Contents

List of Figures

List of Tables

List of Acronyms

AMC	Adaptive Module Controller
ANC	Adaptive Noise Cancelling
BFTF	Block Fast Transversal Filter
BSS	Blind Sources Separation
CPSP	Cross-Power Spectrum Phase
CRANC	Crosstalk Resistant Adaptive Noise Cancellation
DAS	Delay And Sum
DAWSAS	Delay And Weighted Sum And Selection
DFT	Discrete Fourier Transform
DSP	Digital Signal Processor
FBF	Fixed BeamForming
FFT	Fast Fourier Transform
GCC	Generalized Cross-Correlation
GSC	Generalized Sidelobe Canceling
GSVD	Generalized Singular Value Decomposition
HSP	Having Speech Period
ISDS	Improved Shared Distorted Signal
ISS	Improved Spectral Subtraction
LCMV	Linearly Constrained Minimum Variance
LMS	Least Mean Square
LSLL	Least Square Lattice-Ladder
MANC	Multichannel Adaptive Noise Cancellation
MCRANC	Multichannel Crosstalk Resistant Adaptive Noise Cancellation
MCRASC	Multichannel Crosstalk Resistant Adaptive Signal Cancellation
MDS	Multichannel Distorted Signals

MGSC	Modified Generalized Sidelobe Canceling
MMSE	Minimum Mean Square Error
MSR	Multiple Sampling Rates
NSP	Non Speech Period
NLMS	Normalized Least Mean Square
ONSP	Overall Non Speech Period
PDA	Personal Digital Assistant
PDS	Power Density Spectrum
PF	Post-Filtering
RLS	Recursive Least Square
SDS	Shared Distorted Signal
SE	Speech Enhancement
SNR	Signal to Noise Ratio
SS	Spectral Subtraction
VAD	Voice Activity Detector
VoIP	Voice over Internet Protocol
WPF	Weiner Post-Filtering

Chapter 1 Introduction

1.1 Speech enhancement

Speech is the most effective and most convenient tool for human communication. It plays a very important role in our daily life.

However, “we live in noisy world” [11]! Speech signals are usually degraded by noise. For example, when using a telephone, recorder, hearing aid, computer interface and many other speech tools, the desired speech signal is usually degraded by environmental noise and the apparatus internal noise. It is necessary to suppress or cancel the noise in the corrupted speech signal before we play, transfer, restore or understand it.

So-called speech enhancement aims to improve the quality and intelligibility of the degraded speech signal [11, 18]. It has very wide applications. In speech communication, the applications include, but are not limited to, hand-free telephony, mobile phone, voice over IP (VoIP), hearing aids, local and long distant telecommunications, voice-controlled machines, automatic speech recognition, speaker recognition and teleconference, etc.

However, speech enhancement is also a quite complicated and difficult objective for researchers [11]. Research work in this area began in the 1960s. Up to now a lot of work has been done and many approaches have been proposed. However, the approaches are still far from satisfactory. The problem remains largely open.

Many algorithms employ only one channel of signal for speech enhancement. They can not improve the quality and the intelligibility of the speech signal at the same time. In fact, recent research work has proved that the reduction of noise can only be achieved at the cost of speech distortion if only one microphone channel of signal is employed [11]. In other words, we can not avoid speech distortion while the noise is suppressed.

As a result, the effectiveness of single-channel speech enhancement approaches is quite limited although some of these approaches have been used in practical applications.

To improve the effectiveness of speech enhancement, one method is to employ more microphones or a microphone array. Obviously a microphone array may achieve better performance since it provides us with more than one channel of signal for processing. It not only provides us with temporal but also spatial information. In recent years it has been theoretically proved that a microphone array may suppress noise with minimal speech signal distortion [11]. Microphone array based algorithms have become a research hot spot in the speech enhancement area.

Most of the methods or algorithms for microphone array based speech enhancement employ quite large arrays. The arrays have more microphones and the apertures of the arrays are usually big. A big aperture may greatly limit the applications of the microphone array. If an array is used in a telephone, mobile phone, hearing aid or PDA, the array should be small enough to be embedded in these small devices. Therefore, the study of speech enhancement methods or algorithms using a small microphone array has great importance and it apparently has great value. However, we find the achievement in small microphone array based algorithms is still quite limited.

In this thesis, we call these microphone arrays, which can be embedded in a telephone, mobile phone, hearing aid or PDA, small microphone arrays. Their apertures are generally less than 8cm and they generally employs less than 8 microphones. Sometimes they employ only 2 or 3 microphones and the apertures are less than 5cm. For most microphone arrays nowadays the aperture is much bigger than 5cm. Some of them even have an aperture of several meters and employs hundreds of microphones. For instance, the microphone array built in Delta Smart House by MIT for speech enhancement and speaker location employs 1020 microphones and takes up a whole sidewall of the laboratory [141].

This thesis will concentrate on the study of the methods or algorithms for speech enhancement using a small microphone array. For the consideration of real-time implementation, this thesis mainly develops the algorithms with low computational complexity.

1.2 Research history of speech enhancement

The research of speech enhancement is regarded as beginning in the 1960s for practical requirements. Figure 1.2.1 shows the number of papers with 'speech enhancement' in their titles and published in journals of the IEEE from 1977 to 2008. The graph shows that research into speech enhancement has been expanding gradually and a new research upsurge has formed in recent years.

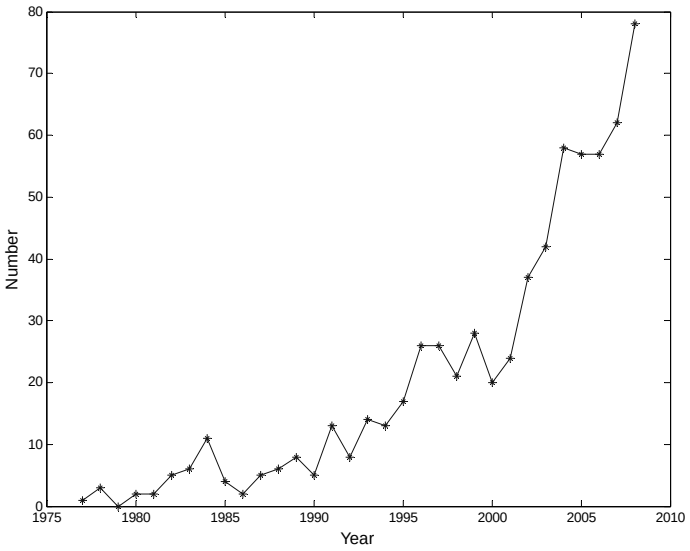

Figure 1.2.1 Number of papers with 'speech enhancement' in their titles and published in journals of the IEEE from 1977 to 2008

In the past 40 years, a large number of speech enhancement algorithms have been proposed [18, 11, 74]. There are different classification methods for these algorithms. However, according to the number of the employed microphones, the algorithms can clearly be classified as single-channel (or one-microphone) speech enhancement algorithms and multichannel (or microphone array) speech enhancement algorithms.

Before the 1980s, most of the algorithms dealt mainly with single-channel speech enhancement. Among these algorithms, the spectral power subtraction, Weiner filtering and the statistical-model-based algorithm are the most promising ones. In the 1980s, with the development of digital processors, the implementation of the algorithms gained a great deal of research interest. Since the 1990s, multichannel or microphone array

speech enhancement methods have flourished and many corresponding algorithms have been proposed. These algorithms include Delay And Sum (DAS) beamforming, Linear Constrain Minimum Variance (LCMV) beamforming, Generalized Sidelode Canceling (GSC), Post-filtering (PF), Generalized Singular Value Decomposition (GSVD), Blind Sources Separation (BSS) and so on. In addition, wavelet, neural networks and subspace algorithms, applied both for single-channel and multichannel speech enhancement, have also been studied. Most of these algorithms will be briefly introduced in the next section.

In recent years, many universities and research institutes have become involved in this research area. New algorithms for speech enhancement are constantly presented, both in single-channel and multichannel speech enhancement.

Some speech enhancement product examples are digital hearing aids, super directive microphones, noise resistant mobile phones and telecommunication networks, robust computer speech recognition systems, etc.

In telecommunication networks or mobile phones, the algorithms employed are mainly based on single-channel speech enhancement and they are implemented by software in the mobile phones or in the interchangers of the telecommunication networks [138, 142]. Widrow's research group in Stanford University designed the necklace microphone array for the digital hearing aid [143]. The super directive microphone was announced by Audio-Technica in 2004, in which five inner microphones are used to form the beamforming [137]. Microsoft Corporation announced a microphone array for the desktop computer in 2005, which may offer better speech quality and increase the speech recognition rate [140]. Microsoft Corporation also announced a microphone array for teleconferencing in 2005 [140].

1.3 Introduction to speech enhancement algorithms

Many speech enhancement algorithms have been proposed. They can be classified into two categories: single-channel speech enhancement and multichannel (or microphone array) speech enhancement algorithms. The main algorithms in each category are briefly introduced as follows.

1.3.1 Single-channel speech enhancement algorithms

Traditional speech enhancement algorithms mainly involve single-channel processing. These algorithms need only one microphone and thus they can be easily embedded in many audio devices such as telephones, mobile phones, computers, etc. They do not need an extra microphone or extra signal acquiring circuit. Comparatively, they also have lower computational complexity.

Another reason for single-channel speech enhancement is that each output of multichannel speech enhancement algorithms can be regarded as a single-channel speech signal and can be further enhanced by single-channel algorithms.

There have been many achievements in single-channel speech enhancement and many algorithms have been proposed. The main algorithms include:

- Short-time spectrum based algorithms
- Statistical model based algorithms
- Hearing model based algorithms
- Speech generating model based algorithms
- Subspace algorithms
- Wavelet algorithms
- Single-channel speech separation algorithms

The short-time spectrum based method has the most abundant content in single-channel speech enhancement [15, 12, 8, 40, 47, 54, 65, 76, 99]. It includes several algorithms such as spectral subtraction, improved spectral subtraction, Weiner filtering, etc. In 1979 Boll proposed a simple but effective algorithm called Spectral Subtraction (SS) [15]. It finds a section of pure noise signal and computes its spectrum. Next, it uses the spectrum as an estimation of the noise spectrum in noisy speech. It then subtracts the estimated noise spectrum from the spectrum of the noisy speech to get the estimation of the spectrum of clean speech. Finally, it transfers the estimated speech spectrum into a time-domain signal to get the enhanced time-domain speech signal. The main drawback of this SS algorithm is that the algorithm can inevitably cause so-called

“music noise” in the enhanced speech. The reason for this is that the real noise spectra are not exactly the same as the estimated spectra. However, if the power of the noise is much less than the power of the speech, this kind of music noise is very light and even cannot be perceived by human ears. Another drawback of the SS algorithm is that it needs a good Voice Activity Detector (VAD). Otherwise, the wrongly detected section of the noise causes serious damage to the enhanced speech. Besides, the SS algorithm can only deal with stationary noise since it uses the spectrum of the pure noise section as the estimation of the noise spectrum for the following sections. It also cannot deal with noise similar to human speech because it cannot distinguish the desired speech from the noise in this circumstance. Since then, many improved SS algorithms have been proposed such as the average amplitude algorithm, the power spectral subtraction algorithm, the nonlinear spectral subtraction algorithm, the iterative spectral subtraction algorithm and the multi-band spectral subtraction algorithm. The introductions and comparative advantages and disadvantages of these algorithms can be found in Loizou’s book [74].

Statistical model based algorithms use a statistical estimation framework to estimate the spectrum of the clean speech in noisy environments. The algorithms make use of probabilistic-based estimators of the speech spectrum such as the maximum-likelihood estimator, the minimum mean-square error estimator, and a posteriori estimators [32, 69, 74].

Virag proposed a speech enhancement algorithm based on the masking properties of human auditory system [118]. Then perceptual filter based algorithms were also studied and developed in [4, 23, 55].

The speech-generating model based algorithm makes use of the model whereby speech is generated through a linear time-variant filter excited by a source signal. It estimates the parameters of the filter and then generates the enhanced speech through the estimated parameters [92, 46].

The subspace algorithm separates the desired signal subspace and noise subspace by eigenvalue decomposing of the noisy speech. It then rebuilds the clean speech signal in the desired signal subspace [33, 7, 56].

The wavelet denoising algorithm first takes wavelet transform to the noisy speech. Then it discards the small coefficients in the wavelet transform according to different characteristics between the coefficients of the speech and the coefficients of the noise. It then takes the inverse wavelet transform to restore the clean speech [9, 119, 120].

Single-channel speech separation has become another potential research subject in recent years [90, 100]. It employs only one channel of noisy speech to separate the clean speech from the noises. In common Blind Sources Separation (BSS) it is required that sensors must be more than the signals. So, the common BSS algorithms cannot be directly used for single-channel speech enhancement. Some improvements or other novel algorithms must be studied. The time-frequency method appears to be one of the promising approaches.

1.3.2 Microphone array speech enhancement algorithms

Unlike single-channel speech enhancement, microphone array based speech enhancement can make use of space-domain information except for time-domain and frequency-domain information. Therefore, microphone array speech enhancement can certainly achieve better results. Generally speaking, the more microphones employed, the better enhancement achieved.

The main algorithms dealing with microphone array speech enhancement include

- Delay And Sum (DAS)
- Linear Constrained Minimum Variance (LCMV)
- Adaptive Noise Canceling (ANC)
- Post-filtering (PF)
- Generalized Sidelobe Canceling (GSC)
- Blind Sources Separation (BSS)
- Subband Processing

A microphone array can suppress noise and thus enhance speech through beamforming. The simplest beamforming algorithm is Delay And Sum (DAS) beamforming [49, 123]. However, it has low efficiency. Even under the most ideal

conditions (the acquired noise signals are completely uncorrelated), to get 20 dB enhancement we must employ at least 100 microphones.

The Linear Constrained Minimum Variance (LCMV) algorithm, especially the LCMV proposed by Frost [36], makes use of not only the present signal samples but also the delayed samples to construct the beamforming. It may achieve much better speech enhancement result than the DAS algorithm.

The Adaptive Noise Canceling (ANC) algorithm published by Widrow in 1975 has wide applications [127, 128, 129]. It is suitable for canceling highly correlated noise. It has the advantage of less complexity and it may deal with many kinds of noises. But, if the speech signal is leaked into its referential channel, the speech will also be partially canceled and thus the speech quality may degrade.

The Post-filtering (PF) algorithm published by Zelinski [131] employs a Weiner filter to further suppress the noise for the enhanced speech signal by the DAS algorithm. However, the estimation for the coefficients of the Weiner filter is processed through multichannel, rather than single-channel, noisy speech signals.

The Generalized Sidelobe Canceling (GSC) algorithm proposed by Griffths and Jim [45] is in fact another form of the LCMV algorithm. It has become one of the most important algorithms for speech enhancement using a microphone array. It consists of a fixed beamformer, a blocking matrix and an adaptive noise canceller. Its fixed beamformer may suppress uncorrelated noise, while its adaptive noise canceller together with the blocking matrix may cancel the correlated noise. Therefore, GSC may suppress correlated and uncorrelated noises. This makes it a more practical application since in most practical situations the noise field is diffused, which means the noise is partially correlated and partially uncorrelated. The main drawback of GSC lies in its blocking matrix for it cannot completely block the speech signal and thus make the partial cancellation of the speech in the enhanced speech. To overcome this drawback, some improved GSC algorithms have been proposed [37, 39, 53, 122, 27].

The Generalized Singular Value Decomposition (GSVD) algorithm is based on singular value decomposition and then it is converted to an optimal filtering problem. It may offer the best speech enhancement effect. But it has high computational complexity and it is used for uncorrelated and stationary noises [31].

The Blind Sources Separation (BSS) algorithm can be used for speech

enhancement [5, 105, 108]. It does not need any transcendental knowledge about speech and noises. It separates the speech and noises on the condition that the speech and all noises are independent. However, because of the complexity of the propagations of the speech and the noises, the elements of the mixed matrix in BSS are time-variable vectors. So, it must be very difficult to find the separation matrix.

The subband algorithm decomposes all acquired noisy speech signals into a group of subband signals for processing. It has more flexibility and may offer better speech enhancement results [33, 41, 42, 72, 92, 3]. In every subband, the array signals fall in a comparatively narrow frequency band. So, a more accurate beamformer may be performed, and the order or length of the adaptive filter for noise cancellation could be shorted. As a result, the enhancement of the speech can be improved and the total complexity of the algorithm might be reduced.

1.4 Evaluation of the enhanced speech

There are two ways to evaluate the quality of the enhanced speech: subjective evaluation and objective evaluation [74, 11, 57, 58, 59].

Subjective evaluation involves comparisons of the original and enhanced speech signals by a group of listeners who are asked to rate the quality of speech along a predetermined scale. Mean Of Scores (MOS) is the commonly used subjective evaluation. Its rating scale is defined in table 1.4.1 [74].

Table 1.4.1 MOS rating scale

Rating	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible, and slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying and objectionable

In terms of objective evaluation, we may measure the enhanced speech by

observing its waveform in the time-domain or the spectrogram in the frequency domain. In particular, we may use a real number for measurement such as SNR (Signal-to-Noise Ratio). Since real number measurement has a consentaneous standard, this evaluation is thus widely used. In fact, when we evaluate the quality of the enhanced speech, the measurement method should be related with the application problem. If the speech enhancement is used for speech recognition, the recognition rate would be the proper standard for measurement. If it is used for teleconferencing or telecommunication, the intelligibility and quality of enhanced speech become the important factors.

A commonly used objective evaluation is the Signal-to-Noise Ratio (SNR). Suppose clean speech signal $s(k)$ can be acquired, and then the SNR of the noisy speech is defined by

$$\text{SNR}=10\log_{10}\frac{\sum_{k=1}^{K}s^2(k)}{\sum_{k=1}^{K}\left[x^2(k)-s^2(k)\right]} \tag{1.4.1}$$

where $x(k)$ is the noisy speech, k is the time index and K is the number of the total samples.

It should be noticed that the high SNR defined by (1.4.1) does not necessarily mean high quality for the enhanced speech. If the speech sections take only a small portion in the whole signal concerned, the SNR can be very high when the noise and speech are both highly depressed. However, the quality of the enhanced speech is not good because the speech is also greatly depressed.

The above SNR can be calculated only if clean speech or pure noise can be acquired. However, in many practical applications clean speech or pure noise is actually unknown to us. So we usually take the following SNR for practical uses.

$$\text{SNR}=10\log_{10}\frac{\sum_{k=1}^{K}\left[x^2(k)-n^2(k)\right]}{\sum_{k=1}^{K}n^2(k)} \tag{1.4.2}$$

In this SNR definition (1.4.2), the noise is supposed to be stationary. So we may use a section of the noise to estimate the whole noise signal $n(k)$.

In order to be used for non-stationary noise and to be more accurate, the above definition (1.4.2) is modified in this thesis as follows

$$\mathrm{SNR} = 10\log_{10} \frac{\alpha \sum_{k \in T_s} x^2(k) - \sum_{k \in T_n} x^2(k)}{\sum_{k \in T_n} x^2(k)} \tag{1.4.3}$$

where $x(k)$ is the noisy speech; T_s is the sample set containing the speech; T_n is the sample set without the speech (pure noise); and $\alpha = m(T_n)/m(T_s)$, where $m(T_n)$ and $m(T_s)$ are the numbers of the samples in T_n and T_s respectively.

Of course, the subjective evaluation is not necessarily related to the objective evaluation. However, generally speaking, they are in accordance in most of the measurements.

There are two reasons that make speech enhancement a difficult research objective [11]. One is the complexity of different kinds of the noise in noisy speech. Another is the complexity of the evaluations for enhanced speech due to the human auditory system. However, this thesis will concentrate on the methods or algorithms for speech enhancement. As to the evaluation of enhanced speech, we use one or several measurements to indicate the effectiveness of the algorithms.

1.5 Strategies and relevant work

As mentioned in section 1.1, single-channel speech enhancement has limited effect because it can not keep speech undistorted while it suppresses environmental noise. Fortunately, a microphone array may break through this limitation.

For many applications, a microphone array should be small enough. The aperture of the array and the number of the employed microphones are greatly limited in a small array. Many algorithms validated for a common array or a big array may have little effect or even no effect at all. They are not suitable for a small array.

In fact, some researchers have noticed the importance of a small array for practical applications. Martin (2001) studied the post-filtering and reverberation suppressing with a small microphone array [69]. However, in his work the aperture of the array was not small, but the number of the microphones employed was small. Spriet (2005) employed

two or three closely placed microphones to form a single-channel hearing aid [109], which is a really small array for speech enhancement. Fortemedia Company (2008) announced a mini array for mobile phone speech enhancement [139], employing only two very closely placed microphones. These achievements are important references for us. However, the algorithms in this thesis are not similar to these achievements.

In a small array, as the microphones are closely placed, the spatial correlation of the noise is usually higher than that in a common array. So, many algorithms that need the uncorrelation of the noise will not perform well, such as Delay And Sum, multichannel Weiner filtering, Subspace and Generalized Singular Value Decomposing, etc. Other algorithms should be employed before these algorithms can be applied. However, in a small array, the requirement for high correlation by some algorithms, such as Adaptive Noise Canceling (ANC) and Multichannel ANC (MANC), can be met well. Therefore, ANC or MANC seems a useful algorithm for a small array. However, ANC or MANC needs referential signals to contain no speech signal. Otherwise, the speech signal will be cancelled with the cancellation of the noise. Therefore, common ANC or MANC is also not a suitable algorithm for a small array. Improvements must be made before it can be used for small array based speech enhancement.

Some improvements for two-channel ANC were made in [86, 136, 67], in which two-channel crosstalk resistant ANC algorithms were proposed. Since in microphone arrays there exists severe crosstalk of speech or noise between any two acquired signals, especially in the small microphone array, crosstalk resistant ANC performs much better than ANC. However, we find these algorithms are not stable. They might diverge from time to time. References [78] and [96] deal with crosstalk resistant ANC for biomedical signal extraction. In [78] Madahavan employed a three-stage adaptive system to extract the desired biomedical signals, in which three adaptive filters are employed. Because these filters do not employ recursive adaptation, they appear quite stable. In this thesis, the three-stage system will be simplified into a two-stage system, in which only two adaptive filters are employed.

The effect of two-channel crosstalk resistant ANC is limited. Multichannel processing may increase the noise cancellation ability. This thesis will also extend the

proposed two-channel two-stage crosstalk resistant ANC algorithm to a multichannel algorithm. A Multichannel Crosstalk Resistant Adaptive Noise Cancellation (MCRANC) algorithm is thus proposed.

Based on MCRANC, the combination algorithms of MCRANC with other speech enhancement algorithms, further improved algorithms to MCRANC, and MCRANC based improvements to an existing powerful algorithm are studied.

1.6 Thesis contributions

This thesis aims to derive new algorithms for small microphone array based speech enhancement. Two effective hybrid algorithms are proposed in this thesis. Each hybrid algorithm employs several new algorithms and methods proposed in different chapters of this thesis. The thesis contributions can be listed as:

(1) The Multichannel Crosstalk Resistant Adaptive Noise Cancellation (MCRANC) algorithm is proposed.

The algorithm is described in detail from two-channel to multichannel. Its principal and its computational complexity are analyzed and calculated. The algorithm employs only two FIR based filters, which gives the algorithm good stability, makes it low computational complexity and ensures few limitations to the type of the noise and the structure of the array. Experimental results indicated that MCRANC is a suitable algorithm for a small array and it can achieve good enhancement results.

(2) Three combined algorithms of MCRANC with other single-channel or array algorithms are proposed.

The combination with single-channel algorithms mainly presents the cascade of MCRANC by the Improved Spectral Subtraction (ISS) algorithm. Theoretical analysis and experimental results prove the cascading algorithm outperforms MCRANC and ISS algorithms if they act alone.

The combination with array algorithms mainly presents the combination with DAS

beamforming and the combination with Weiner post-filtering. Both of the combinations are realized by using MCRANC to pre-enhance every channel of the array signal and then employing the array enhancement algorithms. The pre-enhancement by MCRANC is provided to cancel correlated noise, while the array algorithms are employed to suppress uncorrelated noise.

(3) Four improvements to MCRANC itself are proposed.

An improved MCRANC with multichannel inputs for the second-stage filter is proposed to improve the quality of the final enhanced speech. This improvement is useful when the spatial correlation of the speech signals is not as high as we expect.

Another improvement to MCRANC is employing different sampling rates for the main channel signal and the referential channel signals. It is suggested that the sampling rate for the referential channel signals should be higher or lower, according to the noise type, than the required rate for the output speech.

Fixed Beamforming Partial-channel MCRANC (FBF-P-MCRANC) is also a proposed improvement to MCRANC. For its fixed beamformer, a Delay And Weighted Sum And Selection (DAWSAS) algorithm is also presented.

The fourth improvement is to employ subband menthod to MCRANC. A subband FBF-P-MCRANC is proposed.

(4) Two improved Modified Generalized Sidelobe Canceling (MGSC) algorithms are proposed. It is indicated that the essence of the proposed improved MGSC algorithms is to extend the common blocking matrix to a time-variable vector blocking matrix.

One improved MGSC algorithm uses MCRANC to improve the signal leakage problem in the blocking path of MGSC. The other improved MGSC algorithm deals with the leakage problem by use of a shared distorted signal. It is actually a simplified version of the first improved MGSC algorithm.

(5) Two hybrid algorithms for small microphone array based speech enhancement are presented.

One hybrid algorithm is based on MCRANC. It employs several algorithms and

methods proposed in this thesis and some existing algorithms. It contains MCRANC algorithm, fixed beamforming, DAWSAS algorithm, multiple sampling rates method, partial-channel method, multichannel distorted signal filtering method, subband method and ISS algorithm. The other hybrid algorithm is based on MGSC. It also contains most of the above-mentioned algorithms and methods.

Both of the two hybrid algorithms can be used in different environments. However, the MCRANC based hybrid algorithm acts better when the speech source is very near the array, such as in the case of mobile phones, telephones and PDAs. The MGSC based hybrid algorithm appears to be more effective if the speech source is not very near the array, such as in the case of hearing aids. The proposed algorithms all have low computational complexity and suitable for real-time implementation.

Several publications have been published from this research as listed:

[1] Q. Zeng, W. Abdulla. Speech enhancement by multichannel crosstalk resistant ANC and improved spectrum subtraction. EURASIP Journal on Applied Signal Processing, Vol.2006, Article ID 61214, 10 pages, 2006 (SCI: 1110-8657) (EI: 20064910285991)

[2] Q. Zeng, W. Abdulla. Speech enhancement by multi-channel crosstalk resistant adaptive noise cancellation. IEEE Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006), Toulouse, France, Vol.1, pp.485-488, 2006 (EI: 20071410525407)

[3] Q. Zeng, S. Ouyang. Signal extraction by two-stage space-time adaptive noise cancellation and beamforming. IEEE proceedings of International Conference on Communication, Circuit and Systems (ICCCAS 2006), Guilin, China, Vol.1, pp. 324-328, 2006 (EI: 20081011131732)

[4] Q. Zeng, W. Abdulla. Speech enhancement by MCRANC and post-filtering. Proceedings of Australasian International Conference on Speech Science and Technology, Auckland, New Zealand, pp. 276-280, 2006

[5] Q. Zeng, W. Abdulla. Speech enhancement using GSC with multi-channel crosstalk resistant adaptive signal cancellation. Proceedings of International Conference on

Signal Processing, Pattern Recognition and Applications (ISTEAD 2009), Innsbruck, Austria, pp. 335-339, 2009

[6] Q. Zeng, W. Abdulla. A novel modified GSC and its application to speech enhancement. International Conference on Applied Signal Acquisition and Processing, Kuala Lumpur, Malaysia, accepted, 2009

1.7 Structure of thesis

The first chapter introduces the research history and the main algorithms for speech enhancement. The necessity for the research on small microphone array is pointed out. Chapter 2 presents the Multichannel Crsstalk Resistant Adaptive Noise Cancellation (MCRANC) algorithm. Chapter 3 presents the three combinational algorithms of MARANC with existing algorithms. Chapter 4 describes four improvement methods for MCRANC itself. Chapter 5 gives out two improved Modified Generalized Sidelobe Canceling (MGSC) algorithms. Based on the algorithms and methods from chapter 2 to 5, two more powerful hybrid algorithms are presented in chapter 6 for small microphone array based speech enhancement. Finally, chapter 7 leads to the conclusions of this thesis.

The thesis structure or the relationships of the seven chapters is demonstrated in figure 1.7.1.

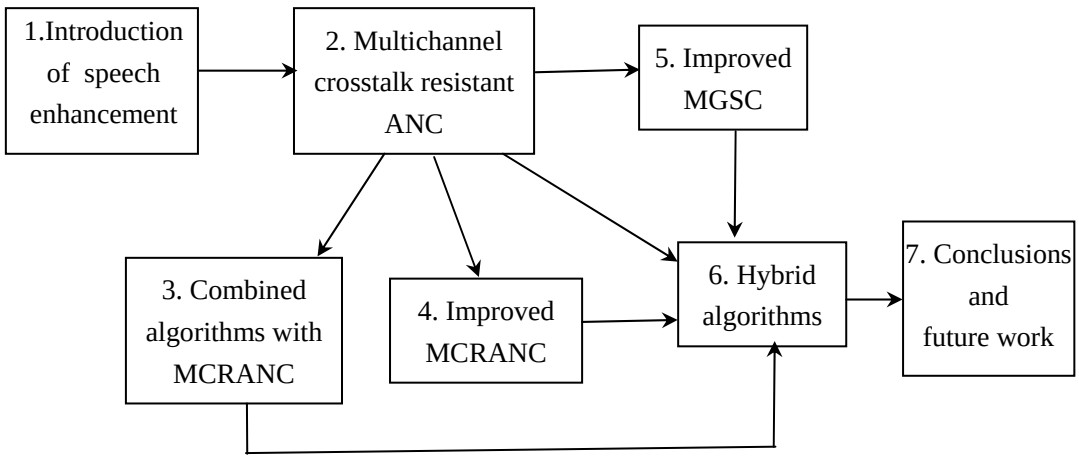

Figure 1.7.1 Structure of thesis

Chapter 2 Multichannel Crosstalk Resistant ANC

Two-channel and Multichannel Crosstalk Resistant Adaptive Noise Cancellation (MCRANC) algorithms are proposed in this chapter after a brief preamble to two-channel and multichannel adaptive noise cancellation algorithms. The roles of these two algorithms in speech enhancement are discussed. Simulation and real environment experiments are presented. The experimental results verify that MCRANC is a proper algorithm for speech enhancement using a small microphone array. It can achieve a significant speech enhancement performance with low computational complexity.

2.1 Introduction

Among many algorithms for speech enhancement, adaptive noise cancellation (ANC) remains one of the most important. It may be used with different kinds of noises and it is easy for real-time implementation as it has a low computational complexity. Unlike many other algorithms that require a large aperture microphone array, it may perform well with a small microphone array.

Due to the propagation complexity of the audio signal, the noises acquired by the main channel and the referential channel may not be completely correlated in a two-microphone based ANC system. As a result, the performance of speech enhancement using two-channel ANC is limited. For better performance, we should employ more microphones to form a microphone array.

In the ANC algorithm, the referential channel should ideally contain only a noise signal. Otherwise, the speech signal in the main channel would be partially or even largely cancelled with the cancellation of the noise [127,128]. The higher level of

speech signal picked by the referential channel, the more speech cancellation occurs in the main channel, and thus the worse the speech enhancement result. However, in the practical environment it is almost impossible for the referential channel to contain only noise. In a small microphone array in particular, the referential channel has almost the same amount of speech signal as the main channel. This problem motivates us to find novel noise cancellation algorithms for a small microphone array.

In this chapter two-channel adaptive noise cancellation and multichannel adaptive noise cancellation are introduced first. Then a two-channel Crosstalk Resistant Adaptive Noise Cancellation (CRANC) algorithm and a Multichannel Crosstalk Resistant Adaptive Noise Cancellation (MCRANC) algorithm are proposed. Their principles for noise cancellation and their computational complexities are also discussed. Finally, simulation and real environment experiments are described. The experimental results verify that the MCRANC is a suitable algorithm for a small microphone array. It can achieve a significant speech enhancement performance and it has a low computational complexity.

2.2 Adaptive noise cancellation

2.2.1 Two-channel adaptive noise cancellation

The research of adaptive noise cancellation was originated from the work of Howells and Applebaum who worked in the General Electric Company from 1957 to 1960. At that time their work was to design an antenna sodelobe canceling system. However, the theoretic work of Adaptive Noise Cancellation (ANC) was completed by Widrow in 1975 [127]. Since then ANC has become a widely used algorithm in many applications.

As indicated in figure 2.2.1, suppose the noise signal $n(t)$ propagates from its source to microphones M_0 and M_1, and the noise signal acquired by microphones

M_0 and M_1 are $n_0(t)$ and $n_1(t)$ respectively. Suppose the speech signal $s(t)$ propagates from its source to the main microphone M_0 and the speech signal acquired by microphone M_0 is $s_0(t)$. We further suppose the speech signal $s(t)$ does not propagate to referential microphone M_1. So the eventual acquired signal by microphones M_0 and M_1 are respectively

$$x_0(t) = s_0(t) + n_0(t) \tag{2.2.1}$$

$$x_1(t) = n_1(t) \tag{2.2.2}$$

We have

$$n_i(t) = h_{ni}(t) * n(t) \qquad i = 0,1, \tag{2.2.3}$$

$$s_0(t) = h_{s0}(t) * s(t) \tag{2.2.4}$$

where $*$ is the convolution sign, $h_{n0}(t)$ and $h_{n1}(t)$ are the impulse responses of the intermediate media between the noise source and the acquiring microphones M_0 and M_1 respectively, $h_{s0}(t)$ is the impulse response of the intermediate media between the speech source and the acquiring microphone M_0.

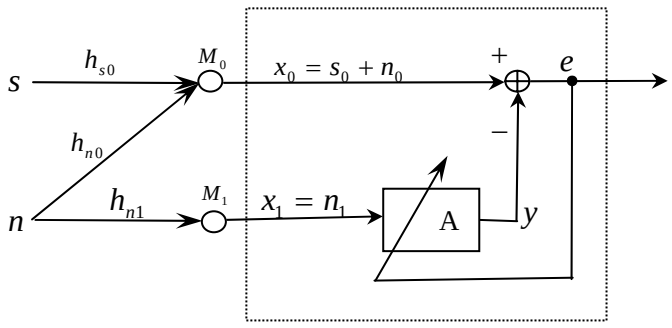

Figure 2.2.1 Two-channel adaptive noise cancellation

In the discrete time domain $t = kT$, where k is the time index integer and T is the period of the sampler. For simplicity, t is replaced by k only. For a continuous

signal $x(t)$, its discrete signal is noted as $x(k)$.

In noisy environments, we can only acquire $x_0(k)$ and $x_1(k)$. We need to use $x_0(k)$ and $x_1(k)$ to extract the speech signal $s(k)$ or the speech signal $s_0(k)$. ANC provides us with a way to extract $s_0(k)$ if the speech signal is uncorrelated with the noise. Its principle can be illustrated by the dotted rectangle in figure 2.2.1.

Figure 2.2.1 obviously shows

$$\begin{aligned} e(k) &= x_0(k) - y(k) \\ &= s_0(k) + n_0(k) - y(k) \end{aligned} \tag{2.2.5}$$

Take the square of both sides of the equation to get

$$e^2(k) = s_0^2(k) + [n_0(k) - y(k)]^2 + 2s_0(k)[n_0(k) - y(k)]$$

Note that $s_0(k)$ is uncorrelated with $n_0(k)$ and $y(k)$, therefore, the mean

$$E[e^2(k)] = E[s_0^2(k)] + E[n_0(k) - y(k)]^2 \tag{2.2.6}$$

Adjust the coefficients of adaptive filter A to minimize the output power of $E[e^2(k)]$ and we have

$$\min E[e^2(k)] = E[s_0^2(k)] + \min E[n_0(k) - y(k)]^2 \tag{2.2.7}$$

From equation (2.2.7) we can see that to minimize $E[e^2(k)]$ means to minimize $E[n_0(k) - y(k)]^2$. So the output $y(k)$ of filter A is the optimal estimation under Minimized Mean Square Error (MMSE) criteria. Thus from equation (2.2.5), the output $e(k)$ of the whole system is the optimal estimation of $s_0(k)$ under the MMSE criteria.

According to equation (2.2.7) ,the minimal power that the system may achieve is $E[e^2(k)] = E[s_0^2(k)]$. When this happens we have $E[n_0(k) - y(k)]^2 = 0$, $y(k) = n_0(k)$ and $e(k) = s_0(k)$, which means the speech in the output of the system is a noise free signal.

In the ANC system filter A may take any formation. However, for simplicity it

usually adopts the FIR formation. In this formation we have

$$y(k) = \mathbf{w}\mathbf{n}_1(k)$$

$$= \sum_{l=-L_1}^{L_2} w_l n_1(k-l)$$

where $\mathbf{w}$ is the coefficient vector of the filter

$$\mathbf{w} = [w_{-L_1}, \cdots, w_{-1}, w_0, w_1, \cdots, w_{L_2}]$$

$$\mathbf{n}_1(k) = [n_1(k+L_1), \cdots, n_1(k+1), n_1(k), n_1(k-1), \cdots, n_1(k-L_2)]^T$$

Since we may delay L_1 samples for the main channel signal, we may take $L_1 = 0$ and note L_2 as L for simplicity.

For the adaptation of the coefficients of the adaptive filter, there are many algorithms such as least mean squares (LMS), normalized least mean squares (NLMS), recursive least squares (RLS), block fast transversal filter (BFTF), least squares lattice-ladder (LSLL) and so on [48]. Among these algorithms LMS is the simplest and most widely used. There are also many improved LMS algorithms [43, 68].

2.2.2 Multichannel adaptive noise cancellation

In a real environment, there might be several noise emission sources. Even if there is only one noise emission source, the noise may take many propagation paths to any acquiring microphone such as a direct path, refraction paths and reflection paths. These propagation facts may lead to a decrease of efficiency in the two-channel ANC algorithm. The reason for this is that one channel only of referential noise is unable to completely cancel the noise in the main channel. One solution is to employ more microphones to get more referential channels of noises and use all of them to cancel the noise in the main channel [127, 129]. Theoretical and experimental results have verified this scheme is feasible and effective.

As indicated in the left side of figure 2.2.2, noises $n_j(t)$ ($j = 1, 2, \cdots, M$)

propagate through impulse responses $h_{n_j 0}(t)$ and $h_{n_j i}(t)$ to arrive at the main microphone M_0 and referential microphones M_i ($i = 1,2,\cdots,N$) while the speech signal $s(t)$ propagates through impulse response $h_{s0}(t)$ to arrive at the main microphone M_0 and $s(t)$ does not propagate to any referential microphones M_i, $i = 1,2,\cdots,N$. As a result, microphone M_0 acquires the signal

$$x_0(t) = s_0(t) + n_0(t)$$

where

$$s_0(t) = h_{s0}(t) * s(t)$$

$$n_0(t) = \sum_{j=1}^{M} h_{n_j 0}(t) * n_j(t)$$

while the referential microphones M_i ($i = 1,2,\cdots,N$) acquire noise signals only

$$x_i(t) = \sum_{j=1}^{M} h_{n_j i}(t) * n_j(t) \qquad (i = 1,2,\cdots,N)$$

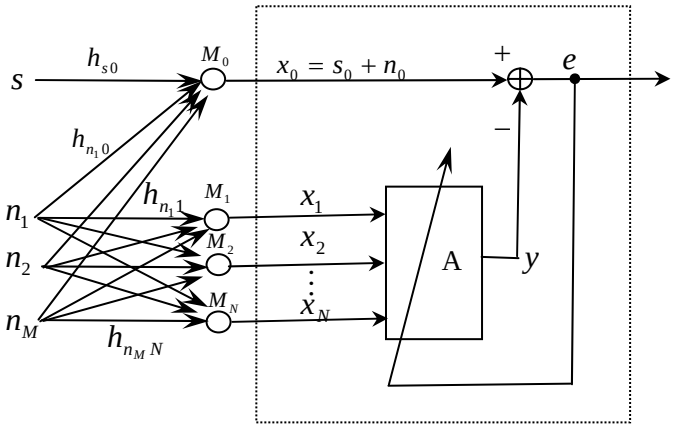

Figure 2.2.2 Multichannel adaptive noise cancellation

Also, in the discrete time domain, a signal $x(t)$ is replaced with $x(k)$.

The inner part of the dotted rectangle in the right side of figure 2.2.2 demonstrates the diagram of the Multichannel Adaptive Noise Cancellation (MANC). Similar to two-channel scheme, we need only adjust the coefficients of filter A to minimize

$E[e^2(k)]$ to get the optimal estimation of the speech $s_0(k)$ under MMSE criteria. $e(k)$ would be the optimal estimation of $s_0(k)$ and $y(k)$ would be the optimal estimation of $n_0(k)$.

Again, filter A may take any formation and it usually adopts FIR formation for simplicity. In FIR formation, we have

$$y(k) = \mathbf{w}\mathbf{x}(k)$$

$$= \sum_{i=1}^{N}\sum_{l=0}^{L} w_{il}x_i(k-l)$$

where $\mathbf{w}$ is the coefficient vector of the filter

$$\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_N]$$

$$\mathbf{w}_i = [w_{i0}, w_{i1}, \cdots, w_{iL}]$$

and

$$\mathbf{x}(k) = [\mathbf{x}_1(k), \mathbf{x}_2(k), \cdots, \mathbf{x}_N(k)]^T$$

$$\mathbf{x}_i(k) = [x_i(k), x_i(k-1), \cdots, x_i(k-L)]$$

Of course, there are also many coefficient adaptation algorithms for multi-input filter A [48].

2.3 Two-channel crosstalk resistant ANC algorithm

In subsection 2.2.1, it is assumed that there is no speech component in the referential signal. Otherwise, the speech in the main channel may be partially cancelled with the noise cancellation in ANC [127, 86]. However, in most environments, the speech signal may inevitably propagate to a referential microphone and cause the referential signal to contain the speech as well, specially when the aperture of microphone array is small. Therefore, the common ANC algorithm is not suitable for speech enhancement under such condition.

If the main channel signal and the referential channel signal contain both the speech and noise, we call this phenomenon the “crosstalk” of the speech signal (or noise signal). In references [67, 78, 86, 96] some algorithms based on two channels of crosstalk signals are proposed for the extraction or enhancement of the desired signal. However, we find the algorithms in references [86, 67] are not stable and they even diverge from time to time. References [78, 96] discuss biomedical signal processing problems and they use a three-stage filtering system to extract desired biomedical signals.

In this section, a two-channel Crosstalk Resistant Adaptive Noise Cancellation (CRANC) algorithm is proposed. It consists of only two filters and it has low computational complexity. In addition, it has good stability and can deal with different kinds of noises.

2.3.1 Algorithm and its principle

As shown in figure 2.3.1, assume speech signal $s(k)$ propagates to microphone M_0 and M_1 through propagation functions $H_{s0}(z)$ and $H_{s1}(z)$ and converts to signals $s_0(k)$ and $s_1(k)$ respectively, while noise signal $n(k)$ propagates to microphone M_0 and M_1 through propagation functions $H_{n0}(z)$ and $H_{n1}(z)$ and converts to signals $n_0(k)$ and $n_1(k)$ respectively. Then the actual signals acquired by microphones M_0 and M_1 are

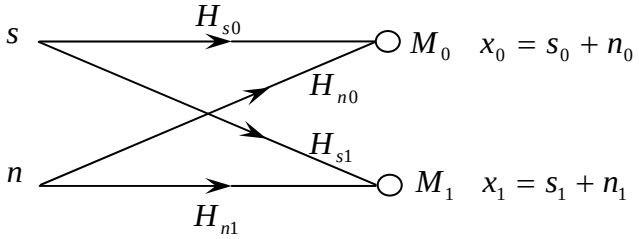

Figure 2.3.1 The crosstalk of speech and noise

$$x_0(k) = s_0(k) + n_0(k)$$

$$x_1(k) = s_1(k) + n_1(k)$$

where both $x_0(k)$ and $x_1(k)$ contain a speech signal and a noise signal. This is the so-called crosstalk of the speech signal (or noise signal) [86, 78].

From figure 2.3.1

$$s_0(z) = H_{s0}(z)s(z) \tag{2.3.1}$$

$$s_1(z) = H_{s1}(z)s(z) \tag{2.3.2}$$

$$n_0(z) = H_{n0}(z)n(z) \tag{2.3.3}$$

$$n_1(z) = H_{n1}(z)n(z) \tag{2.3.4}$$

Note the propagation function from $s_0(k)$ to $s_1(k)$ as $H_{s_0s_1}(z)$, and the propagation function from $n_1(k)$ to $n_0(k)$ as $H_{n_1n_0}(z)$, i.e.

$$s_1(z) = H_{s_0s_1}(z)s_0(z) \tag{2.3.5}$$

$$n_0(z) = H_{n_1n_0}(z)n_1(z) \tag{2.3.6}$$

We may conclude

$$H_{s_0s_1}(z) = \frac{H_{s1}(z)}{H_{s0}(z)} \tag{2.3.7}$$

$$H_{n_1n_0}(z) = \frac{H_{n0}(z)}{H_{n1}(z)} \tag{2.3.8}$$

Figure 2.3.2 demonstrates the structure of Crosstalk Resistant Adaptive Noise Cancellation (CRANC) proposed in this section for speech enhancement. It contains two adaptive filters and a VAD (Voice Activity Detector) [17, 25]. The speech enhancement is achieved by noise cancellation. The noise cancellation is based on the characteristic that the speech signal is an intermittent signal and it can be divided into Having Speech Period (HSP) and Non Speech Period (NSP).

During a NSP time section, microphones M_0 and M_1 acquire pure noise signals

$n_0(k)$ and $n_1(k)$. Use $n_0(k)$ as the main channel signal and $n_1(k)$ as the referential channel signal for the first stage which contains filter A. The purpose of this stage is to get the estimation of noise propagation function $H_{n_1n_0}(z)$. Obviously, we need only to adjust the coefficients of adaptive filter A to minimize the power of $e_1(k)$ in figure 2.3.2 to realize this purpose. After the realization of the minimization, the system function of filter A will be the estimation of $H_{n_1n_0}(z)$.

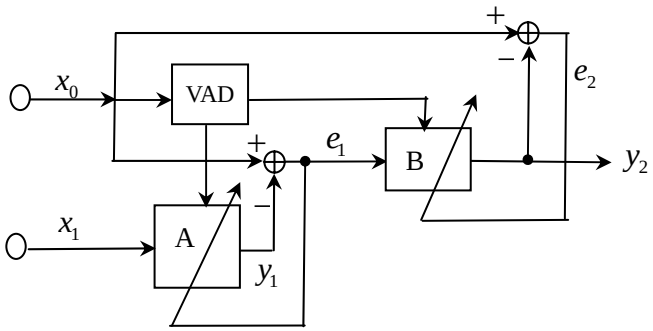

Figure 2.3.2 Structure of crosstalk resistant adaptive noise cancellation algorithm

Then, consider a HSP time section which follows the previous NSP. In this time section, suppose the noisy environment remains almost unchanged (including unchanged or a slight change only), which means the position of the noise source and the position of the microphone array and even the whole propagation environment remain almost unchanged. So, the system function $H_{n_1n_0}(z)$ acquired in the previous NSP will also remain almost unchanged. Thus, we have

$$\begin{aligned} Y_1(z) &= H_{n_1n_0}(z)X_1(z) \\ &= H_{n_1n_0}(z)[S_1(z)+N_1(z)] \end{aligned} \tag{2.3.9}$$

$$\begin{aligned} E_1(z) &= X_0(z)-Y_1(z) \\ &= S_0(z)+N_0(z)-H_{n_1n_0}(z)S_1(z)-H_{n_1n_0}(z)N_1(z) \\ &= S_0(z)-H_{n_1n_0}(z)S_1(z) \\ &= [1-H_{n_1n_0}(z)H_{s_0s_1}(z)]S_0(z) \end{aligned} \tag{2.3.10}$$

If the system function of filter B is $[1-H_{n_1n_0}(z)H_{s_0s_1}(z)]^{-1}$, by equation (2.3.10) we get

$$Y_2(z)=[1-H_{n_1n_0}(z)H_{s_0s_1}(z)]^{-1}E_1(z)=S_0(z) \qquad (2.3.11)$$

$$E_2(z)=X_0(z)-Y_2(z)$$

$$=[S_0(z)+N_0(z)]-S_0(z)=N_0(z) \qquad (2.3.12)$$

From equations (2.3.11) and (2.3.12) we see that speech signal $s_0(k)$ and noise signal $n_0(k)$ in the mixed signal $x_0(k)$ have been separated. So the output signal of filter B is the estimation of the wanted speech signal $s_0(k)$ or the enhanced speech.

