposed noise reduction algorithm in terms of recognition rate. Its performance is further compared with other traditional algorithms. The speech recognition results show that the proposed algorithm outperforms the other algorithms in improving the speech recognition performance in adverse environment.

Publications related to this chapter are [89, 94].

Finally, **Chapter 6** provides the overall conclusions of this research highlighting our contributions achieved in this thesis, discusses some open problems and presents some suggestions for future research.

1.7 Summary

In this chapter, we first demonstrated the effects of speech recognition systems on our daily lives, e.g., including on the writing process, on communication and human-computer interface. Some promising applications are also mentioned in section 1.1.

In section 1.2, we described the importance and challenges of the hands-free robust speech recognition systems in real-world environments.

In section 1.3, possible solutions to the problems associated with hands-free robust speech recognition systems were discussed as well. More attention was paid to noise reduction systems as front-end processors for hands-free speech recognition systems to improve their performance and robustness in real-world environments.

In section 1.4, we discussed the challenges for noise reduction algorithms, which mainly includes the complex and time-varying characteristics of acoustic (speech and noise) signals and acoustic environments.

In section 1.5, we briefly reviewed several widely-used single-channel (e.g., parametric and non-parametric) and multi-channel (beamforming and blind source separation) noise reduction algorithms. More attention was paid to the disadvantages and drawbacks of the traditional single-channel and multi-channel noise reduction algorithms.

In section 1.6, we first showed the research objective of this research, that is, to construct a computationally efficient and practically effective noise reduction system (with small physical size) with the goal of improving the performance of hands-free speech recognition systems in adverse environments. Finally, we demonstrated the chapter-by-chapter overview of this thesis.

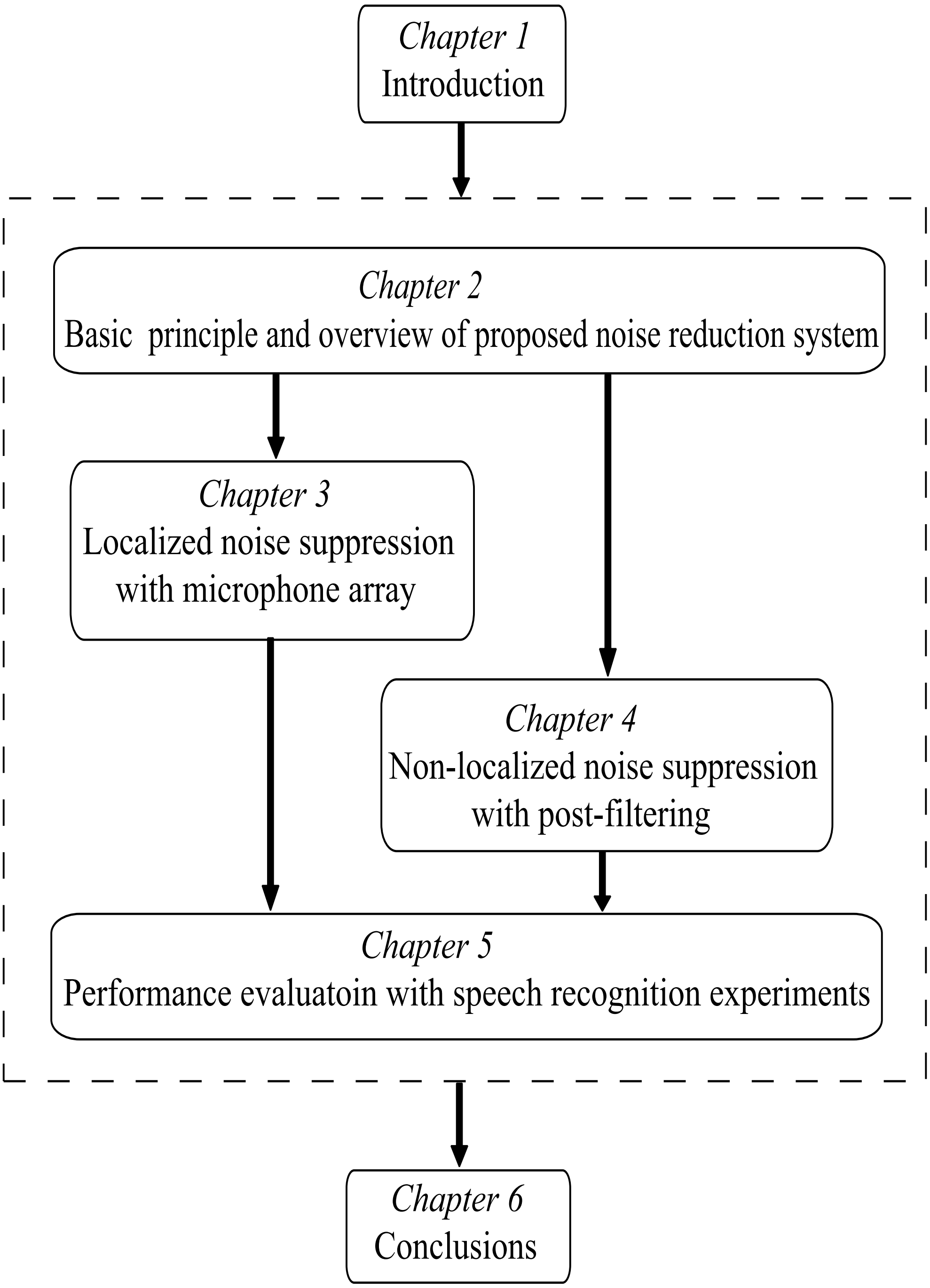

Chapter 1
Introduction
Chapter 2
Basic principle and overview of proposed noise reduction system
Chapter 3
Localized noise suppression with microphone array
Chapter 4
Non-localized noise suppression with post-filtering
Chapter 5
Performance evaluatoin with speech recognition experiments
Chapter 6
Conclusions

Figure 1.1: Schematic overview of the thesis.

Chapter 2

Basic principle and overview of the proposed noise reduction system

Improving the performance and robustness of hands-free speech recognition systems in noisy environments is crucial to put them into use in real-world conditions. Noise reduction algorithms, as front-end processors, have shown great potential in reducing noise and enhancing the performance of speech recognition systems. In addition to the temporal and spectral information of the desired speech signal and noise signal which are widely used in single-channel noise reduction algorithms, multi-channel noise reduction algorithms are able to exploit spatial information of the acoustic sound filed, demonstrating the great ability in reducing noise components and preserving the speech components. Therefore, multi-channel algorithms are becoming more promising and preferred to single-channel algorithms in enhancing the performance of recognition systems. These superiorities motivate us to develop a multi-channel noise reduction system to improve the performance and robustness of hands-free speech recognition systems in adverse environments.

This chapter provides the basic principle and overview of the proposed noise reduction algorithm with is based on microphone array and post-filtering. It is important for deeply understanding the whole system described in the forthcoming chapters.

In section 2.1, some characteristics of speech signal, noise signal and acoustic sound field are firstly introduced, which provides informative cues for designing the proposed noise reduction algorithm. To suppress additive noise signals, section 2.3 introduces the multi-channel Wiener filter, which was proven as an optimal solution to the problem of multi-channel noise reduction for broadband input signals. The multi-channel Wiener filter can further be decomposed into a *minimum variance distortionless responds* (MVDR) beamformer followed by a Wiener post-filter, which provides the theoretical basis for the proposed noise reduction algorithm. In section 2.4, we describe a signal model which is much more reasonable in real-world environments. Based on this signal model, we present a noise reduction system using microphone array and post-filtering which consists of several parts in its implementation. The function of each part is also presented.

2.1 Characteristics of signals and acoustic field

The characteristics of signals (speech signal and noise signal) and acoustic environment have a great influence on the type of noise reduction algorithms to be used and on the performance of the algorithms. In this section, some characteristics and peculiarities of speech signal, noise signal and acoustic field are discussed. Only the characteristics that are important for the algorithms and techniques considered in this thesis are mentioned. More detailed information on speech, noise signals and acoustic environment can be found in [40, 127, 145].

2.1.1 Speech signal

Speech is a *broad-band* signal with frequency components ranging from 100Hz to 8000Hz. According to the steady-state speech production model, a speech signal is not inherently band-limited. For speech understanding, however, mainly the frequencies between 300Hz and 3400Hz are of interest, which is the classical telephony bandwidth. Therefore, in this case, a sampling rate of 8kHz is usually sufficient to obtain an acceptable speech quality. However, the intelligibility of the speech signal with the sampling frequency of 8kHz is considerably lower than that of person-to-person speech, which is due to the loss of high-frequency information and results in the muffling effect of telephone sound. Therefore, higher sampling frequencies (e.g. 12kHz) are used because of the demand for high-quality speech. In this thesis, we will generally use a sampling rate of 12kHz, if there is no clarification.

Speech is a *time-varying* signal with both time envelop and spectrum continuously changing. The energy distribution and spectrum of speech signal is both time- and frequency-dependent. Sometimes speech can be considered as quasi-periodic (e.g. vowels), at other instances it resembles colored noise (e.g. frication) or impulse-like (e.g. plosive). Furthermore, speech pauses usually exist between the words and in a typical conversation more than 50% of the time will consist of silence. This on/off characteristic of speech signal can be exploited by speech enhancement algorithms, e.g. by using a *voice activity detection* (VAD) algorithm which classifies noise-only periods and speech-and-noise periods. Moreover, these speech pauses alternate with high energetic vowels and plosives, which significantly increases the short-time energy.

2.1.2 Noise signal

In comparison of speech signal, in general, less is known about the noise sources. Background noise can originate from a *localized* noise source coming from a certain direction, or can be *diffuse* noise coming from all directions. For instance, in a car environment, noises generated by the engine and the car radio can be considered as localized noises, while noise from the wind passing around the car cabin or from the frication between road and tires can be considered as diffuse noise. This classification method will be used and further discussed later in this thesis.

Some of the noise sources are *stationary* (e.g. fans) or have a slowly time-varying spectral content, whereas other sources can be highly *non-stationary* (e.g. radio). The most difficult problem arises when the noise signals are also speech signals (e.g., concurrent speakers), which are similar in structure to the desired signal. Furthermore, the noise sources can be narrow-band or wide-band, intermittent or persistent, and they may have the same temporal and spectral characteristics as the desired speech signal.

2.1.3 Acoustic environment

The acoustic environment that noise reduction systems perform plays an important role in hands-free interaction systems, affecting both speech intelligibility and the performance of speech recognition systems. Moreover, the performance of most noise reduction and acoustic source localization algorithms is also strongly influenced by the properties of the acoustic environment.

The acoustic environment is generally characterized by various kinds of background noise signals. The background noises might be localized noise or diffuse noise, with the stationary or non-stationary, coherent or incoherent spectral components. Moreover, all those characteristics are always changing with time and environments in a certain unpredictable way, as aforementioned. In this thesis, main attention is paid to deal with various background noise signals.

The acoustic environment is also characterized by *reverberation*, which is caused by the fact that acoustic waves are reflected by room walls and by other objects present in the environment, such that the signals recorded by the microphone array consist of a direct path signal and multiple delayed and attenuated versions. Obviously, the acoustic path is different for each source-microphone pair. Since the positions of the sources are not necessarily fixed and objects can also move around through the environment, acoustic paths are generally time-varying. It appears that the acoustic path can be modelled quite well by a linear transfer function. Although reverberation is a undesired signal to be reduced in practical environments, in this thesis, less attention is paid to deal with reverberation.

2.2 Signal model and problem formulation

This section describes a signal model and the mathematical formulation of multi-channel noise reduction algorithms.

2.2.1 Signal model

Considering a microphone array with M microphones in a noisy acoustic environment, shown in Fig. 2.1. The signal model, referred to as a generalized signal model in this thesis, is formulated in the time domain and the frequency domain in the following sections.

Time-domain representation

As demonstrated in Fig. 2.1, the observed signal $x_m(t)$ on m-th microphone at time instance t is generally composed of two components. The first is the desired speech signal $s(t)$ transformed by the impulse response $a_m(t)$ between the speech source and the m-th sensor. The second is the additive noise $n_m(t)$. Thus, the received signal in the time domain is given by:

$$x_m(t) = a_m(t) * s(t) + n_m(t), \quad m = 1, 2, \cdots, M \tag{2.1}$$

where $*$ represents the convolution operator.

It should be noted that the noise signal $n_m(t)$ might be composed of two components, i.e., localized noises $n_m^c(t)$ coming from certain directions and non-localized noise $n_m^{uc}(t)$ (i.e., diffuse noise) coming from all directions. Hence, the observed noise signal can further be represented as:

$$n_m(t) = n_m^c(t) + n_m^{uc}(t) \tag{2.2}$$

- The localized noise signals $n_m^c(t)$ are generated by some localized noise sources (e.g., fan, radio and competing speakers), which can be fixed or moveable in the space. Some localized noise sources are spectrally stationary or have slowly time-varying spectral properties (e.g., fan), while others are spectrally highly non-stationary (e.g., competing speech and sudden noise).

- The non-localized noise signals $n_m^{uc}(t)$ are generally modelled as diffuse noise (e.g., wind noise in car environments) arriving from all directions in the space. In most situations, these kinds of noise sources are spectrally stationary or have very slowly time-varying spectral properties.

Frequency-domain representation

Since noise reduction algorithms are usually designed and performed in the frequency domain, the frequency-domain representation of the signal model is necessary to be given.

The noisy signal is transformed into the frequency domain by applying a window h_m of size L to a frame of L consecutive samples of $x_m(t)$ and by computing the K-point *short-time Fourier transform* (STFT) on the windowed data. Before the next STFT

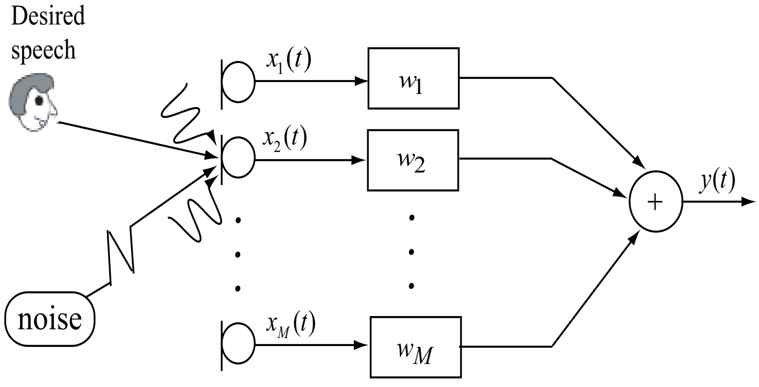


Figure 2.1: Multi-channel noise reduction algorithm.

computation, the window is shifted by R samples, This sliding window STFT analysis results in a set of frequency domain signals, which can be given by:

$$X_m(k,\ell) = \sum_{l=0}^{L-1} x_m(\ell R + l) h_m(l) e^{-\frac{j2\pi kl}{L}}, \quad m = 1, 2, \cdots, M \tag{2.3}$$

where k $(0 \leq k \leq K-1)$ denotes the frequency bin index, and ℓ is the frame index. Furthermore, to facilitate our notation and to avoid normalization factors, we suppose $\sum_{l=0}^{L-1} h_m(l) = 1$. In the following, we use a sampling rate of $f_s = 12000$Hz, $K = L$ and $L = 2R = 512$ and a hamming window.

Applying the STFT to Eq. (2.1), the observed signals on M microphones in the time-frequency domain can be represented as:

$$\mathbf{X}(k,\ell) = \mathbf{A}(k) S(k,\ell) + \mathbf{N}(k,\ell), \tag{2.4}$$

where

$$\mathbf{X}^T(k,\ell) = \big[X_1(k,\ell), X_2(k,\ell), \cdots, X_M(k,\ell)\big], \tag{2.5}$$

$$\mathbf{A}^T(k) = \big[A_1(k), A_2(k), \cdots, A_M(k)\big], \tag{2.6}$$

$$\mathbf{N}^T(k,\ell) = \big[N_1(k,\ell), N_2(k,\ell), \cdots, N_M(k,\ell)\big], \tag{2.7}$$

where $X_m(k,\ell)$, $S(k,\ell)$ and $N_m(k,\ell)$ are the STFTs of the corresponding signals $x_m(t)$, $s(t)$ and $n_m(t)$ respectively. And $A_m(k)$ is the acoustic transfer function between speech source and the m-th microphone, that is, the STFT of the impulse response $a_m(t)$, which is assumed to be time-invariant in the analysis periods. The superscript ^{T} denotes the transpose operator.

2.2.2 Multi-channel noise reduction in the frequency domain

Because most of noise reduction algorithms perform in the frequency domain, we here just give the general frequency-domain representation of the multi-channel noise reduction.

As Fig. 2.1 shows, in the multi-channel noise reduction, the observed signals $X_m(k,\ell)$ on each channel are transformed by the filters $W_m(k,\ell)$, which is either fixed or adaptive filter, and then combined to yield the enhanced speech signal. With the frequency-domain representation of signals, shown in Eqs. (2.4) - (2.7), the output of the multi-channel noise reduction reduction algorithm $Y(k,\ell)$ can be formulated as:

$$Y(k,\ell) = \mathbf{W}^H(k,\ell)\mathbf{X}(k,\ell), \tag{2.8}$$

where $\mathbf{W}^H(k,\ell) = [W_1(k,\ell), W_2(k,\ell), \cdots, W_M(k,\ell)]$ are the gain functions of the filters, the superscript ^{H} denotes the complex conjugative transpose operator.

2.3 Multi-channel Wiener filter

In this section, we introduce the multi-channel Wiener filter which provides an optimal solution to the problem of multi-channel noise reduction for broadband inputs in *minimum mean square error* (MMSE) sense. Moreover, the multi-channel Wiener filter provides the theoretical basis for the proposed noise reduction system.

We start to introduce the multi-channel Wiener filter with the definitions of *power spectral density* (PSD) and cross-PSD of signals. Let define the PSD of the signal $X(k,\ell)$ as:

$$\phi_{xx}(k,\ell) = E\Big[X(k,\ell)X^*(k,\ell)\Big], \tag{2.9}$$

with E denotes the statistical expectation operator, and cross-PSD of the signals $X(k,\ell)$ and $Y(k,\ell)$ as:

$$\phi_{xy}(k,\ell) = E\Big[X(k,\ell)Y^*(k,\ell)\Big], \tag{2.10}$$

and the corresponding cross-PSD vector as:

$$\phi_{xy}(k,\ell) = E\Big[\mathbf{X}(k,\ell)Y^*(k,\ell)\Big]. \tag{2.11}$$

Further, the PSD and cross-PSD matrix are defined as:

$$\mathbf{\Phi}_{xx}(k,\ell) = E\Big[\mathbf{X}(k,\ell)\mathbf{X}^H(k,\ell)\Big], \tag{2.12}$$

$$\mathbf{\Phi}_{xy}(k,\ell) = E\Big[\mathbf{X}(k,\ell)\mathbf{Y}^H(k,\ell)\Big]. \tag{2.13}$$

With the above definitions, the multi-channel Wiener filter is derived in the following. Consider again the multi-channel noise reduction, shown in Fig. 2.1, with the signal definitions in Eqs. (2.4)-(2.7), the system output can be written as:

$$Y(k,\ell) = \mathbf{W}^H(k,\ell)\mathbf{X}(k,\ell) = \mathbf{W}^H(k,\ell)\mathbf{A}(k)S(k,\ell) + \mathbf{W}^H(k,\ell)\mathbf{N}(k,\ell), \tag{2.14}$$

For arbitrary filter coefficients $\mathbf{W}(k,\ell)$, the error between system output $Y(k,\ell)$ and the desired speech signal $S(k,\ell)$ is calculated as:

$$e(k,\ell) = S(k,\ell) - \mathbf{W}^H(k,\ell)\mathbf{X}(k,\ell). \tag{2.15}$$

The square of the error can be calculated as:

$$\begin{aligned}\phi_{ee}(k,\ell) &= E\Big[\Big(S(k,\ell) - \mathbf{W}^H(k,\ell)\mathbf{X}(k,\ell)\Big)\Big(S^*(k,\ell) - \mathbf{X}^H(k,\ell)\mathbf{W}(k,\ell)\Big)\Big] \\ &= \phi_{ss}(k,\ell) - \mathbf{W}^H(k,\ell)\phi_{xs}(k,\ell) - \phi_{xs}^H(k,\ell)\mathbf{W}(k,\ell) \\ &\quad + \mathbf{W}^H(k,\ell)\mathbf{\Phi}_{xx}(k,\ell)\mathbf{W}(k,\ell). \end{aligned} \tag{2.16}$$

Thus, the PSD of the total error in ℓ-frame $\phi_{ee}(\ell)$ is the sum of the errors across all frequency bins, given by:

$$\begin{aligned}\phi_{ee}(\ell) &= \sum_{k=0}^{K-1}\Big[\phi_{ss}(k,\ell) - \mathbf{W}^H(k,\ell)\phi_{ss}(k,\ell) - \phi_{ss}^H(k,\ell)\mathbf{W}(k,\ell) \\ &\qquad + \mathbf{W}^H(k,\ell)\Phi_{xx}(k,\ell)\mathbf{W}(k,\ell)\Big]. \end{aligned} \tag{2.17}$$

Now, the problem that we are facing is to find the weight factors which minimize the squared error $\phi_{ee}(\ell)$. Note that the PSD of the error is exactly real-valued and nonnegative for all sub-bands, hence, the sum can be minimized by the weight factor $\mathbf{W}(k,\ell)$ which provides the minimization of the error power $\phi_{ee}(k,\ell)$ in each k-th sub-band.

Since the error power $\phi_{ee}(k,\ell)$ is a quadratic function of $\mathbf{W}(k,\ell)$, it has a single global minimum. The optimal weight factor $\mathbf{W}(k,\ell)$ which minimizes the error power can be determined by setting the derivation of $\phi_{ee}(k,\ell)$ with respect to $\mathbf{W}(k,\ell)$ to zero vector [17]:

$$\frac{\partial\phi_{ee}(k,\ell)}{\partial\mathbf{W}(k,\ell)} = -2\phi_{xs}(k,\ell) + 2\phi_{xx}(k,\ell)\mathbf{W}(k,\ell) = \mathbf{0}. \tag{2.18}$$

As a sub-band version of the multi-channel Wiener filter, the resulting expression in its general form can be given by:

$$\phi_{xx}(k,\ell)\mathbf{W}(k,\ell) = \phi_{xs}(k,\ell). \tag{2.19}$$

With the assumption that $\phi_{xx}(k,\ell)$ is nonsingular, the optimal weight factor is written as:

$$\mathbf{W}(k,\ell) = \Phi_{xx}^{-1}(k,\ell)\phi_{xs}(k,\ell). \tag{2.20}$$

With the signal model in Eq. (2.4), the optimal weight factor can be rewritten as:

$$\begin{aligned}\mathbf{W}(k,\ell) &= \phi_{xx}^{-1}(k,\ell)\phi_{ss}(k,\ell)\mathbf{A}(k) \\ &= \Big[\phi_{ss}(k,\ell)\mathbf{A}(k)\mathbf{A}^H(k) + \phi_{nn}(k,\ell)\Big]^{-1}\phi_{ss}(k,\ell)\mathbf{A}(k). \end{aligned} \tag{2.21}$$

Using the matrix inversion lemma, the optimal MMSE filter can be transformed into [17]:

$$\begin{aligned}\mathbf{W}(k,\ell) &= \left[\phi_{nn}^{-1}(k,\ell) - \frac{\phi_{ss}(k,\ell)\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)\mathbf{A}^H(k)\phi_{nn}^{-1}(k,\ell)}{1+\phi_{ss}(k,\ell)\mathbf{A}^H(k)\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)}\right]\phi_{ss}(k,\ell)\mathbf{A}(k)\\ &= \left[1 - \frac{\phi_{ss}(k,\ell)\mathbf{A}^H(k)\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)}{1+\phi_{ss}(k,\ell)\mathbf{A}^H(k)\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)}\right]\phi_{ss}(k,\ell)\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)\\ &= \left[\frac{\phi_{ss}(k,\ell)}{1+\phi_{ss}(k,\ell)\mathbf{A}^H(k)\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)}\right]\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)\\ &= \left[\frac{\phi_{ss}(k,\ell)}{\phi_{ss}(k,\ell)+\left(\mathbf{A}^H(k,\ell)\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)\right)^{-1}}\right]\frac{\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)}{\mathbf{A}^H(k)\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)}. \qquad (2.22)\end{aligned}$$

To further investigate the properties of the multi-channel Wiener filter, we consider the power of the desired signal at the output of MVDR beamformer, given by:

$$\phi_{s_o s_o}(k,\ell) = \phi_{ss}(k,\ell)\left|\frac{\mathbf{A}^H(k)\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)}{\mathbf{A}^H(k)\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)}\right|^2 = \phi_{ss}(k,\ell). \qquad (2.23)$$

Obviously, the desired signal component at the MVDR beamformer is distortionless since it is exactly equivalent to that at the beamformer input. Moreover, the power of the noise signal at the MVDR beamformer output is:

$$\phi_{n_o n_o}(k,\ell) = \frac{\mathbf{A}^H(k)\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)}{\left[\mathbf{A}^H(k)\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)\right]^2} = \frac{1}{\mathbf{A}^H(k)\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)}. \qquad (2.24)$$

Substituting Eqs. (2.23) and (2.24) into (2.22), the multi-channel Wiener filter, as an optimal solution to the problem of minimizing the mean square error between the desired speech and its estimation for broadband inputs, can be factorized as [17]:

$$\mathbf{W}_{opt}(k,\ell) = \underbrace{\frac{\phi_{nn}^{-1}(k,\ell)\mathbf{A}(k)}{\mathbf{A}^H(k)\phi_{nn}(k,\ell)\mathbf{A}(k)}}_{\text{MVDR beamformer}}\underbrace{\left[\frac{\phi_{s_o s_o}(k,\ell)}{\phi_{s_o s_o}(k,\ell)+\phi_{n_o n_o}(k,\ell)}\right]}_{\text{Wiener post−filter}}. \qquad (2.25)$$

As Eq. (2.25) demonstrates, the multi-channel Wiener filter can be decomposed into a MVDR beamformer followed by a single-channel Wiener filter, which provides the theoretical basis for the proposed noise reduction system detailed in this thesis. The MVDR beamformer provides a MMSE estimate of the desired signal under the constraint of a distortionless look-direction response, yielding a maximum directivity index for a diffuse noise field. Residual noise at the beamformer output can be further suppressed by a Wiener post-filter, further improving the noise reduction capability and producing an improved output SNR.

2.4 Proposed noise reduction algorithm

2.4.1 Signal model in the proposed system

Consider again Fig. 2.1, which shows the general configuration of multi-channel noise reduction algorithms using M microphones. In the proposed noise reduction algorithm presented in this thesis, a microphone array with three linearly and equidistantly-distributed omnidirectional microphones is considered in a noisy environment, shown in Fig. 2.2. The observed signal on each microphone consists of three components. The first is the desired speech signal $s(t)$ arriving from the direction such that the direction in arrival time between two end microphones is 2ζ. The second is localized noise signals $n_p^c(t)$ arriving from the directions such that the time differences are $2\delta_p(t)$ $(p = 1, 2, \dots, P)$ and the third is non-localized signal $n_{\cdot}^{uc}(t)$ which propagates in all directions simultaneously and is normally modelled as diffuse noise. Thus, the observed signals imposing on three microphones can be given by:

$$x_1(t) = s(t - \zeta) + \sum_{p=1}^{P} n_p^c(t - \delta_p) + n_1^{uc}(t), \tag{2.26}$$

$$x_2(t) = s(t) + \sum_{p=1}^{P} n_p^c(t) + n_2^{uc}(t), \tag{2.27}$$

$$x_3(t) = s(t + \zeta) + \sum_{p=1}^{P} n_p^c(t + \delta_p) + n_3^{uc}(t). \tag{2.28}$$

Note that in the signal model assumed in the proposed algorithm shown in Eqs. (2.26)-(2.28), the observed noise signal is considered to be of two components: localized noises with the certain directions (e.g. fan with fixed direction or other speakers with time-varying direction) and non-localized noise coming from all direction. In this sense, this signal model is an elaborated or refined version of the general signal model given in Eqs. (2.1). This refined signal model is more suitable to design an effective noise reduction algorithm and to show the respective performance of microphone array and post-filter which is dependent on the spatial characteristics of the noise field (i.e., the directionality and non-directionality of the noise sources), which will be seen clearly in the forthcoming chapters.

2.4.2 Overview of the proposed noise reduction algorithm

With the signal model represented by Eqs. (2.26)-(2.28), the task of our work is to reduce both localized noise $n_{\cdot}^c(t)$ and non-localized noise $n_{\cdot}^{uc}(t)$ while preserving the desired speech signal $s(t)$ distortionless. The proposed noise reduction algorithm exploits the beamforming based multi-channel processing technique and a post-filter. The block diagram of the proposed algorithm is shown in Fig. 2.3, which mainly consists of spectral

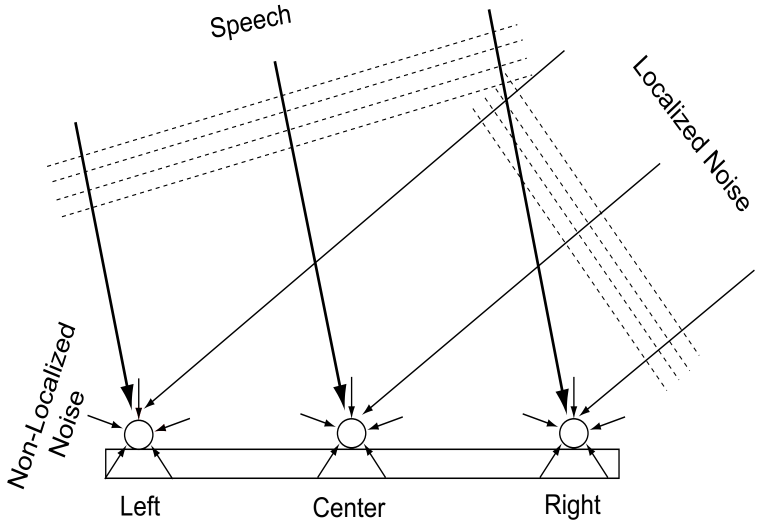


Figure 2.2: Microphone array and signal model assumed in the proposed noise reduction system.

analysis/synthesis, time delay compensation, localized noise suppression and non-localized noise suppression. In the following, we present the brief overview for each part.

Spectral analysis and synthesis

Spectral analysis and synthesis are indispensable components for state-of-the-art noise reduction algorithms which perform in the frequency domain. Although spectra analysis/synthesis are not explicitly plotted in Fig. 2.3 which is just for simplicity without confusion, they are also crucial for the proposed noise reduction algorithm which is a frequency-domain algorithm.

For spectral analysis, a compromise between frequency resolution and time resolution has to be made. High resolution in the frequency domain leads to poor resolution in the time domain and vice versa. Therefore, the highest possible frequency resolution that does not violate the short-term stationarity of speech could be chosen. Furthermore, the minimum error in the time domain is only reached if the filters have non-overlapping frequency regions. Since such filters are physically unrealizable, overlapping of sub-bands can not be avoided. As a result, the suppression of a noise-only sub-band may affect adjacent sub-bands containing desired signal components. In this thesis, we will use the windowing, *short-time Fourier transform* (STFT), *inverse short-time Fourier transform* (ISTFT) and the overlap-and-add method to implement spectral analysis and synthesis.

Time delay compensation

In real-world environments, the sound signals, especially the desired speech signal, are transformed by the impulse responses between the speech source and microphones. As the signal model in Eqs. (2.26)-(2.28) shows, the impulse responses are formulated as amplitude terms and phase terms. Time delay compensation aims to compensate the propagation effect (i.e., both amplitude and phase) between speech source and microphones on the desired speech signal. Special attention is paid to compensate the time delay difference between the desired speech signals on each microphone since the amplitude differences on each microphones are small enough to be ignored with the assumption that the speech is in the far-filed of the array after steering the main beam to the direction of desired speech signal.

Time delay estimation is currently also a hot research topic in array signal processing, a variety of algorithms, generally using the phase information present in signals picked up by spatially distributed microphones, have been published so far [21, 68, 117]. Given the geometry of a microphone array, the time delays between difference microphone pairs are dependent on the *direction of arrival* (DOA) of signal. There are three main categories of methods that process this information to estimate time delay or DOA.

The first is the steered beamformer based methods. Beamformers enhance the signals coming from a certain "look" direction by the use of the available signals observed on all (or partial) microphones. Thus, if a signal is present in the "look" direction, the array output power is high, while if there is no signal in the look-direction, the array output power is low. Therefore, the array can be used to construct beamformers that "look" in all possible directions and the direction that gives the maximum output power can be considered an estimate of the DOA (i.e., time delay).

The second is the subspace based methods. These methods divide the cross-correlation matrix of the array signals into signal and noise subspaces using *singular value decomposition* (SVD) to perform time delay (i.e., DOA) estimation. These methods are able to distinguish multiple sources that are located very close to each other much better than the steered beamformer based methods because the metric that is computed generally yields much sharper peaks at the correct locations. Both the steered beamformer based methods and the subspace based methods involve a high computational cost.

The third category consists of two steps, the time delays are first estimated for each pair of microphones in the array, and the geometry information of array is then combined to yield the best estimate of DOA. A variety of techniques were reported to compute pair-wise time delays. The *generalized cross correlation* (GCC) method and the narrow-band filtering followed by phase difference estimation of sinusoids are two examples. The *phase transform* (PHAT) is the most commonly used pre-filter for the GCC. The estimated time-delay for a pair of microphone is assumed to be the delay that maximizes the GCC-PHAT function for that pair. Fusing of the pair-wise *time delay estimates* (TDEs) is usually done in the least square sense by solving a set of linear equations to minimize the least squared error.

In this thesis, for the explanation simplicity, we assume that the *time delay compensation* (TDC) has been performed in advance. Further, suppose that for the respective signals the same symbols are used for notational simplicity, which is done by setting $\zeta = 0$ in Eqs. (2.26)-(2.28).

Localized noise suppression with microphone array

In practical environments, desired speech signal is generally corrupted by noises originated from various kinds of sources. Some noises might come from the sources (e.g., computer fan and radio) which are fixed or moveable in the environment. These noises are referred to as localized noise in the thesis.

To suppress the localized noises, the basic idea of our method is that the spectra of localized noises are first estimated and then subtracted from the those of the observed noisy signals.

In general, noise spectrum estimation is a crucial component for most noise reduction and speech enhancement algorithms. For localized noises, the multi-channel estimation algorithm is preferable to the single-channel estimation algorithm by the use of the spatial (directional) characteristics of localized noises. To estimate noise spectrum, a subtractive beamformer based estimation algorithm was proposed before with the problem of its failure in some frequencies and directions [1, 3]. To mitigate this problem, we propose a hybrid noise estimation algorithm which combines a multi-channel estimation technique and a single-channel estimation technique in a parallel structure. The multi-channel estimation technique was implemented using the subtractive beamformer based method since it yields much more accurate spectral estimates for localized noises at most instances. And the single-channel estimation technique was implemented using a soft-decision based noise estimation technique due to its ability in estimating the spectrum of non-stationary signal. The combination between the multi-channel and single-channel estimation algorithms are done in an effective way by the use of a novel *robust and accurate speech absence probability* (RA-SAP) estimator. This RA-SAP estimator considers the strong correlation of speech presence between adjacent frequency bins and consecutive frames and makes full use of the high estimation accuracy of the multi-channel estimation approach. Therefore, the final estimation accuracy for localized noises is greatly enhanced by the suggested RA-SAP estimator. After obtaining the more accurate noise spectral estimates, they are reduced from the spectra of observed noisy signals using the non-linear spectral subtraction.

In addition, we present a generalized expression of the subtractive beamformer in an arbitrary noise environment. This generalized expression is actually a natural extension of the original subtractive beamformer previously presented by extending the assumption of a localized (directional) noise field to that of an arbitrary noise field. Furthermore, we prove that the multi-channel technique (i.e., two subtractive beamformer based approaches) is the optimal solution for minimizing the system output with the constrain of distortionless response in “look” direction, by considering the generalized subtractive beamformer. That is, the subtractive beamformer is a MVDR beamformer in theory. Some experimental results are also presented to show the superiority of the generalized subtractive beamformer in car noise conditions.

Non-localized noise suppression with post-filtering

In practical environments, desired speech signal is generally corrupted by noises originated from various kinds of sources. Some noises might come from all directions (e.g., wind and the friction between tyre and road in car environments). Such noise environment is widely modelled as a diffuse noise field, such as in a car or in a room. These noises are referred to as non-localized noise in the thesis.

In addition to localized noises, non-localized noise (e.g., background noise) also degrades the quality of speech observed on microphones. After suppressing localized noises, non-localized noise has to be further suppressed. With the assumption of a diffuse noise field, we present a hybrid post-filter for microphone arrays, which exploits a modified Zelinski post-filter in the high frequencies, and a single-channel post-filter in the low frequencies.

For the modified Zelinski post-filter, we consider and make full use of the correlation of noises on different microphones to improve the noise reduction with minimum speech distortion. The implementation of the modified Zelinski post-filter consists of four steps: determine the transient frequencies according to the microphone array geometry; determine the microphone pairs on which noise is mutually uncorrelated for each frequency; compute the spectral densities of the desired and noisy signals; compute the gain function of the modified Zelinski post-filter. The first two steps can be done beforehand since they are only dependent on the microphone array geometry and independent on the input signals. Thus, the computational cost will greatly be reduced.

For the single-channel post-filter, we adopted a single-channel Wiener filter in the low frequencies. The *a priori* SNR, a crucial parameter used in the Wiener filter, is updated with the decision-directed mechanism which introduces less “musical noise”.

The proposed hybrid post-filter has some advantages: it is a Wiener filter in theory, hence, following the theoretical principle of the multi-channel Wiener filter which can be decomposed into a MVDR beamformer followed by a Wiener filter; in practice, the superiority of the proposed post-filter is verified by experiments using real-world multi-channel recordings.

2.5 Summary

In this chapter, we discussed the characteristics of signals and acoustic environment, the signal model and the multi-channel Wiener filter, and presented an overview of our proposed noise reduction algorithm.

