SPEECH ENHANCEMENT USING FIBER ACOUSTIC SENSOR

Miao Wang

A thesis
in
The Department
of
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Applied Science
Concordia University
Montréal, Québec, Canada

April 2020

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Miao Wang

Entitled: Speech Enhancement Using Fiber Acoustic Sensor

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Electrical and Computer Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. J. Cai

_____ External Examiner
Dr. R. Sedaghati (MIAE)

_____ Internal Examiner
Dr. J. Cai

_____ Supervisor
Dr. W.-P. Zhu

Approved by: _____
Dr. Y.R. Shayan, Chair
Department of Electrical and Computer Engineering

_____ 20___

_____
Dr. Amir Asif, Dean,
Faculty of Engineering and Computer Science

Abstract

Speech Enhancement Using Fiber Acoustic Sensor

Miao Wang

With the development of IoT (Internet of Things) services and devices, the voice command becomes a more and more important tool for human computer interaction. However, the audio signal recorded by the conventional omni-directional microphone is easy to be corrupted by the environmental noise like interference speech. Although the conventional beamforming techniques are able to point the main lobe of beam pattern at the desired speaker, it requires several omni microphones to form a microphone array, which will occupy large space on an IoT device. Many researchers are devoting their efforts to inventing a microphone of small size that can create directional beam pattern. Recently, researchers get inspirations from the spider's way to sense the acoustic wave. They invented a new small-size acoustic sensor made of spider silks. This acoustic sensor has a frequency-independent dipole beam pattern for wideband audio signal. Utilizing this fiber acoustic sensor, two compact microphone arrays and corresponding speech enhancement systems can be constructed. The first microphone array consists of one omni-microphone collocated with one fiber acoustic sensor. And the second one consists of two collocated fiber acoustic sensors with orthogonal dipole beam patterns.

By using the first microphone array, a first-order adaptive beamformer is designed in this thesis to reduce speech interference effects and separate speeches. In this design, an adaptive first-order beam pattern is formed by means of normalized least mean square method. Considering a scenario where the desired speech and interference speech are present at the same time, this adaptive beamformer is able to point the null angle of beam pattern at the undesired speaker to achieve speech interference reduction. In order to verify this idea, numerical simulations are conducted in an ideal condition (clean speech without reverberation) and real scenario (clean speech corrupted by white noise and reverberation). The results show that this design is able to improve speech quality significantly in ideal case. Under the condition suffering

from white noise and reverberation, the improvement is achieved as well but at a much smaller scale.

By using the second collocated microphone array, a speech enhancement system is proposed to make the collocated fiber acoustic sensors be able to capture the speech from any direction and suppress the white noise. This system includes three parts. The first part conducts DOA (direction of arrival) estimation empowered by a machine learning method. Here the inter-channel level difference is employed to compute raw DOA estimates in the presence of white noise and reverberation. After obtaining the raw DOA estimates, the machine learning method (generalized wrapped Gaussian mixture model) is used to give a more accurate DOA estimation. This proposed method is robust to both white noise and reverberation with a low computational complexity and solves the phase ambiguity problem between 0 and π for DOA estimation. In the second part, by using the orthogonality of the dipoles of the two collocated fiber acoustic sensors (one is $sin\theta$ and the other is $cos\theta$), along with the DOA (θ) estimated by the generalized wrapped Gaussian mixture model, a steerable dipole beam pattern is generated to point its main lobe at the speaker. In the third part, a noise reduction procedure is applied to the output signal of the steerable beamformer. The proposed method is based on a time-frequency mask, which is used to filter out time-frequency bins of white noise and keep those of speech signal. In order to verify the effectiveness of the designed system, numerical simulations are conducted in the existence of both white noise and reverberation. The result shows that the proposed DOA estimation method is robust to both white noise and reverberation. It implies that this type of microphone array is able to obtain precise speaker spatial information. Meanwhile, the audio quality of the output signal of this system is improved by at least 50%.

Acknowledgments

The work included in this thesis is completed under the help from many people. I would like to express my thankfulness here.

Firstly, I would like to show my appreciation to Prof. Wei-Ping Zhu, who is my supervisor, for his providing me with unique research opportunity and financial assistance. Under his guidance and encouragement, I explored the speech processing world and experienced the beauty behind science and technology. This journey will be a treasure in my lifetime. I am also grateful to him for getting me involved in his research projects. During this period, my knowledge about signal processing is extended and the soft skills are enriched.

Secondly, I would like to thank Frederic Lepoutre, Iman Moazzen, Stephane Leahy, Lucas Carneiro, Zhiheng Ouyang and Abrar Hussain. It is a good time to work with all of them. Thanks for their help and discussion with me, from which I have learned and benefited a lot. Here I should also give special thanks to Frederic Lepoutre for giving me technical insights about the knowledge in audio processing. I wish him a great success in his entrepreneurship.

Thirdly, I am also grateful to Concordia University for providing me with a great environment to study. All the staffs at Concordia are awesome. It is my fortunate period to conduct graduate study here.

At last, I would like to thank my family members (杨俊霞, 王福祥, 王玉荣, 王志忠 and 把明宇) for their love and support. My self-motivated power is endless and comes from their endless love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

With the increasing demand for recording high quality audio signals in IoT (Internet of Things) devices, modern audio system has shifted from single channel recording to multiple channel recording format like stereo type. Many multi-channel audio capturing systems are utilizing microphone array techniques to obtain high quality audio signal since these techniques provide more spatial information of speakers to enable a better speech processing even in a noisy environment.

1.1.1 Microphone Array Techniques

Researchers have developed different kinds of microphone arrays, which can be divided into two main categories [1]:

(a) Distributed arrays where microphone capsules are geometrically distributed.

(b) Collocated microphone arrays where the microphone capsules are arranged such that there is no time delay among the sounds reaching the capsules in the array.

Traditionally, microphone arrays are used for DOA estimation to localize sound sources [2]. According to recent studies [3], microphone arrays have the ability to enhance audio quality, remove background noise and even separate speeches.

The Uniform Linear Microphone Array

The uniform linear array is the most common microphone array in literature. In this type of microphone array, several omni microphones are arranged along a straight

line with a separation distance d between each other [1]. This arrangement is shown in Fig.1, where M omni microphones are distributed uniformly along a straight line. Given speech source signal $s(t)$, the audio signal received by the mth microphone in an uniform linear array is expressed in equation (1) below,

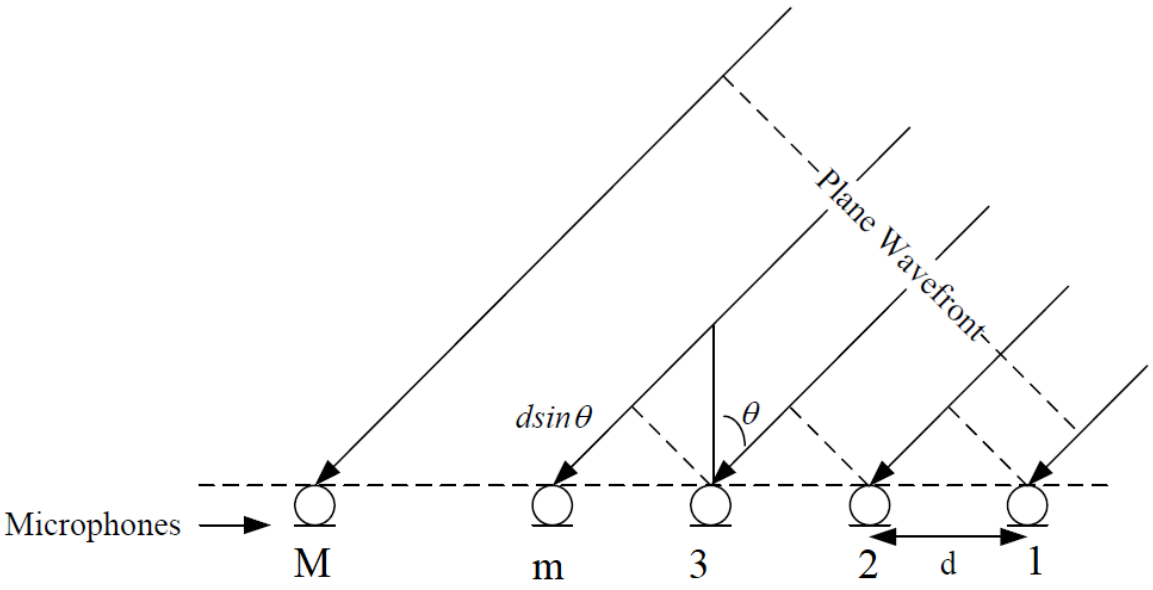

Figure 1: Uniform linear array [1]

$$x_m(t) = h_m(t) * s(t - \tau_m) + n_m(t) \tag{1}$$

where $x_m(t)$ represents the signal recorded by the mth omni microphone; h_m the impulse response from speech source to the mth microphone; $s(t - \tau_m)$ the delayed version of original speech $s(t)$ with τ_m representing the time delay; and $n_m(t)$ the noise existing in the mth omni microphone. As shown in Fig.1, the relationship between DOA θ and time delay τ is given by,

$$\tau = \frac{d \cdot sin\theta}{c} \tag{2}$$

where d denotes the distance between two adjacent microphones; c the sound speed in the air; and τ the time delay between signals received by two adjacent microphones.

Based on this signal model of uniform linear array where a time delay exists, the generalized cross correlation method [4] and its derivatives [5]–[10] are invented to localize sound sources, enhance speech quality and even separate speeches. Using the

generalized cross correlation, the time difference of arrival (TDOA) τ between two adjacent omni microphones $x_m(t)$ and $x_{m+1}(t)$ can be estimated by,

$$\tau^* = \text{argmax}\{\text{Real}\{R(\tau)\}\} = argmax\{\text{Real}\{\int_{-\infty}^{\infty} X_m^*(w)X_{m+1}(w)e^{jw\tau}dw\}\} \quad (3)$$

where $X_m(w)$ and $X_{m+1}(w)$ denote the Fourier transforms (FT) of $x_m(t)$ and $x_{m+1}(t)$ respectively; w is the angular frequency; $[\cdot]^*$ stands for the complex conjugate operation; and $R(\tau)$ denotes the cross correlation between $x_m(t)$ and $x_{m+1}(t)$.

This approach is able to give a reasonable result under the condition of low noise and reverberation level. Researchers find that using phase transformation with generalized cross correlation method is able to give a more accurate result that is also more robust to noise and reverberation [7]. This method is named as GCC-PHAT (generalized cross-correlation phase transform). Equation (4) shows how to estimate TDOA τ by applying the Fourier transforms of the $x_m(t)$ and $x_{m+1}(t)$ with unit magnitude,

$$\tau^* = \text{argmax}\{\text{Real}\{\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{|X_m^*(w)X_{m+1}(w)|} X_m^*(w)X_{m+1}(w)e^{jw\tau}dw\}\} \quad (4)$$

Besides the generalized cross correlation methods, another type of DOA estimation method is the steered power response (SPR) approach. This approach processes the audio signals recorded by all microphones together by using maximum likelihood estimation to obtain the maximum signal energy from a given direction. The delay and sum beamformer is one of the simplest beamforming methods using SPR [1]. For a uniform linear microphone array, the delay and sum beamformer is expressed below,

$$o(t) = \sum_{i=1}^{M} x_i(t - \tau_i) \quad (5)$$

where $x_i(t)$ is the signal recorded by ith microphone; τ_i is the time delay in relation to the reference microphone ($i = 1$); and $o(t)$ is the output signal from this delay and sum beamformer. However, the effectiveness of this beamformer is limited since it is not able to enhance the desired speech under the condition of moderate levels of noise and reverberation [1]. So based on this delay and sum beamformer, people invent an advanced version called filter and sum beamformer. Note that these beamformers could be used in both distributed and collocated microphone arrays. For a uniform

linear microphone array shown in Fig 1, the filter and sum beamformer is expressed as,

$$O(w) = \sum_{i=1}^{M} H_i(w) X_i(w) e^{-jw\tau(i-1)} \tag{6}$$

where $H_i(w)$ is the filter, $X_i(w)$ the signal received by the ith microphone in frequency domain; and τ is the TDOA between signals received by two adjacent microphones. Furthermore, the procedure to estimate the TDOA is simplified as finding an optimal value of τ to maximize the signal energy $\int_{-\infty}^{\infty} |O(w)|^2 dw$, namely,

$$\tau^* = \text{argmax} \int_{-\infty}^{\infty} |O(w)|^2 dw \tag{7}$$

where τ^* represents the optimal estimate of τ. Similarly with GCC-PHAT, an enhanced version of the filter and sum beaforming method is proposed in [11] and named as steered power response with phase transform (SPR-PHAT). Since then, it has become one of the most widely used algorithms for sound source localization. This enhanced method of TDOA estimation is expressed as,

$$\tau^* = \text{argmax} \sum_{i}^{M} \sum_{j}^{M} \text{Real}\{ \int_{-\infty}^{\infty} \frac{X_i^*(w) X_j(w) e^{jw\tau}}{|X_i^*(w) X_j(w)|} dw \} \tag{8}$$

This method is considered as one of the most robust DOA estimation methods as long as the number of microphones is large enough [1]. However, for the collocated microphone arrays like acoustic vector sensor, this method cannot work any more since there is no time delay between any pair of microphones in a collocated microphone array.

In order to minimize the computation cost for equation (8), in 2007, researchers provided a fast implementation using coarse-to-fine region contraction to locate sound sources [12]. After two years, the authors proposed a new method which uses stochastic particle filtering to localize sound sources with less computational cost [13]. The authors claim that this approach makes SPR-PHAT more practical for real-time applications [13]. In the work of [14], the authors proposed a method to localize multiple simultaneous talkers using SPR-PHAT. Another improvement for SPR method is given by [15], where the researchers claim that their method is fast enough for the development of real-time source localization applications.

There is a type of DOA estimation methods without dependency on the geometric structure of microphone array. This approach is named as spectral estimation based method [1]. Among many spectral estimation methods, MUSIC (multiple signal classification) and ESPRIT (estimation of signal parameters via rotation invariance techniques) are the top two methods used widely for DOA estimation.

MUSIC method utilizes eigenvalue decomposition to split output signals of microphone array into speech and noise parts [16]. It is proposed initially for narrow band signal processing like DOA estimation for radar signals. And then researchers in the field of audio processing borrow this idea and apply it to sound source localization. Since it requires eigenvalue decomposition, the computational cost of MUSIC is a little heavy. With the development in past several decades, many variants of MUSIC method are proposed like cyclostationary MUSIC [17], spatial smoothing MUSIC [18], [19] and root MUSIC algorithm [20]. The spatial smoothing MUSIC method is able to estimate DOA values even if the speeches are correlated [1]. Note that MUSIC method has the ability to do DOA estimation with both distributed and collocated microphone arrays.

Unlike MUSIC method, the ESPRIT as another spectral estimation based method needs to do exhaustive search among all possible steering vectors for DOA estimation [21], [22]. Furthermore, the ESPRIT method requires more eigenvalue decomposition operations and matrix manipulations, meaning that this method is of a large computational cost.

The Circular Microphone Array

Another type of commonly used distributed microphone array is the circular microphone array. As shown in Fig.2, there are M omni microphones distributed along a circle with radius q. Assuming that there are P active sound sources near the circular microphone array, the signal received at the mth microphone $x_m(t)$ is given below,

$$x_m(t) = \sum_{k=1}^{P} a_{mk} s_k(t - t_m(\theta_k)) + n_m(t) \tag{9}$$

where s_k is the kth sound source; a_{mk} and $t_m(\theta_k)$ are, respectively, the attenuation factor and time delay from the kth source to the mth microphone; θ_k is the DOA of source s_k observed with respect to the x-axis in Fig.2; and $n_m(t)$ is an additive white

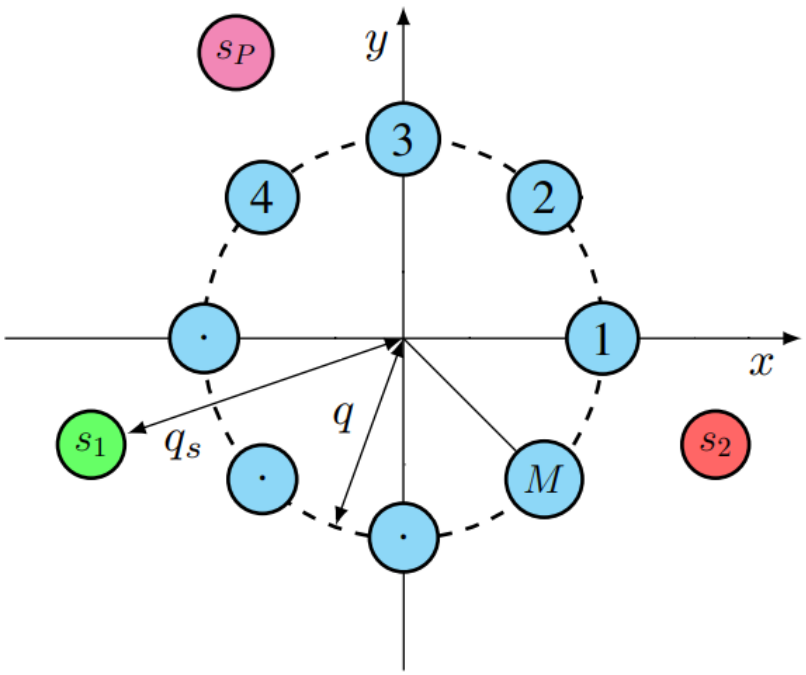

Figure 2: Circular microphone array [23]

noise at the mth microphone, which is uncorrelated with $s_k(t)$ and the white noises from other microphones.

Circular microphone array is commonly used for sound source localization. Compared to uniform linear microphone array, the circular array is able to contain more microphone capsules within a smaller space. For example, in paper [24], 32 omni microphones are arranged uniformly on a 0.5m diameter circle. And in paper [25], the authors constructed a circular microphone array by distributing 288 microphone capsules uniformly on a circle with radius of 1 meter. Since there is displacement between any pair of microphone capsules, the generalized cross correlation is adapted to estimate DOA using circular microphone array [26]. In paper [27], the authors proposed a DOA estimation method using circular microphone array, called circular integrated cross spectrum (CICS). Following this idea, the authors in papers [23], [28]–[30] have developed several real-time implementation schemes to localize multiple sound sources and determine the number of active sound sources. These techniques are able to achieve a promising result for applications like speech enhancement and speech separation.

The Spherical Microphone Array

The spherical microphone array has three-dimensional symmetry [31] and has been often used for spatial audio recordings. It is able to capture the sound information in three dimensional space as accurately as possible [1]. And it is also used to extract DOA information for beamforming and speech enhancement. In paper [32], the authors utilized the spherical microphone array to suppress reverberation. One big advantage of spherical microphone array is that it is able to contain large amount of microphones in a relatively small space [33]. The microphone capsule could be either omni microphone or cardioid microphone. And the position of each capsule is arranged randomly in paper [34] or optimally for specific purpose in paper [35].

The Acoustic Vector Sensor

The acoustic vector sensor is a type of collocated microphone array, which is able to capture audio signal with spatial information within a small space. This property enables such device to be deployed on a small mobile or embedded device. As one important type of collocated microphone array, acoustic vector sensor is able to record both sound pressure and particle velocity signal of sound waves simultaneously within a small size. Usually, it includes two or three directional microphones pointing at different orthogonal directions and one omni microphone capturing sound pressure signal. At the beginning, the acoustic vector sensor is used to detect electromagnetic waves and measure seismic data for underwater acoustic applications [1]. Researchers have done theoretical derivations of the performance of acoustic vector sensor [36]. They have shown a Cramer-Rao lower bound for localizing sound sources using acoustic vector sensor [36], [37]. This lower bound indicates that the performance of DOA estimation using acoustic vector sensor is better than those using other microphone arrays with comparable number of microphones and size [1].

The signal model of a 3D acoustic vector sensor is expressed in equation (10), under assumptions that the acoustic wave is traveling in a homogeneous space and

the signal is considered to be a plane wave at the location of the sensor [37].

$$\begin{pmatrix} s_o[n] \\ s_x[n] \\ s_y[n] \\ s_z[n] \end{pmatrix} = \begin{pmatrix} 1 \\ sin\phi cos\theta \\ sin\phi sin\theta \\ cos\phi \end{pmatrix} p[n] + \begin{pmatrix} w_o[n] \\ w_x[n] \\ w_y[n] \\ w_z[n] \end{pmatrix} \tag{10}$$

where $s_i[n], i \in \{o, x, y, z\}$ are called intensities in acoustic vector sensor, and represent, respectively, the signals received by the omni microphone, directional microphone along x axis, directional microphone along y axis and that along z axis; θ denotes the azimuth angle; ϕ denotes the elevation angle; $p[n]$ denotes the sound pressure signal; $w_o[n]$ denotes the additive white noise signal of omni microphone; $w_x[n]$,$w_y[n]$, and $w_z[n]$ denote the additive white noises of three corresponding directional microphones. In a simpler case where the directional microphone along z axis is omitted, the signal model of such arrangement is given below,

$$\begin{pmatrix} s_o[n] \\ s_x[n] \\ s_y[n] \end{pmatrix} = \begin{pmatrix} 1 \\ cos\theta \\ sin\theta \end{pmatrix} p[n] + \begin{pmatrix} w_o[n] \\ w_x[n] \\ w_y[n] \end{pmatrix} \tag{11}$$

In the following summary about acoustic vector sensor, only the methods for the structure with three directional microphones are demonstrated. However, the methods are also suitable for the structure with two directional microphones as long as a model simplification step is proceeded.

In real applications, in order to describe the reverberation of the room acoustic, the signal model of the acoustic vector sensor which considers the room impulse response is given as,

$$\begin{pmatrix} s_o[n] \\ s_x[n] \\ s_y[n] \\ s_z[n] \end{pmatrix} = \begin{pmatrix} (p * h_o)[n] \\ (p * h_x)[n] \\ (p * h_y)[n] \\ (p * h_z)[n] \end{pmatrix} + \begin{pmatrix} w_o[n] \\ w_x[n] \\ w_y[n] \\ w_z[n] \end{pmatrix} \tag{12}$$

where $*$ denotes the convolution operation; h_i, $i \in \{o, x, y, z\}$ represents the impulse response of corresponding microphone [37].

