# SINGLE CHANNEL SIGNAL SEPARATION USING PSEUDO-STEREO MODEL AND TIME-FREQUENCY MASKING

Naruephorn Tengtrairat

BEng

MSc

**A thesis submitted to the university of Newcastle for the degree of**

**Doctor of Philosophy**

School of Electrical, Electronic and Computer Engineering

Faculty of Science, Agriculture and Engineering

August 2013

# ABSTRACT

In many practical applications, one sensor is only available to record a mixture of a number of signals. Single-channel blind signal separation (SCBSS) is the research topic that addresses the problem of recovering the original signals from the observed mixture without (or as little as possible) any prior knowledge of the signals. Given a single mixture, a new pseudo-stereo mixing model is developed. A “pseudo-stereo” mixture is formulated by weighting and time-shifting the original single-channel mixture. This creates an artificial resemblance of a stereo signal given by one location which results in the same time-delay but different attenuation of the source signals. The pseudo-stereo mixing model relaxes the underdetermined ill-conditions associated with monaural source separation and begets the advantage of the relationship of the signals between the readily observed mixture and the pseudo-stereo mixture. This research proposes three novel algorithms based on the pseudo-stereo mixing model and the binary time-frequency (TF) mask. Firstly, the proposed SCBSS algorithm estimates signals’ weighted coefficients from a ratio of the pseudo-stereo mixing model and then constructs a binary maximum likelihood TF masking for separating the observed mixture. Secondly, a mixture in noisy background environment is considered. Thus, a mixture enhancement algorithm has been developed and the proposed SCBSS algorithm is reformulated using an adaptive coefficients estimator. The adaptive coefficients estimator computes the signal characteristics for each time frame. This property is desirable for both speech and audio signals as they are aptly characterized as non-stationary AR processes. Finally, a multiple-time delay (MTD) pseudo-stereo mixture is developed. The MTD mixture enhances the flexibility as well as the separability over the originally proposed pseudo-stereo mixing model. The separation algorithm of the MTD mixture has also been derived. Additionally, comparison analysis between the MTD mixture and the pseudo-stereo mixture has also been identified. All algorithms have been demonstrated by synthesized and real-audio signals. The performance of source separation has been assessed by measuring the distortion between original source and the estimated one according to the signal-to-distortion (SDR) ratio. Results show that all proposed SCBSS algorithms yield a significantly better separation performance with an average SDR improvement that ranges from 2.4dB to 5dB per source and they are computationally faster over the benchmarked algorithms.

# ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude to my supervisors Dr. Wai Lok Woo and Professor Satnam Dlay for giving me the opportunity to pursue my Ph.D and the unconditional support during my research. I am very appreciated all their contribution of time, guidance, support, encouragement and patient to make my PhD a remarkable achievement. They have not only taught me to understand fundamental knowledge of signal separation and advance signal processing but the joy and passion on their research has also been an excellent role models for my research root.

I would also like to thank my thesis examination committee members for their time, constructive criticism, and feedback. This thesis is much the better because of them.

I gratefully acknowledge Payap University for being the source of funding that made my Ph.D. work possible.

My sincere thanks my research colleague Dr. Bin Gao, who guided me at the beginning of my research tenure. I would like to give the warmest thank to Phetcharat Parathai, who is not only my colleague but also my best friend, for all her support and encouragement. She has always been beside me during the happy and hard moments.

My heartfelt thanks to my family for their faith in me. They have provided me with love, generous care and support all my life. I have been truly blessed to have them as my parents.

Above all, I would like to pay my highest regards and genuine thanks to my Almighty God; Jesus, for granting me the pursuit of my Ph.D. With His mercy and unconditional love, He also grants me the wisdom to complete this journey.

# LIST OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF SYMBOLS

| | |
|---|---|
| $x(t)$ | mixture in time domain |
| $s(t)$ | original signal in time domain |
| $t$ | time index |
| $\hat{s}(t)$ | estimated signal in time domain |
| $w_i$ | mixture weights |
| $\mu_i$ | mean vector |
| $q_t$ | probabilities associated with the state of HMM |
| $a_{kl}$ | state transition probabilities |
| $S_l$ | current state of the HMM model |
| $S_k$ | previous state of the HMM model |
| $A$ | matrix of the state transition probability distribution |
| $B$ | matrix of the observation probability distribution |
| $v_k$ | observation symbols |
| $\alpha_t(l)$ | forward algorithm |
| $o_t$ | observed sequences |
| $\mathbb{R}^N$ | Real number |
| $D_J$ | maximum AR order |
| $e_j(t)$ | independent identically distributed (i.i.d.) random signal with zero mean and variance |
| $a_{s_j}(m;t)$ | $m^{th}$ order AR coefficient of the $j^{th}$ signal at time $t$ |
| $\gamma$ | weight parameter of the pseudo-stereo mixture |
| $\delta$ | time-delay of the pseudo-stereo mixture |
| $a_j(t;\delta,\gamma) \Leftrightarrow a_j(t)$ | mixing attenuation of the $j^{th}$ signal in time domain |
| $r_j(t;\delta,\gamma) \Leftrightarrow r_j(t)$ | residue of the $j^{th}$ signal in time domain |
| $F^W(\cdot)$ | Fourier transform |

| | |
|---|---|
| $W(\cdot)$ | window function |
| $S_j(\tau,\omega)$ | original signal in time-frequency domain |
| $f_{max}$ | maximum frequency present in the signals |
| $f_s$ | sampling frequency |
| $R_j(\tau,\omega)$ | residue of the $j^{th}$ signal in time-frequency domain |
| $C_j(\tau,\omega)$ | coefficient of the residue |
| $X(\tau,\omega)$ | mixture in time-frequency domain |
| $L_j(\tau,\omega)$ | Gaussian likelihood function |
| $C$ | normalizing constant |
| $S_j^{ML}(\tau,\omega)$ | maximum likelihood estimated signal in time-frequency domain |
| $\bar{a}_j(\tau,\omega)$ | mixing coefficient |
| $j$ | $j^{th}$ signal |
| $J(\tau,\omega)$ | cost function |
| $\kappa_j(\tau,\omega)$ | Maximum value of $C_j(\tau,\omega)$ |
| $\hat{\bar{a}}_j$ | Mixing coefficient estimator |
| $e_{interf}$ | Interference error from other signals |
| $e_{artif}$ | Artifact error |
| $B_l$ | Block $l$ |
| $\bar{a}_j^{(r)}(\tau,\omega)$ | Real part of the mixing coefficient of the $j^{th}$ signal |
| $\bar{a}_j^{(i)}(\tau,\omega)$ | Imaginary part of the mixing coefficient of the $j^{th}$ signal |
| $\hat{\bar{a}}_j^{(r)}$ | Real part of the mixing coefficient estimator |
| $\hat{\bar{a}}_j^{(i)}$ | Imaginary part of the mixing coefficient estimator |
| $\hat{a}_j$ | Mixing attenuation estimator |
| $\hat{C}_j$ | Residual coefficient estimator |
| $H_k(\tau,\omega)$ | Maximum likelihood cost function |
| $M_j(\tau,\omega)$ | Time-frequency masking |
| $\tilde{S}_j(\tau,\omega)$ | Estimated signal in time-frequency domain |

| | |
|---|---|
| $\hat{\hat{a}}_{jl}$ | Estimate of $\hat{\hat{a}}_j$ from the $l^{th}$ block |
| $L$ | Number of blocks |
| $N$ | Sample size of a signal |
| $N_s$ | Number of signals |
| $N_\emptyset$ | Number of frequency-shifts |
| $W_l$ | Length of the STFT window |
| $N_\emptyset$ | Number of frequency-shifts |
| $N_\tau$ | Time-shift |
| $K$ | Number of SCICA blocks |
| $I_{SNMF2D}$ | Number of iterations for SNMF2D |
| $I_{SCICA}$ | Number of iterations for SCICA |
| $\tilde{X}(\tau,\omega)$ | Signal-presence mixing model |
| $\hat{\tilde{X}}(\tau,\omega)$ | Noise-reduced mixture |
| $\tilde{\tilde{a}}_j(\tau)$ | Adaptive mixing attenuation estimator |
| $n_1(t)$ | Additive uncorrelated noise |
| $n_2(t;\delta,\gamma)$ | Noise by weighting and time-shifting of the additive noise |
| $H_0(\tau,\omega)$: | Signal absence hypothesis |
| $H_1(\tau,\omega)$ | Signal presence hypothesis |
| $S(\tau,\omega)$ | Sum of original signals |
| $q_\omega$ | Prior probabilities ratio of signal presence per signal absence hypotheses |
| $p(H_0)$ | Prior probabilities of the signal absence hypothesis |
| $p(H_1)$ | Prior probabilities of the signal presence hypothesis |
| $T_L$ | Local threshold |
| $T_G$ | Global threshold |
| $p_e$ | Probability of error |
| $X_{T_L}(\tau,\omega)$ | Threshold boundary between source absence and presence |
| $\widetilde{N}(\tau,\omega)$ | Residual noise |
| $A(\tau,\omega)$ | Spectral amplitude of signals |

| | |
|---|---|
| $\hat{A}(\tau,\omega)$ | Proposed improved mean square error short-time spectral amplitude (iMMSE-STSA) estimator |
| $\tilde{A}(\tau,\omega)$ | Conventional MMSE-STSA estimator |
| $I_0(\cdot)$ | Modified Bessel functions of zero$^{\text{th}}$ order |
| $I_1(\cdot)$ | Modified Bessel functions of firth order |
| $E[\cdot]$ | Expectation function |
| $\zeta_\xi$ | Weighing factor of *a priori* SNR estimator |
| $\hat{S}(\tau,\omega)$ | Estimated spectra of the mixture in time-frequency domain |
| $\hat{\hat{S}}_j(\tau,\omega)$ | Estimated signal from the noise-reduced mixture in time-frequency domain |
| $a_{ij}(\cdot)$ | $i^{th}$ mixing attenuation of the $j^{th}$ signal |

## Greek Symbols

| | |
|---|---|
| $\sigma_i^2$ | covariance matrix |
| $\pi_k$ | initial state distribution |
| $\gamma_t(k)$ | posterior probability of HMM components at time $t$ |
| $\beta_t(k)$ | backward algorithm |
| $\lambda$ | HMM model parameters |
| $\phi$ | Maximum time-delay (shift) associated with the Fourier transform with an appropriate window function |
| $\mathit{\Omega}_j(\tau,\omega)$ | Active area of the $j^{th}$ signal at $(\tau,\omega)$ unit |
| $\theta$ | Distinguishability function of the mixing attenuation |
| $\alpha_j$ | Symmetric mixing coefficient |
| $\mathit{\Lambda}$ | Set of the selected $\gamma$ and $\delta$ |
| $\psi$ | ARR threshold |
| $\Delta_{\alpha^{(r)}}$ | Maximum value of real part of symmetric mixing coefficient for histogram |
| $\Delta_{\alpha^{(i)}}$ | Maximum value of imaginary part of symmetric mixing coefficient for histogram |
| $\zeta^{(r)}$ | Number of bins of real part of symmetric mixing coefficient for histogram |

| | |
|---|---|
| $\zeta^{(i)}$ | Number of bins of imaginary part of symmetric mixing coefficient for histogram |
| $\Lambda(\tau,\omega)$ | Likelihood ratio of the signal presence and signal absence at $(\tau,\omega)$ units |
| $\zeta_N$ | Smoothing parameter of the noise power estimate |
| $\hat{\sigma}_N^2(\tau,\omega)$ | Noise power estimate |
| $\sigma_S^2(\tau,\omega)$ | Signal power spectral density |
| $\sigma_N^2(\tau,\omega)$ | Noise power |
| $\xi_f$ | Proposed fixed a priori SNR |
| $\theta\omega$ | Complex exponential of the noisy phase |
| $\alpha\omega$ | Complex exponential of the signal phase |
| $\Gamma(\cdot)$ | gamma function |
| $\gamma_{SNR}(\tau,\omega)$ | *a posteriori* SNR |
| $\xi(\tau,\omega)$ | *a priori* SNR |
| $\hat{\gamma}_{SNR}(\tau,\omega)$ | *a posteriori* SNR estimator |
| $\hat{\xi}(\tau,\omega)$ | *a priori* SNR estimator |
| $\hat{\xi}_f$ | Optimal $\xi_f$ |
| $\zeta_M$ | Smoothing parameter of the adaptive mixing attenuation estimator. |

## ABBREVIATIONS/ACRONYMS

| | |
|---|---|
| BSS: | Blind Source Separation |
| ICA | Independent Component Analysis |
| SCSS: | Single Channel Source Separation |
| DUET | Degenerate Estimation Technique |
| TF | Time Frequency |
| SCBSS | Single Channel Blind Source Separation |
| SNMF | Sparse Non-negative Matrix Factorization |
| ASR | Automatic speech Recognition |
| EMG | Electromagnetic |
| EEG | Electroencephalogram |
| BCI | Brain Computer Interfacing |
| SDR | Signal-to-Distortion Ratio |
| MTD | Multi-Time Delay |
| GMM | Gaussian Mixture Model |
| HMM | Hidden Markov Model |
| EM | Expectation-Maximization |
| MAP | Maximum A Posteriori |
| CASA | Computational auditory scene analysis |
| A/U/V | Accompaniment/ Unvoiced singing voice/ Voiced singing voice |

| | |
|---|---|
| STFT | Short Time Fourier Transform |
| SCICA | Single - Channel Independent Component Analysis |
| LS | Least Square |
| KL | Kullback-Leibler |
| AR | Autoregressive |
| WDO | Windowed - Disjoint Orthogonality |
| ML | Maximum Likelihood |
| MAD | Mixing Attenuation Distinguishability |
| ARR | Attenuation-to-Residue Ratio |
| SNR | Signal-to-Noise Ratio |
| SS | Speech and Speech |
| MM | Music and Music |
| SM | Speech and Music |
| 2D | 2-Dimensional |
| SOLO | Single Observation Likelihood estimatiOn |
| SIR | Signal-to-Interference Ratio |
| IBM | Ideal Binary Mask |
| AAD | Audio Activity Detection |
| PDF | Probability Density Function |
| LSAP | Local Source Absence Probability |
| GSAP | Global Source Absence Probability |
| VAD | Voice Activity Detection |
| iMMSE-STSA | Improved Mean Square Error Short-Time Spectral |

| | |
|---|---|
| | Amplitude |
| MMSE-STSA | Mean Square Error Short-Time Spectral Amplitude |
| SPP | Source Presence Probability |
| MSE | Mean - Square Error |
| PESQ | Perceptual Evaluation of Speech Quality |
| MTD | Multiple Times Delay |

# LIST OF PUBLICATIONS

- N. Tengtrairat, Bin Gao, W.L. Woo, S.S. Dlay, “Single Channel Blind Separation using Pseudo-Stereo Mixture and Complex 2D Histogram”, *IEEE Transactions on Neural Networks and Learning System,* vol.24, no.11, pp. 1722 -1735, Nov. 2013.

- N. Tengtrairat and W.L. Woo, “Extension of DUET to Single-Channel Mixing Model and Separability Analysis,” *Signal Processing*, vol.96, pp.261 – 265, Mar. 2014.

- N. Tengtrairat, Bin Gao, W.L. Woo, S.S. Dlay, “Online Noisy Single-Channel Blind Separation by Spectrum Amplitude Estimator and Masking”, *submitted to IEEE Transactions on Signal Processing (as under review).*

# CHAPTER 1

# INTRODUCTION TO THE THESIS

## 1.1 Background of Blind Signal Separation

In natural auditory environments, a number of people are talking simultaneously in a scene and a listerner is trying to follow one of the conversations by separating the mixed speech into individual speech signal corresponding to each speaker [1]. This classical example is known as the “cocktail party” problem. The sounds in an auditory scene all sum together through the recording sensors which can be expressed mathematically as:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_M(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MN} \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_N(t) \end{bmatrix} \tag{1.1}$$

where $[x_1(t)x_2(t) \cdots x_M(t)]^T$ denotes a set of the recording sensors which are random processe as a mixture of underlying source signals $[s_1(t)s_2(t) \cdots s_N(t)]^T$, and $\{a_{mn}\}$ $\Leftrightarrow A, \forall m \in M$and $\forall n \in N$ denotes the unknown mixing matrix of dimension $M \times N$ and $t = 1, 2, \ldots, T$ is the time index. Eq.(1.1) introduces the “cocktail party problem” [2, 3] by means of statistical methods under the name of blind signal sepation (BSS). A typical BSS process is illustrated in Fig.1.1.

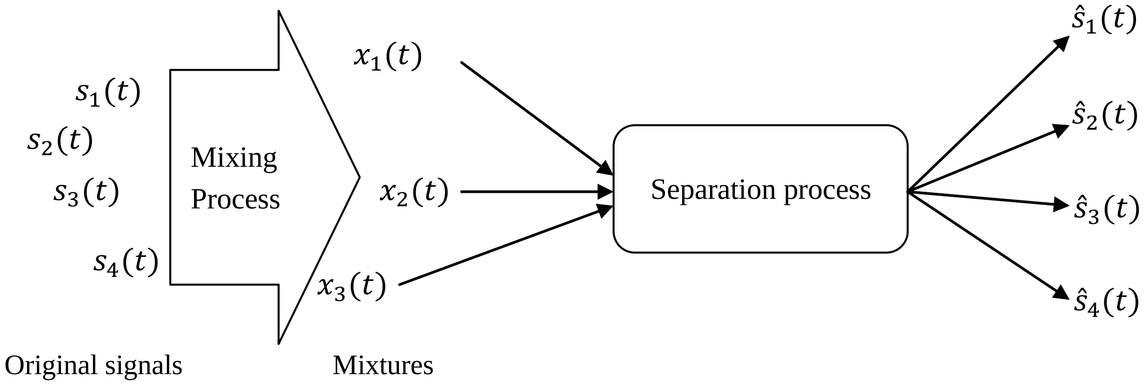


Figure 1.1: Typical blind source separation process.

The technique of BSS aims to estimate both the original signals $[s_1(t)s_2(t)\cdots s_N(t)]^T$ and the mixing matrix $A$ using only the observerd mixture $[x_1(t)x_2(t)\cdots x_M(t)]^T$. The BSS problem reveals two major challenging tasks in recovering the original signals: first, identifying components of the mixture that belong to each original signal. Second, partition the mixture that corresponds to the same component. In a realistic scenario, the cocktail party takes place with background noises which interfere with the source signals; especially in places where the source signals has lower energy than the noise. The signals are thus physically marked by noise. This problem results in increased difficulty to distinguish the original signals from the noisy mixture in a BSS process.

Blind signal separation is the process of recovering underlying source signals from an unknown mixing given only the sensor signals [4-6]. BSS has interested many researchers during the last decade because of its potential to solve problems in a ubiquitous range of disciplines. In the last decade, promising results have been obtained in the solutions of BSS. The solutions of the BSS problem depend on several factors as follows [7]: linearity of a mixture, time characteristic of mixing process, mixing operation, sensors’ quality, and relation between number of signals and number of measurements.

In the early BSS era, independent component analysis (ICA) was first proposed as a solution [8]. The ICA approach aims to recover the unknown mixing matrices from a number of observed mixtures for extracting a number of signals. The ICA method is based on the critical assumption that the original signals are non-Gaussian [9] and mutually independent. BSS using ICA approaches is straightforward and has been used in many applications with great success [10, 11]. Existing approaches have been successful in different conditions of the BSS problem. However, in the case of a single channel sensor, none of them are yet satisfactory for an application.

Typically, biological auditory system can efficiently solve the BSS problem which is known as a ‘binaural BSS problem’. The binaural approach required two microphones for recording signals from scenes with more than two signals. In a binaural BSS method, the Degenerate Unmixing Estimation Technique (DUET) [12] and its variants [13, 14] have been proposed as a separating method using binary time-frequency (TF) masks. A major advantage of DUET is that the estimates from two channels are combined inherently as a part of the clustering process. DUET algorithm has been demonstrated to recover the underlying sparse signals given two anechoic mixtures in the TF domain. However, the DUET algorithm has been practically handicapped to separate signals when only one recording channel is available. Additionally, determining the masks blindly from only one mixture is still an open problem. In practical applications, this crux problem has not yet developed enough to make its way out of laboratories.

### *1.1.1 Single Channel Blind Signal Separation (SCBSS) Problem*

In practice, it may not be able to provide a sensor for individual signal because of limited spaces, high cost of sensors, violation of assumptions, and so forth [15]. For these reasons, the number of sensors is mostly less than the number of source signals. Furthermore, there is a case where only one sensor is available which corresponds to the extreme case of the underdetermined BSS problem. Under this circumstance, most conventional BSS methods fail to recover the original signal from the single channel observation. This leads to a research avenue of single channel blind signal separation (SCBSS) problem. SCBSS represents the separation of mixed signal from a single sensor. Mathematically, it can be treated as one mixture of $N$ unknown original signals:

$$x(t) = s_1(t) + s_2(t) + \cdots + s_N(t) \tag{1.2}$$

where $t = 1, 2, \dots, T$ denotes time index and the goal is to estimate the signals $s_n(t)$, $\forall n \in N$ of length $T$ when only the observation signal $x(t)$ is available. In (1.2), the number of source signals $\{s_n(t)\}_{n=1,2,\dots,N}$ is more than the number of the observed mixture $x(t)$, this becomes the underdetermined SCBSS problem. Recently, new SCBSS approaches have been proposed to solve the problem. In general, they can be categorized into two groups i.e. model-based and data-driven methodologies. A “model-based” separation approach requires prior knowledge from the training datasets to estimate the unknown signals. Model-based SCBSS methods have been dominantly illuminated by computational auditory scene analysis, and hidden Markov models methods. The data-driven SCBSS methods perform signal separation without any recourse to the training information. The popular method in this category is the sparse non-negative matrix factorization (SNMF). More details of the above methods will be reviewed in Chapter 2.

### *1.1.2 Applications of SCBSS*

Single-channel signal separation has been an exciting approach of engineering research in the last two decades because its derivative techniques have played a prominent role in both academic and industry areas. In the case of a sole recording sensor, its practical applications are listed below:

- Automatic Speech recognition (ASR) is for command and control applications with a single microphone. The performance of ASR systems relies on quality and volume of the target subject. In the presence of acoustic interferences with background noise, the ASR performance dramatically sinks. This ASR problem can be alleviated by using the SCBSS technique, if the target signal can be segratated from the noisy mixture to provide the ASR system with a clean target signal.
- Automatic music transcription of polyphonic music is one of the challenging problems to separate individual instrument from the musical mixture. Musical instruments have a wide range of sound production mechanisms, and the observed mixture have thus a wide range of spectral and temporal characteristics. Extracting information of each signal, for example: a signal from guitar and piano will be useful to indicate of the key of a song. Subsequently, the musical signals can be transcribed individually.

- In the analysis of electromagnetic (EMG) brain signals, there are instances where only one sensor is available. Neurophysiologically information is required to be segregated from the observerd mixture for example the analysis of the epileptic electroencephalogram (EEG) or the interpretation of brain computer interfacing (BCI). The SCBSS solution will be useful to distinguish, reveal and track neurophysiologically signals underlying the single EEG or BCI mixture. [16]

## 1.2 Objectives of Thesis

The aim of this thesis is to solve the SCBSS problem without resorting to the training information of the original sources. To pursue this goal, existing SCBSS approaches based on a single mixture have been reviewed and investigated. In particular, the thesis work will develop new framework to study and tackle the SCBSS problem efficiently. The objectives of the thesis are listed as follows:

i). To present a unified perspective of the widely used the state-of-the-art separation approaches when only one channel is available. The theoretical aspects of SCBSS are presented to provide sufficient background knowledge relevant to the thesis.

ii). To develop new algorithms that simulate the human auditory sensory which creates an artificial stereo mixtures from a sole observed mixture.

iii). To develop new algorithms based on the artificial stereo mixtures for unveiling the original time-varing signals.

iv). To carry out rigorous mathematical derivations and analysis, and compare the separation performance of the proposed algorithm with the existing state-of-the-art

SCBSS methods using objective as well as perceptual evaluation of audio quality such as Signal-to-Distortion ratio (SDR).

## 1.3 Thesis Outline

This thesis focuses principally on unsupervised separation of single channel mixtures. The thesis comprises an introductory chapter, the main contents, and concludsion. Three novel methods for SCBSS constitute the main contribution of the thesis. The thesis outline is as follows:

In Chapter 2, an overview of single-channel signal separation is introduced. A comprehensive review of recent SCBSS approaches is by classifying into two separation themes i.e. model-based and data-driven approaches. Model-based and data-driven separation approaches are sequentially presented and analysised in this chapter.

In Chapter 3, a novel ‘pseudo-stereo’ mixture is proposed to model of an artificial stereo model. The impetus behind this is that the parameter estimation of the signal from two mixtures is combined inherently as part of the clustering process. The pseudo-stereo mixture is formulated by weighting and time-shifting the original single-channel mixture. Separability analysis of the pseudo-stereo model has also been derived to verify that the pseudo-stereo model is separable.

In Chapter 4, a novel method in a single audio recording for blind signal separation is developed by incoorperating the proposed pseudo-stereo mixture (as detailed in Chapter 3). The proposed method is based on the estimation of mixing coefficients of the signal.

The original signals are assumpted that can be modelled by the autoregressive process. Thus, the coefficient domain is introduced by taking the advantage on the difference of AR coefficients between the two mixtures. Additionally, a binary time-frequency mask was built by evaluating a proposed cost function. Experimental testing of the proposed method yields superior performance and is computationally very fast compared with existing methods

In Chapter 5, a novel method to solving SCBSS in noisy environment is proposed. The new method was developed by reformulating the pseudo-stereo model and the speech enhancement problem into a joint SCBSS problem. The mixture enhancement is introduced to degrade noise and extract signal information from the noisy mixture. Henceforth, a separation process decomposed the original signals by multiplying a mask on the noise-reduced mixture. Experimental results showed that the proposed method yielded a superior separation performance especially in low input SNR as compared with existing SCBSS methods.