In section 2.1, the characteristics of speech signal, noise signal and acoustic environment were discussed. Speech is generally characterized by the wide frequency range (e.g., from 100Hz to 8000Hz) and the spectrum which is time- and frequency-dependent. Noise signal is generally characterized by stationarity or non-stationarity, narrow-band or broad-band, directional (i.e., localized noise) or non-directional (i.e., non-localized noise). Moreover, acoustic environment is generally characterized by reverberation and acoustic echo, which also disturb the desired speech signal.

In section 2.2, the signal model and the problem to solve are presented based on the discussions shown in section 2.1. In the signal model, the observed signal on each microphone consist of a transformed speech signal which is filtered by the acoustic impulse response between speech source and microphones, and noise signal which is composed of directional (localized) and non-directional (non-localized) noise. The frequency representation of this signal model is also given. Subsequently, the general problem of multi-channel noise reduction is discussed in the frequency domain since most noise reduction algorithms perform in the frequency domain.

In section 2.3, we described the derivation of the multi-channel Wiener filter, which is an optimal solution to the problem of multi-channel noise reduction for broad-band inputs in MMSE sense. The multi-channel Wiener filter can further be decomposed into a MVDR beamformer followed by a Wiener filter, which provides the theoretical basis of the noise reduction algorithm we proposed in this thesis.

In section 2.4, the signal model described in section 2.1 is re-formulated to make the explanation of our proposed noise reduction algorithm easy to understand. The proposed noise reduction system consists of spectral analysis/synthesis, time delay compensation, localized noise suppression and non-localized noise suppression. Each component is described with an brief overview. The main idea of the proposed algorithm is: localized noise components are first estimated using a hybrid noise estimation technique which combines a multi-channel subtractive beamformer based estimation approach and a single-channel soft-decision based estimation approach, and then reduced from the observed signals by the non-linear spectral subtraction; non-localized noise are further reduce with a hybrid Wiener filter which exploits a modified Zelinski post-filter in the high frequencies and a single-channel Wiener post-filter in the low frequencies. Two main parts, localized noise suppression and non-localized noise suppression will be further discussed in detail in the forthcoming chapters 3 and 4.

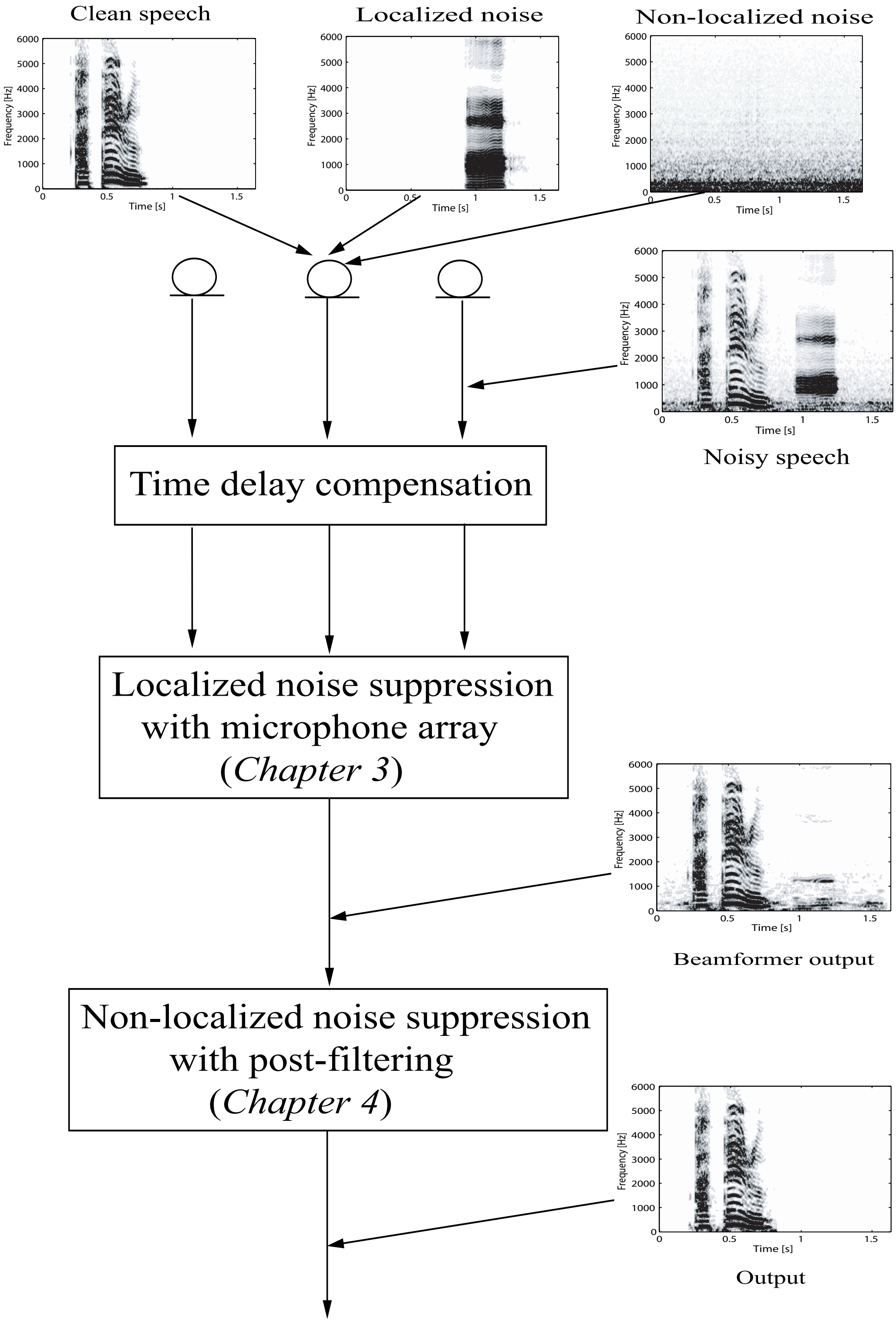


Figure 2.3: Proposed noise reduction algorithm.

Chapter 3

Localized noise suppression with microphone array

In real-world environments, the desired speech signal is usually corrupted by various kinds of noise signals (e.g., localized noise and non-localized noise). As aforementioned in chapter 2, localized noises are referred to as the noises originated from some (fixed or moving) point sources, that is, having the determinable directions. In this chapter, we deal with the problem of suppressing localized noise components with microphone array.

Considering the spatial characteristics of localized noises, the directionality provides some informative cues for designing an effective noise reduction algorithm. Especially when multiple microphones are available, in addition to temporal and spectral information of signals, the spatial information can be fully exploited to improve the noise reduction performance. Therefore, compared with single-channel noise reduction algorithms, multi-channel algorithms yield much better noise reduction performance with minimum speech distortion.

In this chapter, we deal with localized noise components using microphone array. The basic idea of our algorithm is that the spectra of localized noises are first estimated and then subtracted from those of the observed noisy signals. To accurately estimate the spectra of localized noises, we propose a hybrid noise estimation technique which combines a subtractive beamformer based multi-channel estimation approach and a soft-decision based single-channel estimation approach. The combination of the multi- and single-channel estimation approaches is greatly reinforced with a *robust and accurate speech absence probability* (RA-SAP) estimator. After obtaining the spectra of localized noises, the estimated spectra are then reduced from those of noisy signals using non-linear spectral subtraction. Furthermore, we extend the subtractive beamformer to a generalized subtractive beamformer by changing the assumption of a localized noise field to one of an arbitrary noise field. The generalized subtractive beamformer prove that the subtractive beamformer using in our noise reduction algorithm is a MVDR beamformer theoretically. Moreover, some experimental results are also presented to verify the superiority of the proposed generalized subtractive beamformer in car environments.

3.1 Introduction

In comparison to single-channel algorithms, multi-channel algorithms have demonstrated a substantial superiority in reducing noise due to their spatial filtering capability to suppress interfering signals arriving from directions other than the specified “look” direction [17]. Therefore, multi-channel algorithms (e.g., beamformering based algorithms) have attracted great research interest in recent years.

A variety of beamforming based algorithms have been proposed in the literature [7, 12, 23, 39, 40, 49, 52, 62, 79, 83, 103, 145]. The beamforming algorithms include fixed beamforming and adaptive beamforming, which are brief discussed in the sections 3.1.1-3.1.2. Special attention is paid to the disadvantages of the existing beamforming algorithms which motivate our research and highlight the advantages of the proposed algorithm.

3.1.1 Fixed beamforming

The simplest beamformer, referred to as *delay-and-sum* (DS) beamformer, enhances the desired speech signal by summing the in-phase microphone signals after compensating for the arrival time differences of the desired sound signal to each microphone by inserting delays after each microphone, that is, the array is first electronically steered to the look-direction. The advantages of the DS beamformer are that it is very simple to implement and that it minimizes the noise sensitivity and hence provides a high robustness against errors in the assumed signal model. However, a large number of microphones are normally needed to obtain an acceptable performance in real-world environments.

The superdirective beamformer is another widely studied fixed beamformer [39]. The supdirective beamformer maximizes the directivity index in the direction of the speech source for a diffuse noise field. Actually, the superdirective beamformer minimizes the power of the beamformer output subject to distortionless response for the “look” direction, hence, it is also an MVDR beamformer. The implementation simplicity of the superdirective beamformer leads to its widely use in some known noise field. However, its data-independent property results in that only limited noise reduction performance can be obtained in practical time-varying environments.

3.1.2 Adaptive beamforming

In real-world environments, the characteristics of signals (speech signal and noise signal) and acoustic condition vary with time, spectrum and space. Adaptive beamformers exploit all available information of signals and noise field in the way that combines the spatial focusing of fixed beamformers with adaptive noise reduction suppression, such that they are able to adapt to time-varying acoustic environments and generally exhibit a higher noise reduction performance than fixed beamformers.

The linear constrained adaptive beamformer is first presented by Frost [52]. In Frost beamformer, the adaptive filters are computed such that the power of the beamformer output is minimized. Furthermore, some linear constraints are combined to avoid the speech distortion. Therefore, this filter is also referred to as *linear constrained minimum variance* (LCMV) beamformer. With the assumption of zero correlation between the speech signal and the noise signal, the constraint of the constrained beamformer output power minimization corresponds to the constraint of the constrained noise output power minimization. Hence, MVDR beamformer is a special case of LCMV beamformer in this sense.

A *generalized sidelobe canceller* (GSC) beamformer, as an alternative implementation structure of the Frost beamformer and first presented by Griffiths and Jim, has also been widely researched [62]. In the GSC beamformer, the constrained minimization problem of Frost beamformer is reformulated as an unconstrained minimization problem which is more simple for implementation. The GSC beamformer consists of three parts: a fixed beamformer which electronically steers the microphone array to the direction of interest (i.e., the speech source) and generates the so-called *speech reference* signal, a block matrix which steers the spatial nulls to the direction of speech source and generates the so-call *noise reference* signals, and a *multi-channel noise canceller* which suppress the residual noise components in the speech reference signal by using a multi-channel adaptive filter.

In the adaptive (e.g., Frost and GSC) beamformers, adaptive signal processing (e.g., LMS) is normally used to avoid cancellation of the desired speech signal [52, 62]. However, adaptive signal processing systems do not show a high enough convergence rate and a high stability in practical environments. Moreover, the adaptive beamformers only perform well and provide acceptable performance when the number of interfering noise sources is less than that of the microphones. Their performance will be greatly degraded by the reverberation effect and in the scenario where more noise sources exist (e.g., larger than the number of sensors).

3.2 Proposed localized noise suppression algorithm

To suppress localized noises, as mentioned above in 3.1, many beamforming (e.g., fixed or adaptive beamforming) based algorithms have been reported with the drawbacks of a large physical size (DS beamformer), the limited performance for time-varying acoustic environment (fixed beamformer) and the low performance in multiple-noise-source scenarios (adaptive beamformer). These disadvantages make the existing beamforming algorithms difficult to be put into use in practical acoustic environments.

In this section, we propose a novel localized noise suppression algorithm which deals

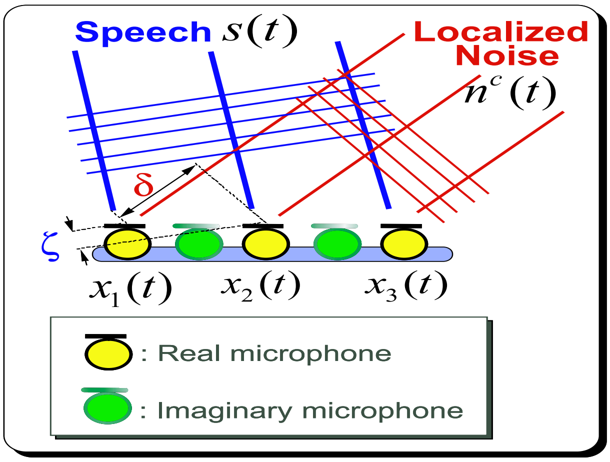


Figure 3.1: Microphone array for localized noise suppression.

with the drawbacks of the conventional beamformers and is able to reduce highly non-stationary (e.g., sudden) noise, multi-source noise and near field noise with only three microphones. This proposed algorithm enhances the desired speech signal by subtracting the spectral estimates of localized noises which are computed using a hybrid noise estimation technique from those of the noisy signals on each channel.

3.2.1 Overview of the proposed algorithm

To deal with localized noises, an array with three microphones is assumed in a noisy environment, as shown in Fig. 3.1. The time-aligned signals on all channels at the time delay compensation output are then further processed to estimate the spectra of localized noises using a hybrid noise estimation technique. The estimated spectra of localized noises are then reduced from those of the observed noisy signals using spectral subtraction. The block diagram of the proposed localized noises suppression algorithm is plotted in Fig. 3.2. As Fig. 3.2 shows, the noise spectrum on each channel is individually estimated and then individually subtracted from the spectrum of the noisy signal on the corresponding channel. Hence, three enhanced speech signals (localized-noise-suppressed signals) are outputted, which will be further processed to suppress non-localized noise components by the post-filter detailed in chapter 4.

3.2.2 Hybrid noise estimation technique

As Fig. 3.2 demonstrates, the localized noise spectrum is estimated and subtracted in each channel. Since the estimation-and-reduction mechanism performs in each channel in a similar or same way, to simplify the explanation, we just pay attention to the estimation-and-reduction mechanism in one channel (e.g., the second channel) to show how the

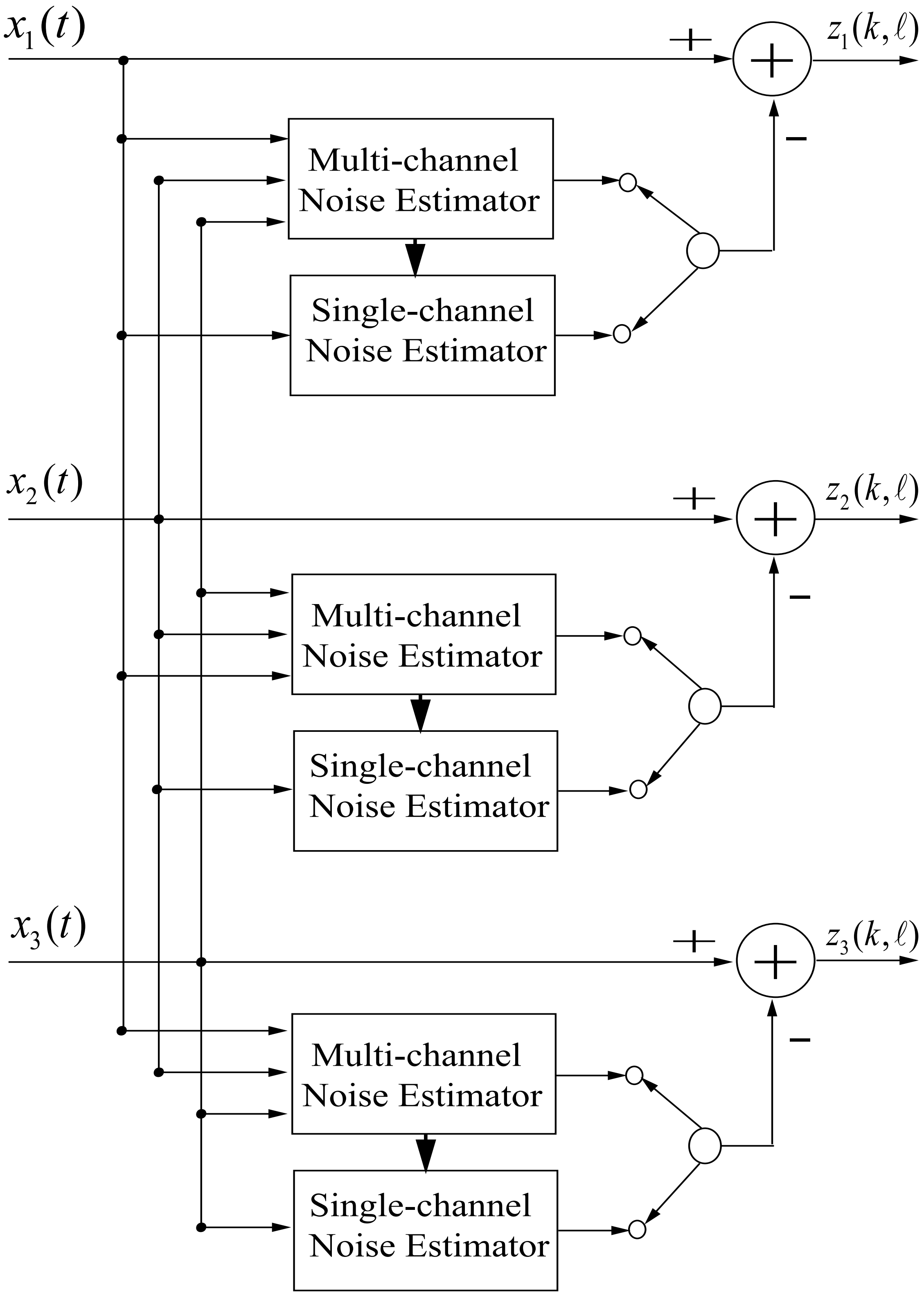

$x_1(t)$
$z_1(k,\ell)$
+
−
Multi-channel Noise Estimator
Single-channel Noise Estimator
$x_2(t)$
$z_2(k,\ell)$
+
−
Multi-channel Noise Estimator
Single-channel Noise Estimator
$x_3(t)$
$z_3(k,\ell)$
+
−
Multi-channel Noise Estimator
Single-channel Noise Estimator

Figure 3.2: Block diagram of the proposed algorithm for localized noise suppression.

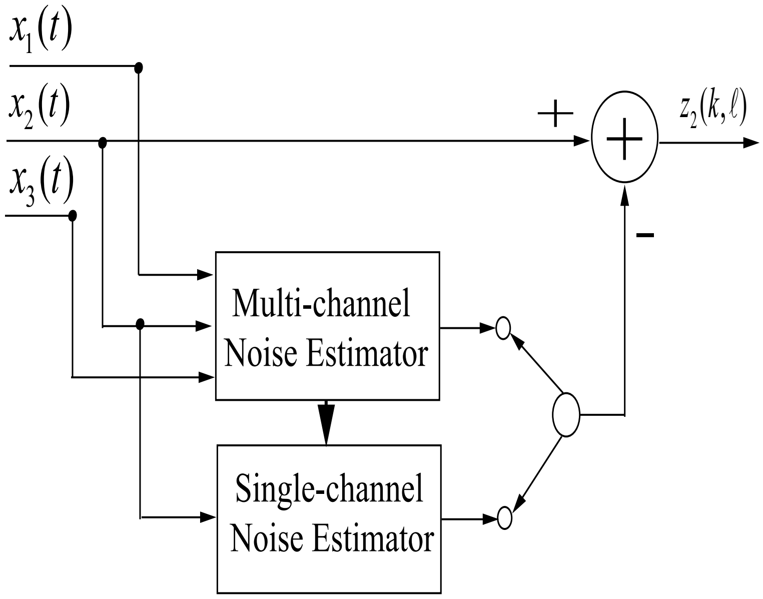


Figure 3.3: Block diagram of the proposed algorithm for localized noise suppression on the second channel.

proposed algorithm works. The proposed algorithm on the second channel is re-plotted in Fig. 3.3 again for explanation simplification.

The spectra of localized noises are estimated using a hybrid estimation technique which combines a multi-channel (subtractive beamformer based) estimation approach and a single-channel (soft-decision based) estimation approach. By integrating a novel *speech absence probability* (SAP) estimator, this combination between multi- and single-channel estimation approaches becomes very close and tight, yielding much more accurate spectral estimates for localized noises. Furthermore, the accurate noise spectral estimates are then subtracted from those of the noisy observations, producing the enhanced signal with less speech distortion.

In the following, we begin to introduce the hybrid noise estimation technique with the description of multi-channel localized noise estimation approach on which the proposed hybrid estimation technique is based. The multi-channel estimation approach was first presented by Akagi and Mizumachi [1, 2, 3, 114, 116] and its performance is further improved by integrating a single-channel soft-decision based noise estimation approach. Furthermore, a novel *robust and accurate speech absence probability* (RA-SAP) estimator is then presented to improve the estimation accuracy of the hybrid estimation approach for localized noises.

A. Multi-channel noise estimation approach [1, 2, 3]

Considering a three-microphone array in a noisy environment, the observed signals are composed of the desired speech signal, localized noise signal and non-localized noise signal. Assume that the *time delay compensation* (TDC) has been performed in advance. At the TDC output, by setting $\zeta = 0$ in Eqs. (2.26) - (2.28) the signals can be rewritten as:

$$x_1(t) = s(t) + \sum_{p=1}^{P} n_p^c(t - \delta_p) + n_1^{uc}(t), \tag{3.1}$$

$$x_2(t) = s(t) + \sum_{p=1}^{P} n_p^c(t) + n_2^{uc}(t), \tag{3.2}$$

$$x_3(t) = s(t) + \sum_{p=1}^{P} n_p^c(t + \delta_p) + n_3^{uc}(t), \tag{3.3}$$

and their STFTs in the time-frequency domain are:

$$X_1(k,\ell) = S(k,\ell) + \sum_{p=1}^{P} N_p^c(k,\ell)e^{-2jk\pi\delta_p} + N_1^{uc}(k,\ell), \tag{3.4}$$

$$X_2(k,\ell) = S(k,\ell) + \sum_{p=1}^{P} N_p^c(k,\ell) + N_2^{uc}(k,\ell), \tag{3.5}$$

$$X_3(k,\ell) = S(k,\ell) + \sum_{p=1}^{P} N_p^c(k,\ell)e^{2jk\pi\delta_p} + N_3^{uc}(k,\ell), \tag{3.6}$$

where $X_{\cdot}(k,\ell)$, $S(k,\ell)$, $N_{\cdot}^c(k,\ell)$ and $N_{\cdot}^{uc}(k,\ell)$ are the STFTs of the corresponding signals.

The multi-channel noise estimation approach consists of three steps: desired signal cancellation, noise direction estimation and noise spectral estimation, shown in Fig. 3.4. The desired speech signal is first cancelled by steering the spatial nulls to the direction of the desired speech signal, which yields the noise-only outputs. The directions of localized noises are then estimated based on the noise-only outputs using the *generalized cross-correlation* (GCC) based direction estimation method in each sub-band. The spectra of localized noises can further be calculated using the estimated directions and the noise-only outputs in each sub-band and finally are combined across all sub-bands. Each step will be described in the following in detail.

1. ***Desired signal cancellation***

   Model of the system used to estimate/reduce localized noises is constructed based on the knowledge about auditory physiology and/or psychoacoustics [1, 2, 3]. A model of neural-cancellation system is used to design our filters. In the original cancellation method, a periodical desired signal with periodicity of T is subtracted. Considering the delay time T to be the *interaural time difference* (ITD) for spatial

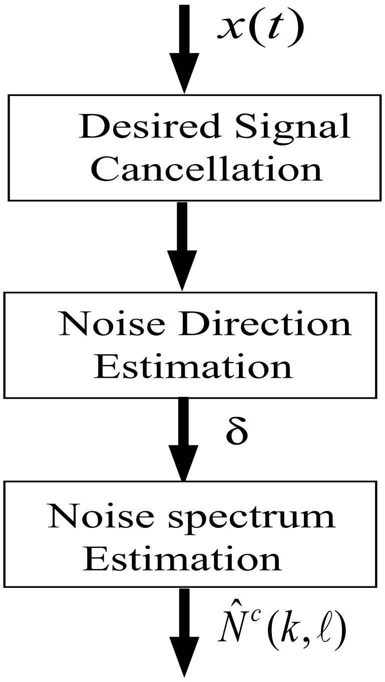

$x(t)$
Desired Signal Cancellation
Noise Direction Estimation
δ
Noise spectrum Estimation
$\hat{N}^c(k,\ell)$

Figure 3.4: Multi-channel noise estimation approach.

filtering, we design the method form the engineering point of view, shown in Fig. 3.5.

According to the basic circuit shown in Fig. 3.5 and with the signal models in Eqs. (3.1)-(3.3), the time-aligned microphone signals $x_1(t)$, $x_2(t)$ and $x_3(t)$ are first shifted $\pm\tau$ in the time domain ($\tau \neq 0$) and two beamformers in the time domain are constructed as [1, 2, 3]:

$$u_{13}(t) = \frac{1}{4}\Big\{\Big[x_1(t+\tau) - x_1(t-\tau)\Big] - \Big[x_3(t+\tau) - x_3(t-\tau)\Big]\Big\}, \quad (3.7)$$

$$u_{23}(t) = \frac{1}{4}\Big\{\Big[x_2(t+\tau) - x_2(t-\tau)\Big] - \Big[x_3(t+\tau) - x_3(t-\tau)\Big]\Big\}. \quad (3.8)$$

In order to simplify the implementation, the differences of non-localized noises at different microphones are assumed to be small enough to be ignored, which is normally satisfied in a diffuse noise field. The two beamformers in the frequency domain can then be calculated as:

$$U_{13}(k,\ell) = \sin\Big(2k\pi\tau\Big)\sum_{p=1}^{P} N_p^c(k,\ell)\sin\Big(2k\pi\delta_p\Big), \quad (3.9)$$

$$U_{23}(k,\ell) = \sin\Big(2k\pi\tau\Big)\sum_{p=1}^{P} N_p^c(k,\ell)e^{jk\pi\delta_p}\sin\Big(k\pi\delta_p\Big), \quad (3.10)$$

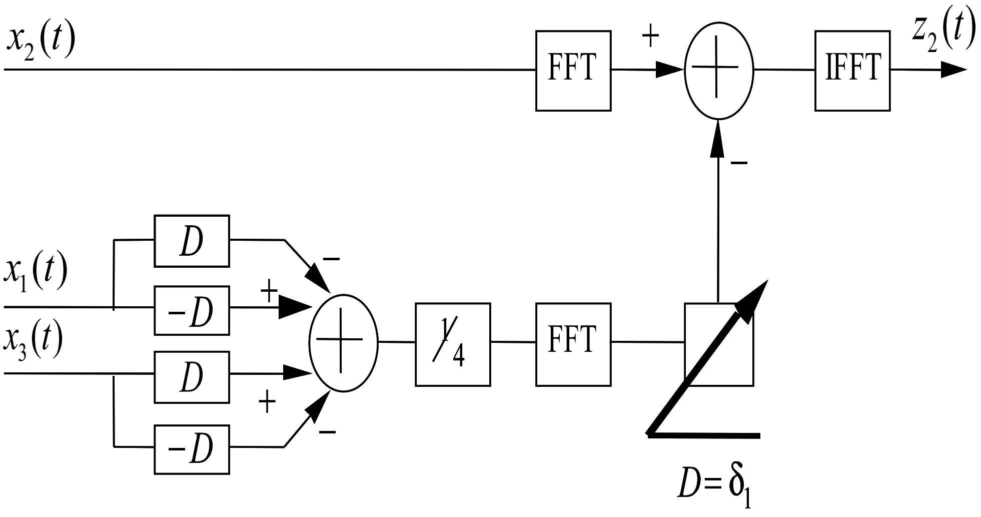


Figure 3.5: Basic circuit of the multi-channel noise estimation/reduction system [1, 3].

where $N_p^c(k,\ell)$ is the STFT of the noise $n_p^c(t)$. Note that the outputs of two beamformers do not include the desired speech components, which have been cancelled successfully.

2. ***Division of frequency band into sub-bands***

Since that the sum of sinusoidal waves becomes a sinusoidal wave in a narrow sub-band [3], we divide the full frequency band into several narrow sub-bands. Then we can further assume that the multiple noise sources can be regarded as one integrated interfering noise source in each sub-band. Assuming that $k_{i-1} \le \tilde{k} < k_i$, $k_i - k_{i-1} < \varepsilon$ $(i = 1, 2, \ldots, I)$, $k_0 = 0$ and ε is a very small value. Consequently, the beamformer output signals in each sub-band can be represented as:

$$\sum_{p=1}^{P} N_p^c(\tilde{k},\ell)\sin\left(2\tilde{k}\pi\delta_p\right) = N_i^c(\tilde{k},\ell)\sin\left(2\tilde{k}\pi\delta_i\right), \quad i = 1, 2, \ldots, I \tag{3.11}$$

$$U_{13}(\tilde{k},\ell) = \sin\left(2\tilde{k}\pi\tau\right)\sin\left(2\tilde{k}\pi\delta_i\right)N_i^c(\tilde{k},\ell), \tag{3.12}$$

$$U_{23}(\tilde{k},\ell) = \sin\left(2\tilde{k}\pi\tau\right)\sin\left(\tilde{k}\pi\delta_i\right)N_i^c(\tilde{k},\ell)e^{j\tilde{k}\pi\delta_i}. \quad k_{i-1} \le \tilde{k} < k_i \tag{3.13}$$

Note that $N_i^c(\tilde{k},\ell)$ is the integrated noise spectrum in the i-th sub-band and δ_i is the virtual time difference when assuming that the number of noise source is one.

3. ***Noise direction estimation***

Eqs. (3.12) and (3.13) show that the spectrum of the integrated noise can be estimated from that of the speech-cancelled signals $U_{13}(\tilde{k},\ell)$ and $U_{23}(\tilde{k},\ell)$, and that the noise direction information δ_i in each sub-band is a "must" for estimating the spectrum of localized noise.

To estimate the virtual noise direction δ_i in i-th sub-band, another subtractive beamformer is defined:

$$\begin{aligned} u_{12}(t) &= \frac{1}{4}\left\{\Big[x_1(t+\tau) - x_1(t-\tau)\Big] - \Big[x_2(t+\tau) - x_2(t-\tau)\Big]\right\}, && (3.14) \\ U_{12}(\tilde{k},\ell) &= \sin\left(2\tilde{k}\pi\tau\right)N_i(\tilde{k},\ell)e^{-j\tilde{k}\pi\delta_i}\sin\left(\tilde{k}\pi\delta_i\right). \quad k_{i-1} \le \tilde{k} < k_i && (3.15) \end{aligned}$$

With the outputs of two beamformers, the virtual arrival time difference δ_i in i-th sub-band is automatically estimated using the GCC direction estimation technique frame by frame in each sub-band, given by:

$$\delta_i = \arg\max_t \left[\text{IFFT}\left[\frac{U_{12}(\tilde{k},\ell)U_{23}^*(\tilde{k},\ell)}{\left|U_{12}(\tilde{k},\ell)\right|\left|U_{23}^*(\tilde{k},\ell)\right|}\right]\right], \quad k_{i-1} \le \tilde{k} < k_i. \tag{3.16}$$

The value δ_i, $(i = 1, 2, \ldots, I)$ is the half of the difference in the virtual arrival time difference between two microphones x_1 and x_3. Moreover, it should be noted that since $U_{23}(\tilde{k},\ell)$ in Eq. (3.13) and $U_{12}(\tilde{k},\ell)$ in Eq. (3.15) do not include the target signal at all, the desired speech signal has no effect on the estimation of noise direction.

4. ***Noise spectral estimation***

Eq. (3.12) shows that the spectrum of the integrated noise can be estimated from that of the speech-cancelled signal $U_{13}(\tilde{k},\ell)$, since the speech-cancelled signal does not contain any desired speech signal. With the virtual DOA of the integrated noise signal, the spectral estimate of the integrated noise $\hat{N}^c_{m,i}(\tilde{k},\ell)$ in i-th sub-band can be calculated from the speech-cancelled signals. Then, the spectral estimate of the integrated noise $\hat{N}^c_m(k,\ell)$ can be calculated over the entire frequency region as $\left(\tau = \delta_i \text{ in Eq. (3.9) and } \tau = \frac{\delta_i}{2} \text{ in Eq. (3.10)}\right)$:

$$\hat{N}^c_{mul,i}(\tilde{k},\ell) = \begin{cases} U_{13}(\tilde{k},\ell)\Big/\sin^2\left(2\tilde{k}\pi\delta_i\right), & \sin^2\left(2\tilde{k}\pi\delta_i\right) > \varepsilon_1 \\ U_{23}(\tilde{k},\ell)e^{-j\tilde{k}\pi\delta_i}\Big/\sin^2\left(\tilde{k}\pi\delta_i\right), & \sin^2\left(2\tilde{k}\pi\delta_i\right) < \varepsilon_1 \text{ and } \sin^2\left(\tilde{k}\pi\delta_i\right) > \varepsilon_2 \\ U_{23}(\tilde{k},\ell)\Big/\varepsilon_2, & \text{otherwise} \end{cases} \tag{3.17}$$

$$\hat{N}^c_{mul}(k,\ell) \;=\; \sum_{i}^{I} \hat{N}^c_{mul,i}(\tilde{k},\ell), \quad k_{i-1} \leq \tilde{k} < k_i \tag{3.18}$$

where $\hat{N}^c_{mul}(k,\ell)$ indicates that the spectral values are estimated by the multi-channel technique; ε_1 and ε_2 are two threshold values determined experimentally ($\varepsilon_1 = 0.5$ and $\varepsilon_2 = 0.1$ in this work).

B. Proposed hybrid noise estimation approach

To improve the estimation accuracy of the multi-channel estimation approach, in common sense, it is necessary to first know what are problems and what are the sources which cause these problems. Therefore, in this part, we first discuss the problems of the multi-channel estimation approach and then describe the findings about the causes of these problems, and finally present a hybrid noise estimation technique, as a solution to these problems, which succeed in mitigating the estimation errors and further increasing the estimation accuracy for the localized noises.

1. ***Problems of multi-channel estimation approach***

   The multi-channel noise estimation approach has a great ability to estimate the spectra for localized noises which arrive from some certain determinable directions, providing much more accurate spectral estimates for localized noises and further high performance in reducing localized noises by the non-linear spectral subtraction [1, 2, 3, 114, 116, 118]. However, for the multi-channel noise estimation approach, there still is a large room to improve since its estimation accuracy is expected to be greatly degraded in some cases.

   To demonstrate the problems which degrade the estimation accuracy of the multi-channel noise estimation approach, we provide a typical example shown in Fig. 3.6. In this example, the clean speech signals are corrupted by the band-limited (low-frequency) highly non-stationary intermittent localized noises; the noisy signals are then processed by the multi-channel noise estimation approach and the non-linear spectral subtraction [3]. Both the noisy (noise-corrupted) signal and the enhanced signal are plotted in Fig. 3.6. From Fig. 3.6, it is obvious to see that the localized noises are remained in some frequency bins in the enhanced signal after processed by the multi-channel estimation approach based noise reduction method [3]. This problem of residual noise components is caused by the fact that localized noise components can not be accurately estimated by the multi-channel noise estimation approach in those frequency bins. As a result, the low estimation accuracy of the multi-channel noise estimation approach results in the low performance of the multi-channel estimation approach based noise reduction algorithm in reducing localized noises.

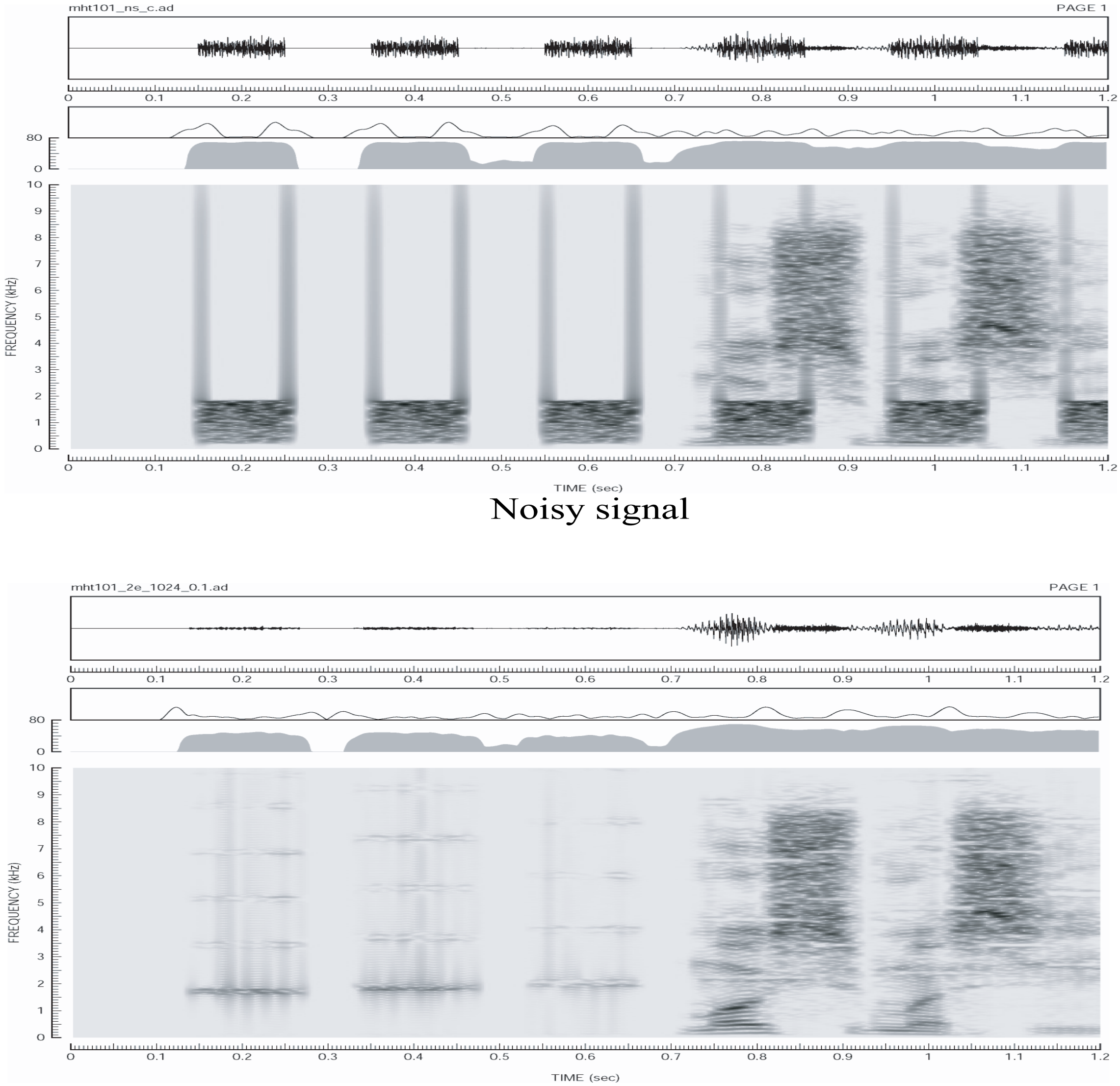
Keeping the problems of the multi-channel noise estimation approach in mind, we turn to look for the causes/reasons of these problems. We again look at Eq. 3.17 which is used to estimate the localized noise spectra by the multi-channel estimation approach. In theory, the first two equations yield exactly accurate localized noise spectra, while the estimation accuracy will be greatly degraded if the third equation as an approximation is used. And this approximation should be used only when the values of both $\sin^2\left(2\tilde{k}\pi\delta\right)$ and $\sin^2\left(\tilde{k}\pi\delta\right)$ are very small (smaller than a certain threshold). In a further step, it is of interest to note that this approximation problem corresponds to the satisfaction of the condition $\tilde{k}\delta_i = integer$. That is, when the condition $\tilde{k}\delta_i = integer$ holds, the approximation is used which further results in the considerable estimation error for localized noises. Moreover, the satisfaction of the condition $\tilde{k}\delta_i = integer$ demonstrates that the multi-channel estimation approach fails for the localized noises in some frequencies and some DOAs. That is, the estimation accuracy of the multi-channel noise estimation approach is dependent on the frequencies and the DOAs of localized noises. Furthermore, it should be noted that, in principle, the problem of the failure of the multi-channel estimation approach in estimating localized noise spectra (i.e., the satisfaction of the condition $\tilde{k}\delta_i = integer$) corresponds to the grating sidelobes of the microphone arrays with small physical size.