The methods for DOA estimation using acoustic vector sensor have been developed in the past decades. Among these methods, four methods are the most representative

ones. The first approach is based on the signal intensity difference between different channels in acoustic vector sensor [36]. The second one utilizes the velocity covariance matrix [36]. The third one is named as maximum power method [38]. And the fourth one is based on maximum likelihood estimation [39]. The common part among these four methods is that a vector $\mathbf{u} = [u_x, u_y, u_z]^T = [sin\phi cos\theta, sin\phi sin\theta, cos\phi]^T$, is estimated firstly, and then θ and ϕ are calculated by,

$$\begin{pmatrix} \phi \\ \theta \end{pmatrix} = \begin{pmatrix} tan^{-1}(\frac{\sqrt{u_x^2+u_y^2}}{u_z}) \\ tan^{-1}(\frac{u_y}{u_x}) \end{pmatrix} \tag{13}$$

The DOA estimation approach based on intensity difference is expressed as,

$$\begin{gathered} \tilde{\mathbf{u}} = \frac{\mathbf{v}}{||\mathbf{v}||} \\ \mathbf{v} = \frac{1}{N}\sum_{n=1}^{N} \begin{pmatrix} s_o[n] \cdot s_x[n] \\ s_o[n] \cdot s_y[n] \\ s_o[n] \cdot s_z[n] \end{pmatrix} \end{gathered} \tag{14}$$

where $\tilde{\mathbf{u}}$ is the estimate of $\mathbf{u}$; $\mathbf{v}$ is the mean value of the product between the pressure signal $s_o[n]$ and the vector containing signals received by the three directional microphones [37]. And the DOA estimation method using velocity covariance matrix applies the eigenvalue decomposition method to the covariance matrix to locate sound sources. This method is given by,

$$\begin{gathered} \mathbf{R} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{d_v}[\mathbf{n}]\mathbf{d_v^T}[\mathbf{n}] \\ \tilde{\mathbf{u}} : \text{Eigenvector corresponding to the largest eigenvalue of } \mathbf{R} \end{gathered} \tag{15}$$

where $\mathbf{d_v}[\mathbf{n}] = [s_x[n], s_y[n], s_z[n]]^T$; $\mathbf{R}$ is the covariance of the signals recorded by directional microphones. The third most representative approach is designed to search for one direction which is able to maximize the power of the first-order beamformer $\mathbf{t}[\mathbf{n}]$, which is given by [38],

$$\mathbf{t}[\mathbf{n}] = \alpha \cdot s_o[n] + (1-\alpha) \cdot \tilde{\mathbf{u}}_{\mathbf{steer}}^{\mathbf{T}} \mathbf{d_v}[\mathbf{n}] \tag{16}$$

where $\tilde{\mathbf{u}}_{\mathbf{steer}}^{\mathbf{T}}$ is the unit norm steering vector and $\alpha \in [0, 1]$ controls the weight between the omni-directional and directional sensor response [37]. The maximum power method finds the unit vector which has the largest beamforming response as the source

direction. This procedure is given by,

$$\begin{gathered}\tilde{\mathbf{u}} = \text{argmax } T(\tilde{\mathbf{u}}_{steer}) \\ \text{subject to } \tilde{\mathbf{u}}_{\mathbf{steer}}^{\mathbf{T}}\tilde{\mathbf{u}}_{\mathbf{steer}} = 1, \text{with } T(\tilde{\mathbf{u}}_{steer}) = \frac{1}{N}\sum_{n=1}^{N}\mathbf{t^2}[\mathbf{n}]\end{gathered} \tag{17}$$

Similar with the third approach discussed above, the fourth method also finds the estimate of vector $\mathbf{u}$ by solving an optimization problem. However, the objective function is changed as a probability density function, which is explained with details in paper [39].

1.1.2 Fiber Acoustic Sensor

As one part of acoustic vector sensor, the directional microphone plays an important role to the performance of DOA estimation and the recorded speech quality. Recently, researchers have invented a directional microphone with a frequency-independent beam pattern according to their investigations on spider silks. Spiders use their silks to weave a web. The web is very important to spiders. With sticky web, they not only can catch other small insects as food but also sense the environment according to the vibration of spider silks. More specifically, spider silks are sensitive enough to surrounding acoustic waves. Utilizing this special function of spider silks, several researchers in Binghamton University developed a new directional microphone made of spider silks [40]. They found that spider silks move with the same velocity of surrounding air [40]. Based on this finding, a new microphone prototype has been developed. This directional microphone is named as fiber acoustic sensor, which is one type of particle velocity microphone. In 2018, after some further investigations, the prototype microphone is refined with an instinct wideband directive beam pattern [41]. Furthermore, this device is able to sense acoustic velocity by fibers in a nano scale with a frequency-independent dipole beam pattern. With this unique acoustic property, this new microphone will benefit numerous applications like acoustic vector sensor that require directional microphones for sound detection, identification, and localization [41].

Components

Fiber acoustic sensor consists of the following four parts: fiber, chip, battery and magnetic field. The fiber is used to sense acoustic waves. Once the fiber starts

vibrating with the acoustic wave in magnetic field, a voltage signal is generated. This voltage signal is amplified and transmitted by the chip. Note that a battery is used to power the chip.

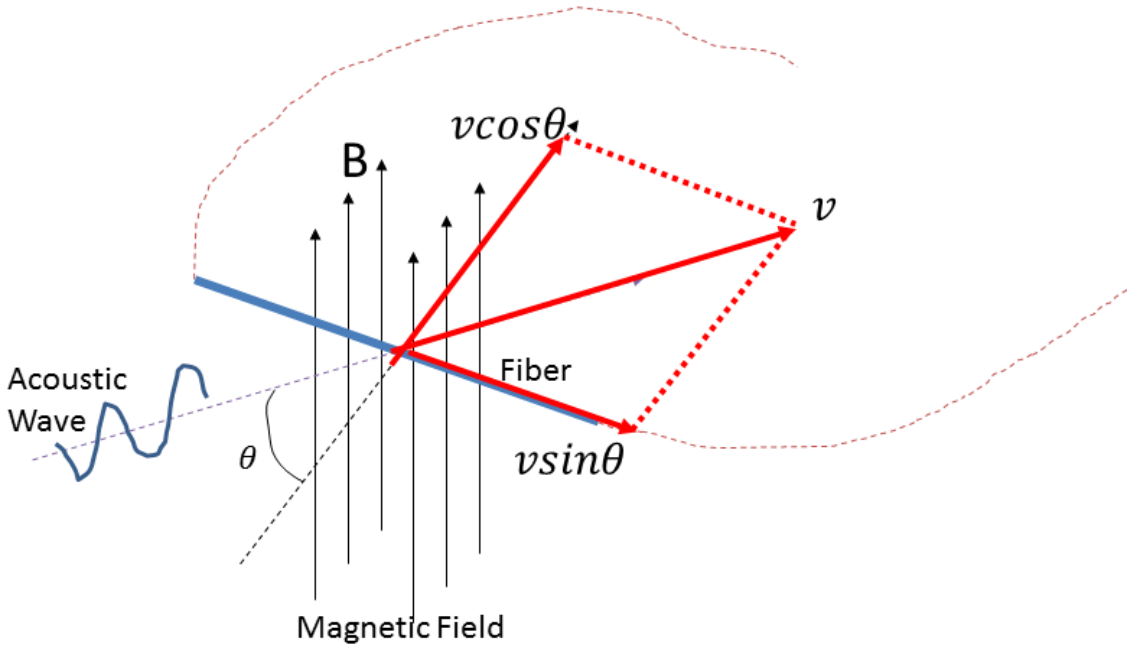

Figure 3: Fiber sensor mechanism

Mechanism

The core idea behind fiber acoustic sensor is based on the electromagnetic induction law. As shown in Fig.3, the acoustic wave propagates with a velocity v and a coming angle θ. Once the acoustic wave reaches the fiber, the fiber starts moving with a same velocity v. The velocity can be decomposed into 2 orthogonal components: $v \sin\theta$ (along the fiber) and $v \cos\theta$ (orthogonal to the fiber) as shown in Fig.3. Because this movement is located inside a magnetic field, a voltage between the two ends of fiber is generated according to the electromagnetic induction law. This weak voltage signal is detected and amplified by the chip. Then, the amplified voltage signal is transmitted to computer through a USB port. This is the entire process of sensing acoustic wave by the fiber sensor.

Furthermore, according to the electromagnetic induction law, this kind of sensing process could be quantified. Given the intensity of magnetic field B, the length of

fiber L, the velocity of acoustic wave $v(t)$ and speech arriving angle θ, the voltage signal $u(t)$ is determined by the following formula:

$$u(t) = B \cdot L \cdot v(t) \cdot \cos\theta \tag{18}$$

According to the equation above, the information about the direction of sound source is automatically encoded into the audio signal $u(t)$.

Beam Pattern

Note that the parameters B and L are constants in equation (18), and thus this equation can be simplified as,

$$u(t) = k \cdot v(t) \cdot \cos\theta$$

where k is a constant. Applying Fourier transform to both ends of the equation above, a beam pattern equation is obtained as given by,

$$B(\theta) = |\frac{U(e^{jw})}{V(e^{jw})}| = k \cdot |\cos\theta| \tag{19}$$

Note that the instinct beam pattern of the fiber sensor is dipole. Also it is frequency-independent since it is only impacted by the arriving angle of acoustic wave θ.

As shown in Fig.4, the beam pattern of fiber sensor is plotted under different frequencies. When the frequency changes from 100Hz to 10KHz, the dipole beam pattern doesn't change. However, it is observed that the fiber sensor has a good beam pattern (perfect dipole) in low frequency around 100Hz and the beam pattern gets a little distortion when frequency reaches around 10KHz. The slight degradation in directivity at very high frequencies can be removed by decreasing the fiber length [41]. In general, the instinct beam pattern of the fiber sensor is a frequency independent dipole. This property makes this type of sensor a promising product for high quality audio recording.

Directivity Factor of Fiber Sensor

Directivity factor is used to measure whether a microphone is more sensitive to particular directions. Given beam pattern $B(\theta)$ and a particular direction θ_0, it is calculated by,

$$G(\theta_0) = \frac{B^2(\theta_0)}{\frac{1}{\pi}\int_0^\pi B^2(\theta)d\theta} \tag{20}$$

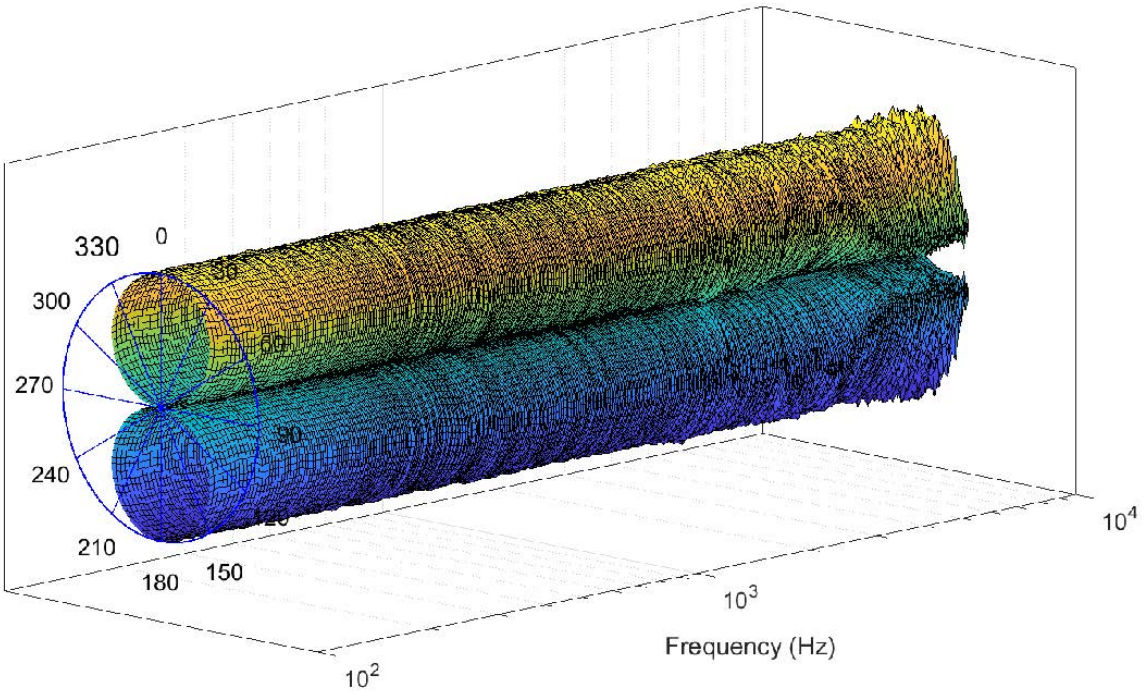

Figure 4: Dipole beam pattern of spider silk [41]

Usually people have more interests in the main lobe of beam pattern. So the directivity factor G [42] is also computed by,

$$G = \frac{\max\{B^2(\theta)\}}{\frac{1}{\pi}\int_0^\pi B^2(\theta)d\theta} \tag{21}$$

Since the beam pattern of fiber sensor is dipole, which can be expressed by $cos\theta$, then the term $\max\{cos^2(\theta)\}$ equals 1 and $\int_0^\pi cos^2(\theta)d\theta$ equals 0.5π. So the directivity factor of fiber sensor is calculated as 2.

Directivity Index of Fiber Sensor

The definition of directivity index $D(\theta_0)$ in specific direction θ_0 is given by,

$$D(\theta_0) = 10log_{10}(G(\theta_0)) \tag{22}$$

Given that θ_0 equals 0, the maximal directivity index of the fiber sensor is equal to 3.01.

Signal to Noise Ratio (SNR)

The definition of SNR here is based on the energy ratio between clean speech and noise signal, namely,

$$SNR = 10log_{10}(\frac{E_s}{E_n}) \tag{23}$$

where E_s denotes the energy of clean speech; and E_n denotes the energy of noise only audio. For the fiber acoustic sensor, since the noise here is the white noise generated by the circuits, there is no way to record a clean speech. In practice, the SNR is calculated by,

$$SNR = 10log_{10}(\frac{E_s - \sigma_n^2}{\sigma_n^2}) \tag{24}$$

where $E_s = \frac{1}{N}\sum_{i=1}^{N} s^2(i)$ denotes the average energy of noisy speech signal $s(n)$; and σ_n^2 denotes the variance of the white noise. For example, if the variance of the white noise in the circuits is 4.3023e-05 and the energy of noisy speech is 0.0043316, then the SNR of the fiber acoustic sensor is 19.9861 dB.

1.2 Objective of the Research

According to the discussion in the previous section, it is clear that the collocated microphone array like acoustic vector sensor is more suitable than those distributed microphone arrays to be deployed on a mobile or embedded device like cellphones and audio devices, due to its small size as opposed to distributed arrays. Usually, an acoustic vector sensor consists of four microphones: one omni-directional microphone and three directional microphones. It is obvious that a microphone array is able to be constructed within a smaller size by using less microphone capsules. Simplifying an acoustic vector sensor like removing several microphone capsules but retaining the functions of speech enhancement and DOA estimation is still an open question. There is a lack of investigations about this simplified version of acoustic vector sensors.

Meanwhile, the inventors of this fiber acoustic sensor have granted the corresponding technology to a startup company, which has applied for a patent about this novel directional microphone. The startup company aims to develop prototypes for speech enhancement by using this fiber acoustic sensor. In collaboration with the company,

the work done in this thesis is based on an NSERC (National Sciences and Engineering Research Council of Canada) Engage project which aims to develop a simpler collocated microphone array using the newly invented fiber acoustic sensor. More specifically, two simplified versions of acoustic vector sensor are investigated in this thesis: the first one is built by one omni microphone collocated with a fiber acoustic sensor; and the second one consists of double collocated fiber acoustic sensors that are pointing at orthogonal directions. For each array, the corresponding speech enhancement system is proposed. Required by the NSERC, results achieved in this thesis have been transferred to the company and are protected as the intelligent property of the company as well.

1.3 Organization of the Thesis

The rest of this thesis is organized as follows.

In chapter 2, a simplified version of acoustic vector sensor using one omni microphone and one fiber acoustic sensor is constructed. Inspired by the first-order adaptive differential microphone, this collocated microphone array is utilized to form an adaptive beamformer. This adaptive beamformer is designed to reduce the impact of speech interference during audio recording by achieving the beam pattern whose null angle always points at the interference speaker. Compared with the first-order differential microphone array which is used to obtain the first-order adaptive beamformer, the proposed collocated microphone array has the advantages of simpler structure and independence of frequency. Numerical simulations are conducted in an anechoic environment, showing that this method is able to suppress speech inference and improve the speech quality significantly. On the other hand, when considering the reverberation and white noise in the fiber acoustic sensor, our numerical simulations show that the effectiveness of this method for speech interference reduction encounters a big degradation. Further investigations about dereverberation and white noise removal are needed for this collocated microphone array.

In chapter 3, using double collocated fiber acoustic sensors, another collocated microphone array is constructed. Based on the signal intensity difference among these double directional microphones, a new method of calculating DOA estimates is proposed. However, limited by the symmetric property of dipole beam pattern, a

phase ambiguity problem between 0 and 180 degrees is introduced. In order to solve this problem, a histogram based wrapped Gaussian mixture model with less computational complexity is proposed to estimate DOA. After obtaining reliable DOA estimates, a steerable beamforming method is utilized here to point the main lobe of dipole beam pattern at the speaker and enhance the quality of recorded audio. This approach enables the collocated microphone array to record the speech from any direction. However, the white noise in fiber acoustic sensors corrupts the recorded audio quality a lot. In order to suppress the white noise in the output signal of steerable beamformer, a spectral subtraction method is proposed to reduce the negative impact of white noise on the audio quality. The aforementioned schemes constitute a comprehensive speech enhancement system using the collocated microphone array. In order to investigate the performance of this system under room reverberation and the condition of there being white noise in fiber acoustic sensors, a virtual acoustic environment is constructed. Numerical simulations are conducted inside the virtual acoustic room. Our results show that the proposed system is able to estimate DOA accurately and is robust to reverberation and white noise of fiber acoustic sensor. Furthermore, the PESQ (Perceptual Evaluation of Speech Quality) score, as the performance metric, of the output signal of this speech enhancement system is improved by at least 50%. It can be concluded that this speech enhancement system is able to boost the quality of recorded audio significantly by using the collocated double fiber acoustic sensors.

In chapter 4, the work of this thesis is summarized firstly. Then, several potential aspects of the future work are pointed out.

1.4 Contributions

The main contributions of this thesis are highlighted as follows.

1. A collocated microphone array which consists of one omni microphone and a fiber acoustic sensor, is designed to form an adaptive first-order beamformer to achieve speech interference reduction and speech separation.

2. A collocated microphone array which consists of two fiber acoustic sensors pointing at orthogonal directions, is designed to capture the speech from any direction and extract DOA information.

3. The wrapped Gaussian mixture model is generalized to handle angular values with any period. In particular, it can successfully estimate the probability distribution of angular values with a period of π.

4. A novel DOA estimation method is proposed by using the double collocated fiber acoustic sensors, which is robust to white noise and reverberation and solves the phase ambiguity problem between 0 and 180 degrees for DOA estimation.

5. A novel and simple method is developed by using the double collocated fiber acoustic sensors to enhance audio quality through suppressing the white noise in the circuits.

6. A speech enhancement system which consists of 3 procedures (DOA estimation, steerable beamformer and white noise suppression), is developed by using the double collocated fiber acoustic sensors. The numerical simulation results show that this system is able to enhance speech signal significantly.

7. One patent based on the results in this thesis has been applied for.

Chapter 2

Speech Enhancement using Collocated Fiber Sensor and Omni Microphone

As mentioned in the previous chapter, an acoustic vector sensor includes four microphones: one omni microphone and three directional microphones pointing at x, y and z axes respectively. This type of acoustic vector sensor has been fully investigated. However, there are not enough investigations about a simplified version of the acoustic vector sensor, which only consists of one omni microphone and one directional microphone. In this chapter, by using one omni microphone and a fiber acoustic sensor, a new collocated microphone array is constructed. More specifically, the knowledge about the first-order differential microphone array and the first-order adaptive differential microphone array is reviewed at first. Then, similar with the design of the first-order adaptive differential microphone array, an adaptive beam pattern is formed by using the proposed collocated microphone array. At last, numerical simulations are conducted to investigate the performance of the proposed microphone array for speech enhancement.

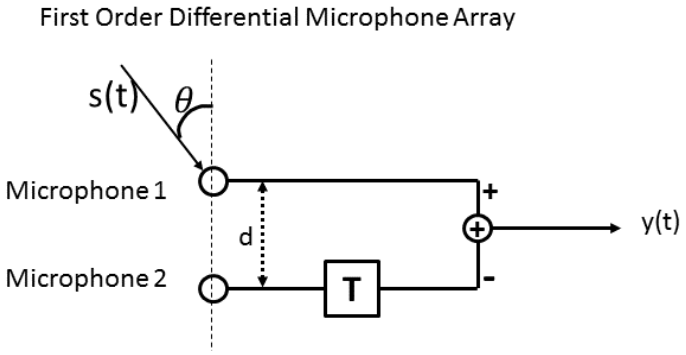

Figure 5: First-Order differential microphone array

2.1 Principle of First-Order Differential Microphone Array

In 2001, the first-order full band differential microphone array is invented by a team led by Gary Elko [43]. In their approach, two omni microphones are adopted to form the first-order differential array. And the first-order adaptive beamformer is performed by the normalized least mean square (NLMS) method [44].

2.1.1 The First-Order Differential Microphone Array

As shown in Fig.5, the first-order differential microphone array consists of two omni microphones and one time delay T. The distance between two microphones is d. An acoustic wave coming from angle θ can reach 2 omni microphones respectively. Due to the displacement d between these 2 omni microphones, there is a time delay generated among the audio signals recorded by these 2 microphones. The difference of the two received signals $y(t)$ is expressed as,

$$y(t) = s(t) - s(t - T - \frac{d\cos\theta}{c}) \tag{25}$$

where c represents the speed of sound; and T represents the time delay. By applying Fourier transform to both ends of equation (25), a directive pattern of this differential microphone is achieved, which is given by,

$$\frac{Y(e^{jw})}{S(e^{jw})} = 1 - e^{-jw(T+\frac{d\cos\theta}{c})} \tag{26}$$

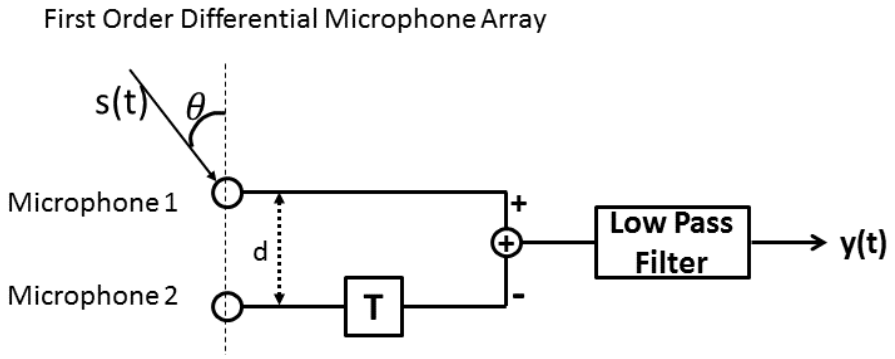

Figure 6: First-Order differential microphone array with low pass filter

where w denotes the frequency and j equals $\sqrt{-1}$. Note that the exponential term in equation (26) can be replaced with its Taylor series expansion. The definition about Taylor series expansion is given in Appendix A.1. Utilizing this series expansion, the beam pattern of this differential microphone is rewritten as,

$$\frac{Y(e^{jw})}{S(e^{jw})} = 1 - (1 - jw(T + \frac{d\cos\theta}{c}) + ...) \approx jw(T + \frac{d\cos\theta}{c}) \tag{27}$$

Note that this beam pattern is linearly dependent on the frequency w as seen in equation (27) above. In order to compensate for this frequency dependency, a fist-order low pass filter $\frac{1}{jw}$ is usually used. By applying this low pass filter, the beam pattern can be expressed as,

$$|\frac{Y(e^{jw})}{S(e^{jw})}| \approx |(T + \frac{d\cos\theta}{c})| \tag{28}$$

This design is shown in Fig.6, where the value of T can be specified to achieve different first-order beam patterns. Several possible values of T are listed in Table.1, where $\tau_0 = \frac{d}{c}$, to form the corresponding beam patterns. Meanwhile, with the help of low pass filter, the beam pattern becomes more flat along frequency axis. However, this approach of differential approximation is known to suffer from high white noise gain [39], since the gain of low pass filter $\frac{1}{jw}$ becomes infinity when w is close to 0. Furthermore, the accuracy of a spatial derivative requires that the distance between sensors to be small with respect to the acoustical wavelength [39], [45]. At high frequencies, however, this assumption fails, leading to distorted beam patterns [39].