In Chapter 6, an extention of the pseudo-stereo mixture is developed. The new pseudo-stereo mixture introduced multi-time delay (MTD) the single audio recording. Separability of the MTD mixture was analysed and shown that the MTD mixture is separable. Hence, the MTD mixture can be separated to unveil the original sources. Furthermore, comparison analysis between the pseudo-stere mixture and the MTD mixture is presented. By comparing with the pseudo-stereo mixture where employing the same separation method, experimental results illustrates that the MTD mixture leads to a significantly improvement of separation performance.

This thesis is concluded with Chapter 7. This chapter exhibits the closing remarks as well as future avenues for research.

## 1.4 Contribution

This thesis contributes three novel solutions for the SCBSS problem. The proposed methods deals with the constraints related the recent SCBSS approaches. The contributions in this thesis are summarised the following:

i). A unified approach was developed for the existing SCBSS methods based on the linear instantaneous mixing model.

ii). A novel artificial mixture was developed that relaxes the under-determined ill-conditions associated with monaural signal separation and path the way for binaural signal separation approaches to solve monaural mixture.

iii). A novel algorithm was developed based the artificial mixture (the pseudo-stereo mixture and the MTD mixture) is proposed:

- It is executed in “one-go” without the need of iterative optimization. Hence, the method works very fast and does not require any parameter tuning. This should be contrasted with other SCBSS methods such as SNMF and underdetermined-ICA SCBSS which require many iterative optimization of the solution.
- It is independent of initialization condition, i.e. it has no need for random initial inputs or any predetermined structure on the sensors. This renders robustness to

the proposed method.

- It has low computational complexity and does not exploit high-order statistic.

iv). A novel framework to solve SCBSS in noisy background environment is proposed.

- It is online adaptive separation, where the observed mixture is segmented into small frames. The separation process is then executed adaptively frame-by-frame. This online separation reduces the computational complexity of the whole observed mixture. Thus, it yields a low computational cost. Batch methods usually suffer from a large storage requirement and a high computational complexity when the observed mixture is large scale. Hence, the robustness of the proposed algorithm can benefit for real-time signal processing applications.
- It is an adaptive parameters estimation method. The parameters are adaptively estimated from two consecutive frames. The self-adaptive property is preferred for time-varying signals especially speech and highly nonstationary noise.
- It is independent of parameters initialization, i.e. no need for random initial inputs or any predetermined structure on the sensors. This renders robustness to the proposed method.
- It has computational simplicity and does not exploit high-order statistic. Hence this yields the benefit of ease of implementation.

v). A novel MTD mixture is proposed:

- It enhances the accuracy of the signal-signature estimator by increasing the distinguishability of the mixing attenuation between signals and reducing AR coefficients residues. Thus, the coefficient domain distinctively reveals the coefficients of the signals. This significantly advances the separation performance over the pseudo-stereo mixture.

# CHAPTER 2

# OVERVIEW OF SINGLE CHANNEL BLIND SIGNAL SEPARATION

This chapter gives an overview of the existing methods for SCBSS which have proven to produce separable results in the case of audio signals. In general, the SCBSS approaches have been classified into two main categories i.e. the model-based approach and the data-driven approach which are illustrated in Fig.2.1.

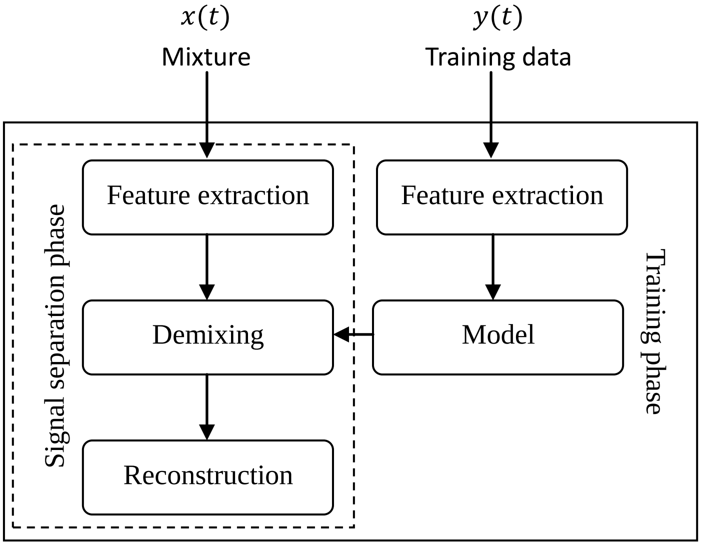


Figure 2.1: Overview of SCBSS approach.

A general framework of SCBSS consists of two main phases as shown in Fig. 2.1. The mixture is fed into the signal separation phase without any training data. This solution is regarded as the data-driven approach. Otherwise, a solution which contains the training phase is considered as the model-based approach. The details of the both approaches are discussed in Sections 2.1 and 2.2.

## 2.1 *Model-Based Approach*

The term “model-based” separation approach requires prior knowledge from the training datasets to estimate the unknown signals. This method is supervised by the training data from some or all of the original signals. The signal separation phase directly performs based on *a priori* knowledge of the signals from the training phase. For the training phase, modeling methods have been dominantly illuminated by Gaussian mixture model (GMM) and hidden Markov models (HMMs). These modeling methods are introduced in Section 2.1.1. Next, examples of the model-based SCBSS algorithms are presented in Section 2.1.2.

### *2.1.1 Modeling Methods*

#### 2.1.1.1 Gaussian Mixture models

A Gaussian Mixture Model is a parametric probability density function represented as a weighted sum of Gaussian component densities [17]. Technically, a Gaussian mixture model can be expressed as

$$p(s|w_i,\mu_i,\sigma_i^2) = \sum_{i=1}^{M} w_i\, g(s|\mu_i,\sigma_i^2) \qquad i = 1,\dots,M \tag{2.1}$$

$$g(s|\mu_i,\Sigma_i) = \frac{1}{(2\pi)^{D/2}|\sigma_i^2|^{1/2}} exp\{-\frac{1}{2\sigma_i^2}(s-\mu_i\}^T(s-\mu_i)\} \tag{2.2}$$

where $s$ is a data vector, $w_i$, $\mu_i$, and $\sigma_i^2$ are the mixture weights, a mean of the vector, and its covariance matrix, respectively. The term $g(s|\mu_i,\sigma_i^2)$ is the Gaussian density function, $D$ denotes a $D$ -dimensional continuous data vector, and $T$ is a transpose operation. The mixture weight satisfy the constraint that $\sum_{i=1}^{M} w_i = 1$. GMM generally updates its parameters $w_i$, $\mu_i$, ,$\sigma_i^2$ via the iterative expectation-maximization (EM) algorithm. To begin with an initial model $\theta = \{w_i, \mu_i, \sigma_i^2\}$, a new model $\bar{\theta}$ is estimated to satisfy the condition $p(s|\bar{\theta}) \geq p(s|\theta)$. For the next iteration, the new model then becomes the initial model. The process is repeated until a convergence threshold is achieved. The re-estimation based on EM algorithm can be expressed as the following [17]:

$$\overline{w}_i = \frac{1}{T}\sum_{t=1}^{T} P(i|s(t), \theta) \qquad (2.3)$$

$$\bar{\mu}_i = \frac{\sum_{t-1}^{T} P(i|s(t),\theta)\ s(t)}{\sum_{t=1}^{T} P(i|s(t),\theta)} \qquad (2.4)$$

$$\bar{\sigma}_i^2 = \frac{\sum_{t-1}^{T} P(i|s(t),\theta)\ s^2(t)}{\sum_{t=1}^{T} P(i|s(t),\theta)} - \bar{\mu}_i^2 \qquad (2.5)$$

The *a posteriori* probability for the component $i$ is given by

$$P(i|s(t), \theta) = \frac{w_i\ g(s(t)|\mu_i,\sigma_i^2)}{\sum_{k=1}^{M} w_k\ g(s(t)|\mu_k,\sigma_k^2)} \qquad (2.6)$$

### 2.1.1.2 Hidden Markov Models

Hidden Markov parameters [18] are initialized as $\lambda = (A, B, \pi)$ where: $\lambda$ represents the complete set of the HMM parameters for convenience. The term $A$ is the matrix of the state transition probability distribution. HMMs is based on a change of state according to a set of probabilities associated with the state $q_t$ at time $t = 1, 2, \ldots, T$. Thus the state transition probabilities $a_{kl}$ is truncated to just the current state $S_l$ and the previous state $S_k$ state. The state transition probability distribution can be expressed as:

$$A = a_{kl} \triangleq (q_t = S_l | q_{t-1} = S_k) \qquad 1 \leq k, l \leq N \qquad (2.7)$$

where $N$ is the number of states in the model as a result of the squared matrix $A$, with the state transition coefficients having the properties: $a_{kl} \geq 0$, and $\sum_{l=1}^{N} a_{kl} = 1$. Secondly, the term $B$ is the matrix of the observation probability distribution $P(O|\lambda)$ in state $l$, $B = \{b_l(o_t)\}, 1 \leq l \leq N$ where $o_t \Leftrightarrow O$ denotes the observed sequences.

$$b_l(o_t) = P(v_k\ at\ t | q_t = S_j) \qquad 1 \leq j \leq N, 1 \leq k \leq M \qquad (2.8)$$

where $M$ denotes the number of distinct observation symbols per state for example for the coin toss: the observation symbols are heads or tails $V = \{v_1, v_2\}$ and for a mixture of three signals, the observation symbols are the original signals $V = \{v_1, v_2, v_3\}$. The probability of the observation sequence given the model $P(O|\lambda)$ can be express as:

$$\begin{aligned} P(O|\lambda) &= \sum_{all\ Q} P(O|Q, \lambda) P(Q|\lambda) \\ &= \sum_{q_1, q_2, \ldots, q_T} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \ldots\ a_{q_{T-1} q_T} b_{q_T}(o_T) \qquad (2.9) \end{aligned}$$

The probability of the observation sequence $P(O|\lambda)$ can be solved by computing the forward part of the forward-backward algorithm. The forward algorithm can be expressed as:

$$\begin{aligned} \alpha_t(l) &= P(o_1\, o_2\, \cdots\, o_t, q_t = S_l | \lambda) \\ &= [\sum_{k=1}^{N} \alpha_{t-1}(k)\, a_{kl}] b_l(o_t), \qquad (2.10) \end{aligned}$$

where $1 \leq t \leq T-1$ and $1 \leq l \leq N$

The last parameter $\pi = \{\pi_k\}$ denotes the initial state distribution at time $t = 1$ and is defined as

$$\pi_k \triangleq P(q_1 = S_k) \qquad\qquad 1 \leq k \leq N \qquad\qquad (2.11)$$

Additionally, the posterior probability of HMM components $\gamma_{k_i}(t)$ can be expressed as

$$\begin{aligned}\gamma_t(k) &\triangleq P(q_t = S_k | O, \lambda) \\ &= \frac{\alpha_t(k)\beta_t(k)}{\sum_{k=1}^{N} \alpha_t(k)\beta_t(k)}\end{aligned} \qquad (2.12)$$

where $\gamma_t(k)$ is the probability of being in state $S_i$ at time $t$, given by O and $\lambda$. The model parameters $\lambda = (A, B, \pi)$ is reestimated by maximizing the observed probability $P(O|\lambda)$ by the iterative Baum-Welch algorithm which is known as the forward-backward algorithm. The backward algorithm $\beta_t(k)$ can be express as:

$$\begin{aligned}\beta_t(k) &\triangleq P(o_{t+1}\, o_{t+2} \cdots o_T, q_t = S_k | \lambda) \\ &= \sum_{l=1}^{N} a_{kl} b_l(o_{t+1})\, \beta_{t+1}(l),\end{aligned} \qquad (2.13)$$

where $t = T - 1, T - 2, \cdots, 1$ and $1 \leq k \leq N$.

### *2.1.2 Model-based SCBSS*

#### 2.1.2.1 CASA Model-based SCBSS

The human auditory system has an impressive capability to distinguish sounds from different signals without much difficulty, even monaurally. Computational auditory scene analysis (CASA) replicates the process of human auditory system by using signal processing approaches and grouping them into auditory streams using psycho-acoustical cues. The CASA approach aims to form the separation systems given by one or two recordings. This approach has attracted the interest of researcher in broadly disciplinary i.e. machine learning, signal processing and computational models. Many signal separation based on CASA methods have been proposed in the last few years. The overview of CASA framework is presented in Fig.2.2.

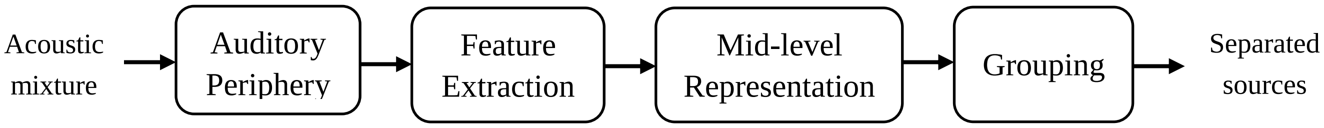


Figure 2.2: Overview of the CASA framework

A mixture is firstly transformed from time representation into time-frequency (TF) representation of auditory activity i.e. cochleagram. A gammatone filterbank derived from psychophysical observations of the auditory periphery and is a typical model of cochlear filtering. The gammatone filter is an approximation to the physiologically-recorded impulse responses of auditory nerve fibres and this filterbank is a standard model of cochlear filtering. The parameters of the gammatone filterbank (the filter order, bandwidth and frequency spacing) are usually chosen to provide a match to neuromechanical transduction in the cochlea [19]. Secondly, implicit signal features are then extracted for examples: pitch (which is quantified a frequency), periodicity i.e. a fundamental frequency, onsets, offsets, amplitude modulation, and frequency modulation (harmonic). Note that sounds with definite pitch have harmonic frequency spectra [20]. Next, mid-level representation, i.e. segmentation and pitch tracking, is formed unvoiced and voiced representation based on the extracted features. Unvoiced representation can be obtained from onset/offset analysis. Voiced representation mainly depends on the pitch estimation and pitch contours. In Grouping, the TF units are labeled into groups corresponding to the extracted feature. A number of groups correspond to the number of signals. Finally, a mask is then constructed by the labeled TF units. The mask can be binary or real-valued [21]. Many CASA algorithms employ the time-frequency (TF) mask for their separation process [22]. The original signals are then estimated by masking the TF plane of the mixture.

Recently, CASA based training methods have been introduced as in [23-25]. The work in [23] was proposed to separate the unvoiced signing voice from the song mixture. The input mixture is transformed into 128 channels using the gammatone filterbank where center frequencies are quasi-logarithmically spaced from 80Hz to 5kHz. The impulse response of gammatone filter and its frequency response are given by [25]. A HMM is trained to decode the mixture into Accompaniment/ Unvoiced singing voice/ Voiced singing voice (A/U/V). The HMM of A/U/V detection is illustrated in Fig. 2.3

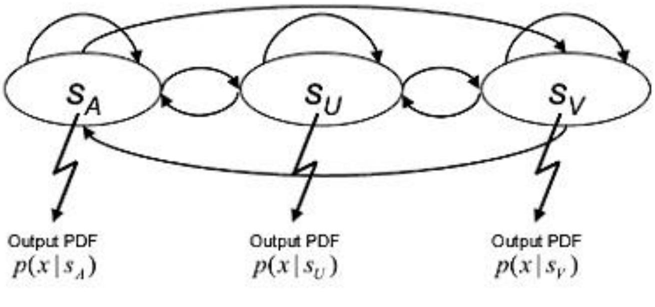


Figure 2.3: Hidden Markov model for detecting A/U/V [23].

where $x$ denotes the observed mixture, and $S_A$ , $S_B$ , and $S_C$ denote states for accompaniment, voiced and unvoiced frames, respectively. The most likely a sequence of A/U/V states given by $X = \{x_0, \ldots, x_t, \ldots\}$ was defined as $\hat{S} = \{s_0, \ldots, s_t, \ldots\}$

$$\hat{S} = \text{argmax}_S\{p(x_0|s_0)\prod_t[p(x_t|s_t)p(s_t\ s_{t-1})]\} \tag{2.14}$$

where $p(x\ s)$ is the output probability density function (pdf) of a state $s$, and $p(s_t\ s_{t-1})$ denotes the state transition probability from stage $s_{t-1}$ to $s_t$. The pitch contours is estimated to identify the voice units of the TF plane where its local periodicity matches the estimated pitch of the frame. While the unvoiced components are determined by using GMMs. GMMs are used to build a binary mask by comparing the energy of voice and music accompaniment. If the voice is larger, the TF unit is labeled as unvoiced-dominant. This method can separate the unvoiced singing voice satisfactorily. However, this methods requires trained models of voice for HMM and GMMs. With the model-based method, this causes the expense of high computational complexity.

To sum up, innovations in CASA methods emphasizes the signal representation such as the cochleagram for presenting the mixture in the well-defined TF domain. The trend of CASA methods have focused on multi-pitch tracking, feature-based processing, signal grouping, and model-based separation. These techniques are key issues for further research. The CASA methods can be used for applications for example to align the lyrics with singing voice on the lyric alignment system and to automate the melody transcription on karaoke application. The performance of the CASA method will be best when the interferer is tonal or locally narrowband. On the other hand, the CASA performed poorly in conditions where there is substantial spectral overlap between the speech and interferer. The drawbacks of CASA are summarized as follows: CASA methods cannot replicate the entire process performed in the auditory system since the process beyond the auditory nerve is not well studied. In addition, it is difficult to separate of the unvoiced speech from the background interference by using pitch tracking.

### 2.1.2.2 Soft Masking Model-based SCBSS

In the soft masking model-based SCBSS method [26], the algorithm assumes that both signals have Gaussian centered priors, with diagonal covariance matrices. HMMs are taken into account to pattern the structure of signals through the covariance matrices in the TF domain via short time Fourier transforms (STFT) for training phase. In separation, HMMs provide the posterior probability of the observation mixture sequences, given the model. Then Wiener filters are established from the covariance matrices and the posterior probability to estimate the original signals. This method can be categorized into 2 main phases which are training sources, signal separation, sequentially as illustrated in Fig 2.4.

This method proposed the SCBSS solution by assuming that both signals have Gaussian centered priors $S_s(t,f) \sim \mathcal{N}(0, diag\left(\sigma_{k_s}^2(f)\right))$, with diagonal covariance matrices $\Sigma_s = diag(\sigma_s^2(f))$. Where $\sigma_s^2(f)$ denotes the covariance matrices. In the traning phase, HMMs have been used to model the structure of sources through the covariance matrices. Additional, GMM is used to estimate the covariance matrices which requires for computing the intitial observation probability of HMM. HMM parameters were initialized as $\lambda = (A, B, \pi)$.

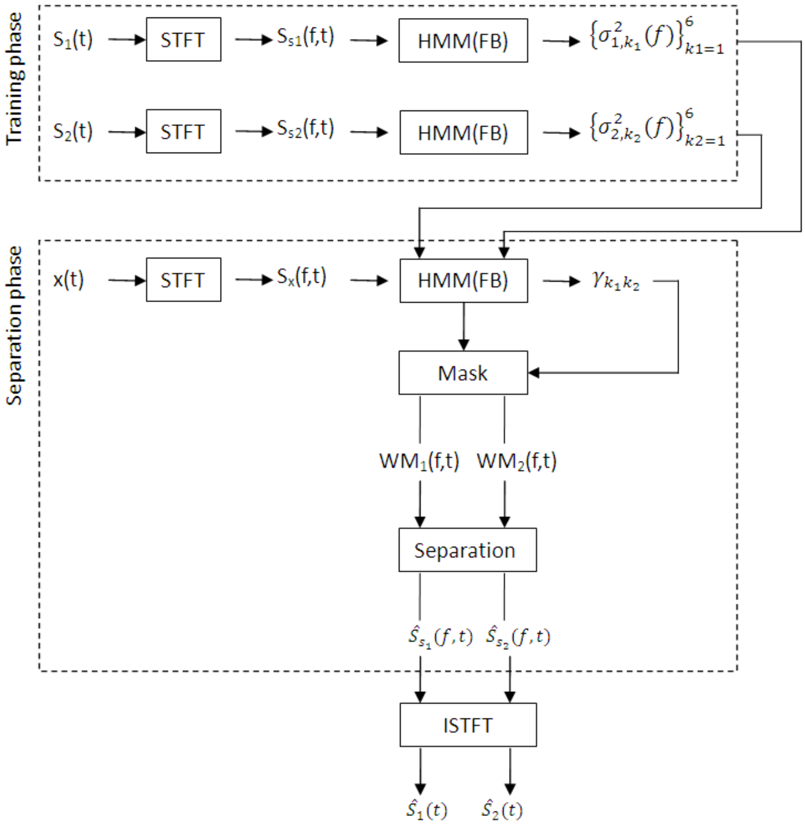


Figure 2.4: Overview of Soft masking based on GMM and HMM training.

The Gaussian probability density $p_G\left(S_i(t,f)\middle|\{\sigma_{k_i}^2(f)\}\right)$function of $S_i(t,f)$ given by $\sigma_s^2(f)$ can be expressed as below:

$$p_G\left(S_{s_i}(t,f)\middle|\{\sigma_{k_i}^2(f)\}\right) = \frac{1}{(2\pi)^{d/2}\prod_f \sigma_{k_i}(f)}\, exp\left[-\frac{1}{2}\sum_f \frac{\left|S_{s_i}(t,f)\right|^2}{\sigma_{k_i}^2(f)}\right] \qquad (2.15)$$

$$\sigma_{k_i}^2(f) = \frac{\sum_{t=1}^{T}\gamma_{k_i}(t)|S_s(t,f)|^2}{\sum_{t=1}^{T}\gamma_{k_i}(t)} \qquad (2.16)$$

where $d$ denotes the dimensional and equals to the number of frequency components, $\gamma_{k_i}(t)$ denotes Gaussian mixture components. The posteriori probability of $\sigma_{k_i}^2(f)$ is estimated by maximizing (2.9) based on the EM algorithm.The term $\gamma_{k_i}(t)$ estimated from HMM that reckons as the probability of being in state $S_i$ at time $t$, given the observation sequence $(O)$ and the model $(\lambda)$ i.e. $\gamma_t(k) = \frac{\alpha_t(k)\beta_t(k)}{\sum_{k=1}^{N}\alpha_t(k)\beta_t(k)}$ with the constraint that $\sum_{k=1}^{N}\gamma_t(k) = 1$. Thus the posterior probability to maximize $\sigma_{ki}^2$ of each speech source signal can be expressed as:

$$\sigma_{k_i}^2(f) = \frac{\sum_{t=1}^{T} \gamma_{k_i}(t) \left| S_{s_i}(t,f) \right|^2}{\sum_{t=1}^{T} \gamma_{k_i}(t)} \tag{2.17}$$

In separation, HMMs provide the posterior probability of the observation mixture sequences, given the model. The observed process $S_x(t,f) = S_{s_1}(t,f) + S_{s_2}(t,f)$ is centered Gaussian distributed $S_x(t,f) \sim \mathcal{N}(0, diag\left(\sigma_{k_1}^2(f) + \sigma_{k_2}^2(f)\right)$ with covariance matrix diag $(\sigma_{k_1}^2(f) + \sigma_{k_2}^2(f))$. The posterior probability of the observation mixture sequence $\gamma_{k_1,k_2}(t) \triangleq P_t(q_1 = S_k, q_2 = S_k | S_x(t_1, f), \ldots, S_x(t_N, f))$ is calculated by forward and backward algorithm form HMM. Finally, Wiener filters are established from the covariance matrices and the posterior probability to estimate the source signals. The summarized formulas of the estimated signals by Wiener filters are shown in equations below:

$$\hat{S}_{s_1}(f,t) = \left[ \sum_{k_1=1}^{N} \sum_{k_2=1}^{N} \frac{\sigma_{k_1}^2(f)}{\sigma_{k_1}^2(f) + \sigma_{k_2}^2(f)} \gamma_{k_1,k_2}(t) \right] S_x(f,t) \tag{2.18}$$

$$\hat{S}_{s_2}(f,t) = \left[ \sum_{k_1=1}^{N} \sum_{k_2=1}^{N} \frac{\sigma_{k_2}^2(f)}{\sigma_{k_1}^2(f) + \sigma_{k_2}^2(f)} \gamma_{k_1,k_2}(t) \right] S_x(f,t) \tag{2.19}$$

This Wiener filter based GMM and HMM showed good results when a priori information of the signals was sufficiently provided for the training phase. Conversely, HMMs obstacle to recover original signals when given by *a* low *priori* information of the signals and substantial overlap between the signals. However, the drawbacks of the system are the computational time consumption not only for the training but also the separation process.

### *2.2 Data-Driven Approach*

For data-driven approach, these methods perform signal separation without any recourse to the training information. The signal characteristics are not provided *a priori* knowledge for this approach. Thus, the data-driven method directly computes the parameters of the signals from the observed mixture which is known as a single-channel blind signal separation (SCBSS) algorithm. Blind means the separation process of a mixture, without any information of the original signals or the mixing process.

#### *2.2.1 Underdetermined-ICA time SCBSS*

Single-channel independent component analysis (SCICA) is an adaptation of the ICA algorithm to one observed sensor [27]. The SCICA approach in [28] applies the standard ICA to separate the independent signals from a single mixture. The special structure induced by mapping the observed mixture into a multi-channel model. The signals can then be separated by only employing the standard ICA. SCICA can be expressed in vector-matrix form as

$$\boldsymbol{x} = \boldsymbol{AS} \tag{2.20}$$

where $\boldsymbol{x}$ denotes a sequence of vectors $x(t) \in \mathbb{R}^N$, $\boldsymbol{A} = [a_1, \dots, a_N]$ is a mixing matrix, without prior knowledge of signals, which assumed that is invertible. The term $\boldsymbol{S}$ is the independent signals. Generally, the original signals can be distinguished from $\boldsymbol{x}$ by $\boldsymbol{S} = \boldsymbol{Wx}$ where $\boldsymbol{W} = \boldsymbol{A}^{-\boldsymbol{1}}$. For SCICA, the observed mixture $\boldsymbol{x}$ is broken up into a sequence of contiguous blocks $k$ with length $N$. These are treated as a sequence of vector mixtures:

$$\boldsymbol{x}(k) = [x(k\tau), \dots, x(k\tau + N - 1)]^T \tag{2.21}$$

where $k, k = 1,2, \dots, K$ is the block index. The matrix $\boldsymbol{X}$ is then formed as a set of mixtures $\boldsymbol{x}(k)$ as the following:

$$\boldsymbol{X} = [x(1), \dots, x(K)]^T \tag{2.22}$$

The FastICA algorithm can then be applied to $\boldsymbol{X}$ to compute the mixing and unmixing matrix $\boldsymbol{A}$ and $\boldsymbol{W}$. For a perfect reconstruction decompoisition, the separation process performs in the mixture domain where each signal is discovered via $\boldsymbol{A}$ and $\boldsymbol{W}$ as:

$$\boldsymbol{x}_{\boldsymbol{s}}^{(j)} = \boldsymbol{A}_{(:,j)} \boldsymbol{W}_{(j,:)} \boldsymbol{x} \tag{2.23}$$

where $\boldsymbol{x}_{\boldsymbol{s}}^{(j)}$ is the $j^{th}$ original signal in the mixture domain i.e. $\boldsymbol{x} = \sum_{\boldsymbol{j}} \boldsymbol{x}_{\boldsymbol{s}}^{(j)}$. The $j^{th}$ signal is consecutively estimated and subtracted from $\boldsymbol{x}$ one by one where the subtracted $\boldsymbol{x}$ is redefined as a new obtaiained mixture $\boldsymbol{x}$. The algorithm repeats to extract the second signal and so on which is presented in Table 2.1.