In order to get the above system function $[1-H_{n_1n_0}(z)H_{s_0s_1}(z)]^{-1}$ by filter B, we need only adjust the coefficients of adaptive filter B to minimize the power of $e_2(k)$ in the second stage of CRANC. This is because

$$\|e_2(k)\|^2=\|x_0(k)-y_2(k)\|^2$$

$$=\|s_0(k)+n_0(k)-y_2(k)\|^2$$

$$=\|n_0(k)\|^2+\|s_0(k)-y_2(k)\|^2+2n_0(k)[s_0(k)-y_2(k)] \qquad (2.3.13)$$

Take expectation to both sides of the above equation and notice that $n_0(k)$ is uncorrelated with $s_0(k)$, then

$$E[e_2^2(k)]=E[n_0^2(k)]+E\{[s_0(k)-y_2(k)]^2\} \qquad (2.3.14)$$

So, minimizing the power of $e_2(k)$ implies minimizing $E\{[s_0(k)-y_2(k)]^2\}$. Since the input signal $e_1(k)$ for filter B is correlated with $s_0(k)$, the minimization is feasible. As a result, the output $y_2(k)$ of filter B will be the optimal estimation of speech signal $s_0(k)$ under MMSE criteria.

The level of the residual noise in the enhanced speech during HSP maybe is different from the level of the residual noise during the nearby NSP as the environment for noise propagation maybe changes a little bit after a short time section. This fact may

cause the residual noise in the enhanced speech somewhat fluctuate in different time sections. To overcome this fluctuation, we may adjust filter B all the time to overcome the fluctuation.

To sum up, in figure 2.3.2 we need only to adjust the coefficients of adaptive filter A to minimize the power of $e_1(k)$ during NSP and adjust the coefficients of adaptive filter B to minimize the power of $e_2(k)$ all the time, then the output $y_2(k)$ of filter B will be the enhanced speech. This is the two-channel CRANC algorithm.

The system functions of filter A and B are $H_{n_1n_0}(z)$ and $[1-H_{n_1n_0}(z)H_{s_0s_1}(z)]^{-1}$ respectively. $H_{n_1n_0}(z)$ is achieved by filter A during NSP and $[1-H_{n_1n_0}(z)H_{s_0s_1}(z)]^{-1}$ by filter B during the followed HSP. If during the next NSP the noise environment remains unchanged, the system function of filter A will not changed either in that time section. If during the next HSP the speech environment also remains unchanged (this means the position of the speaker and the position of the microphones and the speech propagation environment remain the same), the system function of filter B will not change either during that HSP. Otherwise, they will change themselves to follow the change of the noise environment and the speech environment. So, we may employ only a common VAD in a noise environment for the CRANC system. We adjust the coefficients of filter A only when a NSP time section is assured by VAD and freeze its coefficients all other times.

2.3.2 Adaptation algorithm

In figure 2.3.2 the adaptive filters A and B should converge to system functions $H_{n_1n_0}(z)$ and $[1-H_{n_1n_0}(z)H_{s_0s_1}(z)]^{-1}$ respectively. For simplicity, we usually choose FIR formation for filters A and B. There are many adaptation algorithms for these two filters such as LMS, NLMS, RSL and so on [48, 69, 70]. However, the adaptation algorithm will also affect the enhanced speech. If the adaptation algorithm offers smaller error and converges faster, the residual noise in the enhanced speech will be

smaller. Unfortunately, the adaptation algorithms that offer smaller error and converge faster usually have a higher computational complexity. Therefore, they might not be ideal for real-time implementation. Here the least squares lattice-ladder (LSLL) algorithm [69] is recommended for the adaptation of the filters for it is a trade-off of residual noise and computational complexity. Of course, if the enhancement of the speech does not need to be real-time implemented we may choose other algorithms such as RLS which may offer smaller residual noise [48].

2.4 Multichannel crosstalk resistant ANC

This section will extend two-channel CRANC to a multichannel processing. The extended CRANC is called Multichannel Crosstalk Resistant Adaptive Noise Cancellation (MCRANC).

In microphone array speech enhancement, the crosstalk effect of the speech signal in different channels of the acquired signals is very serious, especially for a small-size microphone array. Although a two-channel CRANC algorithm was proposed in the section 2.3, its noise cancellation ability is usually quite limited due to the multiple propagation paths and multiple noise sources. To increase the noise cancellation, multichannel CRANC is proposed as follows. Like CRANC, it also employs only two adaptive filters and has a low computational complexity. In particular, it has no strict limitations for the structure of the microphone array and the types of noises.

2.4.1 MCRANC algorithm

Suppose a speech $s(k)$ and noise (or noises) $n(k)$ are generated by independent sources, as indicated in figure 2.4.1. These signals arrive at microphone M_i through multi-paths and convert to $s_i(k)$ and $n_i(k)$. The impulse responses of the intermediate

media between the speech and noise sources and the acquiring microphone M_i are $h_{si}(k)$ and $h_{ni}(k)$ respectively. The audio signal acquired by microphone M_i can be represented by

$$x_i(k) = s_i(k) + n_i(k) \qquad i = 0,1,\cdots,N \tag{2.4.1}$$

N+1 is the number of microphones employed; k is the discrete time index. Since the acquired signals by the microphones contain noise and speech concurrently, crosstalk between noise and speech happens.

From figure 2.4.1 we have

$$s_i(k) = h_{si}(k) * s(k) \qquad i = 0,1,\cdots,N \tag{2.4.2}$$

$$n_i(k) = h_{ni}(k) * n(k) \qquad i = 0,1,\cdots,N \tag{2.4.3}$$

where $*$ is the convolution operator.

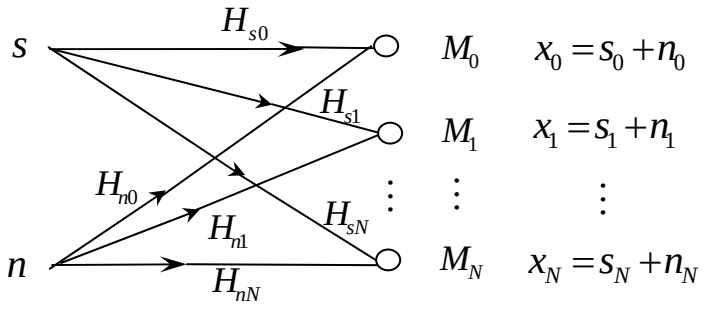

Figure 2.4.1 Speech and noise propagation between the emitting sources and the acquiring microphones

Let the impulse response of the intermediate environment between the input signal $s_i(k)$ and the output signal $s_j(k)$ be $h_{s_i s_j}(k)$, and the impulse response of the intermediate environment between the input signal $n_i(k)$ and the output signal $n_j(k)$ be $h_{n_i n_j}(k)$. Then

$$s_j(k) = h_{s_i s_j}(k) * s_i(k) \qquad i, j = 0,1,\cdots,N \tag{2.4.4}$$

$$n_j(k) = h_{n_i n_j}(k) * n_i(k) \qquad i, j = 0,1,\cdots,N \tag{2.4.5}$$

Through (2.4.4)-(2.4.5) we have

$$H_{s_i s_j}(z) = \frac{H_{sj}(z)}{H_{si}(z)} \qquad i, j = 0,1,\cdots,N \tag{2.4.6}$$

$$H_{n_i n_j}(z) = \frac{H_{nj}(z)}{H_{ni}(z)} \qquad i, j = 0,1,\cdots,N \tag{2.4.7}$$

where $H_{si}(z)$ is the z-transform of $h_{si}(k)$ and so forth for other notations.

In the practical environment, noise emitted from a certain source may propagate to microphone M_i through multiple paths including direct propagations, reflections and refractions. The noise may also be emitted from multiple sources. We consider those noises are from a combined source and all propagation paths are included in the combined transfer function $H_{ni}(z)$, which has an impulse response $h_{ni}(k)$.

Take $x_0(k)$ as the main channel signal acquired by microphone M_0, and others $x_i(k)$ $(i = 1,\cdots,N)$ as the referential signals acquired by other *N* microphones. Assume that the main-channel signal is correlated with the referential-channel signals, which is usually a valid assumption if the microphones are located in close proximity. As the referential signals contain both speech and noise, the common Multichannel ANC (MANC) method will not be an appropriate method for the speech enhancement. That is because the crosstalk effect violates the working conditions and consequently both speech and noise will be cancelled out.

MCRANC algorithm is shown in figure 2.4.2. It consists of a VAD and two adaptive filters A and B. It takes use of the characteristic that for a speech signal the time index can be divided into a series of Non Speech Periods (NSP) and Having Speech Periods (HSP).

During a Non Speech Period (NSP), microphones $M_0, M_1,\cdots,M_N$ acquire only noise $n_0(k)$, $n_1(k),\cdots,n_N(k)$. Take $n_0(k)$ as the main channel noise and $n_1(k),\cdots,n_N(k)$ as the referential channel noises. Input all referential channel noises to adaptive filter A for the first stage of MCRANC to cancel the main channel noise

$n_0(k)$.

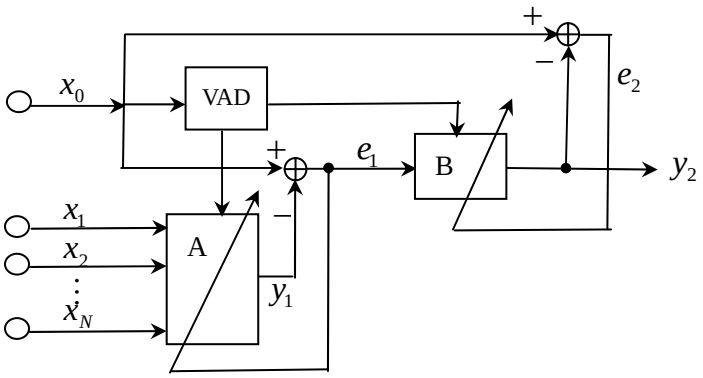

Figure 2.4.2 Structure of MCRANC

In two-channel ANC, we use only one channel of referential noise $n_i(k)$ to cancel $n_0(k)$ and we usually use the FIR type for the adaptive filter A, i.e.

$$n_0(k) = \mathbf{w}_i \mathbf{n}_i(k) + e_{i1}(k) \tag{2.4.8}$$

$$\mathbf{w}_i = (w_{i0}, w_{i1}, \cdots, w_{iL'}) \tag{2.4.9}$$

$$\mathbf{n}_i(k) = [n_i(k), n_i(k-1), \cdots, n_i(k-L')]^T \tag{2.4.10}$$

where $\mathbf{w}_i$ is the coefficients of filter A ; $L'+1$ is the length of filter A; $e_{i1}(k)$ is the prediction error by using $\mathbf{n}_i(k)$ to predict $n_0(k)$. The error power $P[e_{i1}(k)]$ has great affinity to the final enhanced speech. However, in the real environment, no matter how we optimize $\mathbf{w}_i$ and even how we increase L' or choose a proper L', the minimal error power $P[e_{i1}^*(k)]$ in equation (2.4.8) is usually not small enough. This might result from the complicated propagations of the noises in the intermediate media and the mismatch of the microphones. The noise signal acquired from M_i is not highly correlated with the noise signal from M_0. In the audio situation, it is found that the farther the distance between two microphones M_i and M_0, the weaker the correlation between the signals acquired by them and thus the greater the minimal error power $P[e_{i1}^*(k)]$.

If multiple referential signals $n_1(k),\cdots,n_N(k)$ are used as input to the FIR filter A to cancel $n_0(k)$, we have

$$n_0(k) = \mathbf{w}\mathbf{n}(k) + e_1(k) \tag{2.4.11}$$

$$\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_N) \tag{2.4.12}$$

$$\mathbf{w}_i = (w_{i0}, w_{i1}, \cdots, w_{iL}) \tag{2.4.13}$$

where $\mathbf{w}$ is the coefficients of the filter and it is a row vector with $N(L+1)$ dimension, and

$$\mathbf{n}(k) = [\mathbf{n}_1(k), \mathbf{n}_2(k), \cdots, \mathbf{n}_N(k)]^T \tag{2.4.14}$$

$$\mathbf{n}_i(k) = [n_i(k), n_i(k-1), \cdots, n_i(k-L)] \tag{2.4.15}$$

where $\mathbf{n}(k)$ is a column vector with $N(L+1)$ elements, and $e_1(k)$ is the prediction error.

If we take $\mathbf{w} = (\mathbf{0}, \cdots, \mathbf{0}, \mathbf{w}_i, \mathbf{0} \cdots, \mathbf{0})$, where $\mathbf{0}$ represent a row vector with $L+1$ zeros, we have

$$e_1(k) = n_0(k) - \mathbf{w}\mathbf{n}(k)$$

$$= n_0(k) - \mathbf{w}_i\mathbf{n}_i(k) = e_{i1}(k)$$

So, the minimal error powers satisfy the following inequation (2.4.16) if the coefficient vectors $\mathbf{w}$ and $\mathbf{w}_i$ are optimized.

$$P[e_1^*(k)] \le P[e_{i1}^*(k)] \tag{2.4.16}$$

In particular, $P[e_1^*(k)]$ is usually much smaller than $P[e_{1i}^*(k)]$ if the noises are from multiple sources or the noises have more than one propagation path to the microphone array. This means the residual noise in the main channel, after the noise cancellation by use of N referential noises $n_1(k),\cdots,n_N(k)$, is much smaller than the residual noise by use of only one referential noise $n_i(k)$. This fact indicates that increasing the referential microphones may increase the correlation between the main channel noise

and the referential noises.

However, a too bigger N (this means too many microphones employed in the array) and too bigger L will make the optimization of $\mathbf{w}$ more difficult and inaccurate in practical computations. This may lead inequation (2.4.16) to be untrue. So, a proper values for microphone number N and sample delay number L are actually needed in practice.

Denote the optimal coefficient vector of filter A for minimizing the error power $e_1(k)$ in (2.4.11) as

$$\begin{aligned}\mathbf{w}^* &= (\mathbf{w}_1^*, \mathbf{w}_2^*, \cdots, \mathbf{w}_N^*) \\ &= (w_{10}^*, w_{11}^*, \cdots, w_{1L}^*, w_{20}^*, w_{21}^*, \cdots, w_{2L}^*, \cdots, w_{N0}^*, w_{N1}^*, \cdots, w_{NL}^*)\end{aligned} \tag{2.4.17}$$

The corresponding minimal power is noted as $P[e_1^*(k)]$. To get $\mathbf{w}^*$, we need only to adjust the coefficients $\mathbf{w}$ of filter A to minimize the power of $e_1(k)$ in figure 2.4.2.

Then during a Having Speech Period (HSP), which follows the above-mentioned NSP time section, we assume the environment will not change or only change slowly for the noise propagations. As a result, the noise impulse response $h_{n_i n_0}(k)$ in this HSP section would be almost the same as that in the previous NSP section. Thus

$$\begin{aligned}y_1(k) &= \mathbf{w}^* \mathbf{x}(k) \\ &= \mathbf{w}^* [\mathbf{s}(k) + \mathbf{n}(k)] \\ &= \mathbf{w}^* \mathbf{s}(k) + \mathbf{w}^* \mathbf{n}(k) \\ &= \mathbf{w}^* \mathbf{s}(k) + [n_0(k) - e_1^*(k)]\end{aligned} \tag{2.4.18}$$

Here $\mathbf{s}(k)$ and $\mathbf{x}(k)$ are represented in a similar way to $\mathbf{n}(k)$ described by (2.4.14) and (2.4.15). For example,

$$\mathbf{s}(k) = [\mathbf{s}_1(k), \mathbf{s}_2(k), \cdots, \mathbf{s}_N(k)]^T$$

$$\mathbf{s}_i(k) = [s_i(k), s_i(k-1), \cdots, s_i(k-L)]$$

So we have

$$\begin{aligned} e_1(k) &= x_0(k) - y_1(k) \\ &= [s_0(k) + n_0(k)] - [\mathbf{w}^* \mathbf{s}(k) + n_0(k) - e_1^*(k)] \\ &= s_0(k) - \mathbf{w}^* \mathbf{s}(k) + e_1^*(k) \\ &= p(k) + e_1^*(k) \end{aligned} \tag{2.4.19}$$

where

$$p(k) = s_0(k) - \mathbf{w}^* \mathbf{s}(k) \tag{2.4.20}$$

Take z-transform to both sides of equation (2.4.20) to get

$$\begin{aligned} P(z) &= S_0(z) - Z[\sum_{i=1}^{N}\sum_{l=0}^{L} w_{il}^* s_i(k-l)] \\ &= S_0(z) - Z[\sum_{i=1}^{N}\sum_{l=0}^{L} w_{il}^* h_{s_0 s_i}(k-l) * s_0(k-l)] \\ &= S_0(z) - \sum_{i=1}^{N}\sum_{l=0}^{L} w_{il}^* z^{-2l} H_{s_0 s_i}(z) S_0(z) \\ &= [1 - \sum_{i=1}^{N}\sum_{l=0}^{L} w_{il}^* z^{-2l} H_{s_0 s_i}(z)] S_0(z) \\ &= \tilde{H}(z) S_0(z) \end{aligned} \tag{2.4.21}$$

where

$$\tilde{H}(z) = 1 - \sum_{i=1}^{N}\sum_{l=0}^{L} w_{il}^* z^{-2l} H_{s_0 s_i}(z) \tag{2.4.22}$$

So we see that $p(k)$ is actually a distorted speech and it is correlated with $s_0(k)$.

Furthermore, the power of $p(k)$ is usually not so small as the power of the error $e_1^*(k)$. This means, to the noisy speech, the power of the speech will not decrease as much as the power of the noise does. This is because the speech signal has different propagation paths from the propagation paths of noises. It is actually affected by the propagation environment, the positions of the microphone array and the sources of the speech and the noises. It may also be regarded as the zero-point forming technique in array signal processing. After the first-stage of processing with filter A, the microphone

array forms the zero points to the directions of noises from which the noises propagate to the array by directional, refractive and reflective paths. However, the directions of propagation paths for the speech signal will not all fall into these zero points.

The SNR of $e_1(k)$ is usually greatly improved compared with the noisy signal $x_0(k)$, where the signal is $p(k)$ and noise is $e_1^*(k)$. However, signal $p(k)$ in $e_1(k)$ is not the approximation of $s_0(k)$, but a distorted signal of $s_0(k)$. The distortion usually becomes more serious with the increase of the microphones. The second-stage with filter B is used to change the distorted speech $p(k)$ into the desired speech $s_0(k)$.

For this purpose, we need only adjust the coefficients of filter B to minimize the power of $e_2(k)$ in figure 2.4.2 under the assumption that the speech signal is not correlated with the noises. The reason is totally the same as that described in section 2.3. The higher the SNR of $e_1(k)$, the smaller the error between $y_2(k)$ and $s_0(k)$.

Similarly, to overcome the fluctuation of the remaining noise in the enhanced speech, it is best to adjust filter B all the time to minimize the power of $e_2(k)$ while adjusting filter A to minimize the power of $e_1(k)$ only during NSP. Then the output $y_2(k)$ of filter B will be the optimal estimation of the speech $s_0(k)$ under MMSE criteria.

It is obvious that the system function of filter B approaches to $\tilde{H}^{-1}(z) = [\tilde{H}(z)]^{-1}$. By figure 2.4.2 and equations (2.4.19) and (2.4.20), we have

$$\begin{aligned} Y_2(z) &= \tilde{H}^{-1}(z)E_1(z) \\ &= \tilde{H}^{-1}(z)[\tilde{H}(z)S_0(z) + E_1^*(z)] \\ &= S_0(z) + \tilde{H}^{-1}(z)E_1^*(z) \end{aligned} \tag{2.4.23}$$

Thus

$$y_2(k) = s_0(k) + \tilde{h}^{-1}(k) * e_1^*(k) \tag{2.4.24}$$

where $\tilde{h}^{-1}(k)$ is the inverse z-transform of $\tilde{H}^{-1}(z)$.

Again, similar to section 2.3, we may employ only a common VAD in a noisy environment for our MCRANC system. We may adjust the coefficients of filter A only when NSP is assured by VAD and freeze its coefficients all other times.

2.4.2 Computational complexity

For filter A and B in MCRANC as shown in figure 2.4.2, we may employ adaptation algorithms such as LMS, NLMS, RLS, BFTF and LSLL, etc. [48, 69, 70]. The algorithm may be selected according to the calculation ability of DSP and the requirement of the applications. If the LSLL algorithm is used, it can be found that the floating point operations in every sampling interval is less than

$$24(L_{max}+1)$$

where L_{max} is the maximum length of filter A and B. Here we did not take the operations of VAD into account. In practice, we may usually take $L_{max}<100$. So, if the sampling rate is 8K (the commonly-used sampling rate for a speech signal), the computational complexity in a second will be less than

19.2 MFLOPS

If a simple LMS algorithm is used, the corresponding complexity will drop to

3.2 MFLOPS

So, MCRANC as proposed in this section is quite suitable for real-time implementation. For example, DSP TMS320VC33 from Texas Instruments Company has the computation ability of 150 MFLOPS.

2.5 Experimental results

One simulation experiment and two experiments in real environments will be presented in this section. The experimental results will show that CRANC and

MCRANC are quite effective for small microphone array speech enhancement. The algorithms may be used with different kinds of noises.

2.5.1 Simulation experiment

This simulation experiment will indicate the effectiveness of two-channel CRANC for speech enhancement. In this experiment, the speech signal and the noise signal are real signals while the propagation transfer functions for speech and noise are simulative.

First of all, a section of speech and a section of music are recorded respectively with a computer. The music is viewed as noise. The sampling rate for speech and noise is 8 kHz.

For simulation, discretionarily assume the transfer functions for speech from its source to the main microphone and referential microphone are respectively

$$H_{s_0}(z)=[0.0408\ 0.0817\ 0.1633\ 0.1225\ 0.0408\ 0.2042\ 0.2450\ 0.0817]\,\mathbf{z} \quad (2.5.1)$$

$$H_{s_1}(z)=[0.1293\ 0.1293\ 0.0970\ 0.2587\ 0.0647\ 0.0323\ 0.0970\ 0.1617]\,\mathbf{z} \quad (2.5.2)$$

where $\mathbf{z}=[1\quad z^{-1}\quad z^{-2}\quad z^{-3}\quad z^{-4}\quad z^{-5}\quad z^{-6}\quad z^{-7}]^{T}$, while the transfer functions for noise from its source to the main microphone and referential microphone are respectively

$$H_{n_0}(z)=[0.1187\ 0.2969\ 0.1781\ 0.0594\ 0.0000\ 0.0000\ 0.2375\ 0.0594]\,\mathbf{z} \quad (2.5.3)$$

$$H_{n_1}(z)=[0.1309\ 0.2182\ 0.2618\ 0.0000\ 0.0873\ 0.0000\ 0.0436\ 0.2182]\,\mathbf{z} \quad (2.5.4)$$

Thus, the speech and noise acquired by the main microphone will be $s_0(k)=h_{s_0}(k)*s(k)$ and $n_0(k)=h_{n_0}(k)*n(k)$, where $h_{s_0}(k)$ and $h_{n_0}(k)$ are the impulse response corresponding to transfer functions $H_{s_0}(z)$ and $H_{n_0}(z)$ respectively. Similarly, we can get the speech and noise acquired by the referential microphone $s_1(k)$ and $n_1(k)$.

The waveforms of clean speech $s_0(k)$ and pure noise $n_0(k)$ are depicted in figure 2.5.1 (a) and (b) respectively.

The mixed signals of speech and noise actually acquired by the main microphone and referential microphone are noisy speech $x_0(k) = s_0(k) + n_0(k)$ and $x_1(k) = s_1(k) + n_1(k)$ respectively. They are depicted in figure 2.5.1 (c) and (d).

In figure 2.5.2, (a) and (b) are the waveforms of pure speech $s_0(k)$ and noisy speech $x_0(k)$ respectively. (c) is the waveform of the enhanced speech $\hat{s}_0(k)$ by use of a common two-channel ANC algorithm when taking $x_0(k)$ as the main signal and $x_1(k)$ as the referential signal. (d) is the waveform of the enhanced speech $\tilde{s}_0(k)$ by use of the two-channel CRANC algorithm proposed in section 2.3 when taking $x_0(k)$ as the main signal and $x_1(k)$ as the referential signal.

In figure 2.5.3, (a), (b), (c) and (d) are the spectrograms of the corresponding signals depicted in figure 2.5.2.

The SNRs of the noisy speech $x_0(k)$, enhanced speech $\hat{s}_0(k)$ by common ANC algorithm and the enhanced speech $\tilde{s}_0(k)$ by proposed CRANC algorithm are SNR(x_0)= 0.31 (dB), SNR($\hat{s}_0$)= 7.63 (dB) and SNR($\tilde{s}_0$)= 14.32 (dB) respectively. Here SNR takes the common definition given in equation (1.4.1) because the pure speech signal $s_0(k)$ is available in this simulation experiment.

In the ANC processing for the enhanced speech $\hat{s}_0(k)$, the filter is a FIR filter with length L=32 and a LSLL adaptation algorithm is employed. In the CRANC processing for the enhanced speech $\tilde{s}_0(k)$, the FIR filters A and B have lengths L=32 and L_B=48 and a LSLL adaptation algorithm is again employed.

From the SNRs of the noisy speech $x_0(k)$, enhanced speech $\hat{s}_0(k)$ and enhanced speech $\tilde{s}_0(k)$, we can find the advantages of the proposed CRANC algorithm.

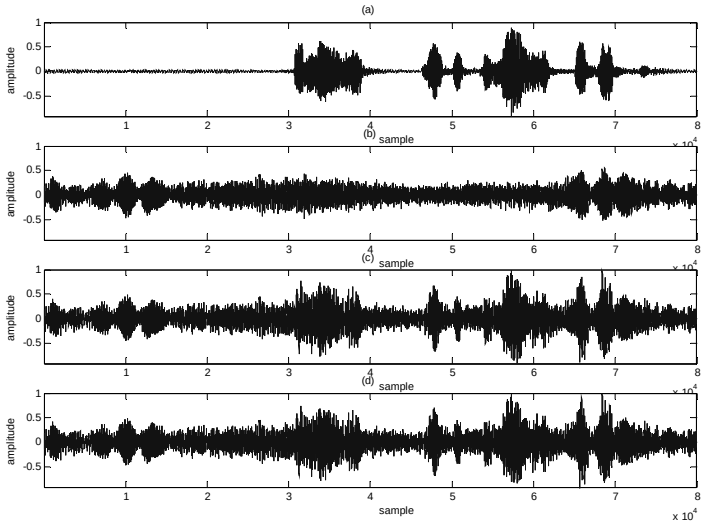

Figure 2.5.1 Speech, noise and their mixed signals

(a) Clean speech

(b) Pure noise

(c) Acquired noisy speech by main microphone

(d) Acquired noisy speech by referential microphone

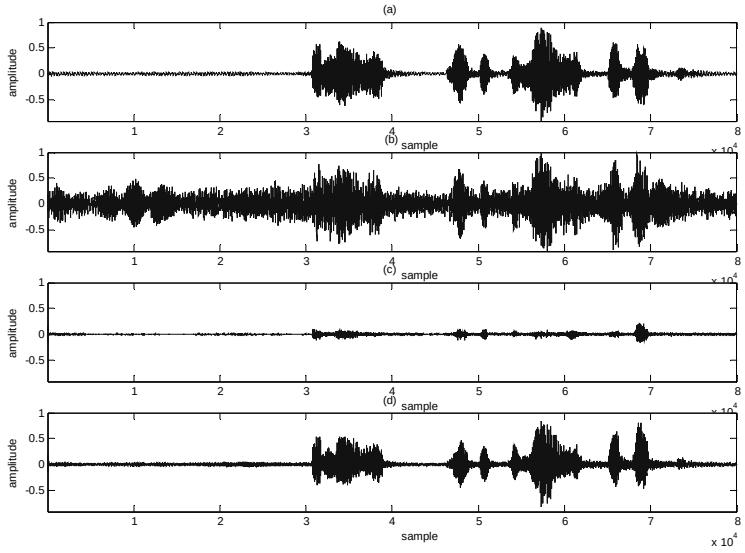

Figure 2.5.2 Simulation results

(a) Clean speech

(b) Acquired noisy speech by main microphone

(c) Enhanced speech by ANC

(d) Enhanced speech by CRANC

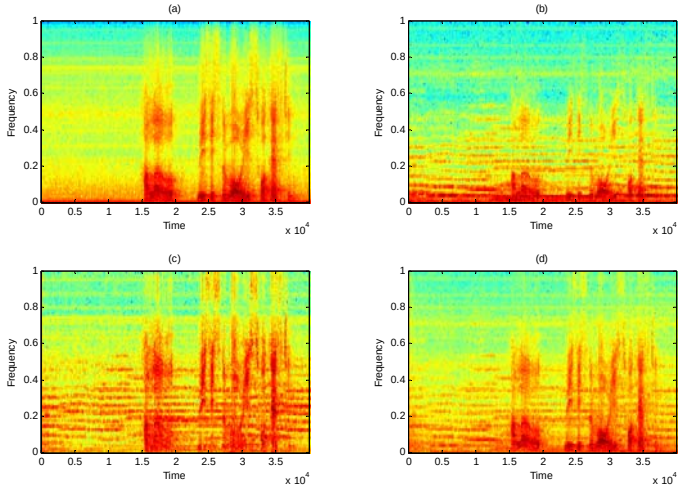

Figure 2.5.3 Spectrograms for the signals in figure 2.5.2

(a) Spectrogram of clean speech

(b) Spectrogram of acquired noisy speech by main microphone

(c) Spectrogram of enhanced speech by ANC

(d) Spectrogram of enhanced speech by CRANC

2.5.2 Experiments in real environments

Several experiments were made in real environments. The experimental results verify the effectiveness and advantages of the MCRANC algorithm.

Experiment 1

This experiment was carried out in a common research room with dimensions of 8x5x3m. Four microphones $M_0, M_1, \cdots, M_3$ were closely placed in a quite free formation. The maximum distance between any two microphones was only 2cm. The noise was generated from an improperly tuned radio 1m away from the microphones. The speech was from a person 0.5m away from the microphones. The sampling rate was 8 kHz.

Figure 2.5.4 shows the results. (a) is the noisy speech signal $x_0(k)$ acquired by the main microphone. Its SNR is 2.75 dB. The signals acquired by the referential microphones look like almost the same as $x_0(k)$. (b) is the enhanced speech signal by the ordinary Multichannel ANC (MANC) algorithm. The SNR improvement is 9.1 dB but the speech is seriously corrupted. (c) is the enhanced speech by the two-channel

CRANC algorithm proposed in section 2.3, with SNR improvement 8.6 dB. (d) is the enhanced speech by MCRANC proposed in section 2.4, with SNR improvement 17.8 dB. Here SNR is calculated by definition (1.4.3) since the clean speech signal $s_0(k)$ is unavailable. By listening, one may find that both (c) and (d) have better quality than (a) and (b), and (d) has the best quality.

Figure 2.5.5 is a zoom section of figure 2.5.4. It deals with only the pure noise section. From it we may see that MANC and MCRANC have a high noise cancellation ability.

Figure 2.5.6 is also a zoom section of figure 2.5.4, however it deals with only the noisy speech section. From it we may see that MANC creates great damage to the speech signal and that both CRANC and MCRANC create less damage to the speech signal. But MCRANC outperforms CRANC for MCRANC contains less noise.

In the processing of MANC to get the output as shown in figure 2.5.4 (b), the length of the FIR filter is 99 and the Normalized Least Mean Square (NLMS) algorithm is employed with the learning rate $\mu = 0.01$. In the processing of the two-channel CANC to get the output as shown in figure 2.5.4 (c), the length of FIR filter A is 99, which means the sample delay for the referential signal is 98, and the length of FIR filter B is 49. Both filters employ a NLMS adaptation algorithm with learning rate $\mu = 0.01$. In the processing of MCANC to get the output as shown in figure 2.5.4 (d), the length of FIR filter A is 99, which means the sample delay for every referential signal is 32, and the length of FIR filter B is 49. Both filters also employ a NLMS adaptation algorithm with learning rate $\mu = 0.01$.

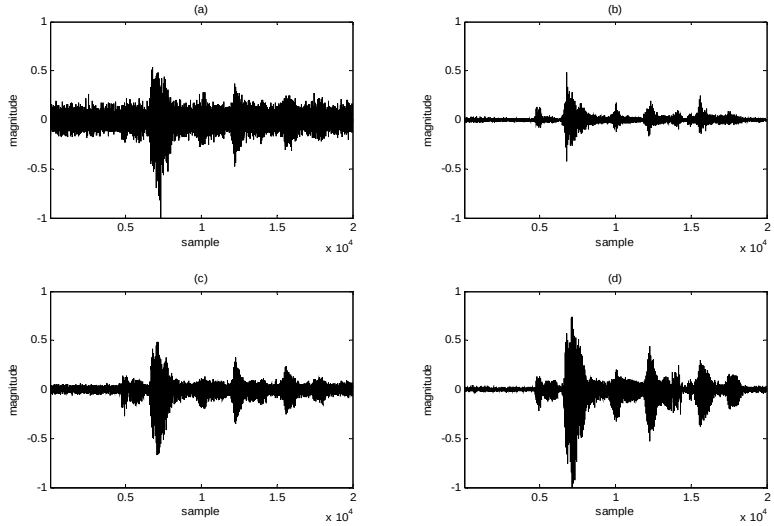

Figure 2.5.4 Experimental results

(a) Noisy speech signal
(b) Enhanced speech by common MANC
(c) Enhanced speech by two-channel CRANC
(d) Enhanced speech by proposed MCRAN

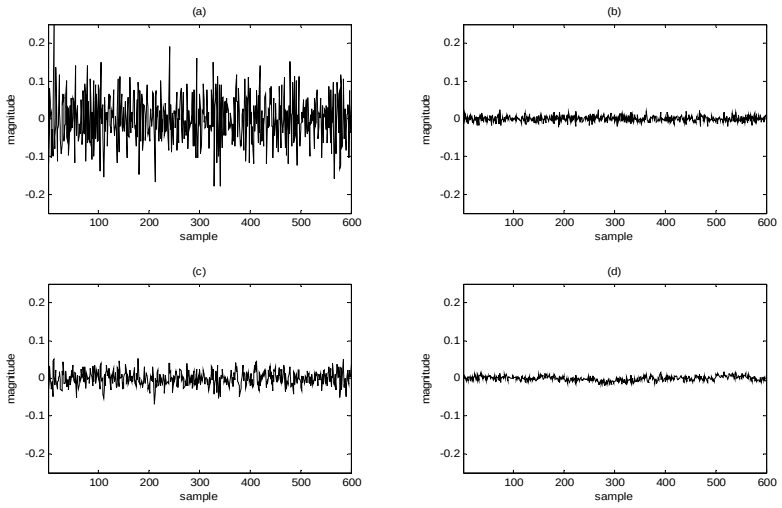

Figure 2.5.5 A section of figure 2.5.4 (pure noise)

(a) Pure noise
(b) Output noise by common MANC
(c) Output noise by two-channel CRANC
(d) Output noise by proposed MCRANC

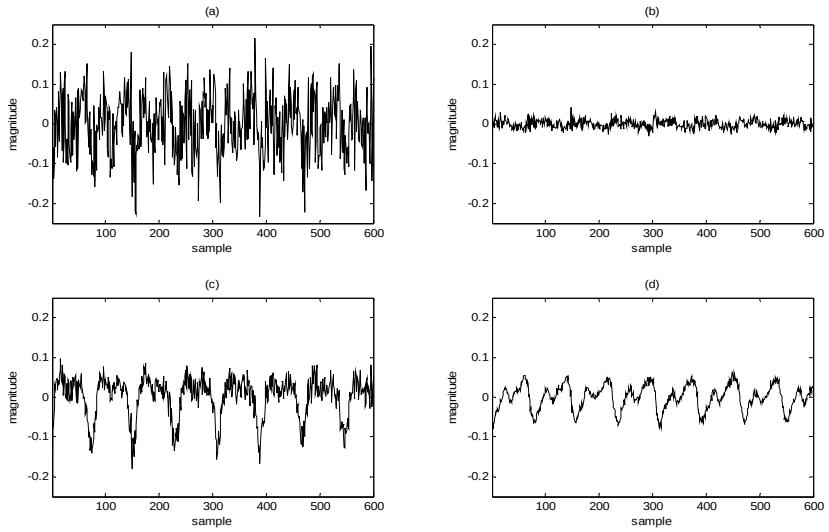

Figure 2.5.6 A section of figure 2.5.4 (noisy speech)

(a) Noisy speech
(b) Enhanced speech by common MANC
(c) Enhanced speech by two-channel CRANC
(d) Enhanced speech by proposed MCRANC

Experiment 2

This experiment was made in a common study room of dimensions 5x4x2.8m. The array was put on a desk. The center of the array was 1.4m from the front wall, 1.8m from the left wall and 1.23m from the floor. There were two sofas, a cabinet and two desks in the room. The room had two glass windows and a wooden door, all of which were closed.

Four small microphones were set up in a planar array with an aperture of less than 5cm as shown in figure 2.5.7. The speech and the noises were generated concurrently by loudspeakers from different locations. The speech data was from a section of recorded speech on a computer and the noise data was from database NoiseX92. The sampling rate used to digitize the acquired signals was 8 kHz.

One of the experiment cases is shown in figure 4.3.10. For simplicity, the figure is a planar one since the loudspeakers emitting speech and noises have almost the same height from the floor as the array used in the experiment. In this case, the speech loudspeaker is placed 30cm in front of the microphone array at (0,30). Noise

loudspeakers concurrently emit Volvo, Leopard, Factory2 and White noises. They are positioned at (-100,100), (50,50), (200,250) and (0,100)cm respectively.

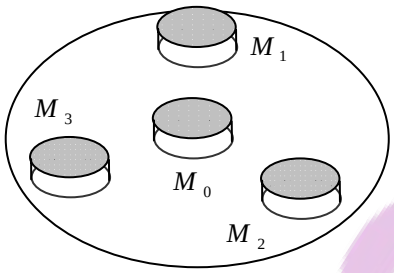

Figure 2.5.7 Employed 4-microphone planar array

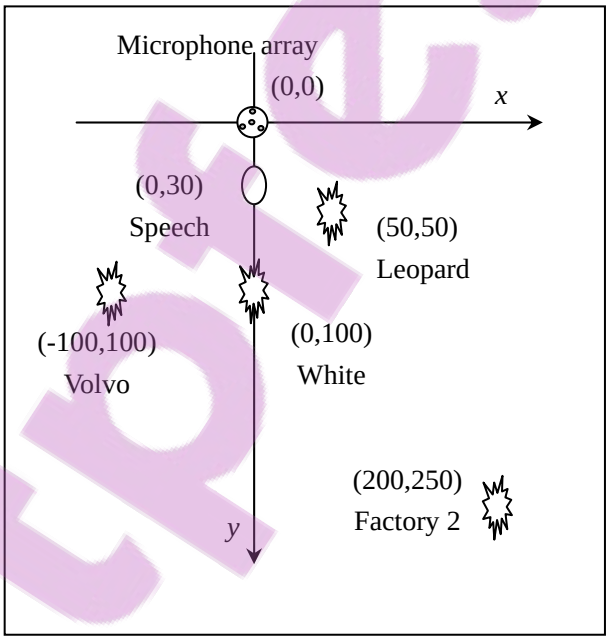

Figure 2.5.8 One of the experiment environments

Table 2.5.1 presents the SNRs of the original noisy speech and the enhanced speech obtained by use of the MCRANC algorithm when the speech source is at location (0,30) and different types of noises at different source locations. Here the SNR is calculated according to formula (1.4.3).

In this table, an outstanding algorithm named Modified Generalized Sidelobe Canceling (MGSC) is used for comparison [44, 72]. The table also presents the SNRs

of the enhanced speeches obtained by use of the MGSC algorithm.

In a real environment, MGSC is one of the excellent algorithms for canceling noise using a microphone array. It will be introduced in section 5.2 of chapter 5. MGSC is an improved GSC by adding a signal delayer z^{-d} and a voice activity detector VAD to the Generalized Sidelob Canceling (GSC) structure. Although there have been more advanced and complicated algorithms published in recent years, such as the Transfer Function GSC (TF-MGSC) proposed by Gannot in 2004 [38], MGSC still have more SNR improvements. That is why MGSC is used for comparison.

However, both MCRANC and MGSC need a VAD. TF-GSC has other advantages over MGSC and MCRANC because it does not need a VAD.

The combined algorithms, such as those to employ GSC and then cascaded with a single-channel speech enhancement algorithm, is not used for comparison because MGSC or MCRANC can also be cascaded by other speech enhancement algorithms. This will be presented in next chapter.

Figure 2.5.9 depicts the three lines of SNRs in table 2.5.1 for the original noisy speech, the speech enhanced by MGSC and the speech enhanced by MCRANC. Every line presents the SNRs in 56 different cases containing different noises and locations.

Table 2.5.2 gives out the SNRs and MOS scores of the original noisy speech and the enhanced speeches by MGSC and MCRANC under multiple noise sources and different locations of the speech source.

Figure 2.5.10 depicts the three lines of SNRs in table 2.5.2 for the original noisy speech, the speech enhanced by MGSC and the speech enhanced by MCRANC. Every line presents the SNRs in eight different cases containing different numbers of noises in different locations to emit together, along with different locations of speech sources. Figure 2.5.11 depicts the 3 lines of MOS scores corresponding to the above 8 cases.

From the experimental results in the above tables and figures, we may find the proposed MCRANC algorithm gives more SNR improvement than MGSC in small microphone array based speech enhancement.

Table 2.5.1 The SNRs (dB) of original noisy speech and the enhanced speech by MGSC and MCRANC when the speech source is at (0,30) and different noises are at different source locations

Noise Type / Noise Location / Algorithm	Babble x 0.5	Bucaneer1 x 0.5	Factory2	Leopard	Pink	Volvo	White	Volvo x 5
(200,250)								
Original	14.60	15.76	14.01	14.39	18.36	22.18	21.31	11.48
MGSC	20.80	16.10	14.65	17.82	18.20	22.92	20.87	12.54
MCRANC	20.83	21.46	21.23	19.86	22.71	24.22	23.18	21.30
(-100,100)								
Original	13.05	14.54	11.93	11.74	16.62	20.71	20.13	10.25
MGSC	13.06	13.90	11.32	14.25	15.47	22.68	19.87	11.55
MCRANC	20.48	20.34	19.23	19.47	21.33	24.06	22.67	20.23
(0,100)								
Original	11.16	11.63	10.22	8.97	13.78	19.31	16.55	9.32
MGSC	11.29	11.75	10.42	13.31	13.90	22.44	17.38	12.64
MCRANC	17.70	18.75	19.03	18.10	20.14	23.72	20.95	20.62
(-50,50)								
Original	13.62	12.49	9.40	8.52	15.03	23.13	17.49	10.51
MGSC	12.90	12.01	8.22	11.86	14.54	22.73	18.22	7.51
MCRANC	18.84	9.10	17.49	14.53	20.38	23.58	21.54	19.45
(0,50)								
Original	10.59	8.62	6.72	5.06	11.53	18.80	15.94	8.04
MGSC	9.65	7.69	5.45	9.23	9.95	21.10	14.97	9.68
MCRANC	19.11	15.92	16.36	15.68	16.77	23.12	20.17	18.12
(30,30)								
Original	5.12	4.75	4.40	-0.97	8.37	18.07	10.84	6.40
MGSC	7.26	6.29	8.05	6.76	8.88	19.88	11.48	9.97
MCRANC	15.96	14.03	16.89	14.56	17.25	23.22	16.57	18.12
(0,20)								
Original	4.30	2.08	-0.56	-0.08	4.48	15.56	6.74	2.55
MGSC	2.42	0.66	-2.43	2.23	1.40	15.85	6.75	2.32
MCRANC	16.61	-4.54	10.27	10.49	10.18	21.64	8.23	10.38

- Different rows present different locations of the noises. Different columns present different types of noises. In each row, Original directs to SNRs of the original noisy speeches while MGSC and MCRANC direct to SNRs of the enhanced speech by MGSC and MCRANC respectively.

- Noise x 0.5 presents the noise, which has only half the amplitude of the original noise. Noise x 5 presents the noise, which has the amplitude 5 times the amplitude of the original noise.

Table 2.5.2 The SNRs (dB) and MOS scores of the original noisy speech and the enhanced speech by MGSC and MCRANC under multiple noise sources and different locations of the speech source

Sources of speech and noises / SNR(dB) / MOS / Algorithms	S(0,20) L(30,30)+V(-100,100)	S(0,20) V(-100,100)+F(200,250)	S(0,20) L(30,30)+V(-100,100)+F(200,250)	S(0,20) L(30,30)+V(-100,100)+F(200,250)+W(0,100)	S(-50,50)x2 L(30,30)+V(-100,100)	S(-50,50)x2 V(-100,100)+F(200,250)	S(-50,50)x2 L(30,30)+V(-100,100)+F(250,250)	S(-50,50)x2 L(30,30)+V(-100,100)+F(200,250)+W(0,100)
SNR(dB)								
Original	-1.23	13.48	-1.23	-1.24	-20.12	8.48	-20.02	-23.67
MGSC	6.46	14.57	5.60	4.68	-0.64	9.17	-1.16	-1.28
MCRANC	13.48	21.32	15.71	14.85	6.50	12.46	6.04	4.91
MOS								
Original	3.0	4.6	3.0	3.0	0.2	4.2	0.2	0.2
MGSC	3.8	4.4	3.8	3.6	3.0	4.2	2.8	2.8
MCRANC	4.4	5.0	4.6	4.6	4.0	4.6	3.8	3.6

- In the first row the different columns present different speech locations and different noises and their locations. The second row gives out the SNRs, where Original directs to SNRs of the original noisy speeches while MGSC and MCRANC direct to SNRs of the enhanced speeches by MGSC and MCRANC respectively. The third row gives out the corresponding MOS scores.
- S(0,30) presents the speech source at (0,30). L(30,30)+V(-100,100) presents the

Leopard noise source at (30,30) and the Volvo noise source at (-100,100). They emit from these sources simultaneously. So forth for other notations.

- Speech x 2 presents the speech, which has the amplitude 2 times the amplitude of the original speech.