Beam Pattern	Value of T
Dipole	0
Cardioid	τ_0
Hyper-Cardioid	$0.5\tau_0$
Super-Cardioid	$\frac{\sqrt{2}-1}{2-\sqrt{2}}\tau_0$

Table 1: Beam pattern of first-order differential microphone array [46]

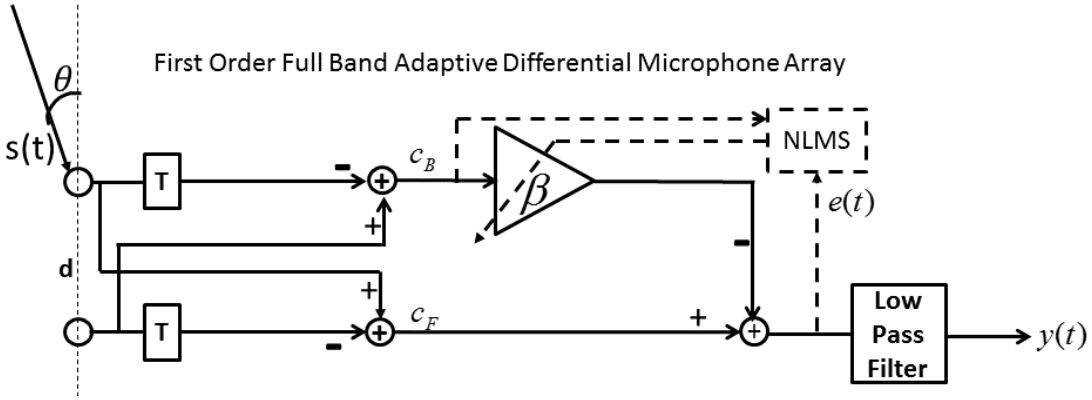

Figure 7: First-Order full band adaptive differential microphone array [43]

2.1.2 Adaptive First-Order Differential Microphone Array

Although the time delay T in Fig.6 provides the chance to generate different beam patterns, it is not suitable for a real-time implementation which requires time-varying beam patterns. Utilizing forward-facing and backward-facing cardioid, the research team led by Elko proposed a structure of first-order adaptive differential microphone array [43].

As shown in Fig.7, where T is a fixed time delay and its value is equal to $\frac{d}{c}$, this structure consists of three main parts: fixed beamformer, adaptive beamformer and low pass filter. The signal $c_F(t)$ (forward-facing cardioid) and $c_B(t)$ (backward-facing cardioid) are the outputs of the fixed beamformer. And the Fourier transforms of these two signals are as follows:

$$C_F(w,\theta) = jwT(1+cos\theta)S(e^{jw})$$

$$C_B(w,\theta) = jwT(1-cos\theta)S(e^{jw})$$

The derivation processes of the forward-facing and backward-facing cardioids,

which utilize the Taylor series expansion as well, are given in Appendixes A.2 and A.3 respectively. And these two signals ($c_F(t)$ and $c_B(t)$) are fed into the adaptive beamformer where a real value β_t is updated at each time step by the NLMS method to achieve the adaptive beamformer, which is given by,

$$e(t) = c_F(t) - \beta_t c_B(t)$$

$$\beta_t = \beta_{t-1} + \mu \frac{e(t)c_B(t)}{||c_B(t)||^2 + \delta}$$

where μ is the step size; and δ is a small constant. At last, since the cardioid beam pattern is generated by using differential methods, a low pass filter is required to compensate for the output signal from the adaptive beamformer.

2.2 Fiber Sensor with Omni Microphone Model

The main idea behind the first-order adaptive differential microphone array is using an adaptive filter to find an optimal beam pattern, whose null angle always points at the interference speaker.

Since the beam pattern of omni microphone is a constant 1 and that of fiber acoustic sensor is $\cos\theta$, which depends on the speech coming angle θ and is frequency independent, combining one fibre sensor with one omni microphone is able to achieve cardioids ($1 + \cos\theta$ and $1 - \cos\theta$) as the fixed beamformers discussed in section 2.1.2. So the fixed beamformers in section 2.1.2 could be replaced with one omni microphone and one fiber sensor. Based on this assumption, a collocated microphone array is designed in this section by using one omni microphone and one fiber acoustic sensor. Note that the instinct dipole beam pattern is generated by the fiber acoustic sensor, and there is no need to use a low pass filter to do compensation.

2.2.1 Components and Structure

In order to form an adaptive first-order beam pattern easily, the arrangement of the fiber sensor and the omni microphone is shown in Fig.8, where the omni microphone is at the center of fiber sensor. When the speech signal $s(t)$ comes from θ angle, it is recorded by the omni microphone and the fiber sensor at the same time. The signal recorded by omni microphone $s_m(t)$ should be the same as the original signal $s(t)$. And the signal recorded by fiber sensor $s_f(t)$ should be the product of $cos\theta$ and $s(t)$.

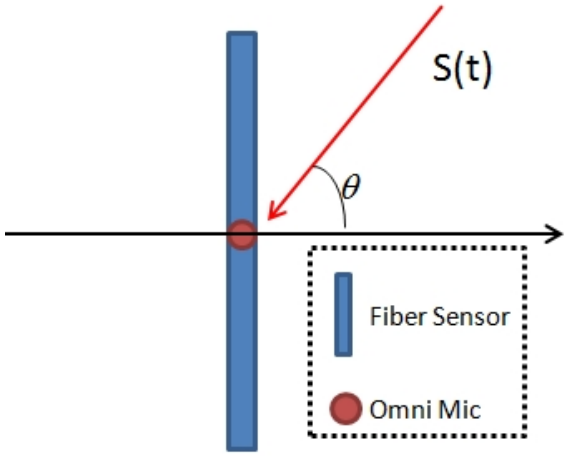

Figure 8: Arrangement of fiber sensor and omni microphone

So the beam patterns, forward facing cardioid $c_F(t,\theta)$ and backward facing cardioid $c_B(t,\theta)$, are computed by the equation (29) and (30) respectively,

$$c_F(t,\theta) = s_m(t) + s_f(t) = (1 + cos\theta) \cdot s(t) \tag{29}$$

$$c_B(t,\theta) = s_m(t) - s_f(t) = (1 - cos\theta) \cdot s(t) \tag{30}$$

where $s_m(t)$ represents the audio signal recorded by the omni microphone and $s_f(t)$ represents the audio signal recorded by the fiber acoustic sensor.

The signal flow graph in Fig.9 shows how an adaptive beamformer is constructed by using the omni microphone and fiber acoustic sensor. The NLMS method is used to find the optimal β in a time-varying situation. And β is a real value gain, which is updated for each time step t. The update formula of NLMS can be written as,

$$\beta_{t+1} = \beta_t + \mu \frac{y(t)c_B(t)}{||c_B(t)||^2 + \delta} \tag{31}$$

where μ is the step size; δ is a small constant to avoid division by zero; and $y(t)$ equals $c_F(t) - \beta c_B(t)$. Note that the beam pattern of the fiber sensor is independent of frequency, and thus there is no need to add a low pass filter for compensation before $y(t)$.

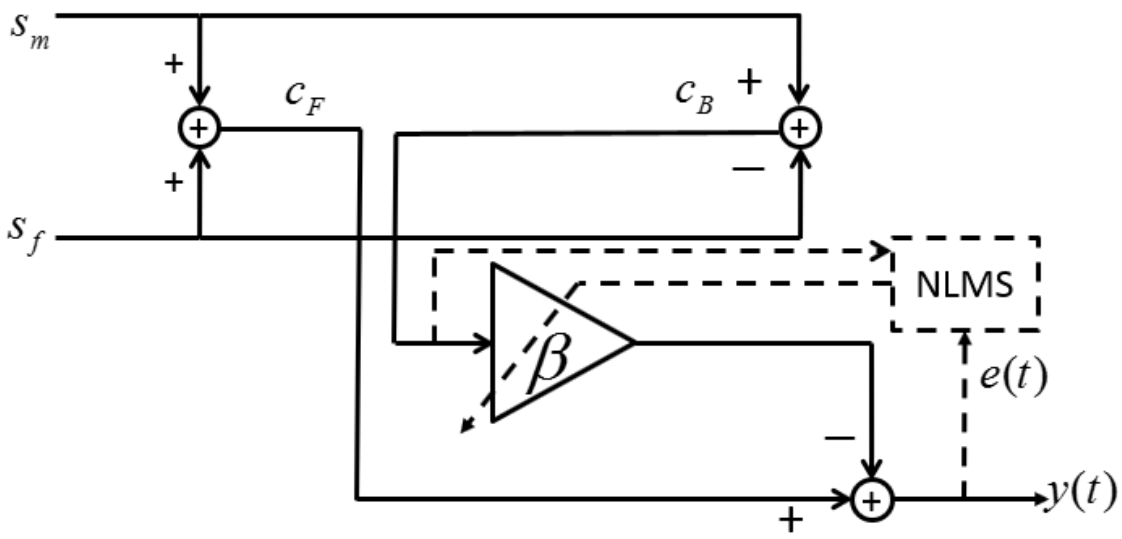

Figure 9: Adaptive beamformer achieved by the proposed method

2.2.2 Generalized Adaptive First-Order Beamformer

It should be noticed that there are two constraints during the usage of the adaptive first-order beamformer. Firstly, the interference speech should come from the back half plane. In other words, the interference speech should come from any angle in the range from 90 to 270 degrees. Secondly, the maximal gain of this beamformer can only be achieved in 0 degree. It implies that people should always point this microphone array to the desired speaker at 0 degree. These two constraints should be taken into account in the usage of this adaptive first-order beamformer.

As discussed in equation (31), where $y(t) = c_F(t) - \beta c_B(t) = ((1 + cos\theta) - \beta(1 - cos\theta))s(t)$, the gain of beamformer can be extracted as,

$$g(\theta, \beta) = (1 + cos\theta) - \beta(1 - cos\theta) \tag{32}$$

Since the purpose of the adaptive beamformer is to reduce the interference, the gain should be 0 when an interference speech comes from null angel θ_{null}.

$$g(\theta_{null}, \beta) = (1 + cos\theta_{null}) - \beta(1 - cos\theta_{null}) = 0 \tag{33}$$

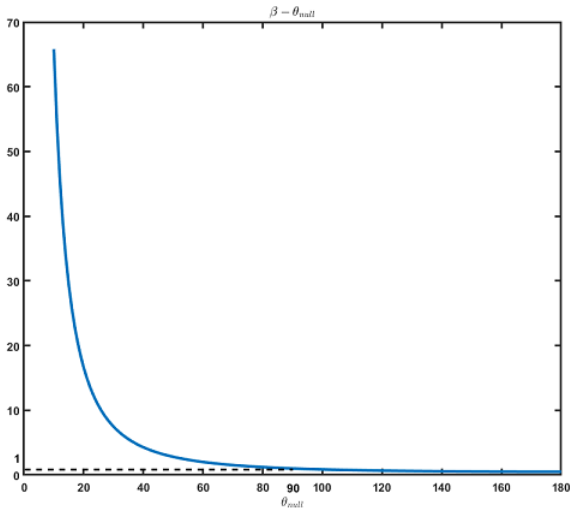

Figure 10: $\beta - \theta_{null}$ forward facing

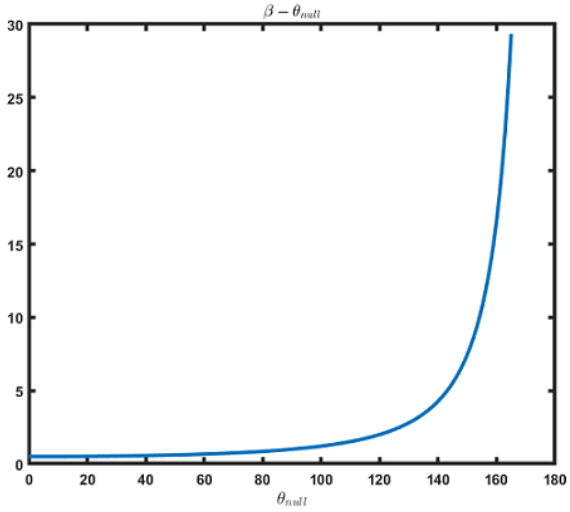

Figure 11: β - θ_{null} backward facing

By solving equation (33), the relationship between β and θ_{null} is expressed below.

$$\beta = \frac{1 + cos\theta_{null}}{1 - cos\theta_{null}} \tag{34}$$

As shown in Fig.10, the value of β diverges when θ_{null} is close to 0. However, when θ_{null} ranges from 90 to 180 degrees, β becomes stable and it ranges from 1 to 0. This feature leads to that the coming angle of interference speech is limited in the back half plane. Furthermore, it can be concluded that the generalized adaptive first-order beamformer is described by the following formulas.

$$y(t) = c_1(t) - \beta c_2(t) \tag{35}$$

$$\beta_{t+1} = \beta_t + \mu \frac{y(t)c_2(t)}{||c_2(t)||^2 + \delta} \tag{36}$$

where the parameters μ and δ remain the same definitions as in equation (31). In this generalized beamformer, signal $c_1(t)$ can be $c_F(t)$ or $c_B(t)$; and signal $c_2(t)$ can be $c_F(t)$ or $c_B(t)$ as well.

Forward Facing Adaptive First-Order Beamformer

In equation (35), when $c_1(t)$ is equal to $c_F(t)$ and $c_2(t)$ is equal to $c_B(t)$, the forward facing adaptive first-order beamformer is obtained. This beamformer suppresses the signal coming from back half plane and enhances that coming from 0 degree. At the

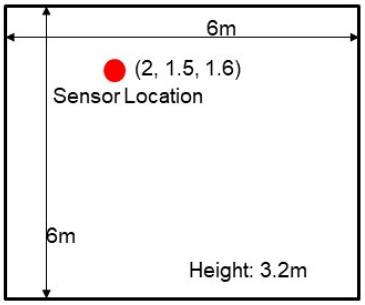

Figure 12: Reverberant room

same time, it requires that the coming angle of interference speech should be in the back half plane.

Backward Facing Adaptive First-Order Beamformer

Similarly, in equation (35), the backward facing adaptive first-order beamformer is obtained, when $c_1(t)$ is equal to $c_B(t)$ and $c_2(t)$ is equal to $c_F(t)$. It always suppresses the signal coming from forward half plane and enhances that coming from 180 degrees. Note that the relationship between β and θ_{null} is modified in this case as,

$$\beta = \frac{1 - cos\theta_{null}}{1 + cos\theta_{null}} \tag{37}$$

According to equation (37), the value of β gets infinity when θ_{null} is close to 180 degrees and becomes stable as θ_{null} remains in the front half plane. This feature, depicted in Fig.11, requires that the coming angle of interference speech should be in the forward half plane.

2.3 Experimental Results

In this section, based on the adaptive first-order beamformer implemented by the collocated fiber acoustic sensor and omni microphone, two numerical simulations are conducted to demonstrate the speech enhancement performance of this design: the first one is speech interference reduction and the second one is speech separation.

2.3.1 Virtual Reverberant Room

A virtual reverberant room with the room size and the fiber acoustic sensor location is shown in Fig.12. We use the image method [47] to simulate the reverberation in the

Acoustic Parameter	Corresponding Value
Room Size	$6m \times 6m \times 3.2m$
Omni Microphone and Fiber Sensor Location	$2m \times 1.5m \times 1.6m$
Fiber Sensor Orientation	0
Reverberation Time T_{60}	400 ms (milliseconds)
Sound Speed	340 m/s

Table 2: Acoustic parameters of virtual reverberant room

real life environment. Table.2 gives the detailed room size and acoustic parameters of a virtual reverberant room setup. In our numerical experiments, the distance between each speaker and the microphone array is fixed as 1.5 meters.

Given the audio from desired speaker $s_d(n)$ and that from the interference speaker $s_i(n)$, the signal received by the omni microphone $s_o(n)$ and that received by the fiber acoustic sensor $s_f(n)$ in a reverberant environment are expressed in equations (38) and (39), respectively,

$$s_o(n) = \mathrm{RIR}_{od}(\theta_d) * s_d(n) + \mathrm{RIR}_{oi}(\theta_i) * s_i(n) \tag{38}$$

$$s_f(n) = \mathrm{RIR}_{fd}(\theta_d) * s_d(n) + \mathrm{RIR}_{fi}(\theta_i) * s_i(n) + s_n(n) \tag{39}$$

where $\mathrm{RIR}_{od}(\theta_d)$ denotes the room impulse response for omni microphone from the desired speaker; $\mathrm{RIR}_{fd}(\theta_d)$ the room impulse response for the fiber sensor from the desired speaker; RIR_{oi} the room impulse response for omni microphone from the interference speaker; RIR_{fi} the room impulse response for the fiber sensor from the interference speaker; θ_d the angular position of the desired speaker; and θ_i the angular position of the interference speaker. The SNR of the fiber sensor is set to be less than 20dB, considering that there is a white noise in the circuits which is denoted as $s_n(n)$ in equation (39).

2.3.2 Speech Interference Reduction

Ideal Case

We first consider the ideal case where the acoustic environment is anechoic and noise-free (without reverberation and noise) and the audio recorded by fiber sensor is not

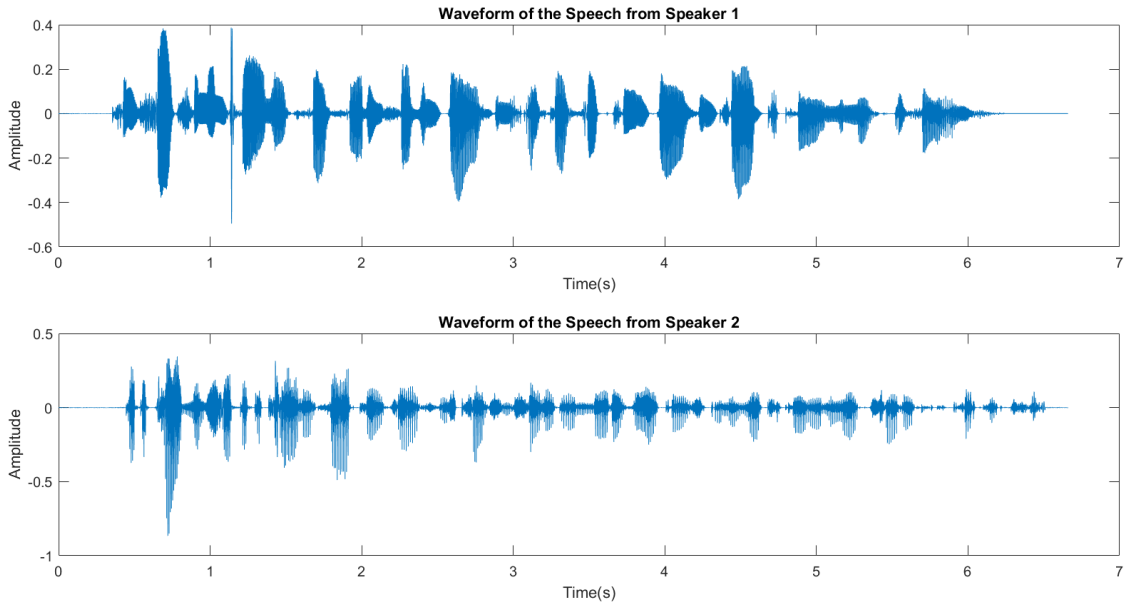

Figure 13: Speech waveform

corrupted by white noise. As shown in Fig.13, two audio samples are used as inputs representing the speeches of the desired and interference speakers respectively.

More specifically, the arrangement of speakers and microphone array is shown in Fig.14, where the speech of speaker 1 from 0 degree and that of speaker 2 from 135 degrees are recorded by the fiber sensor and omni microphone, suppose that the desired speaker is speaker 1. Note that in this setup, the main lobe of the dipole beam pattern of the fiber sensor should always point at the desired speaker.

As the interference speech comes from back half plane, this adaptive beamformer finds an optimal beam pattern by NLMS method to suppress the speech. In this case, the gain of beam pattern at 135 degrees should be 0, since the first-order adaptive beamformer gives the entire beam pattern as displayed in Fig.15. By applying this beam pattern, the generated audio signal is shown in Fig.16, where the waveform is very close to the speech of the desired speaker. We have also conducted more experiments in the cases when the interference speaker is talking at 90 and 180 degrees, and measured the effectiveness of the speech enhancement system using wideband PESQ (Perceptual Evaluation of Speech Quality) score as the objective performance metric. The PESQ score can range from 1 to 4.5. A bigger PESQ score implies a higher speech quality.

The PESQ scores of the signals recorded by omni microphone and output signal

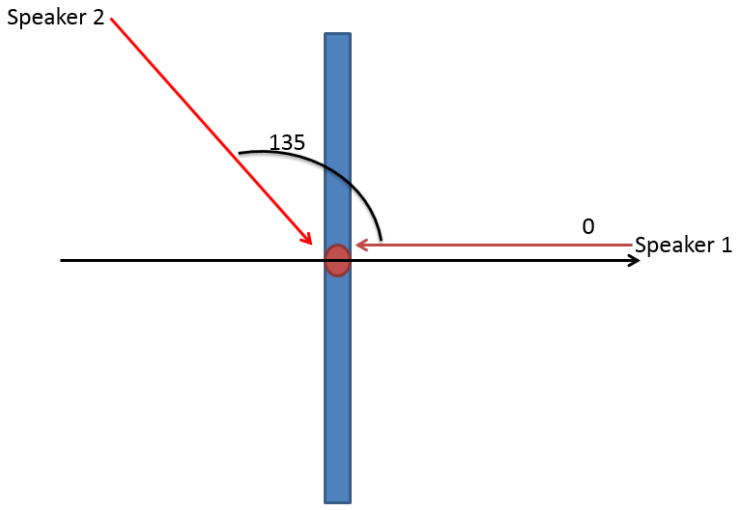

Figure 14: Simulation setup for speech interference reduction

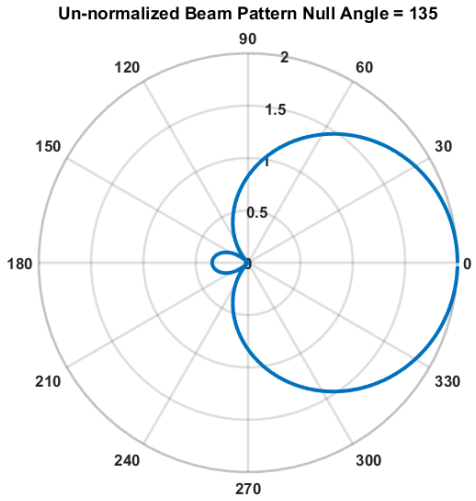

Figure 15: Beam pattern achieved by the adaptive beamformer when the interference speech comes from 135 degrees

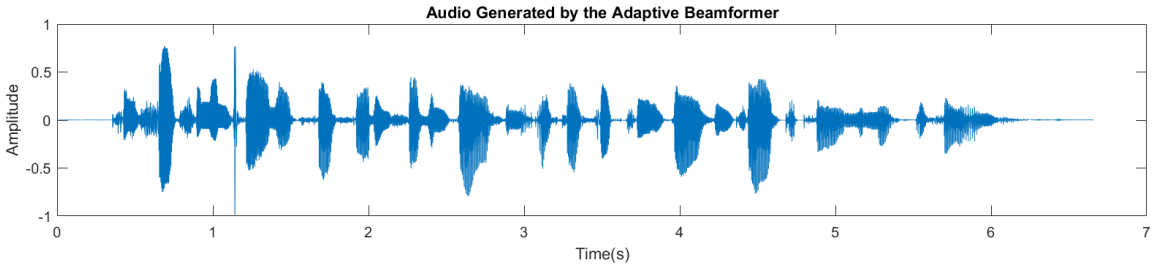

Figure 16: Audio generated by adaptive beamformer

	PESQ (wideband)		
Interference Angle	Omni	Adaptive Beamformer	Improvement
90	1.0821	2.6162	141.77%
135	1.0821	2.8425	162.68%
180	1.0821	3.3587	210.39%

Table 3: PESQ score (ideal case)

	PESQ (wideband)		
Interference Angle	Omni	Adaptive Beamformer	Improvement
90	1.0788	1.2389	14.84 %
135	1.0845	1.2348	13.86 %
180	1.0870	1.1918	9.64 %

Table 4: PESQ score (real case)

from adaptive beamformer are listed in Table.3, where the big improvements on PESQ scores indicate that the adaptive beamformer achieved by the proposed collocated microphone array is able to reduce speech interference significantly. However, these improvements are achieved in the ideal cases without considering white noise and reverberations. In reality, the wall reflections of signals make speech interference reduction more challenging.