Table 2.1: Algorithm of SCICA

1. Break up an observed mixture $\boldsymbol{x}$ into a sequence of adjacent blocks $\boldsymbol{X}$

2. Apply the FastICA algorithm to this matrix, to compute the unmixing matrix $\boldsymbol{W}$

3. Extract the particular signal $\boldsymbol{s}(i)$ of interest by filtering the mixture $\boldsymbol{x}$ with the corresponding row of the matrix $\boldsymbol{W}$

4. Recover the original signal $\boldsymbol{x}_{\boldsymbol{s}}^{(j)}$ by multiply the extracted signal $\boldsymbol{s}(i)$ with the $i^{th}$ column of the matrix $\boldsymbol{M}$

5. Subtract the recovered signal $\boldsymbol{x}_{\boldsymbol{s}}^{(j)}$ from the mixture $\boldsymbol{x}$, redefine the substracted mixture as $\boldsymbol{x}$, and repeat the steps from 1 – 4 to further extract the remaining signals.

This algorithm has certain limitation. For example, signals are assumed to be independent signals which are invariable in time. Secondly mixtures compose of nonoverlap spectrum-desity signals.

### *2.2.2 Nonnegative matrix factorization based SCBSS*

Nonnegative matrix factorization (NMF) was firstly introduced by Lee and Seung [29] to decomponse a matrix $\boldsymbol{X}$ into the product of two matrices $\boldsymbol{D}$ and $\boldsymbol{H}$ under the constraint that all element in $\boldsymbol{D}$ and $\boldsymbol{H}$ must be equal to or greater than zero as

$$\boldsymbol{X} \approx \boldsymbol{DH} \;, \qquad\qquad \boldsymbol{D}, \boldsymbol{H} \geq 0 \tag{2.24}$$

where $\boldsymbol{X} \in \mathbb{R}_{+}^{F \times T_s}$ is the TF representation of a mixture , $\boldsymbol{D} \in \mathbb{R}_{+}^{F \times I}$ is a matrix containing only a set of spectral basis vectors, and $\boldsymbol{H} \in \mathbb{R}_{+}^{I \times T_s}$ is an encoding matrix which describes the amplitude of each basis vector at each time unit [14]. In general NMF form, $\boldsymbol{D}$ and $\boldsymbol{H}$ can be computed by selecting the arguments that minimize a cost function $C(\cdot)$ or the divergence between $\boldsymbol{X}$ and $\boldsymbol{DH}$. This can be expressed as:

$$\{\boldsymbol{D}, \boldsymbol{H}\} = \operatorname{argmin}_{D,H \geq 0} C(\boldsymbol{X}; \boldsymbol{D}, \boldsymbol{H}) \tag{2.25}$$

Thus, NMF algorithms aim to find a local minimum of the divergence. Commonly used cost functions for NMF are Least Square (LS) distance and the generalized Kullback-Leibler (KL) divergence which have been introduced in [30]. LS distance corresponds to the assumption that the residual is independent and identically Gaussian distributed. On the other hands, KL divergence measures the relative entropy between the data $\boldsymbol{X}$ and the approximate factorization $\boldsymbol{DH}$, if $\boldsymbol{X}$ can be considered as an unnormalized discrete probability distribution.

One reason for this popularity is that NMF codes naturally favor sparse. The decompositions are computed by minimizing a cost function augmented by penalty or regularization terms that account for these constraints on the factors, $\boldsymbol{X} \approx \boldsymbol{D}\boldsymbol{H}$ where $H$ is sparse i.e. most of its elements are zero. A sparseness constraint introduced in [31] can be added to minimize the penalized cost functions. This method was termed as sparse NMF (SNMF) where the penalty term is given by a sparsity function and a control parameter. In [31], the SNMF method has been proposed to determine a set of basis for each speaker and a mixture is mapped onto the joint bases of the speakers. This technique is a powerful linear model which has the advantage of simplicity. It requires no assumption on signals such as statistical independent and non-Gaussian distribution and no grammatical model. However, the SNMF method does not model the temporal structure at all and it requires large amount of computation to determine the speaker independent basis. Moreover, it is essential to consider the temporal variation that underlies human speech. The acoustic signal and high-level temporal parameters should be mapped not only into corresponding low-level durational variations, but also into modifications of fundamental frequency and intensity [32]. To integrate these features into the SNMF, a two-dimensional model leading to the SNMF2D has thus been developed in [33]. The SNMF2D uses a double convolution to model both spreading of spectral basis and variation of temporal structure inherent in the signals. Some success has already been reported in recent literature [34, 35] to show the validity of SNMF2D in separating single channel mixture. While these approaches increase the accuracy of matrix factorization, it only works when large sample dataset is available. Moreover, it consumes significantly high computational complexity at each iteration to adapt the parameters

## 2.3 Summary

Various methods for SCBSS have been reviewed in this chapter. The methods can be generally classified as model-based and data-driven solutions. The typical difference lies in providing a priori information of signals or without any a priori knowledge of signals for model-based and data-driven methods, respectively. Given no prior information of sources, the data-driven approaches are the preferred solution to the single-channle blind signal separation problem. The SCBSS methods can incorporate advanced model of the source signals with the signal separation process. Thus, the SCBSS approaches generally deliver high quality of separation performance. However, the training process requires rigorious criterion for producing a good model such as adequate data of sources and choosing appropriate statistical models to represent the characteristic of sources. This causes high computational complexity of SCBSS methods. On the other hand, in most practical applications, only observed signal is available where lack of a prior knowledge of the source model. The SCBSS approaches are required for separating the mixture by extracting the source information from the sole observed mixture. This scenario has drawn much research interests to solve the SCBSS problem. However, current proposed methods still have constraints to make the way out of laboratories. Therefore, the SCBSS problem is still an open problem.

# CHAPTER 3

# THEORY OF PSEUDO – STEREO MIXING MODEL

In this chapter, a novel pseudo-stereo mixing model is proposed. The pseudo-stereo mixing model has an artificial resemblance of the stereo signal concept given by a single observed mixture. The proposed mixing model comprises an observed mixture and a proposed ‘pseudo-stereo’ mixture. The proposed ‘pseudo-stereo’ mixture is formulated by weighting and time-shifting the observed mixture, where the original signals are modeled by the autoregressive (AR) process. This model takes an advantage of the relationship between the readily available mixture and the pseudo-stereo mixture model to estimate the signature parameter of the original signals. Separability analysis of the proposed model has also been derived to verify that the proposed mixing model is separable. Therefore, this model relaxes the under-determined ill-conditions associated with monaural signal separation and paves the way for binaural signal separation approaches to solve monaural mixture.

The chapter is organized as follows: Section 3.1 introduces the background of AR model and a concept of stereo channels. In Section 3.2, the proposed pseudo – stereo mixture is derived in both time and frequency domains, respectively. Separability of the proposed mixing model is analyzed in Section 3.3. Section 3.4 is presented how to determine the value of the pseudo – stereo parameters. Finally, Section 3.5 concludes the chapter.

## 3.1 Background

### *3.1.1 Autoregressive Model*

Autoregressive models are Markov processes with dependence of higher order than lag-1 for univariate time series. Mathematically, AR model can be expressed as follow [36]:

$$s(t) = -\sum_{m=1}^{D} a_s(m)s(t-m) + e_j(t) \tag{3.1}$$

where $s(t)$ is a random signal, $a_s(m)$ denotes the $m^{th}$ order AR coefficient, $D$ is the maximum AR order, and $e_j(t)$ is an independent identically distributed (i.i.d.) random signal with zero mean and variance $\sigma_e^2$. In signal processing, a random signal $s(t)$ can be obtained by a linear regression from its previous time i.e. $s(t-m)$, …, $s(t-D)$ thus this is called 'autoregressive' process. AR process is considered as a practical process for time series analysis and decomposition of processes into components. Due to the fact that AR process can handle both stationary and nonstationary AR signals. For nonstationary AR signals, an online adaptive sliding-window method can be employed to update the AR process for each lastest sample [37].

### *3.1.2 Concept of Stereo Channels*

The human ears hear sound in stereo, and the brain uses the subtle differences in sound entering each ear to perform localization, mainly time, level and spectral differences between the channels [38]. Stereo recording is recording onto two separate channels, one channel for the left sound input and the other channel for the right sound input. With stereo, recording on the two channels are independent of each other. The two channels must be properly positioned to accurately capture a stereo image, and speakers must also be spaced properly to re-create a stereo image accurately. Psychoacoustic research has quantified the time and level differences adequate for directional imaging to any position on the line between left and right loudspeaker in a standard loudspeaker setup as shown in Fig.1 [39].

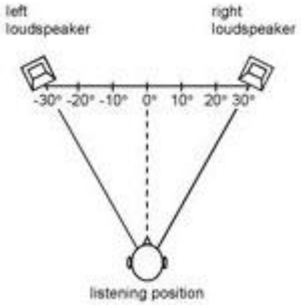


Figure 3.1: Stereo recording model.

These are the relative level (or loudness) difference between the two channels $\Delta L$, and the time delay difference in arrival times for the same sound in each channel $\Delta t$.

## 3.2 Proposed Pseudo – Stereo Mixture Model

In this chapter, for simplicity we consider the case of a mixture of two signals in time domain as

$$x_1(t) = s_1(t) + s_2(t) \tag{3.2}$$

where $x_1(t)$ is the single channel mixture, and $s_1(t)$ and $s_2(t)$ are the original signals which are assumed to be modeled by the AR process [36]:

$$s_j(t) = -\sum_{m=1}^{D_j} a_{s_j}(m;t)s_j(t-m) + e_j(t) \tag{3.3}$$

where $a_{s_j}(m;t)$ denotes the $m^{th}$ order AR coefficient of the $j^{th}$ signal at time $t$, $D_j$ is the maximum AR order. This model is particularly interesting in signal separation; firstly, many audio signals satisfy this process and secondly, it enables us to formulate a virtual mixture by weighting and time-shifting the single channel mixture $x_1(t)$ as

$$x_2(t) = \frac{x_1(t) + \gamma x_1(t-\delta)}{1+|\gamma|} \tag{3.4}$$

In (3.4), $\gamma \in \Re$ is the weight parameter, and $\delta$ is the time-delay. The mixture in (3.2) and (3.4) is termed as "pseudo-stereo" because it has an artificial resemblance of a stereo signal except that it is given by one location which results in the same time-delay but different attenuation of the source signals. To show this, we can express (3.4) in terms of the source signals, AR coefficient and time-delay as

$$\begin{aligned} x_2(t) &= \frac{x_1(t) + \gamma x_1(t-\delta)}{1+|\gamma|} \\ &= \frac{s_1(t)+s_2(t)+\gamma[s_1(t-\delta)+s_2(t-\delta)\,]}{1+|\gamma|} \\ &= \frac{-\sum_{m=1}^{D_1} a_{s_1}(m)s_1(t-m) + e_1(t)}{1+|\gamma|} + \frac{\gamma s_1(t-\delta)}{1+|\gamma|} + \frac{-\sum_{m=1}^{D_2} a_{s_2}(m)s_2(t-m) + e_2(t)}{1+|\gamma|} + \frac{\gamma s_2(t-\delta)}{1+|\gamma|} \\ &= \frac{\left(-a_{s_1}(\delta)+\gamma\right)}{1+|\gamma|} s_1(t-\delta) + \frac{\left(-a_{s_2}(\delta)+\gamma\right)}{1+|\gamma|} s_2(t-\delta) + \\ &\quad \frac{e_1(t)-\sum_{\substack{m=1\\ m\neq\delta}}^{D_1} a_{s_1}(m)s_1(t-m)}{1+|\gamma|} + \frac{e_2(t)-\sum_{\substack{m=1\\ m\neq\delta}}^{D_2} a_{s_2}(m)s_2(t-m)}{1+|\gamma|} \end{aligned} \tag{3.5}$$

Define

$$a_j(t;\delta,\gamma) = \frac{-a_{s_j}(\delta;t)+\gamma}{1+|\gamma|} \tag{3.6}$$

$$r_j(t;\delta,\gamma) = \frac{e_j(t)-\sum_{\substack{m=1\\ m\neq\delta}}^{D_j} a_{s_j}(m;t)s_j(t-m)}{1+|\gamma|} \tag{3.7}$$

where $a_j(t;\delta,\gamma)$ and $r_j(t;\delta,\gamma)$ represent the mixing attenuation and residue of the $j^{th}$ signal, respectively. Note that the parameterization of $a_j(t;\delta,\gamma)$ and $r_j(t;\delta,\gamma)$ depends on $\delta$ and $\gamma$ although this is not shown explicitly. By comparing with the single channel mixture, the pseudo-stereo mixture $x_2(t)$ contains extra information i.e. $a_j(t),\delta,r_j(t)$ which are used to construct the complex 2D histogram for estimating the signals. Using (3.6) and (3.7), the overall proposed mixing model of the SOLO can now be formulated in terms of the signals as

$$\begin{aligned} x_1(t) &= s_1(t) + s_2(t) \\ x_2(t) &= a_1(t;\delta,\gamma)s_1(t-\delta) + a_2(t;\delta,\gamma)s_2(t-\delta) + r_1(t;\delta,\gamma) + r_2(t;\delta,\gamma) \end{aligned} \quad (3.8)$$

### *3.2.1 Model Assumption*

***Assumption 1***: The source signals satisfy the local stationarity of the time-frequency representation. This refers to the approximation of $S_j(\tau-\phi,\omega) \approx S_j(\tau,\omega)$ where $\phi$ is the maximum time-delay (shift) associated with $F^W(\cdot)$ with an appropriate window function $W(\cdot)$. If $\phi$ is small compared with the length of $W(\cdot)$ then $W(\cdot-\phi) \approx W(\cdot)$ [40]. Hence, the Fourier transform of a windowed function with shift $\phi$ yields approximately the same Fourier transform without $\phi$. For the proposed method, the pseudo-stereo mixture is shifted by $\delta$ and by invoking the local stationarity this leads to

$$\begin{aligned} s_j(t-\delta) &\xrightarrow{STFT} e^{-i\omega\delta}S_j(\tau-\delta,\omega) \\ &\approx e^{-i\omega\delta}S_j(\tau,\omega) \quad , \quad \forall\delta, |\delta| \le \phi \end{aligned} \quad (3.9)$$

Thus, the STFT of $s_j(t-\delta)$ where $|\delta| \leq \phi$ is approximately $e^{-i\omega\delta}S_j(\tau,\omega)$ according to the local stationarity.

***Assumption 2***: The source signals satisfy the windowed-disjoint orthogonality (WDO) condition where different signals are approximately orthogonal to each other [41]:

$$S_i(\tau,\omega)S_j(\tau,\omega) \approx 0, \qquad \forall i \neq j\,,\ \forall \tau,\omega \tag{3.10}$$

where $S_j(\tau,\omega)$ is the STFT of $s_j(t)$ defined as

$$\begin{aligned} S_j(\tau,\omega) &= F^W\big[s_j(t)\big](\tau,\omega) \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} W(t-\tau)\, s_j(t) e^{-i\omega t} dt \end{aligned} \tag{3.11}$$

and $W(t)$ is the window function. The STFT is performed on the signal frame-by-frame and thus, $\tau$ represents the window shift.

**Assumption 3**: Phase Ambiguity. The factor $e^{-i\omega\delta}$ is only uniquely specified if $|\omega\delta| < \pi$, otherwise this would cause phase-wrap [42]. Selecting improper time-delay $\delta$ will lead to phase-wrap if the maximum frequency of the signal is exceeded. In order to avoid phase ambiguity, we must satisfy

$$|\omega_{max}\delta_{max}| < \pi \tag{3.12}$$

where $\omega_{max} = \frac{2\pi f_{max}}{f_s}$ , $\delta_{max}$ is the maximum time delay, $f_{max}$ is the maximum frequency present in the signals and $f_s$ is the sampling frequency. Hence, $\delta_{max}$ can be determined from (3.12) according to

$$\delta_{max} < \frac{f_s}{2f_{max}} \tag{3.13}$$

As long as the delay parameter is less than $\delta_{max}$, there will not be any phase ambiguity. For example, for a maximum frequency $f_{max} = 3.5\,kHz$, and a sampling frequency $f_s = 16\,kHz$, one obtains $\delta_{max} < 2.28$ using (3.13). Therefore, phase ambiguity can be avoided provided $\delta$ is selected to be either 1 or 2. Additionally, for a maximum frequency $f_{max} = 8\,kHz$ the maximum delay $\delta_{max}$ is limited to 1 only. This condition will be used to determine the range of $\delta$ in formulating the pseudo-stereo mixture.

### *3.2.2 Frequency Domain*

Based on the above assumptions, the TF representation of the mixing model is obtained using the STFT of $x_j(t),\ j = 1,2$ as

$$X_1(\tau,\omega) = S_1(\tau,\omega) + S_2(\tau,\omega)$$

$$X_2(\tau,\omega) = a_1(\tau)e^{-i\omega\delta}S_1(\tau-\delta,\omega) + a_2(\tau)e^{-i\omega\delta}S_2(\tau-\delta,\omega) - \left(\sum_{\substack{m=1\\m\neq\delta}}^{D_1}\frac{a_{s_1}(m;\tau)}{1+|\gamma|}e^{-i\omega m}S_1(\tau-m,\omega) + \sum_{\substack{m=1\\m\neq\delta}}^{D_2}\frac{a_{s_2}(m;\tau)}{1+|\gamma|}e^{-i\omega m}S_2(\tau-m,\omega)\right) \tag{3.14}$$

for $\forall\tau,\omega$. In (3.14), we have used the fact that $e_j(t) \ll s_j(t)$, thus the TF of $r_j(t)$ in (3.7) simplifies to

$$R_j(\tau,\omega) = -\sum_{\substack{m=1\\m\neq\delta}}^{D_j}\frac{a_{s_j}(m;\tau)}{1+|\gamma|}e^{-i\omega m}S_j(\tau-m,\omega) \tag{3.15}$$

To facilitate further analysis, we also define

$$C_j(\tau,\omega) = \frac{1}{1+|\gamma|}\sum_{\substack{m=1\\ m\neq\delta}}^{D_j} a_{s_j}(m;\tau)e^{-i\omega(m-\delta)} \tag{3.16}$$

which forms part of $R_j(\tau,\omega)$ without the contribution of the signal $S_j(\tau,\omega)$. Assuming that the $j^{th}$ signal is dominant at a particular TF unit, (3.14) can be simplified by using (3.6) and (3.16) as follows:

$$\begin{aligned} X_1(\tau,\omega) &= S_j(\tau,\omega) \\ X_2(\tau,\omega) &= a_j(\tau)e^{-i\omega\delta}S_j(\tau-\delta,\omega) - \sum_{\substack{m=1\\ m\neq\delta}}^{D_j}\frac{a_{s_j}(m;\tau)}{1+|\gamma|}e^{-i\omega m}S_j(\tau-m,\omega) \quad (3.17) \\ &\approx \left[a_j(\tau) - C_j(\tau,\omega)\right]e^{-i\omega\delta}S_j(\tau,\omega) \quad , \quad (\tau,\omega)\in\Omega_j \end{aligned}$$

for $\delta$ and $m \quad \leq \phi$ , and $\Omega_j$ is the active area of $S_j(\tau,\omega)$ defined as $\Omega_j := \{(\tau,\omega): S_j(\tau,\omega) \neq 0,\ \forall k \neq j\}$. From (3.17), it can be seen that the pseudo-stereo mixture comprises three components i.e. $a_j e^{-i\omega\delta}$, $C_j(\tau,\omega)$ and $S_j(\tau,\omega)$. A careful analysis of (3.17) will reveal that even if $S_j(\tau,\omega)$ is unknown, the signature of each signal can be extracted directly from $X_1(\tau,\omega)$ using only information of $a_j e^{-i\omega\delta}$ and $C_j(\omega)$. Thus, this constitutes the separability of the proposed mixing model which will be analyzed in the following section.

### 3.3 Analysis of Separability of Pseudo – stereo mixing model

The separability of the proposed mixing model can be examined from the pseudo-stereo mixture by considering $a_j(t;\delta,\gamma)$ and $r_j(t;\delta,\gamma)$ in the following three cases. Case I refers to identical signals mixed in the single channel, Case II represents different signals but setting $\gamma$ and $\delta$ for the pseudo-stereo mixture such that $a_1(t;\delta,\gamma) = a_2(t;\delta,\gamma)$, and Case III corresponds to the most general case where the signals are distinct, and $\gamma$ and $\delta$ are selected arbitrarily such that the mixing attenuations and residues are also different. The above cases are evaluated by the maximum likelihood (ML) cost function for the $j^{th}$ signal which is derived from the ML framework. Firstly, the Gaussian likelihood function is formulated by using (3.17) as

$$L_j(\tau,\omega) := p(X_1(\tau,\omega), X_2(\tau,\omega) | S_j(\tau,\omega), \bar{a}_j(\tau,\omega), \delta)$$

$$= C \cdot \ exp\left( -\frac{1}{2} \sum\nolimits_{(\tau,\omega)\in\Omega_j} \frac{\left|X_1(\tau,\omega)-S_j(\tau,\omega)\right|^2}{\sigma_1^2(\tau,\omega)} + \frac{\left|X_2(\tau,\omega)-\bar{a}_j(\tau,\omega)e^{i\omega\delta}S_j(\tau,\omega)\right|^2}{\sigma_2^2(\tau,\omega)} \right) \quad (3.18)$$

where $C$ is a normalizing constant, $X_1(\tau,\omega)$ and $X_2(\tau,\omega) \in \Omega_j$. Maximizing (3.18) is equivalent to maximizing the following:

$$L_j(\tau,\omega) = -\sum\nolimits_{(\tau,\omega)\in\Omega_j} \frac{\left|X_1(\tau,\omega)-S_j(\tau,\omega)\right|^2}{\sigma_1^2(\tau,\omega)} + \frac{\left|X_2(\tau,\omega)-\bar{a}_j(\tau,\omega)e^{i\omega\delta}S_j(\tau,\omega)\right|^2}{\sigma_2^2(\tau,\omega)} \quad (3.19)$$

Secondly, the Gaussian likelihood function is maximized with respect to $S_j(\tau,\omega)$. The ML of $S_j(\tau,\omega)$ is obtained by solving $\partial L(\tau,\omega)/\partial S_j(\tau,\omega) = 0$ for $\forall(\tau,\omega) \in \Omega_j$ as below:

$$\frac{\partial L(\tau,\omega)}{\partial S_j(\tau,\omega)} = \frac{\partial}{\partial S_j(\tau,\omega)} \left( \frac{\left(X_1(\tau,\omega)-S_j(\tau,\omega)\right)\left(X_1^*(\tau,\omega)-S_j^*(\tau,\omega)\right)}{\sigma_1^2(\tau,\omega)} \right.$$

$$\left. - \frac{\left(X_2(\tau,\omega)-\bar{a}_j(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega)\right)\left(X_2^*(\tau,\omega)-\bar{a}_j(\tau,\omega)e^{i\omega\delta}S_j^*(\tau,\omega)\right)}{\sigma_2^2(\tau,\omega)} \right)$$

$$= \frac{-\left(X_1(\tau,\omega)-S_j(\tau,\omega)\right)}{\sigma_1^2(\tau,\omega)} - \frac{\left(X_2(\tau,\omega)-\bar{a}_j(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega)\right)\bar{a}_j(\tau,\omega)e^{i\omega\delta}}{\sigma_2^2(\tau,\omega)} \quad (\tau,\omega) \in \Omega_j \quad (3.20)$$

Equating the above to zero, $S_j^{ML}(\tau,\omega)$ can be derived as

$$\frac{-\left(X_1(\tau,\omega)-S_j(\tau,\omega)\right)}{\sigma_1^2(\tau,\omega)}=\frac{\left(X_2(\tau,\omega)-\bar{a}_j(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega)\right)\bar{a}_j(\tau,\omega)e^{i\omega\delta}}{\sigma_2^2(\tau,\omega)}$$

$$\frac{\sigma_2^2(\tau,\omega)+\sigma_1^2(\tau,\omega)\bar{a}_j^2(\tau,\omega)}{\sigma_1^2(\tau,\omega)\sigma_2^2(\tau,\omega)}S_j(\tau,\omega)=\frac{\sigma_2^2(\tau,\omega)X_1(\tau,\omega)+\sigma_1^2(\tau,\omega)\bar{a}_j e^{i\omega\delta}X_2(\tau,\omega)}{\sigma_1^2(\tau,\omega)\sigma_2^2(\tau,\omega)}$$

$$\therefore \quad S_j^{ML}(\tau,\omega)=\frac{\sigma_2^2(\tau,\omega)X_1(\tau,\omega)+\sigma_1^2(\tau,\omega)\bar{a}_j e^{i\omega\delta}X_2(\tau,\omega)}{\sigma_2^2(\tau,\omega)+\sigma_1^2(\tau,\omega)\bar{a}_j^2(\tau,\omega)}\ ,(\tau,\omega)\in\Omega_j \quad (3.21)$$

Subsequently, the obtained result (3.21) is substituted into the Gaussian likelihood function (3.19) and assuming that $\sigma_1^2(\tau,\omega)\approx\sigma_2^2(\tau,\omega)=\sigma^2(\tau,\omega)$, we then have

$$F_j(\tau,\omega)=-\sum_{(\tau,\omega)\in\Omega_j}\frac{\left|\bar{a}_j(\tau,\omega)e^{-i\omega\delta}X_1(\tau,\omega)-X_2(\tau,\omega)\right|^2}{1+\bar{a}_j^2(\tau,\omega)} \quad (3.22)$$

Maximizing (3.22) is equivalent to minimizing the following:

$$G(\tau,\omega)=\underset{k}{argmin}\left|\bar{a}_k(\tau,\omega)e^{-i\omega\delta}X_1(\tau,\omega)-X_2(\tau,\omega)\right|^2 \quad (3.23)$$

Using the proposed pseudo-stereo mixture, the mixture can be expressed in term of the $j^{th}$ signal as $x_2(t)=\frac{s_j(t)+\gamma s_j(t-\delta)}{1+|\gamma|}$ in time domain where the TF representation of this mixture is:

$$X_2(\tau,\omega)=\frac{S_j(\tau,\omega)+\gamma e^{-i\omega\delta}S_j(\tau-\delta,\omega)}{1+|\gamma|} \quad ,\ (\tau,\omega)\in\Omega_j \quad (3.24)$$

for $\delta\leq\phi$. Invoking the local stationary condition for (3.24), $X_2(\tau,\omega)$ can be now expressed as:

$$X_2(\tau,\omega)\approx\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)X_1(\tau,\omega) \quad (3.25)$$

In this light, the proposed ML cost function can finally be formulated based on the single mixture $X_1(\tau,\omega)$ by substituting this relation into $G(\tau,\omega)$ in (3.23). The proposed ML cost function then obtains:

$$J(\tau,\omega) = \underset{k}{argmin}\left|\bar{a}_k(\tau,\omega)e^{-i\omega\delta}X_1(\tau,\omega) - \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)X_1(\tau,\omega)\right|^2 \tag{3.26}$$

where

$$\bar{a}_k(\tau,\omega) = \frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}e^{i\omega\delta} = a_k(\tau) - C_k(\tau,\omega)$$

$$a_k(\tau) = \frac{-a_{s_k}(\delta;\tau)+\gamma}{1+|\gamma|}$$

$$C_k(\tau,\omega) = \sum_{\substack{m=1\\ m\neq\delta}}^{D_k}\frac{a_{s_k}(m;\tau)e^{-i\omega(m-\delta)}}{1+|\gamma|}$$

Technically, the ML cost function (3.26) partitions the TF plane of the mixed signal into $k$ groups of $(\tau,\omega)$ units by evaluating the cost function. For each TF unit, the $k^{th}$ argument that gives the minimum cost will be assigned to the $k^{th}$ signal. Technically, this function partitions the TF plane of the mixed signal into $k$ groups of $(\tau,\omega)$ units by evaluating the cost function. For each TF unit, the $k^{th}$ argument that gives the minimum cost will be assigned to the $k^{th}$ signal.