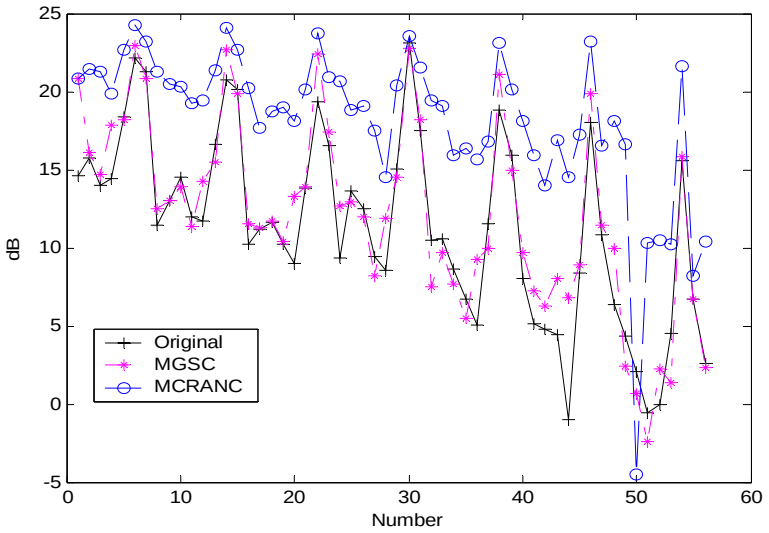

Figure 2.5.9 The three lines of SNRs in table 2.5.1 for the original noisy speech, the speech enhanced by MGSC and the speech enhanced by MCRANC

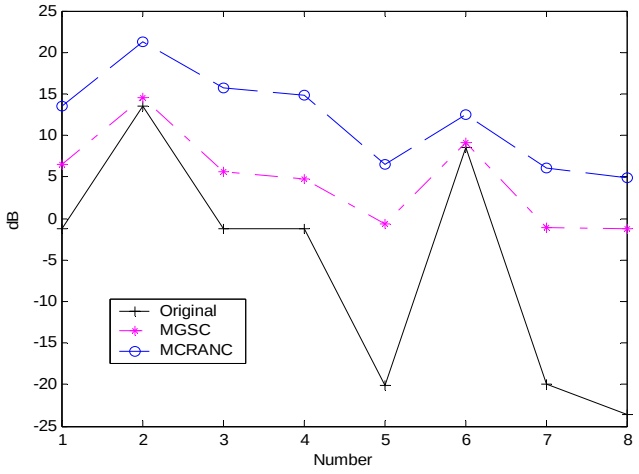

Figure 2.5.10 The three lines of SNRs in table 2.5.2 for the original noisy speech, the speech enhanced by MGSC and the speech enhanced by MCRANC in eight cases.

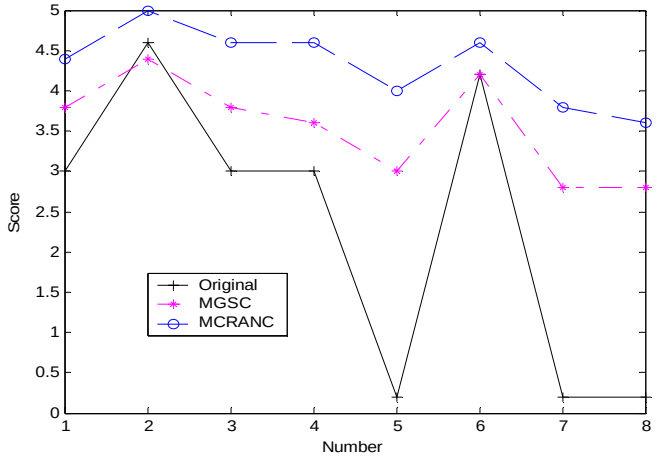

Figure 2.5.11 The three lines of MOS scores in table 2.5.2 for the original noisy speech, the speech enhanced by MGSC and the speech enhanced by MCRANC in eight cases.

2.6 Summary

In this chapter two-channel Crosstalk Resistant Adaptive Noise Cancellation (CRANC) and Multichannel Crosstalk Resistant Adaptive Noise Cancellation (MCRANC) are proposed. Their principles for noise cancellation and the computational complexity of MCRANC are presented. A simulation experiment and two experiments in real environments are also presented for speech enhancement using a small microphone array. MCRANC employs only two adaptive filters and a Voice Activity Detector (VAD). It has low computational complexity and it can be used with different kinds of noises. It also has no strict limitations to the structure of the microphone array. Experimental results indicate that the MCRANC is a suitable algorithm for speech enhancement using a small microphone array. It outperforms the outstanding Modified Generalized Sidelobe Canceling (MGSC) algorithm and may achieve significant speech improvement performance.

Chapter 3 Combined Algorithms With MCRANC

Three combined algorithms employing Multichannel Crosstalk Resistant Adaptive Noise Cancellation (MCRANC) and other existing algorithms are proposed in this chapter. The first combined algorithm is the combination of MCRANC with the single-channel Improved Spectrum Subtraction (ISS). The second is the combination with Delay And Sum (DAS) beamforming which is a basic algorithm to deal with sensor array signals. The third is the combination with multichannel Weiner Post-Filtering (WPF). Theoretic analysis and experimental results verify that the combined algorithms achieve better speech enhancement performances.

3.1 Introduction

The Multichannel Crosstalk Resistant Adaptive Noise Cancellation (MCRANC) algorithm proposed in chapter 2 may be combined with other existing algorithms to achieve better speech enhancement performances.

In one way, the output of MCRANC is an enhanced single-channel speech and it inevitably retains some residual noise. Therefore, it still can be viewed as single-channel noisy speech, and thus it can be further enhanced with single-channel speech enhancement algorithms. For example, it can be further enhanced with the spectral subtraction algorithm [15], the masking properties based algorithm [118], the wavelet denoising algorithm [9] and so on. Since MCRANC generally improves the Signal to Noise Ratio (SNR) of noisy speech, the enhanced speech by MCRANC is more likely to meet the requirements and conditions for many single-channel speech enhancement algorithms. As a result, the combined algorithm may achieve a better speech enhancement performance than MCRANC or a single-channel algorithm alone.

In another way, MCRANC can also be combined with other array signal enhancement algorithms to get more powerful algorithms. For example, it may be combined with the Delay And Sum (DAS) algorithm [49, 123], the Post-Filtering (PF) algorithm [131, 79] and so forth.

In this chapter, the combination of MCRANC with Improved Spectral Subtraction (ISS) and the combinations with DAS and PF are presented respectively. All of these combinational algorithms may achieve better speech enhancement performances than any algorithm alone.

3.2 Combined MCRANC with improved spectral subtraction

3.2.1 Description

The combined algorithm of MCRANC with ISS is indicated in figure 3.2.1. In this figure, MCRANC is on the left of the dotted line while ISS is on the right. From now on, VAD in MCRANC will not be depicted for simplicity. As MCRANC has been detailed in chapter 2, only ISS will be briefly described in this section.

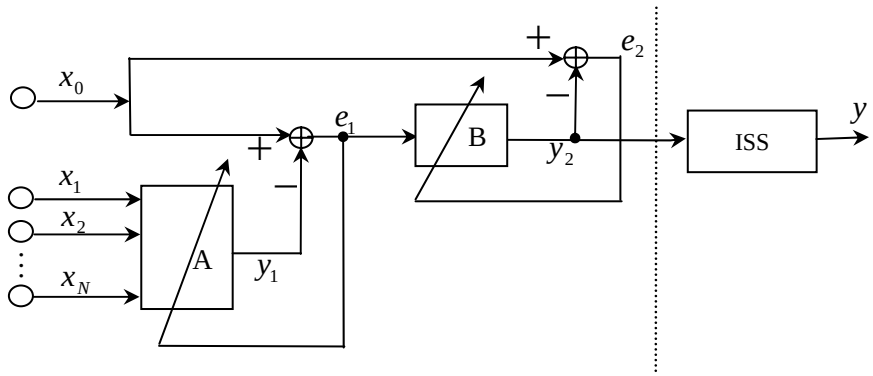

Figure 3.2.1 Structure of combined MCRANC with ISS

ISS is a traditional and widely used algorithm for single-channel speech enhancement [13]. ISS subtracts the noise spectrum from the noisy speech spectrum and then converts the frequency-domain signal to a time-domain signal.

The weakness of ISS is the so-called “music noise” problem. If the SNR of the

noisy speech is low, its music noise level will be high. However, if the SNR of the noisy speech is high, its music noise level will be low and the enhanced speech becomes acceptable. The enhanced speech by MCRANC meets this requirement better than the original noisy speech. From equation (2.4.24), the enhanced speech by MCRANC is

$$y_2(k) = s_0(k) + e(k) \tag{3.2.1}$$

where $e(k) = \tilde{h}^{-1}(k) * e_1^*(k)$ is the residual noise. The enhanced speech $y_2(k)$ usually has higher SNR than original noisy speech $x_0(k)$. Therefore, the ISS algorithm is more suitable for the enhanced speech $y_2(k)$ than the original noisy speech $x_0(k)$.

3.2.2 Improved spectral subtraction

The ISS algorithm applying to the MCRANC enhanced speech $y_2(k)$, as expressed by equation (3.2.1), is introduced as follows.

Divide $y_2(k)$ into overlapped frames. A window is used to smooth each frame and to reduce spectrum leakage. Then apply Discrete Fourier Transform (DFT) to each frame to obtain the power spectrum estimation of $y_2(k)$

$$|Y_2(l)|^2 \approx |S_0(l)|^2 + |E(l)|^2 \tag{3.2.2}$$

where

$$Y_2(l) = \sum_{k=0}^{K-1} y_2(k) e^{-j\frac{2\pi lk}{N}} = |Y_2(l)|\, e^{j\varphi(l)} \tag{3.2.3}$$

where K is the length of the frame, and $\varphi(l)$ is the phase of $Y_2(l)$.

As the power spectrum $|E(l)|^2$ of the residual noise can not be obtained directly, use the weighted average $|\tilde{E}(l)|^2$ of several frames of the residual noise power spectrum during NSP (Non Speech Period) as the estimation of $|E(l)|^2$. Since the residual noise is uncorrelated with speech, the power spectrum of the speech can be estimated as

$$|\tilde{S}_0(l)|^2 = |Y_2(l)|^2 - |\tilde{E}(l)|^2 \tag{3.2.4}$$

where the power spectrum estimation is derived by subtracting the noise spectrum estimation from noisy speech. Because of the difference between the noise power and its estimation, the right side of equation (3.2.4) might be negative. To avoid a negative

power spectrum, we replace it with 0. This process is called half-wave rectification. After this rectification, the speech spectrum estimation is

$$|\hat{S}_0(l)|^2=\begin{cases}|\tilde{S}_0(l)|^2 & if \quad |\tilde{S}_0(l)|^2\geq 0\\ 0 & others\end{cases} \tag{3.2.5}$$

By use of the phase of the noisy speech, the speech signal $\hat{s}_0(k)$ in the time-domain can be estimated by IDFT (Inverse DFT) transform

$$y(k)=\hat{s}_0(k)=\text{IDFT}(|\hat{S}_0(l)|e^{j\varphi(l)}) \tag{3.2.6}$$

where $\varphi(l)$ is the phase of the noisy speech. $\varphi(l)$ is used as the phase of speech signal since the human auditory system is insensitive to the phase of the speech signal.

The main drawback of the above spectral subtraction algorithm is that the algorithm causes the so-called “music noise” problem. To alleviate the music noise, Berouti proposed the ISS algorithm which employs an over-subtraction factor and a spectrum base [13]. In the ISS algorithm, the expression of $|\hat{S}_0(l)|^2$ is

$$|\tilde{S}_0(l)|^2=|Y_0(l)|^2-\alpha|\tilde{E}(l)|^2 \tag{3.2.7}$$

where α is the over-subtraction factor and is expressed by

$$\alpha=\alpha_0-\frac{3}{20}SNR \qquad -5dB\leq SNR\leq 20dB \tag{3.2.8}$$

where α_0 is the value of the over-subtraction factor α when SNR=0 dB. Usually we take $\alpha_0=3$. Other equations of α are

$$\alpha=\begin{cases}4.75 & SNR<-5dB\\ 4-0.15SNR & -5dB\leq SNR\leq 20dB\\ 1 & SNR>20dB\end{cases} \tag{3.2.9}$$

and

$$\alpha=\alpha(k)=1+\frac{sd(|N(k)|)}{E[|N(k)|]} \tag{3.2.10}$$

where $E(|N(k)|)$ and $sd(|N(k)|)$ is the expectation and standard deviation of the noise at frequency k.

The rectification to (3.2.7) is

$$|\hat{S}_0(l)|^2=\begin{cases}|\tilde{S}_0(l)|^2 & if \quad |\tilde{S}_0(l)|^2\geq \beta|\tilde{E}(l)|^2\\ \beta|\tilde{E}(l)|^2 & others\end{cases} \tag{3.2.11}$$

where β is a small positive number called spectrum base. $0<\beta<<1$ and its typical value is 0.1.

The over-subtraction factor is actually a time-variable factor and is used to control the extent of the noise subtraction. Spectrum base β is used to prevent the power spectrum of the enhanced speech being lower than $\beta|\tilde{E}(l)|^2$. Its purpose is to use the wide-band noise to conceal the music noise.

Other efforts have also been made in recent years to alleviate the music noise. But no matter what improvement is made, the music noise can not be completely eliminated. Despite this problem, ISS is widely used on noisy speech since it is simple, effective and easy for implementation.

3.2.3 Experimental results

Several experiments were conducted to benchmark the performance of the proposed algorithm against some commonly used algorithms.

Experiment 1

Our first experiment was carried out in a common research room with dimensions of 8x5x3m. In the experiment, four small microphones $M_0, M_1, \cdots, M_3$ were employed and closely placed on a cylindrical shape structure with 1cm radius as shown in figure 3.2.2. M_0 was placed onto the top surface of the cylinder while the referential microphones were embedded into the side surface. The noise was generated from an improperly tuned (no station) radio located at about 1.5m from the microphone array, as shown in figure 3.2.3. The speech came from a person at 0.5m from the microphones. The sampling rate was 8 kHz.

In the processing of speech enhancement, the NLMS algorithm is employed to adapt the coefficients of filters A and B in MCRANC. For filter A, the tapped delay line per channel is L=32 and hence filter A has 99 coefficients. The number of coefficients of filter B is selected to be 48. VAD is energy and zero-crossing rate based one. If the

VAD fails, we use only the beginning 0.3s signal as the pure noise.

In ISS, the window frame length is K=256 and the windows is 50% overlapped. A Hamming window is employed for smoothing. We average the power spectra over 3 frames of pure noise during the NSP to estimate the residual noise power spectrum $|E(l)|^2$. The over-subtraction factor shown in equation (3.2.8) is selected as $\alpha_0 = 3$ and the spectrum base factor of equation (3.2.11) $\beta = 0.1$.

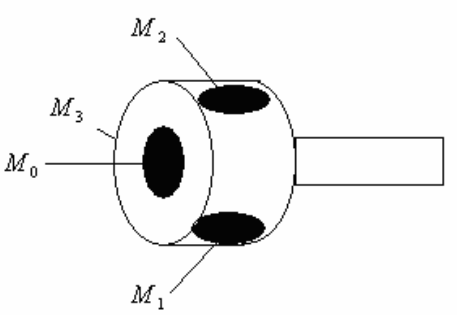

Figure 3.2.2 A solid microphone array

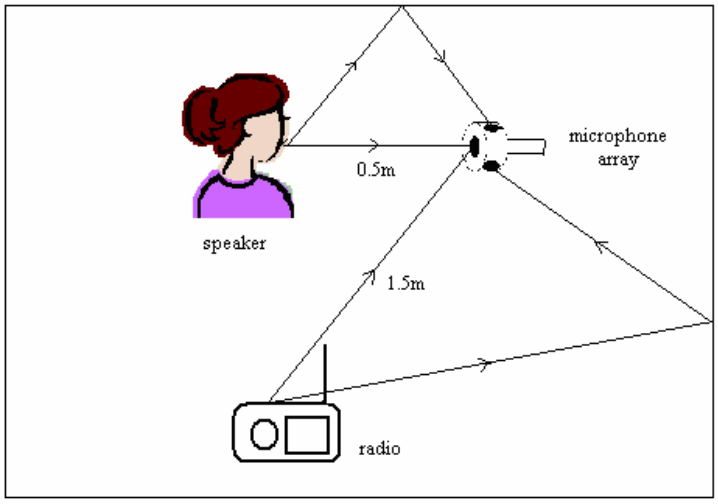

Figure 3.2.3 A scenario of a noisy speech environment

Figure 3.2.4 shows visually the performance of the proposed speech enhancement system. Figure 3.2.4 (a) shows the noisy speech signal $x_0(k)$ acquired by the main microphone with a SNR of 2.8 dB. Signals acquired by the referential microphones are visually similar to $x_0(k)$ and they do not need to be replicated. Figure 3.2.4 (b) is the

enhanced speech using the two-channel CRANC algorithm, with a SNR improvement of 9.2 dB. Figure 3.2.4 (c) is the enhanced speech using MCRANC algorithm with a SNR improvement of 18.0 dB. Figure 3.2.4 (d) is the enhanced speech using the

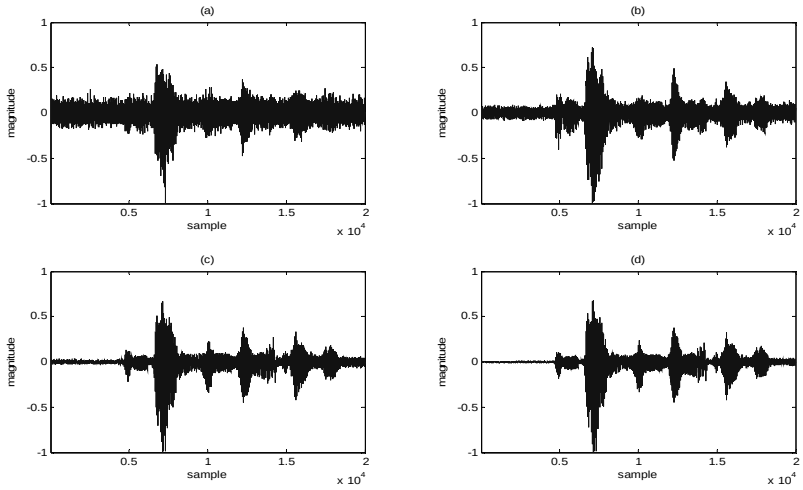

Figure 3.2.4 Results of experiment 1

(a) Noisy speech signal
(b) Enhanced speech by two-channel CRANC
(c) Enhanced speech by MCRANC
(d) Enhanced speech by MCRANC and ISS

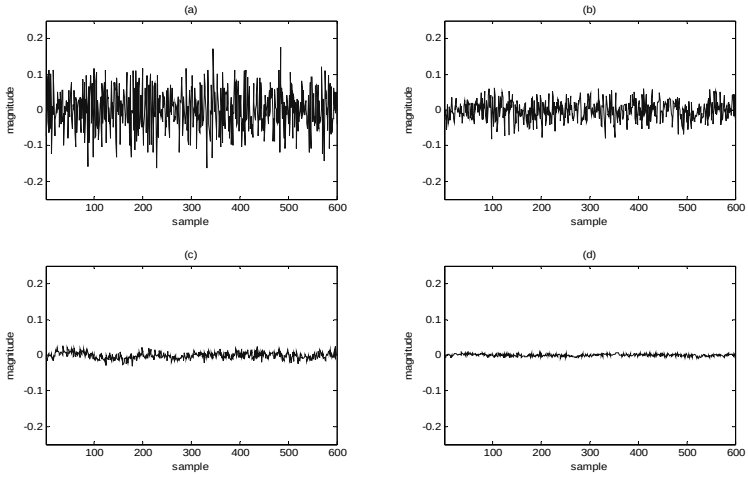

Figure 3.2.5 Zoomed view of a short noise segment from figure 3.2.4 (pure noise)

(a) Pure noise segment
(b) Output noise by two-channel CRANC
(c) Output noise by MCRANC
(d) Output noise by MCRANC and ISS

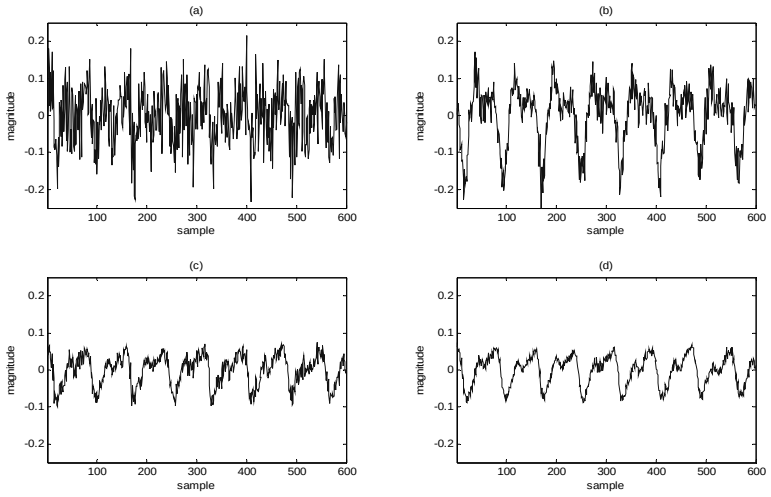

Figure 3.2.6 Zoomed view of a short speech segment from figure 3.2.4 (noisy speech)

(a) Noisy speech segment
(b) Enhanced speech by two-channel CRANC
(c) Enhanced speech by MCRANC
(d) Enhanced speech by MCRANC and ISS

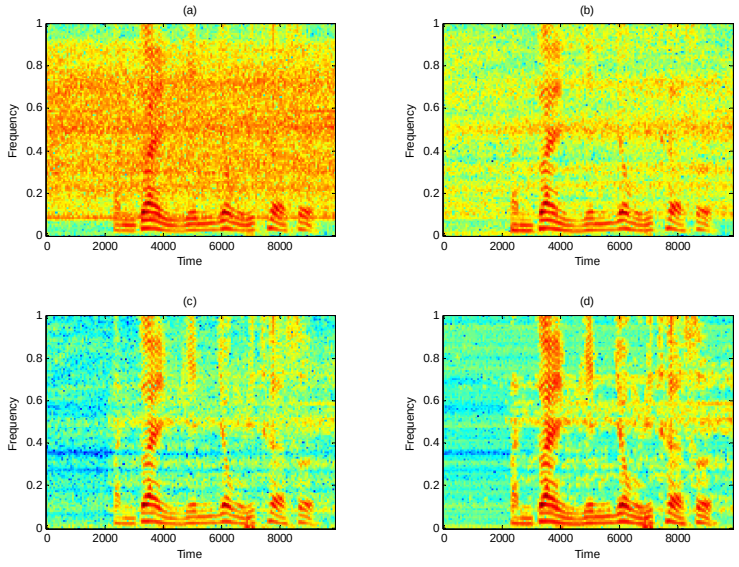

Figure 3.2.7 Spectrograms for the signals in figure 3.2.4

(a) Spectrogram of noisy speech signal
(b) Spectrogram of enhanced speech by two-channel CRANC
(c) Spectrogram of enhanced speech by MCRANC
(d) Spectrogram of enhanced speech by MCRANC and ISS

algorithm of MCRANC augmented with ISS, which achieves a SNR improvement of 27.0 dB. Since it is impossible to get a clean speech signal in this experiment, the SNR here is computed by using formula (1.4.3).

Figure 3.2.5 shows a zoomed view of a short noise segment from figure 3.2.4. Figure 3.2.6 shows also a zoomed view of a short speech segment from figure 3.2.4.

Figure 3.2.7 (a), (b), (c) and (d) show the spectrograms of the corresponding signals depicted in figure 3.2.4.

From the experimental results, the proposed combined algorithm of MCRANC and ISS outperforms CRANC or MCRANC algorithm.

Experiment 2

The second experiment was carried out in a Mitsubishi ETERNA car. A uniform linear array with four microphones was placed in front of the driver. Small microphones were collinearly placed with each neighboring microphone separated by 3cm. The aperture of the array was about 13cm. One of the two microphones near the center of the array was used as the main microphone while the rest were considered as referential microphones. The car engine, air conditioning, and the car radio generated the coexisting noises. The noise from the radio was a piece of a musical song. The speech was from the driver about 60cm away from the microphone array. The sampling rate was also 8 kHz.

For MCRANC and ISS used in the enhancement process, all parameters are set as in experiment 1. The NSP was detected with the samples $[1,2,\cdots,10500)$ and $[27001,27002,\cdots,30000)$.

Figure 3.2.8 shows the results of enhancements obtained from this experiment. Figure 3.2.8 (a) is the noisy speech signal $x_0(k)$ acquired by the main microphone, with a SNR = -8.4 dB. Figure 3.2.8 (b) is the enhanced speech using the ISS algorithm only and giving a SNR improvement of 14.5 dB. Figure 3.2.8 (c) is the enhanced speech obtained by using the MCRANC algorithm, with a SNR improvement of 15.1 dB. Figure 3.2.8 (d) is the enhanced speech obtained by combining the MCRANC and ISS algorithms, which offers a SNR improvement of 25.4 dB. The SNR is also estimated by applying formula (1.4.3).

From this experiment, we may also find the combination of MCRANC and ISS performs better than MCRANC or ISS algorithm alone.

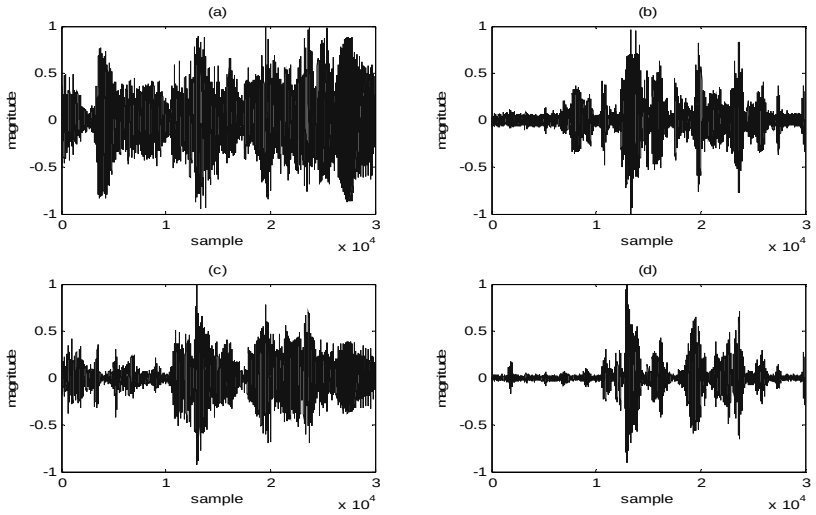

Figure 3.2.8 Results of experiment 2

(a) Noisy speech
(b) Enhanced speech by ISS
(c) Enhanced speech by MCRANC
(d) Enhanced speech by MCRANC and ISS

3.2.4 Conclusions

In this section a combined algorithm is presented for speech enhancement, in which the MCRANC algorithm is used to obtain a primary enhancement of a noisy speech signal, and then followed by the ISS to further improve the enhancement performance.

The MCRANC partially cancels out the introduced noise in the acquired speech signal to get SNR improvement for the noisy speech signal. Thus, it provides a more appropriate signal for the ISS algorithm and leads the introduced spectral subtraction byproduct (music noise) to a lower level.

The speech enhancement based on the proposed combinational algorithm uses a small-size microphone array, and achieves better speech enhancement performance than the ISS, CRANC or MCRANC algorithms alone.

3.3 Combined MCRANC with delay and sum beamforming

In this section, a combined algorithm of MCRANC and Delay And Sum (DAS) beamforming is presented for speech enhancement. It first employs MCRANC for every channel of the noisy speech to get enhanced speeches with higher SNR and less correlated residual noises. Then the enhanced speeches are input to a DAS beamformer to get further enhancement.

3.3.1 Delay and sum beamforming

The DAS algorithm is the most basic algorithm for beamforming. It aligns the array signals and then sums the aligned signals to get the output. Its structure is indicated in figure 3.3.1.

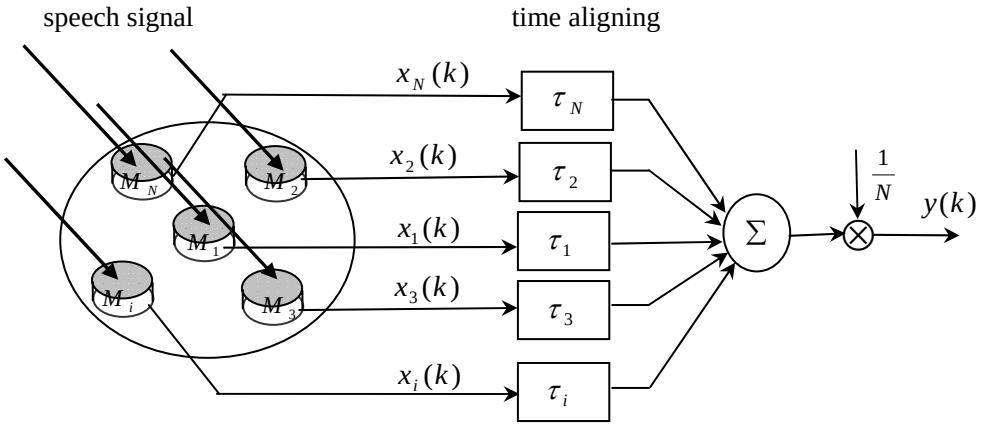

Figure 3.3.1 Delay And Sum beamforming

Suppose the audio signal arrives at all N microphones $M_1, M_2, \cdots, M_N$ in an array without difference in amplitude, and microphone M_1 is used as the referential microphone for time alignment. If the speech signal acquired by microphone M_1 is $s_1(k)$ and the time delay for the speech signal in microphone M_i is τ_i with reference to microphone M_1, the speech signal acquired by microphone M_i will be $s_i(k) =$

$s_1(k-\tau_i)$. So the time compensation should be τ_i for the speech signal from microphone M_i. The output of the DAS beamforming is

$$\begin{aligned} y(k) &= \frac{1}{N}\sum_{i=1}^{N} x_i(k+\tau_i) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left[s_i(k+\tau_i)+n_i(k+\tau_i)\right] \\ &= s_1(k)+\frac{1}{N}\sum_{i=1}^{N} n_i(k+\tau_i) \end{aligned} \tag{3.3.1}$$

where $n_i(k)$ is the noise signal acquired by microphone M_i. Since the noise signal is random, the power of $e(k)=\frac{1}{N}\sum_{i=1}^{N} n_i(k+\tau_i)$ is usually less then the power of $n_1(k)$, especially when $n_i(k)$ $(i=1,2,\cdots,N)$ are Gaussian white noises. Therefore, DAS beamforming enhances the speech signal.

Estimation for time delay

The estimation of the time delay is the key factor for DAS beamforming. A commonly used algorithm for time delay estimation is the Generalized Cross-Correlation (GCC) algorithm since it is simple and it can be used to deal different kinds of noises [66]. We introduce it as follows.

Assume the noisy speech acquired by microphone M_i and microphone M_j are

$$\begin{aligned} x_i(k) &= s(k)+n_i(k) \\ x_j(k) &= s(k-\tau_{ij})+n_j(k) \end{aligned} \tag{3.3.2}$$

where $s(k)$ is the speech signal, and $n_i(k)$ and $n_j(k)$ are the noises acquired by the microphone M_i and microphone M_j respectively. It is also assumed that $s(k)$, $n_i(k)$ and $n_j(k)$ are mutually uncorrelated, i.e.

$$R_{sn_i}(\tau-\tau_{ij}) = R_{sn_i}(\tau) = R_{n_in_j}(\tau) = 0 \tag{3.3.3}$$

Where τ_{ij} is the time delay of the speech signal between microphone M_i and microphone M_j. From equations (3.3.2) and (3.3.3), the cross-correlation

$$R_{ij}(\tau) = E[x_i(k)x_j(k-\tau)] = R_{ss}(\tau-\tau_{ij}) \tag{3.3.4}$$

$R_{ss}(\cdot)$ will be maximum if $\tau - \tau_{ij}$=0. Therefore, we may search for $\hat{\tau}_{ij}$ to make $R_{ss}(\tau - \tau_{ij})$ maximal and use it as the estimation of time delay τ_{ij}.

Since the noises in different channels might not be completely uncorrelated and the statistic average can only be estimated by a time-limited average in practice, improved algorithms are introduced. Consider

$$R_{ij}(\tau) = \int_{-\infty}^{+\infty} \phi_{ij}(f) e^{j2\pi f\tau} df \tag{3.3.5}$$

where $\phi_{ij}(f)$ is the cross-power spectrum between microphone signals $x_i(k)$ and $x_j(k)$. To sharpen the peak value of $R_{ij}(\tau)$, a weight function $\psi_{ij}(f)$ can be used to suppress the effect of the noise and reverberation, i.e.

$$R_{ij}(\tau) = \int_{-\infty}^{+\infty} \psi_{ij}(f)\phi_{ij}(f) e^{j2\pi f\tau} df \tag{3.3.6}$$

This cross-correlation function is called the GCC function. $\psi_{ij}(f)$ should be selected according to the noise type.

If we take the weight $\Psi_{ij}(f) = \left|\phi_{ij}(f)\right|^{-1}$, we get the Cross-Power Spectrum Phase (CPSP) algorithm. $\phi_{ij}(f)$ is the cross-power spectrum of signals $x_i(k)$ and $x_j(k)$. From equation (3.3.2), we have

$$\phi_{ij}(f) = \phi_{ss}(f) e^{-j2\pi f\tau_{ij}} \tag{3.3.7}$$

Thus, equation (3.3.6) becomes

$$R_{ij}(\tau) = \delta(\tau - \tau_{ij}) \tag{3.3.8}$$

where δ is the delta function.

Further improvement is made by taking

$$\Psi_{ij}(f) = \left|\phi_{ij}(f)\right|^{-\lambda} \tag{3.3.9}$$

The corresponding algorithm is called the Modified Cross-Power Spectrum Phase (M-CPSP) algorithm.

If λ =0 in equation (3.3.9), the corresponding algorithm is the common cross-correlation algorithm. If λ =1, the corresponding algorithm is the CPSP algorithm. After experiments in different environments, it was suggested to take λ =0.75.

Capability of DAS beamforming

In the ideal condition that the noises are completely uncorrelated and the time alignments are precise, it can be proved that the SNR improvement provided by DAS beamforming is

$$\mathrm{SNR}_{\mathrm{improved}}=10\log_{10}(N) \tag{3.3.10}$$

where N is the number of the microphones in the array. Since

$$10\log_{10}(2N)\approx 3+10\log_{10}(N)$$

the SNR will increase about 3 dB as the number of the microphones doubles. However, the SNR improvement will greatly decrease as the noise correlation increases. In fact, DAS will not provide any SNR improvement if the noise correlation reaches its maximum value 1.

For a small microphone array, the noises in the microphone signals are more highly correlated and fewer microphones can be employed. As a result, DAS can provide only very limited SNR improvement to small microphone array. To get better SNR improvement, it should be used with other algorithms.

3.3.2 Combined MCRANC with DAS beamforming

As shown in figure 3.3.2, the combination scheme of MCRANC with DAS consists of N subsystems of MCRANC and a DAS beamformer, where N is the number of microphones employed in the array. Every dot-lined frame in figure 3.3.2 contains an N-input and one-output MCRANC subsystem. The output of any MCRANC subsystem is actually a primarily enhanced speech signal. These enhanced speech signals are input to DAS beamformer to get further enhancement.

Figure 3.3.3 indicates the i-th MCRANC subsystem, in which the i-th channel of signal is used as the main signal and the other N-1 channels of signals are used as referential signals in a MCRANC.

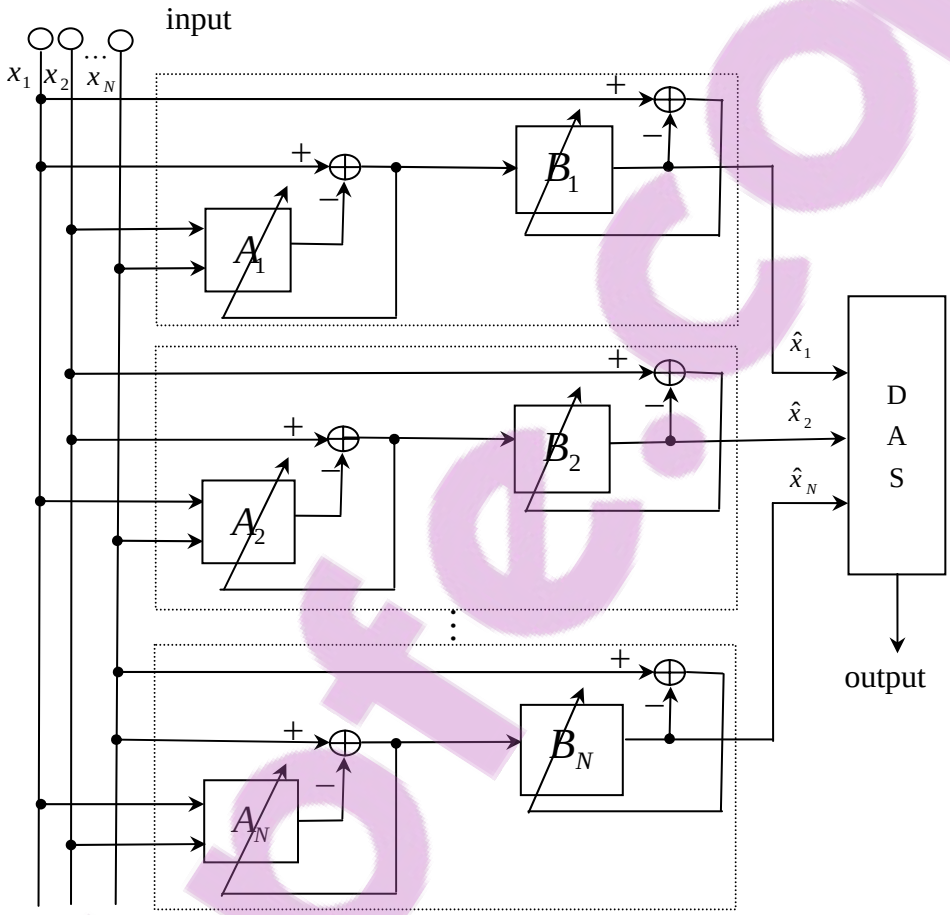

Figure 3.3.2 Combined structure of MCRANC with DAS

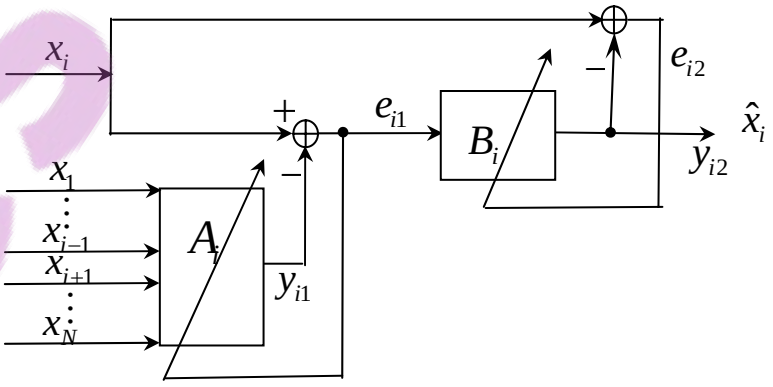

Figure 3.3.3 The i-th MCRANC subsystem

In MCRANC subsystems, the VAD is omitted for simplicity. A VAD based Adaptation Mode Controller (AMC) is used to control the filters in these subsystems when to adapt their coefficients and when to freeze the coefficients.

3.3.2.1 MCRANC subsystem

The subsystem as shown in figure 3.3.3 is a MCRANC as introduced in chapter 2. However, the index notations for the subsystem are somewhat different from those in chapter 2. Unlike chapter 2, there are only N channel noisy signals in this section and every channel of the signal is used in turn as a main channel signal while the other N-1 channels are treated as referential signals. Because these MCRANC subsystems will also be referred to in the next sections and the following chapters, they are described again as follows.

Suppose speech $s(k)$ and noise $n(k)$ are generated by independent sources. They arrive at microphone M_i through multi-paths and are acquired by M_i as $s_i(k)$ and $n_i(k)$ respectively. The impulse responses of the intermediate media between the speech and noise sources and the microphone M_i are $h_{si}(k)$ and $h_{ni}(k)$ respectively. As indicated in figure 3.3.4, the actual signal acquired by microphone M_i can be represented by $x_i(k) = s_i(k) + n_i(k)$, where $i = 1,2,\cdots,N$ and $k = 0,1,2,\cdots,$ N is the number of microphones employed in the array, and k is the discrete time index.

We have

$$x_i(k) = s_i(k) + n_i(k) \tag{3.3.11}$$

$$s_i(k) = h_{si}(k) * s(k) \tag{3.3.12}$$

$$n_i(k) = h_{ni}(k) * n(k) \qquad i = 1,2,\cdots,N \tag{3.3.13}$$

where $*$ is the convolution sign.

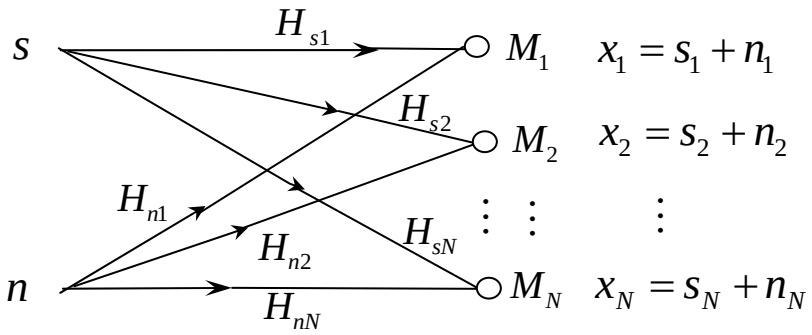

Figure 3.3.4 Speech and noise propagation between the emitting sources and the acquiring microphones

Note the impulse response of a system with input s_i and output s_j as $h_{s_i s_j}(k)$, and the impulse response of a system with input n_i to n_j as $h_{n_i n_j}(k)$, i.e.

$$s_j(k) = h_{s_i s_j}(k) * s_i(k) \tag{3.3.14}$$

$$n_j(k) = h_{n_i n_j}(k) * n_i(k) \qquad i, j = 1,2,\cdots,N \tag{3.3.15}$$

From equations (3.3.12)-(3.3.15) we have

$$H_{s_i s_j}(z) = \frac{H_{sj}(z)}{H_{si}(z)} \tag{3.3.16}$$

$$H_{n_i n_j}(z) = \frac{H_{nj}(z)}{H_{ni}(z)} \qquad i, j = 1,2,\cdots,N \tag{3.3.17}$$

where $H_{si}(z)$ is the z-transform of $h_{si}(k)$ and so forth for other notations.

In the i-th MCRANC subsystem indicated in figure 3.3.3, the signal $x_i(k)$ from microphone M_i is regarded as the main signal while the other N-1 signals $x_j(k)$ ($j = 1,\cdots,i-1,i+1,\cdots,N$) are used as referential signals. In this subsystem, two adaptive filters A_i and B_i are employed.

A VAD based AMC is used to detect the special Overall Non Speech Periods (ONSP), which are time segments containing only pure noises for all channels of the signals. The AMC and the special ONSP will be presented in next subsection 3.3.2.2.

In the special ONSP, we have $s_i(k)=0$ ($i=1,2,\cdots,N$). Thus, from expression

$$x_i(k)=y_{i1}(k)+e_{i1}(k) \tag{3.3.18}$$

we have

$$n_i(k)=\mathbf{w}_i\mathbf{n}_i(k)+e_{i1}(k) \tag{3.3.19}$$

where $x_i(k)=n_i(k)$, $y_{i1}(k)=\mathbf{w}_i\mathbf{n}_i(k)$ is the output of filter A_i and $\mathbf{w}_i$ is a $1\times(N-1)(L+1)$-dimension coefficient vector of filter A_i, i.e.

$$\mathbf{w}_i=(\mathbf{w}_{i1},\cdots,\mathbf{w}_{i(i-1)},\mathbf{w}_{i(i+1)}\cdots,\mathbf{w}_{iN}) \tag{3.3.20}$$

where $\mathbf{w}_{ij}=(w_{ij0},w_{ij1},\cdots,w_{ijL})$, $\mathbf{n}_i(k)$ is a $(N-1)(L+1)\times 1$-dimension noise vector

$$\mathbf{n}_i(k)=\left[\mathbf{n}_{i1}(k),\cdots,\mathbf{n}_{i(i-1)}(k),\mathbf{n}_{i(i+1)}(k),\cdots,\mathbf{n}_{iN}(k)\right]^T \tag{3.3.21}$$

where $\mathbf{n}_{ij}(k)=[n_j(k),n_j(k-1),\cdots,n_j(k-L)]$, L is the number of the sample delay for every referential signal. In equation (3.3.19) $e_{i1}(k)$ is the prediction error. Let the minimal error power be denoted by $P[e_{i1}^*(k)]$ and the corresponding optimal coefficient vector by

$$\begin{aligned}\mathbf{w}_i^* &=(\mathbf{w}_{i1}^*,\cdots,\mathbf{w}_{i(i-1)}^*,\mathbf{w}_{i(i+1)}^*,\cdots,\mathbf{w}_{iN}^*)\\ &=(w_{i10}^*,w_{i11}^*,\cdots,w_{i1L}^*,\cdots\cdots,w_{i(i-1)0}^*,w_{i(i-1)1}^*,\cdots,w_{i(i-1)L}^*,\cdots\cdots,\\ &\quad w_{i(i+1)0}^*,w_{i(i+1)1}^*,\cdots,w_{i(i+1)L}^*,\cdots\cdots,w_{iN0}^*,w_{iN1}^*,\cdots,w_{iNL}^*)\end{aligned} \tag{3.3.22}$$

We only need to adjust the coefficients of filter A_i to minimize the power of $e_{i1}(k)$ in figure 3.3.3 to obtain $\mathbf{w}_i^*$.