Real Case with Reverberation and White Noise

In this case, numerical simulations are conducted in the virtual reverberation room. We assume that the white noise only exists in the audio recorded by the fiber acoustic sensor. The SNR of the audio recorded by the fiber sensor is set to 15dB. Similar with the settings in the ideal cases, the simulations are conducted when the interference speaker is talking at 90, 135 and 180 degrees. The wideband PESQ score is used here as the performance metric of audio quality as well.

The results shown in Table.4 indicate that the proposed method is able to improve audio quality by reducing the negative effects of speech interference. However, due to the presence of reverberation and white noise, the proposed adaptive beamformer cannot obtain significant improvements on the recorded audio quality, comparing to the result in ideal cases, which is shown in Table.3. In order to get more effectiveness

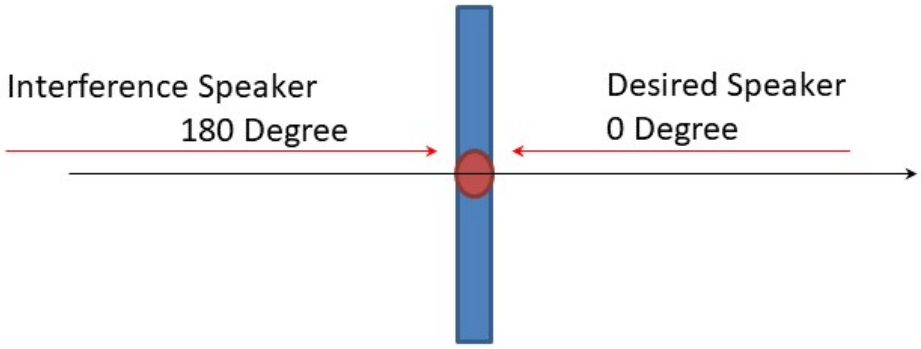

Figure 17: Simulation setup for speech separation

on speech interference reduction to tackle white noise and reverberation, it seems that more investigations are required, such as designing a fiber sensor with higher SNR and a procedure for dereverberation.

2.3.3 Speech Separation

Ideal Case

As shown in Fig.17, speaker 1 from 0 degree and speaker 2 from 180 degrees are talking simultaneously. Their speeches are recorded by the fiber acoustic sensor and omni microphone. Using the generalized adaptive beamformer, which is discussed previously in section 2.2.2, it is apparent that the forward and backward facing adaptive beamformer could extract the speech of speaker 1 and 2 separately.

Speech	Speech 1	Speech 2
PESQ (Recorded by Omni)	1.0821	1.1122
PESQ (Recorded by Generalized Adaptive Beamformer)	**3.3587**	**3.5475**

Table 5: The improvement of PESQ score (ideal case)

After these 2 speeches are separated, the wideband PESQ scores of these recovered signals are calculated. At the same time, the PESQ scores of the speeches recorded by omni microphone are also computed for comparison. These PESQ scores are listed in Table.5, where results show that the proposed method using the collocated microphone array can improve audio quality by speech separation.

Real Case with Reverberation and White Noise

Here, we consider the room reverberation and the white noise existing in the fiber acoustic sensor for our numerical simulations. The forward-facing and backward-facing adaptive beamformer are used to separate the two simultaneous speeches as well. After the simulations, the wideband PESQ scores are listed in Table.6. Comparing to the results in the ideal cases, the performance of speech separation is degraded a lot due to the presence of reverberation and white noise. It can be concluded that the proposed speech separation method is able to enhance the audio quality on a smaller scale. However, in reality, a dereverberation procedure and a fiber sensor with better SNR are demanded, especially in a reverberation intensive environment.

Speech	Speech 1	Speech 2
PESQ (Recorded by Omni)	1.0870	1.0635
PESQ (Recorded by Generalized Adaptive Beamformer)	**1.1918**	**1.1038**

Table 6: The improvement of PESQ score (real case)

2.4 Conclusion

In this chapter, the design of the first-order adaptive differential microphone array has been presented firstly. Then, inspired by this differential microphone array and the beam pattern of fiber acoustic sensor, a more compact microphone array, which is constructed by one omni microphone collocated with a fiber acoustic sensor, has been proposed to form a first-order adaptive beam pattern. This design can be regarded as a simplified version of acoustic vector sensor as well. In order to investigate the speech enhancement performance of this collocated microphone array, numerical experiments are conducted in both ideal case and real case, where the former means an anechoic and noise-free condition and the latter corresponds to the an acoustic environment with both reverberation and white noise. The results of numerical experiments show that this design is able to enhance the audio quality through speech interference reduction and speech separation in the ideal case. However, limited by the environment with strong reverberation and white noise in the fiber sensor, the improvement of audio quality degrades a lot in a real scenario. In the future, more

investigations are required, such as developing a procedure for dereverberation or designing a better fiber acoustic sensor with higher SNR.

Chapter 3

Speech Enhancement using X-Y Collocated Fiber Sensors

As discussed in the previous chapter, a two-dimensional acoustic vector sensor includes one omni microphone and two directional microphones pointing at x and y axes respectively. Utilizing this arrangement, it is able to locate speaker's direction and enhance the quality of recorded speech. In this chapter, in order to construct a simplified version of acoustic vector sensor, we use only two fiber acoustic sensors to build a collocated microphone array and investigate its performance on DOA estimation and speech enhancement. This type of collocated microphone array is named as X-Y collocated fiber sensors.

The beam pattern of each fiber sensor is a dipole one, namely, the angle that delivers the maximal gain is orthogonal to the null angle of the dipole. This feature leads to that the audio signal arriving from null angle is eliminated. Intuitively, the null angle of one fiber acoustic sensor could be compensated easily by aligning another fiber acoustic sensor perpendicularly to the current one. This kind of orthogonal arrangement is the property of X-Y collocated fiber sensors. Here we also call this collocated microphone array as X-Y sensors for simplification. The detailed structure and acoustic property of X-Y sensors are presented in section 3.1.

However, once adding another perpendicular fiber sensor and forming the X-Y sensors, two audio signals from these sensors could be obtained. How to utilize these two audio signals and produce the final output is the key to the success of employing this X-Y sensors. Otherwise, the audio quality of the final output cannot be assured.

In order to solve this problem, a steerable beamforming approach and DOA estimation methods are proposed and explained in detail in section 3.2 and 3.3 respectively.

In the real scenario, the circuits in the fiber acoustic sensor converts the vibration of acoustic waves into voltage signal. Meanwhile, it also produces white noise, as side effect, which degrades the audio quality severely. However, the property of the white noise produced by certain circuits is stable and could be measured as prior knowledge. In order to reduce the audio quality degradation caused by white noise, a spectral subtraction approach is proposed to suppress the white noise. This content is introduced in section 3.4.

At last, in section 3.6, to make the numerical simulation close enough to the real world scenario, the image method [47] is adopted to simulate the room acoustic conditions. Utilizing the DOA estimation techniques discussed in section 3.3, a steerable beam pattern is achieved and followed by noise reduction techniques introduced in section 3.4. This entire procedure is regarded as a speech enhancement system and summarized in section 3.5. In order to evaluate the proposed system, the PESQ score [48] is calculated as the prime comparison metric.

3.1 Principle of X-Y Collocated Fiber Acoustic Sensors

Since the beam pattern of the fiber sensor is directional, the speech coming from null angle will not be captured. In order to compensate for this drawback, one of the most intuitive and easiest ways is adding the second fiber sensor, which is collocated and orthogonal to the first one. These two collocated fiber sensors are named as X-fiber and Y-fiber respectively.

As shown in Fig.18, two orthogonal dipoles are formed by these two collocated fiber acoustic sensors. Assuming that one speaker is talking at 0 degree, the signals recorded by X-fiber and Y-fiber are shown in Fig.19. The audio signal captured by Y-fiber is silence since the speaker is at the null angle of Y-fiber dipole. Meanwhile, the audio signal is fully recorded by X-fiber as the speaker is at the main lobe of X-fiber dipole. It should also be noted that there is white noise presenting in the signals recorded by fiber sensors. This white noise is generated in the circuits, which leads to a degradation of audio quality.

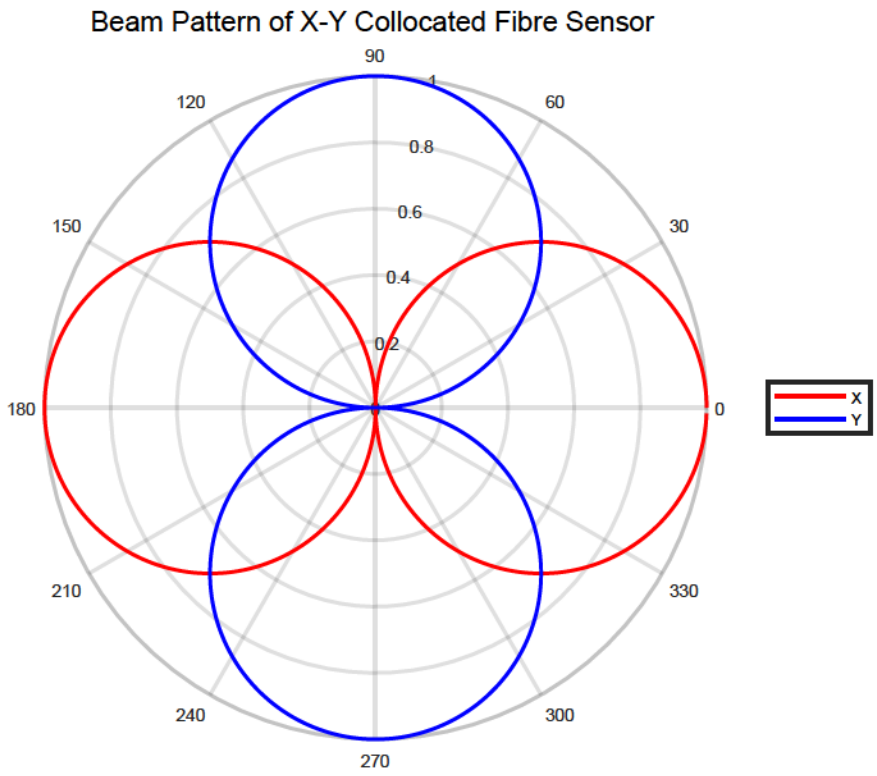

Figure 18: Beam pattern of X-Y collocated fiber sensors

In summary, given a speaker talking at θ degree, the output signals of X-fiber and Y-fiber can be represented by equations (40) and (41), respectively,

$$s_X(n) = cos(\theta) \cdot s(n) \tag{40}$$

$$s_Y(n) = sin(\theta) \cdot s(n) \tag{41}$$

where $s(n)$ denotes the original speech signal of the speaker; $s_X(n)$ and $s_Y(n)$ represent the audio signals recorded by X-fiber and Y-fiber, respectively.

3.2 Steerable Beamforming

As discussed in the previous section, the X-fiber and Y-fiber are able to form sine and cosine dipole beam patterns respectively. Meanwhile, since the cosine and sine functions are orthogonal to each other, it is easy to form a dipole pointing at angle θ

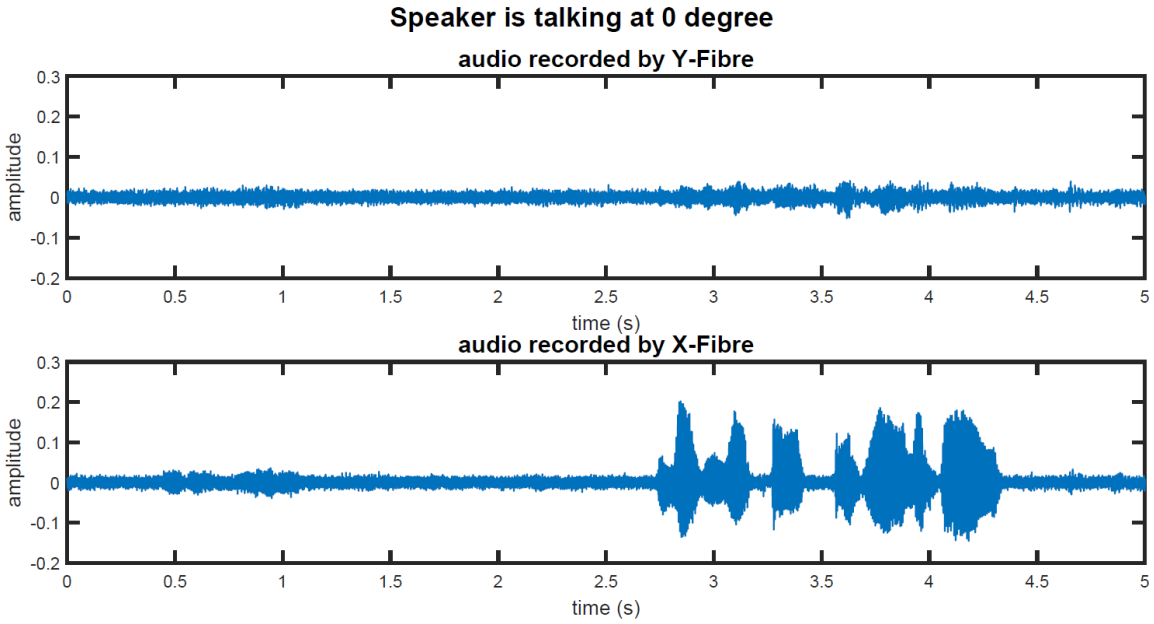

Figure 19: Directional properties of X-Y fiber sensors

in a 2D plane by,

$$\cos(\phi - \theta) = \cos(\phi) \cdot \cos(\theta) + \sin(\phi) \cdot \sin(\theta) \tag{42}$$

where ϕ ranges from $[0,\ 2\pi)$ and θ is the angle pointed by the main lobe of dipole $\cos(\phi - \theta)$. Note that the value of θ could be limited into a smaller range: $[0,\ \pi)$, due to the symmetric property of dipole beam pattern.

Considering the situation where only one speaker is present, steering the main lobe of a dipole to the speaker could be expressed below if the angular location of the speaker is given,

$$p(n) = cos(\theta) \cdot s_X(n) + sin(\theta) \cdot s_Y(n) \tag{43}$$

where θ is the angular location of the speaker; $s_X(n)$ and $s_Y(n)$ represent the signals recorded by X-fiber and Y-fiber respectively; and $p(n)$ is the output signal of steerable beamformer.

3.3 DOA Estimation

Sound localization is an important technique for many applications like target detection, speaker localization and identification [49]–[51]. Recently, sound localization

and DOA estimation are used widely in human computer interaction, hearing aids device in noisy environment and virtual reality [52], [53]. Conventional sound localization methods by acoustic pressure sensing usually need TDOA or sound pressure amplitude difference. Here, for X-Y sensors, a sound localization method needs to be developed to form a steerable beam pattern, which is used to record the speech from any direction.

In order to achieve the steerable dipole and point it at the speaker, a DOA estimation method is demanded. Due to the X-fiber and Y-fiber are collocated, there is no time delay between signals recorded by these two fiber sensors. It means that the typical DOA estimation methods which need inter-element TDOA, such as GCC-PHAT [5], are not suitable for X-Y sensors any more. However, due to the dipole beam patterns formed by X-fiber and Y-fiber with orthogonal orientations, the different amplitudes of the signals recorded by X-fiber and Y-fiber expose the spatial clues about speaker's direction. In this section, a robust DOA estimation method, which is based on inter-channel level difference, is proposed to locate speech arriving angle using this collocated microphone array. This method is also extendable to locate multiple sound sources with small computational complexity.

3.3.1 Inter-Channel Level Difference

Considering a scenario where only one person is speaking, given the angular position of the speaker θ, the output signals $s_X(n)$ and $s_Y(n)$ could be expressed by equations (40) and (41) respectively. On the other hand, it should be possible to estimate θ by using $s_X(n)$ and $s_Y(n)$, since this θ could be calculated by,

$$\theta = tan^{-1}(\frac{|S_Y(e^{jw})|}{|S_X(e^{jw})|}) \tag{44}$$

where $S_X(e^{jw})$ and $S_Y(e^{jw})$ are obtained by applying Fourier transform to $s_X(n)$ and $s_Y(n)$ respectively. This approach, based on the amplitude difference between X-fiber and Y-fiber signals, is also named as inter-channel level difference method.

Rectangular (Bandpass) Filter Bank on Mel Scale

It is more efficient to estimate DOA in the time-frequency domain in the presence of noise and reverberation. In this section, the rectangular filter bank on Mel scale is

used to compute the spectrum. The Mel scale is usually adopted to mimic the non-linear property of human hearing system, which implies that sounds in low frequencies are discriminated by human beings more easily than that in high frequencies. It was devised through human perception experiments and was firstly suggested by Stevens and Volkman in 1937 [54]. The relationship between Mel scale m and frequency f in Hertz in non-linear (logarithmic) form is given by,

$$m = 2595 log_{10}(1 + \frac{f}{700}) \tag{45}$$

or

$$f = 700(10^{\frac{m}{2595}} - 1) \tag{46}$$

Algorithm 1: The Procedure to Compute Bandpass Filter Bank

Input: Fs: sampling frequency;
    N_{fft}: number of FFT points;
    f_{min}: minimal frequency;
    f_{max}: maximal frequency;
    N_{filter}: number of bandpass filters
Procedure:
    $m_{min} = 2595 log_{10}(1 + \frac{f_{min}}{700})$
    $m_{max} = 2595 log_{10}(1 + \frac{f_{max}}{700})$
    $m = linspace(m_{min}, m_{max}, N_{filter} + 1)$
    $f = 700(10^{\frac{m}{2595}} - 1)$
    $b = floor(\frac{f}{Fs} N_{fft}) + 1$
    $f_{start} = b(1 : end - 1)$
    $f_{stop} = b(2 : end)$
    H = zeros(N_{fft}, N_{filter})
    for i = 1 to N_{filter}
        $H(f_{start}(i) : f_{stop}(i), i) = 1$
    endif
    return H

Equation (46) explains how to compute frequency value f once the corresponding Mel value m is given. Furthermore, on the top of Mel scale, several bandpass filters are constructed to exploit the human perceptual feature. Once the bandpass filter bank is obtained, these filters are applied to the spectrums of audio signals from both X and Y fiber sensors. The procedure to implement the bandpass filter bank is given in the Algorithm.1. Note that the bandwidth of each bandpass filter varies. In practice,

a bandpass filter is discarded once its bandwidth is not large enough. Furthermore, since the frequency of human voice usually ranges from 300Hz to 20kHz, the bandpass filters should also focus on this range. As shown in Fig.20, eight bandpass filters are constructed with the frequency ranging from 300Hz to 20kHz under the sampling rate of F_s=48kHz and the FFT (fast Fourier Transform) size of N_{fft}=128.

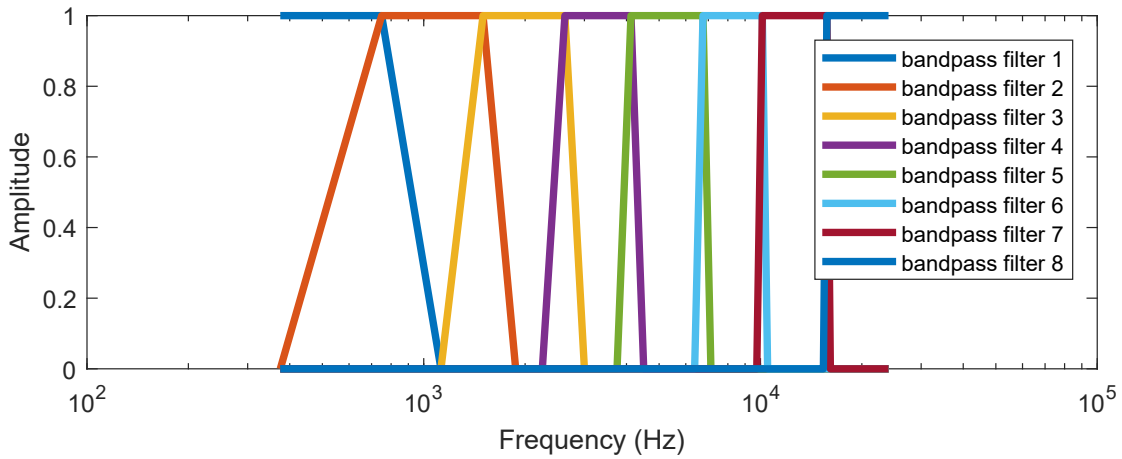

Figure 20: Frequency responses of eight bandpass filters

DOA Calculation

Given double channel signal $s_X(n)$ and $s_Y(n)$, the short time Fourier transform (STFT) is used to obtain power spectrum $|S_X(t,w)|^2$ and $|S_Y(t,w)|^2$, where t is the time frame index and w is the frequency bin. At this step, after applying filter bank $H(w, n_{filter})$ to both $|S_X(t,w)|^2$ and $|S_Y(t,w)|^2$, the DOA map $\theta(t, n_{filter})$ is calculated by,

$$\theta(t, n_{filter}) = \begin{cases} tan^{-1}(\sqrt{\frac{E_Y(t,n_{filter})}{E_X(t,n_{filter})}}) & corr(s_X(n), s_Y(n)) > 0 \\ \pi - tan^{-1}(\sqrt{\frac{E_Y(t,n_{filter})}{E_X(t,n_{filter})}}) & otherwise \end{cases} \tag{47}$$

where *corr* denotes the cross correlation; $E_X(t, n_{filter})$ is equal to $|S_X(t,w)|^2 H(w, n_{filter})$; and $E_Y(t, n_{filter})$ is equal to $|S_Y(t,w)|^2 H(w, n_{filter})$. Note that the DOA 0 is identical to π according to this equation. It implies that the range of DOA estimates is with a period π instead of the conventional 2π. This property is named as phase ambiguity.

As shown in Fig.21, a 2D array of DOA estimates is calculated according to equation (47). However, the dataset of all the DOA estimates is super noisy. It is

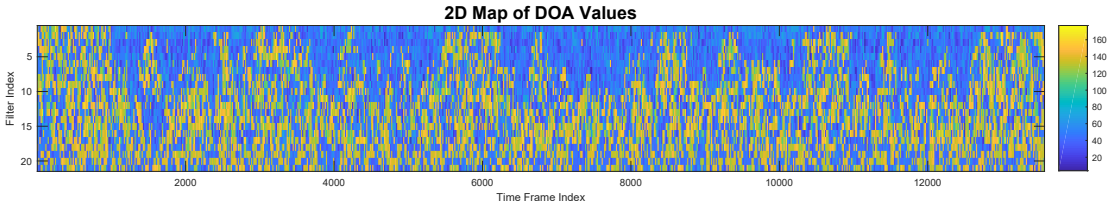

Figure 21: A 2D array of DOA estimates

necessary to develop a metric to select reliable DOA estimates from all these samples.

Time-Frequency Bins Selection

Due to the presence of noise from the fiber acoustic sensor, the X and Y double channel speech quality is degraded. Even if there is no speech present, the spectrums of audio signals $s_X(n)$ and $s_Y(n)$ are not zeros. It is necessary to develop a metric to determine whether a speech signal presents in time-frequency domain. In other words, given a 2D (time-frequency) array where each pixel contains a DOA estimate, a 2D mask, which determines whether a speech is significant enough in each pixel, is used to filter out the unreliable DOA estimates. As a result, by applying this mask to the 2D array of noisy DOA estimates, a reliable and clean DOA dataset is obtained without being contaminated by the white noise inherent in fiber acoustic sensors.