Eq. (3.26) can further be expressed in term of the $j^{th}$ signal by using the observed mixture where $X_1(\tau,\omega) = S_j(\tau,\omega)$ and therefore, (3.26) then becomes

$$J(\tau,\omega) = \underset{k}{argmin}\left|\bar{a}_k(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega) - \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)S_j(\tau,\omega)\right|^2$$

$$= \underset{k}{argmin}\left|\bar{a}_k(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega) - \frac{S_j(\tau,\omega)}{1+|\gamma|} - \frac{\gamma e^{-i\omega\delta}}{1+|\gamma|}S_j(\tau,\omega)\right|^2$$

$$= \underset{k}{argmin}\left|\bar{a}_k(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega) + \sum_{m=1}^{D_j}\frac{a_{s_j}(m;\tau)e^{-i\omega m}}{1+|\gamma|}S_j(\tau-m,\omega) - \frac{\gamma e^{-i\omega\delta}}{1+|\gamma|}S_j(\tau,\omega)\right|^2$$

$$= \underset{k}{argmin}\left|\bar{a}_k(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega) + \sum_{\substack{m=1\\ m\neq\delta}}^{D_j}\frac{a_{s_j}(m;\tau)e^{-i\omega m}}{1+|\gamma|}S_j(\tau-m,\omega)\right.$$

$$\left. - \left(\frac{-a_{s_j}(\delta;\tau)+\gamma}{1+|\gamma|}\right)e^{-i\omega\delta}S_j(\tau,\omega)\right|^2$$

$$= \underset{k}{argmin}\left|a_k(\tau)e^{-i\omega\delta}S_j(\tau,\omega) - C_k(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega) + \right.$$

$$\left.\sum_{\substack{m=1\\m\neq\delta}}^{D_j}\frac{a_{s_j}(m;\tau)e^{-i\omega m}}{1+|\gamma|}S_j(\tau-m,\omega)-a_j(\tau)e^{-i\omega\delta}S_j(\tau,\omega)\right|^2 \tag{3.27}$$

Henceforth, the proposed pseudo-stereo mixing model is considered in the following three cases and evaluated by using (3.27):

**Case I**:

Identical signals mixed in the single channel which can be expressed as follows:

If $a_1(t;\delta,\gamma)=a_2(t;\delta,\gamma)=a(t;\delta,\gamma)$ and $r_1(t;\delta,\gamma)=r_2(t;\delta,\gamma)=r(t;\delta,\gamma)$, then $x_2(t)=\left(\frac{a(t;\delta,\gamma)+\gamma}{1+|\gamma|}\right)x_1(t-\delta)+2r(t;\delta,\gamma)$.

In this case, there is no benefit achieved at all. The second mixture is simply formulated as a time-delayed of the first mixture multiply by a scalar plus the redundant residue. The separability of this case is presented by substituting the pseudo-stereo mixture of Case I into the cost function. Since both residues are equal, then $C_1(\tau,\omega)=C_2(\tau,\omega)=C(\tau,\omega)=\frac{1}{1+|\gamma|}\sum_{\substack{m=1\\m\neq\delta}}^{D}a_s(m;\tau)e^{-i\omega(m-\delta)}$. For Case I, the cost function (3.27) becomes:

$$J(\tau,\omega)=\underset{k}{argmin}\left|a(\tau)e^{-i\omega\delta}S_j(\tau,\omega)-C(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega)+\right.$$

$$\left.\sum_{\substack{m=1\\m\neq\delta}}^{D}\frac{a_s(m;\tau)e^{-i\omega m}}{1+|\gamma|}S_j(\tau-m,\omega)-a(\tau)e^{-i\omega\delta}S_j(\tau,\omega)\right|^2$$

Invoking the local stationarity of the signals $S_j\left(\tau-D_j,\omega\right)=S_j(\tau,\omega)$ for $\left|D_j\right|\leq\phi$, the above leads to

$$J(\tau,\omega)=\underset{k}{argmin}\left|\sum_{\substack{m=1\\m\neq\delta}}^{D}\frac{\left(a_s(m;\tau)e^{-i\omega m}-a_s(m;\tau)e^{-i\omega m}\right)}{1+|\gamma|}\right|^2\left|S_j(\tau,\omega)\right|^2$$

$$=0 \qquad \text{for } \forall k. \tag{3.28}$$

As a result, the cost function $J(\tau, \omega)$ is zero for all $k$ arguments i.e. $J_1 = J_2 = 0$. In this case, the cost function cannot distinguish the $k$ arguments, the mixture is not separable.

**Case II**:

Different signals but setting $\gamma$ and $\delta$ for the pseudo-stereo mixture such that $a_1(t; \delta, \gamma) = a_2(t; \delta, \gamma)$ which can be expressed as follows:

If $a_1(t; \delta, \gamma) = a_2(t; \delta, \gamma) = a(t; \delta, \gamma)$ and $r_1(t; \delta, \gamma) \neq r_2(t; \delta, \gamma)$, then

$x_2(t; \delta, \gamma) = \left(\frac{a(t;\delta,\gamma)+\gamma}{1+|\gamma|}\right) x_1(t - \delta) + r_1(t; \delta, \gamma) + r_2(t; \delta, \gamma)$.

This case is similar to the previous case, but differs only in terms of $r_1(t; \delta, \gamma) \neq r_2(t; \delta, \gamma)$. As each residue $r_j(t; \delta, \gamma)$ is related to the $j^{th}$ signal via $C_j(\tau, \omega)$, the separability of this mixture can be analyzed using (3.27) as

$$J(\tau, \omega) = \underset{k}{argmin} \left| a(\tau)e^{-i\omega\delta} S_j(\tau, \omega) - C_k(\tau, \omega)e^{-i\omega\delta} S_j(\tau, \omega) + \sum_{\substack{m=1 \\ m\neq\delta}}^{D_j} \frac{a_{s_j}(m;\tau)e^{-i\omega m}}{1+|\gamma|} S_j(\tau - m, \omega) - a(\tau)e^{-i\omega\delta} S_j(\tau, \omega) \right|^2$$

$$= \underset{k}{argmin} \left| \sum_{\substack{m=1 \\ m\neq\delta}}^{D_j} \frac{\left(a_{s_j}(m;\tau) - a_{s_k}(m;\tau)\right)}{1+|\gamma|} e^{-i\omega m} \right|^2 \left|S_j(\tau, \omega)\right|^2 \qquad (3.29)$$

It can be deduced from the above that the cost function yields a zero value for $k = j$, and nonzero value for $k \neq j$. Despite the mixing attenuation of both signals are identical, the cost function is still able to distinguish the $k$ arguments by using only the difference of residues. Therefore, the mixture of Case II is separable.

**Case III**:

General case where the signals are distinct, and $\gamma$ and $\delta$ are selected arbitrarily such that the mixing attenuations and residues are also different. Case III can be expressed as follows:

If $a_1(t;\delta,\gamma) \neq a_2(t;\delta,\gamma)$ and $r_1(t;\delta,\gamma) \neq r_2(t;\delta,\gamma)$ (or $r_1(t;\delta,\gamma) = r_2(t;\delta,\gamma)$ )

then $x_2(t) = \left(\frac{a_1(\delta;\tau)+\gamma}{1+|\gamma|}\right) s_1(t-\delta) + \left(\frac{a_2(\delta;\tau)+\gamma}{1+|\gamma|}\right) s_2(t-\delta) + r_1(t;\delta,\gamma) + r_2(t;\delta,\gamma)$

We first treat the situation of $r_1(t;\delta,\gamma) = r_2(t;\delta,\gamma)$. Since the mixing attenuations $a_1(\tau)$ and $a_2(\tau)$ correspond respectively to $s_1(t)$ and $s_2(t)$ then the cost function can be expressed as

$$J(\tau,\omega) = \underset{k}{argmin} \Big| a_k(\tau) e^{-i\omega\delta} S_j(\tau,\omega) - C(\tau,\omega) e^{-i\omega\delta} S_j(\tau,\omega) + \sum_{\substack{m=1 \\ m\neq\delta}}^{D} \frac{a_S(m;\tau)e^{-i\omega m}}{1+|\gamma|} S_j(\tau-m,\omega) - a_j(\tau) e^{-i\omega\delta} S_j(\tau,\omega) \Big|^2$$

$$= \underset{k}{argmin} \left| \big(a_k(\tau) - a_j(\tau)\big) e^{-i\omega\delta} + \sum_{\substack{m=1 \\ m\neq\delta}}^{D} \frac{(a_S(m;\tau) - a_S(m;\tau))}{1+|\gamma|} e^{-i\omega m} \right|^2 \left| S_j(\tau,\omega) \right|^2$$

$$= \underset{k}{argmin} \left| \big(a_k(\tau) - a_j(\tau)\big) e^{-i\omega\delta} \right|^2 \left| S_j(\tau,\omega) \right|^2 \tag{3.30}$$

This cost function yields a nonzero value only for $k \neq j$. In this case, the cost function can separate the $k$ arguments due to the difference between $a_k$ and $a_j$. The case of $r_1(t;\delta,\gamma) \neq r_2(t;\delta,\gamma)$ follows similar line of argument as above where the cost function becomes

$$J(\tau,\omega) = \underset{k}{argmin} \left[ \left| \big(a_k(\tau) - a_j(\tau)\big) e^{-i\omega\delta} \quad + \sum_{\substack{m=1 \\ m\neq\delta}}^{D_j} \frac{\big(a_{s_j}(m;\tau) - a_{s_k}(m;\tau)\big)}{1+|\gamma|} e^{-i\omega m} \right|^2 \left| S_j(\tau,\omega) \right|^2 \right] \tag{3.31}$$

This cost function yields a nonzero value only for $k \neq j$; thus the cost function is able to distinguish the $k$ arguments.

In summary, by considering $a_j(t)$ and $r_j(t)$ with respect to the above three cases, only Case II and Case III are separable. Hence, the proposed pseudo-stereo mixing model can be separated to unveil the original signals by using any binaural blind signal separation methods.

## 3.4 Determination of the values of $\gamma$ and $\delta$

The separability of the proposed method depends on the signals' AR coefficients estimated from the relation of $X_1(\tau,\omega)$, $X_2(\tau,\omega)$ and their residues. The weight $\gamma$ parameter acts as a controlling factor to maintain the difference of the signals' AR coefficients and to control the amount of the residues $r_j(t;\delta,\gamma)$. On the one hand, if $\gamma \gg a_{s_j}(\delta;\tau)$ then the distinguishing ability of the mixing attenuations $a_j(t;\delta,\gamma)$ will tend to be small such that $a_1(t;\delta,\gamma) = a_2(t;\delta,\gamma)$ and thereby we lose the benefit of the pseudo-mixture signal. In addition, it reduces the residues in (6) which subsequently diminishes the contribution of $C_j(\tau,\omega)$ in $\bar{a}_j(\tau,\omega)$. On the other hand, if $\gamma \ll a_{s_j}(\delta;\tau)$ then $x_2(t)$ becomes closer to $x_1(t)$. In the extreme case of $\gamma = 0$ this leads to $x_1(t) = x_2(t)$ where the pseudo-stereo mixture cannot be formulated. Therefore, to this end, we propose the following criterion to balance both extremes.

### *3.4.1 Mixing Attenuation Distinguishability (MAD)*

We define the distinguishability function of the mixing attenuation as

$$\begin{aligned}\theta &= \frac{\left|a_k e^{-i\omega\delta} X_1(\tau,\omega) - X_2(\tau,\omega)\right|^2}{|X_1(\tau,\omega)|^2} \\ &= \frac{\left|a_k e^{-i\omega\delta} S_j(\tau,\omega) - a_j e^{-i\omega\delta} S_j(\tau,\omega)\right|^2}{\left|S_j(\tau,\omega)\right|^2} \\ &= \left|a_k - a_j\right|^2 \qquad (3.32)\end{aligned}$$

for $k \neq j$. The second line of (3.32) is obtained using the pseudo-stereo mixing model (3.17). Larger value of $\theta$ implies that the mixing attenuations between the two signals are distant further from each other. This will yield two distinct peaks in the complex 2D histogram. Alternatively, we can use the concept of symmetric mixing attenuation $\alpha_j$ which is defined as $\alpha_j := a_j - 1/a_j$. In this case, the distinguishability in terms of $\alpha_k$ and $\alpha_j$ takes the form of

$$\theta = \left|\alpha_k - \alpha_j\right|^2 \tag{3.33}$$

### *3.4.2 Attenuation-to-Residue Ratio (ARR)*

$$\mathrm{ARR} = \underset{(\gamma,\delta)}{\operatorname{argmax}} \frac{\theta}{\left|\frac{\kappa_1(\tau,\omega)+\kappa_2(\tau,\omega)}{\kappa_1(\tau,\omega)-\kappa_2(\tau,\omega)}\right|^2} \tag{3.34}$$

where $\kappa_j(\tau,\omega) = \sqrt{\left(D_j - 1\right)\sum_{\substack{m=1 \\ m\neq\delta}}^{D_j}\left|\frac{a_{S_j}(m)}{1+|\gamma|}\right|^2}$ denotes the supremum of $C_j(\tau,\omega)$ which is obtained by applying the Schwarz's inequality to $C_j(\tau,\omega)$. The term $|\kappa_1(\tau,\omega) - \kappa_2(\tau,\omega)|^2$ refers to the maximum difference of residues inspired by the cost function of Case II in Section 3.3: Analysis of Separability of Pseudo – stereo mixing model. On the other hand, the term $|\kappa_1(\tau,\omega) + \kappa_2(\tau,\omega)|^2$ refers to the combined residues inherent in the mixture. In a nutshell, the ARR measures the proportion of distinguishability between the mixing attenuations and AR coefficients residue. The ARR is always positive. In the event where the estimated $\hat{a}_j$ of the two signals are so close together that the peak regions overlap with one another, then this overlap will cause *ambiguity* in identifying the unique peaks. The higher value of ARR represents the larger difference of $\hat{a}_j$ between the signals. Thus, choosing the appropriate $\gamma$ and $\delta$ such that the two peak regions are clearly distinct in the complex 2D histogram is important. As the peaks can be identified more precisely, more accurate mask can therefore be constructed and subsequently yields better separation performance as shown by the the signal-to-distortion ratio (SDR). The SDR is defined as SDR= $10\, log_{10}\left(\|s_{target}\|^2/\|e_{interf}+e_{artif}\|^2\right)$ [43]. Table 3.1 summarizes the steps for determining the range of values for $\gamma$ and $\delta$.

Table 3.1: Determination of $\gamma$ and $\delta$

1. Select an arbitrary range of $\delta$ that satisfies Phase Ambiguity assumption:
2. Calculate the $ARR$ matrices within the range of $\gamma$ and $\delta$ using (3.33) and (3.34).
3. Choose pairs of $\gamma$ and $\delta$ such that the $ARR$ is greater than a threshold i.e.

$$\Lambda = \{\, (\gamma, \delta) | ARR > \psi\} \tag{3.35}$$

where $\Lambda$ is the set of the selected pairs, and $\psi$ is a threshold[1].

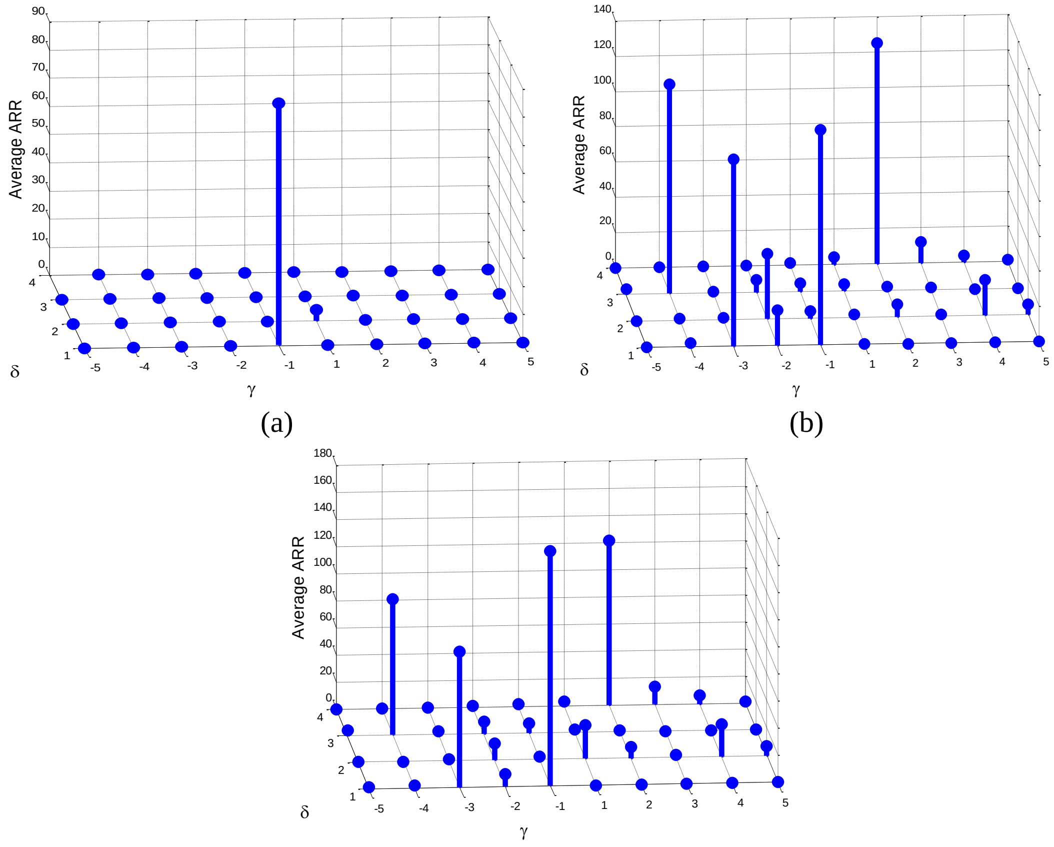
A set of experiments has been conducted to determine the $\gamma$ and $\delta$ pairs by using real-audio signals from TIMIT and RWC [44] databases. 75 types of mixtures were constructed from these databases which were divided into 3 categories: speech and speech (SS), music and music (MM), speech and music (SM). Each type contains two signals and each signal has unit power. All experiments were performed under the following conditions: STFT of 1024-point with 50% overlap [46] and sampling frequency of 16 kHz. Source's AR coefficients were calculated by using Yule-Walker method. A finite range of $\gamma$ and $\delta$ was selected to be [-5, 5] (excluding $\gamma = 0$) and [1,

[1] By means of Monte-Carlo experiments [45], $\psi = 20$ has been experimentally verified to yield satisfactory performance.

4] , respectively. Following the steps in Table 3.2, the set of the pairs $\Lambda$ given the threshold $\psi = 20$ has been found by calculating the average ARR for each category and this is plotted in Fig. 3.2. The results in Fig. 3.2(a) show at least a pair was found for the SS category i.e. $\Lambda = \{\, (\gamma, \delta) | ARR > 20\} = (-1,1)$. The results indicate that only the low order AR coefficients i.e. $\delta = 1$ are beneficial for separation. This is not surprising since speech is mainly characterized by the initial few AR coefficients and these coefficients tend to vary for different speeches. We have also noted the effect of $\gamma$ on the ARR. As $\gamma$ increases in magnitude of both positive and negative directions, $\theta$ and $C_j(\tau, \omega)$ become progressively smaller such that the ARR is almost zero. Fig.3.2(b) shows the results for the MM category with 9 pairs identified as $\Lambda = \{(-4,3), (-3,1), (-2,1), (-2,2), (-1,1), (2,2), (2,4), (3,4), (4,2)\}$. Music signal has AR coefficients that tend to span a large dynamic range and this has therefore contributed to the MM characteristic in Fig.3.2(b). Finally, Fig.3.2(c) shows the results of the SM category with 6 pairs identified as $\Lambda = \{(-4,3), (-3,1), (-1,1), (1,2), (2,4), (4,2)\}$. One may note that both MM and SM categories have broader range of $\gamma$ and $\delta$ than the SS group due to the difference of the AR coefficients at the corresponding order. It is also interesting to observe that several common pairs overlap between the MM and SM categories and these have been tabulated in Table 3.2.

Table 3.2: Common pairs of $(\gamma, \delta)$ in the MM and SM categories

| $\gamma$ | -1 | 4 | 2 |
|---|---|---|---|
| $\delta$ | 1 | 2 | 4 |



(a)

(b)

(c)

Figure 3.2: Set of $\gamma$ and $\delta$ for mixture of (a) speech and speech (SS), (b) music and music (MM), and (c) speech and music (SM).

In the case where the type of signals is unknown, then choosing $(\gamma, \delta) = (-1,1)$ will yield the best possible ARR since this particular pair overlaps with all the three categories. In practice, the AR coefficients of signals are generally unknown. However, if one knows the signal category then $\gamma$ and $\delta$ can be chosen from $\Lambda$. Moreover, if specific information of the signals such as piano or English sentence is known in advance then the AR coefficients can be determined by randomly sample the signals that belong to those groups. Hence, this enables the algorithm to estimate $\delta$ and $\gamma$ for the specific type of signals.

## 3.5 Summary

In this chapter, a novel pseudo-stereo mixture has been proposed by time-delaying and weighting the observed single-channel mixture. The separability of the proposed mixture model is analyzed under three cases:

Case I: $a_1(t) = a_2(t), r_1(t) = r_2(t)$

Case II: $a_1(t) = a_2(t), r_1(t) \neq r_2(t)$

Case III: $a_1(t) \neq a_2(t), r_1(t) \neq r_2(t)$ (or $(r_1(t) = r_2(t))$

From analysis, if at least one parameter of the signals i.e. $a_j(t)$ or $r_{\mathrm{j}}(t)$ has the different values, the artificial stereo model is separable as in Cases II and III. Moreover, $a_j(t)$ and $r_j(t)$ characterise the $j^{th}$ signal. This work overcomes the under-determined system representation associated with monaural signal separation and path the way forward for binaural signal separation approaches to solve monaural mixture. Additionally, the recommended ranges of the $(\gamma, \delta)$ pairs have been provided for all types of audio mixture based on the proposed ARR.

# CHAPTER 4

## SINGLE CHANNEL BLIND SIGNAL SEPARATION USING PSEUDO – STEREO MIXTURE AND COMPLEX 2D HISTOGRAM

In the Chapter 3 Section 3.2, the pseudo-stereo mixture was formulated artificial stereo mixtures given by the sole single-channel mixture. In this chapter, a novel algorithm using the pseudo-stereo mixing model to solve the SCBSS problem is developed. The proposed algorithm is independent of initialization and does not require iterative optimization, and a priori knowledge of the original signals. The proposed algorithm comprises two steps: 1) Estimation of original signal characteristics based on the ratio of AR coefficients between the pseudo-stereo mixtures model. The signal-character estimation will be computed via the proposed complex 2-dimentional histogram. 2) Construction of a binary time-frequency (TF) mask using only the single-channel mixture, the binary TF mask is constructed by evaluating the cost function given by the estimated signals' weighted AR coefficients. Conditions required for a unique mask construction based on the maximum likelihood have also been identified. The proposed algorithm is tested on both synthetic and real-audio signal. As results, the proposed algorithm yields superior performance and is computationaly very fast compared with existing SCBSS methods.