Then, during the time period that follows the special ONSP time section, we may assume the environment remains almost unchanged and accordingly we may keep the optimal weights of filter A_i unchanged, thus

$$\begin{aligned}y_{i1}(k)&=\mathbf{w}_i^*\mathbf{x}_i(k)\\ &=\mathbf{w}_i^*\mathbf{s}_i(k)+\mathbf{w}_i^*\mathbf{n}_i(k)\end{aligned}$$

$$= \mathbf{w}_i^* \mathbf{s}_i(k) + [n_i(k) - e_{i1}^*(k)] \tag{3.3.23}$$

where $\mathbf{x}_i(k)$ and $\mathbf{s}_i(k)$ represent the vectors of noisy speech and clean speech respectively, and may be expressed in a similar way to $\mathbf{n}_i(k)$ in equation (3.3.21). Then from equations (3.3.18) and (3.3.23), we have

$$\begin{aligned} e_{i1}(k) &= x_i(k) - y_{i1}(k) \\ &= [s_i(k) + n_i(k)] - [\mathbf{w}_i^* \mathbf{s}_i(k) + n_i(k) - e_{i1}^*(k)] \\ &= s_i(k) - \mathbf{w}_i^* \mathbf{s}_i(k) + e_{i1}^*(k) \\ &= p_i(k) + e_{i1}^*(k) \end{aligned} \tag{3.3.24}$$

where

$$p_i(k) = s_i(k) - \mathbf{w}_i^* \mathbf{s}_i(k) \tag{3.3.25}$$

Take the z-transform of (3.3.24) and (3.3.25) to get

$$E_{i1}(z) = P_i(z) + E_{i1}^*(z) \tag{3.3.26}$$

$$\begin{aligned} P_i(z) &= S_i(z) - Z[\sum_{m=1,m\neq i}^{N} \sum_{j=0}^{L} w_{imj}^* s_m(k-j)] \\ &= S_i(z) - Z[\sum_{m=1,m\neq i}^{N} \sum_{j=0}^{L} w_{imj}^* h_{s_i s_m}(k-j) * s_i(k-j)] \\ &= S_i(z) - \sum_{m=1,m\neq i}^{N} \sum_{j=0}^{L} w_{imj}^* Z[h_{s_i s_m}(k-j)] Z[s_i(k-j)] \\ &= S_i(z) - \sum_{m=1,m\neq i}^{N} \sum_{j=0}^{L} w_{imj}^* [H_{s_i s_m}(z) z^{-j}][S_i(z) z^{-j}] \\ &= [1 - \sum_{m=1,m\neq i}^{N} \sum_{j=0}^{L} w_{imj}^* z^{-2j} H_{s_i s_m}(z)] S_i(z) \\ &= [1 - \sum_{m=1,m\neq i}^{N} \sum_{j=0}^{L} w_{imj}^0 z^{-2j} \frac{H_{sm}(z)}{H_{si}(z)}] S_i(z) \\ &= \tilde{H}_i(z) S_i(z) \end{aligned} \tag{3.3.27}$$

where

$$\tilde{H}_i(z) = 1 - \sum_{m=1, m\neq i}^{N} \sum_{j=0}^{L} w_{imj}^{*} z^{-2j} \frac{H_{sm}(z)}{H_{si}(z)} \tag{3.3.28}$$

If the system function of filter B_i is $\tilde{H}_i^{-1}(z) = [\tilde{H}_i(z)]^{-1}$, then by using (3.3.27) and (3.3.28) we get

$$\begin{aligned} Y_{i2}(z) &= \tilde{H}_i^{-1}(z) E_{i1}(z) \\ &= \tilde{H}_i^{-1}(z)[\tilde{H}_i(z) S_i(z) + E_{i1}^{*}(z)] \\ &= S_i(z) + \tilde{H}_i^{-1}(z) E_{i1}^{*}(z) \end{aligned} \tag{3.3.29}$$

Thus

$$y_{i2}(k) = s_i(k) + \tilde{h}_i^{-1}(k) * e_{i1}^{*}(k) \tag{3.3.30}$$

where $\tilde{h}_i^{-1}(k)$ is the inverse z-transform of $\tilde{H}_i^{-1}(z)$ and $*$ is the convolution operator. As commonly assumed in ANC, the noise $n_i(k)$ is supposed to be uncorrelated with the signal $s_i(k)$. In order that the system transfer function of filter B_i approximates $\tilde{H}_i^{-1}(z)$, we need only adjust the weights of filter B_i to minimize the square sum of e_{i2}. This is because

$$\begin{aligned} \|e_{i2}(k)\|^2 &= \|x_i(k) - y_{i2}(k)\|^2 \\ &= \|s_i(k) + n_i(k) - y_{i2}(k)\|^2 \\ &= \|n_i(k)\|^2 + \|s_i(k) - y_{i2}(k)\|^2 + 2 n_i(k)[s_i(k) - y_{i2}(k)] \end{aligned} \tag{3.3.31}$$

and

$$E[e_{i2}^2(k)] = E[n_i^2(k)] + E\{[s_i(k) - y_{i2}(k)]^2\} \tag{3.3.32}$$

From (3.3.32), we may conclude that for minimizing $E[e_{i2}^2(k)]$ we also need minimize $E\{[s_i(k) - y_{i2}(k)]^2\}$, which implies minimizing the difference between $y_{i2}(k)$ and $s_i(k)$. For simplicity, filter B_i may also take FIR type.

From the above discussion, we know that the output $\hat{x}_i(k)$ of the i-th subsystem

would be the approach of the signal $s_i(k)$ （$i = 1,2,\cdots,N$）. We may further input these approaches with a DAS beamformer to get a better speech enhancement performance.

3.3.2.2 Adaptive module controller

From equations (3.3.18) to (3.3.23) we know that the impulse response of filter A_i is $\mathbf{w}_i^*$. With filter A_i we can cancel the noise $n_i(k)$ carried with the signal $x_i(k) = s_i(k) + n_i(k)$. If the transfer function or impulse response of the noise is unchanged, which implies the whole environment remains unchanged including the positions of the noise sources, the space and even the air temperature and pressure, the optimal weight $\mathbf{w}_i^*$ would remain unchanged. But unfortunately the transfer function of the noise keeps changing from time to time with the changes of environment such as the opening of a door or the closing of a window. To adapt the system to the dynamical changes of the environment, the weights of filter A_i must be adapted from time to time during NSP time sections to compensate for any change in the noise environment. However, since there are different time delays in the speech signal arriving at the different microphones, The NSP section of one microphone signal will be a little different from the NSP section of another microphone signal. A special Overall NSP (ONSP) should be defined and all MCRANC subsystems can be adapted during this ONSP section

Let $NSP(i)$ denote the NSP of the i-th channel signal $x_i(k)$ acquired by microphone M_i. Obviously $NSP(i)$ consists of a series of discrete time intervals, i.e.

$$NSP(i) = \bigcup_j [k_{ij}^{'}, k_{ij}^{''}] \tag{3.3.33}$$

where discrete time interval

$$[k_{ij}^{'}, k_{ij}^{''}] = \{k_{ij}^{'}, k_{ij}^{'} + 1, \cdots, k_{ij}^{''}\}$$

is the j-th NSP of signal $x_i(k)$. Since the arrival times of the speech signal to different microphones may be different, $NSP(i_1)$ may be different from $NSP(i_2)$, $i_1 \neq i_2$, $i_1, i_2 \in \{1,2,\cdots,N\}$. But $NSP(i_1)$ is only a shift of $NSP(i_2)$ in the time axis.

We define ONSP as

$$\text{ONSP} = \bigcap_{i=1}^{N} NSP(i) \tag{3.3.34}$$

Then we may easily prove that

$$\text{ONSP} = \bigcup_{j} [k_j^{'}, k_j^{''}] \tag{3.3.35}$$

where

$$k_j^{'} = \max_{1 \le i \le N} \{k_{ij}^{'}\}, \qquad k_j^{''} = \min_{1 \le i \le N} \{k_{ij}^{''}\}$$

If $k_j^{''} < k_j^{'}$ we define $[k_j^{'}, k_j^{''}] = \phi$ in equation (3.3.35).

We should not update the weights of filter A_i during periods in which any channel signal input to A_i carries a speech segment. Otherwise, the speech signal will be viewed as a noise signal and thus be cancelled. Therefore, we update the coefficients of filter A_i only during the L-ONSP defined by

$$\text{L-ONSP} = \bigcup_{j} [k_j^{'} + L, k_j^{''}] \tag{3.3.36}$$

where L is the time delay for the referential signals input to filter A_i we used in equation (3.3.20), and

$$[k_j^{'} + L, k_j^{''}] = \{k_j^{'} + L, k_j^{'} + L + 1, \cdots, k_j^{''}\} \tag{3.3.37}$$

If $k_j^{''} < k_j^{'} + L$ we also define $[k_j^{'} + L, k_j^{''}] = \phi$ in (3.3.36).

In L-ONSP, all signals and their delays we used will belong to NSP and carry no speech signals. We update the coefficients of filter A_i only during this L-ONSP. The special ONSP mentioned above and in the second paragraph of section 3.3.2.2 means the segment of L-ONSP as shown in (3.3.36).

There have been many VAD algorithms developed to detect the NSP and HSP. Yet

it is still a difficult task to exactly determine the NSP segments or the HSP segments. Fortunately, in our noise cancellation algorithm we do not need the exact NSP or HSP but only the segments of pure noise periods to train or update the optimal coefficients of filter A_i to compensate the changes of the noise environment. In fact, a simple VAD is enough for deciding when to update the coefficients of filter A_i. Furthermore, for simplicity, we need only detect NSP for only one channel of the signal $x_{i_0}(k)$ instead of all N channels of the signals. We may update the coefficients of all filters A_i only during the $(\Delta, \Delta') - ONSP$ defined as

$$(\Delta, \Delta')\text{-ONSP} = \bigcup_j [k'_{i_0 j} + \Delta, k''_{i_0 j} - \Delta'] \tag{3.3.38}$$

where $[k'_{i_0 j}, k''_{i_0 j}]$ is the discrete time intervals of $NSP(i_0)$ detected by a VAD. Δ' is a positive integer selected to ensure the employed intervals are the pure noise periods, Δ is also a positive integer but

$$\Delta \geq L + \delta + \Delta' \tag{3.3.39}$$

where δ is the maximum time delay for the audio signal arriving to any other microphone with respect to microphone M_{i_0}, i.e.

$$\delta = \frac{f}{c} \max_{1 \leq i \leq N} \{d_i\} \tag{3.3.40}$$

where d_i is the distance between microphone M_{i_0} and microphone M_i, f is the sampling rate for the array signals and c is the propagation speed of the audio signal.

Outside the (Δ, Δ')-ONSP, for every subsystem we keep the coefficients of filters A_i unchanged and do the filtering works with its existing coefficients.

Similarly, in order to adapt to the changes of the transfer functions of the speech signal occurring due to the movements of the speaker and the changes of the environment, we should update from time to time the coefficients of all filters B_i during the HSP sections. However, for the stability of the residual noise in the enhanced speech, in any subsystem we continue the adaptive filtering of B_i all the time. Its

coefficients would always be updated. Since its system function approaches $\tilde{H}_i^{-1}(z)$ as shown by equation (3.3.28), filter B_i may follow the changes of the propagation of the speech signal.

To sum up, we update the coefficients of filter A_i in any subsystem only during (Δ, Δ')-ONSP and update the coefficients of filter B_i all the time. (Δ, Δ')-ONSP may be easily determined by any channel acquired signal with a VAD.

3.3.2.3 Computational complexity

For the general speech enhancement scheme as shown in figure 3.3.2, the main analysis to the computational complexity lies in the MCRANC subsystems, since we may use a simple VAD and an existing DAS. For a MCRANC subsystem, its computational complexity has been presented in chapter 2. It actually depends on what adaptive algorithm is employed for filters A_i and B_i. If we use the LMS adaptive algorithm, the computational complexity of all N MCRANC subsystems would be

$$(2_A + 3_M)[(L+1)(N-1) + (L_B+1)]f \tag{3.3.41}$$

where 2_A means 2 addition operations, 3_M means 3 multiplication operations, L is the delay time used for every referential channel signal as shown in (3.3.21), N is the number of microphones employed in the system, L_B is the order of filter B_i and f is the sampling rate for the array signals.

For example, if we select $L = 24$, $L_B = 48$, $f = 8000$ and employs $N = 5$ microphones, the computational complexity will be less than 5,960,000 FLOPS.

3.3.3 Experimental results

The experiment was carried out in a real environment, a common research room with parameters of around $8 \times 5 \times 2.8$ m and with several desks, computers and air

conditioners in it. Five microphones $M_1, M_2, \cdots, M_5$ were placed closely together as in figure 3.3.5. The distance between the center microphone M_1 and any other microphone M_i ($i = 2,3,4,5$) was only 2cm. The noise was generated from an improperly tuned radio situated at 200cm from the microphone array. The speech was from a male at 50cm from the array. The sampling frequency was 8 kHz.

We selected $L = 24$ and $L_B = 48$. So the orders of filters A_i and B_i were $(24+1)\times 4-1=99$ and 48 respectively. The LMS adaptive algorithm was employed to adjust the coefficients of FIR filters A_i and B_i in every subsystem. In the LMS algorithm the learning factor $\mu = 0.001$. We selected $x_1(k)$ acquired by the microphone M_1 to be the detected signal by VAD. To get $(\Delta, \Delta') - ONVP$ we selected $\Delta' = 100$. VAD is energy and zero-crossing rate based one. The DAS beamforming was simplified as a sum of all five subsystems' outputs since the delay times all are smaller than one sampling interval.

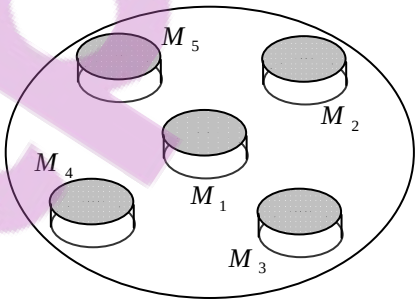

Figure 3.3.5 Small planar array

In figure 3.3.6, (a) shows the noisy speech signal $x_1(k)$ acquired by the microphone M_1. (b) indicates the output $\hat{x}_1(k)$ of the first subsystem, which is one of the primary enhanced speech signals with 18.4 dB SNR improvement. The SNR improvements for other four subsystems are 15.2 dB, 14.9 dB, 13.2 dB and 13.1 dB respectively. (c) depicts the final enhanced speech signal $y(k)$ by the scheme proposed

in this section. The SNR improvement is 22.3 dB with respect to noisy speech $x_1(k)$. Here the SNR is computed by formula (1.4.3).

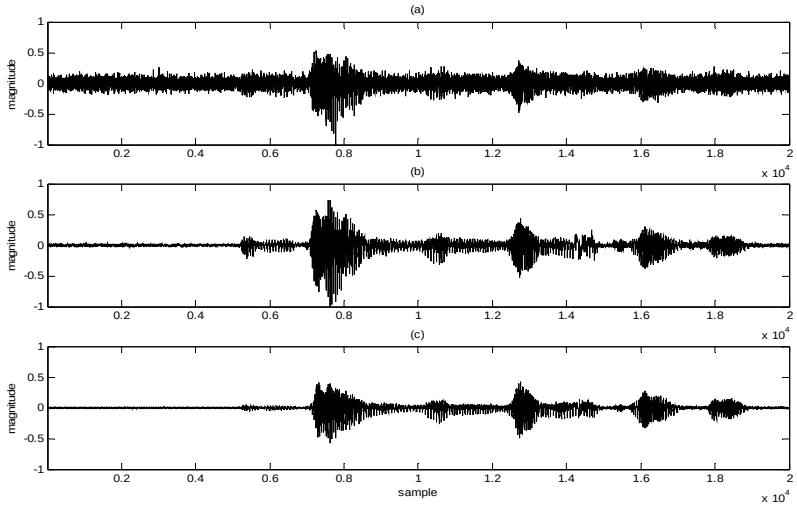

Figure 3.3.6 Experimental results

(a) Noisy speech $x_1(k)$
(b) Primarily enhanced speech $\hat{x}_1(k)$
(c) Final enhanced speech $y(k)$

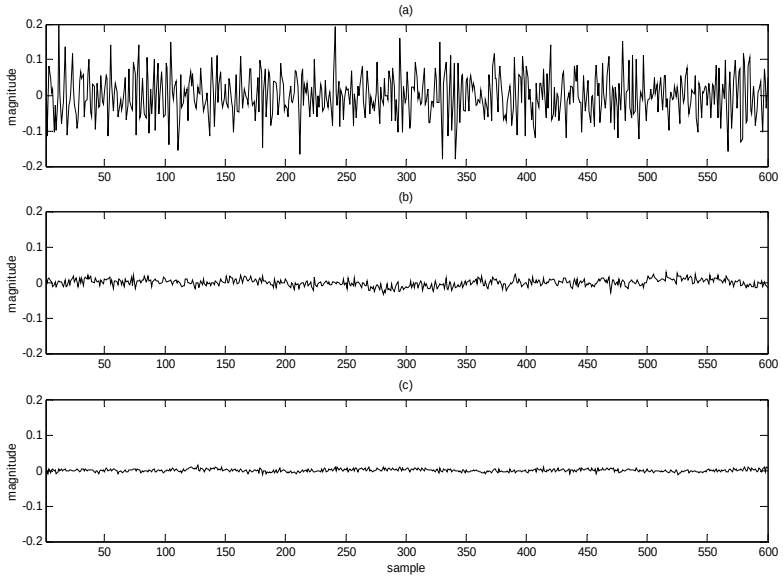

Figure 3.3.7 A zoomed section of NSP from figure 3.3.6

(a) Pure noise section in $x_1(k)$
(b) Residual noise section in $\hat{x}_1(k)$
(c) Residual noise section in $y(k)$

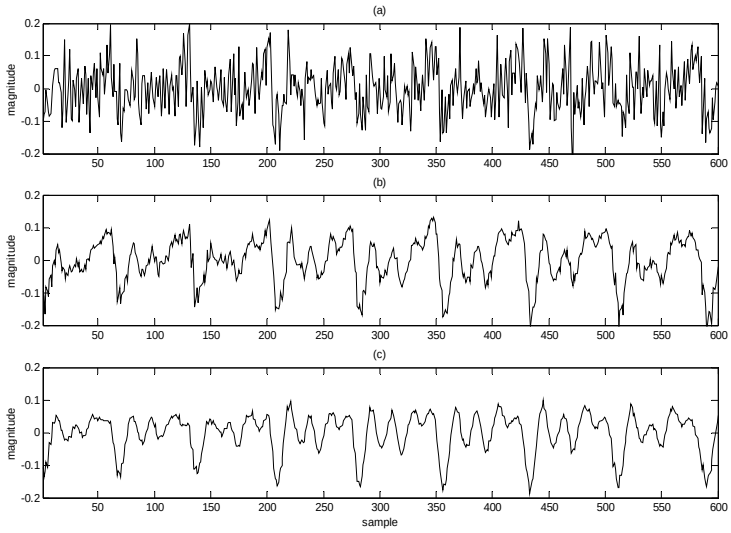

Figure 3.3.8 A zoomed section of HSP from figure 3.3.6

(a) Noisy speech section in $x_1(k)$
(b) Primarily enhanced speech section in $\hat{x}_1(k)$
(c) Final enhanced speech section in $y(k)$

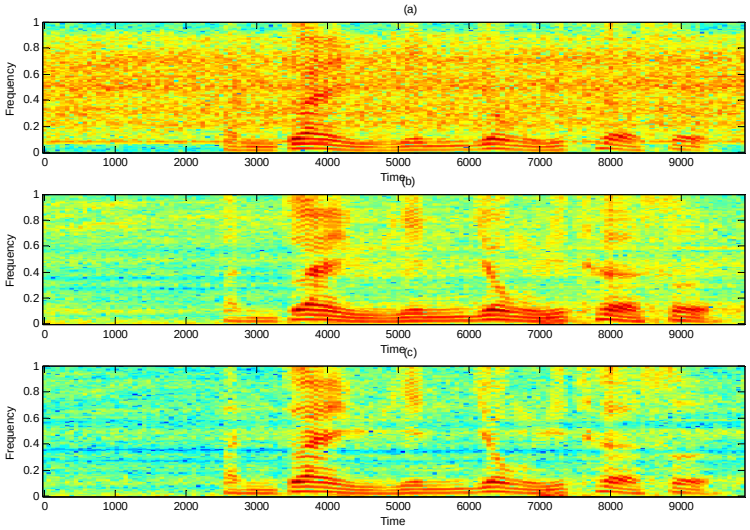

Figure 3.3.9 Spectrograms of the corresponding signals in figure 3.3.6

(a) Spectrogram of noisy speech $x_1(k)$
(b) Spectrogram of primarily enhanced speech $\hat{x}_1(k)$
(c) Spectrogram of final enhanced speech $y(k)$

Figure 3.3.7 shows a zoomed section of NSP from sample 3001 to sample 3600 as shown in figure 3.3.6. From this figure we may see that the proposed scheme has high noise cancellation abilities.

Figure 3.3.8 also illustrates a zoomed section but of HSP from sample 13501 to

sample 14100 as shown in figure 3.3.6. From this figure we may notice that the proposed enhancement system has a better speech enhancement performance.

Figure 3.3.9 shows the spectrograms of the signals in figure 3.3.6.

3.3.4 Conclusions

In this section, a combined scheme of MCRANC and DAS is proposed.

The scheme includes N MCRANC subsystems and a DAS beamformer. In every subsystem, a MCRANC algorithm is used to get a channel of primary enhanced speech. Then all primary enhanced speeches are used as inputs to a DAS beamformer to derive further enhanced speech. The update for the coefficients of the filters in every subsystem is controlled by a VAD-based AMC.

Since the noise signals in a small array are highly correlated with each other, the DAS beamforming has very limited effect for speech enhancement. However, after the MCRANC processing, the noisy speech signals become primarily enhanced speech signals. The residual noises in these primarily enhanced speech signals are much less correlated. So DAS beamforming may take good effect to these signals.

The proposed scheme is adaptive to the changes of the positions of the noise sources and the environment as the filter A_i in the subsystem will update its coefficient vector during the NSP. The proposed scheme is also suitable for many kinds of noises as it employs a noise cancellation method. It performs better than MCRANC algorithm or DAS beamforming alone. In our experiment the achieved SNR improvement reaches 22.3 dB.

3.4 Combined MCRANC with Weiner post-filtering

Weiner post-filtering (WPF) system may be described as a beamformer followed by a Weiner filter. The beamformer offers an enhanced speech and the post-filter gives

further enhancement to the enhanced speech. However, the coefficients of the Weiner filter is not estimated by the enhanced speech, but by the microphone array signals.

This section presents the combinational algorithm of MCRANC with Weiner post-filtering.

3.4.1 Weiner post-filtering

Zelinski's WPF algorithm [131] is the typical Weiner post-filtering algorithm. It contains a DAS beamformer and a Weiner filter. Microphone array signals are input to DAS beamformer to get an enhanced speech. Then a post-filter is cascaded to further enhance the output signal of DAS.

Zelinski's algorithm is indicated in figure 3.4.1. Microphone array signals $x_i(k)$ ($i = 1,2,\cdots,N$) are firstly time-aligned to get signals $\tilde{x}_i(k)$ ($i = 1,2,\cdots,N$)

$$\tilde{x}_i(k) = s_{i_0}(k) + n_i'(k) \tag{3.4.1}$$

where $i_0 \in \{1,2,\cdots,N\}$, and noises $n_i'(k)$ ($i = 1,2,\cdots,N$) are assumed to be mutually uncorrelated. Then the aligned signals $\tilde{x}_i(k)$ ($i = 1,2,\cdots,N$) are summed together and divided by N to get

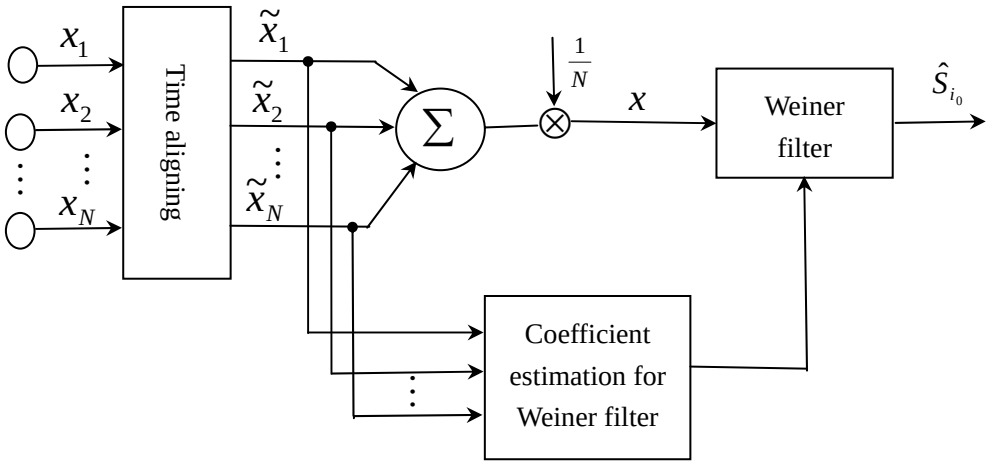

Figure 3.4.1 Zelinski's Weiner post-filtering

$$x(k) = \frac{1}{N}\sum_{i=1}^{N}\tilde{x}_i(k) \tag{3.4.2}$$

$x(k)$ is actually the output of DAS beamformer. At last, the primarily enhanced $x(k)$ is input to Weiner filter to get further enhancement. The coefficients of the Weiner filter may be estimated by aligned signals $\tilde{x}_i(k)$ ($i = 1,2,\cdots,N$).

The Weiner filter with coefficients $w(j)$, defined in the index range $J = \{j : J_1 \le j \le J_2\}$, yields the speech estimate

$$\hat{s}_{i_0}(k) = \sum_{j\in J} w(j)x(k-j) \tag{3.4.3}$$

Minimization of the mean square error $E[(s_{i_0}(k) - \hat{s}_{i_0}(k))^2]$ leads to the Weiner-Hopf equation

$$\sum_{j\in J} w(j)R_{xx}(l-j) = R_{ss}(l) \qquad j \in J \tag{3.4.4}$$

where $R_{xx}(\cdot)$ and $R_{ss}(\cdot)$ are auto-correlations of $x(k)$ and $s_{i_0}(k)$ respectively.

From equation (3.4.4), the coefficients $w(j)$ of the Weiner filter can be obtained if $R_{xx}(\cdot)$ and $R_{ss}(\cdot)$ are known. $R_{xx}(\cdot)$ can be calculated by $x(k)$ directly. Furthermore, if the noises $n_1'(k), n_2'(k), \cdots, n_N'(k)$ and the speech $s_{i_0}(k)$ are mutually uncorrelated, $R_{ss}(\cdot)$ can be estimated by the cross-correlation of $\tilde{x}_i(k)$ and $\tilde{x}_j(k)$

$$\begin{aligned} & E[\tilde{x}_i(k)\tilde{x}_j(k+l)] \\ &= E[(s_{i_0}(k) + n_i'(k))(s_{i_0}(k+l) + n_j'(k+l))] \\ &= R_{ss}(l) \qquad \forall i \ne j \end{aligned} \tag{3.4.5}$$

The convolution computations for estimating $R_{xx}(\cdot)$ and $R_{ss}(\cdot)$ can be carried out in

the frequency domain using Fast Fourier Transform (FFT) with block length L. Each block of $L/2$ consecutive samples $\{\tilde{x}_i(k)\}$ is appended by $L/2$ zeros and transformed into the frequency domain yielding the FFT coefficients

$$\{\tilde{X}_i(m)\} \qquad m = 0,1,\cdots,L-1 \text{ and } i = 1,2,\cdots,N \tag{3.4.6}$$

Then the auto-spectral density

$$A(m) = \frac{1}{N}\sum_{i=1}^{N}\tilde{X}_i(m) \tag{3.4.7}$$

and the cross-spectral density

$$C(m) = \frac{2}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\tilde{X}_i(m)\tilde{X}_j^*(m) \qquad m = 0,1,\cdots,L-1 \tag{3.4.8}$$

where * denotes the conjugate complex value.

The inverse FFTs of $A(m)$ and $C(m)$ lead to the time-domain functions $a(k)$ and $c(k)$, which are the estimates of $R_{xx}(\cdot)$ and $R_{ss}(\cdot)$ respectively. Finally, coefficients $w(j)$ of the Weiner filter may be computed according to equation (3.4.4). Since the frame length L is limited, there must exist estimation error of the cross-spectrum density $C(m)$. This error may result in an audible residual noise in the final enhanced speech $\hat{s}_{i_0}(k)$. This residual noise can be reduced by replacing $C(m)$ with

$$P(m) = \alpha(m)\frac{2}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathrm{Re}(\tilde{X}_i(m)\tilde{X}_j^*(m)) \tag{3.4.9}$$

where

$$\alpha(m) = S_{i_0}^2(m)/[S_{i_0}^2(m) + \frac{2}{N(N-1)}V(m)] \tag{3.4.10}$$

In (3.4.10) $S_{i_0}^2(m)$ and $V(m)$ are estimated by using the following methods.

Define a modified $\tilde{C}(m)$ as

$$\mathrm{Re}\{\tilde{C}(m)\} = \mathrm{Re}\{C(m)\} \quad \text{and} \quad \mathrm{Im}\{\tilde{C}(m)\} = 0 \tag{3.4.11}$$

Replace the negative value of $\tilde{C}(m)$ with 0. Then take a square to the results after the replacement. Finally, take the average of the squared results as the estimation of $S_{i_0}^2(m)$.

To estimate $V(m)$, define

$$\tilde{C}_{ij}(m) = \mathrm{Re}\{\tilde{X}_i(m)\tilde{X}_j^*(m)\} \tag{3.4.12}$$

$$IJ = \{12,13,\cdots,1N,23,24,\cdots,2N,\cdots\cdots,(N-1)N\} \tag{3.4.13}$$

Use all negative value in $\{\tilde{C}_{ij}(m); ij \in IJ\}$ to estimate $V(m)$. Firstly, note

$$\tilde{V}(m) = \frac{1}{M}\sum_{ij}\tilde{C}_{ij}^2(m) \tag{3.4.14}$$

where $ij \in IJ$ and $\tilde{C}_{ij}(m) < 0$, and M is the number of $\tilde{C}_{ij}(m)$ with negative value. Then, use the average of $\{\tilde{V}(m); m = 0,1,\cdots,L-1\}$ as the estimation of $V(m)$.

Post-filtering in the frequency domain

Zelinski's WPF algorithm needs to be processed in the time and frequency domains. For simplicity, Simmer [107] and Fischer [35] proposed the WPF algorithms only in the frequency domain. They also made improvements for estimating the post-filter.

According to Weiner filtering theory, the transfer function of Zelinski's WPF can be described as

$$W_1(e^{j\Omega}) = \frac{\Phi_{ss}(e^{j\Omega})}{\overline{\Phi}_{\tilde{x}\tilde{x}}(e^{j\Omega})} \tag{3.4.15}$$

where $\Phi_{ss}(e^{j\Omega})$ is the Power Density Spectrum (PDS) of the speech signal, and $\overline{\Phi}_{\tilde{x}\tilde{x}}(e^{j\Omega})$ is the average PDS of N aligned signals $\tilde{x}_i$ $(i = 1,2,\cdots,N)$.

Simmer improved the WPF as follows

$$W_2(e^{j\Omega}) = \frac{\Phi_{ss}(e^{j\Omega})}{\Phi_{xx}(e^{j\Omega})} \tag{3.4.16}$$

where $\Phi_{xx}(e^{j\Omega})$ is the PDS of the enhanced speech x by DAS beamforming. Since the post-filtering is processed after DAS beamforming, W_2 is more reasonable than W_1.

The PDS of the speech signal $\Phi_{ss}(e^{j\Omega})$ cannot be obtained directly. However, it can be estimated if it is assumed that the noise signals in different channels are mutually uncorrelated. Under this assumption, the PDS of the speech signal is equal to the PDS of the cross PDS of two noisy speech signals. This is explained as follows.

$$\hat{W}_2(e^{j\Omega}) = \frac{\Phi_{\tilde{x}_i\tilde{x}_j}(e^{j\Omega})}{\Phi_{xx}(e^{j\Omega})} = \frac{\Phi_{ss}(e^{j\Omega}) + \Phi_{n_i'n_j'}(e^{j\Omega})}{\Phi_{xx}(e^{j\Omega})} \tag{3.4.17}$$

where $\Phi_{\tilde{x}_i\tilde{x}_j}(e^{j\Omega})$ is the cross PDS of the aligned signals $\tilde{x}_i$ and $\tilde{x}_j$. Since $\Phi_{n_i'n_j'}(e^{j\Omega}) = 0$, which means the noise signals in $\tilde{x}_i$ and $\tilde{x}_j$ are uncorrelated, $\hat{W}_2$ equals W_2.

The noise signals might not be completely uncorrelated in practical applications. Therefore, it is better to use the average of all cross PDSs as the estimation of $\Phi_{ss}(e^{j\Omega})$. In this way, the WPF will be

$$\hat{W}(e^{j\Omega}) = \frac{\frac{2}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\Phi_{\tilde{x}_i\tilde{x}_j}(e^{j\Omega})}{\Phi_{xx}(e^{j\Omega})} \tag{3.4.18}$$

The above estimation of the WPF in frequency domain can be realized by use of framing, windowing and FFT to the acquired noisy speech signals. Because the speech signal is non-stationary, the length of the window should be limited. Generally, we take 256 as its length if the sampling rate is 8KHz. Therefore, we must estimate the PDS and the cross PDS with limited samples. The short-time spectral estimation algorithm proposed by Nutall and Carter [94] can be applied for the estimates. If a cross PDS has

an imagined part, we use only the real part. Or we may use its mode as the estimate as Fischer supposed, i.e.

$$\hat{W}(e^{j\Omega}) = \frac{\left| \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \Phi_{\tilde{x}_i \tilde{x}_j}(e^{j\Omega}) \right|}{\Phi_{xx}(e^{j\Omega})} \tag{3.4.19}$$

Figure 3.4.2 indicates the structure of post-filtering in the frequency domain.

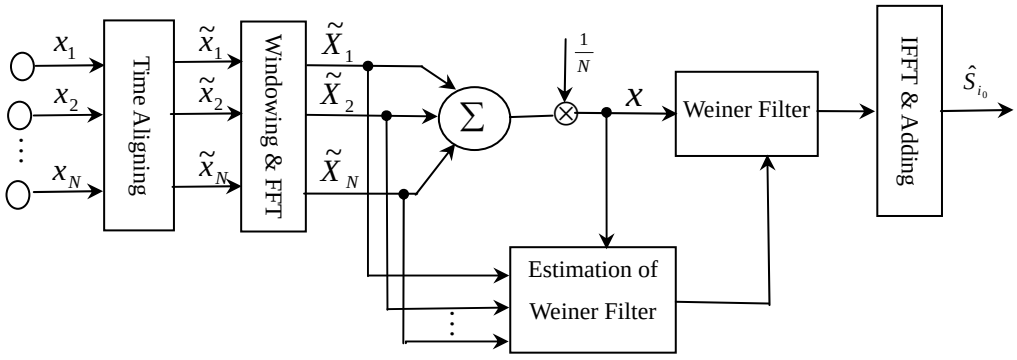

Figure 3.4.2 Structure of the post-filtering in the frequency domain

3.4.2 Combined MCRANC with Weiner post-filtering

In this section, a combined algorithm of MCRANC with Weiner post-filtering will be presented. It follows the same idea as the combined algorithm proposed in section 3.3. That is, it uses MCRANC to primarily enhance every channel of the noisy speech signal. The primarily enhanced speech signals are then cascaded with the post-filtering.

As indicated in figure 3.4.3, the proposed combined algorithm contains a MCRANC module before the post-filtering process. The MCRANC module may enhance the speech signal and make the residual noises in different channels less correlated. So, the output signals of the module would be more suitable for the post-filtering process.

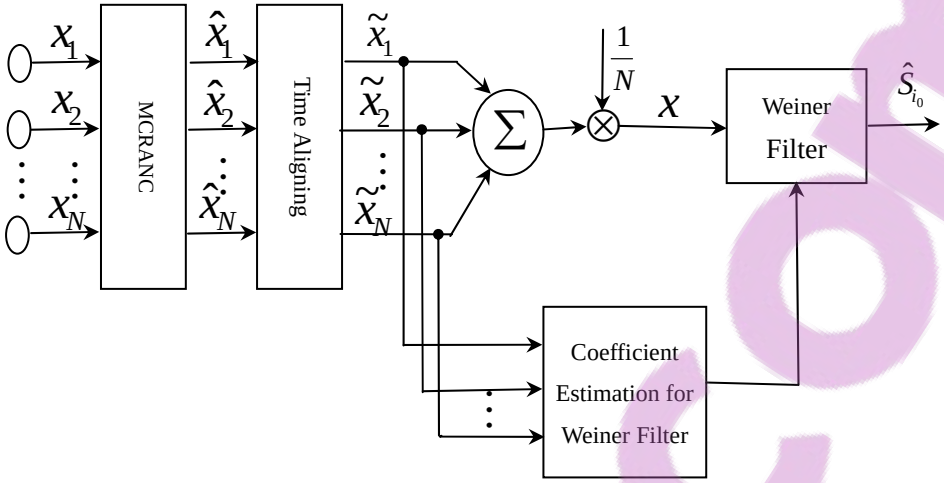

Figure 3.4.3 Combined MCRANC with Weiner post-filtering

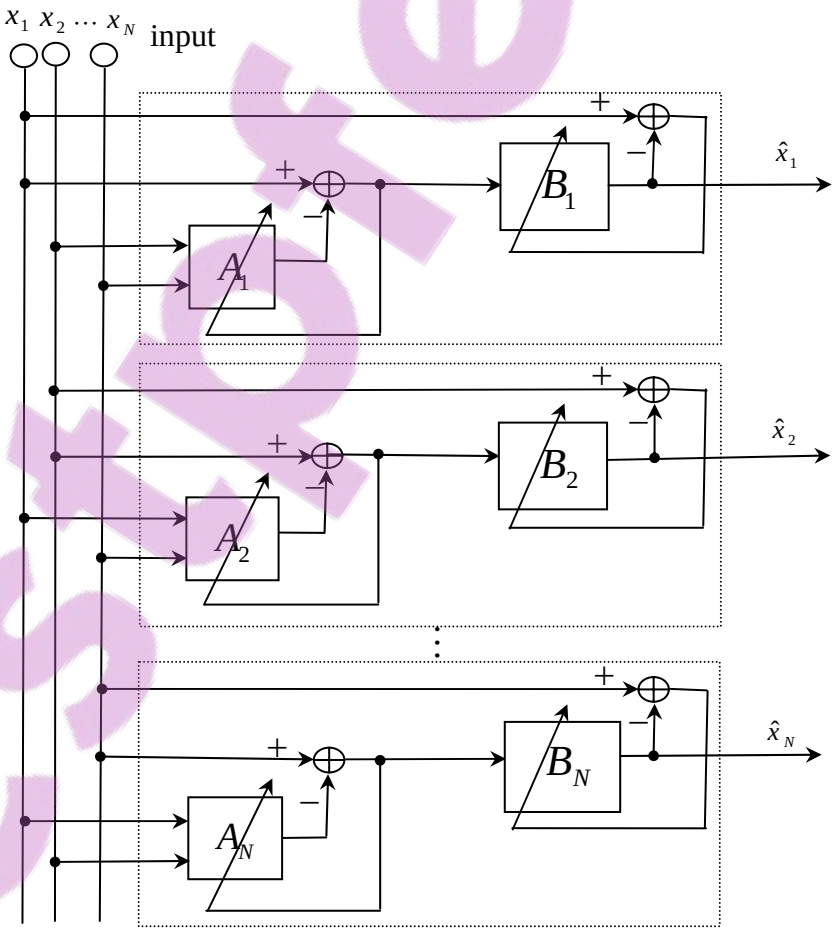

Figure 3.4.4 The MCRANC module

The MCRANC module is similar to that introduced in section 3.3. It is depicted in figure 3.4.4. It contains N subsystems of MCRANC and an AMC is used to control the filters in every subsystem. N is the number of microphones in the array. In the figure, every MCRANC subsystem is in a dotted frame. The details of the i-th MCRANC subsystem have been depicted in figure 3.3.3.

In the i-th subsystem, the i-th channel of the noisy speech signal $x_i(k)$ is used by MCRANC as the main channel signal and the others, $x_j(k)$ ($j = 1, \cdots, i-1, i+1, \cdots, N$), as the referential signals. It has two adaptive filters A_i and B_i. Every subsystem has N channels of input and one channel of output. AMC is used to control the filters in the subsystems, deciding when to adapt their coefficients and when to freeze them.

If the time-aligning can be done in a noisy environment, the MCRANC module may be changed to the rear of the time-aligning module. In this way the AMC will be much simpler since the speech signals have already been aligned.

It is easy to recognize the above algorithm as adding a Weiner filter to the enhanced speech gained by use of the combined algorithm proposed in section 3.3. As a result, the enhanced speech gained by the algorithm proposed in this section will be, generally speaking, better than the enhanced speech gained by the algorithm proposed in section 3.3.

3.4.3 Experimental results

The experiment was the same as that described in subsection 3.3.3. It used the same microphone array and the same 5 channels of the acquired noisy speech signals for evaluation. The processing methods and parameters for the MCRANC module and the time-aligning module were also exactly the same as those in section 3.3.3.

For WPF, the frequency domain method was employed to estimate the Weiner filter. The frame length was 256 and a Hamming window was used with 50% of the neighborhood frame overlapped. Equation (3.4.19) was used to estimate the Weiner filter.

The final enhanced speech has a SNR improvement of 24.1 dB, 1.8 dB more than the final enhanced speech in section 3.3. The figures for depicting the experimental results are very similar to the corresponding figures in subsection 3.3.3. Thus, they will not be depicted here again.

3.4.4 Conclusions

A combined algorithm of MCRANC with Weiner post-filtering is proposed after introducing WPF. It may further improve the enhanced speech gained by use of the combined algorithm of MCRANC and DAS as proposed in section 3.3. The reason for the improvement is that the algorithm proposed in this section is actually adding a Weiner filter to the output of the algorithm proposed in section 3.3. However, the Weiner filter is not estimated by only one channel of signal, but by multiple channels of array signals.

3.5 Summary

In this chapter, three combinational algorithms of MCRANC with existing single-channel or microphone array speech enhancement methods are proposed.

For the combination with a single-channel method, this chapter mainly presents the cascade of MCRANC with ISS. Experimental results prove that the combination algorithm outperforms either MCRANC or the ISS alone. Similarly, MCRANC may be cascaded by other one-channel speech enhancement algorithms.

For the combination with microphone array methods, this chapter mainly discusses the combination with DAS beamforming and the combination with Weiner post-filtering. Both of the combinations are realized by using MCRANC to pre-enhance every channel of the array signal and then employing the existing array algorithms. The pre-enhancement by MCRANC is provided to cancel the correlative part of the noises, while the array algorithms are employed to suppress uncorrelated part of the noises.

Similarly, MCRANC may be combined with other array algorithms, which can suppress uncorrelated noises, to achieve better speech enhancement results.

Chapter 4 Improved MCRANC Methods

This chapter discusses the methods of improving MCRANC itself. The multichannel inputs method to the second stage filter in MCRANC is firstly proposed to get better enhanced speech. Secondly, multiple sampling rates method is studied, in which the main channel signal and referential channel signals employ different sampling rates. It is suggested that the sampling rate for referential channel signals should be higher or lower, according to the noise type, than the required sampling rate for the output speech. Thirdly, fixed beamforming MCRANC, partial-channel MCRANC and their combination called fixed beamforming partial-channel MCRANC are proposed respectively. They may give MCRANC more applications and make it more effective. Delay And Weighted Sum beamforming is also presented for the fixed beamformer. Finally, the subband MCRANC is proposed.

4.1 MCRANC with multichannel distorted signal filtering

In chapters 2 and 3 it is shown that Multichannel Crosstalk Resistant Adaptive Noise Cancellation (MCRANC) has only one channel input signal for its second stage filter B. This one channel input signal is the output of the first stage filter A. It is actually a distorted speech signal, as pointed out in section 2.4 of chapter 2. The function or purpose of filter B is to change the distorted speech signal to the normal speech signal containing in the main channel.

Similar to MANC, if there are more channels of distorted speech signals input to the second stage filter B, the output speech from it would be better. This is the reason to use multichannel distorted signals for MCRANC.

In fact, it is easy to get other channels of distorted speech signals. If we select any reference channel signal in MCRANC as the main channel signal and all other signals, including other reference channel signals and the main channel signal in MCRANC, as the reference signals to form a new MCRANC, the output of filter A in the new MCRANC would be a new channel distorted speech signal.

The above is the principle of MCRANC using Multichannel Distorted Signal (MDS) filtering.

4.1.1 Description of the method

Use all the notations described in subsection 3.3.2.1 of chapter 3 in which MCRANC is detailed.

Figure 4.1.1 indicates a MCRANC system using MDS for its second stage filter. This improved MCRANC is noted as MDS-MCRANC. It selects $x_i(k)$ from microphone M_i as the main channel signal and other signals from other $N-1$ microphones $x_j(k)$ ($j=1,\cdots,i-1,i+1,\cdots,N$) as the referential signals.

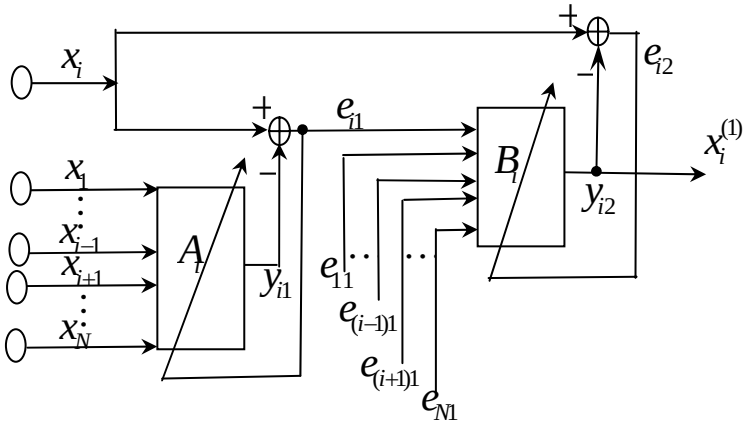

Figure 4.1.1 MCRANC using multichannel distorted speech filtering

Common MCRANC has only one channel of input $e_{i1}(k)$ for its filter B_i. The

improved MCRANC has N channels of inputs $e_{11}(k), e_{21}(k), \cdots, e_{N1}(k)$ for its filter B_i, where $e_{j1}(k)$ is the output of filter A_j in a common MCRANC using $x_j(k)$ as the main channel signal and others as the referential channel signals ($j = 1,2,\cdots,N$). As pointed out in subsection 3.3.2.2, $e_{j1}(k)$ is a distorted speech signal with residual noise, i.e.

$$\begin{aligned} e_{j1}(k) &= x_j(k) - y_{j1}(k) \\ &= p_j(k) + e_{j1}^*(k) \qquad (j = 1,2,\cdots,N) \end{aligned}$$

where

$$p_j(k) = s_j(k) - \mathbf{w}_j^* \mathbf{s}_j(k)$$

$p_j(k)$ is the distorted speech signal and $e_{j1}^*(k)$ is the residual noise. We may get N channel distorted speech signals with residual noises $e_{j1}(k)$ ($j = 1,2,\cdots N$).

If only one $e_{i1}(k)$ is to be input to the filter B_i and the correlation between the distorted speech signal $p_i(k)$ and the speech signal $s_i(k)$ is not very high, the minimal error between $y_{i2}(k)$ and $s_i(k)$ will not be very small. This means the recovered speech signal $y_{i2}(k)$ can not approximate speech $s_i(k)$ as we expect. If we input N channel distorted speech signals $e_{j1}(k)$ ($j = 1,2,\cdots N$) into filter B_i, its output $y_{i2}(k)$ would be more approximate to $s_i(k)$ since all $e_{j1}(k)$ contains distorted speech signals. The reason is similar to what we use multiple inputs for filter A_i in common MCRANC, as described in chapter 2.

4.1.2 Combined with DAS beamforming

As in the common MCRANC algorithm, MDS-MCRANC can also be combined with many speech enhancement algorithms such as ISS, DAS beamforming and Weiner

post-filtering, etc. In this subsection, we only present its combination with DAS beamforming.