2. ***Proposed hybrid noise estimation technique***

After knowing the problems of the multi-channel noise estimation approach and the causes of these problems, it is necessary to deal with the problems with the hope of improving the performance of the multi-channel noise estimation approach.

To deal with the problem of the low estimation accuracy of the multi-channel noise estimation approach, we propose a hybrid noise estimation technique for localized noises based on the discussions described in the last section. In the proposed hybrid noise estimation technique, a single-channel estimation approach is exploited when the multi-channel estimation approach fails. That is, the hybrid noise estimation technique combines the multi-channel estimation approach and the single-channel estimation approach, shown in Fig. 3.2.

In the proposed hybrid estimation technique, the multi-channel estimation approach and the single-channel estimation approach are working at all times continuously. At each time instant, both the output of the multi-channel noise estimation approach and that of the single-channel noise estimation approach are available and might be the final output of this hybrid noise estimation technique. The control condition for choosing the output of the multi-channel estimation approach or that of the single-channel estimation approach is motivated by the causes of the problems of the multi-channel estimation approach, that is, the condition $\tilde{k}\delta_i = integer$, as



Noisy signal

Enhanced signal

Figure 3.6: An example which shows the estimation error of the multi-channel estimation approach. Spectrogram of the noisy speech signal (top) and spectrogram of the enhanced speech signal (bottom).

discussed in last section. Accordingly, in this hybrid estimation technique, the values of $\sin^2\left(2\tilde{k}\pi\delta\right)$ and $\sin^2\left(\tilde{k}\pi\delta\right)$ determine whether the output of the single-channel approach or that of the multi-channel approach should be the final output of this hybrid estimation technique. When the maximum of $\sin^2\left(2\tilde{k}\pi\delta\right)$ and $\sin^2\left(\tilde{k}\pi\delta\right)$ is larger than a threshold ε (an empirical constant), the output of the multi-channel approach is more accurate and preferred to be the final output. Otherwise, the output of the single-channel approach is more accurate and preferred to be the final output. Thus, the final spectral estimates of the localized noises by the proposed hybrid estimation technique can be given by:

$$\left|\hat{N}^c(\tilde{k},\ell)\right| = \begin{cases} \left|\hat{N}^c_{mul}(\tilde{k},\ell)\right|, & \max\left(\sin^2\left(2\tilde{k}\pi\delta\right),\sin^2\left(\tilde{k}\pi\delta\right)\right) > \varepsilon, \\ \left|\hat{N}^c_{sig}(\tilde{k},\ell)\right|, & \text{otherwise.} \end{cases} \tag{3.19}$$
$$\left(k_{i-1} \leq \tilde{k} < k_i, i = 1, 2, \ldots, I\right)$$

where $\left|\cdot\right|$ is the amplitude operator, $N^c_{mul}(\tilde{k},\ell)$ and $N^c_{sig}(\tilde{k},\ell)$ represent the spectral estimates of localized noises by the multi-channel approach, given by Eq. (3.17) and by the single-channel approach detailed in the following section. With this hybrid noise estimation technique, high accurate spectral estimates for localized noises can be expected, and in a general sense, the grating sidelobes of the microphone arrays with small physical size can be expected to be mitigated.

C. Single-channel noise estimation approach

As discussed in 3.2.2 and shown by Eq. (3.19), the spectra of localized noises $\hat{N}^c_{sig}(k,\ell)$ should be computed by a single-channel estimation approach.

Though many single-channel noise estimation methods have been published so far, in this work, we adopt a soft-decision based single-channel estimation approach since it can adaptively update the noise spectra. The soft-decision single-channel approach updates the noise spectral estimates by averaging the noisy speech power spectrum using time and frequency dependent smoothing factors, which are adjusted based on speech absence probability in individual frequency bins. Using this soft-decision based estimation method, the amplitude spectra and power spectra of localized noises can be computed as:

$$\left|\hat{N}^c_{sig}(k,\ell)\right| = \left[\lambda_n(k,\ell)\right]^{\frac{1}{2}}, \tag{3.20}$$

and

$$\lambda_n(k,\ell) = \alpha_n\lambda_n(k,\ell-1) + (1-\alpha_n)E\left[\left|N(k,\ell)\right|^2\middle|X_2(k,\ell)\right], \tag{3.21}$$

where $\alpha_n(0 < \alpha_n < 1)$ is a forgetting factor controlling the update rate of noise estimation. Under the speech presence uncertainty, the second term in the right side of Eq. (3.21) can be estimated as the spectra of observed signal during speech pauses or it holds the values obtained in the previous pauses during speech active periods. Hence, the instantaneous spectral estimates of the noises can be given by:

$$E\left[\left|N(k,\ell)\right|^2\Big|X_2(k,\ell)\right] \;=\; q(k,\ell)\left|X_2(k,\ell)\right|^2+\Big(1-q(k,\ell)\Big)\lambda_n(k,\ell-1), \quad (3.22)$$

where $q(k,\ell)$ is the *speech absence probability* (SAP). As Eq. (3.22) shows, the spectral estimation capability of the single-channel approach is significantly dependent on the successfulness or failure of the SAP estimator [28]. Therefore, its performance is able to be enhanced by integrating a RA-SAP estimator.

D. Further enhance hybrid estimation technique with a RA-SAP estimator

In this part, we further enhance the proposed hybrid noise estimation technique by integrating a *robust and accurate speech absence probability* (RA-SAP) estimator. Considering the strong correlation of speech presence uncertainty between adjacent frequency bins and consecutive frames and making full use of the frequently-perfect high estimation accuracy of the multi-channel approach, a RA-SAP estimator is developed which improves the performance of the hybrid noise estimation technique by combining the multi-channel and single-channel approaches in an effective and tight way.

Assume that the real part and imaginary part of the STFTs of speech and noise signals have the Gaussian distributions. Applying the Bayes rule and total probability theorem, the SAP, which is a conditional probability of speech absent state given noisy observations and denoted by $q(k,\ell)$ for notational simplicity, can be given by [109]:

$$q(k,\ell) \;=\; \left(1+\frac{1-q^{'}(k,\ell)}{q^{'}(k,\ell)}\frac{1}{1+\xi(k,\ell)}\exp\left(\frac{\xi(k,\ell)\gamma(k,\ell)}{1+\gamma(k,\ell)}\right)\right)^{-1}, \quad (3.23)$$

where $q^{'}(k,\ell)$ is the *a priori* speech absence probability; $\xi(k,\ell)=\lambda_s(k,\ell)/\lambda_n(k,\ell)$ and $\gamma(k,\ell)=|X_2(k,\ell)|^2/\lambda_n(k,\ell)$ are the *a priori* SNR and *a posteriori* SNR, as named in [43], and $\lambda_s(k,\ell)$ and $\lambda_n(k,\ell)$ represent the variance of speech signal and noise signal respectively.

Eq. (3.23) demonstrates that for the given *a priori* SAP $q^{'}(k,\ell)$, the speech absence probability $q(k,\ell)$ is greatly dependent on the *a priori* SNR $\xi(k,\ell)$ and *a posteriori* SNR $\gamma(k,\ell)$. It is believed that accurate and robust SAP estimates can be obtained only when $\xi(k,\ell)$ and $\gamma(k,\ell)$ are accurate and robust enough. Consequently, we now turn to the issue of improving the accuracy and robustness of the *a priori* SNR $\xi(k,\ell)$ and *a posteriori* SNR $\gamma(k,\ell)$ estimates.

Taking into account the strong correlation of speech presence uncertainty in adjacent frequency bins and consecutive frames, we propose two estimators for the *a priori* SNR $\xi(k,\ell)$ and *a posteriori* SNR $\gamma(k,\ell)$.

- In the frequency domain, the estimates of $\xi(k,\ell)$ and $\gamma(k,\ell)$ are smoothed by applying a normalized window b (e.g., hamming window) of size $2D+1$, given by:

  $$\begin{aligned} \tilde{\xi}(k,\ell) &= \sum_{\omega=k-D}^{k+D} b(\omega)\xi(\omega,\ell), && (3.24) \\ \tilde{\gamma}(k,\ell) &= \sum_{\omega=k-D}^{k+D} b(\omega)\gamma(\omega,\ell). && (3.25) \end{aligned}$$

  Estimation accuracy of $\xi(k,\ell)$ and $\gamma(k,\ell)$ is improved due to the fact that the noise spectra in adjacent frequencies are likely to be estimated by the multi-channel estimation technique which is able to yield high accurate spectral estimates for localized noises. Furthermore, this frequency-smoothing procedure eliminates fluctuations of the *a priori* SNR $\xi(k,\ell)$ and *a posteriori* SNR $\gamma(k,\ell)$ along the frequency axis on the time-frequency plane, which results in more robust SNR estimates.

- In the time domain, the frequency-smoothed estimates of the *a priori* SNR $\tilde{\xi}(k,\ell)$ and *a posteriori* SNR $\tilde{\gamma}(k,\ell)$ are further processed based on the previous values in an iterative way, given by:

  $$\begin{aligned} \bar{\xi}(k,\ell) &= \alpha_{\xi}\frac{{Z_2}^2(k,\ell-1)}{\eta_n(k,\ell-1)} + (1-\alpha_{\xi})\max[\tilde{\gamma}(k,\ell)-1,0], && (3.26) \\ \bar{\gamma}(k,\ell) &= \tilde{\gamma}(k,\ell), && (3.27) \end{aligned}$$

  where α_{ξ} $(0 < \alpha_{\xi} < 1)$ is a forgetting factor and $\tilde{Z}_2(k,\ell-1)$ is the enhanced signal by spectral subtraction in the previous frame. Actually, Eq. (3.26) is just the decision-directed scheme detailed in [43]. It is of interest to note that the smoothing operation in the time domain is not carried out for the *a posteriori* SNR, since it should be calculated from the current observations and independent on the previous observations.

Based on the time-frequency smoothed *a priori* SNR $\bar{\xi}(k,\ell)$ and the *a posteriori* SNR $\bar{\gamma}(k,\ell)$, the robust and accurate speech absence probability $q(k,\ell)$ can be obtained as:

$$q(k,\ell) = \left(1 + \frac{1-q^{'}(k,\ell)}{q^{'}(k,\ell)}\frac{1}{1+\bar{\xi}(k,\ell)}\exp\left(\frac{\bar{\xi}(k,\ell)\bar{\gamma}(k,\ell)}{1+\bar{\gamma}(k,\ell)}\right)\right)^{-1}. \qquad (3.28)$$

where the *a priori* SAP $q^{'}(k,\ell)$ used in this research is given by [28]:

$$q^{'}(k,\ell) = 1 - P_{local}(k,\ell)P_{global}(k,\ell)P_{frame}(k,\ell). \qquad (3.29)$$

where $P_{local}(k,\ell)$, $P_{global}(k,\ell)$ and $P_{frame}(k,\ell)$ correspond to the speech presence probabilities which are estimated based on the speech energy distribution in a local frequency window, a larger frequency window and neighboring frames in the time-frequency domain respectively, detailed in [28].

A similar time-frequency smoothing procedure has been presented by Cohen *et al.* in a single-channel algorithm to estimate the *a priori* SAP [28]. Cohen's algorithm improves the robustness of the *a priori* SAP estimates, however, it can not improve their estimation accuracy since the estimation accuracy of noise spectra are not improved by the time-frequency smoothing procedure.

In the proposed hybrid estimation technique, both robustness and accuracy of SAP estimates would be improved by exploiting the newly presented time-frequency smoothing scheme, when multi-channel microphones are available. Furthermore, the RA-SAP estimator provides much higher noise estimation accuracy for the hybrid estimation technique, from several aspects. Firstly, as shown in Eq. (3.19), the final spectral estimates are computed by the multi-channel approach which produces the exact high accurate spectral estimates in most cases. Secondly, estimation accuracy of the single-channel approach is significantly improved, which is attributed to the multi-channel estimation approach and this RA-SAP estimator. Since accurate spectral estimates by the multi-channel approach are likely to be distributed around those determined by the single-channel approach, accuracy and robustness of the SAP estimator can be ensured by applying the time-frequency smoothed *a priori* SNR and *a posteriori* SNR which are more accurate. Furthermore, this RA-SAP estimator greatly improves the estimation accuracy of the single-channel approach. Finally, the improved single-channel approach contributes to enhance the final estimation accuracy of the hybrid noise estimation technique.

3.2.3 Localized noise suppression with spectral subtraction

The proposed hybrid noise estimation technique gives accurate spectral estimates for localized noises. Subsequently, the estimated spectra are subtracted from those of the observed noisy signals by spectral subtraction, given by [13]:

$$\big|Z_2(k,\ell)\big| \;=\; \begin{cases} \big|X_2(k,\ell)\big| - \alpha\big|\hat{N}^c(k,\ell)\big|, & \big|X_2(k,\ell)\big| > \alpha\big|\hat{N}^c(k,\ell)\big|, \\ \beta\big|X_2(k,\ell)\big|, & \text{otherwise} \end{cases} \tag{3.30}$$

where α and β are the overestimation factor and spectral floor factor. Since the spectral estimates of localized noises are of high accuracy, $\alpha = 1$ is set to avoid distorting the speech signal. And β is determined experimentally.

3.2.4 Experimental validation

In the experiments, we concentrate on the improvement in estimation accuracy of the proposed hybrid noise estimation technique for localized noises compared to the corresponding single-channel and multi-channel estimation approaches.

Sound data

To objectively evaluate the performance of the proposed hybrid noise estimation technique, 54 clean speech sentences, selected from ATR database and uttered by 3 male and 3 female speakers, are used. The tested noises consist of synthesized noises (white Gaussian noise and pink noise) and real-world noises (car noise, train noise, department noise and exhibition noise). The speech and noise signals are first resampled to 12 kHz and linearly quantized at 16 bits. The noisy signals are generated by mixing the clean speech signals with the localized tested noises with individual DOAs 10-80 degrees (10-degree step) to the right.

Evaluation measure

The performance of the proposed hybrid noise estimation technique is evaluated and further compared to that of the corresponding single-channel and multi-channel approaches in terms of *normalized estimation error* (NEE), defined as:

$$\text{NEE} \;=\; \frac{20}{L}\sum_{\ell=1}^{L}\log_{10}\left(\frac{\displaystyle\sum_{k=0}^{K-1}\left(\left|\hat{N}^c(k,\ell)-N^c(k,\ell)\right|\right)}{\displaystyle\sum_{k=0}^{K-1}\left|N^c(k,\ell)\right|}\right), \quad \text{[dB]} \tag{3.31}$$

where $\hat{N}^c(k,\ell)$ and $N^c(k,\ell)$ are the estimated noise spectrum and "ideal" noise spectrum respectively; K and L are the length of STFT and the number of frames. It should be noted that the smaller NEE represents the more accurate noise estimate obtained by the tested estimation technique.

Evaluation results

The NEE results averaged over the localized noise signals with different DOAs in the tested noise conditions are listed in Table 3.1. In the table 3.1, "single-channel" denotes the single-channel soft-decision based noise estimation approach alone [28], "multi-channel" means the multi-channel subtractive beamformer based noise estimation approach alone [3], "hybrid" represents our proposed noise estimation technique detailed in section 3.2.

Table 3.1 demonstrates that the normalized noise estimation error is consistently decreased for all the tested noise conditions, especially for localized car noise, when the proposed hybrid estimation technique is used. This improvement amounts to 3 dB compared to the single-channel estimation approach alone and 5 dB compared to the multi-channel estimation technique alone in localized car noise environment. Fig. 3.7 illustrates the typical examples of the NEE comparisons of the single-channel, multi-channel and hybrid

Table 3.1: Average NEEs [dB] in various noise conditions

	white	pink	car	train	department	exhibition
single-channel	-5.2842	-4.6423,	-6.7910	-4.8105	-4.5226	-5.4396
multi-channel	-12.8905	-10.2486,	-4.5205	-15.6283	-10.4331	-13.3236
hybrid	-13.3378	-11.1818	-9.7014	-15.5916	-12.3930	-14.0260

noise estimation techniques in localized white and car noise environments. All the observations obtained from Table 3.1 and Fig. 3.7 verify the superiority of our proposed hybrid noise estimation technique compared to the multi-channel estimation alone approach and the signal-channel estimation alone approach. The much more accurate spectral estimates for localized noise should yield higher noise reduction performance with minimum speech distortion by using non-linear spectral subtraction in Eq. (3.30).

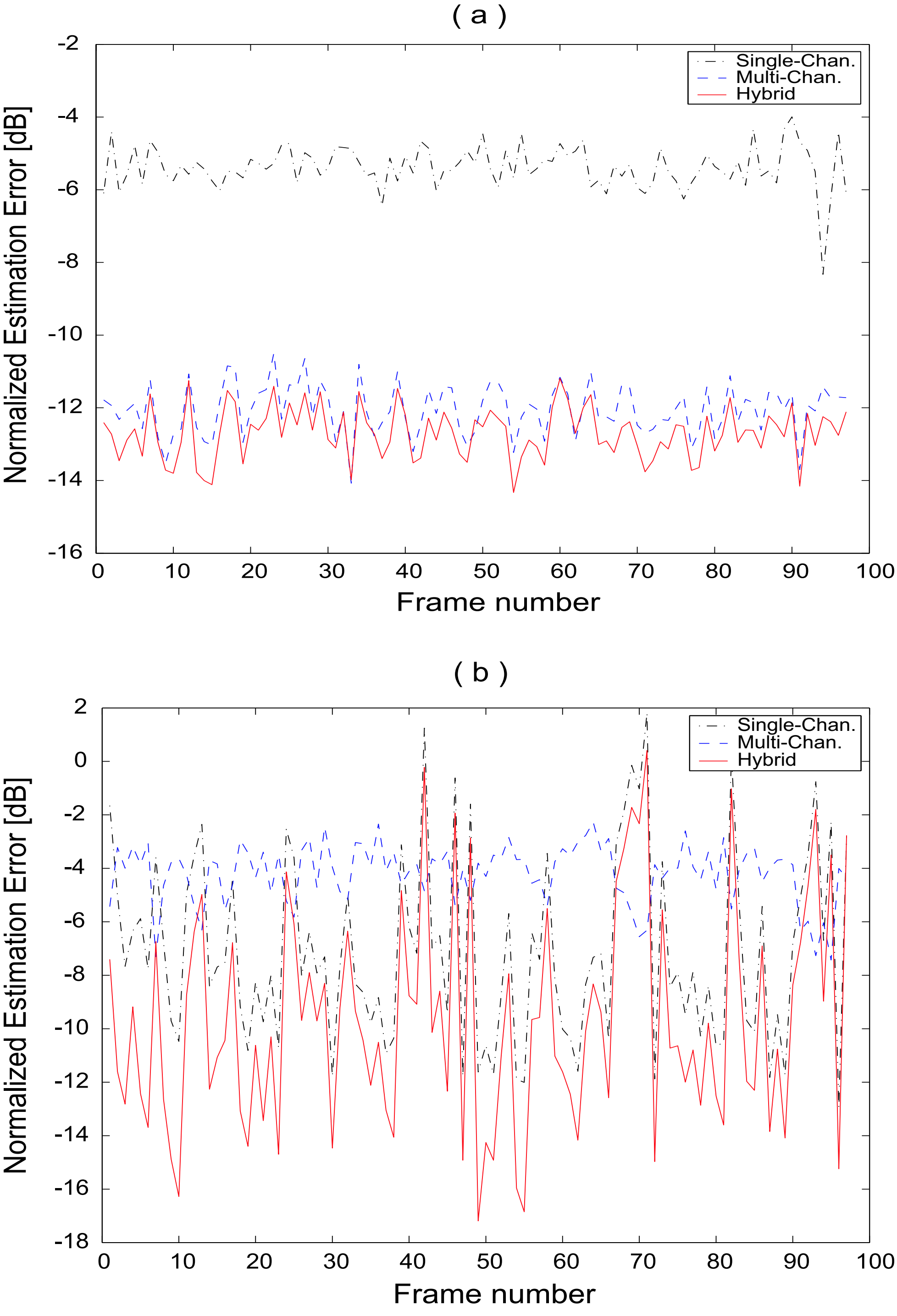


Figure 3.7: Normalized noise estimation error (dB) for signals processed by single-channel technique (dashdot), multi-channel technique (dashed) and hybrid technique (solid) under white noise conditions (a) and car noise conditions (b).

3.3 A generalized subtractive beamformer

Previously, many beamformers were reported with the drawbacks of a large physical size and low performance in time-varying acoustic environments (fixed beamformers) and low performance in multi-noise-source environments and reverberation conditions (for adaptive beamformers). The subtractive beamformer based algorithm described in section 3.2.2 succeeds in dealing with these drawbacks. In this subtractive beamformer based algorithm, noises were analytically estimated based on the arrival time difference between paired microphones instead of exploiting adaptive signal processing. Speech spectra are then enhanced by subtracting the estimated noise spectra from those of the observed noisy spectra. The superiority of this algorithm lies in its high ability to suppress localized noises, especially sudden noise, with only a small number of microphones and without adaptive signal processing techniques (e.g., LMS). The main problem associated with this algorithm is the assumption that only localized noise sources exist in the environment, corresponding to a perfectly coherent noise field. The practical noise condition is generally not a perfectly coherent noise field, e.g., in a car or a reverberant room which can be approximately modelled as a diffuse noise field [86, 108]. Therefore, the performance degradation of the subtractive beamformer based noise reduction algorithm we discussed in section 3.2.2 is expected in these environments.

In the following, we extend the subtractive beamformer described in section 3.2.2 to a generalized expression with the assumption of an arbitrary noise field. For explanation simplicity, whereafter, the subtractive beamformer described in section 3.2.2 is referred to as *original subtractive beamformer* and its generalized expression presented in this section is referred to as *generalized subtractive beamformer*. The generalized subtractive beamformer which has a GSC-like structure includes the original subtractive beamformer as a special case in a coherent noise field when only two microphones are available. Compared to other traditional algorithms, the generalized algorithm have some advantages: exploiting no adaptive signal processing techniques (e.g., LMS); performing well in various kinds of noise conditions due to the assumption of an arbitrary noise field and offering an improved noise reduction ability since much spatial information is considered. The performance of the generalized subtractive beamformer is then analyzed using the noise coherence functions in theoretically defined noise fields. Its superiority is further confirmed by the experiments using multi-channel recordings in car environments.

3.3.1 Problem formulation

To simplify the following explanation and without loss of generality, let us again assume that the time delay compensation has been done in advance. Therefore, the observed noisy signal $x_i(t)$ on i-th microphone is composed of the desired speech signal $s(t)$ and the additive noise $n_i(t)$, described as:

$$x_i(t) = s(t) + n_i(t). \quad i = 1, 2, ..., M \tag{3.32}$$

In the time-frequency domain, we have in the vector form as:

$$\mathbf{X}(k, \ell) = S(k, \ell) + \mathbf{N}(k, \ell), \tag{3.33}$$

where

$$\mathbf{X}^T(k, \ell) = \Big[X_1(k, \ell), X_2(k, \ell), \cdots, X_M(k, \ell)\Big], \tag{3.34}$$

$$\mathbf{N}^T(k, \ell) = \Big[N_1(k, \ell), N_2(k, \ell), \cdots, N_M(k, \ell)\Big]. \tag{3.35}$$

Note that compared with the original subtractive beamformer, this proposed subtractive beamformer has two generalizations: (1) the additive noise signal on each microphone includes all undesired signals, which might be composed of coherent and incoherent components, not only perfectly coherent noise as assumed in the original subtractive beamformer; (2) the number of microphones is M, not only two as in the original subtractive beamformer which is a pair-microphone based algorithm.

3.3.2 Derivation of the generalized subtractive beamformer

The generalized subtractive beamformer based noise reduction algorithm, which has a GSC-like structure, is shown in Fig. 3.8. This method is composed of three components: a *fixed beamformer* (FBF) which constructs the speech reference signal in the upper path, a *blocking matrix* (BM) which blocks the desired speech signal and constructs the noise reference signal, and a *noise canceller* (NC) which suppresses the residual noise by minimizing the power of the system output. The three parts of the generalized subtractive beamformer based algorithm are implemented as follows:

1. *Fixed beamformer.* To be consistent with the original subtractive beamformer based algorithm in section 3.2.2 and make the implementation simple, the FBF of the generalized algorithm is an all-pass filter for the signal on the reference channel (e.g., the first microphone) and blocks the signals from other microphones. Thus, the output of FBF $Y_{FBF}(\omega)$, which is the speech reference signal, is given by:

   $$Y_{FBF}(k, \ell) = X_1(k, \ell) = \mathbf{W}^\dagger \mathbf{X}(k, \ell), \tag{3.36}$$

   where † denotes conjugation transpose and $\mathbf{W}^\dagger = [1, 0, \cdots, 0]$.

   Note, comparatively, in the original GSC beamformer [62], the FBF was usually implemented by the DSBF which introduces some additional "NULLs" in the beam patten of this beamformer, as detailed in [1, 2, 3].

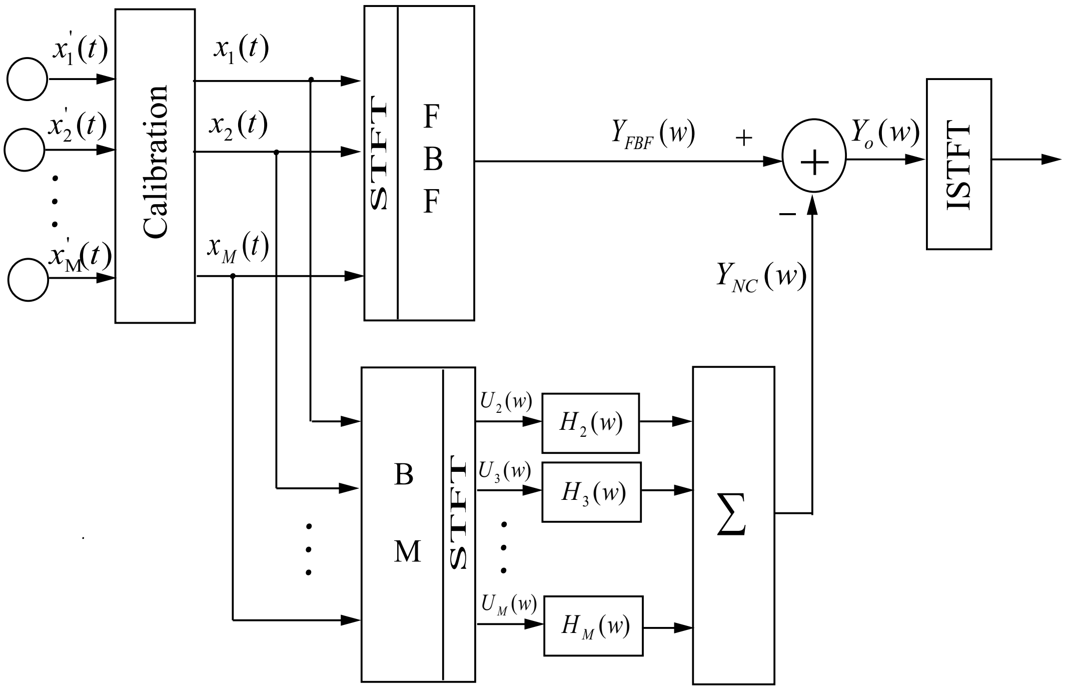


Figure 3.8: Block diagram of the generalized subtractive beamformer.

2. *Blocking matrix.* Since the original subtractive beamformer successfully blocks the desired speech components, the BM part of the generalized algorithm is implemented by using the same mechanism, defined as ($\tau \neq 0$):

$$u_{1i}(t) = \frac{1}{4}\left\{\Big[x_1(t+\tau) - x_1(t-\tau)\Big] - \Big[x_i(t+\tau) - x_i(t-\tau)\Big]\right\}. \quad i = 2, \ldots, M \quad (3.37)$$

With the generalized signal model, shown in Eq. (3.32), the corresponding representation of this beamformer in the time-frequency domain can be described as:

$$\begin{aligned} U_{1i}(k,\ell) &= \frac{1}{2} j \sin\Big(2k\pi\tau\Big)\Big(N_1(k,\ell) - N_i(k,\ell)\Big) \\ &= \frac{1}{2} j \sin\Big(2k\pi\tau\Big)\Big(X_1(k,\ell) - X_i(k,\ell)\Big). \end{aligned} \quad (3.38)$$

That is, we have in vector form as:

$$\mathbf{U}(k,\ell) = \mathbf{B}^{\dagger}(k,\ell)\mathbf{X}(k,\ell), \quad (3.39)$$

where $\mathbf{U}(k,\ell)$ and $\mathbf{B}^{\dagger}(k,\ell)$ are:

$$\mathbf{U}^T(k,\ell) = \Big[U_{12}(k,\ell), U_{13}(k,\ell), \cdots, U_{1M}(k,\ell)\Big], \quad (3.40)$$

$$\begin{aligned}\mathbf{B}^{\dagger}(k,\ell) &= \frac{1}{2}j\sin(2k\pi\tau)\begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix} \\ &\triangleq \frac{1}{2}j\sin(2k\pi\tau)\mathbf{B}_1^{\dagger}. \end{aligned} \tag{3.41}$$

Note, that Eq. (3.38) can not be formulated to Eqs. (3.12), (3.13) and (3.15) as in the original subtractive beamformer, since noise signals $n_1(t)$ and $n_i(t)$ on the two microphones are not directly related and no priori assumption between them is made here. Moveover, in the original GSC beamformers [62], the BM part was implemented by the difference between the observed signals on adjacent sensors, given by:

$$\mathbf{B}_2^{\dagger} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix}, \tag{3.42}$$

which indicates that only limited spatial information was used. Comparatively, the proposed algorithm considers the spatial information not only between adjacent sensors but also other sensor pairs, shown in Eqs. (3.38) and (3.41).

3. *Noise canceller.* The noise canceller output $Y_{NC}(k,\ell)$, which is an estimate of the noise in speech reference signal $Y_{FBF}(k,\ell)$, is constructed by filtering the BM outputs $\mathbf{U}(k,\ell)$ with the filters $\mathbf{H}(k,\ell)$, given by:

$$Y_{NC}(k,\ell) = \mathbf{H}^{\dagger}(k,\ell)\mathbf{U}(k,\ell), \tag{3.43}$$

where

$$\mathbf{H}^{T}(k,\ell) = \Big[H_2(k,\ell), H_3(k,\ell), \cdots, H_M(k,\ell)\Big]. \tag{3.44}$$

With the assumption of zero correlation between speech signal and noise signal, minimizing the mean square error between the speech reference signal $Y_{FBF}(k,\ell)$ and the NC output $Y_{NC}(k,\ell)$ and according to the Wiener theory, the optimal filters $\hat{\mathbf{H}}_{opt}(k,\ell)$ is given by [17, 34]:

$$\hat{\mathbf{H}}_{opt}(k,\ell) = \Phi_{\mathbf{UU}}^{-1}(k,\ell)\Phi_{\mathbf{U}Y}(k,\ell), \tag{3.45}$$

where $\Phi_{\mathbf{UU}}(k,\ell)$ is the *cross-spectral density matrix* of the BM output signals $\mathbf{U}(k,\ell)$, $\Phi_{\mathbf{U}Y}(k,\ell)$ the *cross-spectral density* between the BM output signals $\mathbf{U}(k,\ell)$ and the FBF output signal $Y_{FBF}(k,\ell)$, respectively. They are defined as:

$$\Phi_{\mathbf{UU}}(k,\ell) = E\Big[\mathbf{U}(k,\ell)\mathbf{U}^{\dagger}(k,\ell)\Big], \tag{3.46}$$

$$\Phi_{\mathbf{U}Y}(k,\ell) = E\Big[\mathbf{U}(k,\ell)Y_{FBF}^{*}(k,\ell)\Big], \tag{3.47}$$

After determining the three parts of the proposed algorithm, the output of this algorithm $Y_o(k,\ell)$ is calculated as the difference between the FBF output $Y_{FBF}(k,\ell)$ in the upper path and the NC output $Y_{NC}(k,\ell)$ in the lower path, that is:

$$Y_o(k,\ell) = \mathbf{W}^{\dagger}\mathbf{X}(k,\ell) - \mathbf{H}^{\dagger}(k,\ell)\mathbf{B}^{\dagger}(k,\ell)\mathbf{X}(k,\ell). \tag{3.48}$$

Note that the performance of the proposed algorithm should be only dependent on the characteristics of noise field since the optimal filters $\hat{\mathbf{H}}_{opt}(k,\ell)$ is only determined by the input noise signals under the assumption of zero correlation between desired speech signal and noise signal.

3.3.3 Theoretical analysis

In this subsection, we first define a measure used to show the theoretical noise reduction performance of the proposed algorithm. Then its performance is examined based on the coherence functions in theoretically defined noise fields.

Performance evaluation measure

To examine the performance of the proposed noise reduction algorithm, we define and use a measure which is referred to as *noise reduction performance* (NR). NR is defined as the ratio of PSD of system input $\phi^{(n)}_{XX}(k,\ell)$ and that of system output $\phi^{(n)}_{Y_oY_o}(k,\ell)$ when no desired speech signal is present, given by [54, 108]:

$$\mathrm{NR}(k,\ell) \;=\; \frac{\phi^{(n)}_{XX}(k,\ell)}{\phi^{(n)}_{Y_oY_o}(k,\ell)}, \tag{3.49}$$

where $\phi^{(n)}_{XX}(k,\ell) = E\Big[X(k,\ell)X^*(k,\ell)\Big|N(k,\ell)\Big]$ and $\phi^{(n)}_{Y_oY_o}(k,\ell) = E\Big[Y_o(k,\ell)Y_o^*(k,\ell)\Big|N(k,\ell)\Big]$.

Under the assumptions: (1) desired speech and noise are uncorrelated, (2) PSD of noise on each microphone is identical, we can rewrite $\hat{\mathbf{H}}_{opt}(k,\ell)$ and NR as (see **Appendix A** for detail):

$$\hat{\mathbf{H}}_{opt}(k,\ell) = \Big(\mathbf{B}^{\dagger}(k,\ell)\mathbf{\Gamma}(k,\ell)\mathbf{B}(k,\ell)\Big)^{-1}\mathbf{B}^{\dagger}(k,\ell)\mathbf{\Gamma}(k,\ell)\mathbf{W}, \tag{3.50}$$

and

$$\mathrm{NR}(k,\ell) = \Big[\mathbf{W}^{\dagger}\mathbf{\Gamma}(k,\ell)\mathbf{W} - \mathbf{W}^{\dagger}\mathbf{\Gamma}(k,\ell)\mathbf{B}_1\Big(\mathbf{B}_1^{\dagger}\mathbf{\Gamma}(k,\ell)\mathbf{B}_1\Big)^{-1}\mathbf{B}_1^{\dagger}\mathbf{\Gamma}(k,\ell)\mathbf{W}\Big]^{-1}, \tag{3.51}$$

where $\mathbf{\Gamma}(k,\ell)$ is the coherence function matrix of noise signals on all microphones, given by:

$$\mathbf{\Gamma}(k,\ell) = \begin{bmatrix} 1 & \Gamma_{N_1N_2}(k,\ell) & \cdots & \Gamma_{N_1N_M}(k,\ell) \\ \Gamma_{N_2N_1}(k,\ell) & 1 & \cdots & \Gamma_{N_2N_M}(k,\ell) \\ \vdots & \ddots & \ddots & \vdots \\ \Gamma_{N_MN_1}(k,\ell) & \Gamma_{N_MN_2}(k,\ell) & \cdots & 1 \end{bmatrix}, \tag{3.52}$$

and $\Gamma_{N_\mu N_\nu}(k, \ell)$ is the complex coherence function between the noises $N_\mu(k, \ell)$ and $N_\nu(k, \ell)$, defined as:

$$\Gamma_{N_\mu N_\nu}(k, \ell) = \frac{\phi_{N_\mu N_\nu}(k, \ell)}{\sqrt{\phi_{N_\mu N_\mu}(k, \ell)\phi_{N_\nu N_\nu}(k, \ell)}}. \tag{3.53}$$

Note, as Eqs. (3.50) and (3.51) show, that the optimal NC filters and the noise reduction performance are only determined by the coherence function matrix $\mathbf{\Gamma}(k, \ell)$ of noise signals, corresponding to the characteristics of noise fields.