The chapter is organied as follows: Section 4.1 summarizes the pseudo-stereo mixing model. The proposed algorithm is fully developed in Section 4.2. Experimental results coupled with a series of performance comparison with other SCBSS method are presented in Section 4.3. Finally, Section 4.4 concludes this chapter.

## 4.1 Background

In the Chapter 3 Section 3.2, the pseudo-stereo mixture is formulated by weighting $\gamma$ and time-shifting $\delta$ the single channel mixture $x_1(t)$.

$$x_2(t) = \frac{x_1(t) + \gamma x_1(t-\delta)}{1+|\gamma|} \tag{4.1}$$

where the original signals $s(t)$ are assumed to be modeled by the autoregressive (AR) process i.e. $s_j(t) = -\sum_{m=1}^{D_j} a_{s_j}(m;t)s_j(t-m) + e_j(t)$. As a result, the pseudo-stereo mixing model of two signals; $s_1(t)$ and $s_2(t)$, can be expressed in time domain as

$$\begin{aligned} &x_1(t) = s_1(t) + s_2(t) \\ &x_2(t) = a_1(t;\delta,\gamma)s_1(t-\delta) + a_2(t;\delta,\gamma)s_2(t-\delta) + r_1(t;\delta,\gamma) + r_2(t;\delta,\gamma) \end{aligned} \tag{4.2}$$

where $a_j(t;\delta,\gamma) = \frac{-a_{s_j}(\delta;t)+\gamma}{1+|\gamma|}$ and $r_j(t;\delta,\gamma) = \frac{e_j(t)-\sum_{\substack{m=1\\m\neq\delta}}^{D_j} a_{s_j}(m;t)s_j(t-m)}{1+|\gamma|}$. For TF domain, the mixing model is obtained using the STFT of $x_j(t),\ j = 1,2$ as

$$\begin{aligned} &X_1(\tau,\omega) = S_1(\tau,\omega) + S_2(\tau,\omega) \\ &X_2(\tau,\omega) = a_1(\tau)e^{-i\omega\delta}S_1(\tau-\delta,\omega) + a_2(\tau)e^{-i\omega\delta}S_2(\tau-\delta,\omega) - \\ &\qquad \left(\sum_{\substack{m=1\\m\neq\delta}}^{D_1} \frac{a_{s_1}(m;\tau)}{1+|\gamma|} e^{-i\omega m}S_1(\tau-m,\omega) + \sum_{\substack{m=1\\m\neq\delta}}^{D_2} \frac{a_{s_2}(m;\tau)}{1+|\gamma|} e^{-i\omega m}S_2(\tau-m,\omega)\right) \end{aligned} \tag{4.3}$$

for $\forall\tau,\omega$. The proposed pseudo-stereo mixture is based on the assumptions which were previously stated in Chapter 3 Section 3.2.1 and are summarized as follows: **Assumption 1**: The source signals satisfy the local stationarity of the time-frequency representation; $S_j(\tau - \phi, \omega) \approx S_j(\tau, \omega)$. ***Assumption 2***: The source signals satisfy the windowed-disjoint orthogonality (WDO) condition where different signals are approximately orthogonal to each other i.e. $S_i(\tau, \omega) S_j(\tau, \omega) \approx 0$, $\forall i \neq j$, $\forall \tau, \omega$. **Assumption 3**: phase ambiguity. In order to avoid phase ambiguity, the chosen $\delta$ must satisfies the condition: $\delta_{max} < \frac{f_s}{2 f_{max}}$.

## 4.2 Proposed Separation Method

In this section, a new framework for solving the SCBSS problem is presented by using the pseudo – stereo mixing model. The core concept of developing a separating process is to construct a binary TF mask. The binary mask is constructed by evaluating a proposed cost function given by the estimators of the AR coefficients of the signals. To achieve this, the additional assumption on the source signals is imposed:

**Assumption 4**: The source signals are modelled as quasi-stationary. This refers to the condition where the autoregressive (AR) parameters in AR process i.e. $s_j(t) = -\sum_{m=1}^{D_j} a_{s_j}(m; t) s_j(t-m) + e_j(t)$ are stationary within a block but can change from block to block. Specifically, $s_j(t)$ is partitioned into $L$ contiguous blocks where block $l$ begins at time $t_l$ with length $B_l = t_{l+1} - t_l$, and in this block the AR parameters $a_{s_j}(m; t) = a_{s_j}(m; T_l)$ for $\forall t \in T_l = \{t_l, \dots, t_{l+1} - 1\}$ such that

$$s_j(t) = -\sum_{m=1}^{D_j} a_{s_j}(m; T_l) s_j(t-m) + e_j(t) \quad , \quad \forall t \in T_l \tag{4.4}$$

Stationary AR signals are special case of the above where the AR parameters do not vary with time [47] and this is equivalent to setting $L = 1$ in (4.4).

### *4.2.1 Parameter Estimation using Complex 2D Histogram*

To begin, the $j^{th}$ source signal is assumed to be dominant at a particular TF unit:

$$\begin{aligned} X_1(\tau,\omega) &= S_j(\tau,\omega) \\ X_2(\tau,\omega) &= a_j(\tau)e^{-i\omega\delta}S_j(\tau-\delta,\omega) - \sum_{\substack{m=1\\m\neq\delta}}^{D_j}\frac{a_{s_j}(m;\tau)}{1+|\gamma|}e^{-i\omega m}\,S_j(\tau-m,\omega) \\ &\approx \left[a_j(\tau)-C_j(\tau,\omega)\right]e^{-i\omega\delta}S_j(\tau,\omega) \quad , \quad (\tau,\omega)\in\Omega_j \end{aligned} \tag{4.5}$$

for $\delta$ and $m \leq \phi$, $C_j(\tau,\omega)=\frac{1}{1+|\gamma|}\sum_{\substack{m=1\\m\neq\delta}}^{D_j} a_{s_j}(m;\tau)e^{-i\omega(m-\delta)}$ and $\Omega_j$ is the active area of $S_j(\tau,\omega)$ defined as

$$\Omega_j := \left\{(\tau,\omega): S_j(\tau,\omega)\neq 0,\ \ \forall k\neq j\right\} \tag{4.6}$$

The estimate of $\bar{a}_j(\tau,\omega)=a_j(\tau)-C_j(\tau,\omega)$ associated with the $j^{th}$ signal can be determined as

$$\begin{aligned} \bar{a}_j(\tau,\omega) &= \frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}e^{i\omega\delta} \\ &= a_j(\tau)-C_j(\tau,\omega) \\ &= \bar{a}_j^{(r)}(\tau,\omega)+i\bar{a}_j^{(i)}(\tau,\omega)\ \ ,\quad \forall(\tau,\omega)\in\Omega_j \end{aligned} \tag{4.7}$$

where

$$\bar{a}_j^{(r)}(\tau,\omega)=Re\left[\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}e^{i\omega\delta}\right],\ \bar{a}_j^{(i)}(\tau,\omega)=Im\left[\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}e^{i\omega\delta}\right]$$

are the real and imaginary parts of $\bar{a}_j(\tau,\omega)$, respectively, and $i=\sqrt{-1}$. Although the ratio $X_2(\tau,\omega)/X_1(\tau,\omega)$ seems straightforward, it is difficult to obtain $\bar{a}_j(\tau,\omega)$ directly from this ratio because the term $C_j(\tau,\omega)$ varies with frequency from frame to frame. In the WDO case which one signal is active at each TF unit, a TF plane of $\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}e^{i\omega\delta}$ is labelled by the active $j^{th}$ signal for each $(\tau,\omega)$ units. Thus, the TF plane can be partitioned into $N$ groups (where $N$ is the total number of signals in the mixture) where each group contains the $(\tau, \omega)$ units with identical label. As such, the $N$ groups can be clustered by creating the weighted complex 2-dimensional (2D) histogram. By using $\bar{a}_j^{(r)}(\tau, \omega)$ and $\bar{a}_j^{(i)}(\tau, \omega)$ pairs to indicate the indices into the histogram and using $|X_1(\tau, \omega)X_2(\tau, \omega)|$ for the weight, the cluster of weight will emerge centered on the actual mixing parameter pairs. Therefore, the weighted complex 2D hisgram is employed to determine $\hat{\hat{a}}_j^{(r)}$ and $\hat{\hat{a}}_j^{(i)}$ via identifying peaks in the histogram. The weighted complex 2D histogram estimation method is proposed as a function of $(\tau, \omega)$ with the weight $\sum_{\tau,\omega}|X_1(\tau, \omega)X_2(\tau, \omega)|$ to estimate $\bar{a}_j(\tau, \omega)$ and cluster them into $N$ groups. In particular, the real and imaginary parts of $\bar{a}_j(\tau, \omega)$ can be estimated as

$$\hat{\hat{a}}_j^{(r)} = \frac{\sum_{\tau,\omega\in\Omega_j}|X_1(\tau,\omega)X_2(\tau,\omega)|Re\left[\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}e^{i\omega\delta}\right]}{\sum_{\tau,\omega\in\Omega_j}|X_1(\tau,\omega)X_2(\tau,\omega)|}$$

$$\hat{\hat{a}}_j^{(i)} = \frac{\sum_{\tau,\omega\in\Omega_j}|X_1(\tau,\omega)X_2(\tau,\omega)|Im\left[\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}e^{i\omega\delta}\right]}{\sum_{\tau,\omega\in\Omega_j}|X_1(\tau,\omega)X_2(\tau,\omega)|} \tag{4.8}$$

where $\Omega_j$ is the active area of the $j^{th}$ signal. Eq.(4.8) can then be combined to form the estimate of (4.7) as

$$\hat{\hat{a}}_j = \hat{\hat{a}}_j^{(r)} + i\hat{\hat{a}}_j^{(i)} \tag{4.9}$$

Relating (4.9) with (4.7) based on the similar idea, (4.9) can then be expressed as $\hat{\hat{a}}_j = \hat{a}_j - \hat{C}_j$ where $\hat{a}_j$ and $\hat{C}_j$ are the complex 2D histogram estimates of $a_j(\tau)$ and $C_j(\tau, \omega)$, respectively

### *4.2.2 Construction of Masks*

In this section, the binary TF masks will be established by using $X_1(\tau,\omega)$ alone. The binary TF masks can be constructed by labelling each TF unit with the $k$ argument through maximizing the Gaussian likelihood function. The full detail of the propsed cost function was presented in Chapter 3 Section 3.3 which is recapped here as the following. The Gaussian likelihood function $L_j(\tau,\omega)$ given by $p(X_1(\tau,\omega),X_2(\tau,\omega)|S_j(\tau,\omega),\hat{\hat{a}}_j,\delta)$ can be expressed as:

$$L_j(\tau,\omega) = -\sum_{(\tau,\omega)\in\Omega_j} \frac{\left|X_1(\tau,\omega)-S_j(\tau,\omega)\right|^2}{\sigma_1^2(\tau,\omega)} + \frac{\left|X_2(\tau,\omega)-\hat{\hat{a}}_j(\tau,\omega)e^{i\omega\delta}S_j(\tau,\omega)\right|^2}{\sigma_2^2(\tau,\omega)}. \quad (4.10)$$

To maximize (4.10) with respect to $S_j(\tau,\omega)$, the ML of $S_j(\tau,\omega)$ is obtained by solving $\partial L(\tau,\omega)/\partial S_j(\tau,\omega) = 0$ for $\forall(\tau,\omega) \in \Omega_j$. The ML of the $j^{th}$ signal $S_j^{ML}(\tau,\omega)$ then obtains as:

$$S_j^{ML}(\tau,\omega) = \frac{X_1(\tau,\omega)+\hat{\hat{a}}_j e^{i\omega\delta}X_2(\tau,\omega)}{1+\hat{\hat{a}}_j^2(\tau,\omega)} \qquad ,(\tau,\omega) \in \Omega_j \qquad (4.11)$$

The Gaussian ML function of the $j^{th}$ signal is then created by substitiding $S_j^{ML}(\tau,\omega)$ into $S_j(\tau,\omega)$ in (4.10):

$$F_j(\tau,\omega) = -\sum_{(\tau,\omega)\in\Omega_j} \frac{\left|\hat{\hat{a}}_j(\tau,\omega)e^{-i\omega\delta}X_1(\tau,\omega)-X_2(\tau,\omega)\right|^2}{1+\hat{\hat{a}}_j^2(\tau,\omega)} \qquad (4.12)$$

This process is equivalent to minimizing the following:

$$G(\tau,\omega) = \underset{k}{\operatorname{argmin}} \left|\hat{\hat{a}}_k e^{-i\omega\delta}X_1(\tau,\omega) - X_2(\tau,\omega)\right|^2 \qquad (4.13)$$

Using the proposed pseudo-stereo mixture, the third term of $X_2(\tau,\omega)$ in (4.5) can be expressed as:

$$\begin{aligned}
\sum_{\substack{m=1\\m\neq\delta}}^{D_j} \frac{a_{S_j}(m;\tau)}{1+|\gamma|} e^{-i\omega m} S_j(\tau-m,\omega) &= \frac{1}{1+|\gamma|} \sum_{\substack{m=1\\m\neq\delta}}^{D_j} a_{S_j}(m;\tau)\, e^{-i\omega m} S_j(\tau-m,\omega) \\
&= \frac{1}{1+|\gamma|}\left(-S_j(\tau,\omega) + E_j(\tau,\omega) - a_{S_j}(\delta;\tau) e^{-i\omega\delta} S_j(\tau-\delta,\omega)\right) \\
&\approx -\frac{1}{1+|\gamma|}\left(1 + a_{S_j}(\delta;\tau)e^{-i\omega\delta}\right) S_j(\tau,\omega) + \frac{E_j(\tau,\omega)}{1+|\gamma|} \\
&= -\frac{1}{1+|\gamma|}\left(1 + (\gamma - a_j(\tau)(1+|\gamma|))e^{-i\omega\delta}\right) S_j(\tau,\omega) + \frac{E_j(\tau,\omega)}{1+|\gamma|} \\
&= -\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right) S_j(\tau,\omega) + a_j(\tau) e^{-i\omega\delta} S_j(\tau,\omega) + \frac{E_j(\tau,\omega)}{1+|\gamma|} \qquad (4.14)
\end{aligned}$$

for $\delta \leq \phi$ and by invoking the local stationarity at the second line of (4.14). Eq.(4.14) can then be rearragened to express in terms of the mixtures as

$$\begin{aligned}
&\sum_{\substack{m=1\\m\neq\delta}}^{D_j} \frac{a_{S_j}(m;\tau)}{1+|\gamma|} e^{-i\omega m} S_j(\tau-m,\omega) = -\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right) S_j(\tau,\omega) + a_j(\tau)e^{-i\omega\delta} S_j(\tau,\omega) + \frac{E_j(\tau,\omega)}{1+|\gamma|} \\
&-a_j(\tau)e^{-i\omega\delta} S_j(\tau,\omega) + \sum_{\substack{m=1\\m\neq\delta}}^{D_j} \frac{a_{S_j}(m;\tau)}{1+|\gamma|} e^{-i\omega m} S_j(\tau-m,\omega) - \frac{E_j(\tau,\omega)}{1+|\gamma|} = -\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right) S_j(\tau,\omega) \\
&a_j(\tau)e^{-i\omega\delta} S_j(\tau,\omega) - \sum_{\substack{m=1\\m\neq\delta}}^{D_j} \frac{a_{S_j}(m;\tau)}{1+|\gamma|} e^{-i\omega m} S_j(\tau-m,\omega) + \frac{E_j(\tau,\omega)}{1+|\gamma|} = \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right) S_j(\tau,\omega) \quad (4.15)
\end{aligned}$$

Substituding $X_1(\tau,\omega)$ and $X_2(\tau,\omega)$ in (4.15), Eq.(4.15) then becomes:

$$\begin{aligned}
X_2(\tau,\omega) + \frac{E_j(\tau,\omega)}{1+|\gamma|} &= \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right) S_j(\tau,\omega) \\
X_2(\tau,\omega) &\approx \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right) X_1(\tau,\omega) - \frac{E_j(\tau,\omega)}{1+|\gamma|} \qquad (4.16)
\end{aligned}$$

In this light, the proposed cost function can be formulated based on the single mixture $X_1(\tau,\omega)$ by substituting this relation into the function $G(\tau,\omega)$ in (4.13) which leads to

$$J(\tau,\omega) = \underset{k}{argmin}\, H_k(\tau,\omega) \qquad (4.17)$$

where

$$H_k(\tau,\omega) = \left|\hat{\hat{a}}_k e^{-i\omega\delta} X_1(\tau,\omega) - \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right) X_1(\tau,\omega)\right|^2 \qquad (4.18)$$

Since $e_j(t) \ll s_j(t)$, the term $E_j(\tau,\omega)/(1+|\gamma|)$ is negligible. Hence, $X_2(\tau,\omega) = \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right) X_1(\tau,\omega)$. Below, how the above cost function works is elucidated. First, it is assumed that the $j^{th}$ signal is dominant at $(\tau,\omega) \in \Omega_j$ and then consider the case when $k = j$:

$$
\begin{aligned}
H_{k=j}(\tau,\omega) &= \left|\hat{\hat{a}}_j e^{-i\omega\delta} S_j(\tau,\omega) - \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right) S_j(\tau,\omega)\right|^2 \\
&= \left|\hat{a}_j e^{-i\omega\delta} S_j(\tau,\omega) - \hat{C}_j e^{-i\omega\delta} S_j(\tau,\omega) - a_j(\tau) e^{-i\omega\delta} S_j(\tau,\omega) + \sum_{\substack{m=1 \\ m\neq\delta}}^{D_j} \left(\frac{a_{S_j}(m;\tau) e^{-i\omega m}}{1+|\gamma|}\right) S_j(\tau-m,\omega)\right|^2 \\
&= \left|-\hat{C}_j e^{-i\omega\delta} S_j(\tau,\omega) + \sum_{m=1}^{D_j} \frac{a_{S_j}(m;\tau) e^{-i\omega m}}{1+|\gamma|} S_j(\tau-m,\omega) - \frac{a_{S_j}(\delta;\tau) e^{-i\omega\delta}}{1+|\gamma|} S_j(\tau-\delta,\omega)\right|^2 \\
&= \left|-\hat{C}_j e^{-i\omega\delta} S_j(\tau,\omega) - \frac{S_j(\tau,\omega)}{1+|\gamma|} - \frac{a_{S_j}(\delta;\tau) e^{-i\omega\delta}}{1+|\gamma|} S_j(\tau-\delta,\omega)\right|^2 \\
&= \left|\left(\hat{C}_j + \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right) e^{-i\omega\delta} + \frac{1}{1+|\gamma|}\right|^2 \left|S_j(\tau,\omega)\right|^2 \qquad (4.19)
\end{aligned}
$$

When $k \neq j$, following the above step leads to

$$
H_{k\neq j}(\tau,\omega) = \left|\left(\hat{\hat{a}}_k - a_j(\tau) - \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right) e^{-i\omega\delta} - \frac{1}{1+|\gamma|}\right|^2 \left|S_j(\tau,\omega)\right|^2 \qquad (4.20)
$$

Using (4.19) and (4.20), when the $j^{th}$ signal dominates at $(\tau,\omega) \in \Omega_j$ the cost function will correctly identify the signal if and only if $H_{k=j}(\tau,\omega) < H_{k\neq j}(\tau,\omega)$. This therefore stipulates a condition for $\hat{C}_j$ to ensure that $H_{k=j}(\tau,\omega) < H_{k\neq j}(\tau,\omega)$ is always satisfied. Starting with (4.19) and (4.20), the above condition can be expressed as

$$
\left|\left(\hat{C}_j + \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right) e^{-i\omega\delta} + \frac{1}{1+|\gamma|}\right|^2 < \left|\left(\hat{\hat{a}}_k - a_j(\tau) - \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right) e^{-i\omega\delta} - \frac{1}{1+|\gamma|}\right|^2 \qquad (4.21)
$$

Let $\beta_j = \hat{C}_j + \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}$ and $\beta_l = \hat{\hat{a}}_k - a_j(\tau) - \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}$, then the above becomes

$$
\left|\beta_j e^{-i\omega\delta} + \frac{1}{1+|\gamma|}\right|^2 < \left|\beta_l e^{-i\omega\delta} - \frac{1}{1+|\gamma|}\right|^2 \qquad (4.22)
$$

The left hand side of the above (4.22) is bounded below by

$$\begin{aligned} \left|\beta_j e^{-i\omega\delta} + \frac{1}{1+|\gamma|}\right| &\geq |\beta_j| - \frac{1}{1+|\gamma|} \\ &= \left|\hat{C}_j + \frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right| - \frac{1}{1+|\gamma|} \\ &\geq |\hat{C}_j| - \left|\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right| - \frac{1}{1+|\gamma|} \end{aligned} \tag{4.23}$$

and the right hand side of (4.21) is bounded above by

$$\begin{aligned} \left|\beta_l e^{-i\omega\delta} - \frac{1}{1+|\gamma|}\right| &\leq |\beta_l e^{-i\omega\delta}| + \frac{1}{1+|\gamma|} \\ &= |\beta_l| + \frac{1}{1+|\gamma|} \\ &= \left|\hat{\hat{a}}_k - a_j - \frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right| + \frac{1}{1+|\gamma|} \end{aligned} \tag{4.24}$$

Substituting (4.23) and (4.24) into (4.21) and re-plugging the terms for $\beta_j$ and $\beta_l$, (4.21) results in

$$|\hat{C}_j| < \left|\hat{\hat{a}}_k - a_j(\tau) - \frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right| + \frac{2}{1+|\gamma|} + \left|\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right| \tag{4.25}$$

for $\forall j \neq k$. The proposed cost function (4.17)-(4.18) will correctly assign the $(\tau, \omega)$ unit to the $j^{th}$ signal if the $|\hat{C}_j|$ condition in (4.25) is satisfied across $\Omega_j$. Conversely, if $|\hat{C}_j|$ is larger than the right-hand side of (4.25) then this will lead to wrong assignment of the TF units. Once the cost function is evaluated, the binary TF mask for the $j^{th}$ signal can be constructed as

$$M_j(\tau,\omega) := \begin{cases} 1 & J(\tau,\omega) = j \\ 0 & otherwise \end{cases}. \tag{4.26}$$

The proposed method is termed as Single Observation Likelihood estimatiOn (SOLO) algorithm. The proposed SOLO algorithm is summarized in Table 4.1.

Table 4.1: Pseudo code of SOLO algorithm

```
1. Formulate the pseudo-stereo mixture $x_2(t) = \frac{x_1(t) + \gamma x_1(t-\delta)}{1+|\gamma|}$ with an appropriate $\gamma$ and $\delta$
2. Transform the mixtures into TF domain by using STFT.
3. Generate the weighted complex 2D histogram in terms of $\left(\hat{\hat{a}}_j^{(r)}, \hat{\hat{a}}_j^{(i)}\right)$ according to (4.8) and identify $N$ peaks as the estimated $\hat{\hat{a}}_j$.
4. Formulate the binary TF mask $M_j(\tau, \omega)$ for each pair of $\left(\hat{\hat{a}}_j^{(r)}, \hat{\hat{a}}_j^{(i)}\right)$ using (4.17), (4.18) and (4.26).
5. Separate the observed mixture using
       $\tilde{S}_j(\tau, \omega) = M_j(\tau, \omega) X_1(\tau, \omega).$    (4.27)
6. Convert the estimated signals from TF domain into time domain.
```

### 4.3 Results and Analysis

The performance of SOLO is demonstrated by separating stationary and nonstationary signals. The stationary signals are syntheticed by using the AR process. The chirp signals[2] and real-audio signals were used for the nonstationary signals. The real-audio signals which are inherently non-stationary included voice and music signals. All experiments were conducted under the same conditions as follows: The signals were mixed with normalized power over the duration of the signals. All mixed signals were sampled at 16 kHz sampling rate. The TF representation was computed by using the STFT of

[2] Chirp signal is classified as non-stationary due its time-varying instantaneous frequency.

1024-point Hamming window with 50% overlap. The separation performance was evaluated by measuring the distortion between original signal and the estimated one according to the signal-to-distortion (SDR) ratio and signal-to-interference (SIR) ratio defined as SDR $= 10\, log_{10}\left(\|s_{target}\|^2/\|e_{interf} + e_{artif}\|^2\right)$ and SIR $= 10\, log_{10}\left(\|s_{target}\|^2/\|e_{interf}\|^2\right)$ where $e_{interf}$ represent the interference from other signals and $e_{artif}$ is the artifacts. The proposed approach will be compared with the sparse nonnegative matrix 2-dimensional factorization (SNMF2D) [48], the single-channel independent component analysis (SCICA) [49] and the ideal binary mask (IBM) [50] which represents the ideal separation performance. The SNMF2D parameters are set as follows [51-52]: number of factors was 2, sparsity weight of 1.1, number of phase shift and time shift is 31 and 7, respectively for music. As for speech, both shifts are set to 4. The TF domain used in SNMF2D is based on the log-frequency spectrogram. Cost function of SNMF2D is based on the Kullback-Leibler divergence. As for the SCICA, the number of block is 10 with time delay set to unity. MATLAB is used as the programming platform. All simulations and analyses are performed using a PC with Intel Core 2 CPU 3GHz and 3GB RAM.

### *4.3.1 Stationary Sources*

#### 4.3.1.1 Two Synthetic AR Signals

Two stationary AR signals are synthesized for $s_1(t)$ and $s_2(t)$ using the model (3) with the following the coefficients: $a_{s_1} = [-3.7281, 5.3956, -3.5805, 0.9224]$ and $a_{s_2} = [-0.6070,\ 1.9739, -0.5711, 0.8853]$ and $e_1(t)$ and $e_2(t)$ are zero mean white Gaussian signal with average variances of $6.2 \times 10^{-6}$ and $7.6 \times 10^{-4}$, respectively.