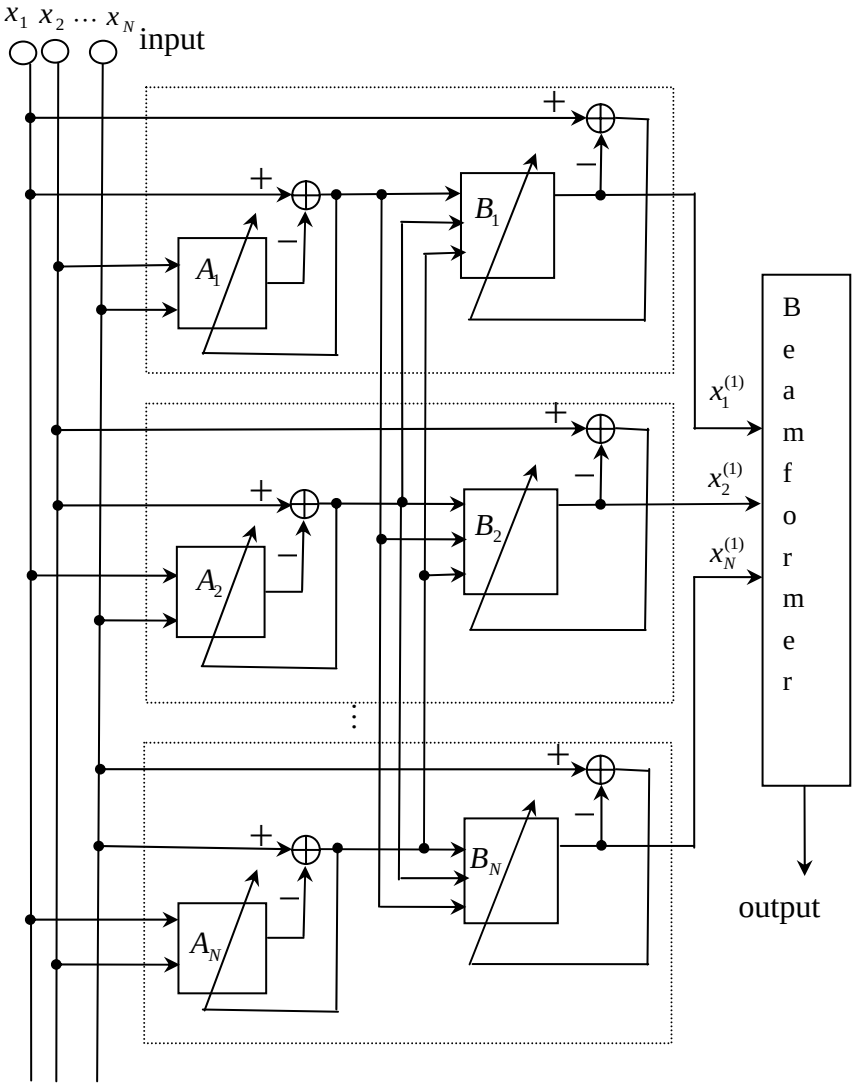

Figure 4.1.2 Combined MDS-MCRANC with DAS beamforming

As shown in figure 4.1.2, the combined algorithm consists of N MDS-MCRANC subsystems and a DAS beamformer. N is the number of microphones employed in the array. Every subsystem is MDS-MCRANC and is presented in a dot-line frame in the figure. The adaptations of filters in the subsystems are controlled by a VAD-basd AMC.

Figure 4.1.2 is similar to figure 3.3.2 in chapter 3. The only difference lies in the N subsystems. For every new subsystem its filter B_i has N inputs, which are the outputs

of all filters A_i ($i=1,2,\cdots N$), other than only one input which is from filter A_i. A AMC is used to control when to adapt the coefficients of the filters in the whole system. In fact, every output of a subsystem is a channel of primarily enhanced speech. All outputs of the subsystems are to be input to a DAS beamformer to get final enhanced speech.

4.1.3 Comments

The proposed MDS-MCRANC in section 4.1.1 has multiple inputs for the second filter for speech signal recovering. It may get better performances than common MCRANC since common MCRANC employs only one input for its second filter. However, common MCRANC is mainly suitable for the applications where spatial correlation of the speech signal is quite high (as is the case if the microphone array size is small and the speech source is near the array). Therefore, further improvement by MDS-MCRANC would usually be minor over the enhanced speech obtained by common MCRANC.

Similarly, the combined algorithm of MDS-MCRANC with DAS proposed in section 4.1.2 will also achieve better performances than the combination of MCRANC with DAS introduced in subsection 3.2. However, the improvement is also minor.

4.1.4 Experimental results

This experiment is exactly the same as described in section 3.3 of chapter 3. All parameters are also the same except the length of filter B_i because now there are five channels, other than one channel, of inputs to filter B_i. The length of B_i is now selected to be 5x24=120.

Finally, the SNR of the final enhanced speech by use of the combined algorithm of MDS-MCRANC with DAS is 23.5 dB, only 1.2 dB more than the SNR of the enhanced speech achieved by the combined algorithm of MCRANC with DAS beamforming.

4.1.5 Conclusions

In this section an improved MCRANC with multichannel distorted signal filtering is proposed first. Then a combination method of MDS-MCRANC with DAS beamforming is presented. The MDS-MCRANC has better speech enhancement performance than common MCRANC, and the combination method has better performance than the combination of common MCRANC with DAS.

4.2 MCRANC using multiple sampling rates

4.2.1 Description of the method

In MCRANC as described in section 2.4 of chapter 2, we need to use the referential channel noise signals $n_1(k),\cdots,n_N(k)$ to cancel the main channel noise $n_0(k)$ through the first stage filter A. The more the noise cancellation, the better the speech enhancement performance.

If we use the same sampling frequency f for all analogue signals $n_i(t)$ ($i=0,1,\cdots,N$) acquired from microphones, we get the discrete signals

$$n_i(k)=n_i(kT)=n_i(t)|_{t=kT} \qquad i=0,1,\cdots,N \tag{4.2.1}$$

where $T=\frac{1}{f}$ is the sampling time interval. Suppose the number of sample delay for $n_i(k)$ in filter A is L, then

$$n_0(k)=\mathbf{w}\mathbf{n}(k)+e_1(k) \tag{4.2.2}$$

where $e_1(k)$ is the prediction error, and

$$\mathbf{w}=(\mathbf{w}_1,\mathbf{w}_2,\cdots,\mathbf{w}_N) \tag{4.2.3}$$

$$\mathbf{w}_i = (w_{i0}, w_{i1}, \cdots, w_{iL}),$$

$\mathbf{w}$ is the coefficient vector of the filter A, and

$$\mathbf{n}(k) = [\mathbf{n}_1(k), \mathbf{n}_2(k), \cdots, \mathbf{n}_N(k)]^T \tag{4.2.4}$$

$$\mathbf{n}_i(k) = [n_i(k), n_i(k-1), \cdots, n_i(k-L)]$$

Denote the power of $e_1(k)$ as $P[e_1(k)]$, and $P[e_1^*(k)]$ as the minimal prediction error power in minimizing $P[e_1(k)]$ through equation (4.2.2). The corresponding optimal coefficient vector of filter A is denoted as $\mathbf{w}^*$.

In noise cancellation, especially in the process of the cancellation for high frequency and wideband noises, the optimal power $P[e_1^*(k)]$ is usually not zero or very small as we expect. This may consequently make the residual noise in final enhanced speech also not small. Even if we greatly increase the sample delay number L in filter A (this makes the length of filter A increase greatly), the optimal power $P[e_1^*(k)]$ can still not be reduced.

In this section, the multiple sampling rates method is proposed to reduce the optimal power $P[e_1^*(k)]$. The method applies a higher or lower sampling rate to the referential channel signals while using the ordinary sampling rate as required for the main channel signal. The following description refers to the case in which a higher sampling rate is applied.

Denote the common sampling rate as f (for example $f = 8\,\text{K Hz}$), and another higher sampling rate as f'. Usually we take

$$f' = pf \tag{4.2.5}$$

where p is a positive integer. If we sample the referential channel noises with sampling rate f', we get

$$n_i'(k') = n_i(t)|_{t=k'T'} \qquad k' = 0,1,2,\cdots \qquad i = 1,\cdots,N \tag{4.2.6}$$

where $T' = \frac{1}{f'} = \frac{1}{pf} = \frac{T}{p}$. Let's consider the p sub-sequences of $n_i'(k')$

$$n_i^{(j)}(k) = \{n_i'(j), n_i'(p+j), \cdots, n_i'(pk+j), \cdots\} \qquad j = 0,1,\cdots, p-1 \tag{4.2.7}$$

Obviously, they all are sequences by sampling $n_i(t)$ with sampling rate f. The differences among them are the beginning time of the sampling. $n_i^{(j+1)}(k)$ has a time delay T' more than $n_i^{(j)}(k)$ in the beginning. This can be viewed as a group of array signals acquired from a linear array with p microphones $M_i^{(0)}, M_i^{(1)}, \cdots, M_i^{(p-1)}$ by use of sampling rate f. So, from this perspective, any microphone M_i with a higher sampling rate can play as a virtual microphone array with a lower sampling rate. Then we have a similar equation to (4.2.2)

$$n_0(k) = \mathbf{w}'\mathbf{n}'(k) + e_1'(k) \tag{4.2.8}$$

where

$$\mathbf{w}' = (\mathbf{w}_1', \mathbf{w}_2', \cdots, \mathbf{w}_N')$$

$$\mathbf{w}_i' = (\mathbf{w}_i^{(0)}, \mathbf{w}_i^{(1)}, \cdots, \mathbf{w}_i^{(p-1)}) \qquad i = 1, \cdots, N$$

$$\mathbf{w}_i^{(j)} = (w_{i0}^{(j)}, w_{i1}^{(j)}, \cdots, w_{iL'}^{(j)}) \qquad j = 0,1,\cdots, p-1 \tag{4.2.9}$$

$\mathbf{w}'$ is the coefficient of filter A. It is a row vector with $Np(L'+1)$ elements. And

$$\mathbf{n}'(k) = [\mathbf{n}_1'(k), \mathbf{n}_2'(k), \cdots, \mathbf{n}_N'(k)]^T$$

$$\mathbf{n}_i'(k) = [\mathbf{n}_i^{(0)}(k), \mathbf{n}_i^{(1)}(k), \cdots, \mathbf{n}_i^{(p-1)}(k)] \qquad i = 1, \cdots, N \tag{4.2.10}$$

$$\mathbf{n}_i^{(j)}(k) = [n_i^{(j)}(k), n_i^{(j)}(k-1), \cdots, n_i^{(j)}(k-L')] \quad j = 0,1,\cdots, p-1 \tag{4.2.11}$$

$\mathbf{n}'(k)$ is a column vector with $Np(L'+1)$ elements. $e_1'(k)$ is the prediction error by use of high sampling rate signal $n_i'(k')$ ($i = 1, \cdots, N$). Denote the optimal coefficient and optimal prediction error in (4.2.8) as w'^* and $e_1'^*(k)$.

If we take $L' = L$ and suppose the optimal coefficient can be found, then

$$P[e_1^{'*}(k)] \le P[e_1^*(k)] \tag{4.2.12}$$

This is because if we take $\mathbf{w}^{'} = (\mathbf{w}^*, \mathbf{0}, \cdots, \mathbf{0})$ we may have

$$\begin{aligned} e_1^{'}(k) &= n_0(k) - \mathbf{w}^{'}\mathbf{n}^{'}(k) \\ &= n_0(k) - \mathbf{w}^*\mathbf{n}^{'}(k) = e_1^*(k) \end{aligned} \tag{4.2.13}$$

From (4.2.13) we see that (4.2.12) is theoretically true.

However, to find the optimal coefficient $\mathbf{w}^{'*}$ through (4.2.8) will take many more computations than to find $\mathbf{w}^*$ through (4.2.2). Although from (4.2.12) the effectiveness of the noise cancellation will theoretically increase with p, the positive p should not be much bigger since the length of a section of pure noise is limited and the accuracy of the computing equipment is also limited. In our experiments we found that $p = 2 \sim 4$ will be suitable if the noise contains high frequency, such as white noise, and the sampling rate for the main channel signal is 8 kHz.

We also found in our experiments that no matter how big the L is, even $L >> \tilde{L} = Np(L^{'}+1)$, the inequation (4.2.12) will usually hold if only $L^{'}$ and p are properly selected. In the case of high frequency and wideband noise, $P[e_1^{'*}(k)]$ would usually be much smaller than $P[e_1^*(k)]$.

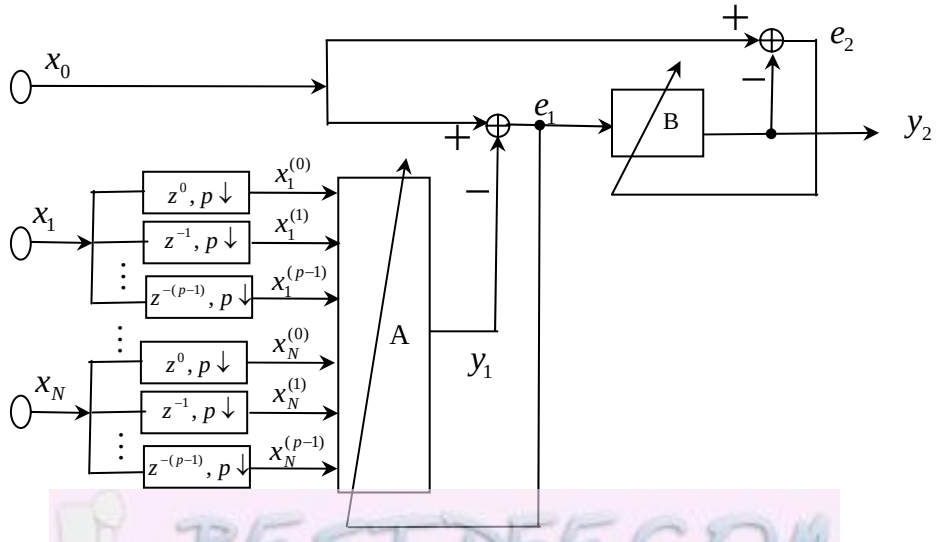

Figure 4.2.1 The structure of multiple sampling rates MCRANC

Multiple sampling rates MCRANC (MSR-MCRANC) method is indicated in figure 4.2.1. It can be summarized as follows.

(1) In a microphone array, select one microphone as the main microphone and others as referential microphones.

For example, if there are 1+N ($N \geq 1$) microphones M_0, $M_1, \cdots, M_N$, select M_0 as the main microphone and $M_1, \cdots, M_N$ as the referential ones.

(2) Sample the signal acquired from the main microphone with the required sampling rate for output speech, and the signals from the referential microphones with a higher sampling rate.

For example, sample signal $x_0(t)$ from the main microphone M_0 with sampling rate $f = 8\,\mathrm{K}$ Hz to get $x_0(k)$, $x_0(k) = x_0(t)|_{t=k/f}$, $k = 0,1,2,\cdots$; and sample referential signal $x_i(t)$ from microphone M_i with sampling rate $f' = pf = 24\,\mathrm{K}$ to get $x_i(t)$, where $p = 3$, $i = 1,\cdots,N$, $x_i(k') = x_i(t)|_{t=k'/f'}$, $k' = 0,1,2,\cdots$.

(3) Down sample every referential signal to get a group of signals which have the same sampling rate as the main channel signal.

For example, for every $x_i(k')$ from microphone M_i, let

$$x_i^{(j)}(k) = \{x_i(j), x_i(p+j), \cdots, x_i(pk+j), \cdots\},\ j = 0,1,\cdots,p-1, i = 1,\cdots,N$$

Then Np referential signals with the same sampling rate as the main channel signal can be obtained.

(4) Use $x_0(k)$ as the main channel signal and $x_i^{(j)}(k)$ ($j = 0,1,\cdots,p-1$, $i = 1,\cdots,N$) as the Np referential signals to process speech enhancement using common MCRANC, where p may be adjusted according to the noise encountered.

As to the realization of multiple sampling rates, one easy way is to employ an over-sampling method. That is, to sample all microphone signals with a higher sampling rate than the required rate for the output speech, including the main channel and referential channel signals. Then, down sample for the signals which does not need the higher sampling rate.

For example, we may use sampling rate $f^{'} = pf$ to sample all microphone signals $x_i(t)$ $(i = 0,1,\cdots,N)$ to get $x_i(k^{'})$ $(i = 0,1,\cdots,N)$. Then down sample $x_0(k^{'})$ to get $x_0(k)$.

4.2.2 Improved MSR-MCRANC

As with filter A, filter B in MSR-MCRANC is best to have its input signal at a higher sampling rate to get better speech enhancement performance. In other words, in figure 4.2.1 e_1 should have a higher sampling rate than y_2, where y_2 has a common sampling rate, say 8 kHz.

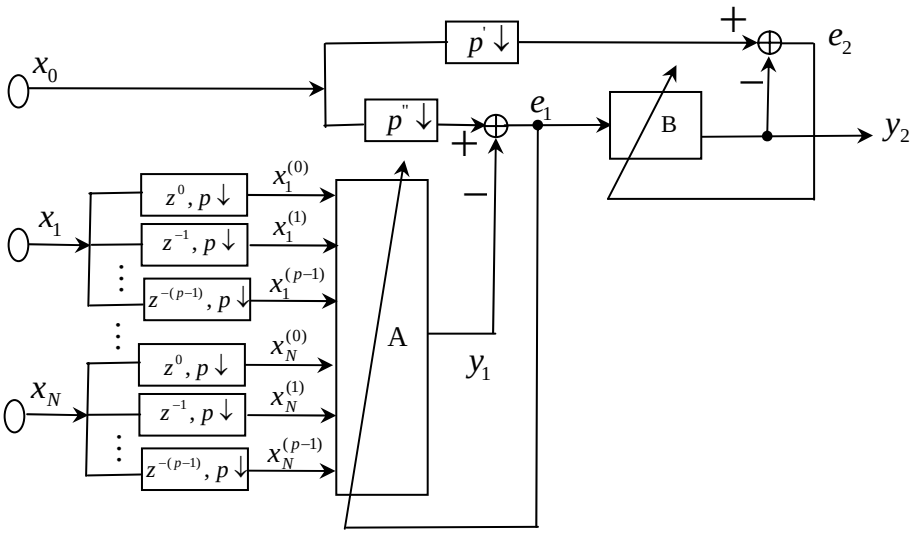

Figure 4.2.2 Improved multiple sampling rates MCRANC

We may realize the above idea by sampling the main channel signal $x_0(t)$ with higher sampling rate $f^{''}$. After we get the high sampling rate signal $x_0(k^{''}) = x_0(t)|_{t=k/f^{''}}$, down sample it into two channel signals for the MCRANC as

shown in figure 4.2.2. One channel is down sampled by picking one sample in every p' samples, and another channel is down sampled by picking one in every p'' samples, where $p' > p''$.

For example, we may sample all microphone signals at sampling rate 32K Hz at first, and then take $p = 4$, $p' = 4$, $p'' = 2$. Thus, e_1 in figure 4.2.2 will have a sampling rate of 16K Hz and the system output y_2 will have a required sampling rate of 8 kHz.

4.2.3 Applied situations

The above multiple sampling rates MCRANC and its improvement may find its applications in an array that employs only a few microphones. For a very small microphone array, say the array in a mobile phone or hearing aid, the array may contain only 2 or 3 microphones. Thus the above algorithms can be considered.

If there are quite a lot of microphones in the array, say more than 5 microphones, the improvement of MSR-MCRANC over common MCRANC will usually be quite limited. Sometimes its enhancement might even become worse. The reason is that too many coefficients in the adaptive filter may cause their optimal solutions to be more inaccurate. So, in this situation, there is no need to employ a higher sampling rate.

Similarly, if the sampling rate for the array signals is already very high, there is also no need to employ a higher sampling rate. We found that if the sampling rate for the array signals is above 32K Hz there is no good to use higher sampling rate.

If the noise is a low frequency noise, there is also no need to employ a higher sampling rate for the referential signals. On the contrary, to cancel the noise efficiently, the sampling rate for the referential signals should be lower than the sampling rate of the main channel signal.

The MSR-MCRANC with a lower sampling rate for referential microphones will not be duplicated here since it is similar to the above descriptions in this section.

4.2.4 Experimental results

The experiment was processed in a common room of 5x4x2.8m. There were desks, chairs and computers in it. The array had only two small microphones, with a distance between them of only 2cm. The speaker was 30cm in front of the array and the noise source, an improperly tuned radio, was 100cm from the array. The noise emitting from the radio was white noise-like. The main microphone was facing the speaker while the referential microphone was facing the radio. The two facing directions formed a 60^0 angle.

The sampling rate required for the output speech is 8 kHz. We used 24KHz as the sampling rate for the array signals.

Figure 4.2.3 (a) shows the noisy speech signal x_0 acquired by the main microphone. It is down sampled to 8 kHz as required. Its SNR is 2.86 dB. By listening, we found that the speech was badly degraded by the noise and we could hardly understand what the speaker said.

Figure 4.2.3 (b) shows the same section of the noisy speech signal x_1 acquired by the referential microphone. Its sampling rate is 24K Hz. Its SNR is 2.73 dB.

Figure 4.2.3 (c) depicts the enhanced speech signal by MCRANC with the same sampling rate of 8 kHz for its main and referential signals. Its SNR is 12.08 dB. In this case, both the main signal and the referential signal are down sampled from 24K Hz to 8 kHz before the speech enhancement process starts.

Figure 4.2.3 (d) depicts the enhanced speech by MSR-MCRANC. Its SNR is 21.30dB. In this case, the main signal is down sampled to 8 kHz as required while the sampling rate for the referential signal is kept at 24K Hz. By listening, we found that the enhanced speech was clear and we could clearly understand what the speaker said.

In the above MCRANC processing and MSR-MCRANC processing, filters A and B are FIR filters with L=32 for A and L_B=48 for filter B. So, in MSR-MCRANC processing the orders of filters A and B are 98 and 48 respectively, and in common MCRANC processing the orders are 32 and 48 respectively.

If we increased the order of filter A from 32 to 98 in common MCRANC

processing, both common MCRANC and MSR-MCRANC processing will have exactly the same computational cost. However, we found the SNR of enhanced speech by common MCRANC had no improvement and even reduced to 11.82 dB. It is much lower than the SNR 21.30 dB achieved by MSR-MCRANC with the same computational cost.

In this experiment, SNR is still computed by use of SNR formula (1.4.3). The LMS algorithm is used for all adaptations of filters A and B, with learning rate $\mu = 0.028$ for filter A and learning rate $\mu = 0.02$ for filter B.

Figure 4.2.4 shows the zoomed figures during a NSP section of figure 4.2.3.

Figure 4.2.5 shows the zoomed figures during a HSP section of figure 4.2.3.

Figure 4.2.6 indicates the spectrograms of the corresponding signals in figure 4.2.3.

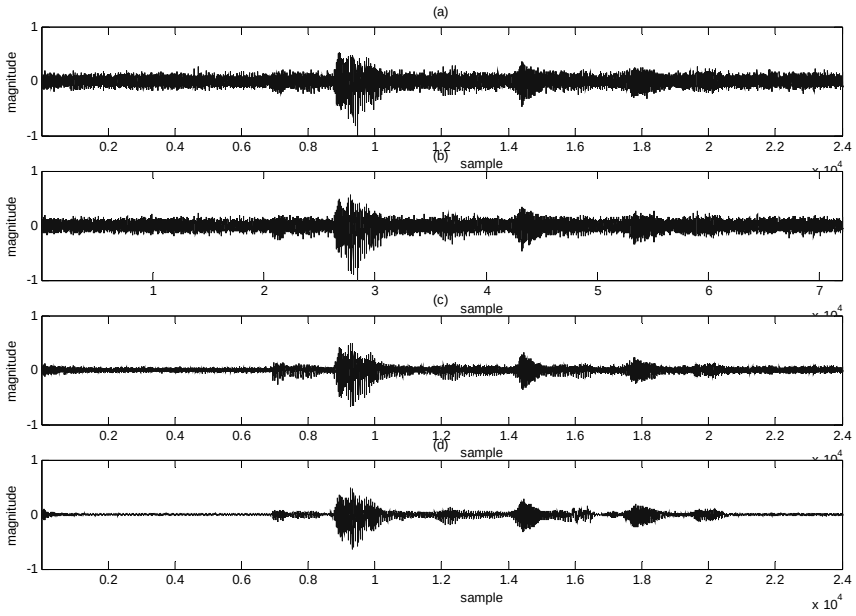

Figure 4.2.3 Experimental results

(a) Noisy speech in main microphone
(b) Noisy speech in referential microphone
(c) Enhanced speech by MCRANC
(d) Enhanced speech by MSR-MCRANC

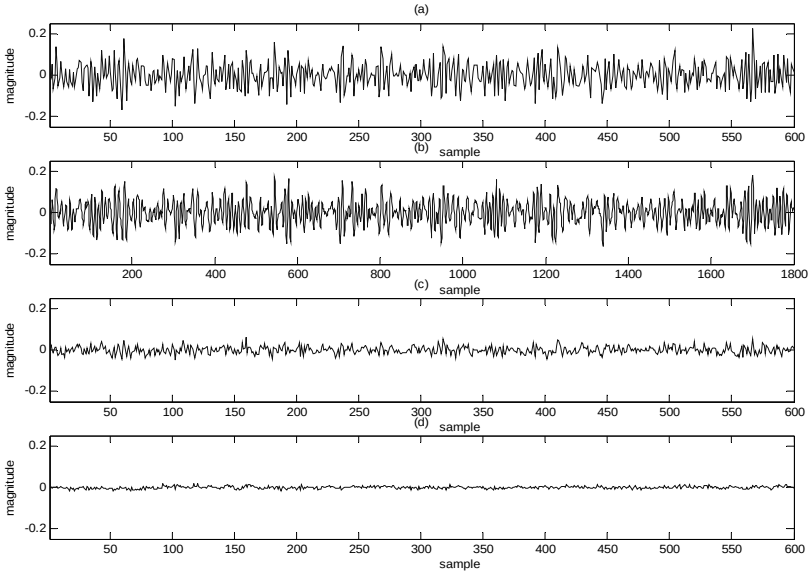

Figure 4.2.4 A zoomed section of figure 4.2.3 (non speech section)

(a) Noise in main microphone
(b) Noise in referential microphone
(c) Residual noise by MCRANC
(d) Residual noise by MSR-MCRANC

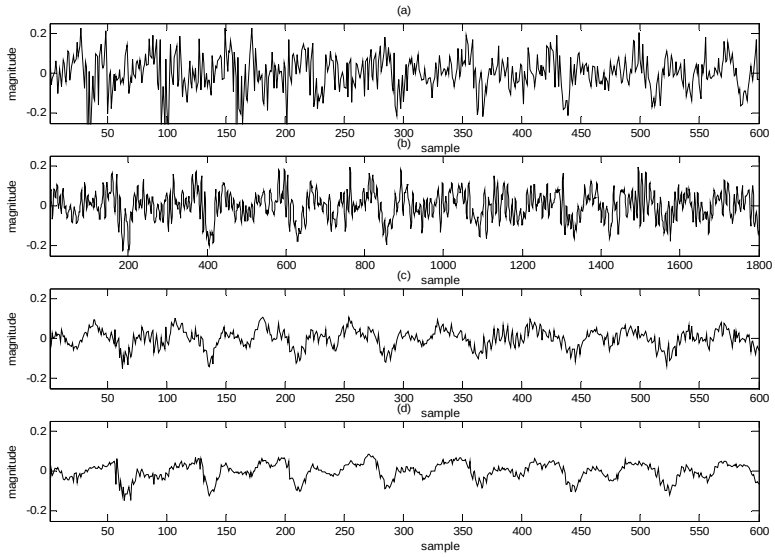

Figure 4.2.5 A zoomed section of figure 4.2.3 (speech section)

(a) Noisy speech in main microphone
(b) Noisy speech in referential microphone
(c) Enhanced speech by MCRANC
(d) Enhanced speech by MSR-MCRANC

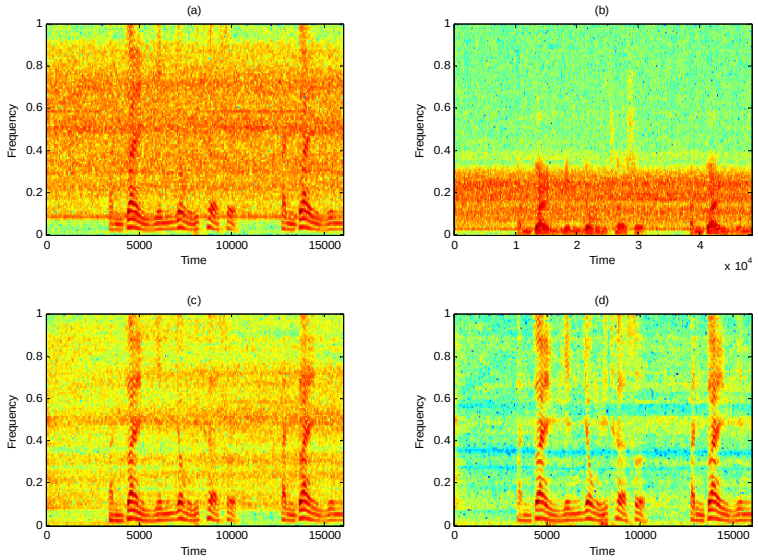

Figure 4.2.6 Spectrograms of the signals in figure 4.2.3

(a) Spectrogram of noisy speech in main microphone.
(b) Spectrogram of noisy speech in referential microphone.
(c) Spectrogram of enhanced speech by MCRANC.
(d) Spectrogram of enhanced speech by MSR-MCRANC

4.2.5 Conclusions

An improvement to MCRANC is proposed by employing different sampling rates for the main channel signal and the referential channel signals. The sampling rate for referential channel signals should be higher or lower, according to the noise frequency, than the required sampling rate for the output speech, which is also the sampling rate for the main cannel signal. If the noises are mainly high frequency-contained, the sampling rate for referential signals should be higher than the normal sampling rate required by the output speech. If the noises are mainly low frequency-contained, the sampling rate for referential signals should be lower than the normal rate. The different rates could be decided by experiment according to the applications to different noisy environments. Experimental results indicate that the MSR-MCRANC may improve the speech enhancement performance obtained by common MCRANC.

4.3 Fixed beamforming partial-channel MCRANC

4.3.1 Fixed beamforming MCRANC

In the MCRANC speech enhancement system, we usually select the acquired signal with highest SNR as the main signal. In this way, the enhanced speech will usually have higher SNR than the enhanced speech by using lower SNR signal as the main signal.

The main idea of fixed beamforming MCRANC (FBF-MCRANC) is to use the output of a fixed beamformer as the main channel signal and use all signals from the array as the referential signals in a MCRANC system. Since the output of the fixed beamformer is an enhanced speech, this enhanced speech usually has higher SNR than the unprocessed main signal in a common MCRANC system. The structure of FBF-MCRANC is shown in figure 4.3.1.

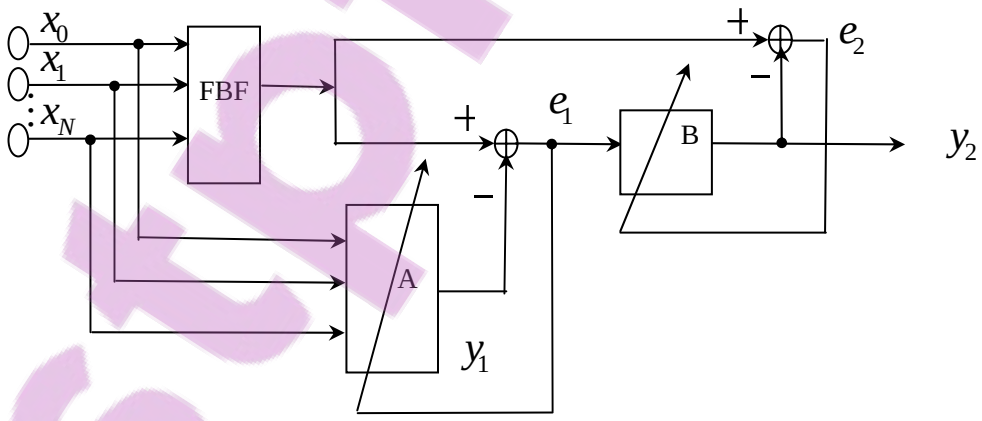

Fig. 4.3.1 Structure of fixed beamforming MCRANC

In figure 4.3.1, FBF indicates the fixed beamformer. There are many fixed beamforming algorithms. The commonly used one is DAS (Delay And Sum) beamforming. The output of DAS is

$$x(k) = \frac{1}{N+1}\sum_{i=0}^{N} x_i(k+\tau_i) \tag{4.3.1}$$

where τ_i is the time delay of the i-th channel signal with respect to the referential

i_0-th channel signal, and $\tau_{i_0} = 0$, $i_0 \in \{0,1,\cdots,N\}$.

However, if the array is a solid microphone array as shown in figure 3.2.2 in chapter 3, or the speaker is very near the microphone array, the SNRs of different channel signals might be quite different. In this circumstance, DAS is not an optimal solution for getting the highest SNR signal. So, an improved DAS algorithm called Delay And Weighted Sum (DAWS) beamforming algorithm will be presented in this section.

In addition, the number of microphones in a small array is usually limited and the noise correlations might not be small. All these facts might cause the SNR of the output signal of DAS or DAWS to be smaller than the SNR of some channel signal in the array. To avoid this happening, we select the highest SNR signal among $\{x, x_0, x_1, \cdots, x_N\}$ to be the main channel signal of MCRANC, where x is the output of the DAWS beamformer. We call this treatment as the DAWS and Selection (DAWSAS) method or algorithm.

4.3.2 Delay And Weighted Sum beamforming

DAWS beamforming is described as

$$x(k) = \sum_{i=0}^{N} \alpha_i x_i(k + \tau_i) \tag{4.3.2}$$

where $\alpha_i \geq 0$ is the weight of the i-th channel signal, and

$$\sum_{i=0}^{N} \alpha_i = 1 \tag{4.3.3}$$

Obviously, if all $\alpha_i = \frac{1}{N+1}$, DAWS becomes to DAS.

If only some weights $\alpha_{i_1}, \alpha_{i_2}, \cdots, \alpha_{i_M}$ are not zero, DAWS becomes partial-channel beamforming since it actually discards the signals with weight 0. This partial-channel beamforming is useful for some arrays with special structures. We may select some signals with high SNR and discard other signals with lower SNR for DAS beamforming,

i.e.

$$\alpha_i = \begin{cases} \dfrac{1}{M} & i \in \{i_1, i_2, \cdots, i_M\} \\ 0 & others \end{cases}$$

What we concern with is how to select the weights to get the highest SNR output of the DAWS when the SNRs of the input signals are different. We may solve this problem as follows.

For simplicity, assume $x_0, x_1, \cdots, x_N$ to be time-aligned and magnitude-aligned signals, i.e.

$$x_i(k) = s(k) + \beta_i n_i(k) \qquad i = 0,1,\cdots,N \qquad (4.3.4)$$

where $n_i(k)$ represents noise and all $n_i(k)$ have the same statistical characteristics. $\beta_i \geq 1$ decides the SNR of i-th signal $x_i(k)$. The smaller the β_i, the greater the SNR of $x_i(k)$. Not losing generality, we assume $\beta_{i_0} = 1$, $i_0 \in \{0,1,\cdots,N\}$, which implies the i_0-th channel signal has the greatest SNR.

Some other formations of the noisy speech signals can be converted to the formation shown in equation (4.1.4). For example, if noisy signals are

$$x_i(k) = \beta_i' s(k - \tau_i) + n_i'(k) \qquad i = 0,1,\cdots,N$$

Then

$$x_i'(k) = \frac{1}{\beta_i'} x(k + \tau_i)$$

$$= s(k) + \beta_i n_i(k)$$

where $\beta_i = \dfrac{1}{\beta_i'}$, $n_i(k) = n_i'(k + \tau_i)$.

Thus, from equation (4.1.4), the output of DAWS is

$$x(k) = \sum_{i=0}^{N} \alpha_i [s(k) + \beta_i n_i(k)]$$

$$= s(k) + \sum_{i=0}^{N} \alpha_i \beta_i n_i(k) \tag{4.3.5}$$

It's auto-correlation is

$$\phi_{xx}(k) = E[x(k)x^*(k)]$$

$$= E\left\{ \left[s(k) + \sum_{i=0}^{N} \alpha_i \beta_i n_i(k) \right] \cdot \left[s^*(k) + \sum_{i=0}^{N} \alpha_i \beta_i n_i^*(k) \right] \right\} \tag{4.3.6}$$

Assume the speech and noise are uncorrelated, by equation (4.3.6) we may get

$$\phi_{xx}(k) = \phi_{ss}(k) + \sum_{i=0}^{N} \alpha_i^2 \beta_i^2 \phi_{nn}(k) + \sum_{i=0}^{N} \sum_{\substack{j=0 \\ j \neq i}}^{N} \alpha_i \alpha_j \beta_i \beta_j real(\phi_{ij}(k)) \tag{4.3.7}$$

where $\phi_{nn}(k)$ is the auto-correlation of noise n_i, $\phi_{ij}(k)$ is the cross-correlation of n_i and n_j, $i, j = 0,1,\cdots,N$. Take the Fourier transform of equation (4.3.7) to get

$$\phi_{xx}(\omega) = \phi_{ss}(\omega) + \sum_{i=0}^{N} \alpha_i^2 \beta_i^2 \phi_{nn}(\omega) + \sum_{i=0}^{N} \sum_{\substack{j=0 \\ j \neq i}}^{N} \alpha_i \alpha_j \beta_i \beta_j real(\phi_{ij}(\omega))$$

$$= \phi_{ss}(\omega) + \sum_{i=0}^{N} \alpha_i^2 \beta_i^2 \phi_{nn}(\omega) + \phi_{nn}(\omega) \sum_{i=0}^{N} \sum_{\substack{j=0 \\ j \neq i}}^{N} \alpha_i \alpha_j \beta_i \beta_j real\left(\frac{\phi_{ij}(\omega)}{\sqrt{\phi_{ii}(\omega)\phi_{jj}(\omega)}}\right)$$

$$= \phi_{ss}(\omega) + \sum_{i=0}^{N} \alpha_i^2 \beta_i^2 \phi_{nn}(\omega) + \phi_{nn}(\omega) \sum_{i=0}^{N} \sum_{\substack{j=0 \\ j \neq i}}^{N} \alpha_i \alpha_j \beta_i \beta_j real(\Gamma_{ij}(\omega)) \tag{4.3.8}$$

So the power spectrum of the output noise

$$\phi_{nn}^{'}(\omega) = \sum_{i=0}^{N} \alpha_i^2 \beta_i^2 \phi_{nn}(\omega) + \phi_{nn}(\omega) \sum_{i=0}^{N} \sum_{\substack{j=0 \\ j \neq i}}^{N} \alpha_i \alpha_j \beta_i \beta_j real(\Gamma_{ij}(\omega))$$

If $n_i(k)$ and $n_j(k)$ are uncorrelated, i.e. $\Gamma_{ij}(\omega) = 0$, then

$$\phi_{nn}^{'}(\omega) = \sum_{i=0}^{N} \alpha_i^2 \beta_i^2 \phi_{nn}(\omega) \tag{4.3.9}$$

From this equation, we may conclude that the optimal weights of DAWS should be the optimal solution of the object function

$$f(\alpha_0, \alpha_1, \cdots, \alpha_N) = \sum_{i=0}^{N} \alpha_i^2 \beta_i^2 \tag{4.3.10}$$

with constraints being equation (4.3.3). To solve this optimal problem, we set up its Lagrange function

$$g(\alpha_1,\alpha_2,\cdots,\alpha_N,\lambda)=\sum_{i=0}^{N}\alpha_i^2\beta_i^2-\lambda(\sum_{i=0}^{N}\alpha_i-1) \tag{4.3.11}$$

Let its partial derivatives

$$\frac{\partial g}{\partial \alpha_i}=2\beta_i^2\alpha_i-\lambda=0 \qquad i=0,1,\cdots,N$$

$$\frac{\partial g}{\partial \lambda}=\sum_{i=0}^{N}\alpha_i-1=0$$

Solve these linear equations to get the optimal weights

$$\alpha_i=\left(\beta_i^2\sum_{j=0}^{N}\beta_j^{-2}\right)^{-1} \quad i=0,1,\cdots,N \tag{4.3.12}$$

Substitute equation (4.3.12) with (4.3.9) to get the minimum power spectrum of the DAWS output noise

$$\phi_{nn}^{'}(\omega)=\left(\sum_{i=0}^{N}\beta_i^{-2}\right)^{-1}\phi_{nn}(\omega) \tag{4.3.13}$$

Since $\beta_i\geq 1, i=0,1,\cdots,N$, $\beta_{i_0}=1, i_0\in\{0,1,\cdots,N\}$, we have $\phi_{nn}^{'}(\omega)\leq\phi_{nn}(\omega)$.

Let's simply note the i-th channel noise power spectrum as $\phi_i(\omega)$. Since $\beta_{i_0}=1$ and $\beta_i\geq 1$, the i_0-th channel noise has the minimum power spectrum $\phi_{i_0}(\omega)=\phi_{nn}(\omega)$. Obviously

$$\frac{\phi_i(\omega)}{\phi_{i_0}(\omega)}=\frac{\beta_i^2\phi_{nn}(\omega)}{\beta_{i_0}^2\phi_{nn}(\omega)}=\beta_i^2 \tag{4.3.14}$$

So, the parameters β_i^2 which decide the optimal weights of DAWS can be easily got by computing the ratio of the i-th channel noise power $\phi_i(\omega)$ over the i_0-th channel noise power.

To sum up, the weights and output of DAWS can be calculated as follows:

Step 1. Calculate the time delays of the speech signal in all channels, and align the noisy speech signals. That is, to get $x_i^{'}(k)=s(k)+n_i^{'}(k), i=0,1,\cdots,N$ ’

Step 2. For every aligned signal in step 1, take the same noise-only sections and calculate its power. That is to calculate $\phi_i=\frac{1}{K_2-K_1}\sum_{k=K_1+1}^{K_2}x_i^2(k)$, $i=0,1,\cdots,N$,

where from K_1+1 to K_2 every channel signal is a noise-only signal.

Step 3. Among the powers calculated in Step 2, find the minimum one. Then calculate the ratio of every sectional noise power over the minimum one, to get the parameters which will decide the optimal weights. That is, to find out $\phi_{i_0}=\min\{\phi_0,\phi_1,\cdots,\phi_N\}$, and then calculate $\beta_i^2=\dfrac{\phi_i}{\phi_{i_0}}$, $i=0,1,\cdots,N$.

Step 4. Calculate the optimal weights by equation (4.3.12) and the output of DAWS by (4.3.2). That is to compute

$$\alpha_i=\left(\beta_i^2\sum_{j=0}^{N}\beta_j^{-2}\right)^{-1},\ i=0,1,\cdots,N$$

and

$$x(k)=\sum_{i=0}^{N}\alpha_i x_i^{'}(k)$$

A simulation experiment and an experiment in the real environment are presented as follows.

Let desired signal

$$s(t)=\begin{cases}0 & 0\le t<\pi,3\pi\le t<4\pi\\ 3\sin(t) & \pi\le t<3\pi\end{cases}$$

Suppose the noisy signals acquired from a two-sensor array are

$$x_1(t)=s(t)+n_1(t)$$

$$x_2(t)=s(t)+3n_2(t)$$

where $n_1(t)$ and $n_2(t)$ are independent white noise in [-1,1]. The sampling time interval is 0.001π. For any signal $x(t)$, note its digitalized signal as $x(k)$.

The DAS output

$$y_{DSB}(k)=\frac{1}{2}\left[x_1(k)+x_2(k)\right]$$

For DAWS, we have $\beta_1=1$ and we may calculate β_2^2 by

$$\beta_2^2=\frac{\sum_{k=1}^{1000}x_2(k)}{\sum_{k=1}^{1000}x_1(k)}$$

So the output of DAWS is

$$y_{DAWS}(k) = \frac{\beta_2^2}{1+\beta_2^2} x_1(k) + \frac{1}{1+\beta_2^2} x_2(k)$$

After simulation by use of MATLAB, we get

$$SNR(y_{DAS}) = 4.37\,\text{dB}$$

$$SNR(y_{DAWS}) = 8.78\,\text{dB}$$

Figure 4.3.2 shows the signals with (a) $s(k)$, (b) $y_{DAS}(k)$ and (c) $y_{DAWS}(k)$.

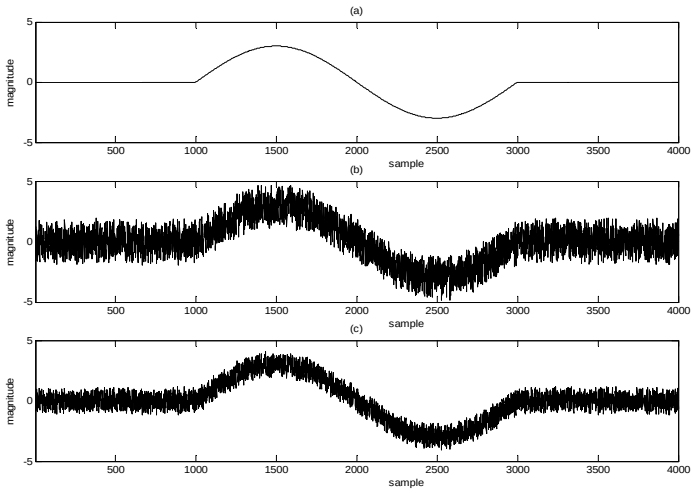

Figure 4.3.2 Simulation results
(a) Clean signal
(b) Enhanced signal by DAS
(c) Enhanced signal by DAWS

The experiment in the real environment was made in a common study room of dimensions 5x4x2.8m. Four microphones were used to construct a 3-dimensional array as shown in figure 3.2.2. The speech and the noises were generated concurrently by loudspeakers from different locations. As shown in figure 4.3.3, the speech loudspeaker was placed 30cm in front of the microphone array at (0,30). The microphone array was placed with M_1 facing the speech loudspeaker directly. The noise loudspeaker emitted white noise at (100,50). The speech data was from a section of speech recorded on a computer and the noise data was from White noise in database NoiseX92. The sampling rate used to digitize the acquired signals was 8 kHz.

Figure 4.3.4 shows the comparative results. (a) shows the noisy speech acquired

from microphone M_1, with $SNR_{original}$=11.16 dB. (b) shows the enhanced speech by DAS which scores a SNR_{DAS}=12.18 dB. (c) shows the enhanced speech by proposed DAWS which scores a SNR_{DAWS} =13.47 dB. So, DAWS algorithm gets higher SNR improvement than DAS. However, from the figure it is hard to find the improvement because of only about 1 dB in difference.

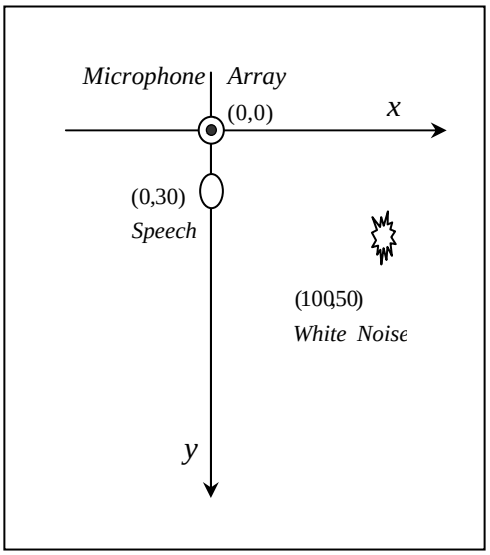

Figure 4.3.3 Experiment environment

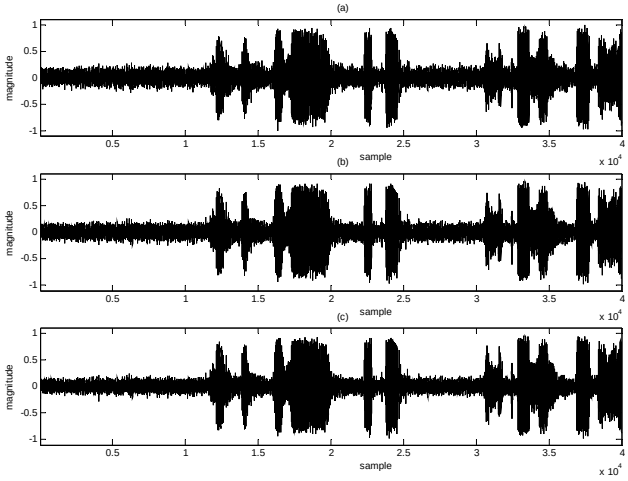

Figure 4.3.4 Speech enhancement results
(a) Noisy speech acquired by main microphone
(b) Enhanced speech by DAS
(c) Enhanced speech by DAWS

4.3.3 Partial-channel MCRANC

If the array contained quite a lot of microphones and we used all of the acquired signals in MCRANC, there would be too many coefficients in its filter A. This may cause great difficulty in finding the optimal coefficients. Although, theoretically speaking, more coefficients may lead to better optimal value of the objective function, it must be based on the fact that there are plenty of pure noise samples and the digital computer has enough computing accuracy. However, this is generally impossible in a practical situation. Too many coefficients will actually make the optimal solution further from the real optimal solution. So, if we use too many channel signals in MCRANC for noise cancellation, the speech enhancement effect might be degraded. In this situation only partial channels should be employed in MCRANC. This is called the Partial-channel MCRANC (P-MCRANC).