Theoretical performance analysis

In the following, we examine the performance of the generalized subtractive beamformer in theoretically-defined noise fields.

1. *Coherent noise field.* In a coherent noise field, e.g., a point sound source in the far field of the microphone array, the coherence function $\Gamma_{N_\mu N_\nu}(k)$ is given by [24, 108]:

   $$\Gamma_{N_\mu N_\nu}(k) = e^{-2jk\pi\delta_{\mu\nu}}, \tag{3.54}$$

   where $\delta_{\mu\nu}$ denotes the time delay between μ-th and ν-th microphones. To find out the relationship between this generalized subtractive beamformer and the original subtractive beamformer [1, 2, 3], let us assume that only two microphones are available and the time delay is δ between them, the optimal solution for the NC filter can be derived as (see **Appendix B**):

   $$\hat{H}^*_{opt}(k) \quad = \quad \frac{1}{e^{jk\pi\delta}\sin\Big(2k\pi\tau\Big)\sin\Big(k\pi\delta\Big)}. \tag{3.55}$$

   Comparing this optimal NC filter in Eq. (3.55) with the "weight factor" in Eq. (3.13), we can note that they are exactly same, which indicates the following:

   (a) The proposed noise reduction algorithm reduces to the previously presented original algorithm in a perfectly coherent noise field;

   (b) The original algorithm is also an optimal solution in *minimum mean square error* (MMSE) sense for reducing coherent noise.

   Putting Eq. (3.54) into (3.51), we can see that the noise reduction performance of this proposed algorithm reaches infinity at all frequencies in a coherent noise field.

2. *Incoherent noise field.* In an incoherent noise field, e.g., the sensor self-noise, the coherence function is zero for all frequencies, $\Gamma_{n_\mu n_\nu}(k) = 0, \forall k$. In this noise field, the noise reduction performance amounts to M, the number of microphones.

3. *Diffuse noise field.* A diffuse noise field has been shown to be a reasonable model for many practical noise environments, such as reverberant rooms and car environments [8, 87]. A diffuse noise field is characterized by the following coherence function [8, 24, 87]:

$$\Gamma(k) = \frac{\sin{(2k\pi d/c)}}{2k\pi d/c}, \tag{3.56}$$

where d and c represent the inter-element spacing and the velocity of sound. Putting Eq. (3.56) into Eq. (3.51), we can find that noise reduction performance depends on the inter-element spacing d and the number of microphones M. Figs. 3.9 and 3.10 plot the noise reduction performance as a function of the frequency for different inter-element spacings d and the different numbers of microphones M. Figs. 3.9 and 3.10 show that the proposed algorithm achieves high noise reduction performance in the moderate and high frequencies, with relatively low ability in the very low frequencies (especially, when the inter-element spacing d is small).

Moreover, with the assumption of identical noise PSD on each microphone, we can derive the same noise reduction performance for our proposed algorithm and the original GSC beamformer [8, 62], as shown in Figs. 3.9 and 3.10. However, in practical environments, the noise PSDs on different microphones are generally not equivalent, which results in that the noise reduction can not be represented as a function of noise coherence function any more, that is, the failure of Eq. (3.51). In this case, the performance of two algorithms is examined by experiments using multi-channel recordings in the following.

3.3.4 Experimental validation

The performance of the proposed noise reduction algorithm based on a generalized subtractive beamformer (PRO-GSBF) was evaluated using multi-channel recordings and its performance was further compared to that of other traditional algorithms: delay-and-sum beamformer (DSBF), the original subtractive beamformer based algorithm (ORG-SBF) [3] and the original GSC beamformer (ORG-GSC) [62, 8, 9], in terms of both objective and subjective evaluation measures.

The proposed algorithm and other traditional algorithms were performed using the *overlap-and-add* (OLA) technique. The window length includes 42.6 ms (512 samples) with an overlap of 21.3ms (256 samples). In our implementations, for the PRO-GSBF and the ORG-GSC [8], the estimated noise component (the output of the NC filter) was subtracted from the output of the upper path in spectral magnitude domain, not in complex spectral domain considered in the theoretical analysis. This is same as the ORG-SBF [1, 2, 3] and different from the ORG-GSC detailed in [8]. We performed this

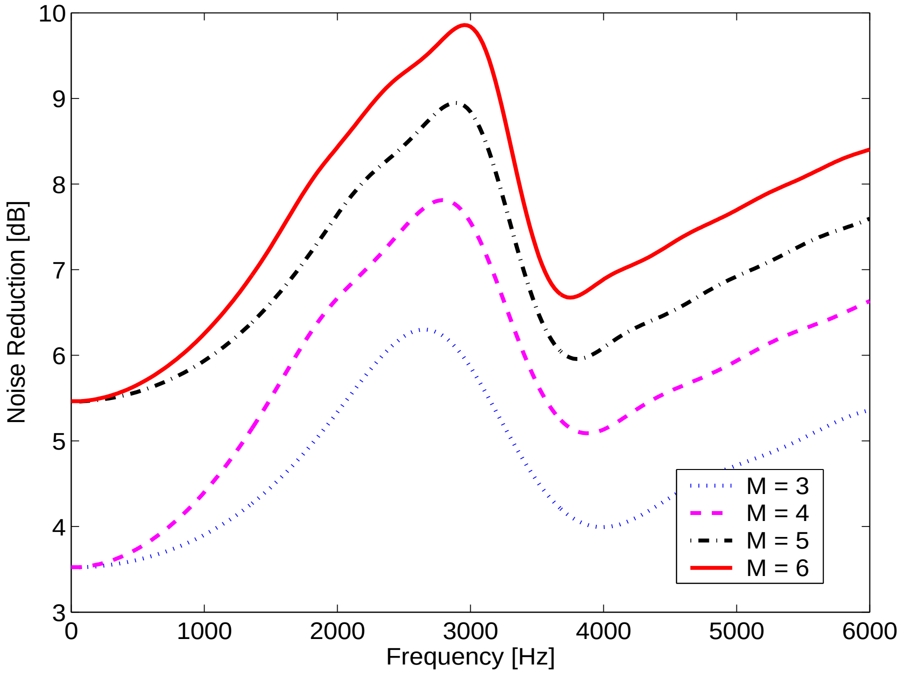


Figure 3.9: Noise reduction performance in a diffuse noise field for different numbers of microphones ($d = 10cm$).

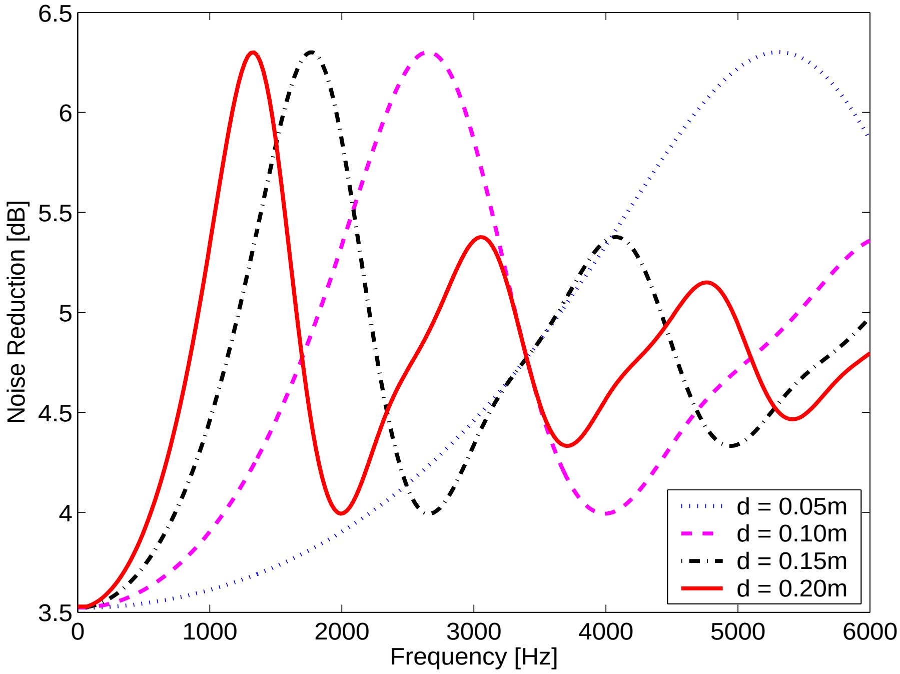


Figure 3.10: Noise reduction performance in a diffuse noise field for different distances between adjacent microphones ($M = 3$).

implementation because of the following considerations: (1) the relative unimportance of phase for speech quality in speech enhancement applications [156]; (2) the amplitude spectra is only important for speech recognition systems [127].

To assess the performance of the studied noise reduction algorithms, an equally-spaced linear array consisting of three microphones with inter-element spacing of 10cm was mounted on the roof near driver's sun-visor in a car. The array was about 50cm away from and directly in front of the driver. Multi-channel noise recordings were performed across all channels when the car was running in two conditions: (1) at the speed of 50km/h without air-condition noise (the air condition is off), (2) at the speed of 100km/h with high-level air-condition noise (the air condition is on). With different types of clean speech signals, we generate the following two different sets of noisy signals to show the performance of the proposed algorithm in various noise conditions.

1. **Pseudo real-world environment**. In this condition, the clean speech signals, consisting of 50 Japanese sentences, were taken from an ATR database [149].

2. **Real-world environment**. In this condition, multi-channel speech recordings were performed across all channels when the car is still/stopped. The speech signals, consisting of 100 Japanese city names, were uttered by two speakers (one male and one female) at the driver's position.

Both speech and noise signals were first re-sampled to 12kHz at 16 bit accuracy. We generated the multi-channel noisy signals by artificially mixing multi-channel real-world speech recordings and multi-channel car noise recordings, and pseudo real-world speech recordings and multi-channel car noise recordings at different global SNR levels [-5, 15] dB. (The calculation of global SNR is detailed in [126].)

Objective evaluation measures

The objective measures used in our performance evluation include *Segmental SNR* (SEGSNR) and *Mel-Frequency Cepstral Coefficient* (MFCC) distance.

Segmental SNR (SEGSNR) is a widely used objective evaluation criterion for speech enhancement or noise reduction algorithms since it is more correlated to subjective results [126]. SEGSNR is defined as the ratio of the power of "ideal" clean speech to that of the noise signal embedded in a noisy signal or in an enhanced speech signal by tested algorithms over all frames, given by:

$$\mathrm{SEGSNR} = \frac{1}{|\Psi|} \sum_{\ell \in \Psi} 10 \log_{10} \left(\frac{\sum_{l=0}^{L-1} [s(\ell L + l)]^2}{\sum_{l=0}^{L-1} [\hat{s}(\ell L + l) - s(\ell L + l)]^2} \right), \quad [\mathrm{dB}] \tag{3.57}$$

where $s(.)$ is the reference speech signal, and $\hat{s}(.)$ represents the noisy signal or enhanced signals processed by the tested algorithms; Ψ denotes the set of frames in the signal and $|\Psi|$ its cardinality; L represents the number of samples per frame (equal to the length of STFT). Note that a higher SEGSNR means higher speech quality of the enhanced signal.

A second evaluation measure, MFCC distance, is defined as the distance between MFCCs of a clean speech signal and those of a noisy signal or enhanced signal, which can be represented as:

$$d_{\mathrm{mfcc}} \quad = \quad \frac{1}{|\Phi|}\sum_{\ell\in\Phi}\sum_{i}\left(mfcc_i - mfcc_i^{'}\right)^2, \tag{3.58}$$

where Φ represents the set of frames in which speech is present and $|\Phi|$ its cardinality; $mfcc_i$ is the 12-order MFCCs of the clean speech signal, and $mfcc_i^{'}$ denotes those of the noisy signal or enhanced speech signals, respectively. Note that a lower MFCC distance level indicates lower speech distortion, corresponding to higher speech quality.

Objective evaluation results

Experimental results of SEGSNR in *pseudo real-world environment* and *real-world environment*, averaged across all sentences under two conditions (50km/h and 100km/h), are respectively plotted in Figs. 3.11 and 3.12. The results demonstrate that the DSBF provides a very limited SEGSNR improvement since only three microphones were used in our experiments, the ORG-SBF does not show sufficient performance improvement due to its unpractical assumption of a coherent noise field. The ORG-GSC beamformer shows higher SEGSNR improvements compared with the DSBF and the ORG-SBF. Furthermore, the PRO-GSBF offers the highest SEGSNR improvements, corresponding to highest speech quality, among the studied algorithms under all test conditions. Moreover, note that the same observations can be obtained in both *pseudo real-world environment* and *real-world environment.*

Experimental results of MFCC distance for *pseudo real-world environment* and *real-world environment* in two noise conditions (50km/h and 100km/h) at various SNRs are plotted in Figs. 3.13 and 3.14. Compared to the noisy inputs, the DSBF and the ORG-SBF algorithms decrease the MFCC distances in all conditions, especially at low SNRs. The ORG-GSC beamformer shows a further decrease in all noise conditions. And the PRO-GSBF method offers the lowest MFCC distance, corresponding to the lowest speech distortion, compared to other algorithms under all conditions. Moreover, note again that the same observations can be obtained in both *pseudo real-world environment* and *real-world environment.*

Subjective evaluation results

Subjective evaluations of the studied algorithms were performed using speech spectrograms. Typical examples of speech spectrograms, in *pseudo real-world environment* corresponding to the Japanese sentence “dozo yoroshiku” and in *real-world environment* corresponding to the Japanese sentence “hatinohe kesennuma yukuhasi”, are plotted in Figs. 3.15 and 3.16, in the car condition at the speed of 100km/h. As Fig. 3.15 (c) shows, the output of the DSBF is characterized by high-level noise since only a small number (3-channel) of microphones was used. The ORG-GSC does not offer sufficient suppression ability for the low-frequency noise, as shown in Fig. 3.15 (d). And as plotted in Fig. 3.15 (e), the ORG-SBF algorithm still shows very limited performance improvement, especially in the low frequency region. Comparatively, Fig. 3.15 (f) demonstrates that the PRO-GSBF algorithm provides a much higher performance improvement, especially in low-frequency region, compared to the other studied algorithms. With regard to other traditional algorithms, the superiority of the proposed PRO-GSBF can also be observed from the spectrograms in *real-world environment* as shown in Fig. 3.16.

3.3.5 Discussions

Based on the experimental results presented in the last subsection, the superiorities of the proposed generalized method with regard to the other algorithms are discussed in the following.

The proposed PRO-GSBF outperforms the DSBF. For the DSBF, many microphones are needed to obtain an acceptable performance. For the proposed method, a small number of microphones are sufficient to achieve the same noise reduction performance.

The proposed PRO-GSBF outperforms the ORG-SBF algorithm. The basic assumption of the ORG-SBF algorithm, a perfectly coherent noise field, is seldom satisfied in real-world environments. While, no priori assumption on noise signals is made in the PRO-GSBF method. That is, the PRO-GSBF algorithm is a natural extension of the ORG-SBF by relaxing the unpractical assumption of an coherent noise field to the one of an arbitrary noise field. Therefore, the improved noise reduction performance can be achieved for the PRO-GSBF algorithm. Moreover, theoretically, the high performance in reducing unstable noise (sudden noise) is expected for the PRO-GSC beamformer because the PRO-GSC is derived based on same ideas as those of the ORG-SBF which has the ability in reducing sudden noise.

The proposed PRO-GSBF outperforms the ORG-GSC algorithm. In theory, with the assumption of identical noise PSD on each microphone, both PRO-GSBF and ORG-GSC show the same noise reduction performance. In practice, the noise PSDs on different microphones are normally different. The PRO-GSBF provides the improved noise reduction performance, especially in reducing low-frequency noise, due to the fact that different inter-element spacings (more spatial information) are used, shown by Eq. (3.41). While, the ORG-GSC beamformer achieves limited performance due to the use of limited spatial information, shown by Eq. (3.42). However, if the desired speech signals received by different microphones are greatly different, both PRO-GSBF and ORG-GSC will introduce some speech distortion, especially for PRO-GSBF which is also because of the use of the sensor pairs with larger spacing.

Therefore, the proposed algorithm provided the highest noise reduction performance with less speech distortion among the studied algorithms in various car environments, both *pseudo real-world environment* and *real-world environment*, at various SNR levels.

3.4 Remarks on two subtractive beamformers

In the last two sections, we first introduced the original subtractive beamformer on which we then presented a generalized subtractive beamformer by relaxing the strict assumption of a perfectly coherent noise field to the one of an arbitrary noise field. Moreover, we proved that the generalized subtractive beamformer reduces to the original subtractive beamformer in a perfectly coherent noise field when only two microphones are available. That is, the generalized subtractive beamformer is a natural extension of the original subtractive beamformer in an arbitrary noise field with arbitrary number of microphones.

Using the multi-channel car noise recordings and different types of speech signals, we conducted some comprehensive experiments. And experimental results confirmed that the generalized subtractive beamformer is superior to the original subtractive beamformer in noise reduction and speech enhancement in car environments. In fact, this conclusion is not surprising since the noise field in car environments is approximately able to be modelled as a diffuse noise, which does not satisfy the coherent noise field assumption of the original subtractive beamformer and in turn degrading the noise reduction performance of the original subtractive beamformer. The generalized subtractive beamformer adaptively estimates the coherence functions based on the input signals, therefore, yielding the improved noise reduction performance in car environments.

However, the original subtractive beamformer should be preferred to the generalized subtractive beamformer in a perfectly coherent noise field. As the theoretical performance analysis results demonstrate, the generalized subtractive beamformer provides infinity noise reduction performance only in a coherent noise field. In this sense, the superiority of the generalized subtractive beamformer is highlighted in a coherent noise field, not other noise fields. Whereas, in a coherent noise field, two subtractive beamformers should be exactly same theoretically. Furthermore, since the DOAs of the coherent noise signals can be determined frame by frame, the original subtractive beamformer is able to reduce the highly non-stationary noise signals (e.g., sudden noise). In theory, the generalized subtractive beamforme can deal with various kinds of noise signals if noise coherence functions are known or can be calculated accurately. In practice, however, in the generalized subtractive beamformer, the noise coherence functions are generally calculated in an iterative way which make the generalized subtractive beamformer difficult to reduce the highly non-stationary noise signals in practice.

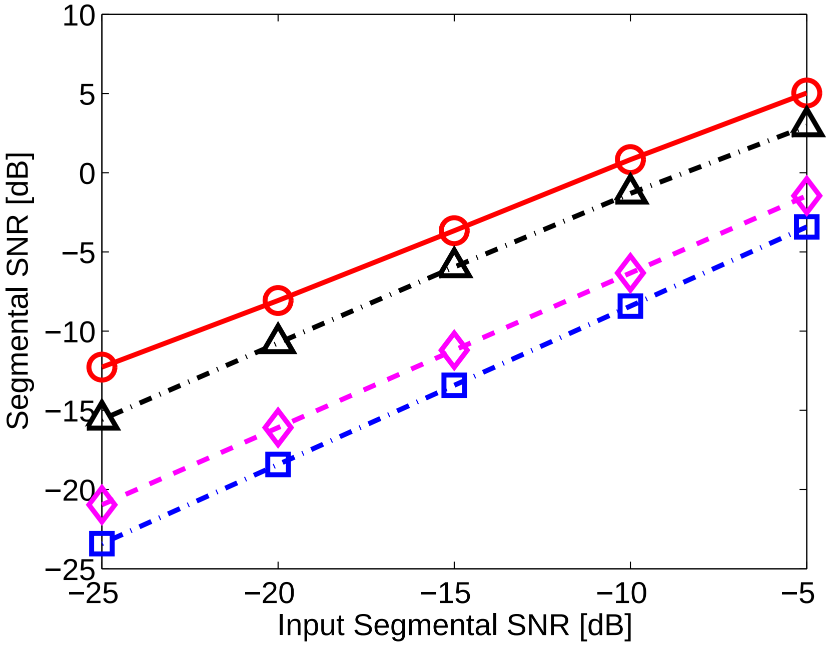
3.5 Summary

In this chapter, we first described the characteristics of the speech signal and the noise field in practical environments. Further, we considered that the noise components in a real-world environments can divided into two types: localized noise signals which have some determinable directions and generally come from some point noise sources; and non-localized noise signals which are from all directions.

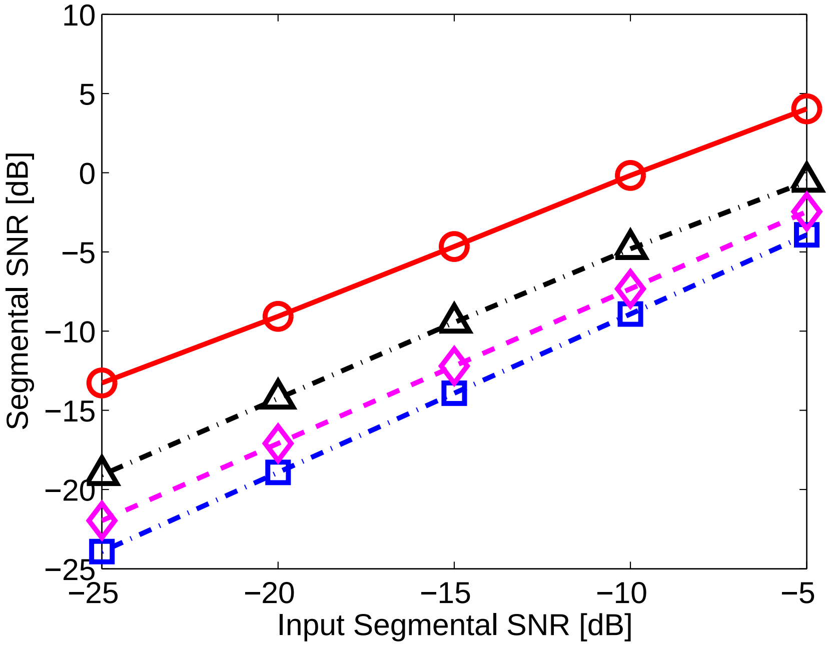
In section 3.2, we discussed the problem of deal with localized noise using a subtractive beamformer based on microphone array. The basic concept of this localized noise suppression algorithm is that the spectra of localized noise are first estimated using a hybrid noise estimation technique and then subtracted from that of the observed noisy signals using non-linear spectral subtraction. The hybrid noise estimation technique combines a single-channel estimation approach and a multi-channel estimation approach. This combination is effectively performed by a *robust and accurate speech absence probability* (RA-SAP) estimator which considers the strong correlation of speech presence between adjacent frequency bins and consecutive frames. With the RA-SAP estimator, this hybrid noise estimation technique provides much more accurate spectral estimate of localized noises, which is then subtracted form those of the observed signals on each microphone to enhance the speech components.

In section 3.3, we extend the original subtractive beamformer we consider before to a generalized expression by reformulating the original subtractive beamformer with an assumption of an arbitrary noise field. The generalized subtractive beamformer has a GSC-like structure, consisting of a *fixed beamformer*, a *blocking matrix* and a *noise canceller*. Theoretical noise reduction performance was also presented, from which we find that the original subtractive beamformer is a special case of this generalized subtractive in a perfectly coherent noise field when only two microphones are available. Moreover, the superiority of this generalized subtractive beamformer was also validated with experimental results in both *pseudo real-world environment* and *real-world environment*.

In section 3.4, we presented some remarks on both the original subtractive beamformer and the generalized subtractive beamformer. we pointed out that in perfectly coherent noise fields, the original subtractive beamformer is superior to the generalized subtractive beamformer to reduce coherent noise field (e.g., sudden noise), while in other noise fields, the generalized subtractive beamformer should be preferred to reduce various kinds of noise components (e.g., diffuse noise).

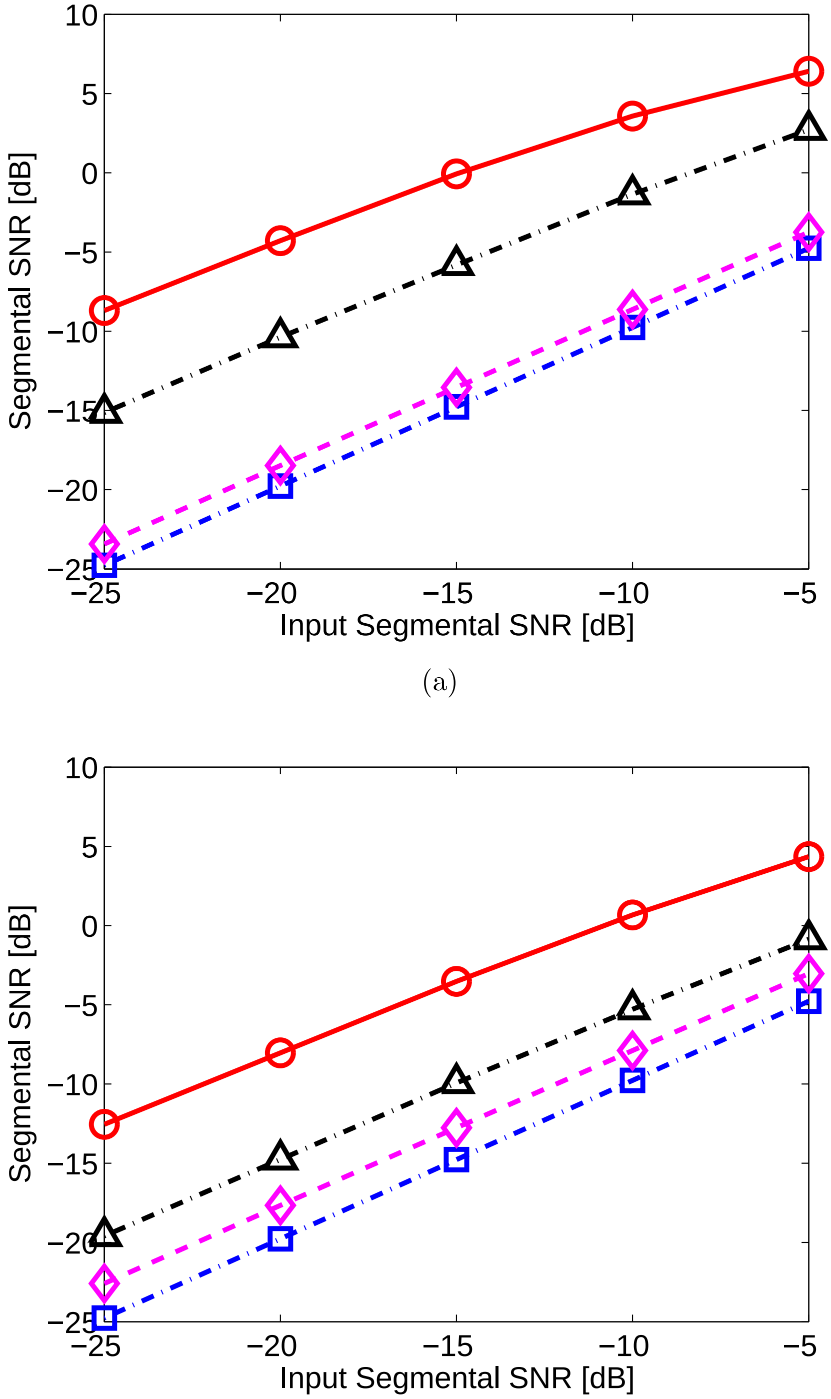


(a)



(b)

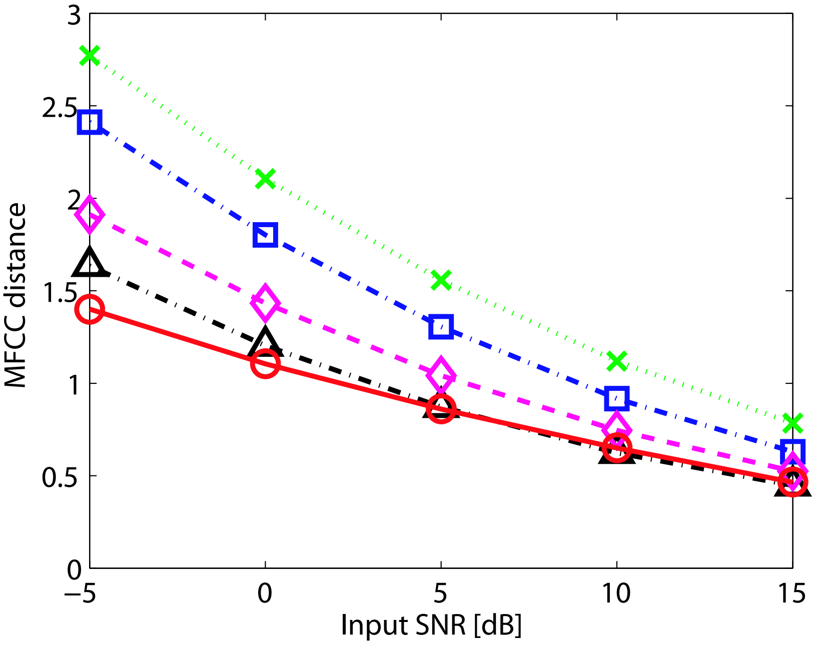
Figure 3.11: Average segmental SNR (SEGSNR) in **pseudo real-world environment** at delay-and-sum beamformer (DSBF) output (□), original GSC beamformer (ORG-GSC) output (△), original subtractive beamformer based (ORG-SBF) algorithm output (♢) and proposed generalized subtractive beamformer based (PRO-GSBF) algorithm output(○), in various noise conditions: speeds of 50km/h (a) and 100km/h (b).



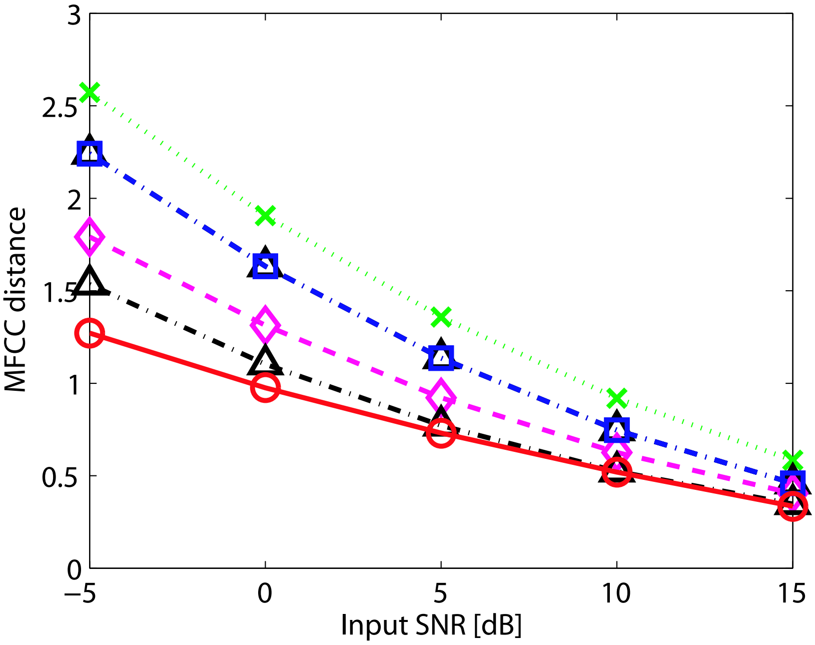
(a)

(b)

Figure 3.12: Average segmental SNR (SEGSNR) in **real-world environment** at delay-and-sum beamformer (DSBF) output (□), original GSC beamformer (ORG-GSC) output (△), original subtractive beamformer based (ORG-SBF) algorithm output (◊) and proposed generalized subtractive beamformer based (PRO-GSBF) algorithm output(○), in various noise conditions: speeds of 50km/h (a) and 100km/h (b).

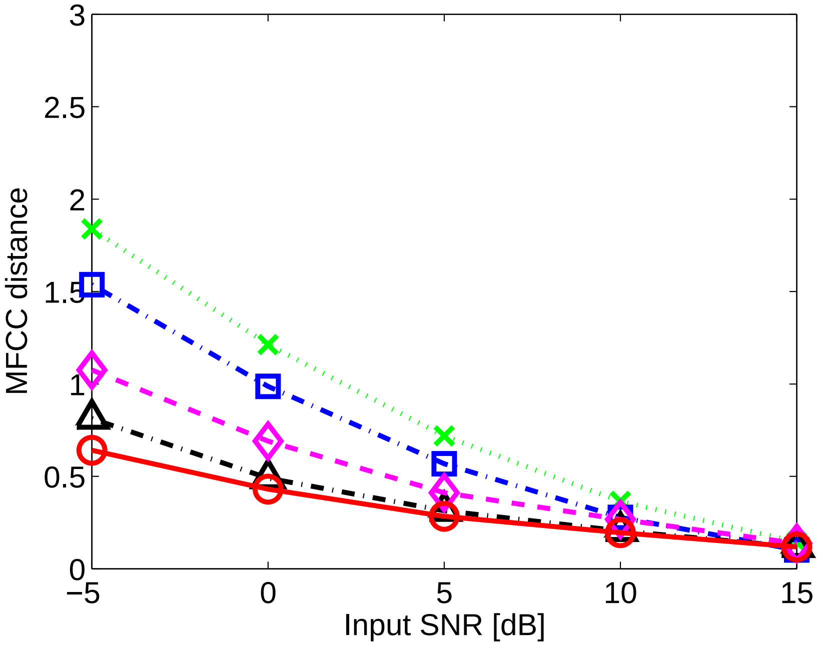


(a)

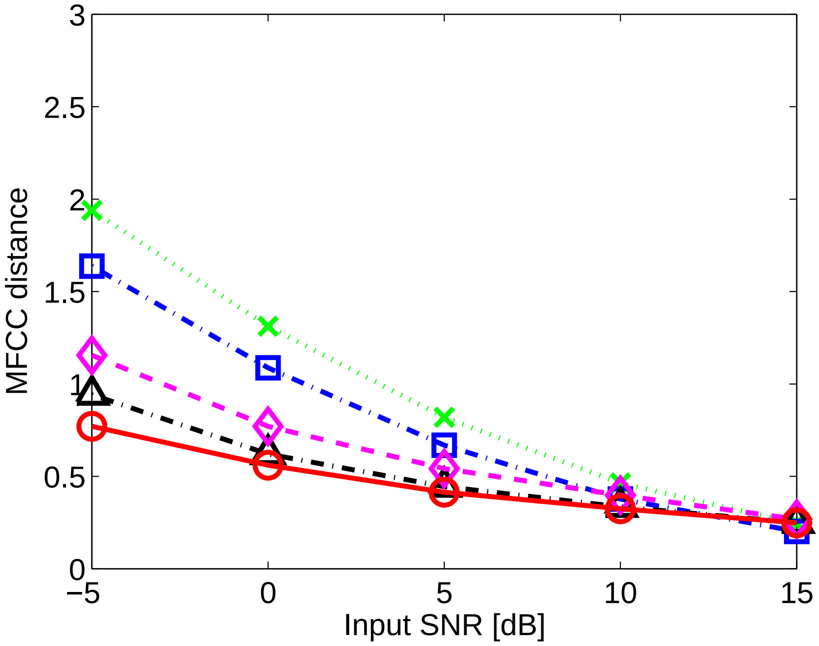


(b)

Figure 3.13: Average MFCC distance in **pseudo real-world environment** at the first microphone (×), delay-and-sum beamformer (DSBF) output (□), original GSC beamformer (ORG-GSC) output (△), original subtractive beamformer based (ORG-SBF) algorithm output (◊) and proposed generalized subtractive beamformer based (PRO-GSBF) algorithm output(○), in various noise conditions: speeds of 50km/h (a) and 100km/h (b).

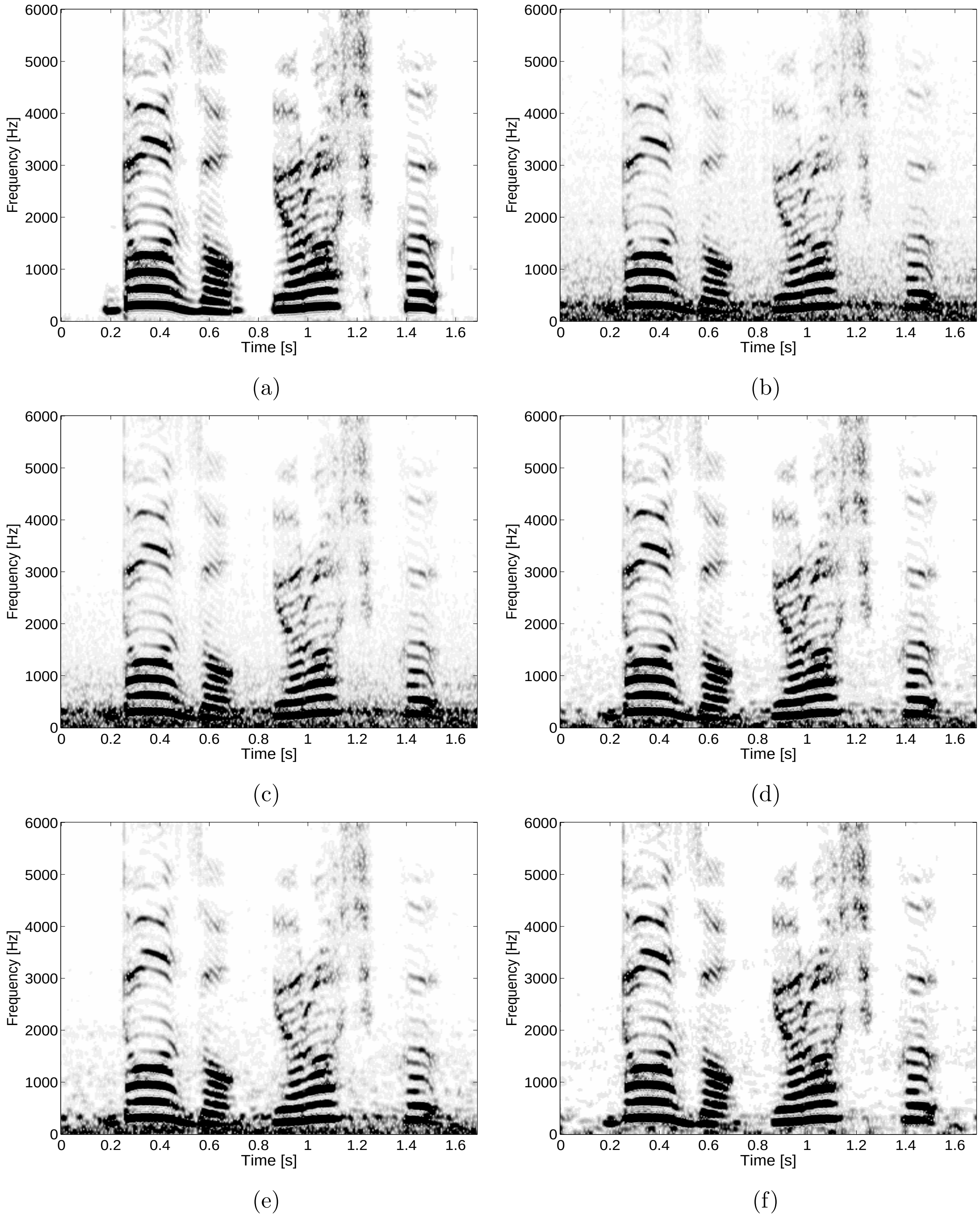


(a)



(b)

Figure 3.14: Average MFCC distance in **real-world environment** at the first microphone (×), delay-and-sum beamformer (DSBF) output (□), original GSC beamformer (ORG-GSC) output (△), original subtractive beamformer based (ORG-SBF) algorithm output (♢) and proposed generalized subtractive beamformer based (PRO-GSBF) algorithm output(∘), in various noise conditions: speeds of 50km/h (a) and 100km/h (b).