Generally speaking, the energy of audio signal becomes larger when a speech signal exists. Similarly, a mask could be calculated by utilizing several thresholding values along frequency axis. More specifically, when there is no speech present, the white noise signal $n(n)$ is recorded. By applying short time Fourier transformation to noise signal $n(n)$ and then applying filter bank $H(w, n_{filter})$ to the power spectrum $|N(t,w)|^2$ in the time-frequency domain, the time-frequency representation $E_N(t, n_{filter})$ is calculated by $E_N(t, n_{filter}) = |N(t,w)|^2 H(w, n_{filter})$, where $N(t,w)$ is obtained by applying STFT to white noise signal $n(n)$. And then the average value of $E_N(t, n_{filter})$ along time axis $E_N^{avg}(n_{filter})$ is computed as well. This value $E_N^{avg}(n_{filter})$ and the corresponding threshold $T_{n_{filter}}$ are used to compute the mask $M(t, n_{filter})$, given $E_X(t, n_{filter})$ and $E_Y(t, n_{filter})$, as described below,

$$M(t, n_{filter}) = \begin{cases} 1 & \frac{max\{E_X(t,n_{filter}), E_Y(t,n_{filter})\}}{E_N^{avg}(n_{filter})} > T_{n_{filter}} \\ 0 & otherwise \end{cases} \tag{48}$$

where t denotes the frame index; n_{filter} represents the index of bandpass filters,

which is discussed in section 3.3.1 above; and $T_{n_{filter}}$ is the thresholding value for the bandpass filter indexed by n_{filter}. In summary, this thresholding metric is based on the prior knowledge about the white noise inherent in fiber acoustic sensors.

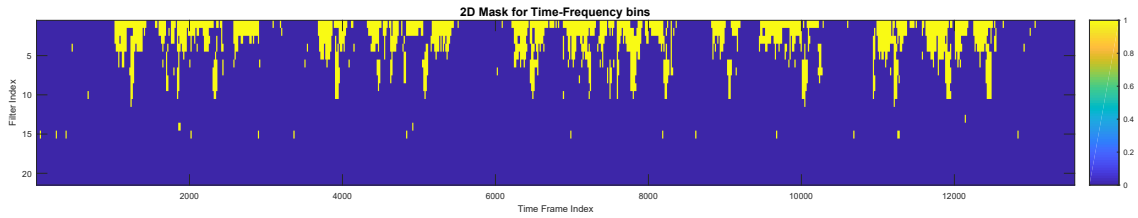

Figure 22: 2D mask for time-frequency bins selection

As shown in Fig.22, using the criteria introduced above, a 2D mask is obtained to select reliable samples from all DOA estimates. Here, the yellow area denotes the locations where DOA estimates are reliable; and on the other hand, the blue area denotes the locations where DOA estimates should be ignored.

The Distribution of DOA Estimates

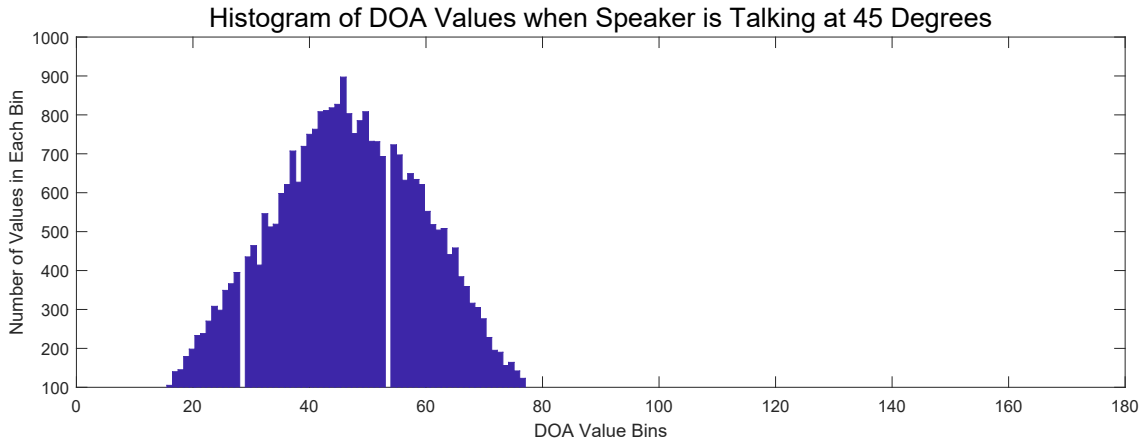

Figure 23: Histogram of DOA estimates when the speech comes from 45 degrees

By applying the 2D mask to the 2D array of raw DOA estimates, a more reliable set of DOA estimates is generated. A histogram is an easy and powerful tool to visualize how these DOA estimates are distributed in angular space. The Fig.23 depicts the distribution of DOA estimates when a speaker is talking at 45 degrees. This distribution could be modeled by the Gaussian distribution. And the DOA

estimate could be represented by the mean value of the Gaussian distribution. So given a set of DOA estimates, the mean value of its Gaussian distribution is the average of all samples. However, it should be noticed that the sample space of the Gaussian distribution includes any real value from $-\infty$ to ∞. Here, the typical sample space of DOA estimates is an angular one where 0 is identical to 2π. More specifically, using the X-Y collocated fiber sensors here, the DOA 0 is identical to π due to the inter-channel level difference method expressed by equation (47). It implies that the phase ambiguity between 0 and π becomes a barrier to obtain an accurate DOA estimate.

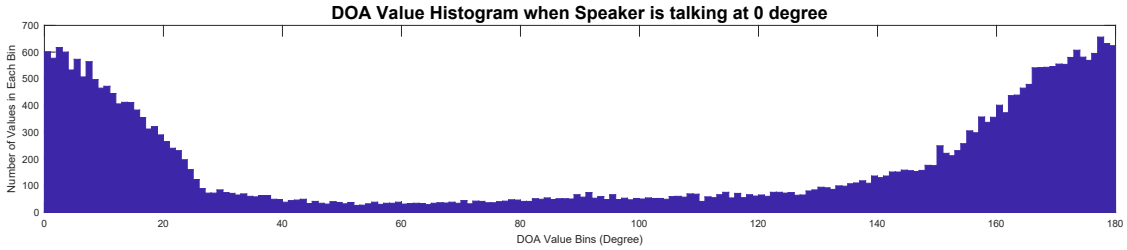

Figure 24: Histogram of DOA estimates when the speech comes from 0 degree

Fig.24 shows a histogram representing the distribution of DOA estimates when the speaker is talking at 0 degree. It is apparent that the distribution cannot be modeled by the Gaussian distribution any more, since the DOA estimates should be in a circular space. Also, using average value as the final DOA estimate is not suitable neither, since the mean value of the dataset shown in Fig.24, where the samples in half dataset are around 0 and samples in the other half dataset are around 180, is 90 but 0. Due to this problem, another distribution, the wrapped Gaussian distribution, is adopted and generalized here to solve the angular ambiguity between 0 and π for DOA estimation.

3.3.2 Wrapped Gaussian Mixture Model

Since the DOA estimates are within an angular space which has a circular property, the approaches for DOA estimation based on Gaussian distribution like Gaussian mixture model (GMM) do not hold any more. However, based on the statistics on circular space, the wrapped Gaussian distribution is suitable to model the distribution of DOA estimates. Furthermore, the wrapped Gaussian mixture model (WGMM) is

able to handle multiple-speaker localization problems.

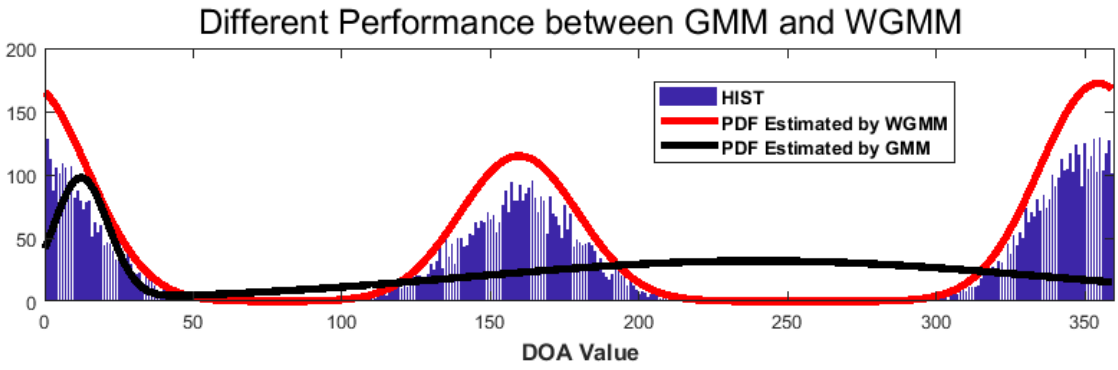

Figure 25: Performance comparison between WGMM and GMM on synthesized angular dataset with means at 160 and 355 degrees

Fig.25 depicts different results of WGMM and GMM methods on estimating the distribution of synthesized DOA samples, which is represented by a histogram. Due to the phase ambiguity problem between 0 and 2π, the GMM method gives a wrong probability density function (PDF), which is highlighted by a black line. It is apparent that the WGMM is able to obtain accurate distribution of DOA samples. The WGMM method was proposed to quantify the harmonic phases in human speech as early as in 2007 [55]. Recently, more and more researchers find it useful for DOA estimation, especially for multiple-speaker localization problems [56], [57]. However, their work only solves the phase ambiguity problem between 0 and 2π. In this section, the WGMM method for any period (2π, π .etc) is derived and a histogram based WGMM method is developed to reduce the computational complexity with minor precision loss. Furthermore, this histogram based WGMM method is suitable to be deployed in embedded systems for real-time applications.

Circular (Directional) Statistics

Circular (Directional) Statistics is a subclass of statistics that deals with directions, axes (lines going through the origin in R^n) and rotations in R^n [58]. More generally, directional statistics deal with observations on compact Riemann manifolds [58]. In this case, the DOA samples calculated by inter-channel level difference method are circular as well, since the 0 degree and 180 degrees are considered identical.

In order to obtain the wrapped Gaussian distribution ($\mathrm{WN}(x;\mu,\sigma^2)$), the linear

normal distribution ($N(x;\mu,\sigma^2)$) is wrapped onto the unit circle [59]. The PDF of normal distribution $N(x;\mu,\sigma^2)$ is given by,

$$N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{49}$$

where x is a circular random variable and confined in interval $[0,\ 2\pi)$; μ is the mean value and σ^2 is the variance. Hence, the PDF of the wrapped Normal distribution is expressed as,

$$\mathrm{WN}(x;\mu,\sigma^2) = \sum_{w\in Z} [\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu-2\pi w)^2}{2\sigma^2}}] \tag{50}$$

where w is an integer; μ and σ^2 are the mean and variance of the wrapped normal distribution respectively.

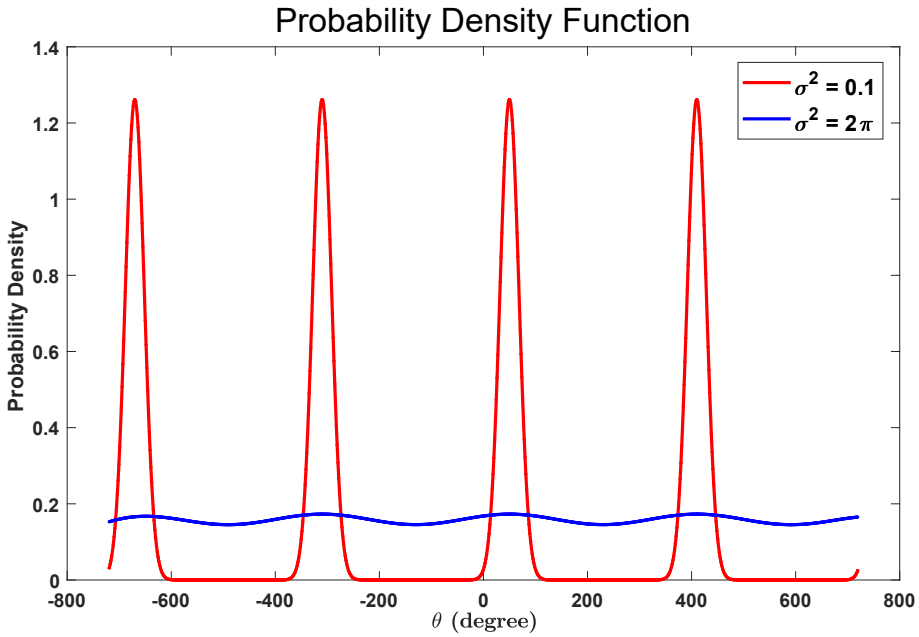

Figure 26: σ^2 impacts on the PDF of wrapped Gaussian distribution

As shown in Fig.26, the PDF of wrapped normal distribution could be approximated by the PDF of linear normal distribution if the variance σ^2 satisfies $\sigma^2 \leq 1$ and by the uniform distribution if the variance σ^2 satisfies $\sigma^2 \geq 2\pi$ [55]. In summary, the PDF of wrapped normal distribution is constructed by infinite wrappings of the PDF of linear normal distribution in the interval $[0,\ 2\pi)$ [55]. In practice, however,

selecting the value of w in equation (50) ranging from -2 to 2 could provide a sufficient approximation even if σ^2 is large [55].

Wrapped Gaussian Mixture Model (WGMM)

According to the DOA calculation procedure stated in section 3.3.1, N DOA estimates could be obtained and regarded as a DOA dataset, where each DOA estimate is marked as x^i. Regarding the clusters as a random variable z, for each DOA estimate x^i, $z^i = k$ if this estimate x^i belongs to cluster k. Assuming there are K clusters in total and each cluster is distributed as wrapped Gaussian distribution, given a DOA estimate x^i, the probability $P(x^i)$ is given by,

$$P(x^i) = \sum_{k=1}^{K} P(x^i|z^i = k)P(z^i = k) = \sum_{k=1}^{K} \mathrm{WN}_k(x^i; \mu_k, \sigma_k^2)\alpha_k \tag{51}$$

where α_k represents the weight of cluster k and $\mathrm{WN}_k(x^i; \mu_k, \sigma_k^2)$ represents the PDF of the kth wrapped Gaussian distribution. Note that μ_k is the mean value of the wrapped Gaussian distribution WN_k and it should range within the interval $[0, 2\pi)$.

Objective Function based on WGMM

Given that the number of components in the mixture model is K, then

$$\gamma = \{\alpha_1, ..., \alpha_k, ..., \alpha_K, \mu_1, ..., \mu_k, ..., \mu_K, \sigma_1^2, ..., \sigma_k^2, ..., \sigma_K^2\}$$

is the parameter set of WGMM. Meanwhile, assuming that the observed DOA is x^i and there are N observations in total, the objective function of the maximum log-likelihood estimation is defined in equation (52), which utilizes a log of the joint probability $P(z^i = k, x^i)$ (the probability when the DOA estimate is x^i and it belongs to cluster k). The procedure about how the objective function is derived is given in Appendix A.4.

$$L = \sum_{i=1}^{N} \sum_{k=1}^{K} P(z^i = k|x^i)[logP(z^i = k, x^i)] \tag{52}$$

Since $P(z^i = k, x^i) = P(z^i = k)P(x^i|z^i = k)$, the objective function could be rewritten as,

$$L = \sum_{i=1}^{N} \sum_{k=1}^{K} P(z^i = k|x^i)[log\alpha_k + log\mathrm{WN}_k] \tag{53}$$

Expectation-Maximization Algorithm

Now, the DOA estimation is rephrased as an optimization problem with the target to find the maximum of objective function L in equation (53) and its corresponding optimal parameters γ. The DOA estimate of cluster k is the mean value μ_k among γ.

In order to find these optimal parameters, the EM (Expectation-Maximization) algorithm is adopted here [60]. The EM algorithm is implemented in iterative manner. Each iteration includes 2 steps: E-step and M-step. On E-step, $P(z^i = k|x^i)$ and $P(x^i)$ are computed. On M-step, the optimal parameter set γ is updated. The equations used for updating γ are obtained based on the following constraints: $\frac{\partial L}{\partial \mu_k} = \frac{\partial L}{\partial \sigma_k^2} = 0$ and $\sum_{i=1}^{K} \alpha_k = 1, \alpha_k > 0$.

Algorithm 2: EM algorithm for DOA estimation based on WGMM

Initialize μ_k, α_k, σ_k^2

Repeat until value of L converges:

1. E-step: Compute $P(z^i = k|x^i)$ and $P(x^i)$

$$P(x^i) = \sum_{k=1}^{K} \alpha_k \mathrm{WN}(x^i; \mu_k, \sigma_k^2)$$

$$P(z^i = k|x^i) = \frac{\alpha_k \mathrm{WN}(x^i; \mu_k, \sigma_k^2)}{P(x^i)}$$

2. M-step: Update α_k, μ_k, σ_k^2. .

$$\alpha_k = \frac{\sum_{i=1}^{N} P(z^i = k|x^i)}{N}$$

$$\mu_k = \frac{\sum_{i=1}^{N} \frac{1}{P(x^i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(x^i - 2w\pi - \mu_k)^2}{2\sigma_k^2}} (x^i - 2w\pi)}{\sum_{i=1}^{N} \frac{1}{P(x^i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(x^i - 2w\pi - \mu_k)^2}{2\sigma_k^2}}}$$

$$\sigma_k^2 = \frac{\sum_{i=1}^{N} \frac{1}{P(x^i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(x^i - 2w\pi - \mu_k)^2}{2\sigma_k^2}} (x^i - 2w\pi - \mu_k)^2}{\sum_{i=1}^{N} \frac{1}{P(x^i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(x^i - 2w\pi - \mu_k)^2}{2\sigma_k^2}}}$$

As shown in Algorithm.2, the procedure about details of EM algorithm is given. The derivation procedure for obtaining equations to update the parameters (α_k, μ_k, σ_k^2) is provided in Appendix A.4.

3.3.3 Generalized Wrapped Gaussian Mixture Model

Usually, the period of the wrapped Gaussian distribution is 2π. Moreover, the corresponding wrapped Gaussian mixture model could deal with DOA samples in a circular space with the period 2π. Furthermore, this type of DOA estimation method becomes more popular in recent research papers like [56] and [57]. However, limited by the symmetric dipole beam pattern of the fiber acoustic sensor, the DOA samples calculated by the inter-channel level difference method are with a period π. In other words, the DOA estimate 0 and 180 degrees are identical in this scenario. So it is necessary to generalize WGMM and make it fit with any possible periods.

Wrapped Gaussian Distribution with Period T

In order to enable the generalized wrapped Gaussian mixture model to be suitable with any period T, the PDF of wrapped Gaussian distribution with period T is defined below,

$$\mathrm{WN}(x;\mu,\sigma^2,T)=\sum_{w\in Z}[\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu-Tw)^2}{2\sigma^2}}] \tag{54}$$

where T is a fixed radian value. In practice, it is not necessary to do a summation over infinite w values. Furthermore, choosing w among a small and finite range could lead to a sufficient approximation. The range of w values depends on the value of T. In practice, we can choose $w \in \{-2,-1,0,1,2\}$ to approximate the PDF of wrapped Gaussian distribution if $T=2\pi$ and choose $w \in \{-1,0,1\}$ if $T=\pi$. These parameters are proved as the optimal ones according to our empirical results.

EM algorithm for Generalized Wrapped Gaussian Mixture Model

The details of EM algorithm for generalized wrapped Gaussian mixture model is given in Algorithm.3. The procedure is similar with EM algorithm for the wrapped Gaussian mixture model except that we need to replace $\mathrm{WN}(x^i;\mu_k,\sigma_k^2)$ with $\mathrm{WN}(x^i;\mu_k,\sigma_k^2,T)$.

Note that the value of T depends on different application scenarios. It is selected as π for the inter-channel level difference method in this thesis, which is introduced in section 3.3.1.

Algorithm 3: EM algorithm for Generalized Wrapped Gaussian Mixture Model

Initialize μ_k, α_k, σ_k^2

Repeat until value of L converges:

1. E-step: Compute $P(z^i = k|x^i)$ and $P(x^i)$

$$P(x^i) = \sum_{k=1}^{K} \alpha_k \mathrm{WN}(x^i; \mu_k, \sigma_k^2, T)$$

$$P(z^i = k|x^i) = \frac{\alpha_k \mathrm{WN}(x^i; \mu_k, \sigma_k^2, T)}{P(x^i)}$$

2. M-step: Update α_k, μ_k, σ_k^2. .

$$\alpha_k = \frac{\sum_{i=1}^{N} P(z^i=k|x^i)}{N}$$

$$\mu_k = \frac{\sum_{i=1}^{N} \frac{1}{P(x^i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(x^i - wT - \mu_k)^2}{2\sigma_k^2}} (x^i - wT)}{\sum_{i=1}^{N} \frac{1}{P(x^i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(x^i - wT - \mu_k)^2}{2\sigma_k^2}}}$$

$$\sigma_k^2 = \frac{\sum_{i=1}^{N} \frac{1}{P(x^i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(x^i - wT - \mu_k)^2}{2\sigma_k^2}} (x^i - wT - \mu_k)^2}{\sum_{i=1}^{N} \frac{1}{P(x^i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(x^i - wT - \mu_k)^2}{2\sigma_k^2}}}$$

3.3.4 Generalized Wrapped Gaussian Mixture Model Based on Histogram

Since most DOA estimation algorithms should run in a real-time scenario, the computational complexity of a real-time algorithm should be feasible under different conditions. The complexity of the generalized wrapped Gaussian mixture model is bounded by 2 factors: N_i (number of iterations) and N_s (number of samples). However, in the real-time scenario, these two values (N_i and N_s) are random. This randomness will cause unpredictable time delay, which degrades the real-time performance of DOA estimation. In order to control this randomness, a maximal number of iterations N_{imax} is used to limit N_i; and for constraining N_s, a histogram is adopted here to set N_s as a fixed value.

Histogram Construction

As shown in Fig.27, a histogram is defined by N_{bin} bins. For the ith bin, b_i denotes the angular value and N_{b_i} represents the number of samples inside the ith bin. Furthermore, the resolution of a histogram is defined as $v_{res} = b_{i+1} - b_i$, where $0 \leq i < N_{bin}$. Given a DOA sample x and the histogram resolution v_{res}, it could be confirmed that DOA sample x belongs to the $\frac{x}{v_{res}}$th bin. A histogram with a proper resolution v_{res}

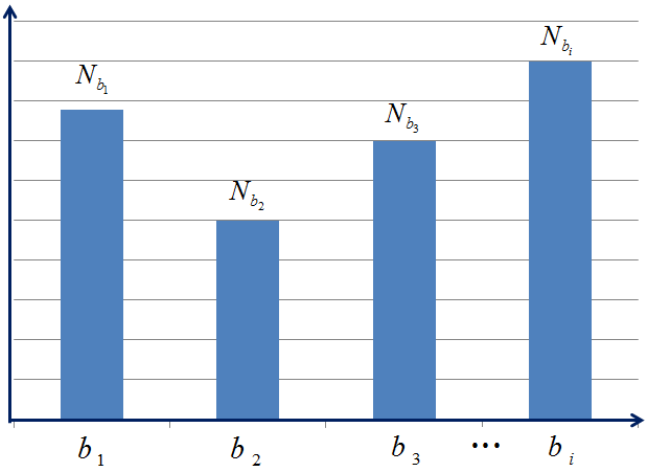

Figure 27: Histogram construction

could obtain a good enough approximation to describe the distribution of DOA samples. At the same time, it reduces the computational complexity of the wrapped Gaussian mixture model as well.