The structure of P-MCRANC is shown in figure 4.3.6, where $x_{i_1}, x_{i_2}, \cdots, x_{i_M} \in \{x_1, x_2, \cdots, x_N\}$, $M \le N$.

The number of the channels and which channel signals should be employed should be decided according to the microphone array structure and the practical problem. Generally, the number of channels employed by MCRANC should not be over five.

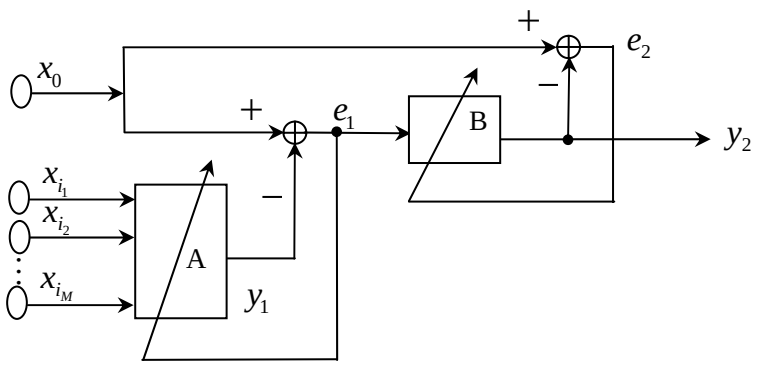

Figure 4.3.5 Structure of P-MCRANC

Similarly, with the improved MCRANC employing multichannel distorted signal

filtering (see section 4.3.1 of this chapter), there is no need to input all channel distorted speech signals into its second filter B. The proper way is to input only partial channels of those signals. Figure 4.3.6 shows the proper structure, where $e_{j_p 1}$ is the distorted speech signal in a common MCRANC, employing the j_p-th channel microphone signal as the main channel signal and other selected signals as referential signals, $j_1, j_2, \cdots, j_P \in \{i_1, i_2, \cdots, i_M\}$, $p = 1,2,\cdots,P$, $P \le M$.

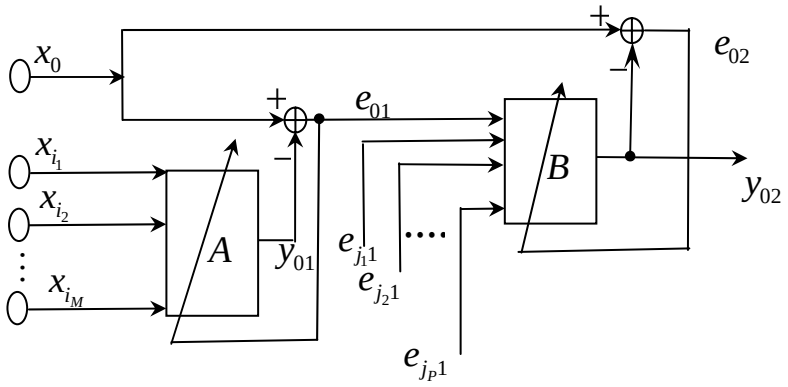

Figure 4.3.6 Structure of P-MCRANC employing partial-channel distorted signals

4.3.4 Fixed beamforming partial-channel MCRANC

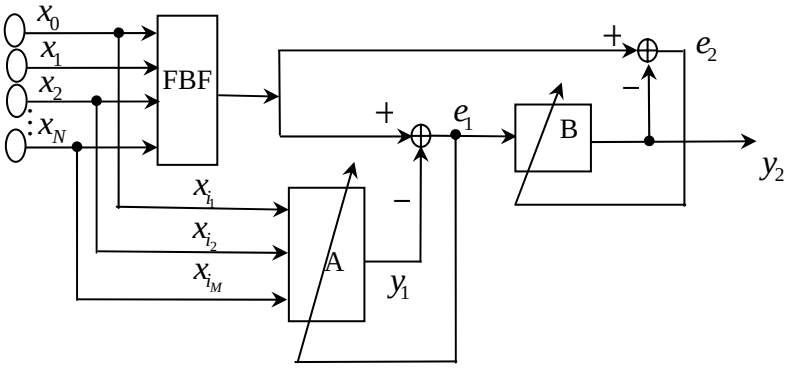

Figure 4.3.7 Structure of FBF-P-MCRANC

The fixed beamforming MCRANC (FBF-MCRANC) and partial-channel MCRANC (P-MCRANC) proposed above both have advantages over common

MCRANC. One way to combine both advantages of FBF-MCRANC and P-MCRANC is the fixed beamforming partial-channel MCRANC (FBF-P-MCRANC). Figure 4.3.7 indicates its structure, where $x_{i_1}, x_{i_2}, \cdots x_{i_M} \in \{x_0, x_1, \cdots x_N\}$, $M \le N+1$. FBF may use the method proposed in section 4.3.1 and section 4.3.2, including the DAWSAS algorithm.

4.3.5 Experimental results

In the experiment seven small microphones were used to construct a planar array with an aperture of about 7cm as shown in figure 4.3.8. The speech and the noises were generated concurrently by loudspeakers from different locations. The speech data was from a section of recorded speech in the computer and the noise data was from the NoiseX92 database. The sampling rate used to digitize the acquired signals is 24K Hz.

The experiment was made in a common study room of dimensions 5x4x2.8m. The array was put on a desk. The center of the array was 1.4m from the front wall, 1.8m from the left wall and 1.23m from the floor. There were two sofas, a cabinet and another two desks in the room. The room had two glass windows and a wooden door, and all of them were closed.

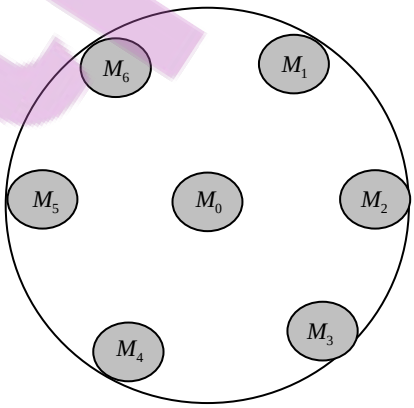

Figure 4.3.8 Planar array with seven microphones

One of the experiment cases is shown as figure 4.3.9. For simplicity, the figure is a

planar one since the loudspeakers emitting speech and noises have almost the same height as the array in the experiment. In this case, the speech loudspeaker was placed 30cm in front of the microphone array at (0,30). Noise loudspeakers concurrently emitted Volvo, Leopard, Factory2 and White noises. They were positioned at (-100,100), (50,50), (200,250) and (0,100)cm respectively. The following cases were considered.

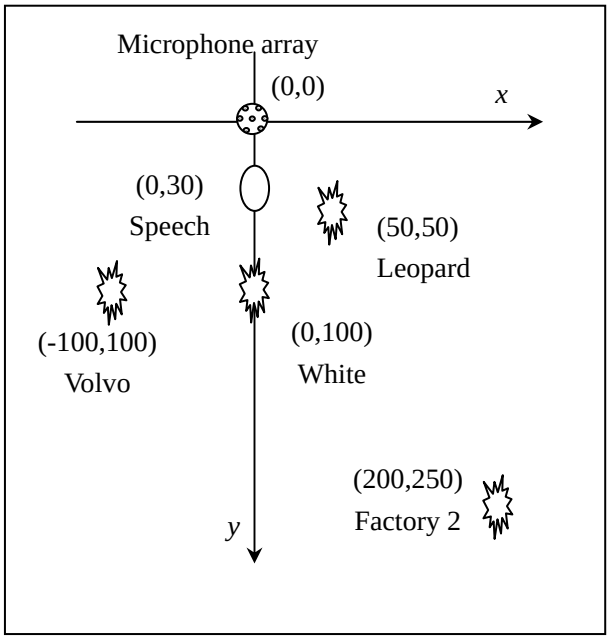

Figure 4.3.9 An experiment environment

Case 1. Speech at (0,30) and very loud White noise at (0,100).

Case 2. Speech at (50,50) and Factory2 noise at (200,250).

Case 3. Speech at (0,30), very loud White noise at (0,100) and Factory2 noise at (200,250).

Case 4. Speech at (0,30), Leopard noise at (50,50), Volvo noise at (-100,100) and White noise at (0,100)

Case 5. Speech at (250,250), Volvo noise at (-100,100).

Case 6. Loud speech at (0,30), Volvo noise at (-100,100), Leopard noise at (50,50), White noise at (0,100) and Factory2 noise at (200,250).

We use the DAWSAS algorithm for the fixed beamformer and a cross-correlation

method to calculate the time delays. For the P-MCRANC, four channels of the signals from microphones M_o, M_2, M_4, M_6 are selected. The lengths of filter A and B in P-MCRANC are 4x32=128 and 48 respectively. Both adaptive filters employ the LMS algorithm with learning rate $\mu = 0.01$.

Table 4.3.1 shows the SNRs and SNR improvements of the original and enhanced speeches by use of different algorithms including the MGSC, MCRANC and the proposed FBF-P-MCRANC. The last two rows are the average SNRs and average SNR improvements. The original signal in the table is the signal $x_0(k)$ from microphone M_0. In this experiment, SNR is calculated by formula (1.4.3).

Table 4.3.1 The SNRs (dB) of original noisy speech and the enhanced speech by MGSC, MCRANC and proposed FBF-P-MCRANC

Case	Original	MGSC	MCRANC	FBF-P-MCRANC
1	6.08	6.42	17.94	22.52
2	3.78	4.93	22.70	24.95
3	7.86	5.78	19.12	22.09
4	1.74	6.20	13.97	27.05
5	7.38	12.15	23.23	30.68
6	2.52	5.23	26.86	29.87
Average	4.89	6.79	20.64	26.19
Improved	0	1.90	15.74	21.30

Figure.4.3.10 shows a comparative system performance under the Case 6 scenario.

Figure.4.3.10 (a) shows the noisy signal $x_0(k)$ acquired by microphone M_0. The acquired speech signal is seriously contaminated with noises at SNR=2.25 dB. The signals acquired by the other microphones are very similar to $x_0(k)$.

Figure.4.3.10 (b) shows the enhanced speech by MGSC algorithm which scores SNR=5.23

dB.

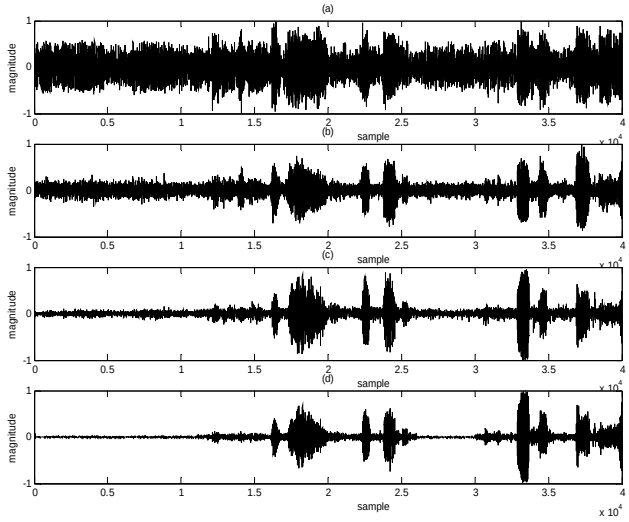

Figure 4.3.10 Speech enhancement results
(a) Noisy speech
(b) Enhanced speech by MGSC
(c) Enhanced speech by MCRANC
(d) Enhanced signal by FBF-P-MCRANC

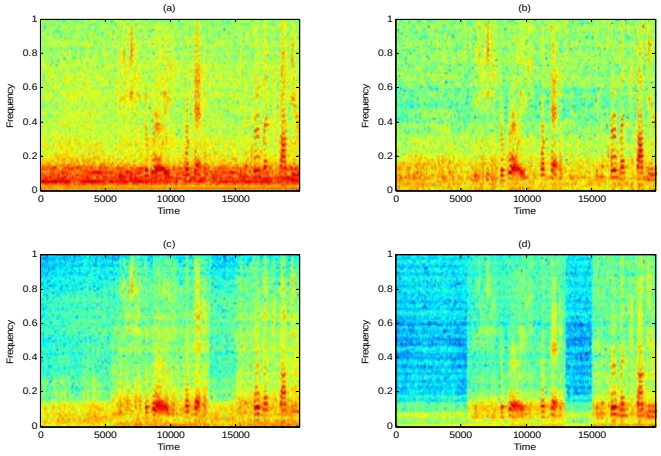

Figure 4.3.11 Spectrograms of the signals in figure.4.3.10
(a) Spectrogram of noisy speech
(b) Spectrogram of enhanced speech by MGSC
(c) Spectrogram of enhanced speech by MCRANC
(d) Spectrogram of enhanced signal by FBF-P-MCRANC

Figure.4.3.10 (c) shows the enhanced speech by using the MCRANC algorithm which scores SNR =26.86 dB.

Figure.4.3.10 (d) is the enhanced speech by the proposed FBF-P-MCRANC

which achieves SNR=29.87 dB.

Figure.4.3.11 shows the spectrograms of the signals in figure.4.3.10.

From table 4.3.1 and figures 4.3.10~11, we can easily find that FBF-P-MCRANC performs better than MCRANC.

4.3.6 Conclusions

Fixed Beamforming Partial-channel MCRANC (FBF-P-MCRANC) is proposed. It may make improvement to MCRANC and broaden its application areas.

Meanwhile, the Delay And Weighted Sum and Selection (DAWSAS) algorithm is presented for the fixed beamformer in FBF-P-MCRANC.

Experimental results indicate the effectiveness of DAWSAS and the SNR improvement by use of the proposed FBF-P-MCRANC.

4.4 Subband MCRANC

4.4.1 Subband MCRANC

A subband system [33, 41, 42, 72, 92, 3] decomposes the wideband input signals into a number of band-limited signals, superficially similar to the treatment the human ear performs on incoming signals. A significant advantage of using subband processing for speech enhancement is that it allows different processing in each subband depending on factors such as signal power, noise power and correlation levels between signals and noises. For instance, if a particular band contains no noise energy, this band could be simply passed through since any processing would actually degrade the speech signal unnecessarily. In addition, the implementation of a classical adaptive noise cancellation

scheme in a number of frequency-limited subbands permits faster convergence of the filter coefficients due to the reduction of signal power and adaptive filter length in each subband.

The subband speech enhancement system is shown in figure 4.4.1.

In figure 4.4.1, every microphone array signal x_i is divided by analysis filter bank into J subband signals $\{x_i^{(j)}, j=1,2,\cdots,J\}$, $i=0,1,,\cdots,N$. Then all the signals in the j-th subband $\{x_i^{(j)}, i=0,1,2,\cdots,N\}$ are used to cause speech enhancement to get $y^{(j)}$, ($j=1,2,,\cdots,J$). Finally, all subband enhanced speech $\{y^{(j)}, j=1,2,\cdots,J\}$ is synthesized by a synthesis filter to form the full-band enhanced speech.

If the speech enhancement method in j-th subband SE(j) is MCRANC ($j=1,2,,\cdots,J$), we call the system as Subband MCRANC.

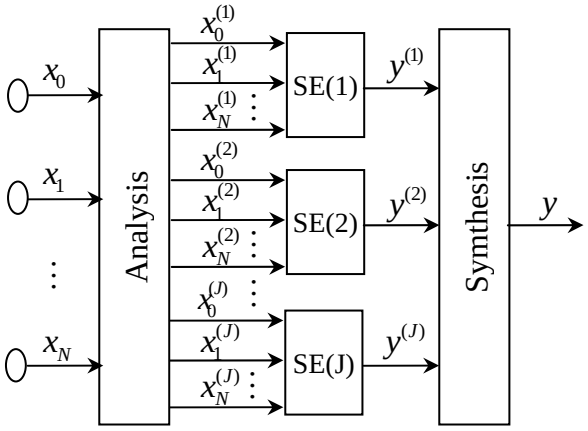

Figure 4.4.1 The structure of subband speech enhancement with microphone array

If the speech enhancement method in j-th subband SE(j) is MDS-MCRANC ($j=1,2,,\cdots,J$), we call the system Subband MDS-MCRANC.

If the speech enhancement method in j-th subband SE(j) is MSR-MCRANC ($j=1,2,,\cdots,J$), we call the system Subband MSR-MCRANC.

If the speech enhancement method in j-th subband SE(j) is FBF-P-MCRANC ($j=1,2,,\cdots,J$), we call the system Subband FBF-P-MCRANC.

Here we only describe subband FBF-P-MCRANC as follows.

4.4.2 Subband FBF-P-MCRANC

The subband FBF-P-MCRANC system is shown in figure 4.4.2. The structure of j-th subsystem FBF-P-MCRANC(j) is indicated in figure 4.4.3. Different subsystems may employ the same FBF-P-MCRANC or different FBF-P-MCRANC.

The following is some experimental results. The setup of the experiment was exactly the same as the experiment in section 4.3. Table 4.4.1 is the extension of table 4.3.1 for it contains one more column to present the SNR results of the enhanced speech by subband FBF-P-MCRANC. It can be found that the results by subband FBF-P-MCRANC are further improved.

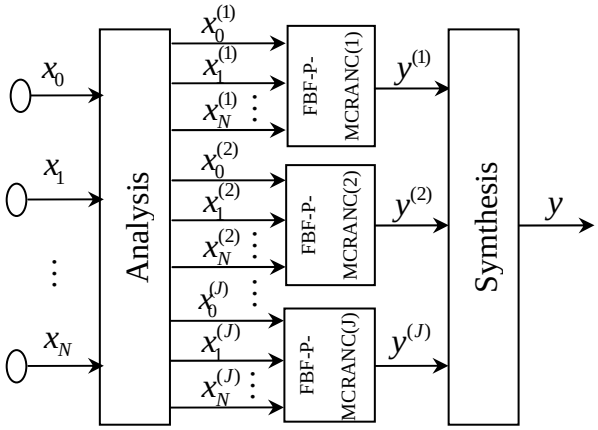

Figure 4.4.2 The structure of subband FBF-P-MCRANC

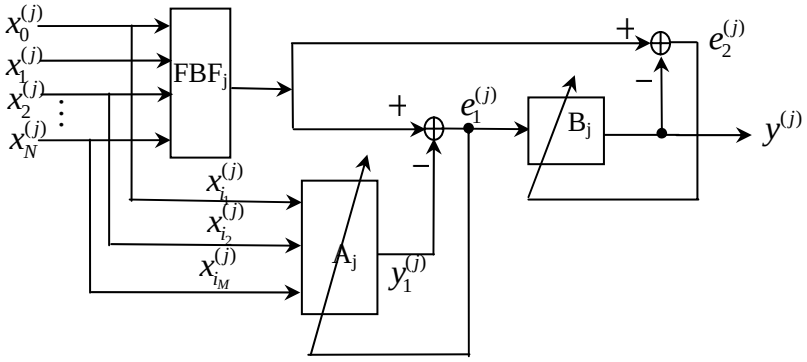

Figure 4.4.3 The structure of subsystem FBF-P-MCRANC(j)

Table 4.4.1 The SNRs (dB) of the noisy speech and the enhanced speech using MGSC, MCRANC, FBF-P-MCRANC and subband FBF-P-MCRANC

Case	Noisy speech	MGSC	MCRANC	FBF-P-MCRANC	Subband FBF -P-MCRANC
1	6.08	6.42	17.94	22.52	23.88
2	3.78	4.93	22.70	24.95	26.53
3	7.86	5.78	19.12	22.09	23.89
4	1.74	6.20	13.97	27.05	27.87
5	7.38	12.15	23.23	30.68	31.56
6	2.52	5.23	26.86	29.87	31.03
Average	4.89	6.79	20.64	26.19	27.46
Improved	0	1.90	15.74	21.30	22.57

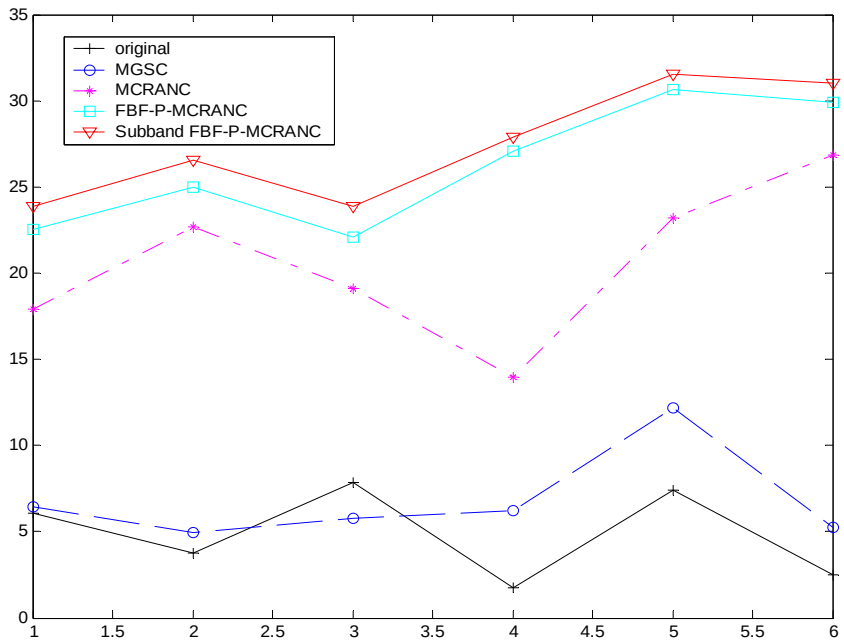

Figure 4.4.4 SNR lines of the noisy speech and the enhanced speech using four different algorithms in six cases

In the subband processing, the full band [300, 4000] Hz is equally devided into 8

subbands $B_1, B_2, \cdots, B_8$. In every subband, FBF-P-MCRANC is employed for subband speech enhancement. All FBF-P-MCRANC employ DAWSAS algorithm for fixed beamforming. Microphone M_0 is used as the reference one for time aligning and the GCC algorithm is employed for finding the time delays. In P-MCRANC, only the signals from microphones M_o, M_2, M_4, M_6 are employed for filter A and the length of FIR filters A and B are 24X4=96 and 32 respectively, and the LMS algorithm is employed for the adaptations of the filters with learning rate $\mu = 0.01$. Multiple sampling rates method is employed for the two highest frequency banks and the two lowest frequency banks. For the two highest frequency banks B_8 and B_7, the sampling rate for referential signals is 24K Hz while the rate for main signal is 8 kHz. For the two lowest frequency banks B_1 and B_2, the sampling rates for referential signals is 4K Hz while the rate for main signal is 8 kHz.

Figure 4.4.4 shows the SNR results of the noisy speech and the enhanced speech using four different methods in six cases. Each line indicates the results of a method. It can be seen that subband FBF-P-MCRANC has the best SNR improvement among them.

4.5 Summary

This chapter presents four methods to improve MCRANC itself. They improve MCRANC from different ways, and all of them may be employed together to make MCRANC more powerful.

Section 4.1 presents the first method, using multichannel distorted signals for the second filter of MCRANC. The distorted speech signals are obtained by taking different channels of signals as the main channel signal and others as referential signals to process the first stage of a common MCRANC. The MDS-MCRANC is useful if the

spatial correlation of the speech signals is not very big. However, the further improvement from the MCRANC is usually minor. In this section, the combinational structure of MDS-MCRANC with DAS is also presented.

In Section 4.2 the second method, using different sampling rates for the main signal and referential signals in MCRANC, is proposed. A higher or lower sampling rate should be used for referential channel signals, according to the types of the noises, while the required sampling rate for output speech is used for the main channel signal. An experiment with two microphones in the array proved the effectiveness of the MSR-MCRANC. The method mentioned in this section has close relations with the subband MCRANC in section 4.4. However, the method in this section does not need to divide the signals into different bands. Therefore, no distortion happens to any signal caused by the subband analyzing and synthesizing filters.

The third method is to add fixed beamformer and partial-channel technique to MCRANC. This addition results in a fixed beamforming partial-channel MCRANC (FBF-P-MCRANC) algorithm, which is detailed in section 4.3. To improve the DAS fixed beamforming, a DAWSAS algorithm is also presented. The third method may broaden the applications of MCRANC in different noise environments. Experiments verified its improvements to MCRANC.

The fourth method presented in section 4.4 is to employ subband processig for MCRANC. It may provide us with flexible treatments for speech enhancement. The structure of subband FBF-P-MCRANC is presented. Experimental results show further SNR improvement by use of this subband method.

Chapter 5 Improved MGSC Algorithms

Two improved algorithms based on Modified Generalized Sidelobe Canceling (MGSC) are proposed in this chapter. One algorithm is to employ Multichannel Crosstalk Resistant Adaptive Signal Cancellation (MCRASC) for the signal blocking module of MGSC to get better performances of speech enhancement. The other is to use a distorted desired signal for the desired signal cancellation in the blocking module. Subband processing of improved MGSC is also proposed in this chapter. Experiments are presented to show the improvements to speech enhancement gained from these proposed algorithms.

5.1 Introduction

Generalized Sidelobe Canceling (GSC) is a widely used algorithm for signal enhancement [45]. The main drawback of conventional GSC is the imperfection of the signal blocking module. GSC cannot efficiently estimate noise through this module, which leads to partial cancellation of the desired signal from the GSC output. Some improved GSC algorithms have been proposed [37, 38, 44, 72, 122, 53]. Among them, Modified GSC (MGSC), which employs a Voice Activity Detector (VAD), is very powerful and very useful for speech enhancement [44, 72]. It adapts the coefficients of the adaptive filter for noise cancellation only during pure noise periods and keeps the coefficients fixed during speech periods, making the speech cancellation partially avoided.

Based on MGSC, two improved algorithms are proposed in this chapter by introducing MCRANC to the blocking module of the MGSC. One algorithm employs MCRANC to block the desired signal for the blocking module of MGSC. A so-called

Multichannel Crosstalk Resistant Signal Cancellation (MCRASC) based MGSC (MCRASC-MGSC) is proposed. Another algorithm is actually a simplified MCRASC-MGSC. It uses a shared distorted signal for signal blocking, and it is named Shared Distorted Signal MGSC (SDS-MGSC). Theoretic analysis indicates that the essence of two proposed algorithms is to extend the blocking matrix in MGSC from a common matrix to a time-variable vector matrix. Experimental results show the advantages of the proposed algorithms for speech enhancement with small microphone array.

5.2 Brief introduction of GSC and MGSC

5.2.1 Generalized sidelobe canceling

GSC was first introduced by Griffiths and Jim in 1982 [45]. Its structure is shown in figure 5.2.1. FBF is a fixed beamformer where DAS (Delay And Sum) is the widely used algorithm for FBF; BM represents a blocking module or a blocking matrix which blocks the desired signal and passes through the noise; MANC is a multichannel adaptive noise canceller.

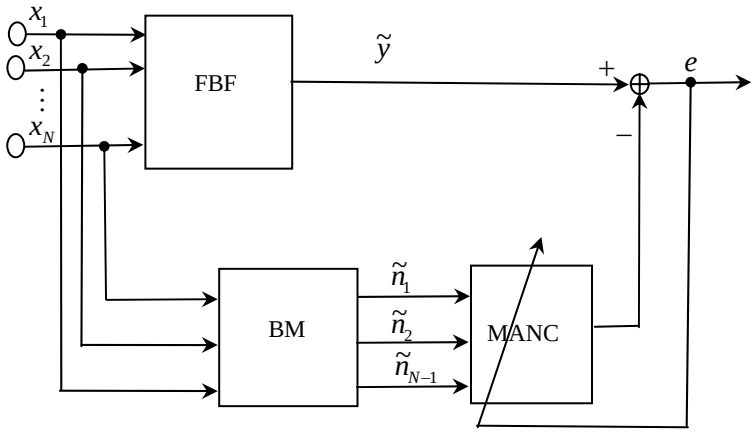

Figure 5.2.1 Structure of GSC

In module BM, the array signals are time-aligned first. Then the aligned signals are processed through a blocking matrix. Different types of matrix B have been proposed. The simplest one is

$$\mathbf{B}=\begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$$

In fact, if a matrix satisfies

$$\sum_{j=1}^{N} b_{ij}=0 \qquad i=1,2,\cdots,N-1 \tag{5.2.1}$$

then B can be used to block the desired signal.

In fact, the above-mentioned selection for matrix B is based on the assumption that the time-aligned signals from the array signals $x_1, x_2, \cdots, x_N$ have the following formation

$$\hat{x}_i(k)=\hat{s}(k)+\hat{n}_i(k) \qquad i=1,2,\cdots,N \tag{5.2.2}$$

where $\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_N$ are the time-aligned signals, $\hat{s}(k)$ is the desired signal and $\hat{n}_i(k)$ is the noise. Under this assumption, the output of the BM module is

$$\tilde{\mathbf{n}}(k)=\mathbf{B}\hat{\mathbf{x}}(k)$$

where $\tilde{\mathbf{n}}(k)=[\tilde{n}_1(k),\tilde{n}_2(k),\cdots,\tilde{n}_{N-1}(k)]^T$, $\hat{\mathbf{x}}(k)=[\hat{x}_1(k),\hat{x}_2(k),\cdots,\hat{x}_{N-1}(k)]^T$. Therefore, by equation (5.2.1) we have

$$\begin{aligned}\tilde{n}_i(k) &= \sum_{j=1}^{N} b_{ij}[\hat{s}(k)+\hat{n}_j(k)] \\ &= \sum_{j=1}^{N} b_{ij}\hat{n}_j(k) \qquad i=1,2,\cdots,N-1 \end{aligned} \tag{5.2.3}$$

So $\tilde{n}_i(k)$ contains only noise components and thus will not cause any desired signal cancellation in the MANC processing.

However, practically we are usually unable to obtain the ideal formation of equation (5.2.2). Instead, we obtain a formation such that

$$\hat{x}_i(k) = \hat{s}_i(k) + \hat{n}_i(k) \qquad i = 1,2,\cdots,N \tag{5.2.4}$$

where $s_i(k) \neq s_j(k)$, if $i \neq j$. As a result, we have

$$\begin{aligned} \tilde{n}_i(k) &= \sum_{j=1}^{N} b_{ij}[\hat{s}_j(k) + \hat{n}_j(k)] \\ &= \sum_{j=1}^{N} b_{ij}\hat{s}_j(k) + \sum_{j=1}^{N} b_{ij}\hat{n}_j(k) \qquad i = 1,2,\cdots,N-1 \end{aligned} \tag{5.2.5}$$

Here, $\tilde{n}_i(k)$ contains desired signal component $\sum_{j=1}^{N} b_{ij}\hat{s}_j(k)$. This component is called the leakage of the desired signal into $\tilde{n}_i(k)$ and it causes cancellation of the desired signal in MANC processing. If only matrix B can reduce the desired signal component and/or increase the noise component in (5.2.5), a better output of GSC can be expected.

When GSC is applied to speech enhancement, the speech leakage in the BM module and thus the speech cancellation in GSC's output usually causes serious problem, especially when the SNR of the noisy speech is high. Sometimes the enhanced speech even becomes worse than the original noisy speech.

5.2.2 Modified generalized sidelobe canceling

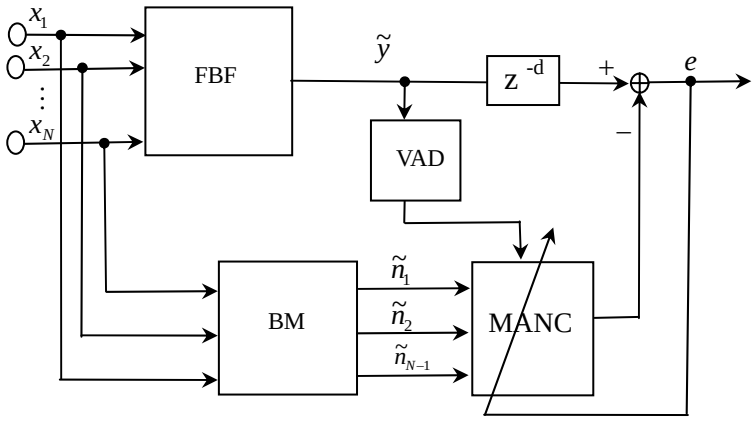

Figure 5.2.2 Structure of MGSC

Figure 5.2.2 shows a MGSC for speech enhancement. It employs a VAD to control

the adaptation of the MANC filter. The coefficients of the MANC filter are adapted only during NSP (Non Speech Period) and are fixed during HSP (Having Speech Period). By this way, the speech cancellation in the final output of GSC will usually be alleviated, making the enhanced speech a higher SNR [44, 72].

In this MGSC, a delay module is also added to ensure the noise cancellation process by MANC is causal. However, according to the research work of Greensberger [44], the delay time should be selected to be shorter than the time needed for the first reflection of the speech signal.

5.3 Proposed MCRASC-MGSC

5.3.1 Description of the algorithm

The structure of the proposed MGSC is shown in figure 5.3.1. It is based on Multichannel Crosstalk Resistant Adaptive Signal Cancellation (MCRASC). MCRASC has close relation with MCRANC and will be presented in subsection 5.3.2. The proposed MCRASC based MGSC (MCRASC-MGSC) has a similar structure as the MGSC. The only difference between them is in the BM module. It employs MCRASC, rather than a common matrix, for the desired signal blocking. The output of the BM module has N channels, rather than N-1 channels, of estimated noises.

In figure 5.3.1, VAD is not only used to control the noise cancellation process in MANC but also used to control the MCRASC process in the BM module.

It will be proved later in this section that the essence of MCRASC is to replace the common matrix with a time-variable vector matrix, in which all elements are real vectors rather than real numbers.

In fact, except for the signal leakage in BM module, the destruction of the noise signals by the blocking matrix also has an important influence on the MGSC output. The common blocking matrix B may reduce the correlation between the noises in the upper

path (output of FBF) and lower path (outputs of BM). This may cause the reduction of the noise cancellation in the final MGSC output.

MCRASC can more effectively retain noise correlation and block speech than the common matrix, ensuring speech enhancement better performances.

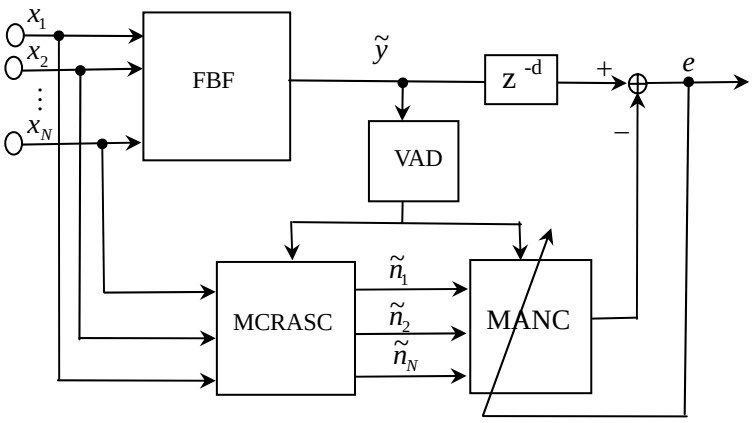

Figure 5.3.1 Structure of MCRASC based MGSC

5.3.2 MCRASC module

MCRASC is actually a MCRANC with different output. It can be easily realized by MCRANC.

As can be seen from section 2.4 of chapter 2 or section 3.3 of chapter 3, if we use signal x_i in $x_1, x_2, \cdots, x_N$ as the main channel signal and all others as the referential channel signals, the output y_{i2} of filter B_i, as shown in figure 5.3.2, would be the estimation $\tilde{s}_i$ of speech signal s_i in x_i, and thus the output of the system e_{i2} would be the estimation $\tilde{n}_i$ of noise signal n_i in x_i, $i = 1,2,\cdots,N$. The structure as shown in figure 5.3.2 is called MCRASC.

Like MCRANC, MCRASC needs a VAD.

The MCRASC module in figure 5.3.1 uses every signal as the main signal and others as referential signals to perform a MCRACSC in order to get noise estimation.

Thus MCRASC would output N channel estimated noises.

To any main signal x_i, we may use only some of the referential signals $x_{i_1}, x_{i_2} \cdots, x_{i_M}$ in a MCRASC, where $x_{i_m} \in \{x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_N\}$, $m = 1,2,\cdots,M$, $M \le N-1$. This usage is similar to partial-channel MCRANC in chapter 4. As to how many partial channels and which channels of the referential signals should be selected, it may be decided by experiment according to the microphone array structure and the application. Generally speaking, there are no strict rules to decide that.

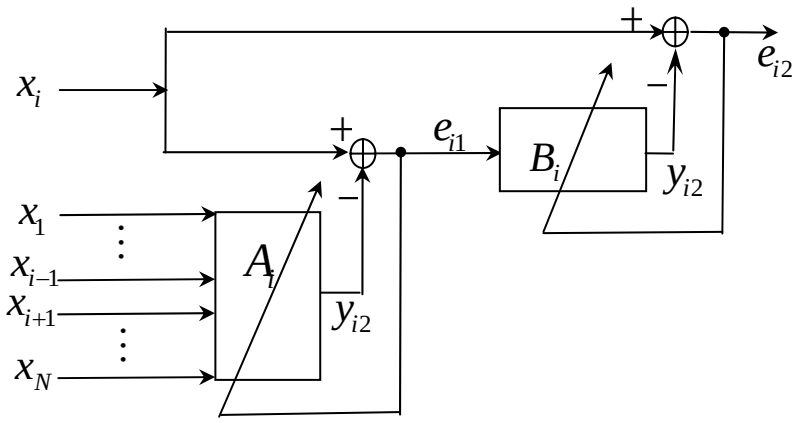

Figure 5.3.2 Structure of MCRASC

In the blocking module we can also employ improved MCRASC to get every channel of estimated noise. The improved MCRASC algorithms are similar to the improved MCRANC algorithms presented in chapter 4.

5.3.3 Vector blocking matrix

This subsection maintains that the essence of the proposed MCRASC-MGSC is to extend the common blocking matrix in MGSC to a vector-valued matrix, in which every element of the matrix can be a vector. If the vectors in the matrix are properly decided, the blocking module can block the desired signal and pass the noise more effectively.

For simplicity of the notations, in this subsection, we still denote the time-aligned array signals as $x_1(k), x_2(k), \cdots, x_N(k)$, and note the z-transform of these aligned

signals as $X_1(z), X_2(z), \cdots, X_N(z)$.

Suppose the employed MCRASC for the BM module is a common MCRASC algorithm as shown in figure 5.3.2. Denote the transfer functions of filters A_i and B_i are $\mathbf{H}_i(z)$ and $G_i(z)$ respectively. Then we have

$$\mathbf{H}_i(z) = [H_{i1}(z) \quad \cdots \quad H_{i(i-1)}(z) \quad H_{i(i+1)}(z) \quad \cdots \quad H_{iN}(z)] \tag{5.3.1}$$

where

$$H_{ij}(z) = \sum_{l=0}^{L_1} w_{ijl} z^{-l} \tag{5.3.2}$$

$$G_i(z) = \sum_{l=0}^{L_2} u_{il} z^{-l} \tag{5.3.3}$$

where w_{ijl} and u_{il} are coefficients of filter A_i and B_i respectively, L_1 is the number of sample delay for every referential signal input to filter A_i and so the length of filter A_i is $(N-1)(L_1+1)$, and (L_2+1) is the length of filter B_i.

The z-transform of $e_{i2}(k) = \tilde{n}_i(k)$ is

$$\begin{aligned} E_{i2}(z) = \tilde{N}_i(z) &= X_i(z) - G_i(z)E_{i1}(z) \\ &= X_i(z) - G_i(z)\mathbf{H}_i(z)\mathbf{X}'(z) \end{aligned} \tag{5.3.4}$$

where $\mathbf{X}'(z) = [X_1(z) \quad \cdots \quad X_{i-1}(z) \quad X_{i+1}(z) \quad \cdots \quad X_N(z)]^T$. So

$$\begin{aligned} \tilde{N}_i(z) = [&-G_i(z)H_{i1}(z) \quad \cdots \quad -G_i(z)H_{i(i-1)}(z) \quad 1 \quad -G_i(z)H_{i(i+1)}(z) \quad \cdots \quad -G_i(z)H_{iN}(z)] \cdot \\ &[X_1(z) \quad \cdots \quad X_{i-1}(z) \quad X_i(z) \quad X_{i+1}(z) \quad \cdots \quad X_N(z)]^T \end{aligned} \tag{5.3.5}$$

Thus, we have

$$\tilde{\mathbf{N}}(z) = \mathbf{B}(z)\mathbf{X}(z) \tag{5.3.6}$$

where

$$\tilde{\mathbf{N}}(z) = [\tilde{N}_1(z) \quad \tilde{N}_2(z) \quad \cdots \quad \tilde{N}_N(z)]^T \tag{5.3.7}$$

$$\mathbf{X}(z) = [X_1(z) \quad X_2(z) \quad \cdots \quad X_N(z)]^T \tag{5.3.8}$$

So, in the blocking matrix $\mathbf{B}(z)=[B_{ij}(z)]_{N\times N}$ we have

$$B_{ij}(z)=\begin{cases}-G_i(z)H_{ij}(z) & i\neq j\\ 1 & i=j\end{cases} \tag{5.3.9}$$

Take the inverse z-transform of $B_{ij}(z)$ to get the time domain element $\mathbf{b}_{ij}$ in the matrix. From equation (5.3.2) and (5.3.3), we may conclude that the length of $\mathbf{b}_{ij}$ is (L_1+L_2+1). By (5.3.5) and the length of $\mathbf{b}_{ij}$, we may deduce that for every signal $\{x_i(k)\}$, its samples $x_i(k),x_i(k-1),\cdots,x_i(k-L_1-L_2-1)$ are used to get $\tilde{n}_i(k)$ in the output $\{\tilde{n}_i(k)\}$ of the BM module, $i=1,2,\cdots,N$.

Since MCRASC can adapt to the change of the environment by adjusting the filters' coefficients, the vector elements in the blocking matrix may change with the adaptation of the filters. So, the vector matrix in the MCRASC module is actually time-variable.

Similarly, we may deduce the vector matrix for partial-channel MCRASC blocking module.

5.4 Proposed SDS-MGSC

5.4.1 Description of the algorithm

The main idea of the MCRASC based MGSC proposed in the previous section is to employ MCRASC to get N channels of the noise estimations in the blocking module of MGSC. Every channel of the noise estimation needs a MCRASC subsystem. And every subsystem contains two adaptive filters. So 2N adaptive filters are used in the blocking module.

However, the MCRASC subsystems have actually close relationships with each

other. In speech enhancement, every e_{i1} in MCRASC subsystems can be regarded as a distorted speech signal and it can be used to cancel the speech component in any noisy speech signal to get a channel of noise estimation. This is feasible especially under the small array circumstance, because in these circumstances the correlation between two speech signals is high. So, in the blocking module all MCRASC subsystems may share a distorted speech signal provided by any MCRASC subsystem. In this way we may employ only N+1 adaptive filters in the blocking module. We called this simplified MCRASC based MGSC as Shared Distorted Signal MGSC (SDS-MGSC). Its structure is indicated in figure 5.4.1. In the figure, the shared distorted signal is provided by the first MCRASC subsystem.

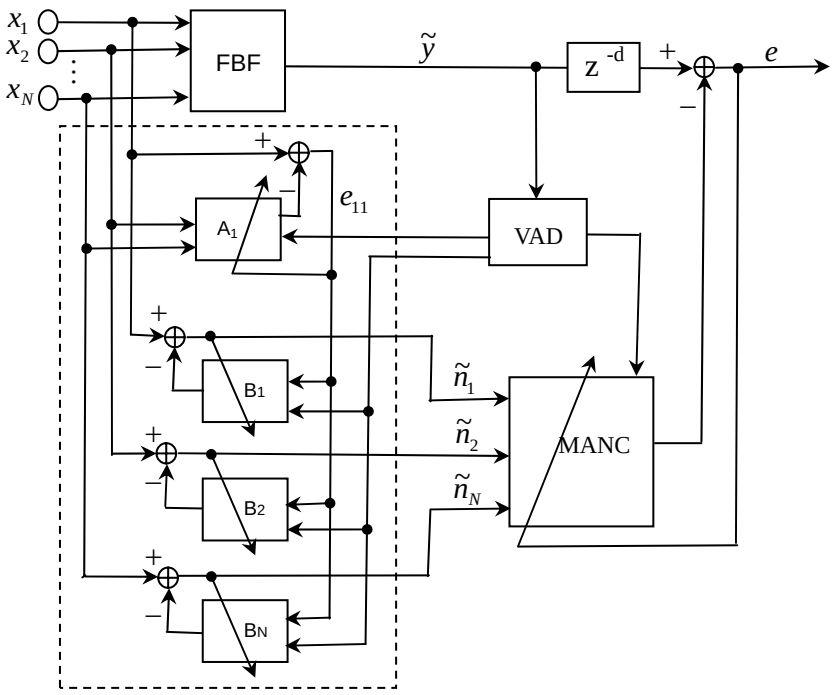

Figure 5.4.1 Structure of shared distorted signal MGSC (SDS-MGSC)

5.4.2 Improved SDS-MGSC

An improvement to SDS-MGSC is to provide the shared distorted signal by using

the output $\tilde{y}$ of the fixed beamformer FBF as the main channel signal and array signals $x_1, x_2, \cdots, x_N$ as the referential signals. Its structure is indicated in figure 5.4.2, and is called Improved SDS-MGSC (ISDS-MGSC).

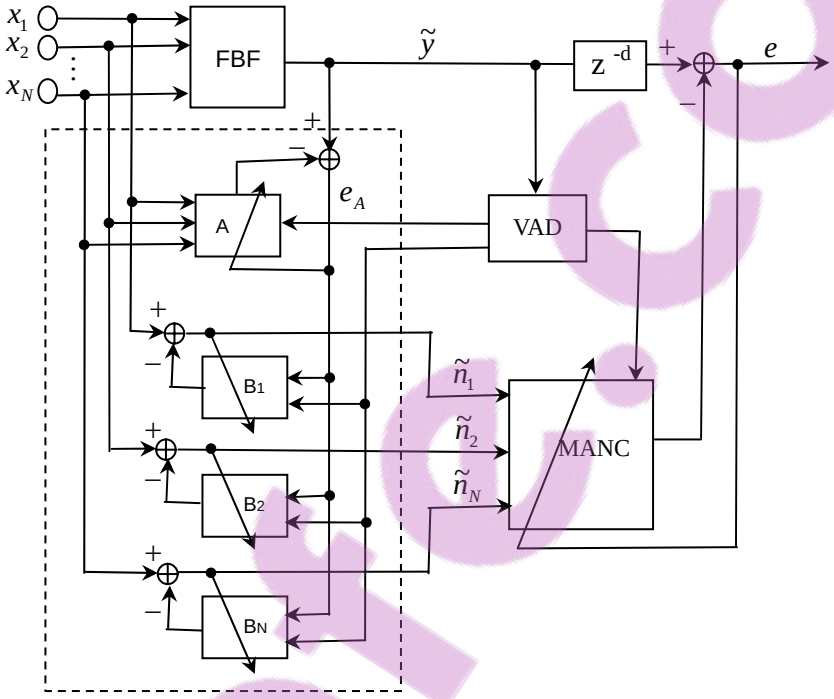

Figure 5.4.2 Structure of improved SDS-MGSC (ISDS-MGSC)

The speech blocking principle shown in figure 5.4.2 is briefly given as follows. It is very similar to that of MCRANC

Not losing generality, we may assume the output of the fixed beamformer is

$$\tilde{y}(k) = s_1(k) + \tilde{n}(k) \tag{5.4.1}$$

where $s_1(k)$ is the speech signal in $x_1(k)$. Thus $x_1(k)$ is used as the standard signal for time alignment. Generally $\tilde{y}(k)$ has a higher SNR than any $x_i(k)$.