(a)

(b)

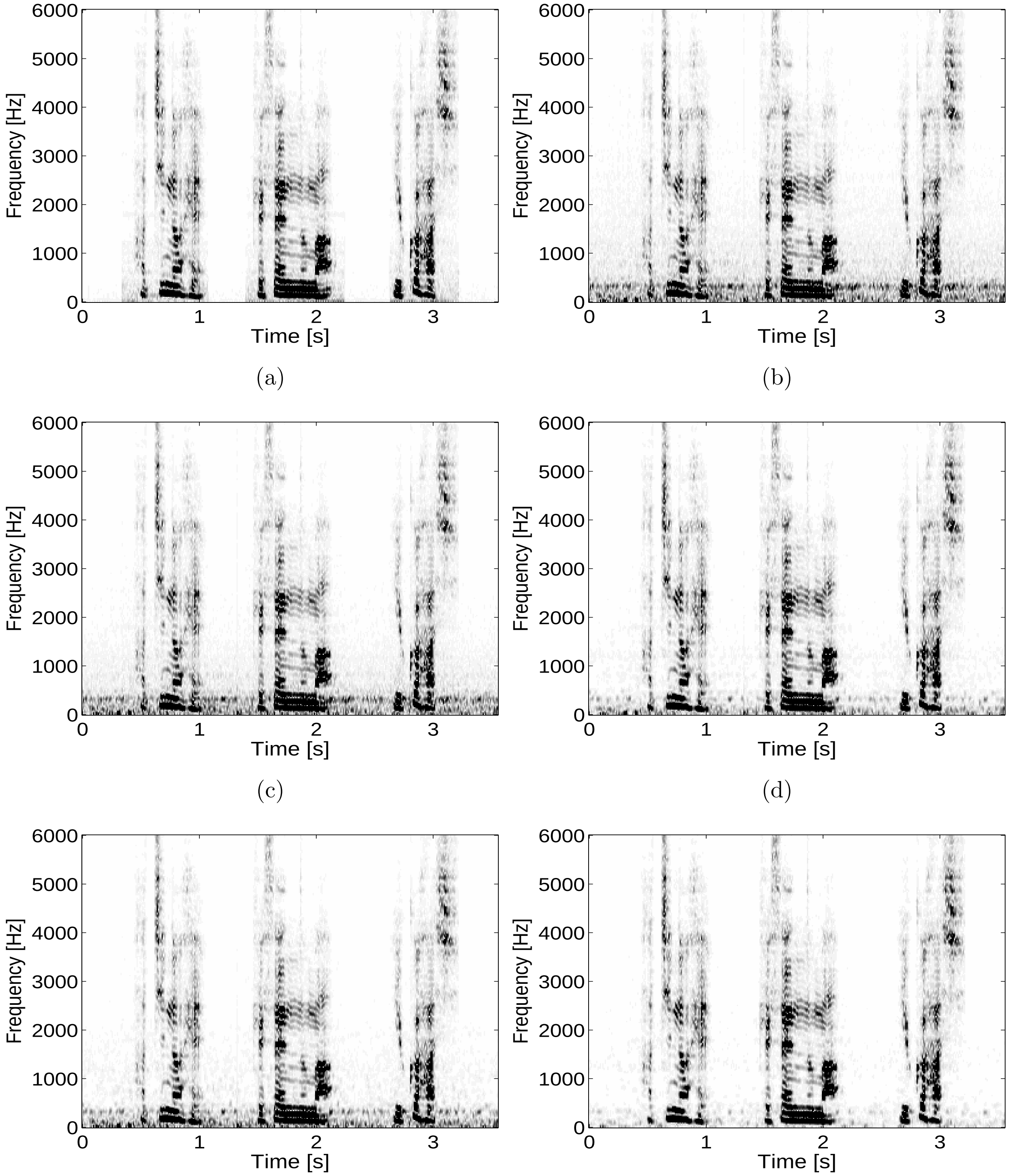
(c)

(d)

(e)

(f)

Figure 3.15: Speech spectrograms in **pseudo real-world environment**. (a) original clean speech signal at the first microphone: “dozo yoroshiku”; (b) noisy signal at the first microphone (SNR = 10 dB); (c) delay-and-sum beamformer (DSBF) output; (d) original GSC beamformer (ORG-GSC) output; (e) original subtractive beamformer based (ORG-SBF) algorithm output; (f) proposed generalized subtractive beamformer based (PRO-GSBF) algorithm output.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.16: Speech spectrograms in **real-world environment**. (a) original clean speech signal at the first microphone: “hatinohe kesennuma yukuhasi”; (b) noisy signal at the first microphone (SNR = 10 dB); (c) delay-and-sum beamformer (DSBF) output; (d) original GSC beamformer (ORG-GSC) output; (e) original subtractive beamformer based (ORG-SBF) algorithm output; (f) proposed generalized subtractive beamformer based (PRO-GSBF) algorithm output.

Chapter 4

Non-localized noise suppression with post-filtering

In real-world environments, undesired noise signals generally originate from various kinds of sound sources (e.g., radio, road and wind). In this research, we divide the undesired noise signals into two categories: localized noise signals with certain directions, and non-localized noise signals with undeterminable directions and modelled as diffuse noises. Previously, we have presented beamforming based techniques which successfully suppress the localized noises with a microphone array. In this chapter, we deal with the problem of suppressing non-localized noise components with post-filtering.

With the use of beamforming techniques exploiting microphone array detailed in section 3.2, localized noise components are first successfully suppressed. At the beamformer output, however, the remaining noise components (especially non-localized noises) are still considerable, which should be further reduced. Therefore, an additional post-filter is generally needed to further improve the noise reduction performance of the multi-channel beamformering techniques, as explained in section 2.3.

In this chapter, we first show a noise field analysis measure, with which the noise field in car environments is examined and further proven to be approximately modelled as a diffuse noise field which actually provides a reasonable model for many practical environments. Therefore, non-localized noises are assumed to be diffuse noise in this research. Considering the spatial characteristics of a diffuse noise field, we present a hybrid post-filter to suppress correlated as well as uncorrelated noise. In the proposed post-filter, a modified Zelinski post-filter, which fully considers and utilizes the correlation characteristics of noise on different microphone pairs, is applied to the high frequencies to suppress spatially uncorrelated noise; a single-channel Wiener post-filter is applied to the low frequencies for cancellation of spatially correlated noise. Experimental results using multi-channel recordings were conducted and experimental results demonstrate the usefulness and superiority of the proposed post-filter with regard to other comparative post-filters in various car environments.

4.1 Introduction

Multi-channel beamforming based algorithms provide high noise reduction performance especially for localized noise, however, only limited noise reduction performance is achieved in a diffuse noise field, as analyzed in the chapter 3. To further suppress non-localized noise (modelled as diffuse noise) at the beamformer output, post-filtering is normally needed to improve the noise reduction performance of microphone arrays in practical environments, as explained in sections 2.3 and 2.4.

A variety of post-filtering techniques have been presented in the literature [9, 17, 34, 55, 108, 113, 163]. One commonly used multi-channel post-filter, which is based on Wiener filter, was first introduced by Zelinski [163]. In this method, the output of a delay-and-sum beamformer is further post-filtered using an adaptive Wiener filter, based on the auto- and cross- spectral densities of the sensor signals. The basic assumption behind this post-filter is that noises on different microphones are mutually uncorrelated, corresponding to a perfectly incoherent noise field. This assumption is, however, seldom satisfied in practical environments, especially for closely-spaced microphones and low frequencies, which are characterized by the high-correlated noise.

To suppress the high-correlated noise, Fischer *et al.* [49] proposed a noise reduction system, which is based on the GSC beamformer. The GSC beamformer reasonably suppresses the spatially coherent noise components, whereas a Wiener filter in the look direction is designed to suppress the the spatially incoherent noise components. However, Bitzer *et al.* pointed out that neither the GSC nor the standard Wiener post-filter performs well at low frequencies in a diffuse noise field [7]. Therefore, they proposed to add a second post-filter at the output of a GSC beamformer with standard Wiener post-filter to reduce the spatially correlated noise components [9]. An alternative solution, presented by Meyer *et al.*, applies the spectral subtraction to suppress the high-correlated noise components in the low frequency region [108]. However, this method introduces the artificial "musical noise" caused by spectral subtraction and fails to deal with non-stationary noise due to the VAD based noise estimation technique. Moreover, a VAD does frequently fail, especially in high noise scenarios. Moreover, Bouquin-Jeannes *et al.* suggested the modification of the cross power spectrum estimation and the Wiener post-filter to take the presence of some correlated noise components into account [16]. The cross power spectrum of the noise signals is first averaged during speech pauses and then subtracted from the cross power spectrum of the sensor signals which is calculated during signal presence. Mamhoudi *et al.* [98, 99] considered a nonlinear coherence filtering in the wavelet domain to improve the performance of the Wiener post-filtering. Instead of the conventional coherence between the individual sensor signals, they used the coherence between the output and the input of the beamformer, which is assumed to be low, even for correlated noise components. Fischer and Kameyer [50] suggested the application of Wiener filter to the output of a broadband beamformer, which is built up by several harmonically nested subarrays. They showed that the resulting noise reduction performance is nearly independent of the correlation properties of the noise field. This structure has been further analyzed by Marro *et al.* [103]. Recently, McCowan *et al.* developed a general expression of the Zelinski post-filter based on the *a priori* coherence function of the noise filed [113]. Although this post-filter was shown to achieve improved speech quality and speech recognition accuracy compared to the Zelinski post-filter using the office room recordings, its performance is expected to be significantly degraded when difference between the "actual" and assumed coherence function exists [113].

Moreover, a single-channel noise suppression algorithm, referred to as *optimally-modified log-spectral amplitude* (OM-LSA) estimator, was presented for minimizing the log-spectral amplitude distortion in non-stationary noise environments [28]. This OM-LSA estimator was also extended to a multi-channel post-filtering approach when multi-channel inputs are available, which was shown effective in reducing highly non-stationary noise components from the desired source components based on the energy-based speech presence probability estimator [34, 55]. Considering the spatially stable characteristics of noise fields, a speech presence probability estimator based on these spatial characteristics was presented to improve the performance of the OM-LSA post-filter [87, 88]. However, the inherent sensitive implementation parameters involved in the variants of the OM-LSA post-filter greatly degrade their performance in practical environments.

Moreover, it has been shown that a diffuse noise field provides a reasonable model for a large number of practical noise environments, such as reverberant rooms and car environments [17, 113, 108]. Among the post-filters, no existing post-filters in theory is based on Wiener filter, and in practice can deal with diffuse noise which is characterized by the low coherence in high frequencies and the high coherence in low frequencies with low speech distortion.

In this chapter, we propose a novel post-filter with a hybrid structure under the assumption of a diffuse noise field. Considering the characteristics of a diffuse noise field, the proposed post-filter applies a multi-channel Wiener post-filter for the high-frequency (low-correlated) noise and a single-channel Wiener post-filter for the low-frequency (high-correlated) noise. In the high frequencies, a modified Zelinski post-filter, which fully considers and utilizes the correlations between noise on different microphone pairs, is presented and used. In the low frequencies, a single-channel Wiener post-filter is adopted which produces less "musical noise" due to the use of the decision-directed SNR estimation mechanism. The merits of the proposed post-filter lie in: in theory, it is a Wiener filter; in practice, it is highly capable of reducing low-correlated as well as high-correlated noise in a diffuse noise field. The superiorities of the proposed post-filter were verified using the multi-channel recordings in various car environments.

4.2 Problem formulation

In a noisy environment, let consider a M-sensor microphone array. The observed signal on each microphone consists of the desired speech signal and additive noise signal which are decomposed into localized and non-localized noise components in this research. The localized noise components have been suppressed by the beamforming based algorithm with the use of a microphone array, as described in the chapter 3. The proposed post-filter is applied to the beamformer outputs to further suppress non-localized noise components and enhance the speech quality. Therefore, after suppressing the localized noise components, the beamformer output signal $Z_m(k, \ell)$ on m-th channel consists of desired speech component $S(k, \ell)$ and beamformer-processed non-localized noise component $D_m(k, \ell)$, which can be represented in the time-frequency domain as:

$$Z_m(k, \ell) = S(k, \ell) + D_m(k, \ell). \quad m = 1, 2, \ldots, M \tag{4.1}$$

Note that the desired speech spectra $S(k, \ell)$ at the beamformer outputs should be identical to those at the beamformer inputs theoretically since the beamformer that we applied to suppress localized noise is a MVDR beamformer as proven in chapter 3. In practice, the speech spectrum $S(k, \ell)$ might be slightly different due to some practical factors, such as steering imperfection and estimation error. Moreover, note that the non-localized noise components at the beamformer output are different from those at the beamformer input since the noise reduction procedure of beamformer was also performed on non-localized noise components. For the notational simplicity, we use the notation $D_m(k, \ell)$ to denote the non-localized noise components at the beamformer output. Moreover, although only three microphones are used in the proposed noise reduction system, in this section, the number of microphones are assumed to be M (not only three) for the generalization.

4.3 Review of existing post-filters

In this section, we briefly review two post-filters, referred to as the Zelinski post-filter and the McCowan post-filter. Our proposed post-filter is based on the Zelinski post-filter and compared to both.

4.3.1 Zelinski post-filter

The Zelinski post-filter approaches a Wiener filter in a perfectly incoherent noise field based on the estimates of the auto- and cross- spectral densities. With the assumptions that the desired signal and noise signal are uncorrelated and that noise on different microphones is also uncorrelated and of identical power spectral density, the auto- and cross-spectral densities of the multi-channel beamformer outputs, $\phi_{z_\mu z_\nu}(k, \ell)$ and $\phi_{z_\mu z_\nu}(k, \ell)$, can be simplified as:

$$\begin{aligned}\phi_{z_\mu z_\nu}(k,\ell) &= \phi_{ss}(k,\ell)+\phi_{dd}(k,\ell), &(4.2)\\ \phi_{z_\mu z_\nu}(k,\ell) &= \phi_{ss}(k,\ell). &(4.3)\end{aligned}$$

where $\phi_{ss}(k,\ell)$ and $\phi_{dd}(k,\ell)$ are the PSDs of speech signal $S(k,\ell)$ and $D(k,\ell)$ at the multi-channel beamformer outputs, respectively.

Based on the simplified expressions of the auto- and cross- spectral densities, the Zelinski post-filter can be formulated as [163]:

$$G_z(k,\ell)=\frac{\frac{2}{M(M-1)}\sum_{\mu=1}^{M-1}\sum_{\nu=\mu+1}^{M}\Re\Big\{\phi_{z_\mu z_\nu}(k,\ell)\Big\}}{\frac{1}{M}\sum_{\mu=1}^{M}\phi_{z_\mu z_\mu}(k,\ell)}, \tag{4.4}$$

where $\Re\{.\}$ is the real operator. Note, to improve the robustness of this post-filter, the averaging operation is performed across all sensor pairs. Moreover, it is of interest to note that the auto- and cross- spectral densities are estimated from the multi-channel inputs (i.e., beamformer outputs). This estimation technique slightly over-estimates the noise spectral density [137]. However, it has been proven to give a high noise reduction performance and also is widely used [103, 113, 137, 163]. A further investigation of this post-filter was also made by Marro *et al.* [103].

4.3.2 McCowan post-filter

As a matter of fact, the basic assumption of the Zelinski post-filter, that noise on each microphone is uncorrelated, is seldom satisfied in practical environments. Considering this fact, McCowan relaxed this practically unreasonable assumption to the one that noise on each microphone is correlated through the coherence function and of identical power spectral densities [113].

With the assumption of zero correlation between the desired speech signal and noise signal and the relaxed assumption, the auto- and cross- spectral densities of multi-channel inputs, $\phi_{z_\mu z_\mu}(k,\ell)$ and $\phi_{z_\nu z_\nu}(k,\ell)$, can be simplified as:

$$\begin{aligned}\phi_{z_\mu z_\mu}(k,\ell) &= \phi_{ss}(k,\ell)+\phi_{dd}(k,\ell), &(4.5)\\ \phi_{z_\mu z_\nu}(k,\ell) &= \phi_{ss}(k,\ell)+\Gamma_{d_\mu d_\nu}(k,\ell)\phi_{dd}(k,\ell), &(4.6)\end{aligned}$$

where $\Gamma_{d_\mu d_\nu}(k,\ell)$ is the complex coherence function, defined as:

$$\Gamma_{d_\mu d_\nu}(k,\ell)=\frac{\phi_{d_\mu d_\nu}(k,\ell)}{\sqrt{\phi_{d_\mu d_\mu}(k,\ell)\phi_{d_\nu d_\nu}(k,\ell)}}. \tag{4.7}$$

Based on these expressions of auto- and cross- spectral densities, the speech power spectral density, which is the numerator term of the Wiener post-filter, can be represented as [113]:

$$\hat{\phi}_{ss}^{(\mu\nu)}(k,\ell) = \frac{\Re\Big\{\phi_{z_\mu z_\nu}(k,\ell)\Big\} - \frac{1}{2}\Re\Big\{\Gamma_{d_\mu d_\nu}(k,\ell)\Big\}\Big(\phi_{z_\mu z_\mu}(k,\ell) + \phi_{z_\nu z_\nu}(k,\ell)\Big)}{1 - \Re\Big\{\Gamma_{d_\mu d_\nu}(k,\ell)\Big\}}. \tag{4.8}$$

Then the McCowan post-filter can be derived as [113]:

$$G_M(k,\ell) = \frac{\frac{2}{M(M-1)}\sum_{\mu=1}^{M-1}\sum_{\nu=\mu+1}^{M}\hat{\phi}_{ss}^{(\mu\nu)}(k,\ell)}{\frac{1}{M}\sum_{\mu=1}^{M}\phi_{z_\mu z_\mu}(k,\ell)}. \tag{4.9}$$

Although the McCowan post-filter has been shown to achieve improved performance compared to the Zelinski post-filter using multi-channel recordings in an office, a significant performance degradation is expected when difference between the actual and assumed coherence functions exists. The performance dependence of the McCowan post-filter on the assumed coherence function was also analyzed in [113].

4.4 Proposed microphone array post-filter

In this section, we first describe the coherence function and its application in analyzing a noise field. Then a hybrid post-filter with the assumption of a diffuse noise field is proposed. Finally, advantages of the proposed post-filter are presented qualitatively.

4.4.1 Analysis of a noise field

To characterize a noise field, a widely used measure is the *magnitude-squared coherence* (MSC) function, simply called coherence function, defined as the magnitude square of the complex coherence function and given by:

$$\mathrm{MSC}_{\mu\nu}(k,\ell) = \frac{|\phi_{d_\mu d_\nu}(k,\ell)|^2}{\phi_{d_\mu d_\mu}(k,\ell)\phi_{d_\nu d_\nu}(k,\ell)}, \tag{4.10}$$

A diffuse noise field, which is one of the underlying assumptions of this paper, has been shown to be a reasonable model for many practical noise environments [17]. A diffuse noise field is characterized by the following MSC function:

$$\mathrm{MSC}(k) = \left|\frac{\sin(2\pi k d/c)}{2\pi k d/c}\right|^2, \tag{4.11}$$

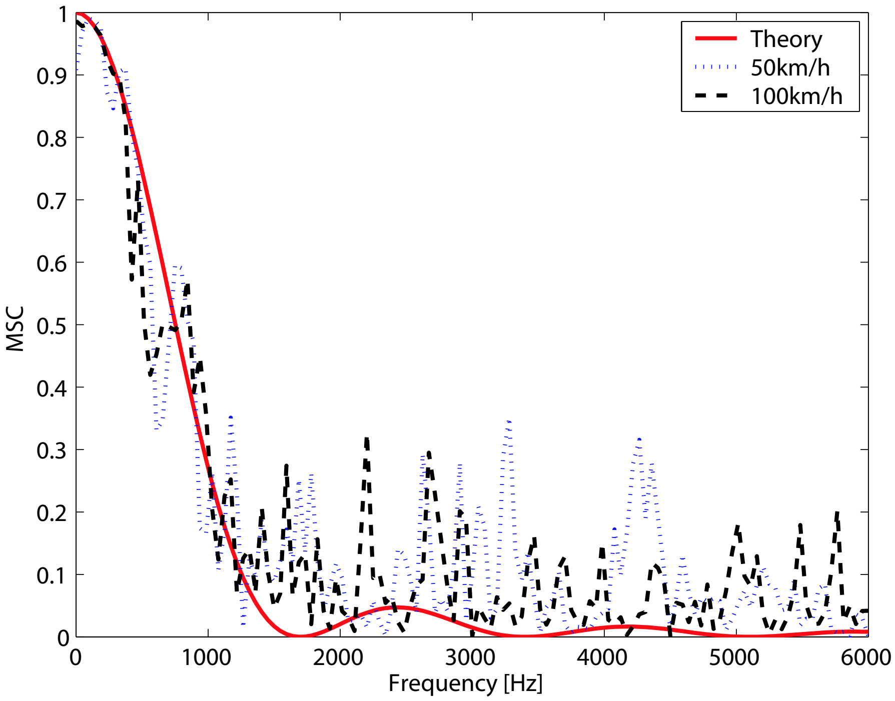


Figure 4.1: Magnitude-squared coherence functions of theoretical diffuse noise field (solid), and in various car environments: 50km/h (dotted) and 100km/h (dashed). (d = 10cm).

where d and c represent the distance between adjacent microphones and the velocity of sound. The MSC function of a perfect diffuse noise field against frequency is plotted in Fig. 4.1. From Fig. 4.1, some characteristics of a diffuse noise field can be easily observed:

1. The MSC function is a frequency-dependent and time-invariant measure;

2. Noise on different microphones is high-correlated in the low frequencies and low-correlated in the high frequencies.

These observations motivate us to divide the spectrum into the low-correlated and high-correlated parts, the transient frequency f_t between two regions is chosen as the first minimum frequency, given by $f_t = c/(2d)$ [108, 137]. Since the velocity of sound c is considered as a constant, the transient frequency is merely determined by the distance d between two microphones, which is a key point for our proposed post-filter.

4.4.2 Proposed post-filter

To formulate the proposed post-filter, let us first give some assumptions on which the proposed post-filter is based:

1. Desired speech signal and noise signal are uncorrelated on each microphone;

2. Noise power spectral density is identical on each microphone;

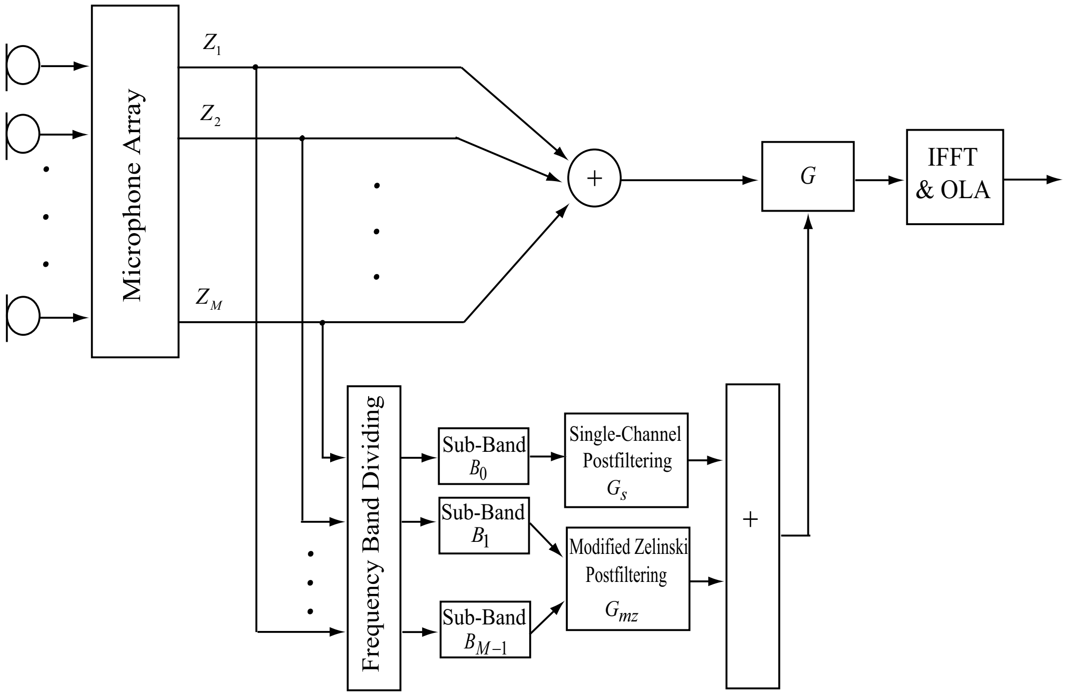


Figure 4.2: Block diagram of the proposed post-filter.

3. Noise on different microphones is diffuse noise.

As a matter of fact, assumption (1) is normally made in speech signal processing, and assumptions (2) and (3) were verified to be fulfilled in a large number of practical noise environments.

In the following discussion, we propose a hybrid post-filter, which applies a modified Zelinski post-filter in the high frequency region and a single-channel Wiener post-filter in the low frequency region, with the hope of enhancing its noise reduction performance. The block diagram of the proposed post-filter along with beamformer is plotted in Fig. 4.2.

A modified Zelinski post-filer in the high frequencies

Based on the assumption that noise on each microphone is mutually uncorrelated, the Zelinski post-filter provides a solution for minimizing the mean-square error between speech and its estimate in an incoherent noise field. As mentioned above, its performance is often significantly degraded when the correlated noise components are involved in estimating the cross-spectral densities of multi-channel inputs. It is, therefore, believed that the performance degradation would be eliminated if the noise, used to estimate the cross-spectral densities of multi-channel inputs, is sufficiently uncorrelated.

As Fig. 4.1 demonstrates, in a diffuse noise field, the spatially weakly correlated noise components on different microphones only exist in the frequencies over the transient frequency f_t. Since the transient frequency is determined by the distance between microphones, microphone pairs with different inter-element spacing are characterized by different transient frequencies. That is, for different microphone pairs with different inter-element spacing, low correlated noise is found in different frequency regions. Furthermore, for a certain frequency, noise is mutually low correlated only on limited microphone pairs, generally not on all pairs. This fact motivates us to propose a modified Zelinski post-filter by calculating the cross-spectral densities of multi-channel beamformer outputs on the corresponding microphone pairs, not on all sensor pairs (as used in the Zelinski and the McCowan post-filters).

The modified Zelinski post-filter is implemented in the following steps:

1. *Determine the transient frequencies according to the microphone array geometry.* Considering a M-sensor array with inter-element spacing $d_{\mu\nu}$ between sensors μ and ν $(\mu, \nu \leq M)$, we have $M(M-1)/2$ microphone pairs which determine $M(M-1)/2$ transient frequencies, each of them can be calculated by $f_{t,\mu\nu} = c/(2d_{\mu\nu})$. Since the inter-element spacings are identical for some microphone pairs, some transient frequencies are identical as well. In principle, if the equidistant microphones are assumed, among $M(M-1)/2$ microphone pairs, only $M-1$ pairs have different inter-element spacings. Correspondingly, we can determine $M-1$ different transient frequencies, denoted by $f_t^1, f_t^2, \cdots, f_t^{M-1}$. Without loss of generality, we further assume the following relationship between transient frequencies $f_t^1 < f_t^2 < \cdots < f_t^{M-1}$.

2. *Determine the microphone pairs on which noise is mutually uncorrelated for each frequency.* As a matter of fact, the $M-1$ different transient frequencies, $f_t^1, f_t^2, \cdots, f_t^{M-1}$, divide the full frequency band into M sub-bands, denoted by $B_0, B_1, \cdots, B_{M-1}$. In each sub-band (expect B_0), some microphone pairs provide low correlated noise components on microphones of the pairs. In principle, the $M(M-1)/2$ microphone pairs can be grouped into $M-1$ sets where some microphone pairs are re-used. Each of $M-1$ sets includes the microphone pairs on which noise signals are mutually weakly correlated for the individual frequency of interest. Corresponding to the transient frequencies $f_t^1, f_t^2, \cdots, f_t^{M-1}$, the $M-1$ microphone pair sets are represented as: $\Omega_1, \Omega_2, \cdots, \Omega_{M-1}$.

3. *Compute the spectral densities of the desired speech signal and the noisy signal.* For each frequency in sub-band $B_m (1 \leq m \leq M-1)$, the noise on the microphone pairs of set Ω_m is weakly correlated. Thus, the spectral densities of the noisy signal and the desired speech signal can be estimated from the auto- and cross- spectral densities of the multi-channel inputs, that is:

$$\begin{aligned} \hat{\phi}_{z_\mu z_\mu}(k,\ell) &= \phi_{ss}(k,\ell) + \phi_{dd}(k,\ell), && (4.12) \\ \hat{\phi}_{z_\mu z_\nu}(k,\ell) &= \phi_{ss}(k,\ell). && (4.13) \end{aligned}$$

4. *Compute the gain function of the modified Zelinski post-filter.* To improve the robustness of the proposed post-filter, estimates of the auto- and cross- spectral densities are averaged across the microphone pairs in the corresponding pair set Ω_m, generally not all microphone pairs. The gain function of the modified Zelinski post-filter is given by:

$$G_{mz}(k,\ell) = \frac{\frac{1}{|\Omega_m(k)|} \sum\limits_{\{\mu,\nu\}\in\Omega_m(k)} \Re\{\hat{\phi}_{z_\mu z_\nu}(k,\ell)\}}{\frac{1}{|\Omega_m(k)|} \sum\limits_{\{\mu,\nu\}\in\Omega_m(k)} \left[\frac{1}{2}\left(\hat{\phi}_{z_\mu z_\mu}(k,\ell) + \hat{\phi}_{z_\nu z_\nu}(k,\ell)\right)\right]}. \tag{4.14}$$

Note that in this modified Zelinski post-filter, the average for the auto- and cross-spectral densities is performed on only limited microphone pairs in the corresponding pair set Ω_m determined in step 2. Since noise on microphones in set Ω_m is weakly correlated, the estimation error caused by the correlated noise components should be mitigated, improving the accuracy and robustness of this modified Zelinski post-filter. While in other post-filters including the Zelinski post-filter and the McCowan post-filter, the average was done across all microphone pairs, involving the correlated noise components in estimating the spectral densities, which introduces the estimation error and further degrades the noise reduction performance.

Moreover, it should be noted that the first two steps should be done in advance, since they are only dependent on the microphone array geometry and independent on the input signals. The limited microphone pairs, involved in the estimation procedure of the auto- and cross- spectral densities, contribute to the decrease of computational cost of this modified Zelinski post-filter.

A single-channel technique in the low frequencies

In the low frequency sub-band (B_0 where $k < f_t^1$), noise on all microphone pairs is high-correlated, indicating that the auto-spectral density of the desired speech signal can not be estimated from the cross-spectral density of multi-channel inputs. Thus, no post-filter that calculates the auto- and cross- spectral densities can perform well in these frequencies.

In the low frequencies ($k < f_t^1$), therefore, we turn to a single-channel technique to estimate a Wiener filter. The gain function of the Wiener filter is rewritten here:

$$G_s(k,\ell) = \frac{E\left[|S(k,\ell)|^2\right]}{E\left[|S(k,\ell)|^2\right] + E\left[|D(k,\ell)|^2\right]} = \frac{SNR_{priori}(k,\ell)}{1 + SNR_{priori}(k,\ell)}, \tag{4.15}$$

where $SNR_{priori}(k,\ell)$ the *a priori* SNR, as named in [43], defined by $SNR_{priori}(k,\ell) = E\left[|S(k,\ell)|^2\right] \Big/ E\left[|D(k,\ell)|^2\right]$. The estimate of the *a priori* SNR, $SNR_{priori}(k,\ell)$, is updated in a decision-directed scheme, as follows [43]:

$$SNR_{priori}(k,\ell) = \alpha_s \frac{\left|S(k,\ell-1)\right|^2}{E\left[\left|D(k,\ell-1)\right|^2\right]} + (1-\alpha_s)\mathrm{max}\left[SNR_{post}(k,\ell)-1,0\right], \quad (4.16)$$

where α_s $(0 < \alpha_s < 1)$ is a forgetting factor and $SNR_{post}(k,\ell)$ is the *a posteriori* SNR, as named in [43], defined by $SNR_{post}(k,\ell) = \left|Z(k,\ell)\right|^2 \Big/ E\left[\left|D(k,\ell)\right|^2\right]$. This decision-directed estimation mechanism for the *a priori* SNR significantly decreases the residual "musical noise", as detailed in [20].

To improve the performance of this single-channel Wiener filter, a crucial point is to estimate noise power spectral density $E\left[\left|D(k,\ell)\right|^2\right]$ with high accuracy. Here, it is implemented by a soft-decision based approach, given by:

$$E\left[\left|D(k,\ell)\right|^2\right] = \beta_s E\left[\left|D(k,\ell-1)\right|^2\right] + (1-\beta_s)E\left[\left|D(k,\ell)\right|^2 \middle| \mathbf{Z}(k,\ell)\right], \quad (4.17)$$

where β_s $(0 < \beta_s < 1)$ is a forgetting factor controlling the update rate of noise estimation. Under speech presence uncertainty, the second term in the right side of Eq. (4.17) can be estimated as:

$$E\left[\left|D(k,\ell)\right|^2 \middle| \mathbf{Z}(k,\ell)\right] = q(k,\ell)\overline{\left|\mathbf{Z}(k,\ell)\right|^2} + (1-q(k,\ell))E\left[|D(k,\ell-1)|^2\right], \quad (4.18)$$

where $q(k,\ell)$ denotes the speech absence probability, $\overline{\left|\mathbf{Z}(k,\ell)\right|^2} = \frac{1}{M}\sum_{m=1}^{M}\left|Z_m(k,\ell)\right|^2$ the average of the individual power spectral density on each sensor. The reason for calculating this average is that considering only one sensor may yield a biased measurement. With the assumption of a complex Gaussian statistic model and applying the Bayes rule and total probability theorem, the speech absence probability conditioned on the observations can be given by [43]:

$$q(k,\ell) = \left(1 + \frac{1-q^{'}(k,\ell)}{q^{'}(k,\ell)} \frac{1}{1+SNR_{priori}(k,\ell)} \mathrm{exp}\left(\frac{SNR_{post}(k,\ell)SNR_{priori}(k,\ell)}{1+SNR_{priori}(k,\ell)}\right)\right)^{-1} \quad (4.19)$$

where $q^{'}(k,\ell)$ is the *a priori* speech absence probability. In the experiments, $q^{'}(k,\ell)$ is set to 0.5 as in [143].

Here, it is of interest to note that the post-filter described above given by Eq. (4.15) is a Wiener filter exactly. This post-filter, which minimizes the *mean square error* (MSE) of spectrum, is also different from the Ephraim-Malah algorithm which is based on the MSE of spectral amplitude [43]. In comparison of the traditional post-filters, this proposed Wiener filter show some advantages: (i) it is able to greatly reduce the "musical noise" due to the use of the decision-directed *a priori* SNR estimation technique [20]; (ii) it is able to deal with the non-stationary noise due to the soft-decision based noise estimation technique [28].

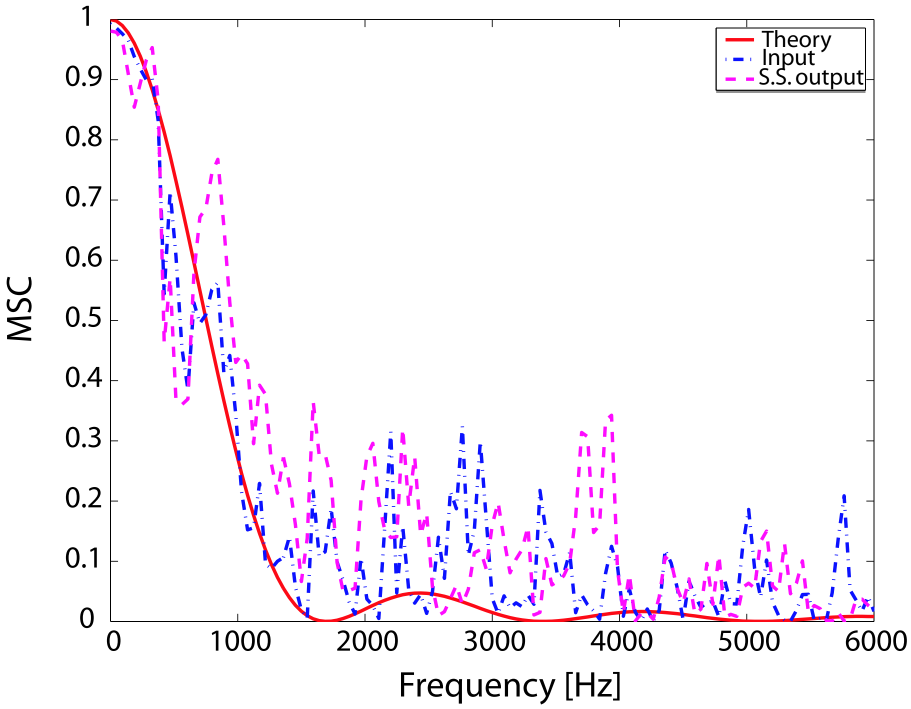


Figure 4.3: Magnitude-squared coherence function of theoretical diffuse noise field (solid), multi-microphone inputs (dash-dotted) and outputs of the localized noise suppression algorithm (dash). (d = 10cm).