The coefficients and the variances are randomly selected. It should be noted that $a_{s_1}(0) = a_{s_2}(0) = 1$ by definition but this has not been included in the above to avoid cluttering the notation. The source signals are shown in Fig.4.2. The pseudo-stereo parameters are selected to be $\gamma = 4$ and $\delta = 2$. The histogram-resolution parameters are set at $\Delta_{\alpha^{(r)}} = 5$, $\Delta_{\alpha^{(i)}} = 50$, $\zeta^{(r)} = 101$ and $\zeta^{(i)} = 3$ where $\Delta_{\alpha^{(r)}}$ and $\Delta_{\alpha^{(i)}}$ are the maximum value of $\alpha^{(r)}$ and $\alpha^{(i)}$, respectively. The term $\zeta^{(r)}$ and $\zeta^{(i)}$ are the number of bins for $\alpha^{(r)}$ and $\alpha^{(i)}$.

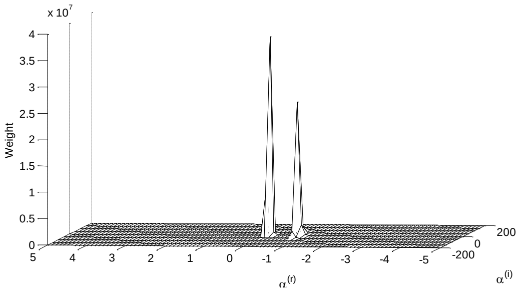


Figure 4.1: A complex 2D histogram corresponding to two signals.

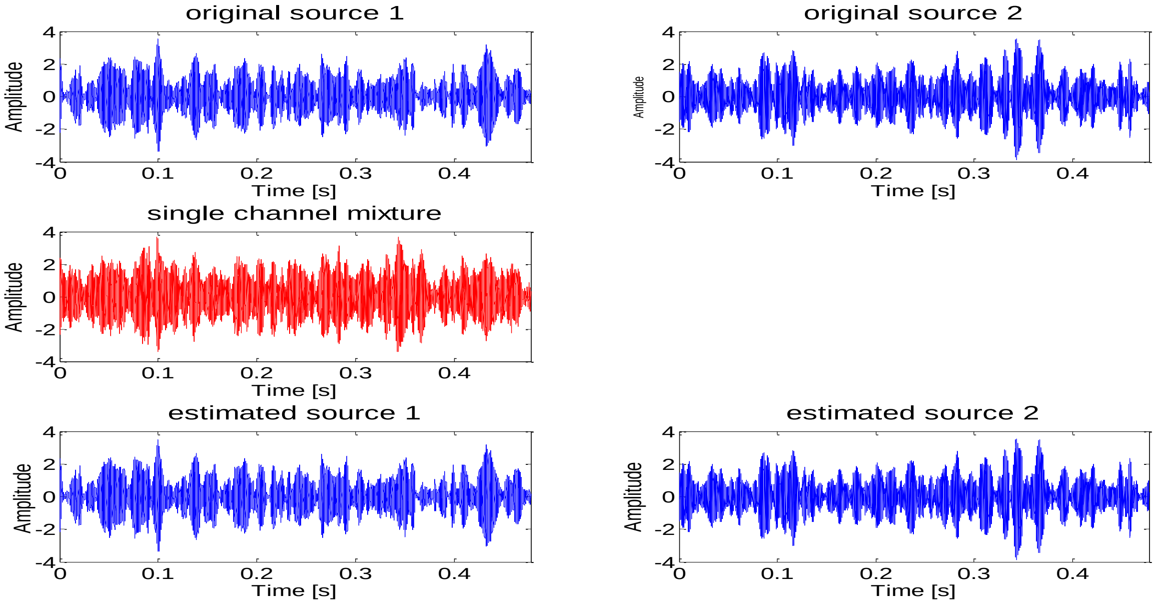


Figure 4.2: Two original signals, observed mixture and two estimated signals.

Fig.4.1 illustrates the clustering of the signals into two peaks which associate with the number of signals in the mixture. Fig.4.2 also shows the mixed signal and the separated signals based on the SOLO method. Visually, it can be seen that the mixture has been very well separated as compared with the original signals. The separation performance is tabulated in Table 4.2 which shows the comparison results of SNMF2D, SCICA, proposed SOLO and IBM. The SDR and SIR results of each method are calculated from the average of 100 experiments under the same mixture. The proposed SOLO method successfully estimated the signals with a high accuracy. In particular, the SOLO method renders an average SDR improvement of $13.4$dB per signal over the SNMF2D and $14.6$dB per signal over the SCICA and an average SIR improvement of $56.1$dB per signal and $53$dB per signal over the SNMF2D and SCICA, respectively.

Table 4.2: Comparison of average SDR and SIR performance on mixture of two AR signals with SNMF2D, SCICA, SOLO and IBM

| Methods | SDR $s_1$ | SDR $s_2$ | SIR $s_1$ | SIR $s_2$ |
|---|---|---|---|---|
| SNMF2D | 7.2 | 5.1 | 17 | 6.8 |
| SCICA | 4.8 | 5.1 | 13.2 | 16.8 |
| SOLO | 19 | 20.1 | 67.8 | 68.2 |
| IBM | 19 | 20.2 | 68.7 | 74 |

Due to the stationarity of the signals, the AR coefficients do not change with $\tau$ and thus $H_{k=j}(\tau,\omega) < H_{k\neq j}(\tau,\omega)$ can be satisfied only when $\left|\hat{C}_j\right| < \left|\hat{\hat{a}}_k - a_j - \frac{a_{s_j}(\delta)}{1+|\gamma|}\right| + \frac{2}{1+|\gamma|} + \left|\frac{a_{s_j}(\delta)}{1+|\gamma|}\right|$ according to (28). For $j = 1$, the term $\left|\hat{C}_1\right| = 1.43$ and $|\hat{\hat{a}}_2 - a_1 - \frac{a_{s_1}(2)}{6}\Big| + \frac{2}{6} + \left|\frac{a_{s_1}(2)}{6}\right| = 2.24$ have been computed in which case it has $1.43 < 2.24$ thus the $\left|\hat{C}_1\right|$ condition is satisfied. For $j = 2$, the term $\left|\hat{C}_2\right| = 1.01$ and $\Big|\hat{\hat{a}}_1 - a_2 -$

$\frac{a_{s_2}(2)}{6}\Big| + \frac{2}{6} + \Big|\frac{a_{s_2}(2)}{6}\Big| = 1.67$ have been computued and therefore $1.01 < 1.67$. Thus, the $|\hat{C}_2|$ condition is also true. Hence the cost function will be able to correctly label all $(\tau, \omega)$ units to their respective original signals. This is clearly evident by the same SDR results between the SOLO and the IBM.

### 4.3.1.2 Separation of more than 2 Synthetic AR Signals

In this evaluation, the proposed method was tested by increasing the number of signals from $j = 2, 3, 4, 5$. Each mixture of 2 to 5 signals is executed 100 times. Five stationary AR signals are synthesized using the model (3) with the following the coefficients:

$$a_{s_1} = [-3.8604, 5.6466, -3.7076, 0.9224]$$

$$a_{s_2} = [-2.6189, 3.5578, -2.4136, 0.8493]$$

$$a_{s_3} = [0.8773, 2.0937, 0.8340, 0.9037]$$

$$a_{s_4} = [2.9132, 3.9841, 2.7128, 0.8672]$$

$$a_{s_5} = [3.8148, 5.5394, 3.6264, 0.9037]$$

and $e_1(t)$ to $e_5(t)$ are zero mean white Gaussian signals with variances $2.16 \times 10^{-7}$, $5.27 \times 10^{-4}$, $4.23 \times 10^{-4}$, $2.32 \times 10^{-4}$ and $8.54 \times 10^{-7}$ , respectively. The coefficients and the variances are randomly selected. All experiments are conducted under the same conditions: $\delta = 1$, $\Delta_{\alpha^{(r)}} = 5$, $\Delta_{\alpha^{(i)}} = 50$, $\zeta^{(r)} = 101$ and $\zeta^{(i)} = 3$.

The SDR performance of higher order mixtures has been tabulated in Table 4.3 and Fig. 4.3 shows the corresponding Box plot. It is noted that the separation performance progressively deteriorates as the number of signals increases. When the signals are not perfectly estimated and become slightly mutually correlated [53], the projection of these signals to the original signal subspace will not be zero and thus, they act as interference. In addition, the noise generated from the windowed-STFT and the excitation signals contribute to the artifacts. Thus, as the number of estimated signals increases, this has inadvertently led to larger values of $e_{interf}$ and $e_{artif}$ , and subsequently decreased the SDR and SIR performance. This explains the result for 5 signals which shows a drop in performance. Although this is the case, the SDR and SIR results are still maintained at a high level. The complex 2D histogram, shown in Fig.4.4, distinctively enumerates five peaks which correspond to the number of signals in the mixture. Figs.4.5 and 4.6 show the original signals, the mixture and the separated signals. One can visually inspect that the separated signals are very similar to the original signals. In this experiment, the signals satisfy the assumptions and the mixing model holds the condition $a_i(t) \neq a_j(t)$ or $r_i(t) \neq r_j(t)$ . As such, the SOLO algorithm has successfully identified and partitioned the mixed signal TF plane into the correct group of signals.

Table 4.3: Average SDR and SIR results for mixture of 2 to 5 signals

| Mixture | $\gamma$ | SDR (dB) | SIR (dB) |
|---|---|---|---|
| $s_1 + s_3$ | 3 | 19.5 | 68 |
| $s_1 + s_2 + s_3$ | 2 | 19.5 | 64.4 |
| $s_1 + s_2 + s_3 + s_4$ | 3 | 19.1 | 61.1 |
| $s_1 + s_2 + s_3 + s_4 + s_5$ | 2 | 18.7 | 57.5 |

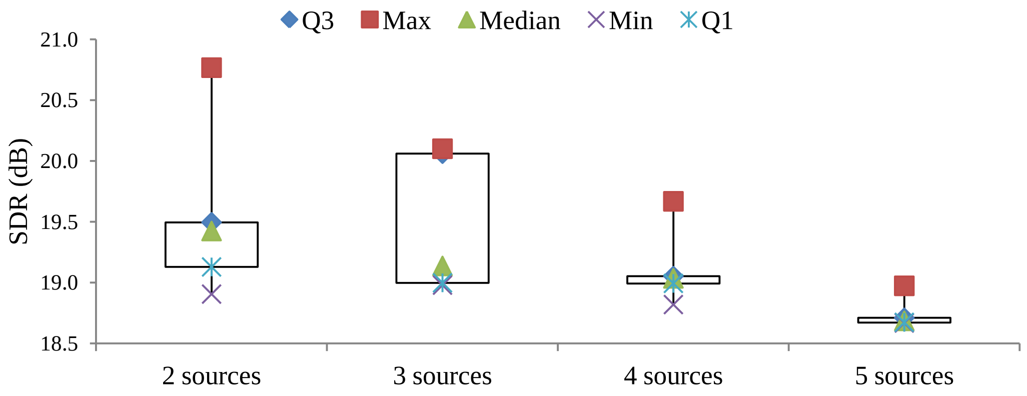


Figure 4.3: Box plot of average SDR results.

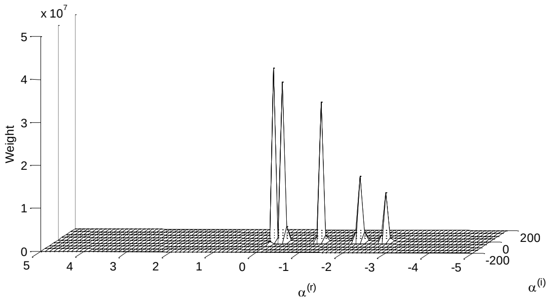


Figure 4.4: The complex 2D histogram of a mixture of five signals.

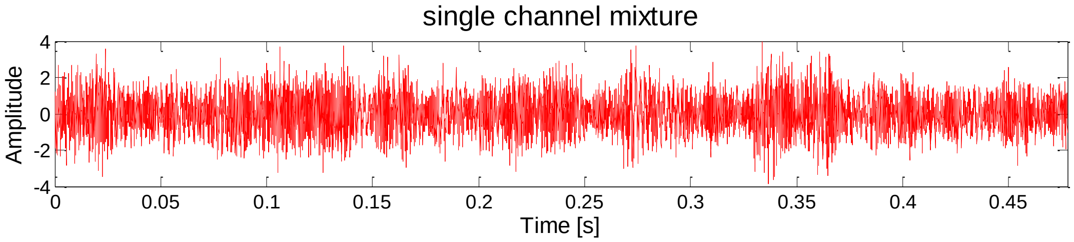


Figure 4.5: Single channel mixture.

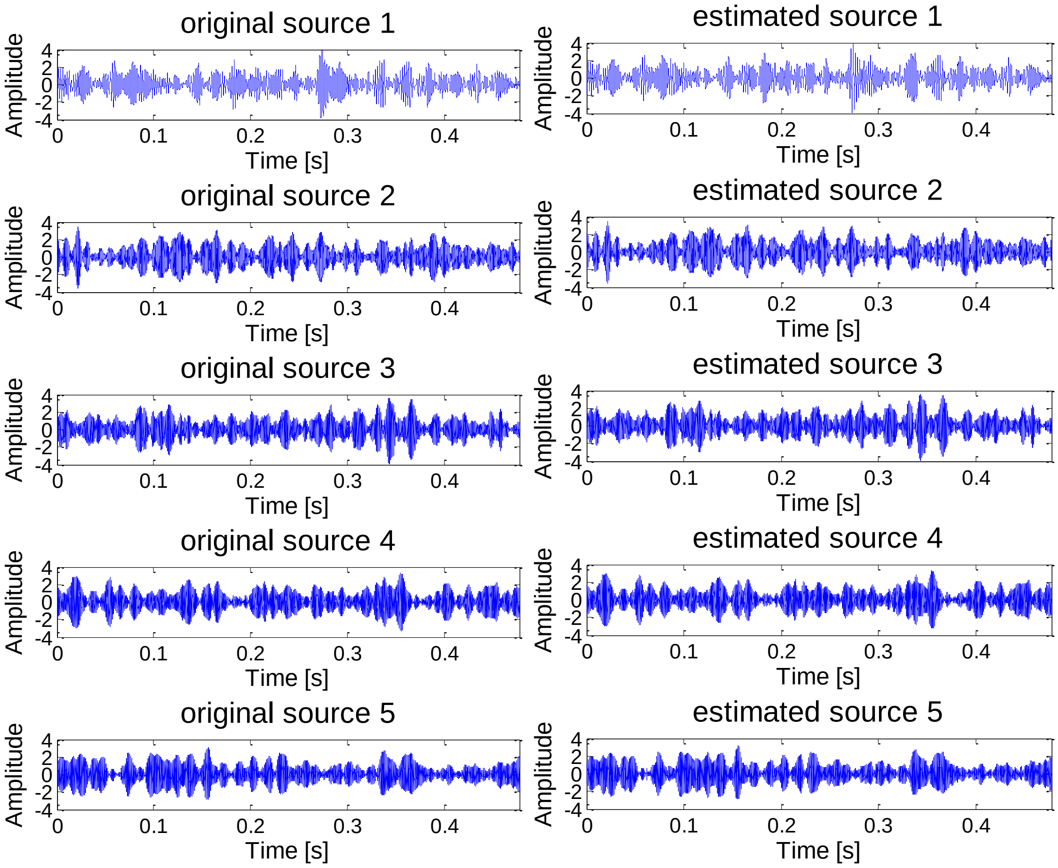


Figure 4.6: Original signals (left) and estimated signals (right).

### *4.3.2 Non-Stationary Source*

Since the proposed method estimates the parameter $\hat{\hat{a}}_j$ from the complex 2D histogram, its result is based on the averaged AR coefficient of each signal. As such, the estimated $\hat{\hat{a}}_j$ befits very well the purpose of separating stationary AR signals. In the case of non-stationary signals, this approach may readily be adapted and invoked the assumption of quasi-stationary. In effect this enables this approach to work under the condition where the AR parameters are stationary within a block but vary from block to block. The idea is then to partition the mixture signal $x_1(t)$ into arbitrary $L$ blocks and use the SOLO on each block to obtain $\hat{\hat{a}}_j = [\hat{\hat{a}}_{j1}\ \hat{\hat{a}}_{j2}\ \dots\ \hat{\hat{a}}_{jL}]$ where $\hat{\hat{a}}_{jl}$ is the estimate of $\hat{\hat{a}}_j$ from the $l^{th}$ block. A mask will subsequently be constructed in exactly the same manner in (29) but using the aggregated $\hat{\hat{a}}_{jl}$ obtained from each block.

#### 4.3.2.1 Chirp Signals

In this example, chirp signals are used to demonstrate the effectiveness of the SOLO method in dealing with non-stationary signals. $s_1$ is a “down-chirp” whose center frequency varies from $3.3 - 2$ kHz. $s_2$ is a quadratic-chirp signal whose center frequency varies from $0.5 - 1.8$ kHz. Both signals are mixed with equal average power over the duration of the signals. The single channel mixture is first divided into $L$ non-overlapping blocks and the parameters of the SOLO are selected to be $\delta = 2$, $\gamma = 3$ , $\Delta_{\alpha^{(r)}}= 5$ , $\Delta_{\alpha^{(i)}}= 50$ , $\zeta^{(r)} = 101$ and $\zeta^{(i)} = 4$ . Fig.4.7 shows the two synthesized chirp signals, the single channel mixture and the separated signals using the SOLO with $L = 5$. From the plots, it is visually evident that the mixture has been separated comparing with the original signals.

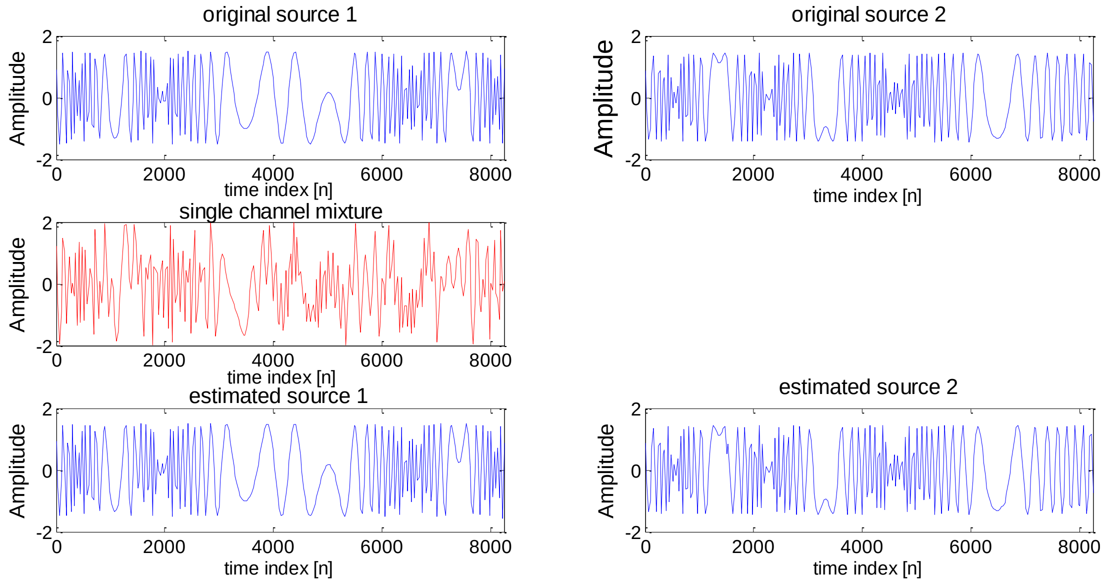


Figure 4.7: Original signals, single channel mixture, and estimated signals using SOLO with $L = 5$.

In Table 4.4, the comparison results have been tabultaed for SNMF2D, SCICA, SOLO with $L = 1,3,5$ and IBM. In general, the SOLO yields far superior separating results compared with the SNMF2D and the SCICA with an average SDR improvement of 9.0dB and 8.3dB per signal, and with an average SIR improvement of 15.3dB and 16.0dB, respectively. With the use of $\hat{\hat{a}}_j = [\hat{\hat{a}}_{j1}\ \hat{\hat{a}}_{j2}\ \ldots\ \hat{\hat{a}}_{jL}]$ partition, SOLO with $L > 1$ has led to substantially better separation performance than the SOLO with $L = 1$. It is clear from Table 4.4 that the average SDR and SIR performance increases by 4dB and 6dB per signal, respectively when $L = 5$.

Because the signals have time-varying instantaneous frequencies, the term $\sum_{\substack{m=1\\ m\neq\delta}}^{D_j}\frac{a_{s_j}(m;\tau)}{1+|\gamma|}e^{-i\omega m}\,S_j(\tau-m,\omega)$ in (7) will change accordingly with $\omega$ and $\tau$. Since $\bar{a}_j(\tau,\omega)$ composes $a_j(\tau)$ and $R_j(\tau,\omega)$, it follows that $\bar{a}_j(\tau,\omega)$ will also vary with $\omega$ and $\tau$. Unfortunately, setting $L = 1$ will mean that $\hat{\hat{a}}_j = \hat{\hat{a}}_{j1}$ which only estimates the global average of $\bar{a}_j(\tau,\omega)$ for all $(\tau,\omega)$. Thus, the obtained result of $\hat{\hat{a}}_{j1}$ can yield significant deviation from the true $\bar{a}_j(\tau,\omega)$. Therefore, the SDR and SIR performance of SOLO with $L = 1$ is not as high as in the previous case of stationary signals. On the other hand, when the mixture signal is divided into $L$ blocks such that each block resembles a mixture of frequency-invariant signals similar to the AR signals, then $\bar{a}_j(\tau,\omega)$ in each block can be treated as constant. As such, the cost function rendered by $\hat{\bar{a}}_j = [\hat{\bar{a}}_{j1}\ \hat{\bar{a}}_{j2}\ \dots\ \hat{\bar{a}}_{jL}]$ will enable all the TF units in each block to be specifically labeled using the estimated $\hat{\bar{a}}_{jl}$ derived from that block. As a result, better separation performance can be obtained as demonstrated in Table 4.4.

Table 4.4: Comparison of SDR and SIR performance on mixture of chirp signals with SNMF2D, SCICA, SOLO and IBM

| Methods | SDR $s_1$ | SDR $s_2$ | SIR $s_1$ | SIR $s_2$ |
|---|---|---|---|---|
| SNMF2D | 3.7 | 6.2 | 9.6 | 12.7 |
| SCICA | 5.1 | 6.3 | 10.1 | 10.8 |
| SOLO ($L = 1$) | 11.0 | 12.9 | 17.4 | 29.5 |
| SOLO ($L = 3$) | 13.4 | 14.6 | 22.1 | 30.8 |
| SOLO ($L = 5$) | 15.8 | 16.0 | 26.4 | 32.6 |
| IBM | 16.1 | 16.1 | 26.9 | 32.9 |

#### 4.3.2.2 Real - Audio Signals

Audio signals can be characterized as non-stationary AR processes since their AR coefficients vary with time. As an example, three type of mixtures were generated, these were: male speech + jazz, female speech + jazz, and male speech + piano. The male and female speeches are randonly selected from TIMIT and music signals from the RWC database. Both signals were mixed with equal power to generate the mixture. This is shown in the first three panels of Fig.4.8. To perform separation, the mixture was firstly divided into $L$ non-overlapping partitions. Two possible choices were available. The first choice was to partition the mixture into equal-length $L$ blocks. The separation performance was investigated by varying $L = 1, 3, 6, 9, 12, 15$. In all cases, the SOLO parameters are set to the followings: $\delta = 2$, $\gamma = 4$, $\Delta_{\alpha^{(r)}} = 2$, $\Delta_{\alpha^{(i)}} = 50$, $\zeta^{(r)} = 101$ and $\zeta^{(i)} = 4$.

The average SDR and SIR results are tabulated in Table 4.5 along with SNMF2D, SCICA and IBM. It is seen that in general the SOLO with increasing the number of blocks shows better separation performance than the SNMF2D and SCICA. From the table, it has also been noted that the performance remains high when using $L = 15$ where the average SDR and SIR results are 7.7dB per signal and 19.7dB per signal, respectively. When $L$ increases, each block becomes progressively narrower and contains less samples. The condition (28) may not be satisfied in some of these blocks particularly those of small amplitudes. In this case, the obtained mask may wrongly assign some of the TF units to the incorrect signal. As a result, the SIR value is slightly decreased. The proposed SOLO method renders an average SDR improvement of 1.2dB and 2.1dB per signal over SNMF2D and SCICA, respectively. Fig.4.9 shows the Box plot corresponding to the above results.

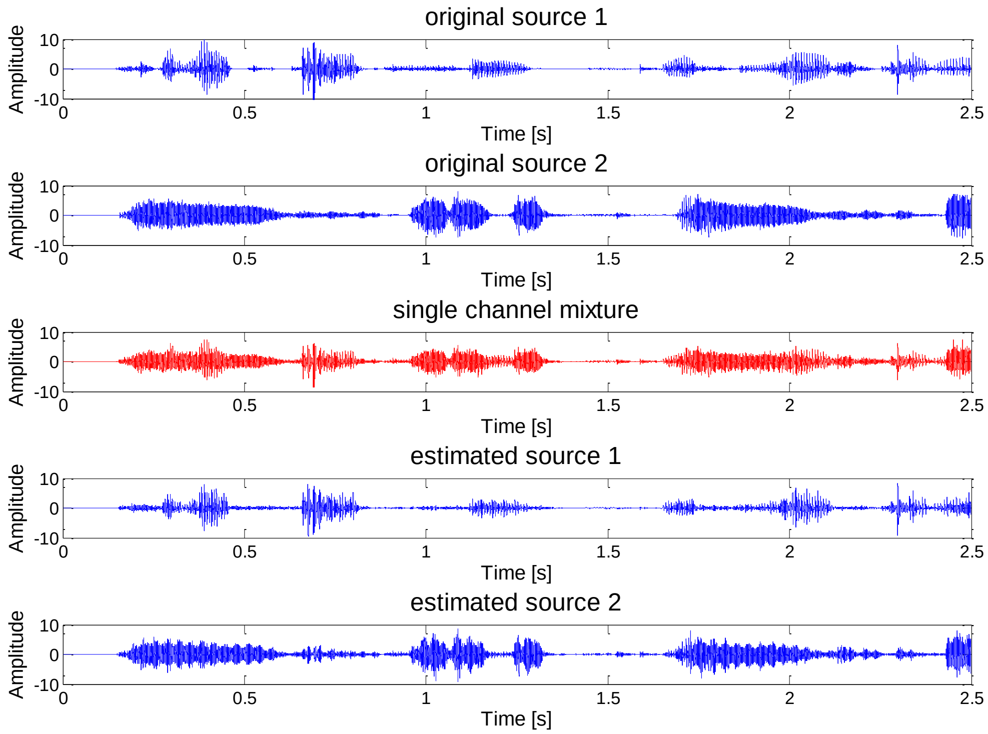


Figure 4.8: Original signals, single channel mixture, and estimated signals in time domain using the SOLO with $L = 6$ non-uniform blocks.