During NSP (Non Speech Period), let us consider

$$e_A(k) = \tilde{n}(k) - \mathbf{wn}(k) \tag{5.4.2}$$

where $\mathbf{wn}(k)$ is the output of filter A, $\mathbf{w}$ is a $1 \times N(L+1)$ row vector coefficient

of filter A

$$\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_N) \tag{5.4.3}$$

$$\mathbf{w}_i = (w_{i0}, w_{i1}, \cdots, w_{iL})$$

and $\mathbf{n}(k)$ is a $N(L+1)\times 1$ column vector

$$\mathbf{n}(k) = [\mathbf{n}_1(k), \mathbf{n}_2(k), \cdots, \mathbf{n}_N(k)]^T \tag{5.4.4}$$

$$\mathbf{n}_i(k) = [n_i(k), n_i(k-1), \cdots, n_i(k-L)]$$

Adjust the coefficient of filter A to minimize the power of $e_A(k)$ to get the optimal coefficient

$$\mathbf{w}^* = (\mathbf{w}_1^*, \mathbf{w}_2^*, \cdots, \mathbf{w}_N^*)$$

$$= (w_{10}^*, w_{11}^*, \cdots, w_{1L}^*, w_{20}^*, w_{21}^*, \cdots, w_{2L}^*, \cdots\cdots, w_{N0}^*, w_{N1}^*, \cdots, w_{NL}^*) \tag{5.4.5}$$

The corresponding output of filter A is denoted by $e_A^*(k)$.

Then during HSP (Having Speech Period) that follows the previous NSP, we assume the environment remains almost unchanged or changes very slowly, and accordingly we may keep the optimal coefficients of filter A unchanged. Thus the output of filter A

$$\mathbf{w}^*\mathbf{x}(k) = \mathbf{w}^*[\mathbf{s}(k) + \mathbf{n}(k)]$$

$$= \mathbf{w}^*\mathbf{s}(k) + [\tilde{n}(k) - e_A^*(k)] \tag{5.4.6}$$

where $\mathbf{x}(k)$ and $\mathbf{s}(k)$ represent the desired signal plus noise and the pure desired signal vectors respectively, which may be expressed in a similar way to $\mathbf{n}(k)$ in equation (5.4.4). Then from equations (5.4.2) and (5.4.5)

$$e_A(k) = \tilde{y}(k) - \mathbf{w}^*\mathbf{x}(k)$$

$$= p(k) + e_A^*(k) \tag{5.4.7}$$

where

$$p(k) = s_1(k) - \mathbf{w}^*\mathbf{s}(k)$$

$$= s_1(k) - \sum_{i=1}^{N}\sum_{j=0}^{L} w_{ij}^{*} h_{s_1 s_i}(k-j) * s_1(k-j) \qquad (5.4.8)$$

According to (5.4.7) and (5.4.8), $e_A(k)$ is a distorted speech signal of $s_1(k)$. It is correlated with $s_1(k)$ and is also correlated with any $s_i(k)$ through the speech propagation impulse response $h_{s_1 s_i}(k)$. So $e_A(k)$ can be used to cancel the speech signal in every $x_i(k)$ to get a noise estimation $\tilde{n}_i(k)$ ($i = 1,2,\cdots,N$).

To sum up, for the blocking process, we only need to adjust the coefficients of filter A in figure 5.4.2 during NSP to minimize the power of $e_A(k)$ and to adjust the coefficients of every filter B_i during HSP to minimize the power of $e_{B_i}(k)$. Then the output $e_{B_i}(k)$ would be the estimation $\tilde{n}_i(k)$ of the pure noise $n_i(k)$.

However, $e_A(k)$ may still contain high level noise due to the incomplete correlation between different channel noises. This fact will cause partial noise cancellation in the estimated noise during HSP and results in the residual noise in the final enhanced signal during HSP being stronger than the residual noise during NSP. To make a steady residual noise in the final output, we can adjust the coefficients of filter B_i not only during HSP but also during NSP. That is, to adjust the coefficients of every filter B_i all the time.

5.4.3 Computational complexity

The difference between ISDS-MGSC and MGSC lies in the blocking module. So, we only need to find the computational cost of the blocking module required for implementation.

The blocking module of ISDS-MGSC consists of FIR filter A with N-channel inputs and N FIR filter B_i ($i = 1,2,\cdots,N$) with only one-channel input. If the LMS algorithm is employed for the adaptations of all FIR filters, it can be calculated that

the maximum numbers of additions and multiplications for all the computations in a second are

(1) maximum number of additions

$$\max\{(2NL+N+1)f, 2N(L_B+1)f\} \tag{5.4.9}$$

(2) maximum number of multiplications

$$\max\{(2NL+2\mathrm{N}+1)f, N(2L_B+3)f\} \tag{5.4.10}$$

where L is the sample delay for every channel of the signal input to filter A and thus the length of filter A is $N(L+1)$; L_B+1 is the length of filter B_i; and f is the sampling rate for array signals.

For example, if the array consists of five microphones and the sampling rate is $f=8000$ Hz; L=24 for filter A; $L_B=48$ for filters B_i ($i=1,2,\cdots,N$), the maximum number of additions per second for the blocking module of ISDS-MGSC is 3,920,000 and the maximum number of multiplications per second is 3,960,000 only.

5.5 Subband partial-channel SC-MGSC

Since both MCRASC-MGSC and SDS-MGSC employ the desired signal cancellation technique, we called them Signal Cancellation based MGSC (SC-MGSC).

If the subband method is used in SC-MGSC and partial channels of signals are employed for the blocking module of SC-MGSC, we call the SC-MGSC as Subband Partial-channel SC-MGSC (SP-SC-MGSC). Obviously SP-SC-MGSC is the extension of SC-MGSC because SP-SC-MCRANC is more flexible.

Figure 5.5.1 is the structure of SP-SC-MGSC, where Analysis and Synthesis represent analysis filter bank and synthesis filter respectively as described in section 4.4.1 of chapter 4, and P-SC-MGSC(j) is denoted for Partial-channel SC-MGSC in j-th subband which employs only partial channels of the array signals for the blocking module. If the SC-MGSC is an ISDS-MGSC as shown in figure 5.4.2, the

P-SC-MGSC(j) is shown in figure 5.5.2.

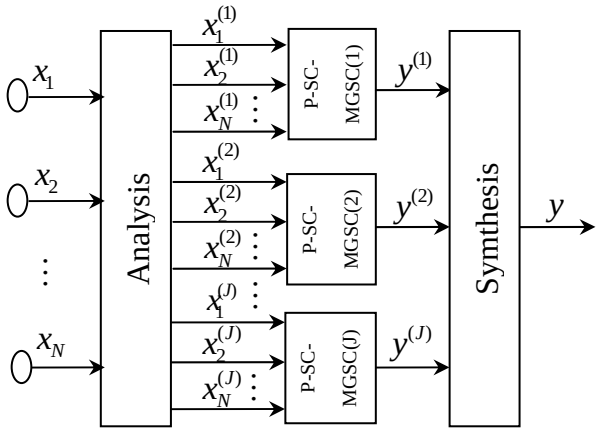

Figure 5.5.1 Structure of Subband Partial-channel SC-MGSC (SP-SC-MGSC)

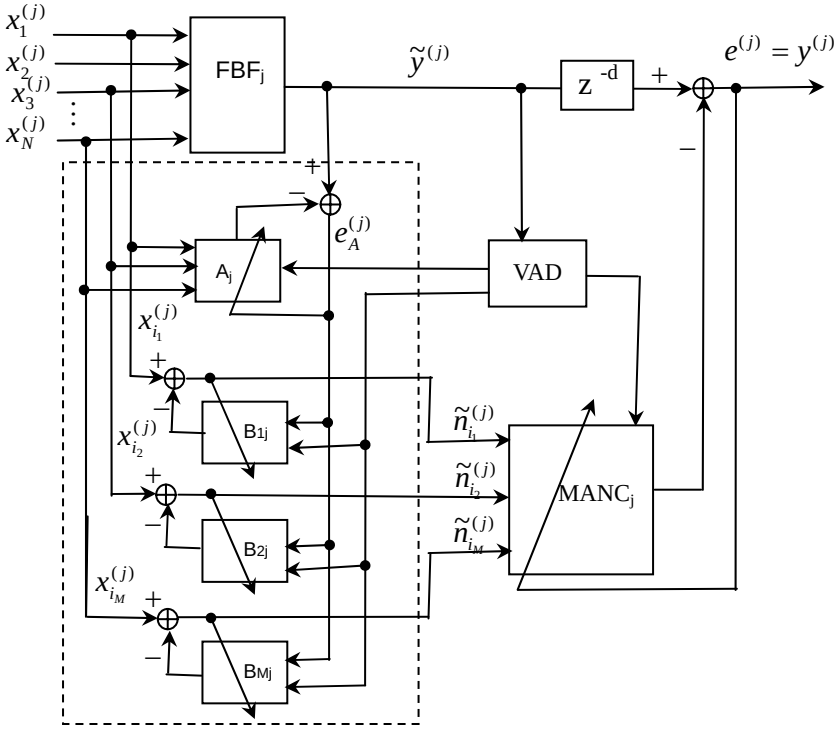

Figure 5.5.2 Structure of P-SC-MGSC used for the j-th subband (P-SC-MGSC(j))

5.6 Experimental results

In the experiment, five small microphones were used to construct a planar array

with an aperture of less than 5cm as shown in figure 5.6.1. The speech and the noises were generated concurrently by loudspeakers from different locations. The speech data was from a section of recorded speech in the computer and the noise data was from the NoiseX92 database. The sampling rate used to digitize the acquired signals was 8 kHz.

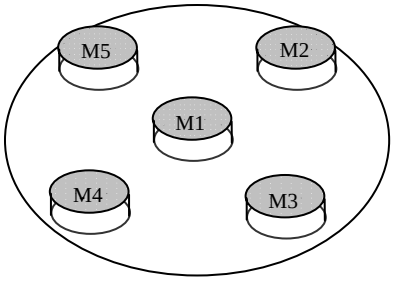

Figure 5.6.1 Employed planar microphone array

The experiment was made in a common study room of dimensions 5x4x2.8m. The array was put on a desk. The center of the array was 1.4m from the front wall, 1.8m from the left wall and 1.23m from the floor. There were two sofas, a cabinet and another two desks in the room. The room had two glass windows and a wooden door, all of them were closed.

One of the experiment cases is shown in figure 5.6.2. For simplicity, the figure is a planar one since the loudspeakers emitting speech and noises have almost the same height from the floor as the array in the experiment. In this case, the speech loudspeaker was placed 30cm in front of the microphone array at (0,30). Noise loudspeakers were concurrently activated to emit Volvo, Leopard, Factory2 and White noises. They were positioned at (-100,100), (50,50), (200,250) and (0,100)cm respectively. The following cases were tested.

Case 1. Speech at (0,30) and Leopard noise at (0,100).

Case 2. Speech at (0,30) and Leopard noise at (200,250).

Case 3. Speech at (0,30) and Volvo noise at (-100,100).

Case 4. Speech at (200,250) and Volvo noise at (-100,100).

Case 5. Speech at (0,30), Volvo noise at (-100,100) and Leopard noise at (50,50).

Case 6 Speech at (0,30), Volvo noise at (-100,100) and Factory2 noise at (200,250).

Case 7 Speech at (0,30), Leopard noise at (50,50) and Factory2 noise at (200,250).

Case 8 Speech at (0,30), Volvo noise at (-100,100), Leopard noise at (50,50) and Factory2 noise at (200,250).

Case 9 Speech at (0,30), Volvo noise at (-100,100), Leopard noise at (50,50), Factory2 noise at (200,250) and White noise at (0,100).

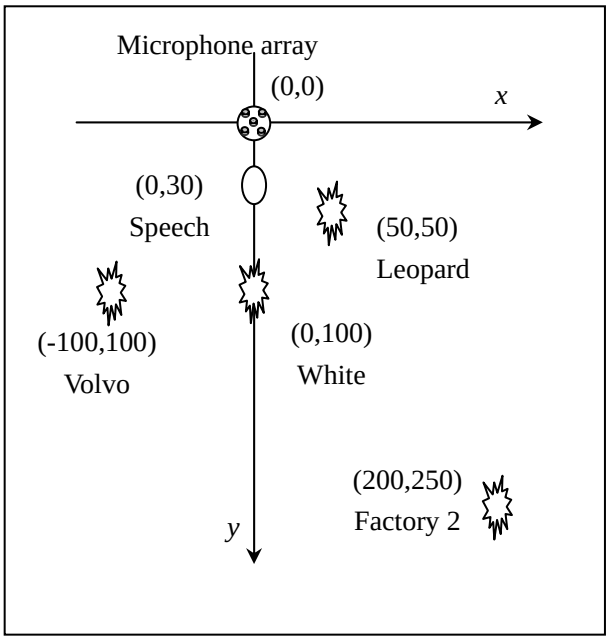

Figure 5.6.2 Case 9 of the experiment environments

Table 5.6.1 shows the SNRs and SNR improvements of the original and enhanced speeches by use of different algorithms including the GSC, MGSC, MCRASC-MGSC, ISDS-MGSC and SP-SC-MGSC. The last two rows are the average SNRs and average SNR improvements.

In table 5.6.1, the original noisy speech signal is $x_1(k)$ acquired from microphone M_1. Other noisy speech signals from other microphones have almost the same SNR as $x_1(k)$ has. Here the SNR is also calculated by equation (1.4.3).

In the processing, we use microphone M_1 as the standard calibrating microphone, a correlation method to calculate the time delays and the DAWSAS algorithm introduced in subsection 4.3.2 for fixed beamformer FBF. VAD employs an energy and zero-crossing rate method. Whenever VAD is failed, use the artificially decided results

about NSP and HSP. The adaptive FIR filter MANC has a length of 120 and a LMS adaptation algorithm with learning rate $\mu = 0.01$.

In MCRASC-MGSC processing, partial-channel MCRASC (corresponding to partial-channel MCRANC in chapter 4) is applied. For microphones M_1, M_2, M_3, M_4 and M_5, we take M=2 and the referential microphones are (M_2, M_3), (M_1, M_3), (M_1, M_4), (M_1, M_5) and (M_1, M_2) respectively. The lengths of filters A_i and B_i are 64 and 48 respectively ($i = 1,2,\cdots,5$). All filters employs the LMS adaptation algorithm with learning rate $\mu = 0.01$

In ISDS-MGSC processing, the length of filter A (see figure 5.4.2) is 120 and all filters B_i ($i = 1,2,\cdots,5$) had the same length of 48. All filters employ the LMS adaptation algorithm with learning rate $\mu = 0.01$.

In SP-SC-MGSC processing, the full frequency band is equally divided into 4 subbands. In every subband, all channels of signals are used and the structure shown in figure 5.5.2 is employed, where ISDS-MGSC is used for computing the outputs of the subband. The length of filter A_j is 80 and the length of B_{ji} is 32, $i = 1,2,\cdots,5$, $j = 1,2,\cdots,4$. All filters employ the LMS adaptation algorithm with learning rate $\mu = 0.01$.

Figure 5.6.3 shows the SNR results of the noisy speech and the enhanced speech using five different methods in nine cases. Each line indicates the results of a method.

Figure5.6.4 shows the signals concerned in case 9.

Figure5.6.4 (a) is the time domain waveform of noisy speech signal $x_1(k)$ from microphone M_1. Its SNR=2.25 dB.

Figure5.6.4 (b) is the enhanced speech by GSC with SNR=1.97dB.

Figure5.6.4 (c) is the enhanced speech by MGSC with SNR=5.25 dB.

Figure5.6.4 (d) is the enhanced speech by ACRACS-MGSC with SNR=16.89 dB.

Figure5.6.4 (e) is the enhanced speech by ISDS-MGSC with SNR=18.32 dB.

Figure5.6.4 (f) is the enhanced speech by SP-SC-MGSC with SNR=19.87 dB

Table 5.6.1 The SNRs (dB) of original noisy speech and the enhanced speech through GSC, MGSC, MCRASC-MGSC, ISDS-MGSC and SP-SC-MGSC

Method Case	Original	GSC	MGSC	MCRASC-MGSC	ISDS-MGSC	SP-SC-MGSC
1	2.64	2.31	6.87	17.11	19.95	21.33
2	13.29	8.99	16.54	23.78	23.59	23.06
3	11.58	5.06	13.89	24.14	23.35	25.87
4	7.38	9.32	12.17	19.10	20.03	22.76
5	2.62	2.34	6.92	17.12	19.66	21.32
6	13.00	7.42	13.49	24.86	23.41	23.96
7	2.56	2.21	5.79	17.43	19.38	23.03
8	2.54	2.15	5.84	16.80	19.03	21.98
9	2.52	1.97	5.25	16.89	18.32	19.87
average	6.46	5.08	9.64	19.69	20.75	22.58
improved	0.00	-1.38	3.18	13.23	14.29	16.12

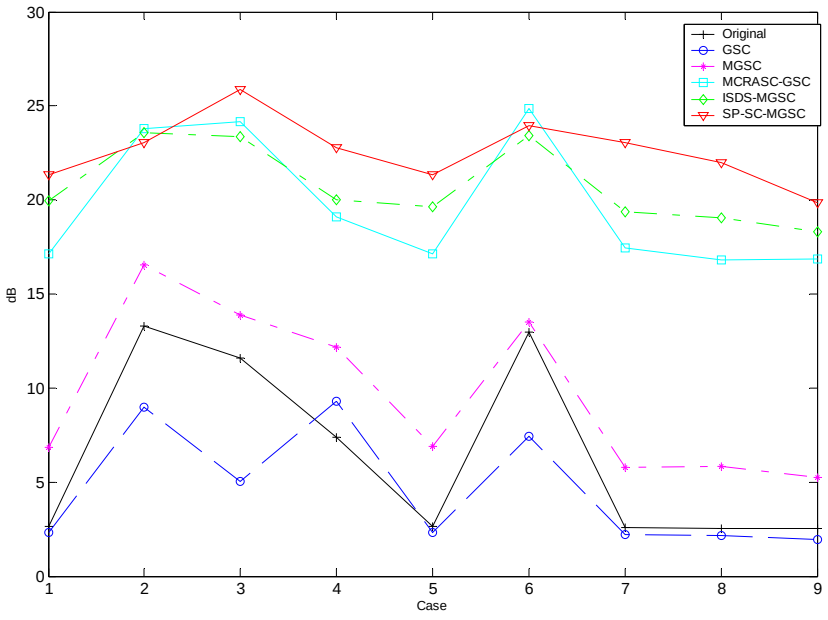

Figure 5.6.3 SNR lines of the noisy speech and the enhanced speech using five different algorithms in nine cases

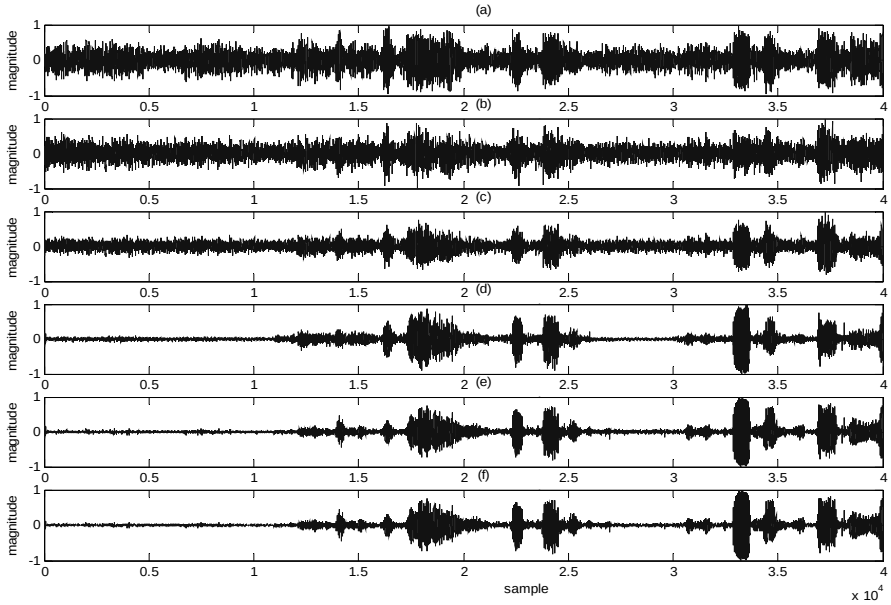

Figure 5.6.4 Speech enhancement results

(a) Noisy speech
(b) Enhanced speech by GSC
(c) Enhanced speech by MGSC
(d) Enhanced speech by MCRASC-MGSC
(e) Enhanced speech by ISDS-MGSC
(f) Enhanced speech by SP-SC-MGSC

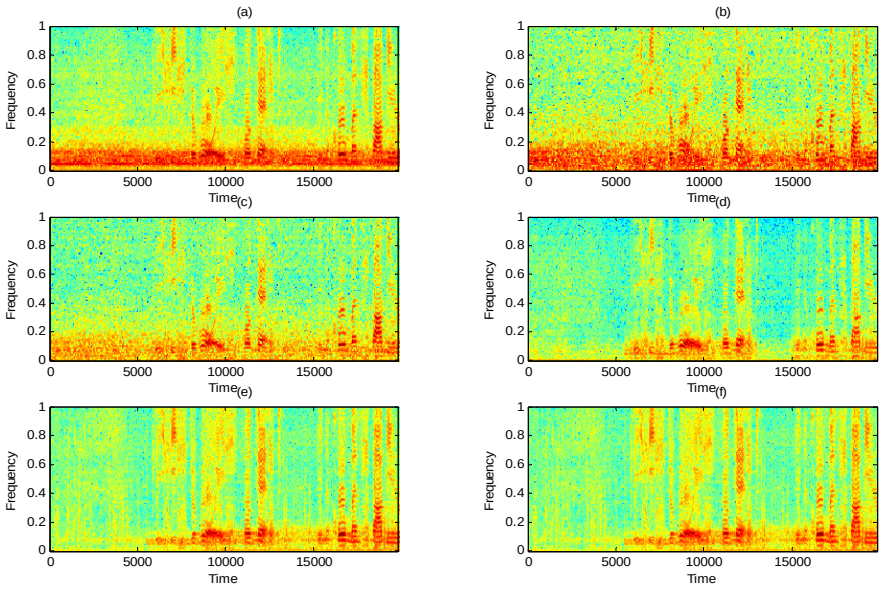

Figure 5.6.5 Spectrograms of the signals in figure.5.6.4

(a) Spectrograms of noisy speech
(b) Spectrograms of enhanced speech by GSC
(c) Spectrograms of enhanced speech by MGSC
(d) Spectrograms of enhanced speech by MCRASC-MGSC
(e) Spectrograms of enhanced speech by ISDS-MGSC
(f) Spectrograms of enhanced speech by SP-SC-MGSC

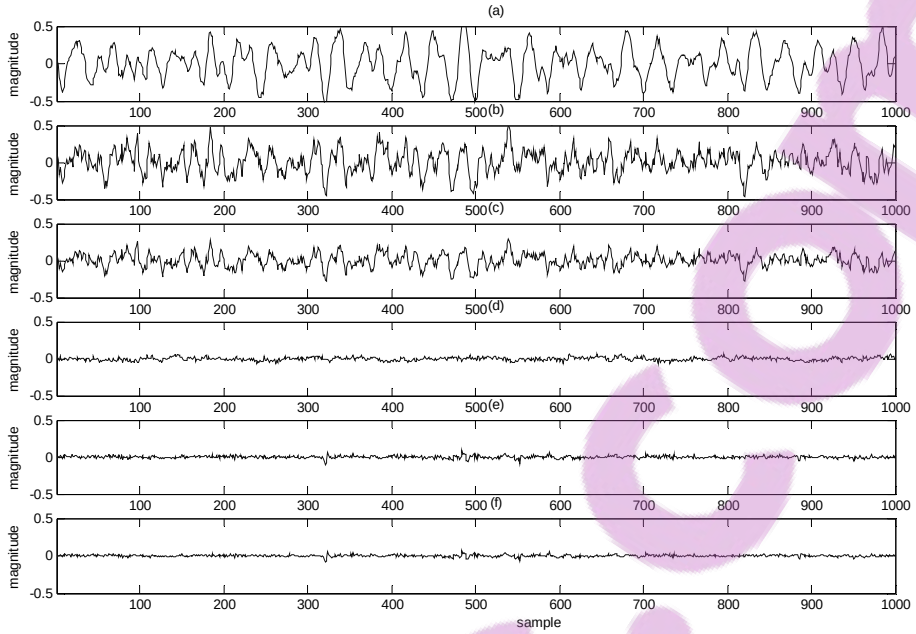

Figure 5.6.6 A zoomed section of figure 5.6.4 (non speech section)

(a) Noisy speech (b) Residual noise by GSC
(c) Residual noise by MGSC (d) Residual noise MCRASC-MGSC
(e) Residual noise by ISDS-MGSC (f) Residual noise by SP-SC-MGSC

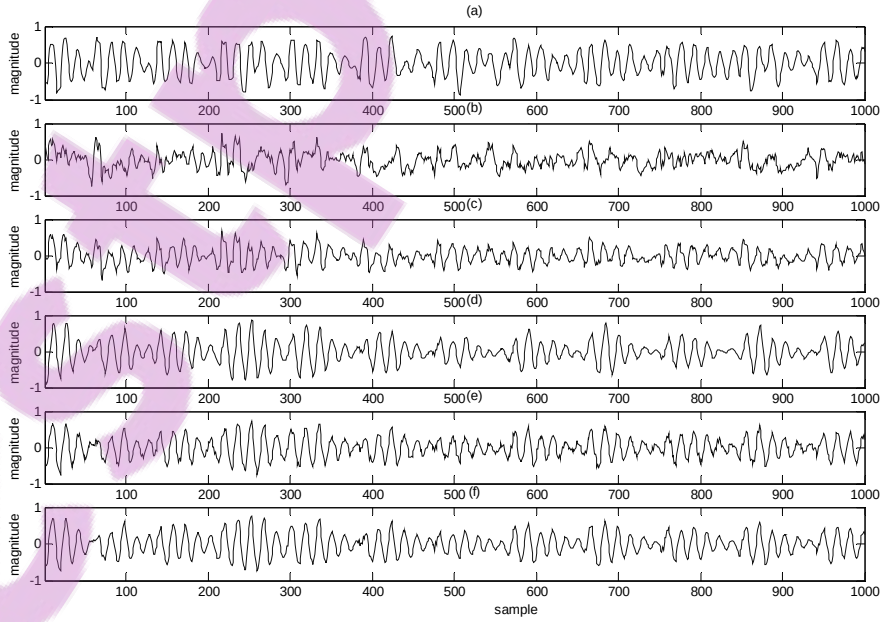

Figure 5.6.7 A zoomed section of figure 5.6.4 (speech section)

(a) Noisy speech (b) Enhanced speech by GSC
(c) Enhanced speech by MGSC (d) Enhanced speech by MCRASC-MGSC
(e) Enhanced speech by ISDS-MGSC (f) Enhanced speech by SP-SC-MGSC

Figure 5.6.5 shows the spectrograms of relevant signals in figure5.6.4.

Figure 5.6.6 shows a zoomed non speech section of relevant signals in figure 5.6.4

Figure 5.6.7 shows a zoomed speech section of relevant signals in figure 5.6.4

From table 5.6.1 and figures 5.6.3~7, we can find all proposed algorithms MCRASC-MGSC, ISDS-MGSC and SP-SC-MGSC achieve much more SNR improvement than conventional GSC and MGSC, with SP-SC-MGSC the most. It should be mentioned that ISDS-MGSC achieves more SNR improvement than MCRASC-MGSC although it has less computational cost than MCRASC-MGSC.

5.7 Summary

Improved MGSC algorithms are proposed in this chapter for speech enhancement. The principle of the improved algorithms is to employ the signal cancellation technique for the blocking module of MGSC. By introducing Multichannel Crosstalk Resistant Adaptive Signal Cancellation (MCRASC) for the desired signal blocking, the improved algorithms can more effectively block the speech signal and pass through the noises. MCRASC is similar to MCRANC and they can convert to each other easily. It is proved that the improved algorithms are actually the extension of MGSC. They extend the common matrix to time-variable vector matrix in the blocking module.

One improved MGSC algorithm is to use every channel of the array signal as the main channel signal and others as the referential signals for MCRASC to get a channel of estimated noise for the blocking module. This improved MGSC is named MCRASC-MGSC.

Another improved MGSC algorithm, named SDS-MGSC, can be viewed as a simplified MCRASC-MGSC. Through setting up a Shared Distorted Signal (SDS), the signal in every channel can be cancelled by use of the SDS to get an estimated noise for the blocking module. The way to set up the SDS is to employ any channel of the array signal as the main channel signal and others as referential signals for an adaptive noise cancellation filter. An improved SDS-MGSC, named ISDS-MGSC, is to employ the

output of the fixed beamformer FBF as the main channel signal for setting up the SDS.

In addition, the proposed algorithms are extended to subband processing and partial-channel processing. A Subband Partial-channel Signal Cancellation MGSC (SP-SC-MGSC) algorithm is also presented for speech enhancement.

Experimental results indicate the proposed algorithms achieve much more SNR improvement than conventional GSC and MGSC. ISDS-MGSC outperforms MCRASC-MGSC since it achieves more SNR improvement while it has less computational cost than MCRASC-MGSC.

Chapter 6 Hybrid Algorithms

Two hybrid algorithms are proposed in this chapter by taking use of the proposed algorithms in previous chapters. One hybrid algorithm is based on Multichannel Crosstalk Resistant Adaptive Noise Cancellation (MCRANC) and employs DAWSAS fixed beamforming, multiple sampling rates method, partial-channel method, multichannel distorted signal filtering method, subband method and ISS algorithm. The other hybrid algorithm is based on Modified Generalized Sidelobe Canceling (MGSC) and also contains most of the above-mentioned algorithms and methods. The proper cases for each hybrid algorithm are suggested. Experimental results verify the advanteges of both hybrid algorithms.

6.1 Introduction

Several algorithms are proposed for speech enhancement using a small microphone array as described in the previous chapters. All of these algorithms can be classified as two groups. One group is based on MCRANC including the algorithms in chapters 2, 3 and 4. The other group is based on MGSC including the algorithms in chapter 5.

To the MCRANC based group, some combined algorithms employing MCRANC and improved algorithms to MCRANC itself have been proposed. All of these algorithms can actually be used to form more powerful hybrid algorithms.

To the MGSC based group, the difference between the proposed algorithms lies in deferent blocking processes. These blocking processes are derived or are developed from MCRANC. They also can be used together for getting more powerful hybrid algorithms.

The DAS or DAWSAS algorithm takes effect when the noise is uncorrelated in the

microphone array, and it does not offer any improvement to the enhanced speech when the noise is completely correlated. The MANC or MCRANC algorithm has great effect when the noise is highly correlated, and it does not offer any improvement to the enhanced speech when the noise is completely uncorrelated. Therefore, any algorithm in the MGSC based group or any algorithm in the MCRANC based group which employs the DAS or DAWSAS subsystem may have the ability to enhance speech in any noise environment, correlated or uncorrelated.

MCRANC needs the speech signals to be highly correlated in the microphone array. If the microphones are closely placed and the speech source is close to the array, this requirement can usually be met well. In this situation, MCRANC performs better than improved MGSC. Otherwise, the improved MGSC algorithm may achieve better results. As a result, MCRANC based group performs better than MGSC based group if speech signals are highly correlated, while MGSC based group achieves better results than MCRANC based group if the speech signals are not highly correlated.

6.2 MCRANC based hybrid algorithm

The MCRANC algorithm can be used with many other speech enhancement algorithms. In chapter 3 some algorithms combining MCRANC with other speech enhancement algorithms are studied. In chapter 4 several improved algorithms to MCRANC are presented. In fact, all of these algorithms may be reasonably employed to construct more powerful MCRANC based hybrid algorithms.

One of these hybrid algorithms is indicated in figure 6.2.1 and figure 6.2.2. It employs subband method, DAWSAS beamforming, partial channels method, multiple sampling rates method, multichannel distorted signal filtering, and single-channel speech enhancement algorithm. These algorithms and methods have been discussed through chapter 2 to 4. Some suggestions about them are as follows.

Subband: This can employ equally-divided or unequally-divided frequency band

algorithms, especially the unequally-divided frequency band algorithms such as Gammatone and Mel frequency bands which make use of human perceptual characteristics.

Beamforming: This may employ the DAS or the DAWSAS algorithm presented in section 4.3.2 of chapter 4. It may also employ other fixed beamforming algorithms.

Partial channels: Only partial channels of the microphone array signals are selected to take part in MCRANC. Notice that we regard all channels as a special case of partial channels because employing all channels may be viewed as a case of partial channels with the maximum number of channels the array may offer.

Multiple sampling rates: Use high sampling rates for referential signals of MCRANC in high frequency bands, and low sampling rates in low frequency bands, as introduced in section 4.2 of chapter 4. When employing multiple sampling rates, over-sampling is necessary for microphone array signals.

Multichannel distorted signal filtering: For the second filter B in MCRANC, multichannel inputs can be employed to improve the recovered speech. This is discussed in section 4.1 of chapter 4.

Single-channel speech enhancement: Any one-channel speech enhancement algorithm is an example of this, such as the improved spectral subtraction algorithm, the Weiner filtering algorithm, and wavelet denoising algorithm, and so on.

In figure 6.2.1, firstly, every noisy speech signal x_i from the microphone array is divided through an analysis filter bank into subband signals $\{x_i^{(j)}, j = 1,2,\cdots,J\}$, $i = 0,1,,\cdots,N$, where J is the number of subbands. Secondly, use the signals in the same band $\{x_i^{(j)}, i = 0,1,2,\cdots,N\}$ to compute $y^{(j)}$ according to figure 6.2.2, $j = 1,2,,\cdots,J$. Thirdly, synthesize all $\{y^{(j)}, m = 1,2,\cdots,J\}$ through a synthesizing filter to get enhanced speech $\tilde{y}$. Finally, use single-channel speech enhancement algorithms, if necessary, to process $\tilde{y}$ to get the final enhanced speech y.

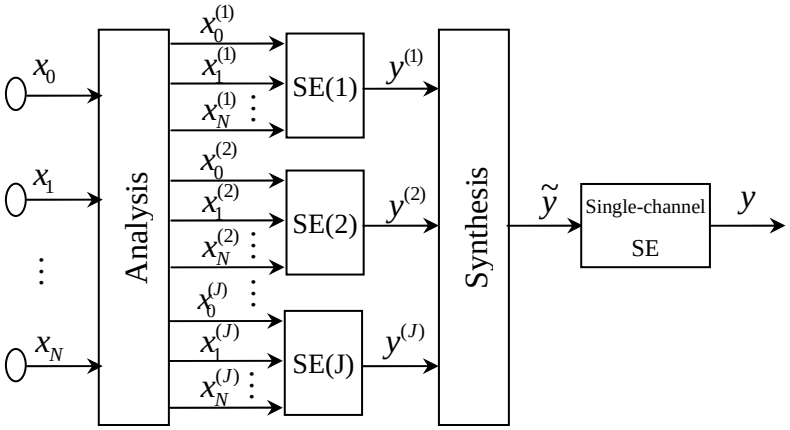

Figure 6.2.1 Structure of MCRANC based hybrid algorithm
(the subsystem SE(j) is indicated in figure 6.2.2)

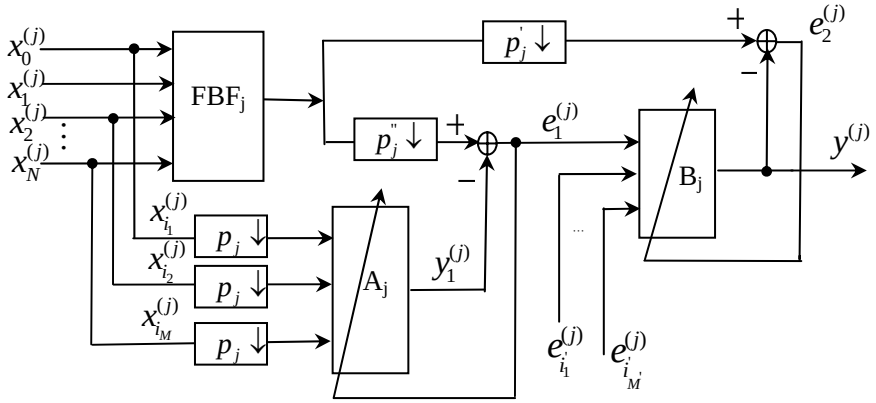

Figure 6.2.2 Structure of subsystem SE(j) in figure 6.2.1

Some suggestions about the subsystem SE(j) as indicated in figure 6.2.2 are as follows.

(1) In any subband, if the SNR of a noisy speech signal for the input, as shown in figure 6.2.2, is much higher, say more than 20 dB, the hybrid algorithm had better take it as the output of the subsystem directly. The subsystem had better not process any processing to avoid causing distortion to the speech since the speech signal already has a much higher SNR.

(2) The MSR (Multiple Sampling Rates) in subsystem SE(j) may employ over-sampling for realization. That is to use a higher sampling rate for array signals than

the required sampling rate for the output enhanced speech. Then down sample the signals through factors p_j, $p_j^{'}$ and $p_j^{''}$ to get some signals with different sampling rates. If the j-th subband is a high frequency band, the three factors should satisfy $p_j \le p_j^{''} \le p_j^{'}$ according to section 4.2 of chapter 4. If the j-th subband is a low frequency band, the factors should be selected to satisfy an opposite relationships. The high or low frequency band can be decided by how many bands the full band is divided. For example, if we divide the full band into eight bands, the first two bands can be regarded as low frequency band while the last two bands can be regarded as high frequency bands.

(3) Since the MDS (multichannel distorted signal) filtering method will greatly increase the computational complexity as the increase of $M^{'}$ in figure 6.2.2 and offer only minor SNR improvement for the enhanced speech, $M^{'}$ should not be big.

6.3 MGSC-based hybrid algorithm

Like the MCRANC algorithm, the improved MGSC algorithms as described in chapter 5 can also employ most of the algorithms or methods described in chapter 4 and chapter 3 to get more powerful MGSC based hybrid algorithms.

A MGSC based hybrid algorithm is indicated in figure 6.3.1 and figure 6.3.2. It employs subband method, DAWSAS beamforming, partial channels method, multiple sampling rates method, and single-channel speech enhancement algorithm. About these algorithms and methods, we have almost the same suggestions as listed in section 6.2 and they will not be duplicated in this section. The only exceptional suggestion is about the MSR (multiple sampling rates) method when it is applied to subsystem SE(j). The suggestion is a little bit different and it is described as follows.

MSR method in subsystem SE(j) may employ over-sampling for realization. It samples the array signals by using higher sampling rate. If the j-th subband is a high

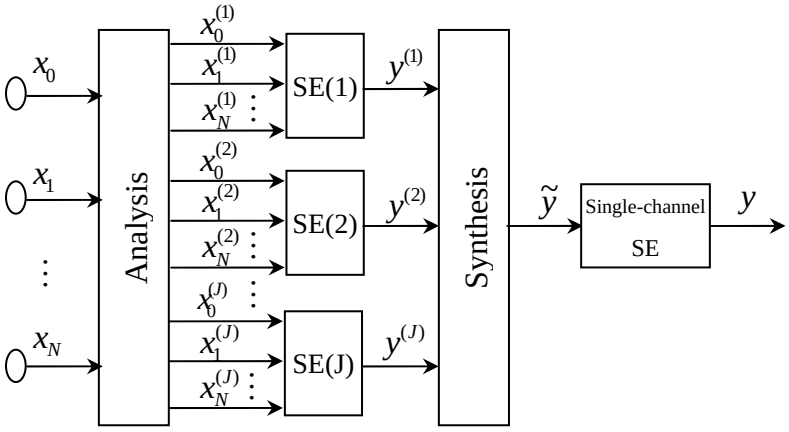

Figure 6.3.1 Hybrid algorithm based on MGSC
(the subsystem is indicated in figure 6.3.2)

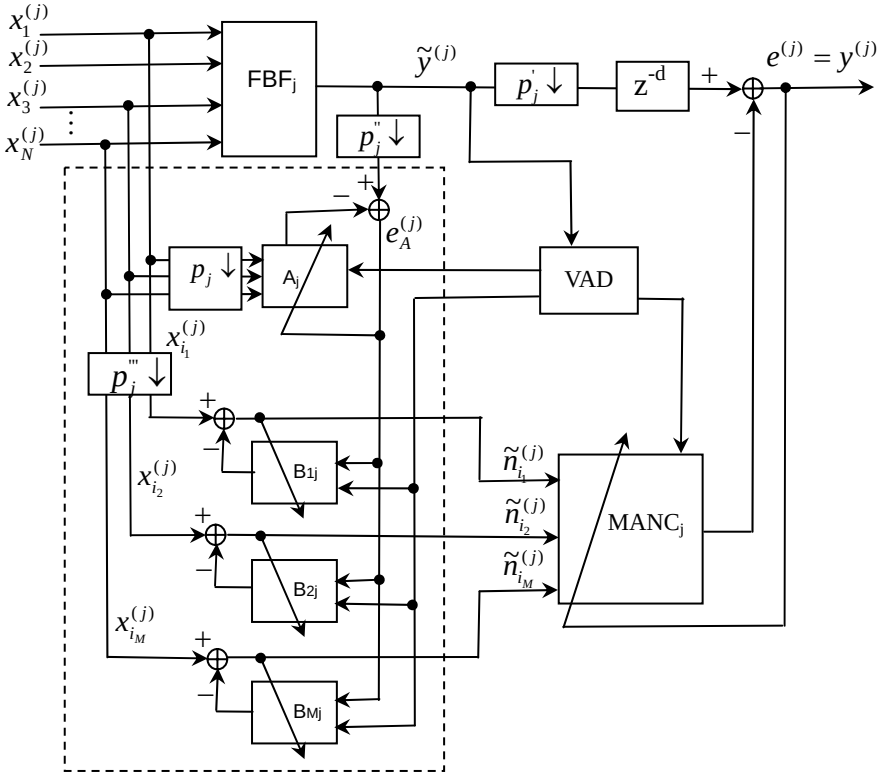

Figure 6.3.2 Structure of subsystem SE(j) in figure 6.3.1

frequency subband, the down sampling factors p_j, $p_j^{''}$ and $p_j^{'''}$ in figure 6.3.2 should be smaller than $p_j^{'}$ to get high sampling rate signals. If the j-th subband is a

low frequency subband, factors p_j, $p_j^{''}$ and $p_j^{'''}$ should be bigger than $p_j^{'}$ to get low sampling rate signals. These factors can be decided with reference to section 4.2 of chapter 4. If we do not employ MSR method, we have $p_j = p_j^{'} = p_j^{''} = p_j^{'''} = 1$.

6.4 Experimental results

In the experiment, a microphone array was constructed with five microphones placed as the array in figure 5.6.1 in chapter 5, with an aperture less than 7cm. The speech and the noises were generated concurrently by loudspeakers from different locations. The speech data was from a section of recorded speech in a computer and the noises data was from the NoiseX92 database. The required sampling rate for the output speech was 8kHZ. The sampling rate used to digitize the acquired signals was 32k Hz.

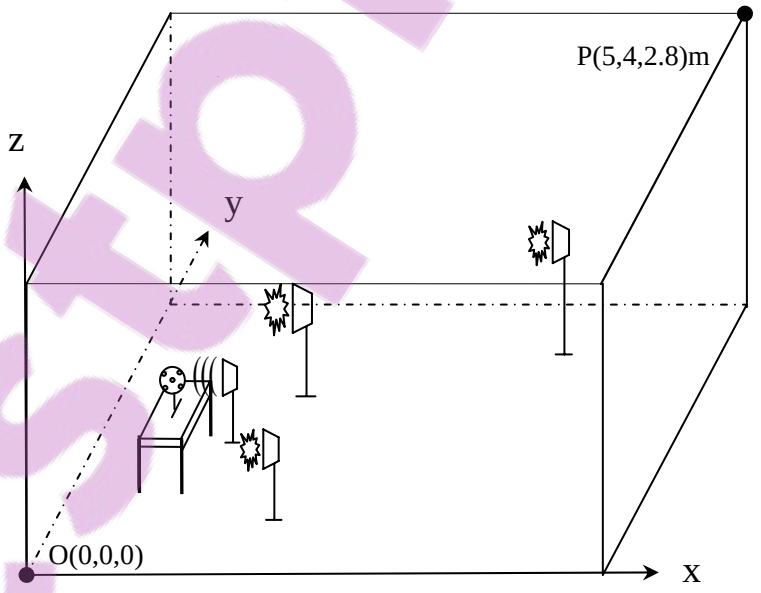

Figure 6.4.1 One of the experiment environments (case 8)

The experiment was made in a common room of dimensions 5x4x2.8m. There were two sofas, a cabinet and two other desks in the room. The room had two glass

windows and a wooden door, and all of them were closed.

In figure 6.4.1, a 3-dimensional coordinate is employed to show the locations of the microphone array and the loudspeakers. The array was put on a desk. The center of the array was at (100, 150, 120) cm in the coordinate. All loudspeakers faced to the microphone array and their positions could be changed. Nine cases were tested.

One of the cases is indicated by figure 6.4.1. In this case, four loudspeakers emitted speech and noises concurrently. The loudspeaker emitting speech was located at (130, 150, 120) cm. Three other loudspeakers emitted Volvo, Leopard and Factory2 noises and they were located at (200,50, 150), (150,200, 120), (350,350, 170) respectively.

Many cases are tested. Among them, nine cases are:

Case 1: Speech at (130,150,120). Leopard noise at (150,200,120).

Case 2: Speech at (130,150,120). Leopard noise at (350,350,170).

Case 3: Speech at (130,150,120). Volvo noise at (-100,100,150).

Case 4: Speech at (350,350,170). Volvo noise at (200,50,150).

Case 5: Speech at (130,150,120). Volvo noise at (200,50,150) and Leopard noise at (150,200,120).

Case 6: Speech at (150,200,120). Volvo noise at (200,50,150) and Factory2 noise at (350,350,170).

Case 7: Speech at (130,150,120). Leopard noise at (150,200, 120) and Factory2 noise at (350,350,170).

Case 8: Speech at (130,150, 120). Volvo noise at (200,50,150), Leopard noise at (150,200,120) and Factory2 noise at (350,350,170).

Case 9: Speech at (130,150,120). Volvo noise at (200,50,150), Leopard noise at (150,200,120), Factory2 noise at (350,350,170) and White noise at (200,150,150).