4.4.3 Analysis of proposed post-filter

In theory, the proposed post-filter is a Wiener post-filter. In the low frequency region, the single-channel post-filter given by Eq. (4.15) is obviously a Wiener filter. In the high frequency region, since noise used to formulate the modified Zelinski expression are weakly correlated, the cross-spectral density of multi-channel signals provides more accurate speech auto-spectral density estimate. Therefore, the modified Zelinski post-filter used in the high frequency region approaches a Wiener filter. Comparatively, although the Zelinski post-filter and the McCowan post-filter have a Wiener-filter structure, performance degradation is expected due to the correlated noise components involved in estimating cross-spectral densities.

It also should note that the proposed post-filter provides a more general expression for the microphone array post-filter. In a perfectly incoherent noise field, the proposed post-filter will reduce to the Zelinski post-filter, just by setting the transient frequencies to zero. And in a perfectly coherent noise field, the proposed post-filter will reduce to the single-channel Wiener post-filter, just by setting the transient frequencies to the highest frequency of interest.

4.5 Experimental validation

To validate the effectiveness of the proposed hybrid post-filter in a diffuse noise field, its performance was investigated and further compared to other conventional post-filters, including the Zelinski post-filter [163], the McCowan post-filter [113] and the single-channel Wiener post-filter alone [5], in various car noise environments. A beamformer was first applied to the multi-channel noisy signals. Then, the beamformer output was further enhanced by the studied post-filters.

To assess the performance of the studied post-filters, some experiments were performed using the sound data same as those used in chapter 3 in *pseudo real-world environment* and *real-world environment*. Their performance was further evaluated in terms of objective (segmental SNR and MFCC distance, defined in chapter 3) and subjective speech quality measures.

4.5.1 Experimental configurations

The effectiveness of the diffuse noise field was investigated by comparing the measured MSC function calculated from real noise recordings with the theoretical function, plotted in Fig. 4.3. It can be seen from Fig. 4.3 that the measured MSC function follows the trend of the theoretical function, which fulfills the assumption of a diffuse noise field used in the proposed post-filter. Moreover, it should be noted that although the beamformer has effect on the input noise components themselves, the spatial characteristics of the noise field at the beamformer output is not changed, that is, the diffuse noise field characteristic is preserved. Therefore, the assumption of a diffuse noise field is still satisfied for the proposed hybrid post-filter.

The beamforming filter was implemented by a superdirective beamformer [39], which is a solution of MVDR beamformer in a diffuse noise field. Note that with the consideration of robustness, the white noise gain constraining procedure was applied during the implementation [39]. The gain function of this superdirective beamformer, which is a function of frequency k, is given by [39]:

$$W_{MVDR}(k) = \frac{\Gamma^{-1}_{diffuse}(k)A(k)}{A^H(k)\Gamma^{-1}_{diffuse}(k)A(k)}, \tag{4.20}$$

where $A(k)$ denotes the acoustic transfer functions between speech source and microphones. The *directivity index* (DI), which shows the noise reduction ability of the array in a diffuse noise field, is given by [17]:

$$\mathrm{DI}(k) = 10 \cdot \log_{10} \left(\frac{\left| W^H_{MVDR}(k)A(k) \right|^2}{W^H_{MVDR}(k)\Gamma_{diffuse}(k)W_{MVDR}(k)} \right), \quad \text{[dB]} \tag{4.21}$$

and shown in Fig. 4.4. Obviously, the superdirective beamformer illustrates the low noise reduction performance for the low-frequency noise components.

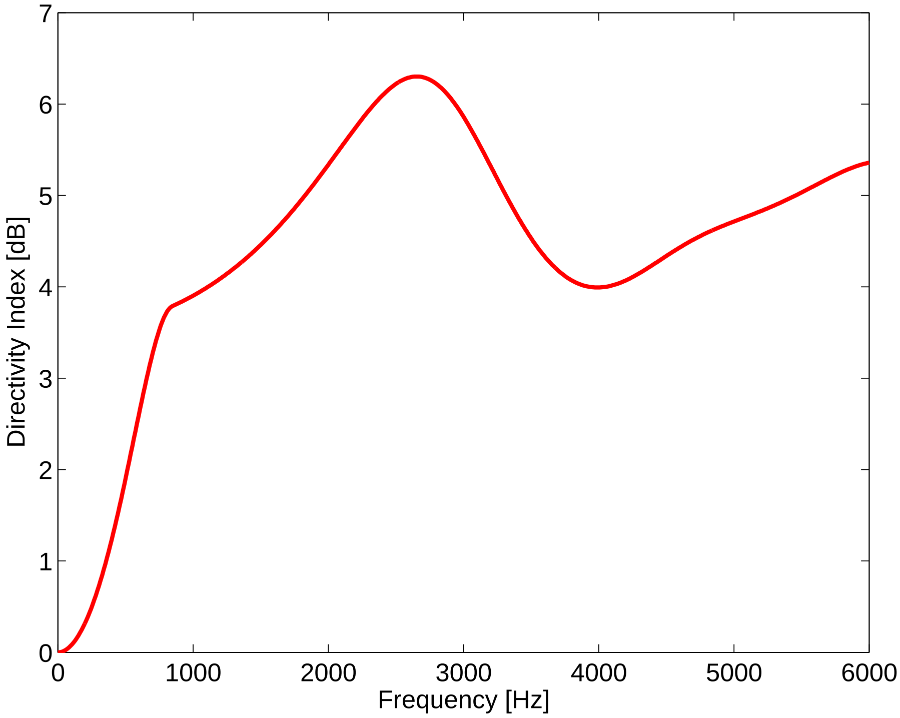


Figure 4.4: Directivity index of the superdirective beamformer (M=3, d=10cm).

4.5.2 Objective evaluation results

Experimental results of the average SEGSNR and MFCC distance calculated for both *pseudo real-world environment* and *real-world environment* in two car noise conditions (50km/h and 100km/h) at various SNR levels, are plotted in Figs. 4.5 - 4.8. The results were averaged across all sentences in each noise condition. The performance was evaluated at the first microphone, the beamformer output and the studied post-filter outputs.

As illustrated in Figs. 4.5 and 4.6, the beamformer shows low SEGSNR improvement due to its low directivity (shown in Fig. 4.4) in the low frequencies. The Zelinski post-filter also only offers limited performance improvement. By integrating an appropriate coherence function of the noise field into the post-filter formulation, the McCowan post-filter shows a great SEGSNR improvement. The single-channel Wiener post-filter shows further SEGSNR improvements compared with the Zelinski and the McCowan post-filters in all noise conditions. The proposed post-filter demonstrates highest performance improvements in SEGSNR sense among the studied post-filters for both *pseudo real-world environment* and *real-world environment* in all tested conditions.

Concerning the results of MFCC distance, plotted in Figs. 4.7 and 4.8, we can readily observe that the beamformer alone and the Zelinski post-filter decrease MFCC distances in all conditions with regard to noisy inputs. Moreover, the single-channel Wiener post-filter shows the lower MFCC distances, especially at low SNRs. The proposed post-filter and the McCowan post-filter offer the lowest speech distortion to an almost same degree at all SNRs, with regard to other post-filters for both *pseudo real-world environment* and *real-world environment* in all noise conditions. Compared to the McCowan post-filter, the proposed post-filter yields less residual noise (eg. "musical noise") which was proven by the informal listening tests.

Taking account of noise reduction and speech distortion, the proposed post-filter demonstrates the great superiority, the highest speech quality and the lowest speech distortion, with regard to other comparative post-filters in all noise conditions.

4.5.3 Subjective evaluation results

Subjective evaluation of the studied post-filters was performed using speech spectrograms and validated by informal listening tests. Typical examples of speech spectrograms in *pseudo real-world environment* corresponding to the Japanese sentence "dozo yoroshiku" and in *real-world environment* corresponding to the Japanese sentence "hatinohe kesennuma yukuhasi", are presented in Figs. 4.9 and 4.10 under the car conditon with a speed of 100km/h. For the *pseudo real-world environment*, Fig. 4.9(d) shows that the output of beamformer is characterized by high-level low-frequency noise due to its weakness in the low frequencies, as shown in Fig. 4.4. The Zelinski post-filter also offers very limited performance in the low frequencies because of the high-coherence characteristics of noise in this region. Fig. 4.9(f) illustrates that the McCowan post-filter does suppress a large amount of noise, even in the low frequency region, and the residual noise exists due to the difference between the assumed and actual coherence values at instantaneous time. The single-channel Wiener post-filter results in speech distortion, as shown in Fig. 4.9(g). Fig. 4.9(h) illustrates the proposed post-filter is able to further suppress the correlated and uncorrelated noises simultaneously, without additional speech distortion. Informal listening tests proved the superiority of the proposed post-filter compared to others. With the other post-filters, the advantage of the proposed post-filter is also confirmed by the observations from the spectrograms in *real-world environments* shown in Fig. 4.10.

4.5.4 Discussions

Compared to the other post-filters, the advantages of the proposed post-filter are discussed in this section from the standpoint of practice based on the experimental results.

The proposed hybrid post-filter is superior to the Zelinski post-filter since the basic assumption of the proposed post-filter (diffuse noise field) is more reasonable than that of the Zelinski post-filter (incoherent noise field) in practical environments. In addition, the Zelinski post-filter fail to reduce the low-frequency (high-correlated) noise components, while the proposed hybrid post-filter is successful for these noise components.

The proposed hybrid post-filter is superior to the McCowan post-filter. The McCowan post-filter is determined based on the coherence function of the noise field itself. Thus, its performance is greatly dependent on the accuracy of the assumed coherence function. The differences between the assumed and actual coherence functions result in its performance significant degradation. However, the proposed hybrid post-filter utilizes the transient frequency only to distinguish correlated and uncorrelated noises, independent on the actual instantaneous values of the coherence function, alleviating the effect caused by the difference between the assumed and actual coherence functions in some sense.

The proposed hybrid post-filter should be superior to the single-channel Wiener filter which is used in the whole frequency band. The single-channel Wiener filter, which is based on measurements of noise characteristics, can hardly be applied for highly non-stationary noise sources even if the soft-decision mechanism is adopted. However, multi-channel technique based on the estimates of the auto- and cross- spectral densities theoretically provides good performance for the highly non-stationary noise. Our proposed modified Zelinski post-filter utilizes this attractiveness fully in each frequency bin in the high frequency region. Additionally, in the low frequency region, both the proposed and the single-channel Wiener post-filters have the same problem in dealing with the highly non-stationary noises.

4.6 Summary

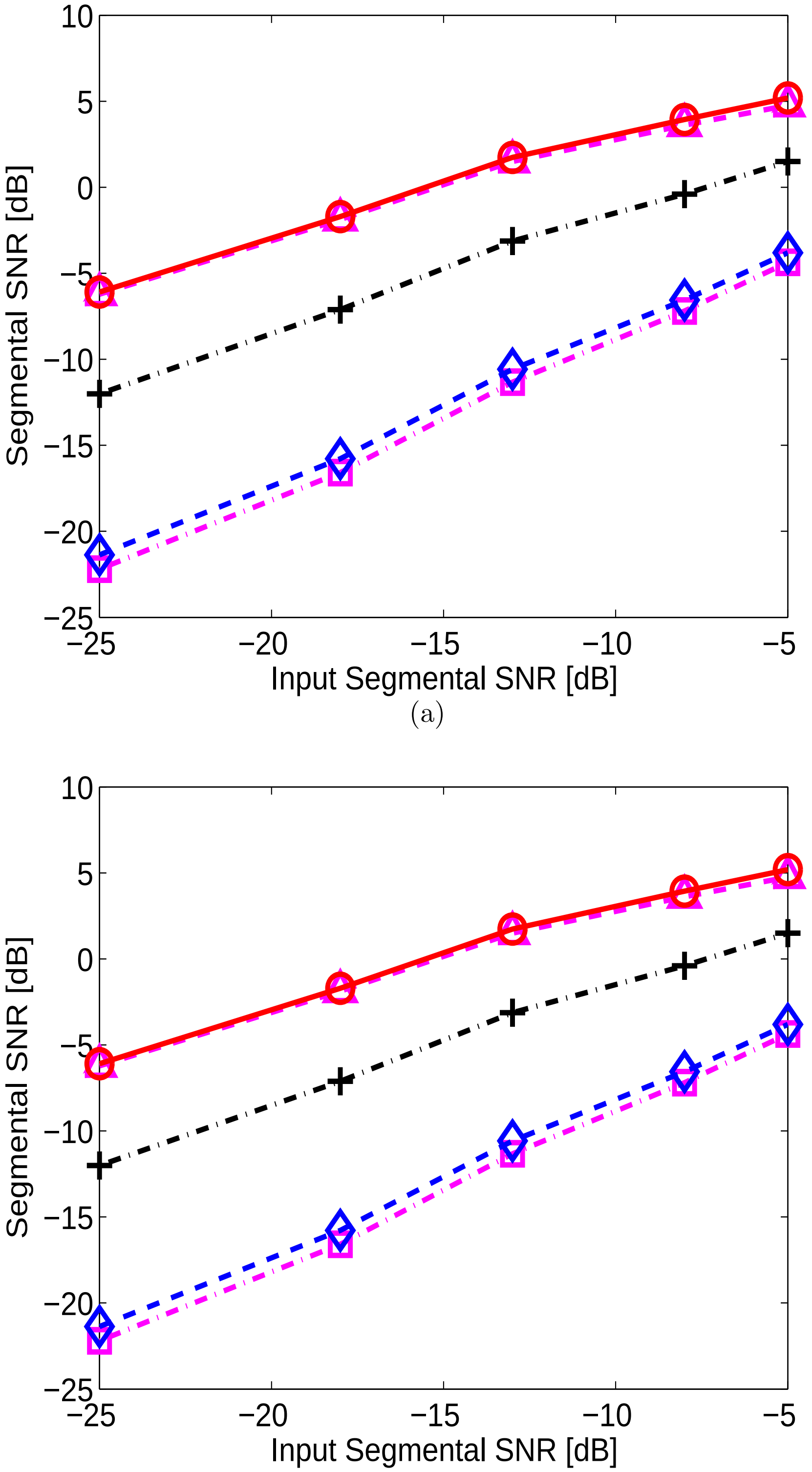
In this chapter, we focused on further dealing with the problem of suppressing non-localized noise components at the beamformer output.

In section 4.1, we briefly reviewed the state-of-the-art microphone array post-filters and more attention was paid to the drawbacks/disadvantages of these existing post-filters.

In section 4.2, we formulated the signal model at the beamformer output, which is composed of desired speech signals and non-localized noise components, since the localized noise components have been suppressed by the beamformer based algorithms.

In section 4.3, we introduced the Zelinski post-filter and McCowan post-filter in brief. The proposed post-filter was based on the former one and compared to both.

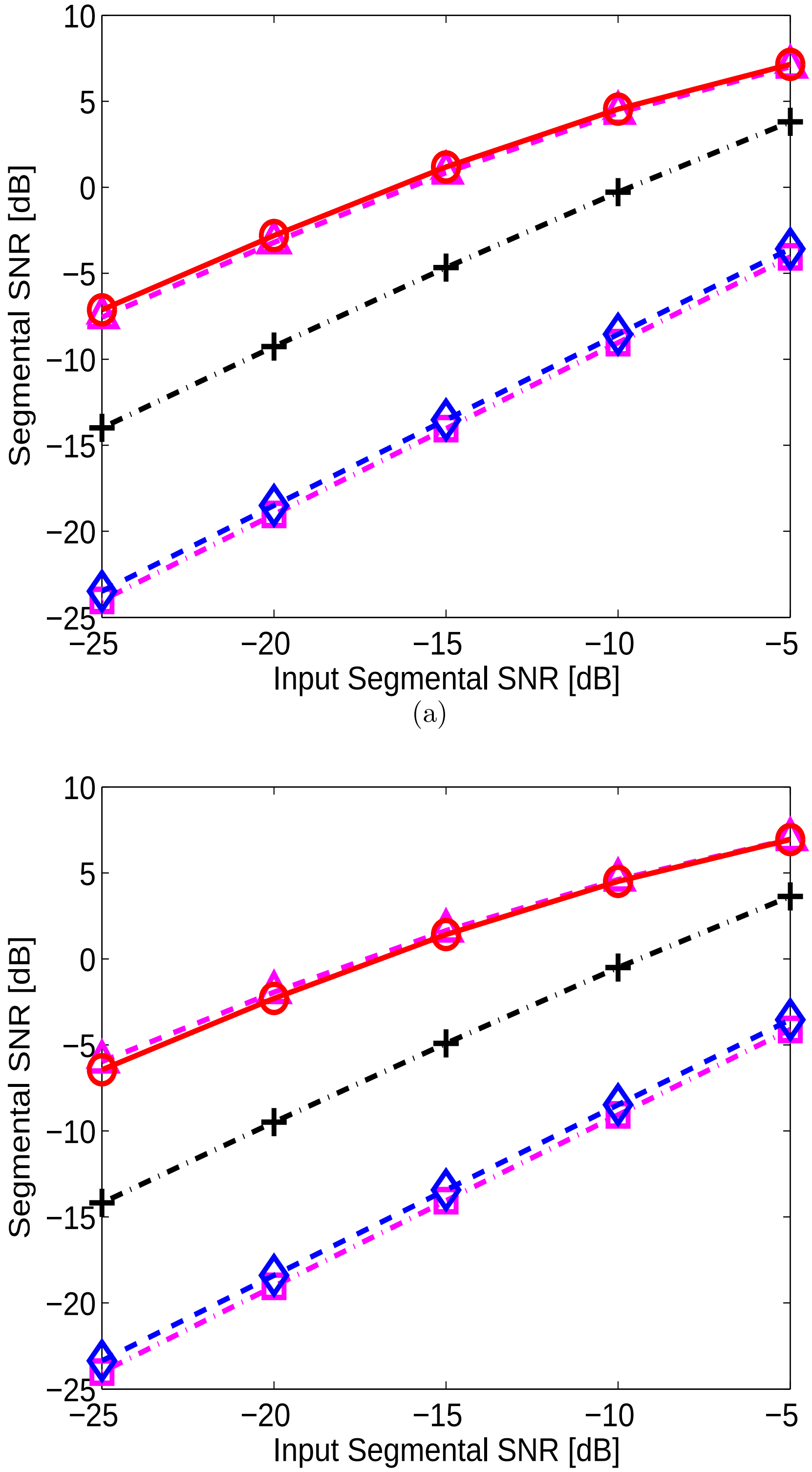
In section 4.4, we proposed a hybrid Wiener post-filter for microphone arrays with the assumption of a diffuse noise field. In this hybrid post-filter, a modified Zelinski post-filter is applied to the low-correlated frequencies in high frequency region; a single-channel Wiener post-filter is applied to the high-correlation frequencies in low frequency region. In theory, the proposed post-filter is a Wiener post-filter, and in practice, it is able to reduce both correlated and uncorrelated noise components in a diffuse noise field. The superiority of the proposed post-filter was verified by the experimental using real-world speech and noise recordings.



(a)

(b)

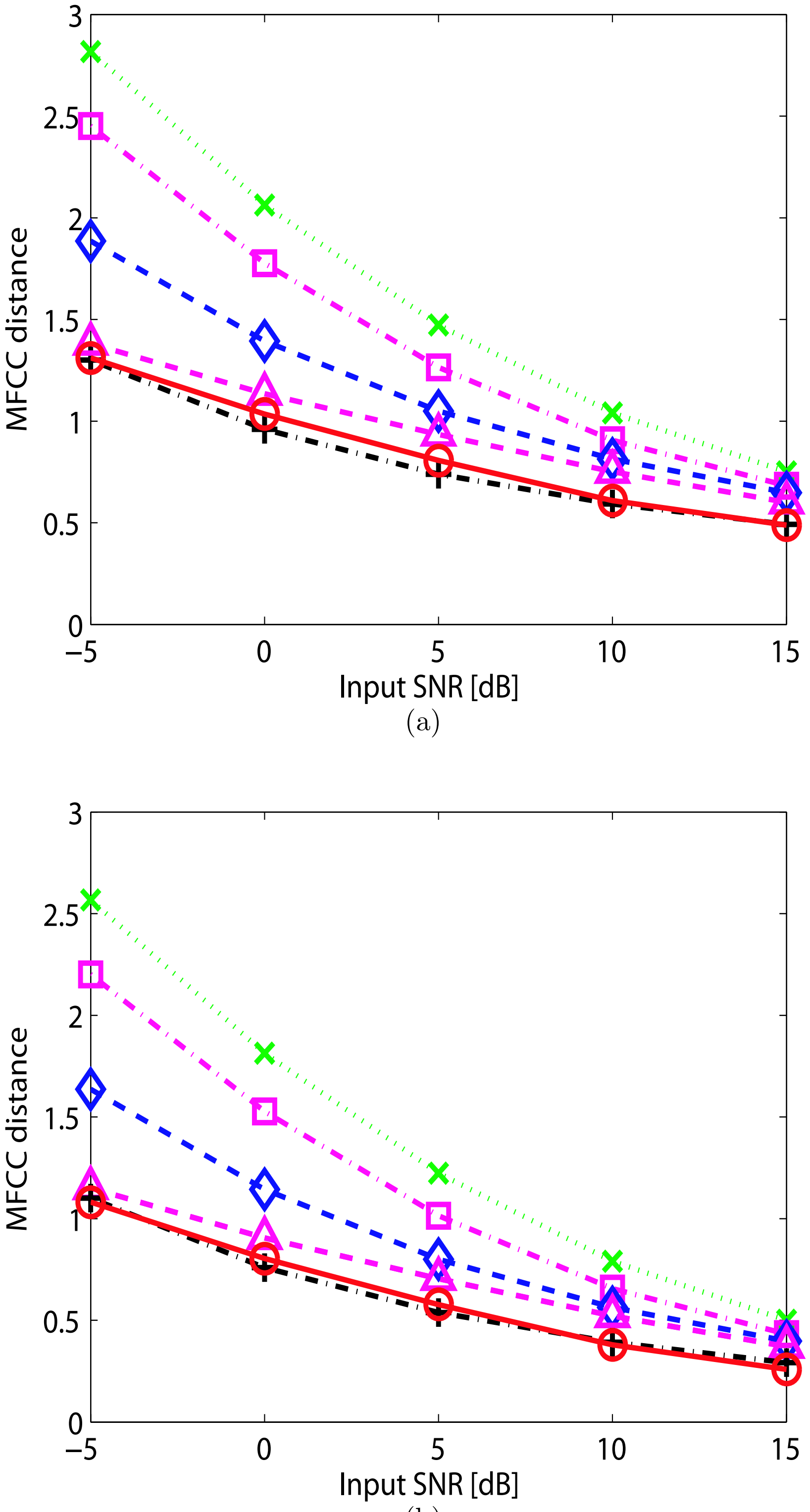
Figure 4.5: Average segmental SNR (SEGSNR) in **pseudo real-world environment** at beamformer output (□), Zelinski post-filter output (◊), McCowan post-filter output (+), single-channel Wiener filter output (△), proposed post-filter output(○), in various noise conditions: 50km/h (a) and 100km/h (b).



(a)

(b)

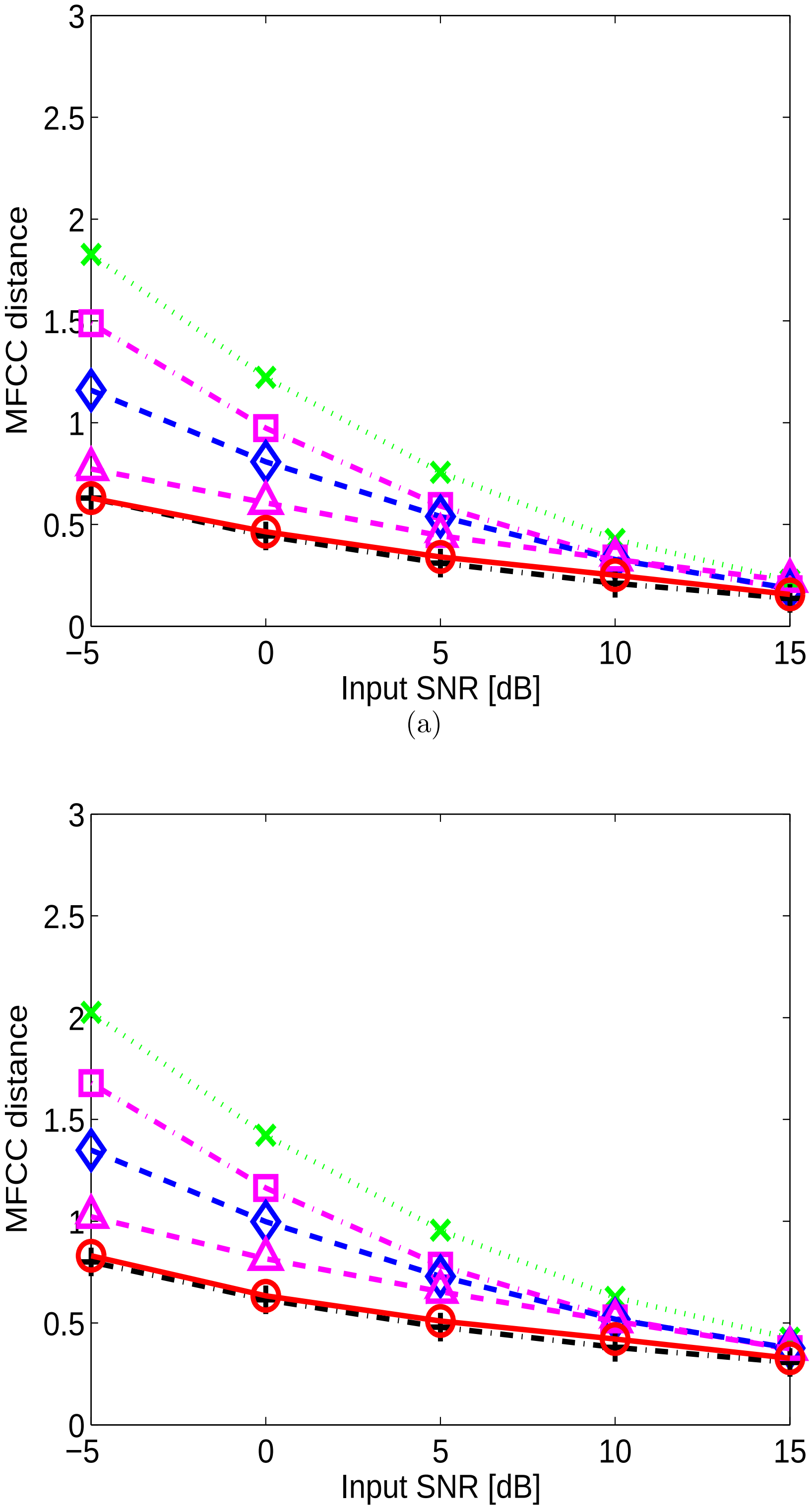
Figure 4.6: Average segmental SNR (SEGSNR) in **real-world environment** at beamformer output (□), Zelinski post-filter output (◊), McCowan post-filter output (+), single-channel Wiener filter output (△), proposed post-filter output(∘), in various noise conditions: 50km/h (a) and 100km/h (b).



(a)

(b)

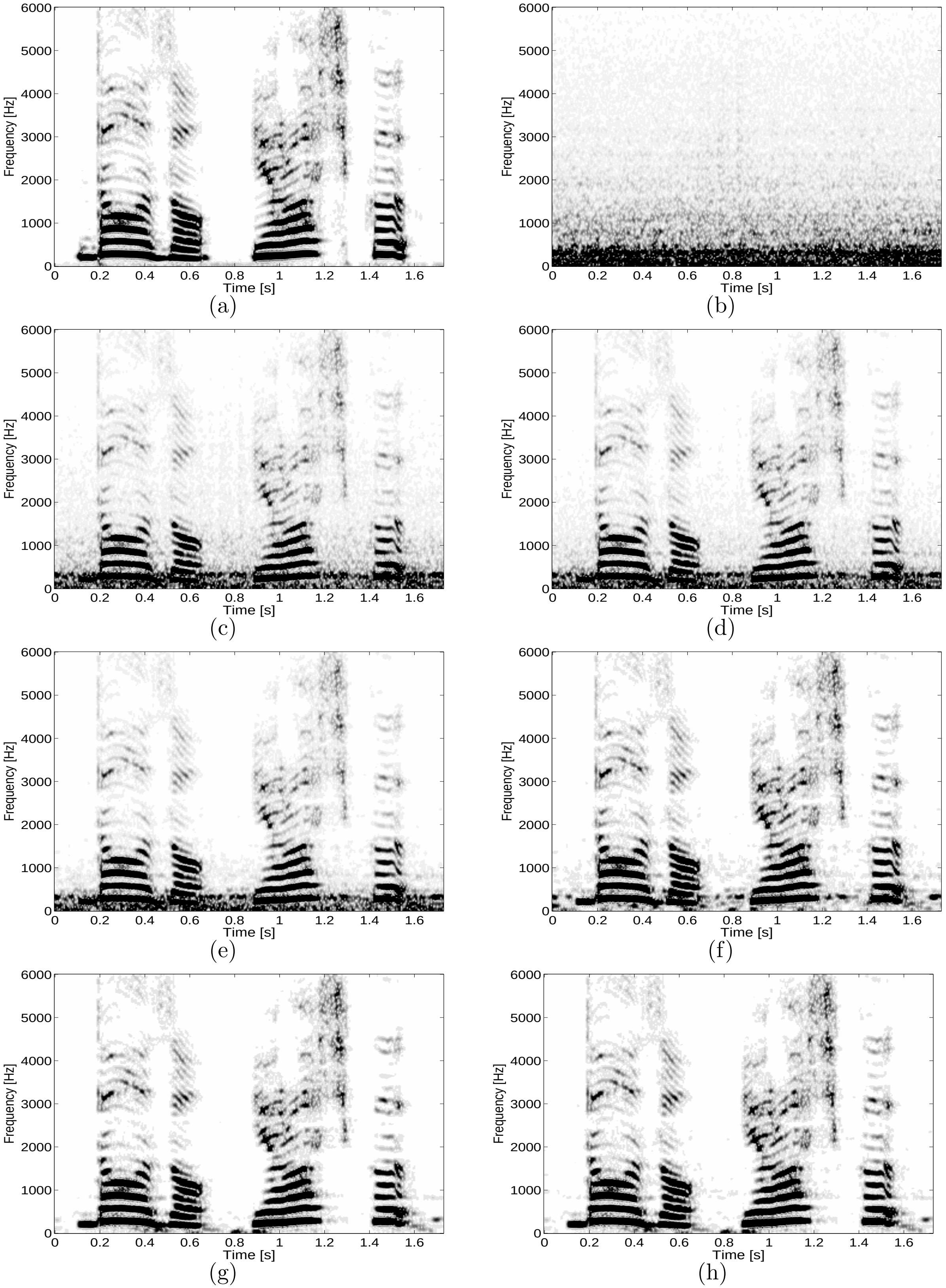
Figure 4.7: Average MFCC distance in **pseudo real-world environment** at the first microphone (×), beamformer output (□), Zelinski post-filter output(◊), McCowan post-filter output(+), single-channel Wiener filter output(△), proposed post-filter output(○), in various noise conditions: 50km/h (a) and 100km/h (b).



(a)

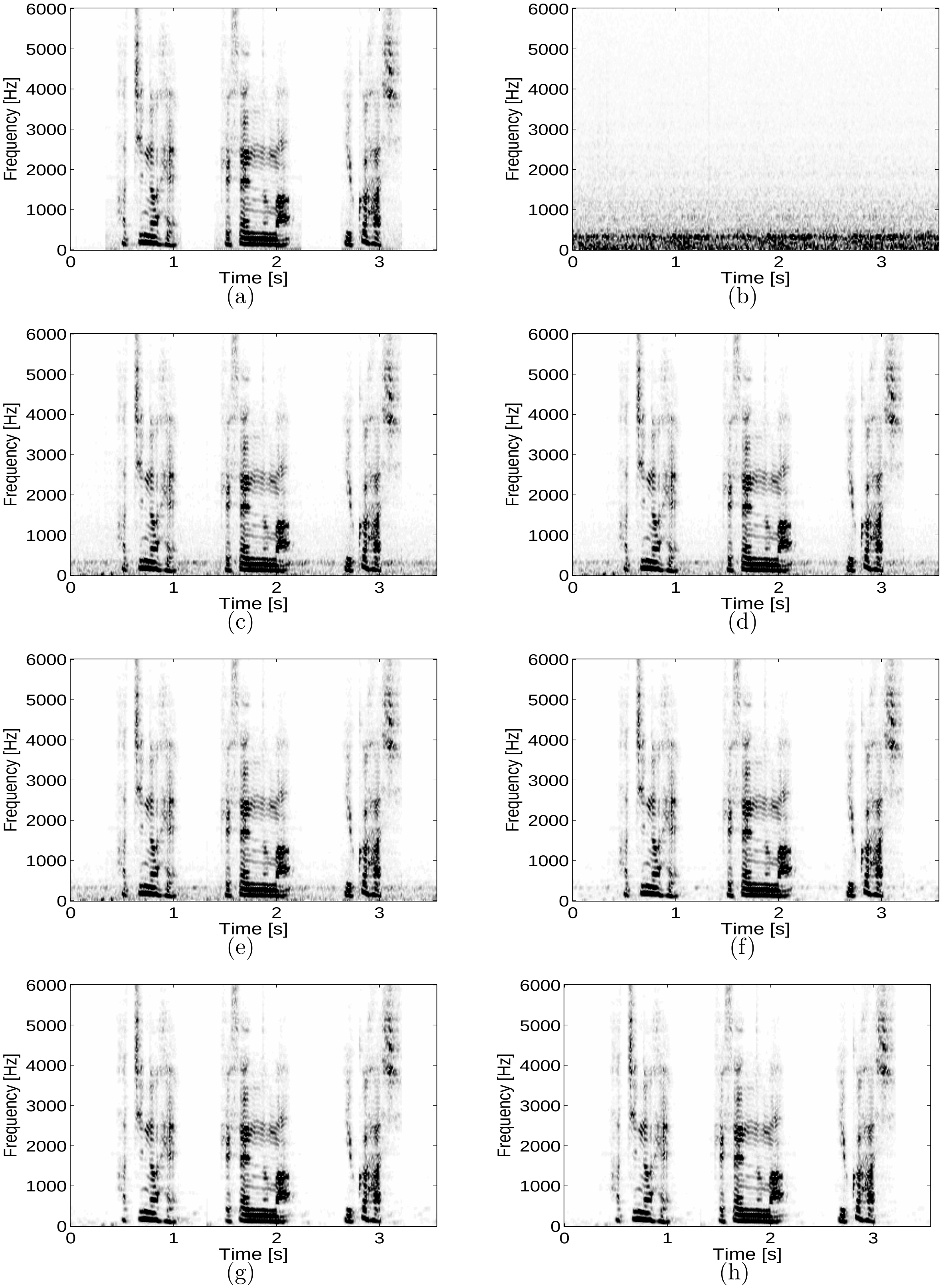
(b)

Figure 4.8: Average MFCC distance in **real-world environment** at the first microphone (×), beamformer output (□), Zelinski post-filter output(◊), McCowan post-filter output(+), single-channel Wiener filter output(△), proposed post-filter output(○), in various noise conditions: 50km/h (a) and 100km/h (b).



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 4.9: Speech spectrograms in **pseudo real-world environment**. (a) Original clean speech signal at the first microphone: "dozo yoroshiku"; (b) Noise signal at the first microphone; (c) Noisy signal at the first microphone (SNR = 10 dB); (d) Beamformer output; (e) Zelinski post-filter output; (f) McCowan post-filter output; (g) Single-channel Wiener post-filter output; (h) Proposed post-filter output.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 4.10: Speech spectrograms in **real-world environment**. (a) Original clean speech signal at the first microphone: “hatinohe kesennuma yukuhasi”; (b) Noise signal at the first microphone; (c) Noisy signal at the first microphone (SNR = 10 dB); (d) Beamformer output; (e) Zelinski post-filter output; (f) McCowan post-filter output; (g) Single-channel Wiener post-filter output; (h) Proposed post-filter output.

Chapter 5

Evaluation of proposed system with speech recognition

Although automatic speech recognition systems have achieved high recognition accuracy in noise-free conditions, their performance significantly degrades in practical environments due to the undesired acoustic noises, which was assumed to include localized and non-localized noise components in this research. The acoustic noises introduce the mismatches between training conditions and testing conditions, resulting in the degradation of the recognition accuracy. It is believed that the performance of recognition system will be improved by suppressing the acoustic noise signals from the observed noisy signals, which eliminates the mismatches between training and testing conditions and in turn improving the recognition accuracy.

In this research, we considered that the undesired noise signal consists of localized noise components and non-localized noise components. Then, we constructed a beamforming based technique to suppress the localized noise and a hybrid post-filter to suppress the non-localized noise. The superiorities of the two noise reduction parts have been verified in chapters 3 and 4 respectively by the experiments in both *pseudo real-world environment* and *real-world environment* in terms of the objective and subjective speech quality measures.

In this chapter, we will further examine the performance of the proposed noise reduction system, consisting of localized noise suppression and non-localized noise suppression as a whole noise-reduction system, using speech recognition experiments. The comprehensive experiments were conducted using the multi-channel noise recordings in various car environments and experimental results shows that the proposed noise reduction system improve the speech recognition accuracy at about 20% improvement averaged across all noise conditions.

5.1 Introduction

Acoustic noise signals degrading the speech quality have been dealt with by using the proposed noise reduction system, yielding the enhanced high-quality speech signals. The effectiveness and superiority of the proposed noise reduction system have been confirmed by a variety of comprehensive experiments in the last two chapters. In this work, we deal with the processing of signals received by a microphone array for inputting into a speech recognition system with the goal of improving speech recognition accuracy. In this chapter, we first briefly review the recognition process by describing how a speech waveform is converted into a sequence of feature vectors and how these feature vectors are then processed by the recognition system in order to generate a hypothesis of the words that were spoken. Then, we pay our main attention to the speech recognition experiments to evaluate the performance of the proposed noise reduction system, which is based on microphone array and post-filtering, and to the discussions based on these experimental results.