EM algorithm for Generalized WGMM based on Histogram

Given a DOA dataset, each sample inside the dataset is mapped into a certain bin of a histogram. Inside the ith bin of the histogram, there are N_{b_i} samples with the same angular value b_i. Under this assumption, the equations of EM algorithm for generalized WGMM are not proper any more. However, the procedure of the EM algorithm keeps the same except the equations for updating parameters, which should be based on the histogram (b_i and N_{b_i}) instead of each DOA sample x^i.

Based on the constructed histogram, the procedure of EM algorithm for generalized WGMM is modified and given in Algorithm.4, where $N = \sum_{i=1}^{N_{bin}} N_{b_i}$; N_{bin} representing the number of bins; and N_{b_i} is the number of samples in the ith bin of the histogram. This modified EM method based on the histogram iterates N_{bin} samples during each iteration. Note that there is one term $\frac{N_{b_i}}{P(b_i)}$ introduced in the algorithm to update the parameters. Its value is defined as NaN (Not a Number) if $N_{b_i} = 0$ and $P(b_i) \approx 0$ during updating parameters. In order to keep the numerical calculation stable, the bins without any DOA samples are omitted during each iteration. This trick reduces the complexity of numerical computation further and ensures numerical stability.

Algorithm 4: EM algorithm for Generalized Wrapped Gaussian Mixture Model based on Histogram

Initialize μ_k, α_k, σ_k^2 (k=1 .. K)

Repeat until value of L converges:

1. E-step: Compute $P(z^i = k|b_i)$ and $P(b_i)$

$$P(b_i) = \sum_{k=1}^{K} \alpha_k \mathrm{WN}(b_i; \mu_k, \sigma_k^2, T)$$

$$P(z^i = k|b_i) = \frac{\alpha_k \mathrm{WN}(b_i; \mu_k, \sigma_k^2, T)}{P(b_i)}$$

2. M-step: Update α_k, μ_k, σ_k^2. .

$$\alpha_k = \frac{\sum_{i=1}^{N_{bin}} N_{b_i} P(z^i{=}k|b_i)}{N}$$

$$\mu_k = \frac{\sum_{i=1}^{N_{bin}} \frac{N_{b_i}}{P(b_i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(b_i - wT - \mu_k)^2}{2\sigma_k^2}} (b_i - wT)}{\sum_{i=1}^{N_{bin}} \frac{N_{b_i}}{P(b_i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(b_i - wT - \mu_k)^2}{2\sigma_k^2}}}$$

$$\sigma_k^2 = \frac{\sum_{i=1}^{N_{bin}} \frac{N_{b_i}}{P(b_i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(b_i - wT - \mu_k)^2}{2\sigma_k^2}} (b_i - wT - \mu_k)^2}{\sum_{i=1}^{N_{bin}} \frac{N_{b_i}}{P(b_i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(b_i - wT - \mu_k)^2}{2\sigma_k^2}}}$$

WGMM Parameter	Value
α_1	0.4
α_2	0.6
μ_1	$260\pi/180$
μ_2	$355\pi/180$
σ_1^2	20 (degree)
σ_2^2	20 (degree)
N_{bin}	360
v_{res}	$\pi/180$

Table 7: The parameters of synthesized DOA dataset

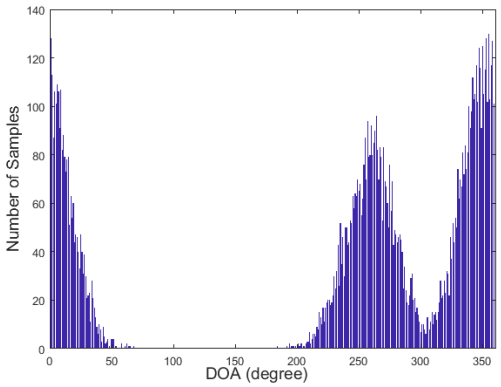

Figure 28: Histogram of synthesized 10000 DOA samples

In order to demonstrate the efficiency of the proposed algorithm, an artificial DOA dataset, which includes 2 wrapped Gaussian distributions, is synthesized, based on the settings shown in Table.7. Fig.28 visualizes the distribution of synthesized DOA samples, where the total number of samples is 10000 and 2 peaks exist near the 2 mean angular values (260 and 355 degrees). Five DOA datasets of different sizes (100, 500, 1000, 5000 and 10000 samples) are synthesized with the same mean values and variances. For each dataset, we run EM methods 20 times and measure the corresponding running time using the same computer. Fig.29 depicts the results of this measurement. With the growth of total number of samples, the running time of WGMM blows up simultaneously. However, the running time of the histogram based WGMM remains the same level no matter how big the DOA dataset is. The histogram based WGMM shows its robustness to deal with uncertain amount of samples within a stable duration. In practice, to make the entire EM algorithm finish within a small time period and fit real-time implementations, a maximal number of iterations is set.

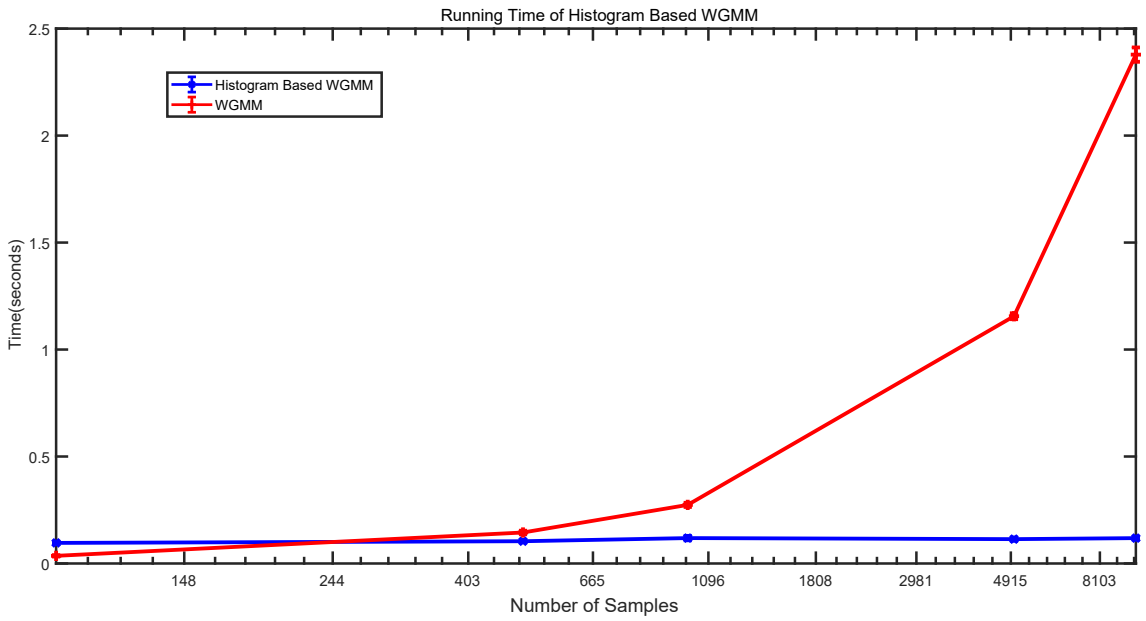

Figure 29: Running time comparison between WGMM and the histogram based WGMM

3.3.5 Proposed DOA Estimation Method

The entire procedure of the DOA estimation method proposed in this chapter is depicted in Fig.30. Here, the double input signals $s_X(n)$ and $s_Y(n)$ are processed by short time Fourier transform (STFT) firstly. Then, a filter bank is used to convert the spectrums of $s_X(n)$ and $s_Y(n)$ into the time-frequency representations $E_X(t, n_{filter})$ and $E_Y(t, n_{filter})$. After that the raw DOA estimates $\mathrm{DOA}(t, n_{filter})$ are calculated by equation (47). Meanwhile, a binary mask is formed by a thresholding method. By applying this mask to the raw DOA estimates, the unreliable DOA samples are filtered out. At last, the remaining reliable DOA samples are used to estimate a precise probability distribution by the generalized WGMM method to overcome the phase ambiguity and negative effects of white noise and reverberation.

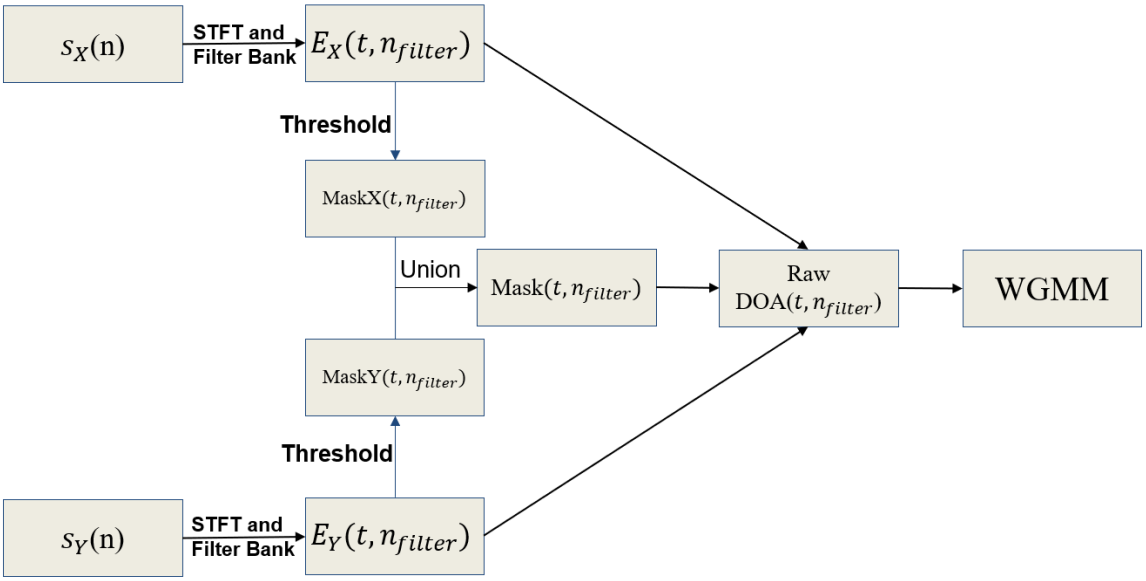

Figure 30: DOA estimation procedure

3.4 Spectral Subtraction for Noise Reduction

The DOA estimated in the previous section is used to point the main lobe of dipole beam pattern at the desired speaker. Meanwhile, the environmental noise or interference like the speeches from other speakers should be attenuated by the beamforming method. However, the white noise existing in the circuits cannot be reduced by the beamformer. It is necessary to design a procedure to do noise reduction with a low

computational complexity. In 2013, a spectral subtraction method is proposed in [42] for speech interference reduction by a research team at Graz University of Technology. However, spectral subtraction algorithms usually create isolated time-frequency blocks which introduce a musical noise into the denoised signal [61]–[63]. This kind of musical noise also degrades the audio quality. In order to reduce the bad effects of spectral subtraction like musical noise, a modified version of spectral subtraction is designed to suppress white noise and enhance audio quality.

The Entire Procedure

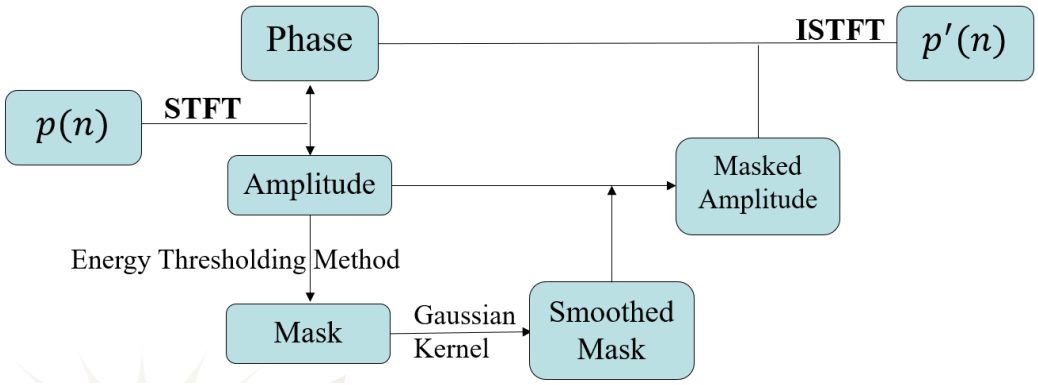

Figure 31: The entire procedure of spectral subtraction

Fig.31 depicts the entire procedure of spectral subtraction for white noise reduction. Here the STFT is used to convert the noisy signal $p(n)$, which is the output signal of steerable beamformer, into the time-frequency domain. This spectral subtraction method deals with the amplitude and phase information of noisy spectrum separately. Based on a set of threshold values, a 2D binary mask is constructed, which is smoothed by a 2D Gaussian kernel afterwards. This mask is designed to constrain the amplitude of the spectrum where white noise is dominant to an ignorable level. Then the mask is applied to the 2D amplitude of noisy speech spectrum. By adding the phase information obtained by STFT at the beginning, the denoised speech spectrum is reconstructed. According to this spectrum, the denoised audio signal $p^{'}(n)$ is generated by the inverse STFT finally.

White Noise

The spectrum of white noise has equal intensity among different frequencies, leading to a constant power spectral density. Since the white noise $n(n)$ from the fiber sensor is able to be measured easily, its time-frequency representation $|N(t,w)|$ at each frame t for each frequency w should be computed as well. Given the 2D array $|N(t,w)|$, it is easy to compute the average value $N_{avg}(w) = \frac{1}{P}\sum_{t=1}^{P}|N(t,w)|$ and standard deviation $N_{\sigma}(w) = \sqrt{\frac{1}{P}\sum_{t=1}^{P}(|N(t,w)| - N_{avg}(w))^2}$ along time frame index. These two values are used to determine the thresholding value and the corresponding 2D noise mask.

Time-Frequency Mask

The white noise is present in the circuits even if the environment is silent. Once acoustic signal is received by the fiber sensor, the amplitude of the spectrum should be above a certain level, based on which we can figure out which time-frequency position should be assigned to speech and which one should be assigned to noise only. After obtaining mean value $N_{avg}(w)$ and standard deviation $N_{\sigma}(w)$, the thresholding value for a 2D noise mask is defined as,

$$T(w) = N_{avg}(w) + c * N_{\sigma}(w) \tag{55}$$

where c is a positive constant to set how far the thresholding value $T(w)$ is greater than mean value $N_{avg}(w)$.

Given the noisy spectrum of the signal from steerable beamformer $S_{noisy}(t,w)$ for each frequency w in each time frame t, a 2D mask for noise time-frequency bins is defined as,

$$M_{noise}(t,w) = \begin{cases} 1 & |S_{noisy}(t,w)| < T(w) \\ 0 & otherwise \end{cases} \tag{56}$$

where $T(w)$ is a preset threshold.

Smoothing Mask by Gaussian Kernel

The noise mask is a binary one which may lead to audio clips after inverse Fourier transform. Here, a 2D Gaussian kernel is used to smooth the binary mask $M_{noise}(t,w)$.

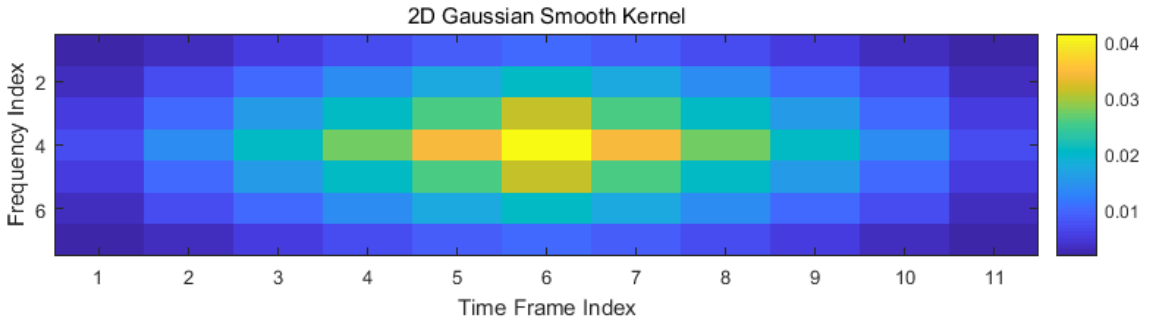

Figure 32: 2D Gaussian smooth kernel

As shown in Fig.32, a rectangular Gaussian kernel is used to smooth the binary mask $M_{noise}(t, w)$ through convolution operations. More specifically, the smooth operation is defined as,

$$M_{n-s}(t, w) = M_{noise}(t, w) * G(t, w) \tag{57}$$

where $M_{noise}(t, w)$ denotes the binary noise mask; $G(t, w)$ the 2D Gaussian smooth kernel; "*" the convolution operation; and $M_{n-s}(t, w)$ the smoothed noise mask.

Applying Smoothed 2D Mask to Spectrum

By applying 2D noise mask $M_{n-s}(t, w)$ to the amplitude of noisy spectrum $|S_{noisy}(t, w)|$, we obtain the amplitude of the denoised spectrum $|S_{denoised}(t, w)|$ as given by,

$$|S_{denoised}(t, w)| = |S_{noisy}(t, w)|(1 - M_{n-s}(t, w)) + Gain_{noise} * M_{n-s}(t, w) \tag{58}$$

where $Gain_{noise}$ is a small positive value. In order to use the inverse Fourier transform for generating the denoised audio, the phase information of the noisy audio ϕ_{noisy} is added to the amplitude of the denoised spectrum. The final reconstructed audio spectrum is expressed as,

$$S_{denoised}(t, w) = |S_{denoised}(t, w)|e^{\phi_{noisy}} \tag{59}$$

3.5 The Entire Speech Enhancement System

Fig.33 depicts the entire speech enhancement system, which includes three parts: DOA estimation, steerable beamforming and spectral subtraction for the white noise reduction. In the system, the STFT is used to convert the input signals $s_X(t)$ and

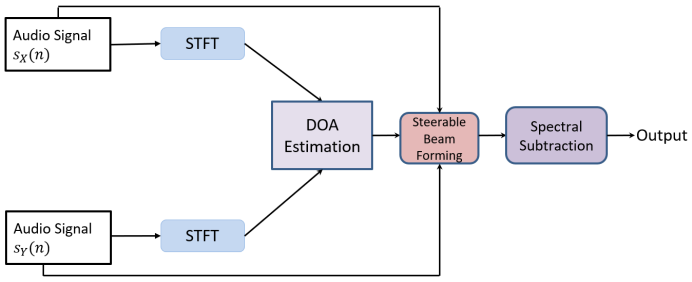

Figure 33: The speech enhancement system for X-Y fiber acoustic sensors

$s_Y(t)$ into the time-frequency representations $S_X(t,w)$ and $S_Y(t,w)$. Then, the DOA estimate is obtained by the histogram based WGMM method, which needs $S_X(t,w)$ and $S_Y(t,w)$ as the inputs. Based on the obtained DOA estimate, the steerable beamformer achieves a dipole beam pattern pointing at the speaker. At last, the spectral subtraction method is applied to the output signal of the beamformer to suppress the white noise inherent in the fiber sensors.

3.6 Experimental Results

In this section, the proposed methods for DOA estimation, steerable beamforming and noise reduction are implemented and tested with the synthesized audio files, which are generated in a virtual reverberant room. The implementation of the proposed algorithm is based on block-wise (frame-wise) processing, making it easy to transfer the simulation codes to the real-time embedded systems. In our simulation-based experiments, we choose the following parameters: sampling frequency 48kHz, hamming window function for segmentation, 128-point FFT (fast Fourier transform) size and 50% overlapping between consecutive frames.

3.6.1 Virtual Reverberant Room

As stated in section 3.3, the DOA estimation is based on the inter-channel level difference between the signals recorded by X and Y fiber sensors. For numerical simulations, the orthogonal X and Y fiber sensors are replaced with two dipole directional microphones. These two dipole microphones are placed in a virtual 3D reverberant room, which is shown in Fig.34. And the environmental acoustic parameters of the reverberant room are listed in Table.8.

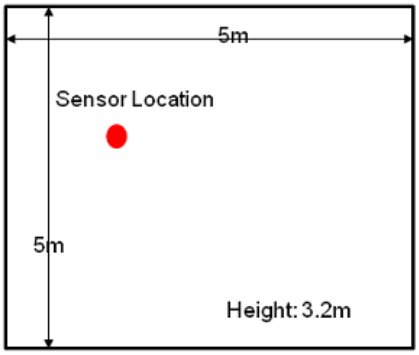

Figure 34: Virtual reverberant room layout

Acoustic Parameter	Corresponding Value
Room Size	$5m \times 5m \times 3.2m$
X-fiber and Y-fiber Location	$2m \times 1.5m \times 1.6m$
X-fiber Orientation	0
Y-fiber Orientation	$\frac{\pi}{2}$
Reverberation Time T_{60}	400 ms (milliseconds)
Sound Speed	340 m/s

Table 8: Acoustic parameters of virtual reverberant room

3.6.2 DOA Estimation Results

In simulations, a speaker is assumed to talk at 0, 45, 90 and 135 degrees respectively. The distance between the speaker and X-Y sensors is 1.5 meters. We also assume that the sound source and X-Y sensors are placed in the same 2D plane. At each location, the room impulse responses for double channel directional microphones are generated and used to synthesize the signals received by X and Y fiber sensors as described by,

$$s_X(n) = h_x(n) * s(n) + n_x(n) \tag{60}$$

$$s_Y(n) = h_y(n) * s(n) + n_y(n) \tag{61}$$

where $s(n)$ denotes the clean speech signal; $h_x(n)$ and $h_y(n)$ denote the room impulse responses for X and Y fiber sensors and $n_x(n)$ and $n_y(n)$ denote the white noises existing in X and Y fiber sensors, respectively.

Simplification of WGMM

If there is only one active speaker, the WGMM can be simplified. The simplified version of WGMM contains the parameters to be estimated i.e., μ_1, σ_1^2. And μ_1 represents the DOA estimate. σ_1^2 usually implies the intensity of reverberation, according to our empirical simulation results. A stronger reverberant environment usually leads to a bigger variance value σ_1^2. The simplified algorithm of the histogram based WGMM method is shown in Algorithm.5.

Algorithm 5: Simplified EM algorithm for the Histogram based WGMM with One Speaker

Initialize μ_1,σ_1^2

Repeat until value of L converges:

1. E-step: Compute $P(b_i)$.

$$P(b_i) = \mathrm{WN}(b_i; \mu_1, \sigma_1^2, \pi)$$

2. M-step: Update μ_1, σ_1^2.

$$\mu_1 = \frac{\sum_{i=1}^{N_{bin}} \frac{N_{b_i}}{P(b_i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(b_i - w\pi - \mu_1)^2}{2\sigma_1^2}} (b_i - w\pi)}{\sum_{i=1}^{N_{bin}} \frac{N_{b_i}}{P(b_i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(b_i - w\pi - \mu_1)^2}{2\sigma_1^2}}}$$

$$\sigma_1^2 = \frac{\sum_{i=1}^{N_{bin}} \frac{N_{b_i}}{P(b_i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(b_i - w\pi - \mu_1)^2}{2\sigma_1^2}} (b_i - w\pi - \mu_1)^2}{\sum_{i=1}^{N_{bin}} \frac{N_{b_i}}{P(b_i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(b_i - w\pi - \mu_1)^2}{2\sigma_1^2}}}$$

DOA Estimation Results using WGMM

In numerical simulations, the WGMM method runs frame by frame for 4 different speaker location settings (0, 45, 90 and 135 degrees). The results of DOA estimation are shown in Fig.35, Fig.36, Fig.37 and Fig.38, where the DOA estimates obtained by the WGMM method are super close to the ground truth highlighted by the red target lines.