Table 6.4.1 lists the SNRs and SNR improvements of the original and enhanced speeches by use of different algorithms including the MGSC, the MCRANC based hybrid algorithm and the MGSC based hybrid algorithm proposed in section 6.2 and section 6.3 respectively. The last two rows are the average SNRs and average SNR

improvements.

In table 6.4.1 the original noisy speech signal is $x_1(k)$ from microphone M_1. Other noisy speech signals from other microphones have almost the same SNR as $x_1(k)$ has.

Here SNR is also calculated by equation (1.4.3), i.e.

$$SNR = 10 \log_{10} \frac{\alpha \sum_{k \in T_s} x^2(k) - \sum_{k \in T_n} x^2(k)}{\sum_{k \in T_n} x^2(k)}$$

where $x(k)$ is the noisy signal concerned; T_s is the sample set containing speech signal; T_n is the sample set without speech signal (pure noise); and $\alpha = m(T_n)/m(T_s)$, where $m(T_n)$ and $m(T_s)$ are the number of samples in T_n and T_s respectively.

In processing the speech enhancement, both hybrid algorithms equally divide the full frequency band (200, 4000) into 4 subbands. In every subband, all channels of signals are employed for beamformers. All beamformers take the DAWSAS algorithm for beamforming and the central microphone M_1 is used as a standard microphone for time-aligning. The time delays are estimated by the GCC method. The single-channel speech enhancement method for both hybrid algorithms is ISS algorithm as introduced in section 3.2 of chapter 3.

In processing the MCRANC based hybrid algorithm, the parameters for the subsystems (see figure 6.2.2) are selected as follows: $M = 3$ and $\{i_1, i_2, i_3\} = \{1,3,5\}$, which means we employed partial-channel MCRANC; $M' = 0$, which implies only one channel distorted signal for the filter in the second stage of MCRANC. The multiple sampling rates method was applied with $p_1 = 8$ for the lowest frequency subband and $p_4 = 1$ for the highest frequency subband and $p_2 = p_3 = 4$ for the other two frequency subbands, and $p_1' = p_2' = p_3' = p_4' = 4$, $p_1'' = 8$, $p_2'' = p_3'' = 4$, $p_4'' = 2$. The lengths of the FIR filters A_j and B_j all were around 120 and 60 respectively ($j = 1,2,3,4$). All filters employed the LMS adaptation algorithm with learning rate $\mu = 0.01$.

In the MGSC based hybrid algorithm, for all of its subband systems (see figure 6.3.2), we took the following parameters: $M = 3$ and $\{i_1, i_2, i_3\} = \{1,3,5\}$, which means we employed partial-channel MGSC. The multiple sampling rates method is applied with $p_1 = 8$ for the lowest frequency subband and $p_4 = 1$ for the highest frequency subband and $p_2 = p_3 = 4$ for the other two frequency subbands, and $p_1^{'} = p_2^{'} = p_3^{'} = p_4^{'} = 4$, $p_1^{''} = 8$, $p_2^{''} = p_3^{''} = 4$, $p_4^{''} = 2$, $p_1^{'''} = 8$, $p_2^{'''} = p_3^{'''} = 4$, $p_4^{'''} = 2$. The length of FIR filters A_j are around 120, the length of FIR filters B_{ij} are around 32 ($i = 1,2,3$), d=5 and the length of FIR filters MANC_j are around 90, for all $j = 1,2,\cdots,4$. All filters employed the LMS adaptation algorithm with learning rate $\mu = 0.01$.

Figure 6.4.2 shows the SNR results of the noisy speech and the enhanced speech using three different algorithms in nine cases. Each line indicates the SNR results of the enhanced speech by use of an algorithm.

Figure 6.4.3 shows the signals concerned in case 8.

Figure 6.4.3 (a) is the time-domain waveform of noisy speech signal $x_1(k)$ from microphone M_1. Its SNR=2.56 dB.

Figure 6.4.3 (b) is the enhanced speech by MGSC algorithm. Its SNR=5.88dB.

Figure 6.4.3 (c) is the enhanced speech by the MCRANC based hybrid algorithm with SNR=26.37 dB.

Figure 6.4.3 (d) is the enhanced speech by the MGSC based hybrid algorithm with SNR=25.25 dB.

Figure 6.4.4 shows the spectrograms of the corresponding signals in figure 6.4.3.

Figure 6.4.5 shows a zoomed non-speech section of the corresponding signals in figure 6.4.3.

Figure 6.4.6 shows a zoomed speech section of the corresponding signals in figure 6.4.3.

From table 6.4.1 and figures 6.4.2~6, we find both hybrid algorithms achieve much more SNR improvements than MGSC. If we compare them with the experimental results described in section 4.4 of chapter 4 and the experimental results described in

section 5.5 of chapter 5, we can also conclude that MCRANC based hybrid algorithm outperforms the improved MCRANC algorithms while MGSC based hybrid algorithm outperforms the improved MGSC algorithms. That is because both hybrid algorithms employ more algorithms and methods.

From table 6.4.1, figure 6.4.2 and many other tested cases, we also find that the MCRANC based hybrid algorithm performs a little bit better than the MGSC based hybrid algorithm when the speech source is near the microphone array, as in the cases 1,2,3,5,7,8,9. The MGSC based hybrid algorithm outperforms the MCRANC based hybrid algorithm when the speech source is not near the microphone array, as in the cases 4 and 6.

Table 6.4.1 The SNRs (dB) of original noisy speech and the enhanced speech through the MGSC, the MCRANC based hybrid algorithm and the MGSC based hybrid algorithm

Algorithm Case	Original	MGSC	MCRANC based hybrid algorithm	MGSC based hybrid algorithm
1	2.82	6.88	25.65	24.13
2	13.32	16.50	28.68	26.56
3	11.60	13.98	28.26	25.87
4	7.48	12.10	21.53	26.36
5	2.65	6.86	26.76	24.51
6	13.08	13.79	23.41	23.98
7	2.50	5.85	28.35	25.03
8	2.56	5.88	27.63	25.98
9	2.53	5.35	27.02	24.87
Average	6.50	9.69	26.37	25.25
Improved	0.00	3.19	19.87	18.75

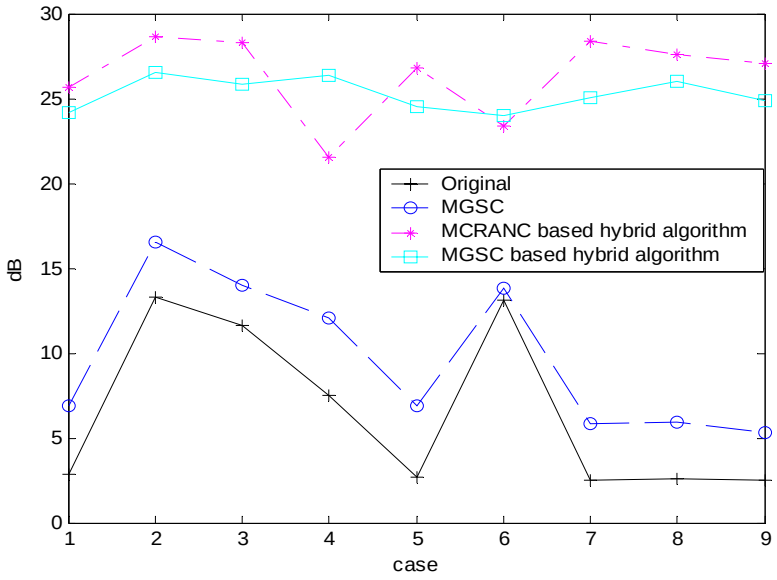

Figure 6.4.2 SNR lines of the noisy speech and the enhanced speech using three different algorithms in nine cases

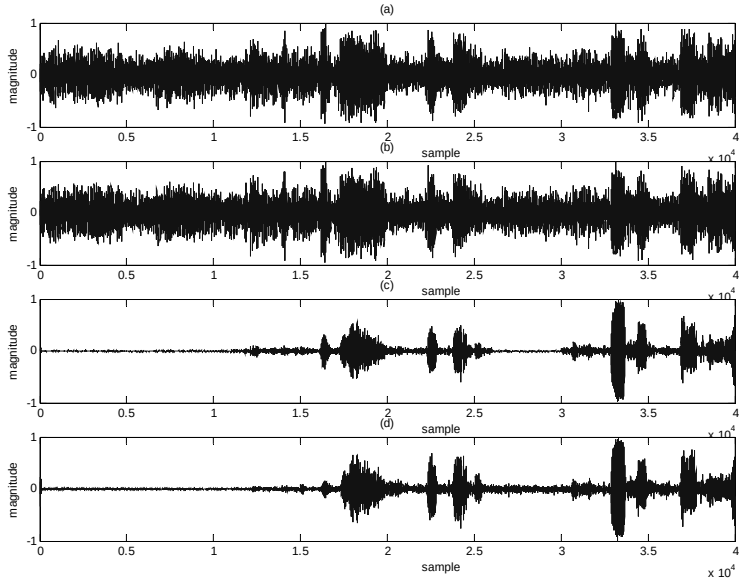

Figure 6.4.3 Speech enhancement results (case 8)

(a) Noisy speech

(b) Enhanced speech by MGSC

(c) Enhanced speech by the MCRANC based hybrid algorithm

(d) Enhanced speech by the MGSC based hybrid algorithm

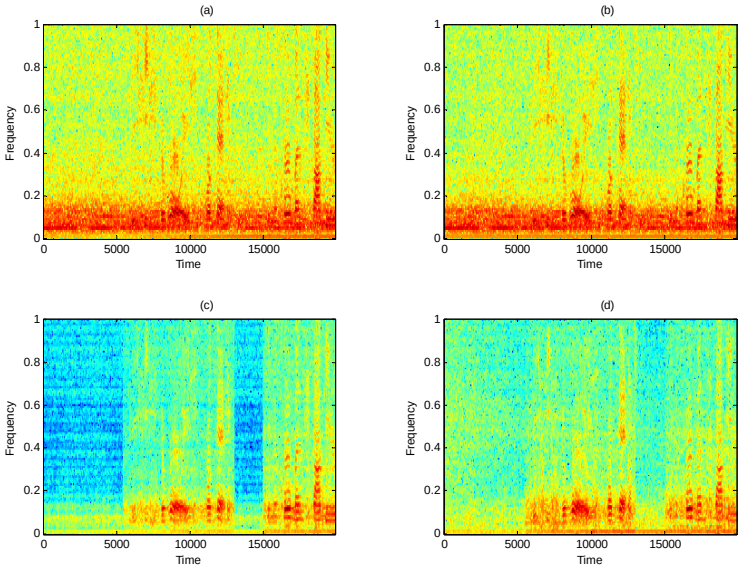

Figure 6.4.4 Spectrograms of the signals in figure 6.4.3
(a) Spectrogram of noisy speech
(b) Spectrogram of enhanced speech by MGSC
(c) Spectrogram of enhanced speech by the MCRANC based hybrid algorithm
(d) Spectrogram of enhanced speech by the MGSC based hybrid algorithm

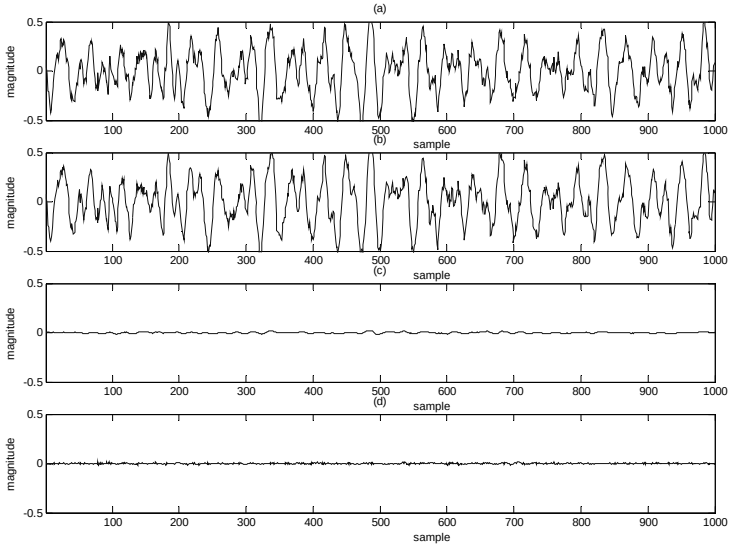

Figure 6.4.5 A zoomed section of figure 6.4.3 (non speech section)
(a) Pure noise
(b) Residual noise by MGSC
(c) Residual noise by the MCRANC based hybrid algorithm
(d) Residual noise by the MGSC based hybrid algorithm

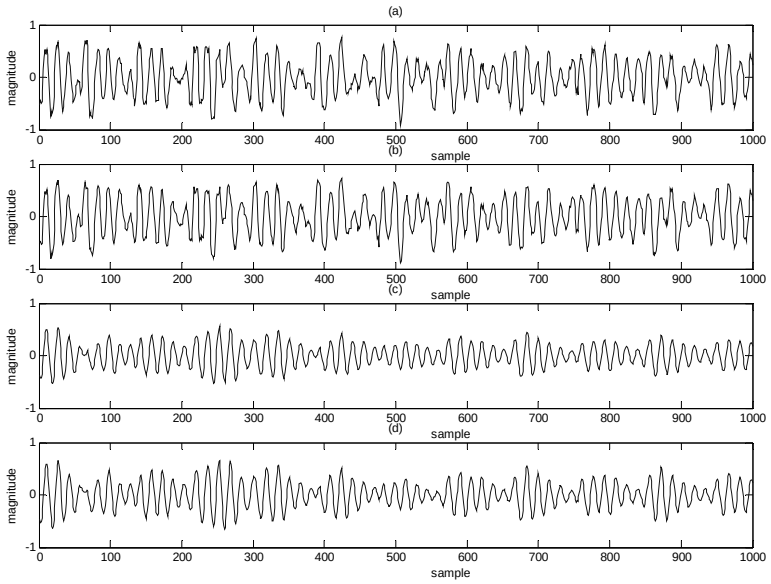

Figure 6.4.6 A zoomed section of figure 6.4.3 (speech section)
(a) Noisy speech
(b) Enhanced speech by MGSC
(c) Enhanced speech by the MCRANC based hybrid algorithm
(d) Enhanced speech by the MGSC based hybrid algorithm

6.5 Summary

Two hybrid algorithms are proposed in this chapter by taking use of the algorithms in the previous chapters.

One hybrid algorithm is based on MCRANC. Besides MCRANC, it also employs subband method, DAWSAS beamforming, partial channels method, multiple sampling rates method, multichannel distorted signal filtering, and single-channel speech enhancement algorithm.

The other hybrid algorithm is based on MGSC. Besides MGSC, it also employs subband method, DAWSAS beamforming, partial channels method, multiple sampling rates method, and single-channel speech enhancement algorithm.

Both hybrid algorithms proposed outperform other algorithms presented in previous chapters. Experimental results verify their advantages.

Both hybrid algorithms can be used for different environments. However,

MCRANC based hybrid algorithm performs somewhat better than the other hybrid algorithm when the speech source is near the microphone array, such as in cases of telephone and mobile phone. The MGSC based hybrid algorithm appears more suitable than MCRANC based hybrid algorithm when the speech source is not near the microphone array such as in the case of hearing aid.

Chapter 7 Conclusions And Future Work

7.1 Conclusions

Speech enhancement has wide applications, yet it is a challenging research field. Among many speech enhancement algorithms, microphone array based algorithms may achieve better performances. However, the common microphone array cannot be embedded in many small devices. Therefore, research on algorithms for a small microphone array has great value. The requirements for small size and fewer microphones make the research more challenging.

In this thesis, two hybrid algorithms are proposed for the speech enhancement using a small microphone array. One is the MCRANC based algorithm and the other is the MGSC based algorithm. Both hybrid algorithms can be used for different environments and both are suitable for real-time implementation.

The MCRANC based algorithm performs better than the MGSC based algorithm when the speech source is near the microphone array, such as in the cases of telephone and mobile phone. The MGSC based algorithm outperforms the MCRANC based algorithm when the speech source is not near the microphone array such as in the case of hearing aid.

The proposed hybrid algorithms are presented in chapter 6. They employ several methods and algorithms which are proposed or introduced in the previous chapters of this thesis, such as MCRANC algorithm, DAWSAS beamforming, subband method, multiple sampling rates method, MDS filtering, partial channels method, and single-channel speech enhancement algorithm. These employed algorithms and methods are presented in chapter 2 to 5, and they can be briefly summarized as follows.

In chapter 2, an algorithm called Multichannel Crosstalk Resistant Adaptive Noise Cancellation (MCRANC) is proposed. It employs only two adaptive filters and has the characteristics of low computational complexity, good stability and significant

enhancement effect. It is suitable for speech enhancement with a small microphone array. A simulation experiment and many experiments in real environments are presented to verify its effectiveness.

Chapter 3 discusses the combinations of MCRANC with other speech enhancement algorithms to get better performances. A combined algorithm with Improved Spectral Subtraction (ISS) in the single-channel algorithms is proposed. Two combined algorithms with Delay And Sum (DAS) beamforming and Weiner Post-filtering (WPF) in the microphone array algorithms are presented respectively. Experimental results are described to indicate the advantages of the combinations.

Chapter 4 gives some improvements to the MCRANC algorithm itself. One of the improvements is to use multichannel distorted signals for the speech recovering filter in MCRANC. Another improvement is to employ different sampling rates for the main channel microphone signal and the referential channel microphone signals. It is suggested that the sampling rate for the referential signals be properly higher than the required sampling rate for the output speech if the noises are high frequency noises, and be lower if the noises are low frequency noises. The third improvement is to add a fixed beamformer to MCRANC and use only partial channels of signals for noise cancellation. This method may broaden the application area as well as improve the enhancement results since it may deal with correlated and uncorrelated noises. The fourth improvement is to use the subband method for MCRANC. A MCRANC based algorithm employing subband, fixed beamforming and partial channels method is proposed.

The Generalized Sidelobe Canceling (GSC) algorithm is well known for signal enhancement. However, it does not perform well when used for a small microphone array. Chapter 5 presents improvements to MGSC (Modified GSC) in order to make it more suitable for a small array. One of the improvements is to use MCRASC and another is to set up a shared distorted signal for the blocking process of MGSC. It is proved that the improvements are actually to extend the blocking matrix of MGSC from a common matrix to a time-variable vector-valued matrix.

Chapter 6 summarizes the algorithms and methods from chapter 2 to 5 and presents two hybrid algorithms, while chapter 1 and chapter 7 are introduction and conclusions respectively.

7.2 Future work

Although some effective algorithms are proposed in this thesis for small microphone array based speech enhancement, these algorithms can further be improved for better enhanced speech and to suit more environmental variability. Some suggestions for future work are:

(a) Improvements can be achieved by including the proposed algorithms in this thesis with the psychoacoustical model-based algorithms, or with the Generalized Singular Value Decomposition (GSVD) algorithms. Also, the frequency domain realizations of the proposed algorithms can be studied.

(b) All algorithms proposed in this thesis need a VAD likewise the other well-known algorithms. However, in a noisy environment a VAD may detect the wrong section of pure noise. Like the spectral subtraction algorithms, the miss detection of noise may negatively affect the performance because the speech signal might partially be canceled with the noise cancellation. So the research on robust VAD in a noisy environment is very useful for these algorithms.

References

[1] W. Abdulla, M. Abdul-Karim. Real-time Spoken Arabic Digits Recognizer. Int. J. Electronics, Vol. 59, No. 5, pp. 645-648, 1985

[2] W. Abdulla, W. Mahmoud. Implementation and Performance of Vector Quantization in Automatic Speech Recognition. J. Electronics and Computers Research, Vol. 3, No. 2, pp. 113-126, 1989

[3] H. R. Abutalebi, H. Sheikhzadeh, R. L. Brennan, G. H. Freeman. A Hybrid Subband Adaptive System for Speech Enhancement in Diffuse Noise Fields. IEEE Signal Processing Letters, Vol.11, No.1, pp. 44-47, 2004

[4] M. J. Alam, D. O'Shaughnessy, S. A. Selouani. A New Perceptual Post-filter for Single Channel Speech Enhancement. ICECE International Conference on Electrical and Computer Engineering, pp.386-390, 2008

[5] S. Araki, S. Makino, R. Aichner, T. Nishikawa, H. Saruwatari. Subband Based Blind Source Separation for Convolutive Mixtures of Speech. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 5, pp. 509-512, 2003

[6] J. Arenas-Garcia, A. R. Figueiras-Vidal, A. H. Sayed. Mean-Square Performance of a Convex Combination of Two Adaptive Filters. IEEE Transactions on Signal Processing, Vol. 54, No. 3, pp.1078-1090, 2006

[7] F. Asano, S. Hayamizu, T. Yamada, S. Nakamura. Speech Enhancement Based on the Subspace Method. IEEE Transactions on Speech and Audio Processing, Vol.8, No.5, pp.497-507, 2000

[8] S. Ayat, M. T. Manzuri, R. Dianat, et al. An Improved Spectral Subtraction Speech Enhancement System by Using an Adaptive Spectral Estimator. IEEE Conference on Electrical and Computer Engineering, Vol. 1, No.1-4, pp. 261-264, 2005

[9] M. Bahoura, J. Rouat. Wavelet Speech Enhancement Based on the Teager Energy Operator. IEEE Signal Processing Letters, Vol.8, No.1 pp.10-12, 2001

[10] A. Bell, T. Sejnowski. An Information Maximization Approach to Blind Separation and Blind Deconvolution. Neural Computing, Vol. 7, pp.1129-1159, 1995

[11] J. Benesty, S. Makino, J. Chen. Speech Enhancement. Berlin: Springer, 2005

[12] J. V. Berghe, J. Wouters. An Adaptive Noise Canceller For Hearing Aids Using Two Nearby Microphones. Journal of Acoustical Society of America. Vol. 103, No. 6, 1998

[13] M. Berouti, R. Schwartz, J. Makhoul. Enhancement of speech corrupted by acoustic noise. Proceedings of IEEE International Conference on ASSP, pp. 208-211, 1979

[14] J. Bitzer, K. U. Simmer, and K. D. Kammeyer. Theoretical Noise Reduction Limits of the Generalized Sidelobe Canceller (GSC) for Speech Enhancement. in Proceedings of International Conference on Acoustics, Speech and Signal Processing, Vol.5, 15-19, pp.2965 - 2968, 1999

[15] S. F. Boll. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 27, No. 2, pp. 113-120, 1979

[16] R. L. Bouquin. Enhancement of Noise Speech Signals: Application to Mobile Radio Communications. Speech Communication, Vol. 18, pp. 3-19, 1996

[17] R. L. Bouquin, G. Faucon. Study of a Voice Activity Detector and its Influence on a Noise Reduction System. Speech Communication, Vol.16, pp. 245-254, 1995

[18] M. Brandstein, D. Ward. Microphone Arrays: Signal Processing Techniques and Applications, Springer-Verlag, 2001

[19] M. S. Brandstein, H. F. Silverman. A Robust Method for Speech Signal Time-delay Estimation in Reverberant Rooms. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol.1, pp.375-378, 1997

[20] K. M. Buckley and L. J. Griffith. An Adaptive Generalized Sidelobe Canceller with Derivative Constraints. IEEE Transactions on Antennas Propagation, Vol. 34, pp. 311-319, 1986

[21] J. F. Cardoso. Blind Signal Separation: Statistical Principles. Proceedings of IEEE, Vol. 86, No. 10, pp. 2009-2025, 1998

[22] D. J. Chapman. Partial Adaptivity for Large Arrays. IEEE Transactions on Antennas Propagation, Vol. 24, pp. 685-696, 1976

[23] N. Cheng, W. Liu, P. Li, B. Xu. Microphone Array Speech Enhancement Based on a Generalized Post-filter and a Novel Perceptual Filter. ICSP International Conference on Signal Processing, pp.370-373, 2008

[24] J. Cho, A. Krishnamurthy. Speech Enhancement using Microphone Array in Moving Vehicle Environment. Proceedings of IEEE Intelligent Vehicles Symposium, pp. 366-371, 2003

[25] Y. D. Cho, 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.

[26] I. Cohen. Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging. IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 5, pp. 466-475, 2003

[27] D. V. Compernolle. Switching Adaptive filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol.2, pp.833-836, 1990

[28] C. Costa, S. L. don, B. G. Aguiar Neto. An Evaluation of an Adaptive Multichannel System for Speech Enhancement with Automatic Phase Alignment. International Conference on Acoustics, Speech, and Signal Processing, Vol.1, pp.844-847, 1995

[29] M. Dahl, I. Claesson, S. Nordebo. Simultaneous Echo Cancellation and Car Noise Suppression Employing a Microphone Array. International Conference on

Acoustics, Speech, and Signal Processing, Vol.1, pp. 239-242, 1997

[30] G. H. Ding, T. Huang, B. Xu. Suppression of Additive Noise Using a Power Spectral Density MMSE Estimator. IEEE Signal Processing Letters, Vol. 11, No.6, pp. 585-588, 2004

[31] S. Doclo, M. Moonen. GSVD-Based Optimal Filtering for Single and Multimicrophone Speech Enhancement. IEEE Transactions on Signal Processing, Vol.50, No.9, 2002

[32] Y. Ephraim et al. Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator. IEEE Transactions on Speech and Audio Processing, Vol. 33, No. 2, pp. 443-445, 1985

[33] Y. Ephraim, H. L. Van Trees. A Signal Subspace Approach for Speech Enhancement. IEEE Transactions on Speech and Audio Processing.Vol.3, No.4, pp. 251-266, 1995

[34] T. H. Falk, W. Chan. Single-ended Speech Quality Measurement Using Machine Learning Methods. IEEE Transactions on Audio, Speech and Language Processing, Vol. 14, No. 6, pp. 1935-1947, 2006

[35] S. Fischer, K. D. Kammeyer. Broadband Beamforming with Adaptive Postfiltering for Speech Acquisition in Noisy Environments, IEEE International Conference on Acoustics, Speech and Signal Processing, Vol.1, pp.359-362, 1997

[36] O. L. Frost. An Algorithm for Linearly Constrained Adaptive Array Processing. Proceedings of IEEE, Vol. 60, No.8 pp. 926-935, 1972.

[37] G. L. Fudge, D. A. Linebarger. A calibrated Generalized Sidelobe Canceller for Wideband Beamforming. IEEE Transactions on Signal Processing, Vol. 42, No. 10, pp. 2871-2875, 1994

[38] S. Gannot, I. Cohen. Speech Enhancement Based on the General Transfer Function GSC and Postfiltering. IEEE Transactions on Speech and Audio Processing,, Vol. 12, No. 6, pp. 561-571, 2004

[39] S. Gannot, D. Burshtein and E. Weinstein. Analysis of the Power Spectral Deviation of the General Transfer Function GSC. IEEE Transactions on Signal Processing, Vol. 52, No. 4, pp. 1115-1121, Apr. 2004.

[40] Z. Goh, K. Tan, B. Tan. Postprocessing Method for Suppressing Musical Noise Generated by Spectral Subtraction. IEEE Transactions on Speech and Audio Processing, Vol.6, No.3, pp.287-292, 1998

[41] N. Grbic, S. Nordholm. Soft Constrained Subband Beamforming for Handsfree Speech Enhancement. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, pp. 885-888, 2002.

[42] N. Grbic, S. Nordholm, A. Cantoni. Optimal FIR Subband Beamforming for Speech Enhancement in Multipath Environments. IEEE Signal Processing Letters, Vol.10, No.11, pp. 335-338, 2003

[43] J. E. Greenberg. Modified LMS algorithms for speech processing with an adaptive noise canceller. IEEE Transactions on Speech and Audio Processing, Vol. 6, No. 4, pp. 338-350, 1998.

[44] J. E. Greenberg, P. M. Zurek, Microphone-array Hearing Aids, in M. Brandstein and D. Ward eds Microphone Arrays: Signal Processing Techniques and Application, Springer, Berlin, pp.229-253, 2001

[45] L. J. Griffiths, C.W. Jim. An Alternative Approach to Linearly Constrained Adaptive Beamforming", IEEE Transactions on Antennas and Propagation, Vol. 30, No. 1, pp. 27-34, 1982.

[46] C. Guan, Y. Chen, B. Wu. Direct Modulation on LPC Coefficients with Application to Speech Enhancement and Improving the Performance of Speech Recognition in Noise. Proceedings of International Conference on Acoustics, Speech and Signal Processing, Vol.2, pp107-110, 1993

[47] H. Gustafsson, S. E. Nordholm and I. Claesson. Spectral Subtraction Using Reduced Delay Convolution and Adaptive Averaging. IEEE Transactions on Speech and Audio Processing, Vol. 9, No.8, pp. 799-807, 2001

[48] S. Haykin. Adaptive Filter Theory, 3rd ed. Englewood Cliffs, NJ: Prentice Hall, 1996

[49] S. S. Haykin, J. H. Justice. Array Signal Processing. Englewood Cliffs, NJ: Prentice-Hall, 1985

[50] M. Harteneck, S. Weiss, and R. W. Stewart. Design of Near Perfect Reconstruction Oversampled Filter Banks for Subband Adaptive Filters. IEEE Transactions on Circuits and Systems, Vol. 46, pp. 1081-1086, 1999.

[51] W. Herbordt, W. Kellermann. Efficient Frequency-domain Realization of Robust Generalized Sidelobe Cancellers. Proceedings of IEEE fourth Workshop on Multimedia Signal Processing, pp. 377-382, 2001

[52] W. Herbordt, W. Kellermann. Analysis of Blocking Matrices for Generalized Sidelobe Cancellers for Non-stationary Broadband Signals. Proceeding of IEEE International Conference on Acoustics, Speech, Signal Processing, Vol.4, p.4178, 2002

[53] O. Hoshuyama, A. Sugiyama, A. HiraNo. A Robust Adaptive Beamformer for Microphone Arrays with a Blocking Matrix Using Constrained Adaptive Filters. IEEE Transactions on Signal Processing, Vol. 47, No. 10, pp. 2677-2684, 1999.

[54] H. T. Hu, F. J. Kuo, H. J. Wang. Supplementary Schemes to Spectral Subtraction for Speech Enhancement. Speech Communication, Vol.36, pp. 205-218, 2002

[55] Y. Hu, P. Loizou. A perceptually Motivated Approach for Speech Enhancement. IEEE Transactions Speech and Audio Processing, Vol.11, No.5, pp. 457-465, 2003

[56] Y. Hu, P. Loizou. A Subspace Approach for Enhancing Speech Corrupted by Colored Noise. IEEE Signal Processing Letters, Vol.9, No.7, pp. 204-206, 2002

[57] Y. Hu, P. Loizou. Subjective Comparison of Speech Enhancement Algorithms. Proceedings of IEEE International Conference on Acoustics., Speech and Signal Processing, Vol. 1, pp. 153-156, 2006

[58] Y. Hu, P. Loizou. Subjective Comparison and Evaluation of Speech Enhancement

Algorithms. Speech Communication, Vol. 49, pp. 588-601, 2007

[59] R. Huber. Objective Assessment of Audio Quality Using an Auditory Processing Model. Ph.D. thesis, University of Oldenburg, Oldenburg, 2003.

[60] A. Hussain. Multi-sensor Adaptive Speech Enhancement Using Diverse Sub-band Processing. International Journal of Robotics & Automation, Vol.15, No.2, pp. 78-84, 2000

[61] S. Jongseo, N. S. Kim, and S. Wonyong. A Statistical Model-based Voice Activity Detection. IEEE Signal Processing Letters, Vol. 6, No.1, pp.1-3, 1999

[62] Y. Kaneda, J. Ohga. Adaptive Microphone-array System for Noise Reduction. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.34, No.6, pp. 1391-1400, 1986

[63] S. Kang, Y. Xiao, Z. Qiu. A Novel Adaptive Noise canceller with Master-slave Structure. IEEE International Conference on Electronics, Circuits and Systems, Vol.1, pp. 487-490, 2000

[64] H. K. Solvang, K. Ishizukabhgfh, M. Fujimoto. Voice Activity Detection Based on Adjustable Linear Prediction and GARCH Models. Speech Communication, 50, pp. 476-486, 2008

[65] W. Kim, S. Kang, H. Ko. Spectral Subtraction Based on Phonetic Dependency and Masking Effect. IEE Proc.-Vis. ISP, Vol.147, No.5, pp. 423-427, 2000

[66] C. H. Knapp, G. C. Carter. The Generalized Correlation Method for Estimation of Time Delay. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-24, No.4, pp. 320-327, 1976

[67] S. M. Kuo, W. M. Peng. Asymmetric Crosstalk-Resistant Adaptive Noise Canceller. Proceedings of IEEE Workshop on Signal Processing Systems, pp. 605-614, 1999

[68] R. H. Kwong, E. W. Johnston. A Variable Step-size LMS Algorithm. IEEE Transactions on Signal Processing, Vol. 40, pp. 1633-1642, 1992.

[69] F. Ling. Numerically Robust LS Lattice-Ladder Algorithms with Direct Updating of Reflection Coefficients. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 34, No.4, pp. 837-845, 1986

[70] F. Ling. Givens Rotation Based Least Squares Lattice and Related Algorithms. IEEE Transactions on Signal Processing. Vol. 39, No.7, pp. 1541-1551, 1991

[71] W. Liu, S. Weiss, L. Hanzo. A Novel Method for Partially Adaptive Broadband Beamforming. Proceedings of IEEE Workshop on Signal Processing Systems, pp. 361-372, 2001

[72] W. Liu, S.Weiss, L. Hanzo. Subband Adaptive Generalized Sidelobe Canceller for Broadband Beamforming. Proceedings of IEEE Workshop on Statistical Signal Processing, pp. 591-594, 2001

[73] T. Liu, L. Wang, B. Xu, A. Xie, H. Zhang. Adaptive Noise Canceler and its Applications for Systems with Time-variant Correlative Noise. Proceedings of the 4th World Congress on Intelligent Control and Automation, Vol.2, pp. 1412 - 1415, 2002

[74] P. C. Loizou. Speech Enhancement: Theory and Practice. CRS Press, Boca Raton, 2007

[75] P. C. Loizou. Speech Enhancement Based on Perceptually Motivated Bayesian Estimators of the Magnitude Spectrum. IEEE Transactions on Speech and Audio Processing, Vol.13, No.5, pp. 857-869, 2005

[76] P. Lockwood, J. Boudy. Experiments with a Nonlinear Spectral Subtraction (NSS), Hidden Markov Models and Projection for Robust Recognition in Cars. Speech Communication, Vol. 20, No. 3, pp. 215-228, 1992

[77] S. Y. Low, S. Nordholm, R. Togneri. Convolutive Bind Signal Separation with Post-processing. IEEE Transactions on Speech and Audio Processing, Vol. 12, No. 5, pp. 439-548, 2004.

[78] G. Madhavan, H. D. Bruin. Crosstalk Resistant Adaptive Noise Cancellation.

Annals of Biomedical Engineering, Vol. 18, pp. 57-67, 1990

[79] C. Marro, Y. Mahieux, K. U. Simmer. Analysis of Noise Reduction and Dereverberation Techniques Based on Microphone Arrays with Postfiltering. IEEE Transactions on Speech and Audio Processing, Vol.6, No.3, pp. 240-259, 1998

[80] R. Martin. Small Microphone Arrays with Postfilters for Noise and Acoustic Echo Reduction. M. Brandstein and D. Ward eds. Microphone Arrays, Springer-Verlag, pp. 255-276, 2001

[81] M. Marzinzik, B. Kollmeier. Speech Pause Detection for Noise Spectrum Estimation by Tracking Power Envelope Dynamics. IEEE Transactions on Speech and Audio Processing, Vol. 10, pp. 109-118, 2002

[82] I. A. McCowan, H. Bourlard. Microphone Array Post-filter for Diffuse Noise Field. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol.1 , pp.905 -908, 2002

[83] I. A. McCowan, H. Bourlard. Microphone Array Post-filter Based on Noise Field Coherence. IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 6, 2003

[84] J. Meyer, K. U. Simmer. Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 1167- 1170, 1997

[85] G. Mirchandani, R. C. Gaus Jr., L. K. Bechtel. Performance Characteristics of a Hardware Implementation of the Cross-talk Resistant Adaptive Noise Canceller. Proceedings of the IEEE international Conference on Acoustics, Speech and Signal Processing, pp. 93-96, 1986

[86] G. Mirchandani, R. L. Zinser, J. B. Evans. A New Adaptive Noise Cancellation Scheme in the Pressence of Crosstalk. IEEE Transactions on Circuits and System, Vol.39, No.10, pp. 681-694, 1992

[87] U. Mittal, N. Phamdo. Signal/noise KLT Based Approach for Enhancing Speech

Degraded by Colored Noise. IEEE Transactions on Speech and Audio Processing, Vol. 8, pp. 159-167, 2000.

[88] T. J. Moir, J. F. Barrett. A Kepstrum Approach to Filtering, Smoothing, and Prediction with Application to Speech Enhancement. Proc. Royal Soc. London, A, Vol. 459, No. 2040, pp.2957-2976, 2003

[89] T. J. Moir. A z-domain transfer function solution to the non-minimum phase acoustic beamformer. Inter. Journal Systems Science, Vol.38, No.7, pp.563-575, 2007

[90] P. Mowlaee, A. Sayadiyan. Performance Evaluation for Transform Domain Model-based Single-channel Speech Separation. IEEE/ACS International Conference on Computer Systems and Applications, pp.935-942, 2009

[91] R. Mucci. Comparison of Efficient Beamforming Algorithms, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 32, No. 3, pp. 548-558, 1984

[92] W. H. Neo, B. Farhang-Boroujeny. Robust Microphone Arrays Using Subband Adaptive Filters. IEE Proc.-Vision, Image Signal Processing, Vol. 149, No. 1, pp. 17-25, 2002.

[93] S. Nordholm, I. Claesson, B. Bengtsson. Adaptive Array Noise Suppression of Handsfree Speaker Input in Cars. IEEE Transactions on Vehicular Technology, Vol. 42, No.4, pp. 514 - 518, 1993

[94] A. H. Nuttall, G. C. Carter. Spectral Estimation Using Combined Time and Lag weighting. Proceedings of IEEE, Vol. 70, No. 9, pp. 1115-1125, 1982.

[95] S. Ouyang, Z. Bao, G. Liao. Robust Recursive Least Squares Learning Algorithm for Principal Component Analysis. IEEE Transactions on Neural Networks, Vol.11, No.1, pp. 215-221, 2000

[96] V. Parsa, P. A. Parker, R. N. Scott. Performance Analysis of a Crosstalk Resistant Adaptive Noise Canceller. IEEE Transactions on Circuits and System, Vol. 43, No.7, pp. 473-481, 1996

[97] P. M. Peterson. Simulating the Response of Multiple Microphones to a Single Acoustic Source in a Reverberant Room. Journal of Acoustics Soc. America, Vol. 80, No. 5, pp. 1527-1529, 1986.

[98] C. Plapous, C. Marro, L. Mauuary, P. Scalart. A Two-step Noise Reduction Technique. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, pp. 289-292, 2004.

[99] J. Poruba. Speech Enhancement Based on Nonlinear Spectral Subtraction. Proceedings of the Fourth IEEE International Caracas Conference on Devices, Circuits and Systems, pp. T031-1-T031-4, 2002

[100] M. H. Radfar, R. M. Dansereau. Single-Channel Speech Separation Using Soft Mask Filtering. IEEE Transactions on Audio, Speech and Language Processing, Vol.15, No. 8, pp2299-2310, 2007

[101] J. Ramirez, J. C. Segura, M. C. Benitez, A. Torre, A. Rubio. A New Kullback-Leibler VAD for Speech Recognition in Noise. IEEE Signal Processing Letters., Vol. 11, No. 2, pp. 666-669, 2004

[102] A. Rezayee, S. Gazor. An Adaptive KLT Approach for Speech Enhancement. IEEE Transactions on Speech and Audio Processing, Vol. 9, pp. 87-95, 2001

[103] J. G. Ryan. Criterion for the Minimum Source Distance at which Plane-wave Beamforming can be Applied. Journal of Acoustic Society of America, Vol. 104, No. 1, pp. 595-598, 1998.

[104] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, K. ShikaNo. Blind Signal Separation Based on a Fast-convergence Algorithm Combining ICA and Beamforming. IEEE Transactions on Audio, Speech and Language Processing, Vol. 14, No. 2, pp. 666-678, 2006.

[105] H. Sawada, S. Araki, R. Mukai, and S. MakiNo. Blind Extraction of a Dominant Source Signal from Mixtures of Many Sources. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 3, pp. 61-64, 2005

[106] M. Schroeder. Models of hearing. Proceedings of IEEE, Vol.63, pp:1332-1350, 1975

[107] K. U. Simmer, A. Wasiljeff. Adaptive Microphone Arrays for Noise Suppression in the Frequency Domain. Second Cost 229 Workshop on Adaptive Algorithms in Communications, Bordeaux, France, pp. 185-194, 1992

[108] P. Smaragdis. Efficient Blind Separation of Convolved Sound Mixtures. Proceedings of IEEE Apps. of Signal Processing to Audio and Acoustics, pp. 19-22, 1997

[109] A. Spriet, M. Moonen, J. Wouters. Stochastic Gradient-based Implementation of Spatially Preprocessed Speech Distortion Weighted Multichannel Wiener Filtering for Noise Reduction in Hearing Aids. IEEE Transactions on Speech and Audio Processing, Vol.13, No.4, pp.487-503, 2005

[110] P. Sun. Speech Enhancement Using Microphone Array. Master Degree Thesis, University of Auckland, Auckland, New Zealand, 2005

[111] S. Takada, S. Kanba, T. Ogawa, K. Akagiri, T. Kobayashi. Sound Source Separation Using Null-Beamforming and Spectral Subtraction for Mobile Devices. Proceedings of IEEE workshop on Applications of Signal Processing to Audio and Acoustics, pp.30-33, 2007

[112] S. G. Tanyer, H. Ozer. Voice Activity Detection Using Cepstral Features. IEEE Transactions s on Speech and Audio Processing, Vol.8, No.4, pp. 478-482, 2000

[113] L. Thorpe, W. Yang. Performance of Current Perceptual Objective Speech Quality Measures. Proceedings of IEEE Speech Coding Workshop, pp. 144-146, 1999

[114] R. M. Udrea, S. Ciochina. Speech Enhancement using Spectral Oversubtraction and Residual Noise Reduction. International Symposium on Signals, Circuits and Systems, Vol.1 , pp.165 - 168, 2003

[115] B. D. Van Veen, R. A. Roberts. Partially Adaptive Beamforming Design via Output Power Minimization. IEEE Transactions on Acoustics, Speech and Signal

Processing, Vol. 35, pp. 1524-1532, 1987

[116] B. D.Van Veen, K. M. Buckley. Beamforming: A Versatile Approach to Spatial Filtering. IEEE ASSP Magazine, Vol.5, No.2, pp. 4-24, 1988

[117] P. P. Vaidyanathan, Multirate Systems and Filter Banks. Englewood Cliffs, NJ: Prentice Hall, 1993

[118] N. Virag. Signal Channel Speech Enhancement Based on Masking Properties of Human Auditory System. IEEE Transactions on Speech and Audio Processing, Vol. 28, No. 3, pp. 126-137, 1999

[119] N. Wang, D. Zheng, S. Xu, S. Zhang. New Algorithm for Speech Enhancement Using Wavelet Packet Transform Based on Auditory Mode. International Conference on Computer Science and Software Engineering, Vol.4, pp.1000-1003, 2008

[120] Y. Y. Wang, W. H. Fang. Wavelet-based Broadband Beamformers with Dynamic Subband Selection. IEICE Transactions on Communication, Vol. E83B, No. 4, pp. 819-826, 2000

[121] E. Warsitz, R. Haeb-Umbach. Blind Acoustic Beamforming Based on Generalized Eigenvalue Decomposition, IEEE Transactions on Audio, Speech and Language Processing., Vol. 15, No. 5, pp. 1529-1539, 2007

[122] E. Warsitz, A. Krueger, R. Haeb-Umbach. Speech Enhancement with a New Generalized Eigenvector Blocking Matrix for Application in a Generalized Sidelobe Canceller. IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vagas, US, pp. 73-76, 2008

[123] C. T. Weibel, G. Fischer. Delay-sum Beamforming Using Delta-sigma Modulated Inputs. Proceedings of IEEE Midwest Symposium on Circuits and Systems, Vol. 3, pp.1198-1201, 2000

[124] S. Weiss, R. W. Stewart, M. Schabert, I. K. Proudler, and M. W. Hoffman. An Efficient Scheme for Broadband Adaptive Beamforming. Conference Record of Thirty-third Asilomar Conference on Signals, Systems and Computers, Vol.1, pp.

496-500, 1999

[125] S. Weiss, R.W. Stewart. Fast Implementation of Oversampled Modulated Filter Banks. IEE Electronics Letters., Vol. 36, No. 17, pp. 1502-1503, 2000

[126] G. Whipple. Low Residual Noise Speech Enhancement Utilizing Time-Frequency Filtering. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol.1, pp. 5-8, 1994

[127] B. Widrow et al. Adaptive Noise Canceling: Principles and Applications. Proceedings of IEEE, Vol. 63, No. 12, pp. 1692-1716, 1975

[128] B. Widrow, K. M. Duvall, R. P. Gooch, W. C. Newman. Signal Cancellation Phenomena in Adaptive Antennas: Causes and Cures. IEEE Transactions on Antennas and Propagation, Vol. 30, No.3, pp. 469-478, 1982.

[129] B. Widrow, S. Stearns. Adaptive Signal Processing. Englewood Cliffs, NJ: Prentice Hall, 1985.

[130] H. Yang, M. A. Ingram. Design of Partially Adaptive Arrays Using the Singular-value Decompostion. IEEE Transactions on Antennas and Propagation, Vol. 45, No.5, pp. 843-850, 1997.

[131] R. Zelinski. A microphone Array with Adaptive Post-filtering for Noise Reduction in Reverberant Rooms. Proceedings of International Conference on Acoustics, Speech and Signal Processing. Vol. 5, pp. 2578-2581, 1988

[132] R. Zelinski. Noise Reduction Based on Microphone Array with LMS Adaptive Post-filtering. IEEE Electronics Letters, Vol.26, No.24, pp.2036-2037, 1990

[133] Y. R. Zheng, R. A. Goubran, M. El-Tanany. A Nested Sensor Array Focusing on Near Field Targets. Proceedings of IEEE Sensors Conference, Toronto, Canada, pp. 843-848, 2003

[134] Y. R. Zheng. Spatial-temporal Subband Beamforming for Near Field Adaptive Array Processing. Ph.D dissertation, Carleton University, Ottawa, Canada, 2002.

[135] Y. R. Zheng, R. A. Goubran, M. El-Tana. Experimental evaluation of a nested microphone array with adaptive noise cancellers. IEEE Transactions on Instrumentation and Measurement, Vol.53, No.3, pp. 777-786, 2004

[136] R. L. Zinser, G. Mirchandani, J. B. Evans. Some Experimental and Theoretical Results Using a New Adaptive Filter Structure for Noise Cancellation in the Presence of Cross-talk. Proceedings of International Conference on Acoustics, Speech and Signal Processing, Vol.3, pp. 1253-1256, 1985

[137]http://mixguides.com/microphones/product_features/audio_audiotechnica_atdspcontrolled_adaptivearray

[138] http://www.chinaunicom.com.cn/profile/xwdt/txjs/file1523.html

[139] http://www.fortemedia.com

[140] http://www.microsoft.com/whdc/device/audio/micarrays.mspx

[141] http://www.smarthome.duke.edu/downloads/microphone_arrays.pdf

[142] http://www.zxbc.cn/html/20070511/16641.html

[143] http://www-isl.stanford.edu/~widrow/papers/j2001amicrophone.pdf