5.2 Principle of automatic speech recognition

The final objective of the speech recognition system is to estimate the correct sequence of words the desired speaker uttered. However, in state-of-the-art recognition systems, speech recognition is performed on a sequence of features extracted from the speech signals in short segments rather than directly on the speech signals.

Suppose that $\mathcal{O}$ represents a sequence of feature vectors extracted from the speech signal, speech recognition systems operate according to the optimal classification equation, formulated as:

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w} \in \boldsymbol{W}} P(\boldsymbol{w}|\mathcal{O}), \tag{5.1}$$

where $\hat{\boldsymbol{w}}$ is the sequence of words hypothesized by the recognizer and $\boldsymbol{W}$ is the set of all possible word sequence that can be hypothesized by the recognition system. However, this expression is not actually computed by recognition systems. Instead, using the Bayes rule Eq. (5.1) can be rewritten as:

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w} \in \boldsymbol{W}} \frac{P(\mathcal{O}|\boldsymbol{w})P(\boldsymbol{w})}{P(\mathcal{O})}, \tag{5.2}$$

where $P(\mathcal{O}|\boldsymbol{w})$ is the acoustic likelihood, representing the probability that feature sequence $\mathcal{O}$ is observed given that word sequence $\boldsymbol{w}$ was spoken, and $P(\boldsymbol{w})$ is the language score, the *a priori* probability of a particular word sequence $\boldsymbol{w}$. This latter term is computed using a language model. Because we are maximizing Eq. (5.2) with respect to the word sequence $\boldsymbol{w}$ for a given sequence of observation $\mathcal{O}$, the denominator term $P(\mathcal{O})$ can be ignored in the maximization, resulting in:

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w} \in \boldsymbol{W}} P(\mathcal{O}|\boldsymbol{w})P(\boldsymbol{w}). \tag{5.3}$$

Therefore, the process of recognizing an utterance of speech can be dividing into two main stages: feature extraction where a speech signal is parameterized into a sequence of feature vectors; and decoding in which the most likely word sequence is hypothesized based on the observed features.

5.2.1 Feature extraction

As mentioned above, state-of-the-art speech recognizers do not perform directly on the speech signal itself, but on a set of feature vectors extracted from the speech signal. In our speech recognition system, the speech signal is first parameterized into a sequence of *mel-frequency cepstral coefficient* (MFCC) vectors. MFCC vectors with their first and second temporal derivatives are then composed into a input vector for the speech recognition recognizer.

The MFCC coefficients are computed in a computationally efficient way, described in the following. The input speech signal is first divided into a sequence of short overlapping frames. Each frame is windowed and then transformed to the frequency domain using a STFT. The magnitude squared of the STFT is computed and then multiplied by a series of overlapping triangular weighting functions. These triangular filter are equally distributed along the mel frequency scale with a 50% overlap between consecutive triangles. These filters are spaced in frequency approximately linearly at low frequencies and logarithmically at high frequencies. MFCC of each frame is computed as a vector whose components represent the energy in each of the mel filters. To approximate human auditory processing more closely, the natural logarithm of each of the elements in the mel spectral vector is then computed, producing the log mel spectrum of the frame. Finally, this vector is converted to mel-frequency cepstra via a *discrete cosine transform* (DCT) and then truncated. This feature extraction procedure is also shown in Fig. 5.1.

5.2.2 Decoding

State-of-the-art speech recognition systems are *Hidden Markov Model* (HMM) based systems, in which each acoustic unite (e.g., word, phoneme) is modeled as an HMM [127]. An HMM can be characterized by:

1. a finite number of states

2. a state-transition probability distribution which describes the probability associated with moving to another state (or the same state) at the next time instant, given the current state

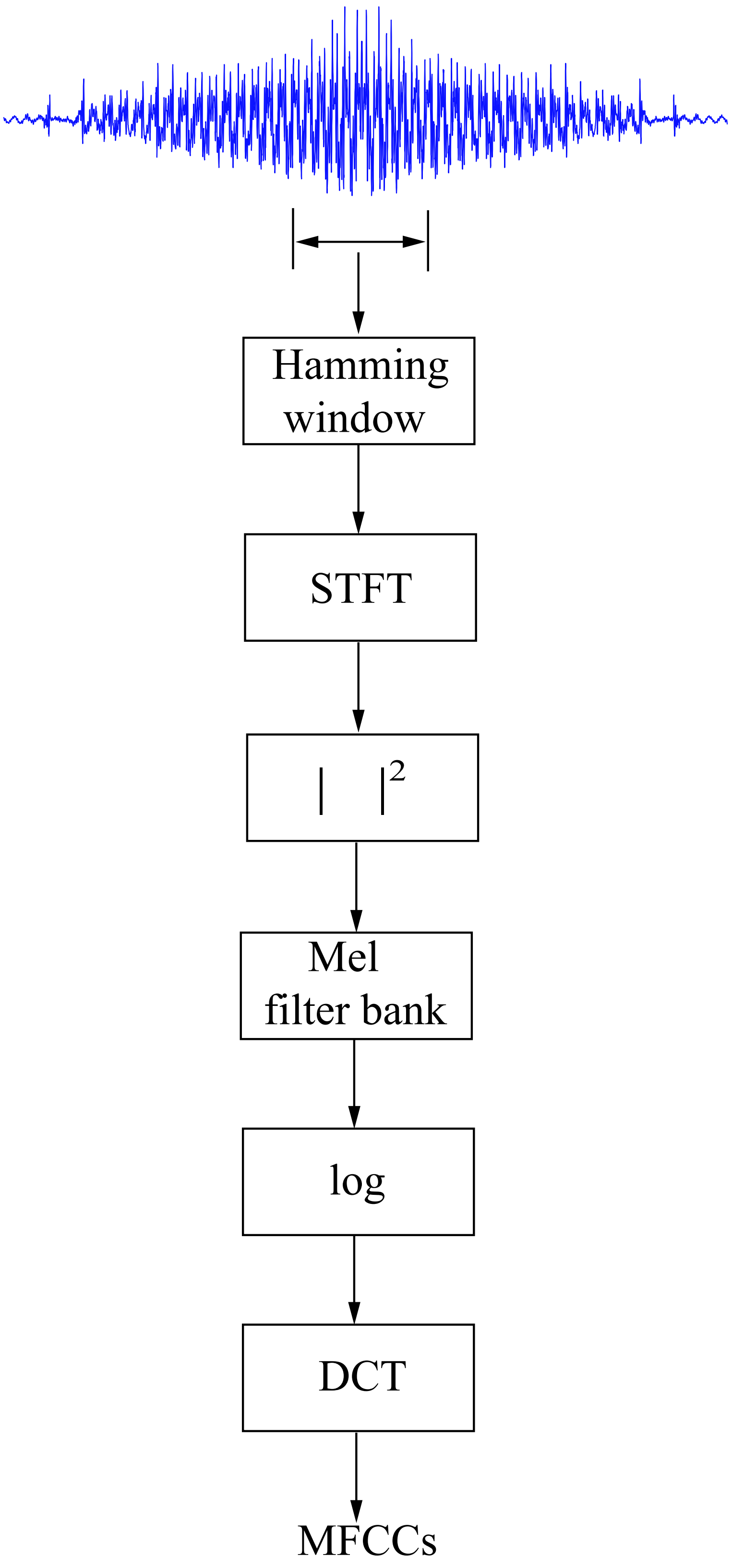


Figure 5.1: Calculation of mel-frequency cepstral coefficients from a frame of speech.

3. an output probability distribution function associated with each state

The statistical behavior of an HMM representing a given word is governed by its state transition probabilities and the output distributions of its consecutive states. For an HMM modeling word $\boldsymbol{w}$, the transition probabilities are represented by a transition matrix, A_w. The elements of this matrix, $a_w(\imath, \jmath)$, represent the probability of transiting to state $\jmath$ at time $t+1$ given that state $\imath$ is occupied time t. Thus, if the HMM for word w has Q states, then:

$$\sum_{\jmath=1}^{Q} a_w(\imath, \jmath) = 1. \tag{5.4}$$

In speech recognition system, the state output probability distribution functions are usually modeled as mixtures of Gaussians. To improve computational efficiency, the Gaussians are assumed to have diagonal covariance matrix. Thus, the output probability of a feature vector $\boldsymbol{o}$ belonging to the state $\imath$ of an HMM for word w, is represented as:

$$b_w(z, \imath) = \sum_{\rho} \alpha_{\imath\rho}^{w} \mathcal{N}\Big(o, \mu_{\imath\rho}^{w}, \sum\nolimits_{\imath\rho}^{w}\Big), \tag{5.5}$$

where $\alpha_{\imath\rho}^{w}$, $\mu_{\imath\rho}^{w}$ and $\sum_{\imath\rho}^{w}$ are the mixture weight, mean vector and covariance matrix associated with the ρ-th Gaussian in the mixture density of state $\imath$ of the HMM of word w. We define B_w as the set of parameters $\left\{\alpha_{\imath\rho}^{w}, \mu_{\imath\rho}^{w}, \sum_{\imath\rho}^{w}\right\}$ for all mixture components for all states in the HMM for word w. We can then define $\lambda_w = (A_w, B_w)$ as the complete set of statistical parameters that define the HMM for word w.

The probability, that a given sequence of feature vector $\mathcal{O} = \{o_1, o_2, \ldots, o_T\}$ given the HMM for the word w, is computed as follows. Let the HMM modelling the word w as HMM_w and $\mathcal{S}$ denote the set of all possible state sequences of length Q through HMM_w, the total probability that HMM_w generated $\mathcal{O}$ can be expressed as:

$$P(\mathcal{O}|w) = \sum_{s \in \mathcal{S}} P(\mathcal{O}|\boldsymbol{s}) P(\boldsymbol{s}|w), \tag{5.6}$$

where $\boldsymbol{s} = \{s_1, s_2, \ldots, s_T\}$ represents a particular state sequence through HMM_w. The expression $P(\boldsymbol{s}|w)$ represents the probability of a particular state sequence and is computed from the state transition matrix A_w. The expression $P(\mathcal{O}|\boldsymbol{s})$ represents the probability of a particular sequence of feature vectors given a state sequence, and is computed from the state output probability distributions with Eq. (5.6). Therefore, Eq. (5.6) can be rewritten as:

$$P(\mathcal{O}|w) = \sum_{s \in \mathcal{S}} \left(\prod_{t=1}^{Q} a_w(s_t, s_{t+1}) \right) \left(\prod_{t=1}^{Q} b_w(o_t, s_t) \right). \tag{5.7}$$

Substituting Eq. (5.7) into Eq. (5.3), we can get the expression used to perform speech recognition as:

$$\hat{w} = \arg\max_{w} \left\{ P(w) \sum_{s\in\mathcal{S}} \left(\prod_{t=1}^{Q} a_w(s_t, s_{t+1}) \right) \left(\prod_{t=1}^{Q} b_w(s_t, s_t) \right) \right\}. \tag{5.8}$$

For computational efficiency, most HMM speech recognition systems estimate the best state sequence (the state sequence with the highest likelihood) associated with the estimated hypothesis. Thus, recognition is actually performed as:

$$\hat{w} = \arg\max_{w, s\in\mathcal{S}} \left\{ P(w) \left(\prod_{t=1}^{Q} a_w(s_t, s_{t+1}) \right) \left(\prod_{t=1}^{Q} b_w(s_t, s_t) \right) \right\}. \tag{5.9}$$

5.3 Speech recognition experiments

Since the final objective of this research is to improve speech recognition performance in adverse environments using microphone array and post-filtering. Therefore, the final evaluation measure of the noise reduction algorithms is recognition performance. In the following, we will examine the performance of the proposed noise reduction algorithms in term of recognition accuracy in various noise environments.

To evaluate the effect of the proposed noise reduction system on the performance improvement of a speech recognition system, it is used as a front-end processor for the speech recognition system which performs in practical noisy environments. In implementation of the proposed noise reduction system, the original subtractive beamformer based technique is used to suppress the localized noise components, instead of the generalized subtractive beamformer presented in chapter 3. The use of the system, the generalized subtractive beamformer followed by the hybrid Wiener filter is one of our future work, as pointed out in chapter 6. The noise reduction system is first performed on the multi-channel observed noisy input signals, yield the enhanced speech signals. The enhanced speech outputs are further input into the speech recognition system for recognizing the utterance. Thus, the performance improvement caused by the noise reduction systems is evaluated based on the recognition rate by comparing with that obtained by using noisy inputs and by other noise reduction algorithms. The architecture of speech recognition systems with noise reduction front-end processor is plotted in Fig. 5.2.

5.3.1 Experimental configuration

To assess the performance of the studied noise reduction algorithms in terms of speech recognition performance, we did comprehensive speech recognition experiments.

In the experiments, the noise signals are recorded using a three-sensor microphone array in car environment (100km/h), same as those used in chapters 3 and 4.

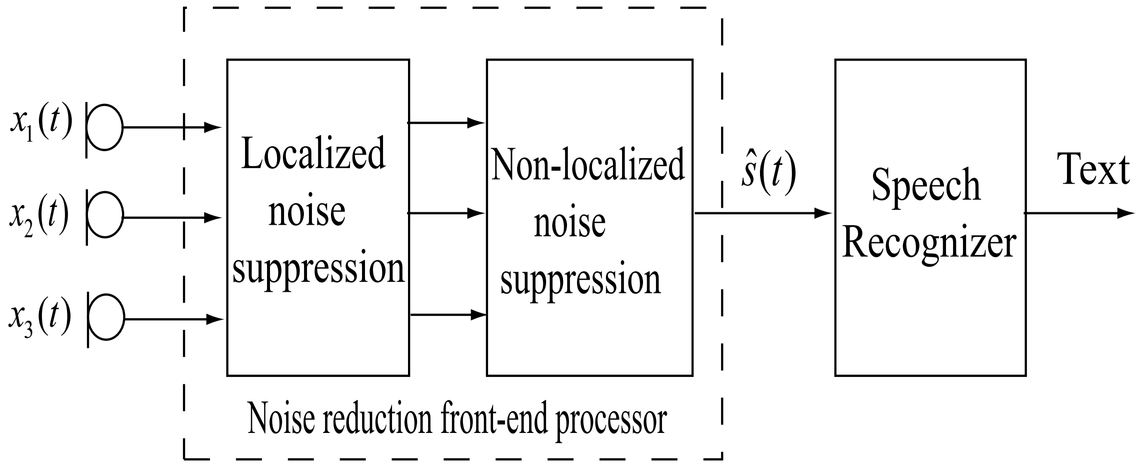

$x_1(t)$
$x_2(t)$
$x_3(t)$
Localized noise suppression
Non-localized noise suppression
$\hat{s}(t)$
Speech Recognizer
Text
Noise reduction front-end processor

Figure 5.2: Block diagram of the speech recognition system with a front-end processor of the proposed noise reduction algorithm.

The training and testing speech signals are selected from AURORA-2J database. AURORA-2J [165] is a Japanese version of AURORA-2 which is a digit strings database. The pronunciations of eleven digits are shown in Table 5.1.

For testing we generated two sets of noise-corrupted data. The first data set, referred to as **data set A**, involved the addition of the randomly selected segments of the multi-channel car noise recordings across 1001 test sentences in AURORA-2J at different SNR levels from 0dB to 20dB in 5dB steps. The second data set, referred to as **data set B**, involved the addition of the multi-channel car noise and a secondary speaker's speech (passenger's interfering noise), which is Japanese digit /ichi/ with DOA of 60 degrees to the right, across 1001 test sentences in AURORA-2J at different SNR levels same as that in **data set A**. Note that data **data set B** corresponds to a more realistic context for a typical car environment where a passenger is speaking. Moreover, it should be noted that since clean speech signals do not include the impluse responses between the speaker and the microphones, therefore, **data set A** and **data set B** correspond to the *pseudo real-world environment*, as considered in the last two chapters. Furthermore, it should be noted that the proposed noise reduction algorithm was proven to consistently outperform other traditional algorithms for both *pseudo real-world environment* and *real-world environment* in chapters 3 and 4. Therefore, it is believed that the experiments using the computer-synthesized data corresponding to *pseudo real-world environment* are enough to confirm the superiority of the proposed noise reduction system for *real-world environment* in the sense of speech recognition performance improvement. For training acoustic model, total (84000) utterances uttered by 110 speakers (55 male and 55 female speakers) are used.

The signals are pre-emphasized with a coefficient of 0.97. A hamming window of 32ms length with 16ms frame rate was used. The first 12 dimensions of de-correlated log compressed Mel energy spectrum was chosen (the zero-th order coefficient was discarded).

Table 5.1: Pronunciations of digits

Digit	AURORA-2	AURORA-2J
1	one	/ichi/
2	two	/ni/
3	three	/sanN/
4	four	/yoN/
5	five	/go/
6	six	/roku/
7	seven	/nana/
8	eight	/hachi/
9	nine	/kyuH/
0(Z)	zero	/zero/
0(O)	oh	/maru/

Combining with the log power energy, we got 13 dimensional static feature vector. Together with their first and second order dynamic values, 39 dimensional feature vectors were formed. The acoustic models consist of ten digits, one silence and short pause models. Each distribution of digit has 18 states with 16 output distributions. Silence model has 5 states with 3 distributions, and short pause model has 3 states with one distribution. Each distribution of digit has 20 Gaussians while that of silence and short pause has 36 Gaussians. Each model was trained as a left-to-right topology with three states (without skip among states) by using Baum-Welch algorithm with a flat-starting embedded training. Standard Viterbi decoding technique was used for recognition. The specification of the speech recognition system is listed in Table 5.2.

5.3.2 Experimental results

The frond-end processors, including *delay-and-sum beamforme with Wiener postfilter* (DSWF) [137], the *microphone array with OM-LSA based post-filtering* (MA-LSA) [88] and the *proposed noise reduction algorithm with microphone array and post-filtering* (PROMAPF), are assessed using two testing data sets: **data set A** and **data set B**. The recognition results for three noise reduction algorithms in two noise conditions are presented in Fig. 5.3 and Fig. 5.4, respectively.

As Fig. 5.3 shows, for **data set A**, all tested noise reduction algorithms provide some degree of performance improvement in speech recognition rate compared with noisy inputs. The average recognition rate improvement achieved by DSWF algorithm amounts to 6.0% with respect to noisy inputs. The MA-LSA provides an average recognition

Table 5.2: Specification of the speech recognition system

Parameter	Value
sampling frequency	12kHz
frame length	42.6ms
frame period	21.3ms
pre-emphasis	1-0.97 z^{-1}
feature vector	12-order MFCCs, power, ΔMFCCs, Δpower, ΔΔMFCCs, ΔΔpower.
HMM	digit: 18 states with 16 distributions; silence: 5 states with 3 distributions; pause: 3 states with 1 distributions. each distribution of digit has 20 Gaussians; each distribution of silence and pause has 36 Gaussians.
Training data	8440 utterances from AURORA-2J database
Testing data	Set A: 1001 utterances corrupted by multi-channel car noises; Set B: 1001 utterance corrupted by multi-channel car noises and passenger's speech (localized noise).

rate improvement of about 13.6%. Whereas, the highest recognition rate improvement of about 18.6% was achieved by the algorithm PRO-MAPF. The recognition rate improvements drastically increase as the noise level increase (SNR decreases). Moreover, in very high SNR conditions, all the tested algorithms provide just slight performance improvement compared with the noisy inputs, which is reasonable since the inputs are "clean" enough and a relatively high recognition rate is achievable in these conditions. In comparison of the algorithm MA-LSA, the algorithm PRO-MAPF provides much higher speech recognition rate in all noise conditions. This superiority is caused by the low speech distortion introduced by the algorithm PRO-MAPF with regard to MA-LSA, although the algorithm MA-LSA was proven to be able to improve the speech quality in subjective evaluations [88].

The recognition results for **data set B** are shown in Fig. 5.4. Concerning the recognition results shown in Fig. 5.4, we can observe that PRO-MAPF also demonstrates highest recognition rate at all SNRs. In this noise condition where the passenger's speech is regarded as localized interfering noise, the recognition rate goes down greatly for unprocessed noisy testing data. Recognition rate improvements of 11.5%, 16.8% and 23.2% were demonstrated by the DSWF, MA-LSA and PRO-MAPF algorithms, respectively. The highest recognition rate of PRO-MAPF can be attributed to the fact that it is successful in dealing with both passenger's interfering speech and diffuse car noises simultaneously with minimum speech distortion, resulting in the higher speech recognition rate.

5.3.3 Discussions

It should be noted that the experimental conditions are slightly different from the real-world environments. For example, in the practical environments, the desired speech signal is corrupted by reverberant noises, especially in a large reverberant room (e.g., conference room). Whereas, the reverberation was disregarded in the experimental conditions. However, it is of interest to note that in the experimental conditions the distance between the speaker and the microphone array is about 50 cm. In this situation, the speech sound via direct propagation path is greatly strong than those via multiple reflected paths. Actually, the reverberation time in a small or mid-size car is below 200 ms, usually about 100 ms. Therefore, the effect of reverberation is very small such that the results in the experimental conditions are reliable in real car environments. Moreover, the results presented in the previous chapters also demonstrated that the proposed noise reduction system yielded the same noise reduction performance in the *pseudo real-world environment* and *real-world environment*.

Based on the speech recognition results presented in this chapter, the superiorities of the proposed noise reduction system (PRO-MAPF) using microphone array and post-filtering are discussed with regard to the other comparative algorithms (DSWF and MA-LSA) in the following.

All studied noise reduction algorithms provide the improvement of recognition accuracy in all tested noise conditions, especially in high noise conditions. In the very low noise conditions (high SNR conditions), all signals (input signals and enhanced signals) are relatively clean such that the obtained recognition results are relatively high. However in the very high noise conditions (low SNR conditions), the input signals are strongly corrupted by interfering noise signals, resulting in low recognition rates. The enhanced signals by the tested algorithms show relative quality improvement, resulting in the improvement of recognition rate.

Furthermore, the PRO-MAPF algorithm is superior to the DSWF algorithm. For the DS beamformer, only very limited noise reduction performance for localized noises can be achieved due to the small physical size (three sensors). Whereas, in the PRO-MAPF algorithm, the hybrid noise estimation technique based noise reduction algorithm could reduce most localized noise components, even all localized noise components in theory, yielding infinite noise reduction performance in a perfect coherent noise field. Concerning the post-filter part, the traditional Wiener filter used in the DSWF algorithm only might give low noise reduction performance because of the involved correlated noise components, especially in the low frequency region. Whereas, in the PRO-MAPF algorithm, the post-filter fully considers and utilizes the spatial correlation characteristics of the noise field, forming a hybrid algorithm which is a combination of a modified Zelinski post-filter in

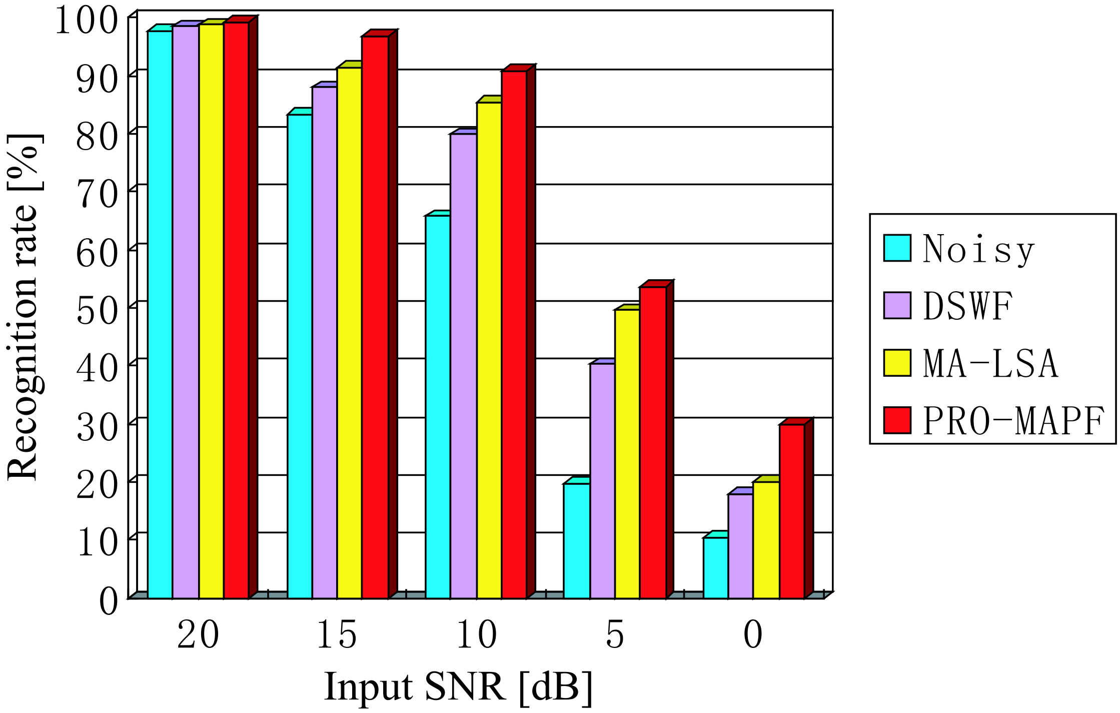


Figure 5.3: Speech recognition results for the **data set A** in which speech signals are corrupted by car noise.

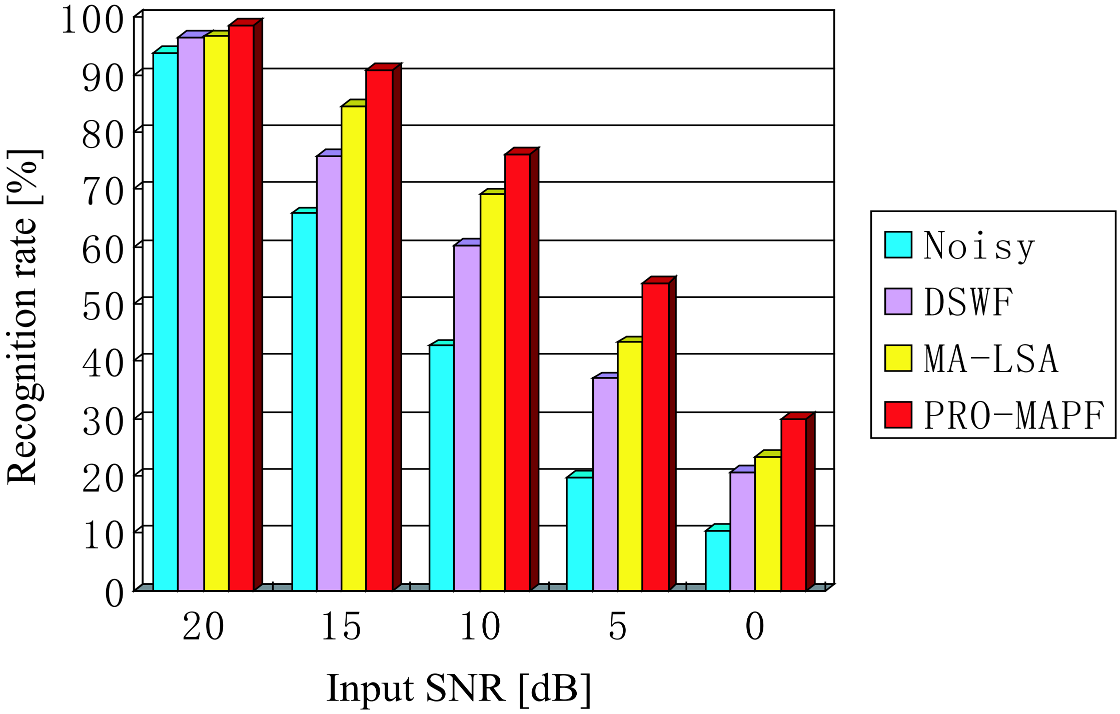


Figure 5.4: Speech recognition results for the data **data set B** in which speech signals are corrupted by car noise and passenger's interfering voice (localized interfering noise).

the high frequency region and a single-channel Wiener filter in the low frequency region. This hybrid post-filter suppresses the non-localized noise components, resulting in the high recognition results as demonstrated in Figs. 5.3 and 5.4.

Moreover, the PRO-MAPF algorithm is superior to the MA-LSA algorithm. For suppressing the localized noise components, both MA-LSA and PRO-MAPF algorithms exploited same noise reduction mechanism. While concerning the post-filter for suppressing non-localized noise components, the MA-LSA algorithm considers a OM-LSA estimator based post-filter. The improved speech quality of the enhanced signals processed by this OM-LSA based method, however, some sensitive implementation parameters are not easy to determine and the non-suitable parameters further deteriorate the speech recognition performance in practical environments. Whereas, the PRO-MAPF algorithm avoids the implementation problems and provides a robust solution for implementing the post-filter to suppress the non-localized noise components. It shows high performance in reducing non-localized noises, which in turn improves the recognition results in practical environments, as demonstrated in Figs. 5.3 and 5.4.

5.4 Summary

In this chapter, we reported the speech recognition results when the studied noise reduction algorithms are used as front-end processors for a speech recognizer.

In section 5.1, we presented the necessity of the speech recognition experiments. Note that the objective of this research is to design noise reduction systems for robust speech recognition in adverse environments.

In section 5.2, we briefly described the basic mechanism of the automatic speech recognition systems, which includes two stages: *feature extraction* and *decoding.* In the feature extraction stage, feature vectors which are then to be recognized by recognition system are extracted from each segments of speech signals. In the decoding stage, the most possible candidate word (word sequence) is determined according some certain criteria.

In section 5.3, we reported the speech recognition results in two noise conditions: multi-channel noises recorded in car environments , and multi-channel car noise and interfering passenger's voice. Compared with other traditional algorithms (DSWF and MA-LSA), the PRO-MAPF algorithm showed highest speech recognition results in all noise conditions. This can be attributed to the fact that the PRO-MAPF algorithm suppresses both localized and non-localized noise components with minimum speech distortion in all considered conditions.

Chapter 6

Conclusions and further research

In this chapter, we summarize the main conclusions of this thesis and provide some suggestions for further research.

6.1 Conclusions

Hands-free technology has demonstrated substantial superiority and desirability for a large variety of speech applications, such as automatic speech recognition systems, due to the convenience and flexibility it provides. In hands-free applications, the received signals on distant microphones are drastically corrupted by acoustic noise, reverberation and acoustic echo signals. These undesirable signals severely deteriorate the recognition performance of hands-free speech recognition systems. To improve the performance and robustness of hands-free speech recognition systems, noise reduction algorithms, as front-end processors for recognition systems, are called for.

In this thesis, main attention is paid to deal with the recognition performance decrease of hands-free speech recognition systems caused by acoustic background noises. For other acoustic disturbance signals (e.g., reverberation and acoustic echo), some suggestions are also presented as the further research work. Therefore, in this sense, the objective of this research was determined to suppress acoustic background noise signals with the goal of improving the performance of hands-free speech recognition systems in adverse environments.

To design an effective noise reduction system, in this thesis, we first deeply considered and analyzed the characteristics of the signals (speech and noise) and acoustic environment where speech recognition systems perform. We proposed that the acoustic noise signals can be decomposed into *localized* and *non-localized* noise components. Considering the spatially directional characteristics of localized noises, a subtractive beamformer based noise estimation-reduction algorithm using a microphone array was presented because of its high performance in reducing various kinds of coherent noises (especially sudden noise). Its performance is further improved by a novel hybrid noise estimation technique which combines the subtractive beamformer based estimation approach and a soft-decision based estimation approach. Furthermore, this combination is reinforced by integrating a *robust and accurate speech absence probability* (RA-SAP) estimator which considers the strong correlations of speech presence uncertainty between adjacent frequency bins and consecutive frames. Moreover, theoretical analysis results showed that the subtractive beamformer we implemented is a *minimum variance distortionless responds* (MVDR) beamformer. Thus, localized noise components are successfully suppressed by this localized noise suppression algorithm with a microphone array in which a hybrid noise estimation technique is followed by non-linear spectral subtraction. Subsequently, the microphone array (e.g., subtractive beamformer) outputs are further processed by a hybrid post-filter to reduce non-localized noise components. In the hybrid post-filter, considering the spatial coherence characteristics of noises on different microphone pairs, in the high frequencies, we presented and applied a modified Zelinski post-filter which is estimated using the signals on the microphone pairs on which noises are low-correlated; in the low frequencies, we applied a single-channel Wiener filter in which the *a priori* SNR is updated in a decision-directed mechanism due to its ability in reducing "musical noise". Moreover, theoretical analysis results showed that the proposed hybrid post-filter is a Wiener filter in theory. As a result, the proposed noise reduction system based on microphone array and post-filter can be decomposed into a MVDR beamformer followed by a single-channel Wiener filter, which is an optimal solution (e.g., multi-channel Wiener filter) to the problem of minimizing the mean square error between desired speech signal and its estimate for broad-band inputs.

Compared with other traditional noise reduction algorithms, the propose noise reduction algorithm has the following advantages: (1) in theory, it provides the optimal solution to the problem of multi-channel noise reduction for broad-band inputs in MMSE sense; (2) it should yield higher performance in dealing with various kinds (localized and non-localized) of noise signals due to the use of all available (temporal, spectral and spatial) information of signals and acoustic environments; (3) it is data-dependent and adaptive algorithm, making it to be able to deal with time-varying noise signals and adapt to time-varying acoustic environments; (4) it does not exploit the adaptive signal processing techniques (e.g., LMS), avoiding the problems of the convergence rate and stability in practical environments; (5) it involves a small-size (three-microphone) microphone array, that is, it is a small-size algorithm with a low computational cost, making it preferred to many additional practical applications, such as in-car applications and hearing aids. Moreover, the success of the proposed noise reduction system in suppressing undesired acoustic noise signals results in its significant ability in improving the performance and robustness of hands-free speech recognition systems, which was verified by the comprehensive experimental results in car environments. Therefore, the proposed noise reduction algorithm satisfies the goals of this thesis — to construct a noise reduction system for improving the performance of hands-free speech recognition systems in adverse environments.

In the following, we briefly conclude the achievements we accomplished in each chapter of this thesis.

In chapter 2, we first described the signal model on which the proposed noise reduction algorithm is based, which consists of desired speech signal, localized noise signal and non-localized noise signals. Subsequently, we presented the brief overview of the proposed noise reduction algorithm, which is based on microphone array and post-filtering. The basic theoretical principle (multi-channel Wiener filter) and practical implementation procedure were described as well.

In chapter 3, we proposed a hybrid noise estimation technique based noise reduction algorithm to deal with localized noise components. The hybrid noise estimation technique combines a single-channel (soft-decision based) estimation approach and a multi-channel (subtractive beamformer based) estimation approach in a parallel way. This combination is greatly improved by a RA-SAP estimator. The much more accurate spectral estimates for localized noise are then subtracted from those of observed noisy signals. The hybrid noise estimation technique, for the first time, mitigates the problems caused by the grating sidelobes of small-size microphone arrays. Experimental results using various localized noises demonstrated that this hybrid noise estimation technique yields much more accurate spectral estimates for localized noise with regard to other estimation approach alone.

The subtractive beamformer was then extended to a generalized expression by relaxing the assumption of a perfectly coherent noise field to the one of an arbitrary noise field. The generalized subtractive beamformer have a GSC-like structure. The theoretical analysis results showed that the generalized subtractive beamformer includes the original subtractive beamformer as a special case in a perfectly coherent noise field when only two microphones are available, and that both subtractive beamformers are optimal solutions in the MMSE sense. The effectiveness and superiorities of the generalized subtractive beamformer were also verified by experimental results in various car environments.

In chapter 4, we proposed to deal with non-localized noise by applying a hybrid post-filter which was applied to the outputs of the microphone array. Considering the characteristics of a diffuse noise field, the hybrid post-filter utilizes a modified Zelinski post-filter in the high frequency region and a single-channel Wiener post-filter in the low frequency region. The theoretical superiority of this hybrid post-filter is the generalized expression for the Zelinski post-filter and the single-channel Wiener post-filter. Its practical effectiveness were also verified by experimental results in various car environments.

In chapter 5, we first introduced the basic principle of automatic speech recognition system. Subsequently, we presented the speech recognition results in car environments with the proposed noise reduction algorithm as a front-end processor. The performance of the proposed noise reduction algorithm was further compared to that of other traditional noise reduction algorithms. The proposed algorithm achieved about 20%, 12% and 5% improvements of speech recognition rate with regards to noisy inputs and the traditional DSWF algorithm and MA-LSA algorithm we suggested before.

6.2 Suggestions for further research

In this thesis, we have proposed a noise reduction algorithm which was designed using microphone array and post-filtering for robust speech recognition in adverse environments. Undesired acoustic noise signals are considered to be composed of localized noise components and non-localized noise components, which are then suppressed by a beamforming technique and a post-filtering, respectively. It has shown that the proposed noise reduction algorithm outperforms many traditional noise reduction algorithms in reducing acoustic noise signals and further improving the recognition performance of speech recognition systems by the comprehensive experiments in various car environments. However, the proposed algorithm could be further improved in the following ways.

In the current implementation, we still used the original subtractive beamformer based technique to reduce localized noise components instead of the generalized subtractive beamforfmer. In chapter 3, the generalized subtractive beamformer has proven to be a natural extension of the original subtractive beamformer in an arbitrary noise field. Therefore, to construct a noise reduction system in which the generalized subtractive beamformer is followed by the hybrid post-filter is suggested in the further research.

Furthermore, in the generalized subtractive beamformer, the input microphone signals were assumed to be perfectly time-aligned in advance, that is, the desired speech signals were assumed to come from the front of the microphone array. In the practical implementation, it is necessary to integrate the transfer functions between desired speech source and microphones into the generalized subtractive beamformer. Thus, the transfer functions should be estimated from the input signals on-line, which might can be done using the system identification techniques (e.g., based on the non-stationarity of signals [54]).

The small physical size of the microphone array used in the proposed noise reduction algorithm makes it more preferable to many applications, such as hearing aids and in-car applications. Because of the small-size microphone array, improving the robustness of the noise reduction algorithm against imperfections, such as the imperfection of microphone positions, is necessary for the real-world implementation, which is suggested as well for further research.