DOA Estimation Results using Histogram Based WGMM

Similar with the simulations for the WGMM method, here we simulate the histogram based WGMM method for DOA estimation. The results are shown in Fig.39, Fig.40, Fig.41 and Fig.42. About the performance of these two DOA estimation methods (WGMM and histogram based WGMM), there is no significant difference observed

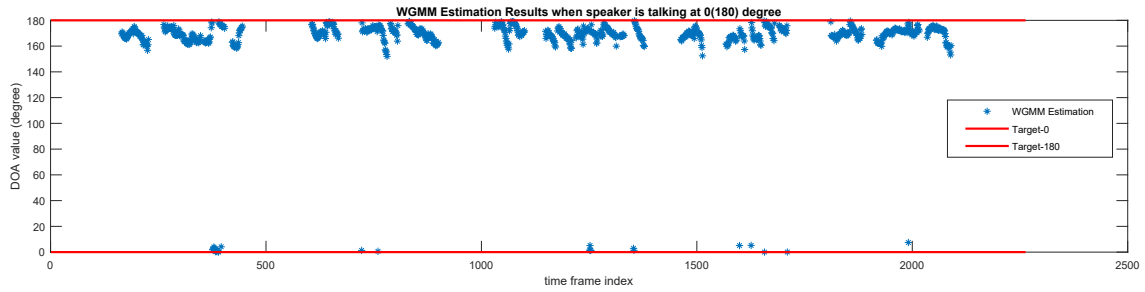

Figure 35: DOA estimation results achieved by WGMM (DOA is 0 degree)

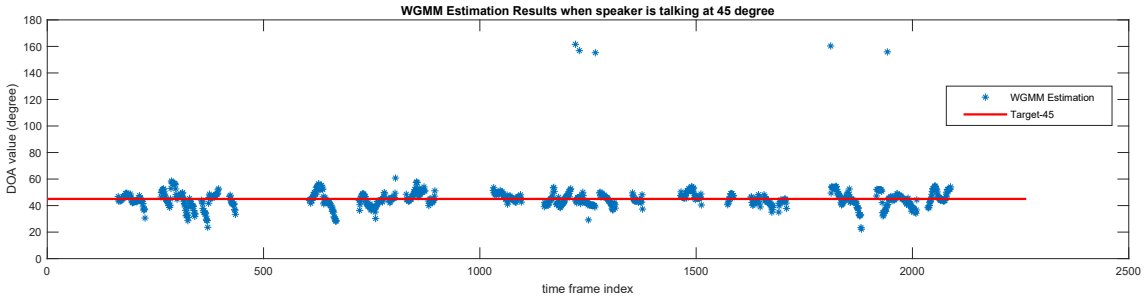

Figure 36: DOA estimation results achieved by WGMM (DOA is 45 degrees)

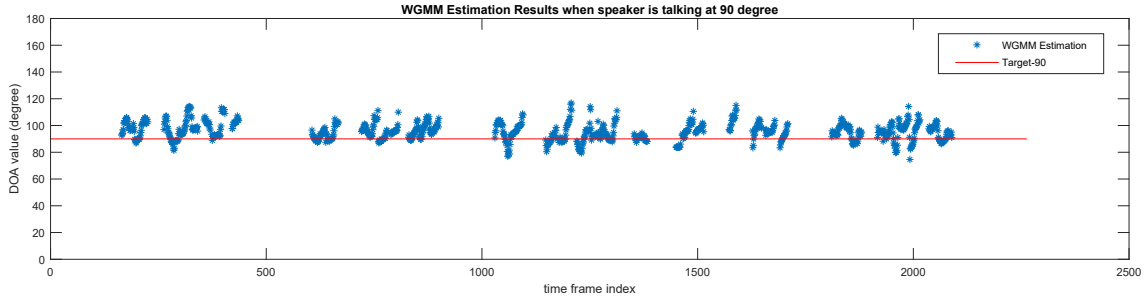

Figure 37: DOA estimation results achieved by WGMM (DOA is 90 degrees)

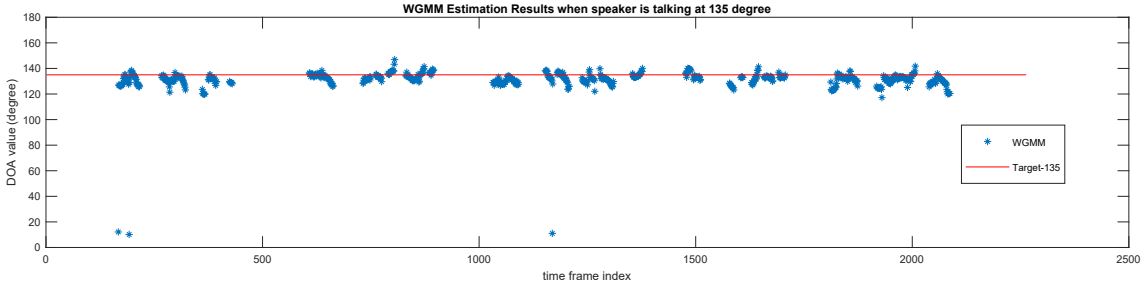

Figure 38: DOA estimation results achieved by WGMM (DOA is 135 degrees)

	MWAE	
DOA	WGMM	histogram based WGMM
0	10.0213	10.4941
45	4.1908	4.5641
90	6.8468	6.5669
135	4.1285	5.1222

Table 9: Performance comparison

from the resulting figures. In order to conduct a numerical comparison, a metric called mean wrapped absolute error (MWAE) is adopted here. The definition of MWAE is given by,

$$\text{MWAE}(\theta_e, \theta_t; T) = min\{abs(\theta_e - \theta_t), abs(\theta_e - \theta_t + T), abs(\theta_e - \theta_t - T)\} \quad (62)$$

where θ_e denotes the DOA estimate; θ_t denotes the ground truth DOA; and T is the angular period fixed as 180 degrees.

The MWAE values are given in Table.9, showing that the performance difference between WGMM and histogram based WGMM is less than 1 degree. It implies that the histogram based WGMM method is a good option for real-time computing with little precision loss.

Effects of Reverberation on DOA Estimation

Here, more numerical experiments are conducted under two different room configurations, 136 true DOA angles ranging from 0 to 135 degrees and 5 different reverberant intensities (wall reflection coefficients: 0%, 25%, 50%, 65%, and 75%). Note that the white noise is not considered in the fiber acoustic sensors. In order to measure how the histogram based WGMM method performs, the MWAE and frame-wise accuracy are used as performance metrics. The frame-wise accuracy is defined as,

$$Acc(\%) = 100\frac{N_c}{N_f}\% \quad (63)$$

where N_f denotes the number of total frames; and N_c the number of frames with a correct DOA estimation. For each frame, the DOA estimate θ_e is considered as correct if $\text{MWAE}(\theta_e, \theta_t) < 5$ degrees, where θ_t is the ground truth.

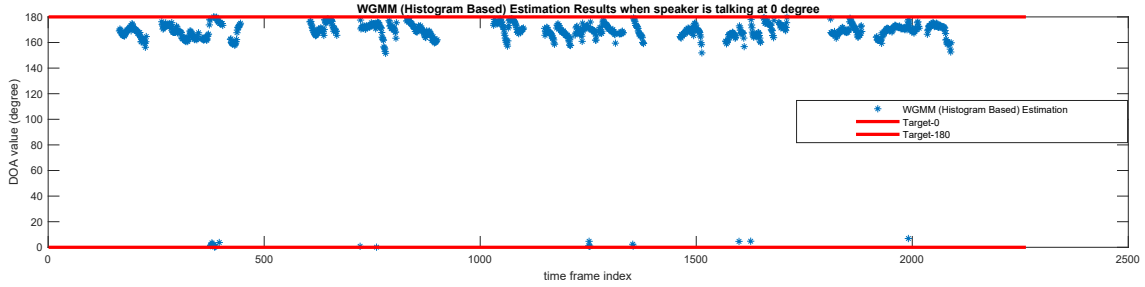

Figure 39: DOA estimation results achieved by the histogram based WGMM (DOA is 0 degree)

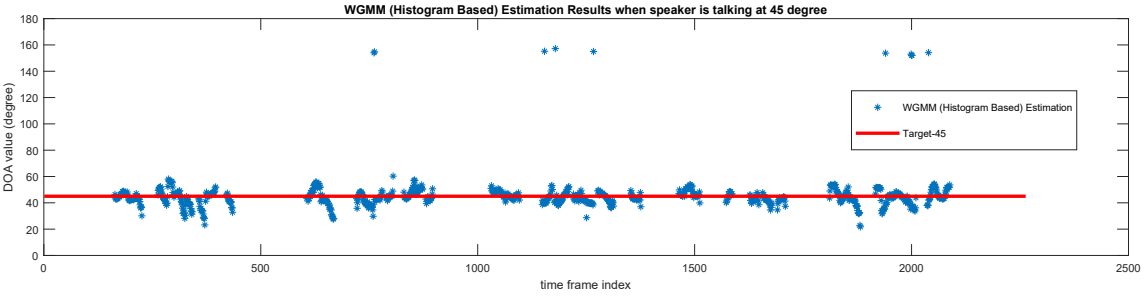

Figure 40: DOA estimation results achieved by the histogram based WGMM (DOA is 45 degrees)

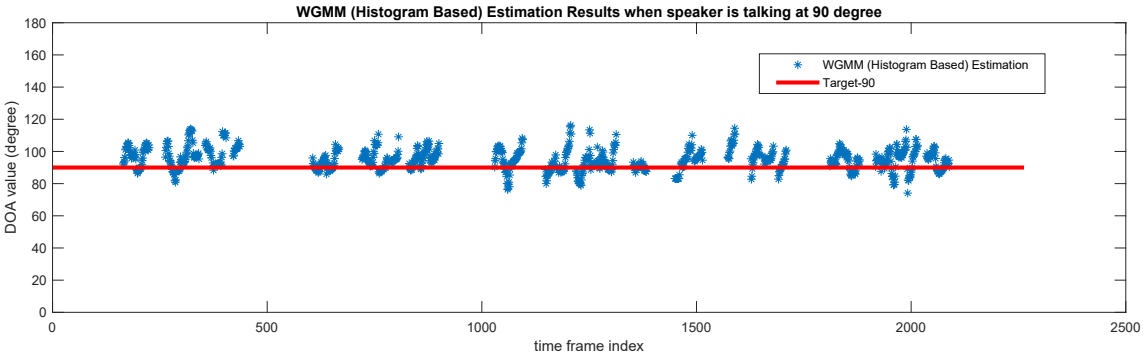

Figure 41: DOA estimation results achieved by the histogram based WGMM (DOA is 90 degrees)

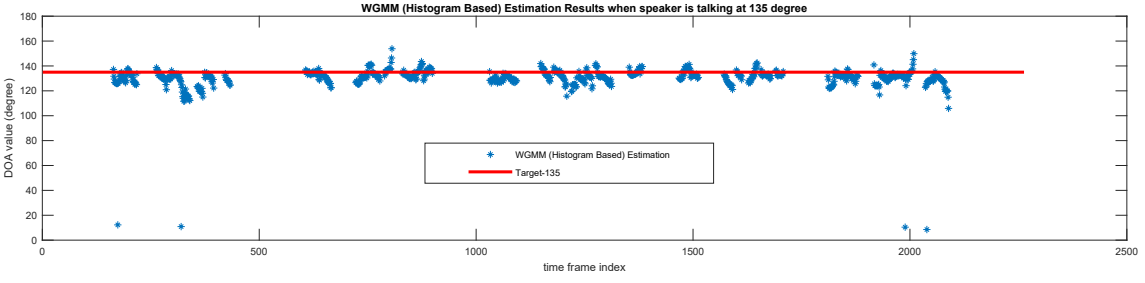

Figure 42: DOA estimation results achieved by the histogram based WGMM (DOA is 135 degrees)

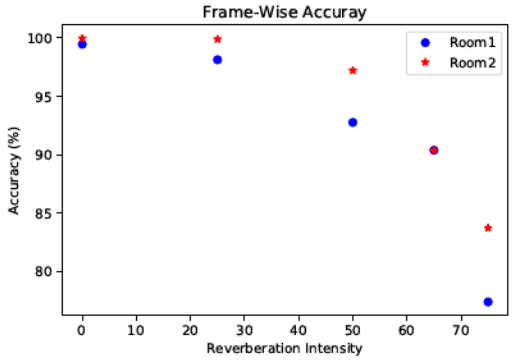

Figure 43: The frame-wise accuracy

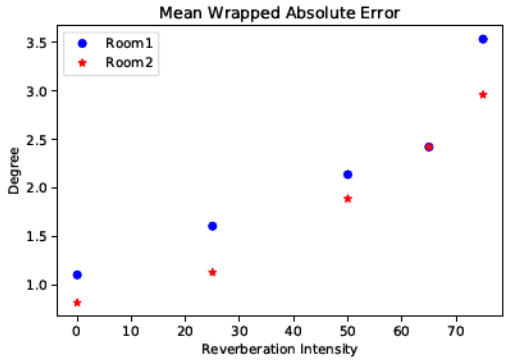

Figure 44: The mean wrapped absolute error

As shown in Fig.43 and Fig.44, with the increasing intensity of reverberation, the MWAE becomes larger and the frame-wise accuracy decreases. It can be concluded that although the reverberation degrades the performance of the histogram based WGMM, the proposed method is still able to achieve 3.5 degrees MWAE and 75% frame-wise accuracy under the highest level of reverberation. This fact shows that the histogram based WGMM is a robust DOA estimation under different reverberant conditions.

3.6.3 Spectral Subtraction Results

Given a DOA θ, the signal received by X-fiber $s_X(n)$ and that by Y-fiber $s_Y(n)$, the output signal $p(n)$ from the steerable beamformer is expressed as,

$$p(n) = cos(\theta) \cdot s_X(n) + sin(\theta) \cdot s_Y(n) \tag{64}$$

Due to the presence of white noise, the signal $s_X(n)$ is corrupted by white noise $n_x(n)$, whose variance is $\sigma^2_{n_x}$. And the signal $s_Y(n)$ is also corrupted by white noise $n_y(n)$, whose variance is $\sigma^2_{n_y}$. Then, the variance of white noise contaminating signal $p(n)$ is derived as,

$$\sigma^2_{n_p} = cos^2(\theta) \cdot \sigma^2_{n_x} + sin^2(\theta) \cdot \sigma^2_{n_y} \tag{65}$$

Since the white noise $n_y(n)$ and $n_x(n)$ are generated in the same circuit, they have the same variance σ^2. Then equation (64) is simplified as $\sigma^2_{n_p} = cos^2(\theta) \cdot \sigma^2_{n_x} + sin^2(\theta) \cdot \sigma^2_{n_y} = [cos^2(\theta) + sin^2(\theta)]\sigma^2 = \sigma^2$. It implies that the white noise of the output signal

$p(n)$ from the steerable beamformer can be measured by using either the X or Y fiber sensor in a silent environment.

With the prior knowledge about the white noise in the output signal $p(n)$, the spectral subtraction method introduced in section 3.4 is applied to $p(n)$ for each simulation. For example, when one speaker is talking at 0 degree, the spectrums of the output signal from beamformer and the denoised signal are shown in Fig.45. And the noisy waveform and denoised waveform are shown in Fig.46. It is apparent that the white noise is suppressed significantly.

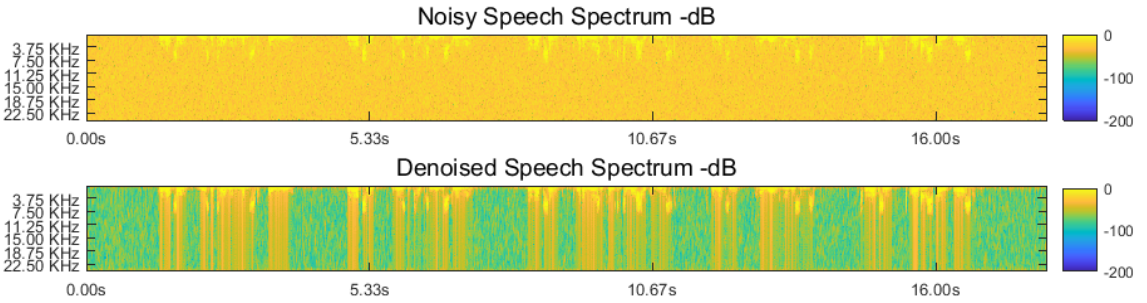

Figure 45: The spectrum of denoised signal (DOA is 0 degree)

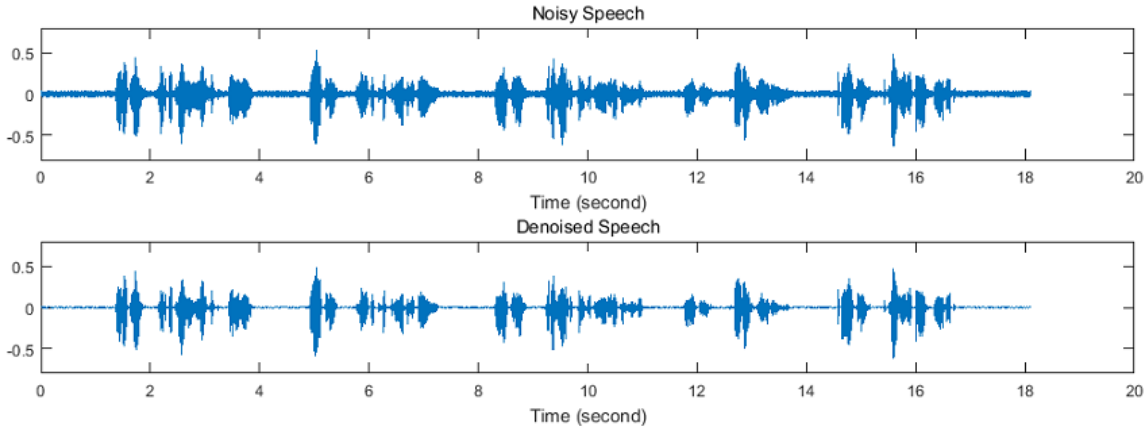

Figure 46: The waveform of denoised signal (DOA is 0 degree)

In order to measure how the audio quality is improved, the wideband PESQ [48] score is used. And the resulting scores are listed in Table.10. Based on the PESQ scores of the denoised signals, the spectral subtraction scheme improves the audio quality by at least 50% .

	PESQ (wideband)			
DOA	X fiber sensor	Y fiber sensor	Denoised	Improvement
0	1.30571198	1.11912858	2.02172279	54.84%
45	1.20795333	1.20616138	1.93385077	60.09%
90	1.13096237	1.23799229	1.95860767	58.21%
135	1.22650313	1.19694507	2.09664202	70.94%

Table 10: PESQ score

3.7 Conclusion

In this chapter, a collocated microphone array, comprised of double collocated fiber sensors, named as X-Y sensors, has been proposed. Based on the properties of X-Y sensors, a speech enhancement system is implemented. The entire procedure of this system includes three parts: DOA estimation, steerable beamformer and spectral subtraction for white noise reduction. In the first part, based on the inter-channel level difference, the DOA estimates are calculated in time-frequency domain. An unsupervised learning method, namely generalized WGMM, is proposed to solve phase ambiguity problem between 0 and 180 degrees for DOA estimation. It has been shown that this method is able to obtain an accurate DOA estimate and is robust to reverberation and white noise. Meanwhile, a histogram based WGMM method is proposed to reduce the computational complexity for large datasets with a minor precision loss. In the second part, once the accurate DOA estimate is obtained, the steerable beamformer points the main lobe of a dipole beam pattern at the speaker. In the third part, the spectral subtraction method is adopted to suppress the white noise generated by the circuits. According to the results of numerical simulations in a virtual reverberant room, the audio quality of the denoised signal is improved by at least 50% overall, indicating a significant enhancement performance resulting from the proposed system.

Chapter 4

Conclusion and Future Work

4.1 Conclusion

In this thesis, several microphone array techniques have been studied. First, a new directional microphone called fiber acoustic sensor is introduced. It senses the acoustic flow by measuring the vibration of a nano-scale fiber, making it possible to design an acoustic particle velocity microphone with a frequency-independent dipole beam pattern. In order to investigate the possible speech enhancement applications using this new microphone, two kinds of collocated microphone array are constructed and corresponding numerical simulations are conducted. These two types of collocated microphone can be regarded as the simplified versions of acoustic vector sensor.

The first type of microphone array consists of one omni microphone collocated with one fiber acoustic sensor. This design is inspired by the conventional first-order adaptive differential microphone array. Since the dipole beam pattern achieved by the fiber acoustic sensor is frequency-independent, the structure of the proposed collocated microphone array is simpler without differential operation and low pass filtering, which is necessary in the first-order adaptive differential microphone array. Numerical simulations are conducted to demonstrate the effectiveness of this design for speech interference reduction and speech separation. However, according to our numerical results, it still needs investigations to make this design more efficient in the presence of reverberation and white noise.

The second type of microphone array designed in this thesis consists of double orthogonal and collocated fiber acoustic sensors. Utilizing this microphone array, a

speech enhancement system, which includes three parts: DOA estimation, steerable beamforming and spectral subtraction, is implemented. More specifically, with these two directive and orthogonal dipole beam patterns, DOA estimates are extracted by inter-channel level difference method in time-frequency domain. Due to the existence of white noise and reverberation, the distribution of DOA samples is modeled by the wrapped Gaussian distribution. Furthermore, in the angular space, the generalized wrapped Gaussian mixture model (WGMM) is proposed to achieve DOA estimation and solve the phase ambiguity problem between 0 and 180 degrees. This proposed method is also able to estimate the DOA of multiple speakers simultaneously. In order to reduce the computational complexity of the WGMM method, a histogram based WGMM approach with a lower time complexity and less memory requirement is proposed. After obtaining the reliable DOA estimates by the WGMM method, a steerable beamfomer is used to point the main lobe of a dipole beam pattern to the desired speaker. As there is white noise existing in the fiber acoustic sensor, the output audio from the beamformer is degraded. So a spectral subtraction method is proposed to suppress the white noise and enhance the quality of output signal from the steerable beamformer. Numerical experiments are conducted in the presence of reverberation and white noise, whose results show that this proposed system can capture the speech from any direction and improve the audio quality by above 50% in terms of the wideband PESQ score as the performance metric.

4.2 Future Work

4.2.1 Algorithm Deployment on Embedded System

The algorithms designed in this thesis are implemented in MATLAB for simulations. However, the potential microphone product will need to run the algorithms on the embedded system. There is still a gap between the codes for simulations running on laptops and that for embedded system. Due to the property of the embedded system, the audio input signal is usually recorded by buffer-wise manner. It means that the digital signal processing code needs to process a fixed-length array of samples at each time. It needs more work if the buffer size is not consistent with the number of FFT (fast Fourier transform) points. This practical implementation issue needs to be addressed before shifting these algorithms to a hardware platform.

4.2.2 Multiple-Speaker Localization and Separation

Although the WGMM used in this thesis has the ability to deal with multiple-speaker localization problems, the number of speakers should be known before using it. This precondition usually limits the deployment of WGMM on a real product. For multiple speakers, the authors in paper [64] proposed a novel sparse source separation method that is able to estimate the number of sound sources. Specifically, the proposed method utilizes a sparse distribution modeled by the Dirichlet distribution as the prior of the WGMM mixture weight [64]. After obtaining the DOA estimates of multiple speakers, the beamforming is used to separate the mixture of speeches. This approach should be promising in the X-Y collocated fiber acoustic sensors as well. The future work may include some investigations about deploying this method on the X-Y fiber sensors.