Table 4.5: Comparison of average SDR and SIR performance on mixture of two audio signals between SNMF2D, SCICA, SOLO and IBM

| Methods | SDR $\mathrm{s}_1$ | SDR $\mathrm{s}_2$ | SIR $\mathrm{s}_1$ | SIR $\mathrm{s}_2$ |
|---|---|---|---|---|
| SNMF2D | 7.5 | 5.5 | 10.3 | 7.3 |
| SCICA | 5.9 | 5.3 | 9.0 | 10.5 |
| SOLO ($L = 1$) | 5.8 | 6.9 | 12.5 | 19.7 |
| SOLO ($L = 3$) | 7.1 | 7.0 | 17.6 | 18.4 |
| SOLO ($L = 6$) | 7.3 | 7.0 | 17.6 | 18.7 |
| SOLO ($L = 9$) | 8.0 | 7.0 | 21.4 | 17.5 |
| SOLO ($L = 12$) | 8.0 | 7.0 | 20.9 | 17.9 |
| SOLO ($L = 15$) | 8.1 | 7.2 | 21.4 | 18.0 |
| IBM | 12.7 | 12.7 | 40 | 35.3 |

Note that $s_1$and $s_2$ refer to speech and music, respectively.

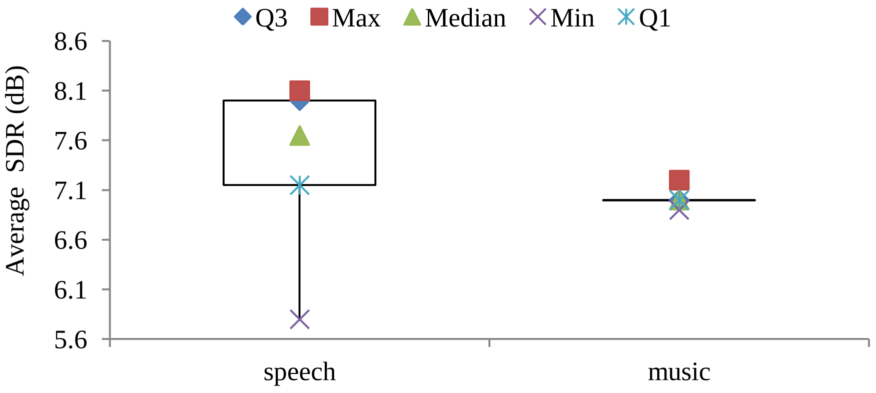


Figure 4.9: Box plot of average SDR results on mixture of two audio signals versus the number of blocks.

The second choice is to examine the characteristics and identify the transition behaviour in the mixture signal. In this case, the window size for each block is not required to be identical. Two examples have been considered here. In the first example, $L = 3$ has been set where it can be observed that the mixture of a male speech and Jazz music shows a transition at time $t = 0.85$s and in the interval around $t = 1.5$s. Thus, this enables the mixture to be partitioned into the following blocks i.e. $T_1 = [0\,, 0.85s]$, $T_2 = (0.85\text{s}\,, 1.5\text{s}]$, and $T_3 = (1.5\text{s}\,, 2.5\text{s}]$. In the second example, the mixture signal is partitioned into $L = 6$ blocks i.e. $T_1 = [0\,, 0.64\text{s}]$ , $T_2 = (0.64\text{s}\,, 0.86\text{s}]$ , $T_3 = (0.86\text{s}\,, 1.06\text{s}]$, $T_4 = (1.06\text{s}\,, 1.38\text{s}]$, $T_5 = (1.38\text{s}\,, 2.18\text{s}]$, and $T_6 = (2.18\text{s}\,, 2.5\text{s}]$. The SDR results are tabulated in Table 4.6. With $L = 3$ non-uniform blocks, the SDR performance gives 7.5dB per signal which matches the case of $L = 9$, and $L = 12$ equal-length blocks. On the other hand, with $L = 6$ non-uniform blocks the SDR performance gives 7.7dB per signal which matches the equal-length partition scheme of $L = 15$. The separated signals are plotted in the last panels of Fig.4.8. Visually, the separated signals resemble closely to the original signals. The IBM results have also been included for comparison purpose. Although all tested methods lag behind the IBM in terms of SDR performance, the proposed SOLO still yields good perceptual qualities of the separated signals.

Table 4.6: Comparison of SDR performance on mixture of two audio signals using SOLO with non-uniform length

| Methods | SDR $\mathrm{s}_1$ | SDR $\mathrm{s}_2$ | SIR $\mathrm{s}_1$ | SIR $\mathrm{s}_2$ |
|---|---|---|---|---|
| SOLO ($L = 3$ with non-uniform blocks) | 7.9 | 7.1 | 20.8 | 17.3 |
| SOLO ($L = 6$ with non-uniform blocks) | 8.1 | 7.3 | 21.7 | 17.5 |

Note that $s_1$ and $s_2$ refer to speech and music, respectively.

The computational complexity has also been calculated for SNMF2D, SCICA, and the proposed SOLO on a function of $N$ sample size of a signal ($N$), number of signals ($N_s$), length of the STFT window ($W_l$), number of frequency-shifts ($N_\emptyset$) and time-shift ($N_\tau$) for the SNMF2D, number of iterations for SNMF2D ($I_{SNMF2D}$) and SCICA ($I_{SCICA}$), and number of SCICA blocks ($K$). This is indicated in Table 4.7.

Table 4.7: Computation complexity of SNMF2D, SCICA, and SOLO

| Methods | Number of operations |
|---|---|
| SNMF2D | $2N\, log_2 W_l + I_{SNMF2D} N_s [3\tau \frac{W_l}{2} + 2N_\phi N_\tau N + N_\phi (4\frac{N}{W_l} + 2N_\tau N)$<br>$+2N_\phi N_\tau (N + \frac{W_l}{2} + N_\tau (N + \frac{W_l}{2} + N_s \frac{W_l}{2}))]$ |
| SCICA | $\big[2K(K+1)(N-K+1)KI_{SCICA} + K^3 + 2\big(K(N-K+1)\big) +$<br>$\big(K^2 + K(K-1)\big)(N-K+1)\big]N_s$ |
| SOLO | $5N + W_l + 4NN_s + 2N\, log_2 W_l$ |

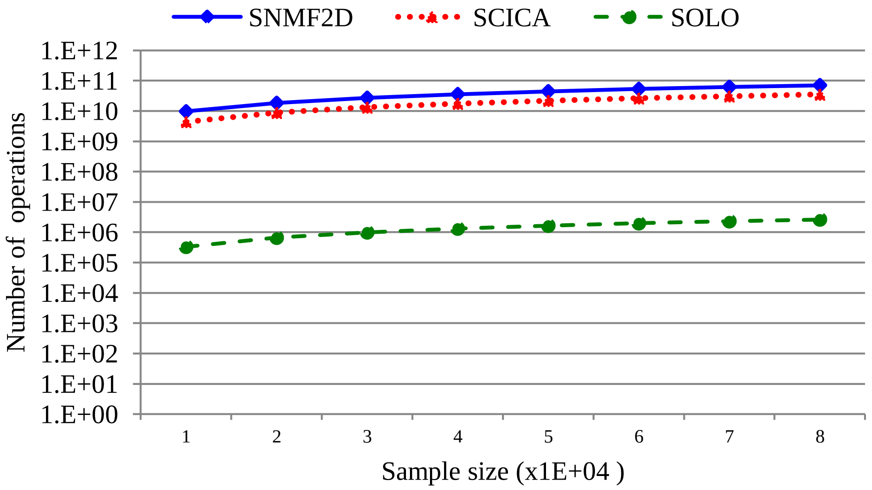


Figure 4.10: Comparison of computational complexity on mixture of two audio signals between SNMF2D, SCICA, and SOLO.

The computation complexity of the above algorithms has been plotted and this is shown in Fig.4.10 with the following parameters: $N_s = 2$, $W_l = 1024$, $N_\emptyset = 31$, $N_\tau = 7$, $I_{SNMF2D} = 100$, $I_{SCICA} = 100$, $K = 10$ and $N$ varies from $1 \times 10^4$ to $8 \times 10^4$. Note that: SOLO is computationally less demanding than SNMF2D and SCICA. The reason is SOLO does not require any iteration for updating parameters. On the other hand, SNMF2D requires updating the spectral basis and the mixing of the signals. As for SCICA, the computational complexity varies gradually with increasing sample size. This result is caused by three major reasons: 1) Complexity of the ICA algorithm within the SCICA grows exponentially with the number of blocks. 2) It requires deflation to remove the contribution of the extracted signal of interest. 3) The steps are repeated until all signals have been extracted. Fig.4.10 shows that the complexity of SCICA is almost identical to SNMF2D in the region of $10^{10}$ operations. Thus, the overall computational complexity associated with both algorithms is significantly high. On the other hand, the proposed SOLO consumes the least computation which renders it very fast and yet yields the best separation performance among the three methods.

## 4.4 Summary

This chapter has presented a novel single channel blind separation algorithm. The proposed method constructs a pseudo-stereo mixture by time-delaying and weighting the observed single channel mixture. The method assumes that the original signals are characterized as AR processes. Experiments have been conducted successfully to separate stationary as well as time-varying AR signals. In this work, the conditions required for a unique mask construction from the maximum likelihood framework have also been identified. The proposed method has demonstrated a high level separation performance for both synthetic and real-audio signals. The proposed method enjoys at least three advantages: Firstly, it does not require *a priori* knowledge of the signals. Secondly, the proposed approach is able to capture the music and speech characteristics and hence, renders robustness to the separation method. Finally, the proposed technique holds a desirable property — neither iterative optimization nor parameter initialization is required and this enables the separation process to be fast and executed in “one-go”.

# CHAPTER 5

# ONLINE NOISY SINGLE-CHANNEL ADAPTIVE BLIND SIGNAL SEPARATION USING SPECTRAL AMPLITUDE ESTIMATOR AND MASKING

In Chapter 4, the proposed SCBSS algorithms are derived for noise-free condition which lacks the potential and robustness to solve the problem in noisy environments since the presence of noise seriously degrades the performance. In a realistic scenario of audio applications, desired signals will be corrupted by an additive background noise. In this chapter, a novel framework to solving SCBSS in noisy environments is proposed. Overview of the proposed framework is illustrated in Fig.5.1. The proposed framework mainly comprises two steps: The first step is mixture enhancement which aims to reduce the additive noise and extracts the signal information. The mixture enhancement classifies the noisy mixture into two non-overlapping TF planes of signal absence or signal presence. The noise-reduced mixture will be then obtained by computing the spectral amplitude on the classified signal presence. The second step is the separation process which isolates the original signals by multiplying a mask on the noise-reduced mixture. The mask is constructed by evaluating the cost function given by each signal-signature estimator.

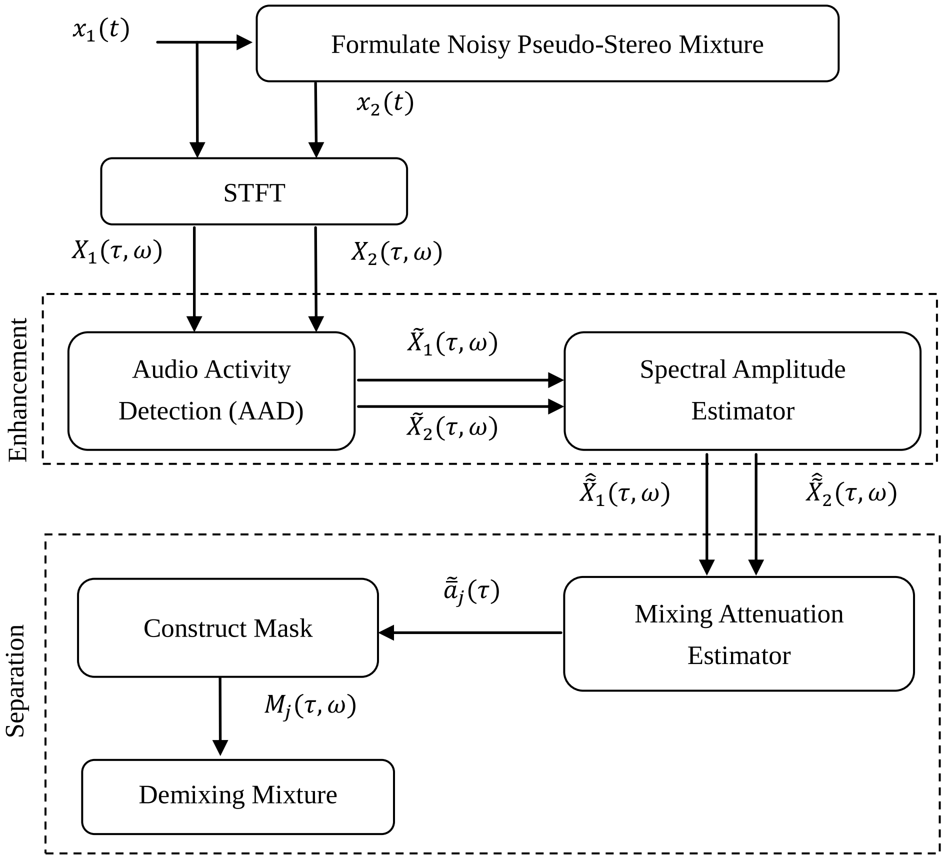

$x_1(t)$
Formulate Noisy Pseudo-Stereo Mixture
$x_2(t)$
STFT
$X_1(\tau,\omega)$
$X_2(\tau,\omega)$
Enhancement
Audio Activity Detection (AAD)
$\tilde{X}_1(\tau,\omega)$
$\tilde{X}_2(\tau,\omega)$
Spectral Amplitude Estimator
$\hat{\tilde{X}}_1(\tau,\omega)$
$\hat{\tilde{X}}_2(\tau,\omega)$
Separation
$\tilde{\bar{a}}_j(\tau)$
Construct Mask
Mixing Attenuation Estimator
$M_j(\tau,\omega)$
Demixing Mixture

Figure 5.1: Overview of the proposed mixture enhancement and separation algorithm in frequency domain.

This chapter is organized as follows: Section 5.1 presents the proposed noisy pseudo-stereo mixing model. The proposed mixture enhancement is articulated in Section 5.2. Next, the proposed signal separation framework is fully expressed in Section 5.3. Experimental results and a series of performance comparison with other existing SCBSS methods are presented in Section 5.4. Finally, Section 5.5 concludes the work of this chapter.

## 5.1 Proposed Noisy Pseudo – Stereo Mixing Model

In this chapter, for simplicity the case of a single-channel noisy mixture of two signals and a noise in time domain is considered as

$$x_1(t) = s_1(t) + s_2(t) + n_1(t) \tag{5.1}$$

where $x_1(t)$ is the single channel mixture, $n_1(t)$ is an additive uncorrelated noise that can be stationary or nonstationary (for generality, this paper will treat it as nonstationary), and $s_1(t)$ and $s_2(t)$ are the original signals which are assumed to be modeled by the autoregressive (AR) process :

$$s_j(t) = -\sum_{m=1}^{D_j} a_{s_j}(m;t)s_j(t-m) + e_j(t) \tag{5.2}$$

where $a_{s_j}(m;t)$ denotes the $m^{th}$ order AR coefficient of the $j^{th}$ signal at time $t$, $D_j$ is the maximum AR order, and $e_j(t)$ is an independent identically distributed (i.i.d.) random signal with zero mean and variance $\sigma_{e_j}^2$. The virtual mixture by weighting and time-shifting the single channel mixture $x_1(t)$ as

$$x_2(t) = \frac{x_1(t) + \gamma x_1(t-\delta)}{1+|\gamma|} \tag{5.3}$$

In (5.3), $\gamma \in \Re$ is the weight parameter, and $\delta$ is the time-delay. The 'noisy pseudo-stereo' mixture (5.3) can be expressed in terms of the source signals, AR coefficient and time-delay as

$$\begin{aligned}
x_2(t) &= \frac{x_1(t) + \gamma x_1(t-\delta)}{1+|\gamma|} \\
&= \frac{s_1(t)+s_2(t)+n(t)+\gamma[s_1(t-\delta)+s_2(t-\delta)+n_1(t-\delta)]}{1+|\gamma|} \\
&= \frac{-\sum_{m=1}^{D_1} a_{s_1}(m)s_1(t-m) + e_1(t)}{1+|\gamma|} + \frac{\gamma s_1(t-\delta)}{1+|\gamma|} + \frac{-\sum_{m=1}^{D_2} a_{s_2}(m)s_2(t-m) + e_2(t)}{1+|\gamma|} + \\
&\quad \frac{\gamma s_2(t-\delta)}{1+|\gamma|} + \frac{n_1(t)+\gamma n_1(t-\delta)}{1+|\gamma|} \\
&= \frac{\left(-a_{s_1}(\delta)+\gamma\right)}{1+|\gamma|} s_1(t-\delta) + \frac{\left(-a_{s_2}(\delta)+\gamma\right)}{1+|\gamma|} s_2(t-\delta) + \frac{e_1(t)-\sum_{\substack{m=1\\m\neq\delta}}^{D_1} a_{s_1}(m)s_1(t-m)}{1+|\gamma|} +
\end{aligned}$$

$$\frac{e_2(t)-\sum_{\substack{m=1\\m\neq\delta}}^{D_2} a_{s_2}(m)s_2(t-m)}{1+|\gamma|}+\frac{n_1(t)+\gamma n_1(t-\delta)}{1+|\gamma|} \tag{5.4}$$

Defining the followings:

$$a_j(t;\delta,\gamma)=\frac{-a_{s_j}(\delta;t)+\gamma}{1+|\gamma|} \tag{5.5}$$

$$r_j(t;\delta,\gamma)=\frac{e_j(t)-\sum_{\substack{m=1\\m\neq\delta}}^{D_j} a_{s_j}(m;t)s_j(t-m)}{1+|\gamma|} \tag{5.6}$$

$$n_2(t;\delta,\gamma)=\frac{n_1(t)+\gamma n_1(t-\delta)}{1+|\gamma|} \tag{5.7}$$

where $a_j(t;\delta,\gamma)$ and $r_j(t;\delta,\gamma)$ represent the mixing attenuation and the residue of the $j^{th}$ signal, respectively, and $n_2(t;\delta,\gamma)$ denotes noise obtained by weighting and time-shifting of the additive noise $n_1(t)$. Using (5.5)-(5.7), the overall proposed noisy mixing model can now be formulated in terms of the signals and the noise as

$$x_1(t)=s_1(t)+s_2(t)+n_1(t)$$

$$x_2(t)=a_1(t;\delta,\gamma)s_1(t-\delta)+a_2(t;\delta,\gamma)s_2(t-\delta)+r_1(t;\delta,\gamma)+r_2(t;\delta,\gamma)+n_2(t;\delta,\gamma) \tag{5.8}$$

This noisy mixing model remains almost similar to the proposed pseudo-stereo mixture in Chapter 3 and differs in terms of the additive noise i.e. $n_1(t)$ and $n_2(t;\delta,\gamma)$.

### *5.1.1 Frequency domain*

Based on the above assumptions, the TF representation of the noisy mixing model is obtained using the STFT of $x_j(t)$, $j=1,2$ as

$$X_1(\tau,\omega)=S_1(\tau,\omega)+S_2(\tau,\omega)+N_1(\tau,\omega)$$

$$X_2(\tau,\omega)=a_1(\tau)e^{-i\omega\delta}S_1(\tau-\delta,\omega)+a_2(\tau)e^{-i\omega\delta}S_2(\tau-\delta,\omega)-$$

$$\left(\sum_{\substack{m=1\\m\neq\delta}}^{D_1}\frac{a_{s_1}(m;\tau)}{1+|\gamma|}e^{-i\omega m}S_1(\tau-m,\omega)\quad+\sum_{\substack{m=1\\m\neq\delta}}^{D_2}\frac{a_{s_2}(m;\tau)}{1+|\gamma|}e^{-i\omega m}S_2(\tau-m,\omega)\right)+N_2(\tau,\omega)$$

(5.9)

for $\forall\tau,\omega$. In (5.9), the fact that $e_j(t)\ll s_j(t)$ has been used, thus the TF of $r_j(t)$ in (5.6) can be simplified to

$$R_j(\tau,\omega)=-\sum_{\substack{m=1\\m\neq\delta}}^{D_j}\frac{a_{s_j}(m;\tau)}{1+|\gamma|}e^{-i\omega m}S_j(\tau-m,\omega)\qquad(5.10)$$

To facilitate further analysis, a term $C_j(\tau,\omega)$ is also defined

$$C_j(\tau,\omega)=\frac{1}{1+|\gamma|}\sum_{\substack{m=1\\m\neq\delta}}^{D_j}a_{s_j}(m;\tau)e^{-i\omega(m-\delta)}\qquad(5.11)$$

which forms a part of $R_j(\tau,\omega)$ without the contribution of the signal $S_j(\tau,\omega)$. From (5.9), it can be seen that the noisy pseudo-stereo mixture comprises four components i.e. $a_je^{-i\omega\delta}$, $C_j(\tau,\omega)$, $N_j(\tau,\omega)$ and $S_j(\tau,\omega)$. The separability of the proposed noise-free pseudo-stereo mixing model in Chapter 3 shows that the proposed noise-free model can be separated when at least $a_j(t;\delta,\gamma)$ or $r_j(t;\delta,\gamma)$ of the signals $j=1$ and $j=2$ are not equal. In the case of a noisy environment, if the signals are extracted from the noisy mixtures such that the remaining noise is small compared to the signals, then this allow the remaining noise to be treated as negligible. Thus, the noisy mixing model then becomes the approximated noise-free mixing model. To achieve this aim, the mixture enhancement method is proposed in the following section.

## 5.2 Proposed Mixture Enhancement

### *5.2.1 Audio Activity Detection*

The audio activity detection (AAD) method enhances the noisy mixture by selecting the TF units that contain original signals and removing those solely of noise. To begin, the two statistical hypotheses are set i.e. $H_0(\tau,\omega)$ and $H_1(\tau,\omega)$ to denote the signal absence and presence, respectively, at $\omega^{th}$ frequency bin of the $\tau^{th}$ frame:

$$\begin{aligned} &H_0(\tau,\omega)\text{: Signal absence:} \quad X(\tau,\omega) = N(\tau,\omega) \\ &H_1(\tau,\omega)\text{: Signal presence:} \quad X(\tau,\omega) = S(\tau,\omega) + N(\tau,\omega) \end{aligned} \tag{5.12}$$

where $X(\tau,\omega)$ is a mixture given by $X_1(\tau,\omega)$ or $X_2(\tau,\omega)$, $S(\tau,\omega)$ is a sum of original signals i.e. $S(\tau,\omega) = S_1(\tau,\omega) + S_2(\tau,\omega)$, and $N(\tau,\omega)$ is the additive noise. $S(\tau,\omega)$ and $N(\tau,\omega)$ are assumed to be complex Gaussian distributed. Source presence at a particular $(\tau,\omega)$ unit is detected by computing a local signal absence probability (LSAP) and selecting the $(\tau,\omega)$ unit that the LSAP is less than a local threshold $T_L$ where $T_L$ can be set by the user. The LSAP can be expressed as

$$\begin{aligned} p\big(H_0(\tau,\omega)\big|X(\tau,\omega)\big) &= \frac{p\big(H_0(\tau,\omega),X(\tau,\omega)\big)}{p(X(\tau,\omega))} \\ &= \frac{p\big(X(\tau,\omega)\big|H_0(\tau,\omega)\big)p(H_0)}{p\big(X(\tau,\omega)\big|H_0(\tau,\omega)\big)p(H_0)+p(X(\tau,\omega)|H_1(\tau,\omega))p(H_1)} \\ &= \frac{1}{1+q_\omega\Lambda(\tau,\omega)} \end{aligned} \tag{5.13}$$

where $p(\cdot)$ denotes a probability density function (PDF), $q_\omega$ is the ratio defined by $q_\omega = \frac{p(H_1)}{p(H_0)}$ , $p(H_0)$ and $p(H_1)$ are the prior probabilities of the respective hypotheses. The term $\Lambda(\tau,\omega) = p\big(X(\tau,\omega)\big|H_1(\tau,\omega)\big)/p\big(X(\tau,\omega)\big|H_0(\tau,\omega)\big)$ is the likelihood ratio of the signal presence and signal absence at $(\tau,\omega)$ units where the likelihood function of the signal presence and absence that can be expressed as: $p\big(X(\tau,\omega)\big|H_1(\tau,\omega)\big) = \frac{1}{\pi(\sigma_S^2(\tau,\omega)+\sigma_N^2(\tau,\omega))} exp(-\frac{|X(\tau,\omega)|^2}{\sigma_S^2(\tau,\omega)+\sigma_N^2(\tau,\omega)})$ and $p\big(X(\tau,\omega)\big|H_0(\tau,\omega)\big) = \frac{1}{\pi\sigma_N^2(\tau,\omega)} exp(-\frac{|X(\tau,\omega)|^2}{\sigma_N^2(\tau,\omega)})$, respectively. In the case of LSAP $\geq T_L$, this particular $(\tau,\omega)$ unit constitutes as noise. In order to update the noise power, a global signal absence probability (GSAP) is used to indicate whether there is a need of an adjustment to the noise power or not. The GSAP computed at the $\tau^{th}$ frame can be expressed as

$$\begin{aligned} p\big(H_0(\tau)\big|X(\tau)\big) &= \frac{p(H_0(\tau),X(\tau))}{p(X(\tau))} \\ &= \frac{p(H_0)\prod_{\omega=1}^{W} p\big(X(\tau,\omega)\big|H_0(\tau)\big)}{p(H_0)\prod_{\omega=1}^{W} p\big(X(\tau,\omega)\big|H_0(\tau)\big)+p(H_1)\prod_{\omega=1}^{W} p\big(X(\tau,\omega)\big|H_1(\tau)\big)} \\ &= \frac{1}{1+q_\omega\prod_{\omega=1}^{W}\Lambda(\tau,\omega)} \end{aligned} \tag{5.14}$$

When the GSAP exceeds a global threshold $T_G$, a noise power estimate is updated. Otherwise, the noise power estimate of the $\tau^{th}$ frame remains the same as in the previous frame. The noise power estimate can be computed as

$$\hat{\sigma}_N^2(\tau,\omega) = \zeta_N\, \hat{\sigma}_N^2(\tau-1,\omega) + (1-\zeta_N\,)|N(\tau,\omega)|^2 \tag{5.15}$$

where $0 < \zeta_N < 1$ is a smoothing parameter of the noise power estimate.