Moreover, in the real-world environments, the performance degradation of hands-free speech recognition systems is caused by not only acoustic background noise, but also reverberation and acoustic echoes. To further improve the performance of speech recognition systems, it is necessary to further deal with reverberation and acoustic echoes by combining the proposed noise reduction algorithm and other advanced dereverberation and echo cancellation techniques. This is an essential and still challenging work for implementing a high-performance speech recognition system in adverse environments.

Finally, the proposed noise reduction was designed and proved to be preferred for improving the performance of speech recognition systems in adverse environments. In addition to speech recognition systems, this proposed algorithm is also very preferable to many other speech applications, such as hearing aids and speech communication systems (e.g., mobile phone) because of its high performance in reducing various kinds of noise signals (especially sudden noise), low computational cost and small physical size. Therefore, the proposed noise reduction algorithm provides a basis/core for many practical applications. Furthermore, considering the specific characteristics of some applications, the proposed noise reduction algorithm can be improved to be an advanced noise reduction system for that specific applications, which is also promising for further research.

Appendix

Appendix A: Derivation of *Noise Reduction Performance*

For the notational simplification, we omit the frequency index k and the frame index ℓ in the following derivation.

To avoid cancellation of the desired speech signal, we calculate the optimal NC filters $\hat{\mathbf{H}}_{opt}$ when desired speech is absent, that is, $\mathbf{X} = \mathbf{N}$. Thus, Eqs. (3.36) and (3.39) can be rewritten as:

$$\begin{aligned} Y_{FBF} = &\quad \mathbf{W}^{\dagger}\mathbf{N}, & (1)\\ \mathbf{U} = &\quad \mathbf{B}^{\dagger}\mathbf{N}. & (2) \end{aligned}$$

Using Eqs. (1) and (2), the PSDs $\Phi_{\mathbf{UU}}$ and $\Phi_{\mathbf{U}Y}$ are calculated as:

$$\Phi_{\mathbf{UU}} = \mathbf{B}^{\dagger}\Phi_{\mathbf{NN}}\mathbf{B}, \tag{3}$$

and

$$\Phi_{\mathbf{U}Y} = \mathbf{B}^{\dagger}\Phi_{\mathbf{NN}}\mathbf{W}, \tag{4}$$

where $\Phi_{\mathbf{NN}} = E\left[\mathbf{N}\mathbf{N}^{\dagger}\right]$. Substituting Eqs. (3) and (4) into Eq. (3.45), the optimal NC filters $\hat{\mathbf{H}}_{opt}$ are:

$$\hat{\mathbf{H}}_{opt} = \Big(\mathbf{B}^{\dagger}\Phi_{\mathbf{NN}}\mathbf{B}\Big)^{-1}\mathbf{B}^{\dagger}\Phi_{\mathbf{NN}}\mathbf{W}. \tag{5}$$

With Eqs. (3.48) and (5), the PSD of the output signal Y_o can be given by:

$$\begin{aligned} \phi_{Y_oY_o} = &\ \mathbf{W}^{\dagger}\Phi_{\mathbf{XX}}\mathbf{W} - \mathbf{W}^{\dagger}\Phi_{\mathbf{XX}}\mathbf{B}\Big(\mathbf{B}^{\dagger}\Phi_{\mathbf{NN}}\mathbf{B}\Big)^{-1}\mathbf{B}^{\dagger}\Phi_{\mathbf{NN}}\mathbf{W} \\ & - \mathbf{W}^{\dagger}\Phi^{\dagger}_{\mathbf{NN}}\mathbf{B}\Big(\mathbf{B}^{\dagger}\Phi^{\dagger}_{\mathbf{NN}}\mathbf{B}\Big)^{-1}\mathbf{B}^{\dagger}\Phi_{\mathbf{XX}}\mathbf{W} + \mathbf{W}^{\dagger}\Phi^{\dagger}_{\mathbf{NN}}\mathbf{B} \\ & \Big(\mathbf{B}^{\dagger}\Phi^{\dagger}_{\mathbf{NN}}\mathbf{B}\Big)^{-1}\mathbf{B}^{\dagger}\Phi_{\mathbf{XX}}\mathbf{B}\Big(\mathbf{B}^{\dagger}\Phi_{\mathbf{NN}}\mathbf{B}\Big)^{-1}\mathbf{B}^{\dagger}\Phi_{\mathbf{NN}}\mathbf{W}. \end{aligned} \tag{6}$$

To determine the theoretical noise reduction performance, we consider the speech absent periods. In this case, the output PSD $\Phi_{Y_oY_o}$ reduces to:

$$\phi^{(n)}_{Y_oY_o} = \mathbf{W}^{\dagger}\Phi_{\mathbf{NN}}\mathbf{W} - \mathbf{W}^{\dagger}\Phi_{\mathbf{NN}}\mathbf{B}\Big(\mathbf{B}^{\dagger}\Phi_{\mathbf{NN}}\mathbf{B}\Big)^{-1}\mathbf{B}^{\dagger}\Phi_{\mathbf{NN}}\mathbf{W}, \tag{7}$$

Under the assumption of identical noise PSD on each microphone, $\Phi_{\mathbf{NN}}$ should be $\Phi_{\mathbf{NN}} = \phi_{NN}\mathbf{\Gamma}$, where $\mathbf{\Gamma}$ denotes the complex coherence function given by Eq. (3.52), and the PSD of input ϕ_{XX} reduces to:

$$\phi_{XX}^{(n)} = \phi_{NN}. \tag{8}$$

Using Eqs. (3.45), (3.49), (7) and (8), we can rewrite the optimal NC filters $\hat{\mathbf{H}}$ and *Noise Reduction Performance* (NR) as :

$$\hat{\mathbf{H}}_{opt} \;=\; \left(\mathbf{B}^{\dagger}\mathbf{\Gamma}\mathbf{B}\right)^{-1}\mathbf{B}^{\dagger}\mathbf{\Gamma}\mathbf{W}, \tag{9}$$

and

$$\text{NR} = \left(\mathbf{W}^{\dagger}\mathbf{\Gamma}\mathbf{W} - \mathbf{W}^{\dagger}\mathbf{\Gamma}\mathbf{B}_1\left(\mathbf{B}_1^{\dagger}\mathbf{\Gamma}\mathbf{B}_1\right)^{-1}\mathbf{B}_1^{\dagger}\mathbf{\Gamma}\mathbf{W}\right)^{-1}. \tag{10}$$

Appendix B: Derivation of the optimal NC filter for coherent noise field

Assuming only two microphones are available, the BM output is a one-channel signal, given by:

$$U(k,\ell) = \frac{1}{2}j\sin\left(2k\pi\tau\right)\left(N_1(k,\ell) - N_2(k,\ell)\right). \tag{11}$$

With the assumption of identical noise PSD on each microphone, the PSDs of $\Phi_{UU}(\omega)$ and $\Phi_{UY}(\omega)$ can be given by:

$$\Phi_{UU}(k,\ell) = \frac{1}{2}\phi_{NN}(k,\ell)\sin^2\left(2k\pi\tau\right)\left(1 - \Re\{\Gamma_{N_1N_2}(k,\ell)\}\right), \tag{12}$$

$$\Phi_{UY}(k,\ell) = \frac{1}{2}j\phi_{NN}(k,\ell)\sin\left(2k\pi\tau\right)\left(1 - \Gamma_{N_2N_1}(k,\ell)\right). \tag{13}$$

Substituting Eqs. (12) and (13) into Eq. (3.45), the optimal NC filter $\hat{H}_{opt}(\omega)$ is obtained as:

$$\hat{H}^*_{opt}(k,\ell) = \frac{-j\left(1 - \Gamma_{N_1N_2}(k,\ell)\right)}{\sin\left(2k\pi\tau\right)\left(1 - \Re\left\{\Gamma_{N_1N_2}(k,\ell)\right\}\right)}. \tag{14}$$

In a coherent noise field, substituting Eqs. (3.54) into (14), the optimal NC filter $\hat{H}_{opt}(\omega)$ in this field is obtained as:

$$\hat{H}^*_{opt}(k,\ell) = \frac{1}{e^{jk\pi\delta}\sin\left(2k\pi\tau\right)\sin\left(k\pi\delta\right)}. \tag{15}$$

Obviously, this optimal filter is exactly identical to the “weight factor” in Eq. (3.13) in our original algorithm.

Bibliography

[1] M. Akagi and M. Mizumachi. Noise reduction by paired microphones. In *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, pp. 335-338, 1997.

[2] M. Akagi, M. Mizumachi, Y. Ishimoto, and M. Unoki. Speech enhancement and segregation based on human auditory mechanisms. In *Proc. IS2000*, pp. 246-253, 2000.

[3] M. Akagi and T. Kago. Noise reduction using a small-scale microphone array in multi noise source environment. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. I 909-912, 2002.

[4] J. An and B. Champagne. GSC realisations using the two-dimensional transform-domain LMS algorithm. In *IEE Proc. - Radar, Sonar, Navig.*, vol. 141, 1994.

[5] A. A. Azirani, R. L. B. Jeannes and G. Faucon, "Speech enhancement using a Wiener filtering under signal presence uncertainty," in *European Signal Processing Conference*, pp. 971-974, 1996.

[6] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 208-211, 1979.

[7] J. Bitzer, K. U. Simmer, and K. D. Kammeyer. Multichannel noise reduction — Algorithms and Theoretical Limits —. In *Proc. European Signal Processing Conf. (EUSIPCO)*, pp. 105-108, 1998.

[8] J. Bitzer, K. D. Kammeyer, and K. U. Simmer. An alternative implementation of the superdirective beamformer", In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1-4, 1999.

[9] J. Bitzer, K. U. Simmer, and K. D. Kammeyer. Multi-microphone noise reduction by post-filter and superdirective beamformer. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 100-103, 1999.

[10] J. Bitzer, K. U. Simmer, and K. D. Kammeyer. Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 100-103, 1999.

[11] J. Bitzer and K. U. Simmer. A new noise model for designing superdirective beamformers for special applications. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 43-46, 2001.

[12] J. Bitzer, K. U. Simmer, and K. D. Kammeyer. Multi-microphone noise reduction techniques as front-end devices for speech recognition. *Speech Communication*, vol. 34, pp. 3-12, 2001.

[13] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-27, no. 2, pp. 113-120, 1979.

[14] R. Le Bouquin. Enhancement of noisy speech signals: application to mobile radio communications, *Speech Communication*, pp. 3-19, 1996.

[15] R. Le Bouquin-Jeannes and G. Faucon. Study of a voice activity detector and its influence on a noise reduction system. *Speech Communication*, pp. 245-254, 1995.

[16] R. L. Bouquin-Jeannes, A. A. Azirani and G. Faucon. Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator. *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 5, pp. 484-487, 1997.

[17] M. S. Brandstein and D. B.Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin: Springer-Verlag, 2001.

[18] K. M. Buckley. Broad-band beamforming and the generalized sidelobe canceller. *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-34, no. 5, pp. 1322-1323, 1986.

[19] D. Burshtein and G. Gannot. Speech enhancement using a mixture-maximum model. *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 6, pp. 341-350, 2002.

[20] O. Cappe. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. In *Proc. IEEE Int. Conf. on Acoustic, Speech, Signal Processsing*, vol. 2, no. 2, pp. 345-349, 1994.

[21] G.C. Carter. Coherence and time delay estimation. *In Proc. of the IEEE*, vol. 75, no. 2, pp. 236-255, 1987.

[22] J. H. Chang and N. S. Kim. Speech enhancement: new approaches to soft decision. *IEICE Trans. Information and System*, vol. E84-D, no. 8, pp. 1231- 1239, 2001.

[23] Y. H. Chen and H. D. Fang. Frequency-domain implementation of Griffiths-Jim adaptive beamformers. *J. Acoust. Soc. Am* 91(6), pp. 3354-3366, 1992.

[24] J. Chen, L. Shue, K. Phua, and H. Sun. Theoretical comparison of dual microphone systems. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. IV-73 - IV-75, 2004.

[25] Z. Cheng and T. T. Tjhung. A new time delay estimator based on ETDE. *IEEE Trans. on Signal Processing*, vol. 51, no. 7, pp. 1859-1869, 2003.

[26] J. T. Chien and P. Y. Lai. Car speech enhancement using a microphone array. *Int. Jouranl of Speech Technolgoy*, pp. 79-91, 2005.

[27] Y. D. Cho and A. Kondoz. Analysis and improvement of a statistical model-based voice activity detector. *IEEE Signal Processing Letters*, vol. 8, no. 10, pp. 276-278, 2001.

[28] I. Cohen and B. Berdugo. Speech enhancement for non-stationary noise environments. *Signal Processing*, vol. 81, pp. 2403-2418, 2001.

[29] I. Cohen and B. Berdugo. Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12-15, Januaray 2002.

[30] I. Cohen. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113-116, 2002.

[31] I. Cohen. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 466-475, 2003,.

[32] I. Cohen. Analysis of two-channel generalized sidelobe canceller (GSC) with post-filtering. *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 684-699, 2003.

[33] I. Cohen, S. Gannot, and B. Berdugo. An integrated real-time beamforming and postfiltering system for non-stationary noise environments. *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1064-1073, 2003.

[34] I. Cohen. Multi-channel post-filtering in non-stationary noise environments. *IEEE Trans. on Signal Processing*, vol. 52, no. 5, pp. 1149-1160, 2004.

[35] I. Cohen. Speech enhancement using a noncausal a priori SNR estimatior. *IEEE Signal Processing Letters*, vol. 11, no. 9, pp. 725-728, 2004.

[36] D. V. Compernolle. Switching adaptive filters for enhancing noisy and reverberant speech from microphne array recordings. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 833-836, 1990.

[37] H. Cox. Spatial correlation in arbitrary noise fields with application to ambient sea noise. *J. Acoust. Soc. Am.*, vol. 54, no. 5, pp. 1289-1301, 1973.

[38] H. Cox, R. M. Zeskind, and T. Kooij. Pratical supergain. *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-34, no. 3, pp. 393-398, 1986.

[39] H. Cox, R. M. Zeskind and M. M. Owen. Robust adaptive beamforming. *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-35, no. 10, pp. 1365-1375, 1987.

[40] S. Doclo. *Multi-microphone noise reduction and dereverberation techniques for speech applications.* PhD thesis, Faculty of Engineering, K.U. Leuven, Belgium, May 2003.

[41] S. Dupont and C. Ris. Assessing local noise level estimation methods. In *Workshop on Robust Methods For Speech Recognition in Adverse Conditions*, pp. 115-118, 1999.

[42] G. W. Elko. Microphone Array systems for hands-free telecommunication. *Speech Communication.* vol.20, no.3-4, pp. 229-240, 1996.

[43] Y. Ephraim and D. Malah. Speech enhancement using minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109-1121, 1984.

[44] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoustic, Speech and Signal Processing*, vol. ASSP-33, no. 2, pp. 443-445, 1985.

[45] Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 251-266, July 1995.

[46] N. W. D. Evans and J. S. Mason. An assessment of local non-linear spectral subtraction for remote speech recognition. In *Proc. of 1st meeting on Speech Technology*, 2000.

[47] M. Feng and K. D. Kammeyer. Modified subspace smoothing technique for multipath direction finding. In *Proc. European Signal Processing Conference (EUSIPCO)*, pp. 181-184, 1998.

[48] S. Fischer and K. U. Simmer. An adaptive microphone array for hands-free communication. In *Proc. Int. Workshop on Acoustic Echo and Noise Control, (IWAENC)*, pp. 44-47, 1995.

[49] S. Fischer and K. U. Simmer. Beamforming microphone arrays for speech acquisition in noisy environments. *Speech Communication*, pp. 215-227, 1996.

[50] S. Fischer and K. D. Kammeyer. Broadband breamforming with adaptive postfiltering for speech acquisition in noisy environments. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21-24, 1997.

[51] J. L. Flanagan, A. C. Surendran and E. E. Jan. Spatially selective sound capture for speech and audio processing. *Speech Communication* 13(1993), pp. 207-222, 1993.

[52] O. L. Frost. An algorithm for linearly constrained adaptive array processing. In *Proc. of the IEEE*, vol. 60, no. 8, pp. 926-935, 1972.

[53] S. Gannot, D. Burshtein, and E.Weinstein. Iterative and sequential Kalman filter-based speech enhancement algorithm. *IEEE Trans. Speech, Audio Processing*, vol. 6, no. 4, pp. 373C385, July 1998.

[54] S. Gannot, D. Burshtein, and E. Weinstein. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Trans. on Singla Processing*, vol. 49, no. 8, pp. 1614-1626, 2001.

[55] S. Gannot and I. Cohen. Speech enhanced based on the General Transfer Function GSC and postfiltering. *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 12, no. 6, pp. 561-571, 2004.

[56] S. Gannot, D. Burshtein, and E. Weinstein. Analysis of the power spectral deviatioin of the general transfer function GSC. *IEEE Trans. on Signal Processing*, vol. 52, no. 4, pp. 1115-1120, 200?

[57] S. Gazor and Y. Grenier, "Criteria for positioning of sensors for a microphone array", *IEEE Trans. on Speech Audio Processing*, vol. 3, no. 4, pp. 294-303, 1995.

[58] S. Gazor and W. Zhang. A soft voice activity detector based on a Laplacian-Gaussian model. *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 498-505, 2003.

[59] S. Gazor and W. Zhang. Speech enhancement employing Laplacian – Gaussian mixture. *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 896- 904, 2005.

[60] M. V. Greening and J. E. Perkins. Adaptive beamforming for nonstationary arrays. *J. Acoust. Soc. Am.* 112(6), pp. 2872-2881, 2002.

[61] Y. Grenier. A microphone array for car environments. *Speech Communication* vol. 12, pp. 25-39, 1993.

[62] , L. J. Griffiths and C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. on Antennas and Progagation*, vol. AP-30, no. 1, pp. 27-34, 1982.

[63] M. M. Goulding and J. S. Bird. Speech enhancement for mobile telephony. *IEEE Trans. on Vehicular Technolgoy*, vol. 39, no. 4, 1990.

[64] H. Gustafsson, S. V. Nordholm, and I. Claesson. Spectral subtraction using reduced delay convolution and adaptive averaging. *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 8, pp. 799-807, 2001.

[65] M. T. Hanna and M. Simaan. Absolutely optimum array filters for sensor arrays. *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-33, no. 6, pp. 1380-1385, 1985.

[66] J. H. L. Hansen and B. L. Pellon. An effective quality evaluation protocol for speech enhancement algorithms. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pp. 2819-2822, 1998.

[67] H. G. Hirsch and C. Ehrlicher. Noise estimation techniques for robust speech recognition. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 153-156, 1995.

[68] Y. Hioka, N. Hamada. DOA estimation of speech signal using microphones located at vertices of equilateral triangle. *IEICE Trans. Fundamentals*, vol. E87-A, no. 3, pp. 559-566, 2004.

[69] M. W. Hoffman, Z. Li, and D. Khataniar. GSC-based spatial voice activity detection for enhanced speech coding in the presence of computing speech. *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 2, pp. 175-178, 2001.

[70] Y. Hu and P. Loizou. Speech enhancement by wavelet thresholding the multitaper spectrum. *IEEE Transactions on Speech and Audio Processing*, vol12, no. 1, pp. 59-67, 2004.

[71] Y. Hu, M. Bhatnagar, and P. C. Loizou. A cross-correlation technique for enhancing speech corrupted with correlated noise. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 673-676, 2001.

[72] H. T. Hu, F. J. Kuo, and H. J. Wang. Supplementary schemes to spectral subtraction for speech enhancement. *Speech Communication* 36(2002), pp. 205-218, 2002.

[73] M. Z. Ikram and D. R. Morgan. Permutation inconststency in blind speech separation: investigation and solution. *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 1, pp. 1-12, January, 2005.

[74] M. Kajala and M. Hamalainen. Filter-and-sum beamformer with adjustable filter characteristics. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2917-2920, 2001.

[75] M. Kallinger, K. D. Kammeyer, and J. Bitzer. Multi-microphone residual echo estimation. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2003.

[76] Y. Kaneda and M. Tohyama. Noise suppression signal processing using 2-point received signals. *Electronics and Communications in Japan*, vol. 67-A, no. 12, 1984.

[77] M. Kato, A. Sugiyama, and M. Serizawa. Noise suppression with high speech quality based on weighted noise estimation and MMSE STSA. *IEICE Trans. Fundamentals*, vol. E85-A, no. 7, pp. 1710-1717, 2002.

[78] A. Kawamura, K. Fujii, Y. Itoh, and Y. Fukui. A nwe noise reduction method using estimated noise spectrum. *IEICE Trans. Fundamentals*, vol. E85-A, no. 4, pp. 784-788, 2002.

[79] K. 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. 2151-2157, 1996.

[80] N. S. Kim and J. H. Chang. Speech enhancement based on global soft decision. *IEEE Signal Processing Letters*, vol. 7, no. 5, pp. 108-110, 2000.

[81] M. Kompis and N. Dillier. Performance of an adaptive beamforming noise reduction scheme for hearing aid applications. I. prediction of the signal-to-noise-ratio improvement. *J. Acoust. Soc. Am.* 109(3), pp. 1123-1133, 2001.

[82] M. Kompis and N. Dillier. Performance of an adaptive beamforming noise reduction scheme for hearing aid applications. II. experimental verification of the predictioins. *J. Acoust. Soc. Am.* 109(3), pp. 1134-1143, 2001.

[83] H. Krim and M. Viberg. Two decades of array signal processing research. *IEEE Signal Processing Magazine*, pp. 67-94, 1996.

[84] H. Kwon, J. Son, and K. Bae. Microphone array with minimum mean-square error short-time spectral amplitude estimator for speech enhancement. *IEICE Trans. Fundamentals*, vol. E87-A, no. 6, pp. 1491-1494, 2004.

[85] J. Li and M. Akagi. Noise reduction using hybrid noise estimation techniques and post-filtering. In *ICSLP2004 - 8th Int. Conf. on Spoken Language Processing*, pp. IV 2705-2708, September 2004.

[86] J. Li and M. Akagi. Suppressing localized and non-localized noises in arbitrary noise environments. In *Joint Workshop on Hands-free Speech Communication and Microphone Array*, March, 2005.

[87] J. Li, X. Lu and M. Akagi. A noise reduction system in arbitrary noise environments and its applications to speech enhancement and speech recognition. In *ICASSP2005 - IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, pp. III 277- 280, March, 2005.

[88] J. Li and M. Akagi. A noise reduction system based on hybrid noise estimation technique and post-filtering in arbitrary noise environments. *Speech Communication*, vol. 48, no. 2, pp. 111-126, 2006.

[89] J. Li, X. Lu and M. Akagi. Noise reduction based on microphone array and post-filtering for robust speech recognition in car environments. In *Workshop on DSP in Mobile and Vehicular Systems*, September 2005.

[90] J. Li and M. Akagi. A hybrid microphone array post-filter in a diffuse noise field. In *Eurospeech2005 - 9th European Conf. on Speech Communication and Technology*, pp. 2313-2316, September 2005.

[91] J. Li and M. Akagi. Theoretical analysis of microphone arrays with post-filtering for coherent and incoherent noise suppression in noisy environments. In *IWAENC2005 - Int. Workshop on Acoustic Echo and Noise Control*, pp. 85-88, September 2005.

[92] J. Li and M. Akagi. A noise reduction method based on a generalized subtractive beamformer. *To appear in Acoustical Science and Technology.*

[93] J. Li and M. Akagi. A hybrid microphone array post-filter in a diffuse noise field. *Submitted to Applied Acoustics.*

[94] J. Li and M. Akagi. Noise reduction based on microphone array and post-filtering for robust speech recognition in car environments. *DSP for In-Vehicle and Mobile Systems, Chapter 13*, Spring Press, May, 2006.

[95] J. Liu, A. B. Gershman, Z. Q. Luo, and K. M. Wong. Adaptive beamforming with sidelobe control: a second-order cone programming approach. *IEEE Signal Processing Letters*, vol. 10, no. 11, pp. 331-334, 2003.

[96] B. Logan and T. Robinson. Adaptive model-based speech enhancement. *Speech Communication* 34(2001), pp. 351-368, 2001.

[97] M. E. Lockwood, D. L. Jones, R. C. Bilger, and C. R. Lansing. Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms. *J. Acoust. Soc. Am.* 115(1), pp. 379-391, 2004.

[98] D. Mahmoudi. A microphone array for speech enhancement using multiresolution wavelet transform. In *Proc. 5th Eur. Conf. Speech, Commun. Technol.*, pp. 339-342, 1997.

[99] D. Mahmoudi and A. Drygajlo. Combined Wiener and coherence filtering in wavelet domain for microphone array speech enhancement. In *Proc. 23th IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 385-388, 1998.

[100] S. Makino, H. Sawada, R. Mukai and S. Araki. Bind source separation of convolutive mixtures of speech in Frequency domain. *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, vol.E88-A, no. 7, pp. 1640-1654, July 2005.

[101] D. Malah, R. V. Cox, and A. J. Accardi. Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 789-792, 1999.

[102] D. Mansour and B. W. Juang. The short-time modified coherence representation and noise speech recognition. *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 37, no. 6, pp. 795-804, 1989.

[103] C. Marro, Y. Mahieux and K.U. Simmer. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 3, pp. 240-259, 1998.

[104] R. Martin. An efficient algorithm to esimate the instantaneous of speech signals. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1093-1096, 1993.

[105] R. Martin. Spectral subtraction based on minimum statistics. In *Proc. European Signal Processing Conf. (EUSIPCO)*, pp.1182-1185, 1994.

[106] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. on Speech and Audio Processing*, vol. 9. no. 5, pp. 504-512, 2001.

[107] R. Martin. Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors. *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 845-856, 2005.

[108] J. Meyer and K. U. Simmer. Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21-24, 1997.

[109] R. J. Mcaulay and M. L. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-28, pp. 137-145, 1980.

[110] I. A. McCowan and S. Sridharan. Multi-channel sub-band speech recognition. *EURASIP Journal on Applied Signal Processing*, 2001(1):45-52, 2001.

[111] I. A. McCowan and S. Sridharan. Adaptive parameter compensation for robust hands-free speech recognition using a dual-beamforming microphone array. In *Proc. Int. Symposium on Intelligent Multimdida, Video and Speech Processing*, 2001.

[112] I. A. McCowan and H. Bourlard. Microphone array post-filter for diffuse noise field. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 905-908, 2002.

[113] I. A. McCowan and H. Bourlard. Microphone array post-filter based on noise field coherence. *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 709-716, 2003.

[114] M. Mizumachi and M. Akagi. Noise reduction by paired-microphones using spectral subtraction. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 1001-1004, 1998.

[115] M. Mizumachi and S. Nakamura. Noise reduction using paired-microphones on non-equally-spaced microphone arrangement. In *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, pp. 585-588, 2003.

[116] M. Mizumachi, M. Akagi, and S. Nakamura. Design of robust subtractive beamformer for noisy speech recognition. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pp. 57-60, 2000.

[117] M. Mizumachi and M. Akagi. Noise reduction method that is equipped for a robust direction finder in adverse environments. In *Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pp. 179-182, 1999.

[118] M. Mizumachi and S. Nakamura. The 2ch hybird subtractive beamformer applied to line sound sources. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pp. 1833-1836, 2002.

[119] C. E. Mokbel and G. F. A. Chollet. Automatic word recognition in cars. *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 5, pp. 346-355, 1995.

[120] D. C. Moore and I. A. McCowan. Microphone array speech recognition: experiments on overlapping speech in meetings. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 497-450, 2003.

[121] H. Nakashima, Y. Chisaki, T. Usagawa, and M. Ebata. Spectral subtraction based on statistical criteria of the speech distribution. *IEICE Trans. Fundamentals*, vol. E85-A, no. 10, pp. 2283-2291, 2002.

[122] S. E. Nordholm and Y. H. Leung. Performance limits of the broadband generalized sidelobe cancelling structure in an isotropic noise field. *J. Acoust. Soc. Am.* 107(2), pp. 1057-1060, 2000.

[123] M. Omologo, P. Svaizer, and M. Matassoni. Environmental condtions and acoutic transduction in hands-free speech recognition. *Speech Communication*, pp. 75-95, 1998.

[124] K. Paliwal and A. Basu. A speech enhancement method based on Kalman filtering. In *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 12, pp. 177C180, Apr. 1987.

[125] L. C. Parra and C. V. Alvino, Geometric source separation: merging convolutive source separation with geometric beamforming, *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, pp. 352-362, 2002.

[126] S.R. Quackenbush, T.P. Barnwell and M.A. Clements. Objective Measures of Speech Quality. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1988.

[127] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.

[128] A. Rezayee and S. Gazor. An adaptive KLT approach for speech enhancement. *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 87C95, Feb. 2001.

[129] P. Scalart and A. Benamar. A system for speech enhancement in the context of hands-free radiotelephony with combined noise reduction and acoustic echo cancellation. *Speech Communication* 20(1996), pp. 203-214, 1996.

[130] V. Schless and F. Class. SNR-dependent flooring and noise overestimation for joint application of soectral subtraction and model combination. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pp. 1495-1497, 1998.

[131] M. L. Seltzer, *Microphone array processing for robust speech recognition*, PhD thesis, Carnegie Mellon University, 2003.

[132] A. Shamsoddini and P. N. Denbigh. A sound segregation algorithm for reverberant conditionis. *Speech Communication* 33(2001) pp. 179-196, 2001.

[133] P. W. Shields and D. R. Campbell. Improvements in intelligibiligy of noisy reverberant speech using a binaural subband adaptive noise-cancellation processing scheme. *J. Acoust. Soc. Am.* 110(6), pp. 3232-3242, 2001.

[134] J. W. Shin, J. H. Chang, and N. S. Kim, Statistical modeling of speech signals based on generalized Gamma distribution, *IEEE Signal Processing Letters*, vol. 12, no. 3, pp. 258-261, 2005.

[135] H. F. Silverman. Some analysis of microphone arrays for speech data acquisition. *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-35, no. 12, pp. 1699-1712, 1987.

[136] K. U. Simmer, P. Kuczynski, and A. Wasiljeff. Time delay compensation for adaptive multichannel speech enhancement system. In *Proc. ISSE-92*, 1992.

[137] K. U. Simmer and A. Wasiljeff. Adaptive microphone arrays for noise suppression in the frequency domain. In *Proc. Workshop on Adaptive Algorithms in Communications*, pp. 185-94, 1992.

[138] B. L. Sim, Y. C. Tong, J. S. Chang and C. T. Tan. A parametric formulation of the generalized specteral subtractin method. *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 4, pp. 328-337, 1998.

[139] K. U. Simmer and J. Bitzer. Multi-microphone noise reduction — theoretical optimum and practical realization. In *Jahrestagung fuer Akustik (DAGA-2003)*, Aachen, 2003.

[140] L. Singh and S. Sridharan. Speech enhancement using critical band spectral subtraction. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pp. 2827-2830, 1998.

[141] J. Sohn, N. S. Kim, and W. Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, Janunary 1999.

[142] I. Y. Soon, S. N. Koh, and C. K. Yeo. Noisy speech enhancement using discrete cosine transform. *Speech Communication*, vol. 24, pp. 249-257, 1998.

[143] I. Y. Soon, S. N. Koh, and C. K. Yeo. Improved noise suppression filter using self-adaptive estimator of probability of speech absence. *Signal Processing*, vol. 75, pp. 151-159, 1999.

[144] I. Y. Soon and S. N. Koh. Low distortion speech enhancement. In *IEE Proc.-Vis. Image Signal Processing*, vol. 147, no. 3, pp. 247-253, 2000.

[145] A. Spriet. *Adaptive filtering techniques for noise reduction and acoustic feedback cancellation in hearing aids.* PhD thesis, Faculty of Engineering, K.U. Leuven, Belgium, September, 2004.

[146] V. Stahl, A. Fischer, and R. Bippus. Quantile based noise estimation for spectral subtraction and Wiener filtering. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1875-1878, 2000.

[147] A. Stephenne and B. Champagne. A new cepstral prefiltering technique for estimating time delay under reverberant conditions. *Signal Processing*, 59(1997), pp. 253-266, 1997.

[148] Y. Suzuki, S. Tsukui, F. Asano, R. Nishimura, and T. Sone. New design method of a Binaural microphone array using multiple constraints. *IEICE Trans. on Fundamentals*, vol. E82-A, no. 4, pp. 587-595, 1999.

[149] K. Takeda, Y. Sagisaka, S. Katagiri, M. Abe and H. Kuwabara. Speech database user's manual. ATR Interpreting Telephony Reserach Laboratories, 1988.

[150] S. G. Tanyer and H. Ozer. Voice activity detection in nonstationary noise. *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 4, pp. 478-482, 2000.

[151] E. Toner and D. R. Campbell. Speech enhancement using sub-band intermittent adaption. *Speech Communication*, 12(1993), pp. 253-259, 1993.

[152] N. Virag. Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 2, pp. 126-137, 1999.

[153] E. Visser, M. Otsuka, and T. W. Lee, A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments, *Speech Communication* 41(2003), pp. 393-407, 2003.

[154] S. A. Vorobyov, A. B. Gershman and Z. Q. Luo. Robust adaptive beamforming using worst-case performance optimization: a solution to the signal mismatch problem. *IEEE Trans. on Signal Processing*, vol. 51, no. 2, pp. 313-232, 2003.

[155] S. A. Vorobyov, A. B. Gershman, Z. Q. Luo, and N. Ma. Adaptive beamforming with joint robustness against mismatched signal steering vector and interference nonstationary. *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 108-111, 2004.

[156] D. L. Wang, J. S. Lim. The unimportance of phase in speech enhancement. *IEEE Trans. Acoust., Speech and Signal process.*, vol.30, 679-681, 1982.

[157] F. M. Wang, P. Kabal, R. P. Ramachandran, and D. O. Shaughnessy. Frequency domain adaptive postfiltering for enhancement of noisy speech. *Speech Communication* 12(1993), pp. 41-56, 1993.

[158] D. B. Ward. Technique for broadband correlated interference rejection in microphone arraya. *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 4, pp. 414-417, 1998.

[159] P. J. Wolfe and S. J. Godsill. Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement. In *Proc. 11th IEEE Workshop on Statistical Signal Processing*, pp. 496C499, 2001.

[160] Y. Wu, H. C. So, and P. C. Ching. Joint time delay and frequency estimation via state-space realization. *IEEE Signal Processing Letters*, vol. 10, no. 11, pp. 339-342, 2003.

[161] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy. Speech enhancement using linear prediction residual. *Speech Communication* 28(1999), pp. 25-42, 1999.

[162] J. Yamauchi and T. Shimamura. Noise estimation using high frequency regions for spectral subtraction. *IEICE Trans. Fundamentals*, vol. E85-A, no. 3, pp. 723-727, 2002.

[163] R. Zelinski. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2578-2581, 1988.

[164] X. X. Zhang and J. H. L. Hansen. CSA-BF: A Constrained Switched Adaptive Beamformer for Speech Enhancement and Recognition in Real Car Environments. *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 733-743, 2003.

[165] http://sp.shinshu-u.ac.jp/CENSREC.

[166] http://www.stanford.edu/ rpropper/e-rhetorics/project/index.html

Publications

[1] Junfeng Li and Masato Akagi, "Noise reduction based on microphone array and post-filtering for robust speech recognition in car environments," Chapter 13 in *Digital Signal Processing for In-Vehicle and Mobile Systems 2*, H. Abut, J.H.L. Hansen and K. Takeda (Eds.), Springer Science, New York, Scheduled for Spring 2006.

Journal papers

[2] Junfeng Li and Masato Akagi, "A hybrid microphone array post-filter in a diffuse noise field", *Submitted to Applied Acoustics.*

[3] Junfeng Li and Masato Akagi, "Noise reduction method based on generalized subtractive beamformer", *To appear in Acoustical Science and Technology.*

[4] Junfeng Li and Masato Akagi, "A noise reduction system based on hybrid noise estimation technique and post-filtering in arbitrary noise environments", *Speech Communication*, vol. 48, no. 2, pp. 111-126, 2006.

International conference papers

[5] Junfeng Li, Masato Akagi and Yôiti Suzuki, "Noise reduction based on a generalized subtractive beamformer for speech enhancement", *Submitted to The 9th Western Pacific Acoustics Conference*, June, 2006.

[6] Junfeng Li and Masato Akagi, "Theoretical analysis of microphone arrays with post-filtering for coherent and incoherent noise suppression in noisy environments", In *IWAENC2005 - Int. Workshop on Acoustic Echo and Noise Control*, pp. 85-88, September 2005.

[7] Junfeng Li and Masato Akagi, "A hybrid microphone array post-filter in a diffuse noise field", In *Eurospeech2005 - 9th European Conf. on Speech Communication and Technology*, pp. 2313-2316, September 2005.

[8] Junfeng Li, Xugang Lu and Masato Akagi, “Noise reduction based on microphone array and post-filtering for robust speech recognition in car environments”, In *Workshop on DSP in Mobile and Vehicular Systems*, September 2005.

[9] Junfeng Li, Xugang Lu and Masato Akagi, “A noise reduction system in arbitrary noise environments and its applications to speech enhancement and speech recognition”, In *ICASSP2005 - IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, pp. III 277- 280, March, 2005.

[10] Junfeng Li and Masato Akagi, “Suppressing localized and non-localized noises in arbitrary noise environments”, In *Joint Workshop on Hands-free Speech Communication and Microphone Array*, March, 2005.

[11] Junfeng Li and Masato Akagi, “Noise reduction using hybrid noise estimation techniques and post-filtering”, In *ICSLP2004 - 8th Int. Conf. on Spoken Language Processing*, pp. IV 2705-2708, September 2004.

Oral presentation

[12] Junfeng Li and Masato Akagi, “A noise reduction method based on a generalized subtractive beamformer”, *ASJ’2005 Fall Meeting*, September, 2005.

[13] Junfeng Li and Masato Akagi, “A noise reduction method based on a generalized subtractive beamformer”, *Technical Report of IEICE,* , August, 2005.

[14] Junfeng Li and Masato Akagi, “Multi-channel post-filtering in diffuse noise environment”, In *ASJ’2004 Fall Meeting*, September 2004.

[15] Junfeng Li and Masato Akagi, “A noise reduction system in localized and non-localized noise environments”, *Technical Report of IEICE, EA2004-34(2004-8)*, August, 2004.

[16] Junfeng Li and Masato Akagi, “A hybrid noise reduction method using single- and multi-channel techniques”, In *ASJ’2004 Spring meeting*, March, 2004.