4.2.3 Multiple Modal Speech Separation

Nowdays, the mobile device like cellphone is usually equipped with multiple microphones and cameras. Apparently, double or even triple cameras become common in the commercial cellphone products like iPhone 11 or higher version. Multi-Camera system makes the 3D vision available. When people are shooting a video by such devices, both camera and audio systems are able to locate the speakers. It is possible to fuse the speaker position information obtained from 3D vision and spatial audio. This technique is useful for speech separation and enhancement. Researchers from Snap Inc call this multi-modal speech processing technique as "Audiovisual Zooming" [65]. This framework is built on the top of the classic ideas of 3D computer vision and audio beamforming and a computational approach is proposed to enhance sound from a single direction using a microphone array [65]. Their results show that using their system, the voice of a desired speaker in the video is enhanced and other interference speech is attenuated. This idea enhances the user experience for a conference even in a noisy environment. Following this idea, it is also possible to build a system using the X-Y collocated fiber sensors with a stereo camera. For this future work, a new coding standard about video recording for audio-visual speech separation should be invented.

Appendix A

Appendix

A.1 Taylor Series Expansion

The Taylor series expansion for exponential function e^x, where x is close to 0, is given by,

$$e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + ... + \frac{1}{n!}x^n + ... \tag{66}$$

A.2 First-Order Forward-facing Cardioid

The first-order forward-facing cardioid is given by,

$$C_F(w,\theta) = \begin{pmatrix} 1 & e^{-jwt_0 cos\theta} \end{pmatrix} \begin{pmatrix} 1 \\ -e^{-jwt_0} \end{pmatrix} = 1 - e^{-jwt_0(1+cos\theta)} \tag{67}$$

where w represents the frequency; and t_0 represents the constant time delay $\frac{d}{c}$ between 2 adjacent omni microphones.

Using Taylor series expansion above and noting that t_0 is an extremely small constant, $C_F(w,\theta)$ can be expanded around 0 as,

$$C_F(w,\theta) = 1 - (1 - jwt_0(1 + cos\theta) + O((-jwt_0(1 + cos\theta))^2) \tag{68}$$

By omitting the extremely small second-order term $O((-jwt_0(1+cos\theta))^2)$ in equation (68), the expression of $C_F(w,\theta)$ is simplified as,

$$C_F(w,\theta) \approx jwt_0(1 + cos\theta) \tag{69}$$

A.3 First-Order Back-facing Cardioid

The first-order back-facing cardioid is given by,

$$C_B(w,\theta) = \begin{pmatrix} 1 & e^{-jwt_0 cos\theta} \end{pmatrix} \begin{pmatrix} -e^{-jwt_0} \\ 1 \end{pmatrix} = -e^{-jwt_0} + e^{-jwt_0 cos\theta} \tag{70}$$

By using the Taylor series expansion in equation (66), the expression of $C_B(w,\theta)$ is expanded as,

$$C_B(w,\theta) = -(1 - jwt_0 + O((wt_0)^2)) + (1 - jwt_0 cos\theta) + O((wt_0 cos\theta)^2) \tag{71}$$

By omitting the extremely small second-order terms $O((wt_0)^2)$ and $O((wt_0 cos\theta)^2)$, $C_B(w,\theta)$ is simplified further as,

$$C_B(w,\theta) \approx jwt_0(1 - cos\theta) \tag{72}$$

A.4 Derivation of Wrapped Gaussian Mixture Model

A.4.1 Wrapped Gaussian Distribution

Regarding the clusters as a random variable z, for each sample x^i, $z^i = k$ if this sample belongs to cluster k.

Assuming there are K clusters and each cluster is distributed as wrapped Gaussian distribution, then the probability density function (PDF) is given by the following equation:

$$\begin{aligned} P(x^i) &= \sum_{k=1}^{K} P(x^i|z^i = k)P(z^i = k) \\ &= \sum_{k=1}^{K} \mathrm{WN}_k(x^i; \mu_k, \sigma_k^2)\alpha_k \end{aligned}$$

where α_k represents the weight of each cluster and $\mathrm{WN}_k(x^i; \mu_k, \sigma_k^2)$ represents the PDF of the wrapped Gaussian distribution with mean μ_k and variance σ_k^2. Note that μ_k ranges within the interval $[0, 2\pi)$.

$$\mathrm{WN}_k(x^i; \mu_k, \sigma_k^2) = \sum_{w \in Z} \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x_i - 2w\pi - \mu_k)^2}{2\sigma_k^2}}$$

A.4.2 Objective Function of Maximum Log-likelihood Estimation

The objective of maximum log-likelihood estimation is defined as,

$$L = \sum_{i=1}^{N} logP(x^i)$$

In order to obtain the optimal parameters θ, which include μ_k, σ_k^2 and α_k, we need to take the derivative of L as follows.

$$\frac{dL}{d\theta} = \sum_{i=1}^{N} \frac{d}{d\theta} logP(x^i)$$

Using $\frac{d}{d\theta} logm = \frac{1}{m}\frac{dm}{d\theta}$ (here log is the natural logarithm) and $P(x^i) = \sum_z P(z, x^i)$, $\frac{dL}{d\theta}$ can be computed by,

$$\frac{dL}{d\theta} = \sum_{i=1}^{N} \frac{1}{\sum_{z'} P(z', x^i)} \sum_z \frac{d}{d\theta} P(z, x^i)$$

Since $\frac{d}{d\theta} P(z, x^i) = P(z, x^i)\frac{d}{d\theta} logP(z, x^i)$, then we have:

$$\frac{dL}{d\theta} = \sum_{i=1}^{N} \frac{\sum_z \frac{d}{d\theta} P(z, x^i)}{\sum_{z'} P(z', x^i)}$$

$$= \sum_{i=1}^{N} \frac{\sum_z P(z, x^i)\frac{d}{d\theta} logP(z, x^i)}{\sum_{z'} P(z', x^i)}$$

$$= \sum_{i=1}^{N} \sum_z \frac{P(z, x^i)}{P(x^i)} \frac{d}{d\theta} logP(z, x^i)$$

$$= \sum_{i=1}^{N} \sum_z P(z|x^i) \frac{d}{d\theta} logP(z, x^i)$$

Hence, the objective function can be equivalently defined as,

$$L = \sum_{i=1}^{N} \sum_{k=1}^{K} P(z^i = k|x^i)[logP(z^i = k, x^i)]$$

Since $P(z^i = k, x^i) = P(z^i = k)P(x^i|z^i = k)$, the objective function can be rewritten as,

$$L = \sum_{i=1}^{N}\sum_{k=1}^{K} P(z^i = k|x^i)[logP(z^i = k) + logP(x^i|z^i = k)]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} P(z^i = k|x^i)[log\alpha_k + log\mathrm{WN}_k]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} R_k^i[log\alpha_k + log\mathrm{WN}_k]$$

where R_k^i is used to represent term $P(z^i = k|x^i)$. So the problem is very much simplified as an optimization problem involving parameters α_k, μ_k and σ_k^2, namely,

$$\underset{\alpha_k,\mu_k,\sigma_k^2}{argmax}\{L\}$$

A.4.3 Estimating Mean

In order to estimate μ_k, we can derive as follows,

$$\frac{\partial L}{\partial \mu_k} = \frac{\partial}{\partial \mu_k}\sum_{i=1}^{N}\sum_{k'=1}^{K} R_{k'}^i[log\alpha_{k'} + log\mathrm{WN}_{k'}]$$

$$= \sum_{i=1}^{N} R_k^i \frac{\partial}{\partial \mu_k}[log\alpha_k + log\mathrm{WN}_k]$$

$$= \sum_{i=1}^{N} R_k^i \frac{\partial}{\partial \mu_k} log\mathrm{WN}_k$$

$$= \sum_{i=1}^{N} P(z^i = k|x^i)\frac{\partial}{\partial \mu_k} logP(x^i|z^i = k)$$

$$= \sum_{i=1}^{N} \frac{P(z^i = k, x^i)}{P(x^i)}\frac{1}{P(x^i|z^i = k)}\frac{\partial}{\partial \mu_k}P(x^i|z^i = k)$$

$$= \sum_{i=1}^{N} \frac{P(z^i = k, x^i)}{P(x^i)}\frac{P(z^i = k)}{P(x^i, z^i = k)}\frac{\partial}{\partial \mu_k}P(x^i|z^i = k)$$

$$= \sum_{i=1}^{N} \frac{P(z^i = k)}{P(x^i)}\frac{\partial}{\partial \mu_k}P(x^i|z^i = k)$$

$$= \sum_{i=1}^{N} \frac{\alpha_k}{P(x^i)} \sum_{w=-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x^i-2w\pi-\mu_k)^2}{2\sigma_k^2}} \frac{(x^i-2w\pi-\mu_k)}{\sigma_k^2}$$

$$= \alpha_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \frac{1}{\sigma_k^2} \sum_{i=1}^{N} \frac{1}{P(x^i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(x^i-2w\pi-\mu_k)^2}{2\sigma_k^2}} (x^i-2w\pi-\mu_k)$$

By setting $\frac{\partial L}{\partial \mu_k} = 0$, we obtain the expression for μ_k as given by,

$$\mu_k = \frac{\sum_{i=1}^{N} \frac{1}{P(x^i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(x^i-2w\pi-\mu_k)^2}{2\sigma_k^2}} (x^i-2w\pi)}{\sum_{i=1}^{N} \frac{1}{P(x^i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(x^i-2w\pi-\mu_k)^2}{2\sigma_k^2}}}$$

A.4.4 Estimating Variance

The way to estimate σ_k^2 is similar with that of estimating μ_k. By calculating the first order derivative of L, we have

$$\frac{\partial L}{\partial \sigma_k^2} = \sum_{i=1}^{N} \sum_{k'=1}^{K} R_{k'}^i \frac{\partial}{\partial \sigma_k^2} log\mathrm{WN}_{k'}$$

$$= \sum_{i=1}^{N} \frac{P(z^i=k)}{P(x^i)} \frac{\partial}{\partial \sigma_k^2} \sum_{w=-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x^i-2w\pi-\mu_k)^2}{2\sigma_k^2}}$$

$$= \sum_{i=1}^{N} \frac{\alpha_k}{P(x^i)} \sum_{w=-\infty}^{\infty} [V_1 + V_2]$$

where

$$V_1 = \frac{1}{\sqrt{2\pi}} (-0.5)(\sigma_k^2)^{-1.5} e^{-\frac{(x^i-2w\pi-\mu_k)^2}{2\sigma_k^2}}$$

$$V_2 = \frac{1}{\sqrt{2\pi}} (\sigma_k^2)^{-0.5} e^{-\frac{(x^i-2w\pi-\mu_k)^2}{2\sigma_k^2}} (-0.5(x^i-2w\pi-\mu_k)^2)(-1)(\sigma_k^2)^{-2}$$

$$\frac{\partial L}{\partial \sigma_k^2} = \frac{0.5\alpha_k}{\sqrt{2\pi}} (\sigma_k^2)^{-2.5} [\sum_{i=1}^{N} \frac{1}{P(x^i)} \sum_{w=-\infty}^{\infty} (e^{-\frac{(x^i-2w\pi-\mu_k)^2}{2\sigma_k^2}} (x^i-2w\pi-\mu_k)^2 - \sigma_k^2 e^{-\frac{(x^i-2w\pi-\mu_k)^2}{2\sigma_k^2}})]$$

Let $\frac{\partial L}{\partial \sigma_k^2}$ equal 0, we can obtain the expression for σ_k^2.

$$\sigma_k^2 = \frac{\sum_{i=1}^{N} \frac{1}{P(x^i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(x^i - 2w\pi - \mu_k)^2}{2\sigma_k^2}} (x^i - 2w\pi - \mu_k)^2}{\sum_{i=1}^{N} \frac{1}{P(x^i)} \sum_{w=-\infty}^{\infty} e^{-\frac{(x^i - 2w\pi - \mu_k)^2}{2\sigma_k^2}}}$$

A.4.5 Estimating Cluster Weight

The following optimization problem is formulated to compute the weight of the kth cluster α_k.

$$\begin{aligned} \underset{\alpha_k}{\text{maximize}} \quad & \sum_{i=1}^{N} \sum_{k=1}^{K} P(z^i = k|x^i) logP(x^i, z^i = k) \\ \text{subject to} \quad & \sum_{k=1}^{K} \alpha_k = 1 \end{aligned}$$

Using the Lagrange multiplier, the problem is rephrased as,

$$\underset{\alpha_k, \lambda}{\text{maximize}} \quad \sum_{i=1}^{N} \sum_{k=1}^{K} P(z^i = k|x^i) logP(x^i, z^i = k) + \lambda(1 - \sum_{k=1}^{K} \alpha_k)$$

Here the cost function $L_c = [\sum_{i=1}^{N} \sum_{k=1}^{K} P(z^i = k|x^i) logP(x^i, z^i = k)] + \lambda(1 - \sum_{k=1}^{K} \alpha_k)$, is regarded as the objective function to obtain optimal solution of α_k. The following procedure shows how the partial derivative of L_c is obtained.

$$\frac{\partial L_c}{\partial \alpha_k} = \frac{\partial}{\partial \alpha_k} [\sum_{i=1}^{N} \sum_{k'=1}^{K} P(z^i = k'|x^i) logP(x^i, z^i = k')] + \frac{\partial}{\partial \alpha_k} [\lambda(1 - \sum_{k'=1}^{K} \alpha'_k)]$$

$$\frac{\partial L_c}{\partial \alpha_k} = \frac{\partial}{\partial \alpha_k} [\sum_{i=1}^{N} \sum_{k'=1}^{K} P(z^i = k'|x^i) [log(\alpha_{k'}) + logP(x^i|z^i = k')]] + \frac{\partial}{\partial \alpha_k} [\lambda(1 - \sum_{k'=1}^{K} \alpha'_k)]$$

$$= [\sum_{i=1}^{N} \frac{P(z^i = k|x^i)}{\alpha_k}] - \lambda$$

Let $\frac{\partial L_c}{\partial \alpha_k}$ equal 0, we can have:

$$\lambda = \frac{\sum_{i=1}^{N} P(z^i = k|x^i)}{\alpha_k}$$

Then we know $\alpha_k \propto \sum_{i=1}^{N} P(z^i = k|x^i)$, it is easy to obtain the formula to calculate α_k.

$$\alpha_k = \frac{\sum_{i=1}^{N} P(z^i = k|x^i)}{\sum_{i=1}^{K}\sum_{i=1}^{N} P(z^i = k|x^i)}$$

$$= \frac{\sum_{i=1}^{N} P(z^i = k|x^i)}{\sum_{i=1}^{N}\sum_{i=1}^{K} P(z^i = k|x^i)}$$

$$= \frac{\sum_{i=1}^{N} P(z^i = k|x^i)}{\sum_{i=1}^{N} 1}$$

$$\alpha_k = \frac{\sum_{i=1}^{N} P(z^i = k|x^i)}{N}$$

Bibliography

[1] M. Shujau, "In air acoustic vector sensors for capturing and processing of speech signals," PhD thesis, 2011.

[2] M. Jian, A. C. Kot, and M. Er, "Doa estimation of speech source with microphone arrays," in *Proceedings of the 1998 IEEE International Symposium on Circuits and Systems (Cat. No. 98CH36187)*, vol. 5, pp. 293–296.

[3] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[4] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[5] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 375–378.

[6] B. Qin, H. Zhang, Q. Fu, and Y. Yan, "Subsample time delay estimation via improved gcc phat algorithm," in *2008 9th International Conference on Signal Processing*, IEEE, pp. 2579–2582.

[7] A. Brutti, M. Omologo, and P. Svaizer, "Comparison between different sound source localization techniques based on a real data collection," in *2008 Hands-Free Speech Communication and Microphone Arrays*, IEEE, pp. 69–72.

[8] S. Lee, Y. Park, and Y.-s. Park, "Cleansed phat gcc based sound source localization," in *International Conference on Control, Automation and Systems*, IEEE, 2010, pp. 2051–2054.

[9] B. Kwon, Y. Park, and Y.-s. Park, "Analysis of the gcc-phat technique for multiple sources," in *International Conference on Control, Automation and Systems*, IEEE, 2010, pp. 2070–2073.

[10] S. U. Wood, J. Rouat, S. Dupont, and G. Pironkov, "Blind speech separation and enhancement with gcc-nmf," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 745–755, 2017.

[11] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," PhD thesis, 2000.

[12] H. Do and H. F. Silverman, "A fast microphone array srp-phat source location implementation using coarse-to-fine region contraction (cfrc)," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 295–298.

[13] ——, "Stochastic particle filtering: A fast srp-phat single source localization algorithm," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 213–216.

[14] ——, "Srp-phat methods of locating simultaneous multiple talkers using a frame of microphone array data," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 125–128.

[15] M. Cobos, A. Marti, and J. J. Lopez, "A modified srp-phat functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 71–74, 2010.

[16] D. Kundu, "Modified music algorithm for estimating doa of signals," *Signal Processing*, vol. 48, no. 1, pp. 85–90, 1996.

[17] W. A. Gardner, "Simplification of music and esprit by exploitation of cyclostationarity," *Proceedings of the IEEE*, vol. 76, no. 7, pp. 845–847, 1988.

[18] J. S. Thompson, P. M. Grant, and B. Mulgrew, "Performance of spatial smoothing algorithms for correlated sources," *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 1040–1046, 1996.

[19] P. Wang, P.-P. Wang, G.-j. Zhang, and J.-j. Xiong, "Spatial smoothing algorithm based on acoustic vector sensor array," in *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, IEEE, vol. 14, pp. V14–27.

[20] H. Hwang, Z. Aliyazicioglu, M. Grice, and A. Yakovlev, "Direction of arrival estimation using a root-music algorithm," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Citeseer, vol. 2, 2008, pp. 19–21.

[21] D. G. Manolakis, V. K. Ingle, S. M. Kogon, *et al.*, *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering, and Array Processing*. McGraw-Hill Boston, 2000.

[22] T. B. Lavate, V. Kokate, and A. Sapkal, "Performance analysis of music and esprit doa estimation algorithms for adaptive array smart antenna in mobile communication," in *2010 Second International Conference on Computer and Network Technology*, IEEE, pp. 308–311.

[23] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[24] Y. Tamai, S. Kagami, Y. Amemiya, Y. Sasaki, H. Mizoguchi, and T. Takano, "Circular microphone array for robot's audition," in *Sensors*, IEEE, 2004, pp. 565–570.

[25] E. Hulsebos, T. Schuurmans, D. de Vries, and R. Boone, "Circular microphone array for discrete multichannel audio recording.," in *Audio Engineering Society Convention 114*, Audio Engineering Society, 2003.

[26] S. Astapov, J. Berdnikova, and J.-S. Preden, "A two-stage approach to 2d doa estimation for a compact circular microphone array," in *2015 International Conference on Informatics, Electronics & Vision (ICIEV)*, IEEE, pp. 1–6.

[27] A. Karbasi and A. Sugiyama, "A new doa estimation method using a circular microphone array," in *2007 15th European Signal Processing Conference*, IEEE, pp. 778–782.

[28] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Source counting in real-time sound source localization using a circular microphone array," in *2012 IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 521–524.

[29] D. Pavlidi, M. Puigt, A. Griffin, and A. Mouchtaris, "Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2625–2628.

[30] A. Griffin, D. Pavlidi, M. Puigt, and A. Mouchtaris, "Real-time multiple speaker doa estimation in a circular microphone array based on matching pursuit," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, IEEE, pp. 2303–2307.

[31] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 135–143, 2004.

[32] P. K. T. Wu, N. Epain, and C. Jin, "A dereverberation algorithm for spherical microphone arrays using compressed sensing techniques," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4053–4056.

[33] Z. Li and R. Duraiswami, "A robust and self-reconfigurable design of spherical microphone array for multi-resolution beamforming," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 4, pp. iv–1137.

[34] B. N. Gover, J. G. Ryan, and M. R. Stinson, "Microphone array measurement system for analysis of directional and spatial variations of sound fields," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 1980–1991, 2002.

[35] C. T. Jin, N. Epain, and A. Parthy, "Design, optimization and evaluation of a dual-radius spherical microphone array," *IEEE/ACM transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 193–204, 2013.

[36] A. Nehorai and E. Paldi, "Acoustic vector-sensor array processing," *IEEE Transactions on Signal Processing*, vol. 42, no. 9, pp. 2481–2491, 1994.

[37] J. Cao, J. Liu, J. Wang, and X. Lai, "Acoustic vector sensor: Reviews and future perspectives," *IET Signal Processing*, vol. 11, no. 1, pp. 1–9, 2016.

[38] D. Levin, S. Gannot, and E. A. Habets, "Direction-of-arrival estimation using acoustic vector sensors in the presence of noise," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 105–108.

[39] D. Levin, E. A. Habets, and S. Gannot, "Maximum likelihood estimation of direction of arrival using an acoustic vector-sensor," *The Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. 1240–1248, 2012.

[40] J. Zhou and R. N. Miles, "Sensing fluctuating airflow with spider silk," *Proceedings of the National Academy of Sciences*, vol. 114, no. 46, pp. 12 120–12 125, 2017.

[41] ——, "Directional sound detection by sensing acoustic flow," *IEEE Sensors Letters*, vol. 2, no. 2, pp. 1–4, 2018.

[42] D.-I. H. Pessentheiner, "Differential microphone arrays," Master's thesis, 2013.

[43] H. Teutsch and G. W. Elko, "First-and second-order adaptive differential microphone arrays," in *International Workshop on Acoustic Echo and Noise Control*, Citeseer, vol. 1, 2001.

[44] P. S. Diniz, *Adaptive Filtering*. Springer, 1997.

[45] G. W. Elko and A.-T. N. Pong, "A simple adaptive first-order differential microphone," in *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Acoustics*, IEEE, pp. 169–172.

[46] J. Chen and J. Benesty, "A general approach to the design and implementation of linear differential microphone arrays," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, IEEE, pp. 1–7.

[47] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[48] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2, pp. 749–752.

[49] W. M. Hartmann, "How we localize sound," *Physics Today*, vol. 52, pp. 24–29, 1999.

[50] B. G. Ferguson, "A ground-based narrow-band passive acoustic technique for estimating the altitude and speed of a propeller-driven aircraft," *The Journal of the Acoustical Society of America*, vol. 92, no. 3, pp. 1403–1407, 1992.

[51] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, vol. 2, pp. 1228–1233.

[52] J. L. Desjardins, "The effects of hearing aid directional microphone and noise reduction processing on listening effort in older adults with hearing loss," *Journal of the American Academy of Audiology*, vol. 27, no. 1, pp. 29–41, 2016.

[53] D. R. Begault and L. J. Trejo, "3d sound for virtual reality and multimedia," Sep. 2000.

[54] B. C. Moore, *Hearing*. Academic Press, 1995.

[55] Y. Agiomyrgiannakis and Y. Stylianou, "Stochastic modeling and quantization of harmonic phases in speech using wrapped gaussian mixture models," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, pp. IV–1121.

[56] A. Brendel, C. Huang, and W. Kellermann, "Stft bin selection for localization algorithms based on the sparsity of speech signal spectra," *Ratio*, vol. 2, p. 6, 2018.

[57] A. Brendel, B. Laufer-Goldshtein, S. Gannot, R. Talmon, and W. Kellermann, "Localization of an unknown number of speakers in adverse acoustic conditions using reliability information and diarization," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7898–7902.

[58] Wikipedia contributors, *Directional statistics — Wikipedia, the free encyclopedia*, https://en.wikipedia.org/w/index.php?title=Directional_statistics&oldid=907978224, [Online; accessed 2-September-2019], 2019.

[59] K. V. Mardia and P. E. Jupp, *Directional statistics*. John Wiley & Sons, 2009, vol. 494.

[60] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[61] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *1979 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 208–211.

[62] S. 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.

[63] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[64] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Stereo source separation and source counting with map estimation with dirichlet prior considering spatial aliasing problem," in *International Conference on Independent Component Analysis and Signal Separation*, Springer, 2009, pp. 742–750.

[65] A. A. Nair, A. Reiter, C. Zheng, and S. Nayar, "Audiovisual zooming: What you see is what you hear," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1107–1118.