In the traditional voice activity detection (VAD) method [54, 55], the likelihood of the presence of the signal requires the signal power spectral density $\sigma_S^2(\tau,\omega)$ which is unknown. Additionally, In the case of low input SNR where source energy $\sigma_S^2(\tau,\omega)$ is low compared with noise power $\sigma_N^2(\tau,\omega)$ i.e. $\sigma_S^2(\tau,\omega) \ll \sigma_N^2(\tau,\omega)$, the likelihood function of the source presence will become $\frac{1}{\pi\,\sigma_N^2(\tau,\omega)} exp\left(\frac{-|X(\tau,\omega)|^2}{\sigma_N^2(\tau,\omega)}\right)$ which is identical to the source absence likelihood. Consequently, a value of $\Lambda(\tau,\omega)$ is equal to 1. As a result, LSAP obtains a value of the prior probability $q_\omega$ ratio. This case causes

LSAP and GSAP to be independent of the mixture. Therefore, LSAP and GSAP cannot correctly identify $(\tau, \omega)$ units of weak source energy in high noise power.

To remedy the ill conditioned LSAP and GSAP, we replace $\sigma_S^2(\tau, \omega)$ by $\xi_f \sigma_N^2(\tau, \omega)$ where $\xi_f$ is the proposed fixed a priori SNR $\xi_f \triangleq \frac{\sigma_S^2(\tau,\omega)}{\sigma_N^2(\tau,\omega)}$ and $\sigma_N^2(\tau, \omega) \triangleq E\{|N(\tau, \omega)|^2|\}$ denotes the short-term spectrum of the noise. The term $\xi_f$ will be set to emphasize the low source energy in high noise-power units and to prevent the noise power estimates from increasing under weak source activity. As the probability $p\big(X(\tau, \omega)\big|H_1(\tau, \omega)\big)$ differs from $p\big(X(\tau, \omega)\big|H_0(\tau, \omega)\big)$, LSAP can then indicate and select the particular TF units which contain weak source components in low input SNR. Hence, most if not all of the information-bearing source data can be preserved for the separation process. The separation performance requires those essential data for accurate estimating the sources' signatures and using it to evaluate the appropriate TF units that belong to the original signals. Additionally, using $\xi_f \sigma_N^2(\tau, \omega)$ instead $\sigma_S^2(\tau, \omega)$ will benefit the decoupling of the noise power estimator and the source spectral amplitude estimator. In this way, both parameters can be individually estimated with better consistency. In this new light, the likelihood function of the observed signal under signal presence can be expressed as

$$p\big(X(\tau, \omega)\big|H_1(\tau, \omega)\big) = \frac{1}{\pi\sigma_N^2(\tau,\omega)\big(1+\xi_f\big)} exp\left\{-\frac{|X(\tau,\omega)|^2}{\sigma_N^2(\tau,\omega)\big(1+\xi_f\big)}\right\} \tag{5.16}$$

The optimal $\xi_f$ is determined by minimizing the integrated probability of error. The decision rule is based on the comparison of $p\big(H_0(\tau, \omega)\big|x(\tau, \omega)\big)$ with the threshold $T_L$: When $p\big(H_0(\tau, \omega)\big|x(\tau, \omega)\big) \geq T_L$ it is decided to be $H_0$ or else decided to be $H_1$. The probability of error $p_e$ can be expressed as

$$
\begin{aligned}
p_e(\xi,\xi_f) &= p(decide\ H_1|H_0)p(H_0) + p(decide\ H_0|H_1)p(H_1) \\
&= Pr\left\{|X(\tau,\omega)| < X_{T_L}(\tau,\omega); H_0\right\}p(H_0) + Pr\left\{|X(\tau,\omega)| \geq X_{T_L}(\tau,\omega); H_1\right\}p(H_1) \\
&= \int_0^{X_{T_L}(\tau,\omega)} \int_0^{2\pi} \frac{1}{\pi\sigma_N^2(\tau,\omega)} exp\left\{-\frac{\left|X_{T_L}(\tau,\omega)\right|^2}{\sigma_N^2(\tau,\omega)}\right\}\left(re^{j\theta}\right)rdrd\theta p(H_0) + \\
&\quad \int_{X_{T_L}(\tau,\omega)}^{\infty} \int_0^{2\pi} \frac{1}{\pi\sigma_N^2(\tau,\omega)(1+\xi)} exp\left\{-\frac{\left|X_{T_L}(\tau,\omega)\right|^2}{\sigma_N^2(\tau,\omega)(1+\xi)}\right\}\left(re^{j\theta}\right)rdrd\theta p(H_1) \\
&= \left(1 - exp\left\{-\frac{\left|X_{T_L}(\tau,\omega)\right|^2}{\sigma_N^2(\tau,\omega)}\right\}\right)p(H_0)\ +\ exp\left\{-\frac{\left|X_{T_L}(\tau,\omega)\right|^2}{\sigma_N^2(\tau,\omega)(1+\xi)}\right\}p(H_1) \\
&= \left(1 - \left(\frac{p(H_0)}{p(H_1)}\left(1+\xi_f\right)\right)^{-\frac{1+\xi_f}{\xi_f(1+\xi)}}\right)p(H_0)\ +\ \left(\frac{p(H_0)}{p(H_1)}\left(1+\xi_f\right)\right)^{-\frac{1+\xi_f}{\xi_f}} p(H_1) \quad (5.17)
\end{aligned}
$$

where $X_{T_L}(\tau,\omega)$ denotes a threshold boundary between source absence and presence, $\xi$ is the true input SNR of a noisy mixture, and $\xi_f$ is a candidate of the optimal $\xi_f$. The optimal $\xi_f$ can be determined from

$$\hat{\xi}_f = \underset{\xi_f}{argmin}\ \int_{\xi_{down}}^{\xi_{top}} p_e(\xi,\xi_f)\ d\xi \qquad (5.18)$$

where $\hat{\xi}_f$ denotes the optimal $\xi_f$ which is determined by selecting $\xi_f$ that yields the minimum value of $\int_{\xi_{down}}^{\xi_{top}} p_e(\xi,\xi_f)\ d\xi$.

The AAD method delivers the TF plane of the signal-presence mixing model i.e. $\tilde{X}_1(\tau,\omega)$ and $\tilde{X}_2(\tau,\omega)$. The noise power estimator will be used to estimate signal spectral amplitude.

### *5.2.2 Mixture Spectral Amplitude Estimator*

Let $\tilde{X}(\tau,\omega)$ denotes the mixture with signal present at $(\tau,\omega)$ units from the AAD method. This consists of the sum of the source signals and the residual noise $\widetilde{N}(\tau,\omega)$, i.e.

$$\tilde{X}(\tau,\omega) = S(\tau,\omega) + \widetilde{N}(\tau,\omega) \tag{5.19}$$

where $\tilde{X}(\tau,\omega) = \left|\tilde{X}(\tau,\omega)\right|e^{i\theta\omega}$, $S(\tau,\omega) = A(\tau,\omega)e^{i\alpha\omega}$ is the sum of the signals (i.e. $S(\tau,\omega) = \sum_{j=1}^{2} S_j(\tau,\omega)$), and $\theta\omega$ and $\alpha\omega$ are the complex exponential of the noisy phase and signal phase, respectively. The residual noise $\widetilde{N}(\tau,\omega)$ refers to the remaining noise in the signal-presence TF units only. This sub-section focuses on the estimation of the spectrum, $S(\tau,\omega)$, by using the proposed improved mean square error short-time spectral amplitude (iMMSE-STSA) estimator $\hat{A}(\tau,\omega)$. This estimator is solely required for estimating the spectral amplitude $A(\tau,\omega)$ from $\tilde{X}(\tau,\omega)$ since it can be proven that the complex exponential estimator is the complex exponential of the noisy phase i.e. $\theta\omega = \alpha\omega$ [56]. The conventional MMSE-STSA estimator is derived from mathematical derivation by minimizing the mean-square error cost function based on statistical independence assumption and models. The MMSE-STSA estimator $\tilde{A}(\tau,\omega)$ of $A(\tau,\omega)$ is obtained as:

$$\begin{aligned}\tilde{A}(\tau,\omega) &= E\{A(\tau,\omega)|\tilde{X}_1(\tau,\omega)\}\\ &= \frac{q_\omega \Lambda(\tau,\omega)}{1+q_\omega \Lambda(\tau,\omega)} \Gamma(1.5) \frac{\sqrt{\upsilon(\tau,\omega)}}{\gamma_{SNR}(\tau,\omega)} \exp\left(-\frac{\upsilon(\tau,\omega)}{2}\right)\left[\left(1+\upsilon(\tau,\omega)\right)I_0\left(\frac{\upsilon(\tau,\omega)}{2}\right)+\upsilon(\tau,\omega)I_1\left(\frac{\upsilon(\tau,\omega)}{2}\right)\right]\left|\tilde{X}_1(\tau,\omega)\right|\end{aligned} \tag{5.20}$$

where $q_\omega \triangleq p(H_1)/p(H_0)$, $\Gamma(\cdot)$ indicates the gamma function, with $\Gamma(1.5) = \frac{\sqrt{\pi}}{2}$, $I_0(\cdot)$ and $I_1(\cdot)$ indicates the modified Bessel functions of zero^th and first order, respectively. $\upsilon(\tau,\omega)$ is defined by $\upsilon(\tau,\omega) = \frac{\xi(\tau,\omega)}{1+\xi(\tau,\omega)}\left(\gamma_{SNR}(\tau,\omega)\right)$, $\gamma_{SNR}(\tau,\omega) \triangleq \frac{|\tilde{X}(\tau,\omega)|^2}{\sigma_N^2(\tau,\omega)}$ and $\xi(\tau,\omega) \triangleq \frac{\sigma_S^2(\tau,\omega)}{\sigma_N^2(\tau,\omega)}$ denote the *a posteriori* SNR and *a priori* SNR, respectively. The efficiency of conventional MMSE-STSA estimator is based on $\hat{\gamma}_{SNR}(\tau,\omega)$ and $\hat{\xi}(\tau,\omega)$ where denote the estimates of $\gamma_{SNR}(\tau,\omega)$ and $\xi(\tau,\omega)$, respectively. The term $\hat{\gamma}_{SNR}(\tau,\omega)$ and $\hat{\xi}(\tau,\omega)$ influence a degree of accuracy of $\tilde{A}(\tau,\omega)$ in (5.20). However, under the case of weak signal components and low input SNR, the conventional *a posteriori* SNR estimator $\hat{\gamma}_{SNR}(\tau,\omega)$ causes deterioration of the weak signal components. This case can be analyzed as follows:

$$\begin{aligned}\gamma_{SNR}(\tau,\omega) &\triangleq \frac{|\tilde{X}(\tau,\omega)|^2}{\sigma_N^2(\tau,\omega)}\\ &= \frac{|S(\tau,\omega)+N(\tau,\omega)|^2}{\sigma_N^2(\tau,\omega)}\end{aligned}$$

Using the subadditivity properties of the absolute value, it is obtained

$$\begin{aligned}E\left[\frac{|S(\tau,\omega)+N(\tau,\omega)|^2}{\sigma_N^2(\tau,\omega)}\right] &\le E\left[\frac{|S(\tau,\omega)|^2+|N(\tau,\omega)|^2}{\sigma_N^2(\tau,\omega)}\right]\\ &= \frac{\sigma_S^2(\tau,\omega)+\sigma_N^2(\tau,\omega)}{\sigma_N^2(\tau,\omega)}\end{aligned}$$

In the case of weak signal components and low inputs SNR i.e. $\sigma_S^2(\tau,\omega) \approx 0$, the term $\gamma_{SNR}(\tau,\omega)$ then have

$$\gamma_{SNR}(\tau,\omega) \le 1$$

The estimation of $\xi(\tau,\omega)$ can be shown to be given by $\hat{\xi}(\tau,\omega) = \zeta_\xi\, \hat{A}^2(\tau-1,\omega)/\sigma_N^2(\tau-1,\omega) + \left(1-\zeta_\xi\right) max\{\hat{\gamma}_{SNR}(\tau,\omega)-1{,}0\}$ which comprises of two terms i.e. the first term represents the scaled *a priori* SNR estimator of its previous frame. The second term is a maximum likelihood estimate of the *a posteriori* SNR $\gamma_{SNR}$ based entirely on the current frame. The term $\zeta_\xi$, $0<\zeta_\xi<1$, is a weighing factor that controls the trade-off between the noise reduction and the transient distortion brought into the signal. At a particular $(\tau,\omega)$ unit of weak signal activity and low input SNR where $\hat{\gamma}_{SNR}(\tau,\omega) \leq 1$, this will cause $\hat{\xi}(\tau,\omega)$ to be solely dominated by the first term i.e. $\zeta_{\xi}\,\tilde{A}^2(\tau-1,\omega)/\sigma_N^2(\tau-1,\omega)$ due to $\left(1-\zeta_{\xi}\right) max\{\gamma_{SNR}(\tau,\omega)-1,0\} = 0$. Thus, $\hat{\xi}(\tau,\omega)$ depends only on the scaling of its previous frame without taking the scaled *a posteriori* SNR estimator into account $\left(1-\zeta_{\xi}\right) max\{\gamma_{SNR}(\tau,\omega)-1,0\}$. The term $\gamma_{SNR}(\tau,\omega)$ is important because it reacts to changes in the signal energy. This property is naturally suited to nonstationary signals such as audio signals. The term $\hat{\xi}(\tau,\omega)$ tends to be stationary and smaller along time frames. The underestimation of $\hat{\xi}(\tau,\omega)$ will cause the spectral amplitude estimator $\tilde{A}(\tau,\omega)$ to be more sensitive to errors. Additionally, $\tilde{A}(\tau,\omega)$ will be intolerably suppressed such that weak source components are also removed as well. Therefore, this leads to the loss of information-bearing source-data which will impact performance of the separation process.

To overcome this issue, the estimation of $\xi(\tau,\omega)$ can be improved by computing the *a posteriori* SNR $\hat{\gamma}_{SNR}(\tau,\omega)$ from the signal presence probability (SPP) with fixed *a priori* $\xi_f$ to guarantee that $\hat{\gamma}_{SNR}(\tau,\omega) > 1$ . the *a posteriori* SNR $\hat{\gamma}_{SNR}(\tau,\omega)$ estimator can be expressed as:

$$\hat{\gamma}_{SNR}(\tau,\omega) = \frac{1}{E} log\left(\frac{1}{q_{\omega}}\left(\frac{1+\xi_f}{p\left(H_1(\tau,\omega)\middle|\tilde{X}(\tau,\omega)\right)^{-1}-1}\right)\right) \tag{5.21}$$

where $E = \xi_f/1+\xi_f$ and $p\left(H_1(\tau,\omega)\middle|\tilde{X}(\tau,\omega)\right)$ denotes a SPP given by the Bayes' theorem:

$$p\left(H_1(\tau,\omega)\middle|\tilde{X}(\tau,\omega)\right) = \frac{p\left(H_1(\tau,\omega),\tilde{X}(\tau,\omega)\right)}{p\left(\tilde{X}(\tau,\omega)\right)}$$

$$= \frac{p\left(\tilde{X}(\tau,\omega)\middle|H_1(\tau,\omega)\right)p(H_1)}{p\left(\tilde{X}(\tau,\omega)\middle|H_0(\tau,\omega)\right)p(H_0)+p\left(\tilde{X}(\tau,\omega)\middle|H_1(\tau,\omega)\right)p(H_1)}$$

$$= \left(\frac{1+\xi_f}{q_\omega}\exp\left\{-E\ \frac{\left|\tilde{X}(\tau,\omega)\right|^2}{\hat{\sigma}_N^2(\tau,\omega)}\right\}+1\right)^{-1} \tag{5.22}$$

Using the $\xi_f$ and $p\left(H_1(\tau,\omega)\middle|\tilde{X}(\tau,\omega)\right) > 0.08$, the *a posteriori* SNR estimator then satisfies $\hat{\gamma}_{SNR}(\tau,\omega) > 1$. Hence, the term $\hat{\xi}(\tau,\omega)$ can be obtained by computing both estimators of the previous and current frames. Therefore, to extract signal information even when signal components are weak in low input SNR, the proposed iMMSE-STSA firstly estimate the *a posteriori* SNR using (5.21) and then using this estimate for computing the spectral amplitude. Finally, the estimated spectra of the mixture can be formulated as

$$\hat{S}(\tau,\omega) = \hat{A}(\tau,\omega)e^{i\theta\omega} \tag{5.23}$$

In conclusion, the proposed mixture enhancement method is to improve the quality of the source signals in $X_j(\tau,\omega)$ , $j = 1,2$. The proposed mixture enhancement will benefit the signal separation by providing the greater degree of signal information by attempting to select the TF units of signal presence and reject the TF units of solely noise. In addition, the remaining noise in $\tilde{X}_j(\tau,\omega)$ which impacts the separation performance especially of low signal energy in low SNR, is suppressed by employing the proposed iMMSE-STSA to extract signal components from $\tilde{X}_j(\tau,\omega)$. Finally, by using (5.23), the noise-reduced mixture can now be modeled as $\hat{\tilde{X}}(\tau,\omega) = \hat{A}(\tau,\omega)e^{i\theta\omega} + \tilde{N}(\tau,\omega)$ (recall that $\tilde{N}(\tau,\omega)$ is the residual noise from (5.19)) which will then be separated by a binary TF mask. The proposed separation algorithm will be articulated in the following section.

## 5.3 Proposed Source Separation

### *5.3.1 Adaptive Mixing Parameter Estimator*

The core concept of our proposed separating algorithm is to construct a cost function to build up a TF mask which requires the estimation of the AR coefficients and time-delay of the signals. Assuming that the $j^{th}$ signal is dominant at a particular $(\tau, \omega)$ unit, the noise-reduced mixture can be more specifically expressed as:

$$\begin{aligned}
\hat{\tilde{X}}_1(\tau, \omega) &= \hat{S}_j(\tau, \omega) + \widetilde{N}_1(\tau, \omega) \\
\hat{\tilde{X}}_2(\tau, \omega) &= a_j(\tau)e^{-i\omega\delta}\hat{S}_j(\tau-\delta, \omega) - \textstyle\sum_{\substack{m=1 \\ m\neq\delta}}^{D_j} \frac{a_{s_j}(m;\tau)}{1+|\gamma|} e^{-i\omega m}\, \hat{S}_j(\tau-m, \omega) + \widetilde{N}_2(\tau, \omega) \\
&\approx \left[a_j(\tau) - C_j(\tau, \omega)\right]e^{-i\omega\delta}\hat{S}_j(\tau, \omega) + \widetilde{N}_2(\tau, \omega), \quad (\tau, \omega) \in \Omega_j
\end{aligned} \tag{5.24}$$

for $\delta$ and $m \leq \phi$. The term $C_j(\tau, \omega)$ is given by (5.11) and $\Omega_j$ is the $j^{th}$ signal presence area defined as $\Omega_j := \{(\tau, \omega) \colon \hat{S}_j(\tau, \omega) \neq 0\}$, $\forall k \neq j$. The estimate of $\bar{a}_j(\tau, w) = a_j(\tau) - C_j(\tau, \omega)$ associated with the $j^{th}$ signal can be determined as

$$\begin{aligned}
\bar{a}_j(\tau, \omega) &= \frac{\hat{\tilde{X}}_2(\tau,\omega)}{\hat{\tilde{X}}_1(\tau,\omega)} e^{i\omega\delta} \\
&= \frac{\left[a_j(\tau)-C_j(\tau,\omega)\right]e^{-i\omega\delta}\hat{S}_j(\tau,\omega)+\widetilde{N}_2(\tau,\omega)}{\hat{S}_j(\tau,\omega)+\widetilde{N}_1(\tau,\omega)} e^{i\omega\delta}
\end{aligned} \tag{5.25}$$

The term $\widetilde{N}_1(\tau, \omega)$ and $\widetilde{N}_2(\tau, \omega)$ can be assumed to be small after the mixture enhancement step (this evidence is shown in Figs. 5.3 and 5.4 in *Section 5.4.2*). In this case, the term $\bar{a}_j(\tau, \omega)$ can be expressed as

$$\begin{aligned}
\bar{a}_j(\tau, \omega) &= \frac{[a_j(\tau)-C_j(\tau,\omega)]e^{-i\omega\delta}\hat{s}_j(\tau,\omega)}{s_j(\tau,\omega)} e^{i\omega\delta} \\
&= a_j(\tau) - C_j(\tau, \omega) \\
&= \bar{a}_j^{(r)}(\tau, \omega) + i\bar{a}_j^{(i)}(\tau, \omega)\ , \quad \forall(\tau, \omega) \in \Omega_j
\end{aligned} \tag{5.26}$$

where $\bar{a}_j^{(r)}(\tau,\omega) = Re\left[\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)} e^{i\omega\delta}\right]$ and $\bar{a}_j^{(i)}(\tau,\omega) = Im\left[\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)} e^{i\omega\delta}\right]$ are the real and imaginary parts of $\bar{a}_j(\tau,\omega)$, respectively, and $i = \sqrt{-1}$. Although the ratio $\hat{\hat{\bar{X}}}_2(\tau,\omega)/\hat{\hat{\bar{X}}}_1(\tau,\omega)$ seems straightforward, it is difficult to obtain $\bar{a}_j(\tau,\omega)$ directly from this ratio because the term $C_j(\tau,\omega)$ varies with frequency from frame to frame. In addition, audio signal is nonstationary and correlated between neighbouring frequencies bins of consecutive frames. To overcome this problem, the adaptive estimate $\bar{a}_j(\tau,\omega)$ frame-by-frame is proposed. Firstly, the complex 2-dimentional (2D) histogram is used to estimate $\bar{a}_j(\tau,\omega)$ frame by frame where the TF units are then clustered into a number of groups corresponding to the number of signals in the mixture. The difference of the complex 2D histogram in this chapter from the first proposed one in Chapter 4.2.2 is that the real and imaginary pair $\left(\hat{\bar{a}}_j^{(r)}(\tau), \hat{\bar{a}}_j^{(i)}(\tau)\right)$ is estimated for each frame basis:

$$\hat{\bar{a}}_j^{(r)}(\tau) = \frac{\sum_\omega \left|\hat{\hat{\bar{X}}}_1(\tau,\omega)\hat{\hat{\bar{X}}}_2(\tau,\omega)\right| Re\left[\frac{\hat{\hat{\bar{X}}}_2(\tau,\omega)}{\hat{\hat{\bar{X}}}_1(\tau,\omega)} e^{i\omega\delta}\right]}{\sum_\omega \left|\hat{\hat{\bar{X}}}_1(\tau\ ,\omega)\hat{\hat{\bar{X}}}_2(\tau,,\omega)\right|}$$

$$\hat{\bar{a}}_j^{(i)}(\tau) = \frac{\sum_\omega \left|\hat{\hat{\bar{X}}}_1(\tau,\omega)\hat{\hat{\bar{X}}}_2(\tau,\omega)\right| Im\left[\frac{\hat{\hat{\bar{X}}}_2(\tau,\omega)}{\hat{\hat{\bar{X}}}_1(\tau,\omega)} e^{i\omega\delta}\right]}{\sum_\omega \left|\hat{\hat{\bar{X}}}_1(\tau,\omega)\hat{\hat{\bar{X}}}_2(\tau,\omega)\right|} \qquad (5.27)$$

Thus, the frame basis estimate of $\bar{a}_j(\tau,\omega)$ can then be formed as

$$\hat{\bar{a}}_j(\tau) = \hat{\bar{a}}_j^{(r)}(\tau) + i\hat{\bar{a}}_j^{(i)}(\tau) \qquad (5.28)$$

where can be expressed as $\hat{\bar{a}}_j(\tau) = \hat{a}_j(\tau) - \hat{C}_j(\tau)$ by relating (5.28) and (5.25). The term $\hat{a}_j(\tau)$ and $\hat{C}_j(\tau)$ are the power weighted estimation of $a_j(\tau)$ and $C_j(\tau,\omega)$, respectively. Secondly, the adaptive mixing attenuation estimator $\tilde{\bar{a}}_j(\tau)$ is obtained by smoothing $\hat{\bar{a}}_j(\tau-1)$ and $\hat{\bar{a}}_j(\tau)$:

$$\bar{\tilde{a}}_j(\tau) = \zeta_M \bar{\tilde{a}}_j(\tau - 1) + (1 - \zeta_M)\hat{\bar{a}}_j(\tau) \qquad (5.29)$$

where $0 < \zeta_M < 1$ is a smoothing parameter of the adaptive mixing attenuation estimator. Determining $\zeta_M$ for optimal tracking will be investigated in *Section 5.4.3*.

### *5.3.2 Construction of Masks*

In this section, the construction of the binary TF masks using sole $\hat{\tilde{X}}_1(\tau,\omega)$ will be presented. The binary TF masks can be constructed by labeling each TF unit with the $k$ argument through maximizing the instantaneous likelihood function. The derivation in this section follows similar steps as Chapter 4.2.2 which taking the residue of the noises into account. The instantaneous likelihood function is derived from the maximum likelihood (ML) framework by first formulating the Gaussian likelihood function $p(\hat{\tilde{X}}_1(\tau,\omega), \hat{\tilde{X}}_2(\tau,\omega)|S_j(\tau,\omega), \bar{\tilde{a}}_j, \sigma^2_{\tilde{\tilde{N}}_j})$ using (5.24), maximizing the likelihood function with respect to $S_j(\tau,\omega)$ and then substituting the obtained result into the Gaussian likelihood function. The resulting instantaneous likelihood function finally takes the following form:

$$\begin{aligned} L_j(\tau,\omega) &:= p(\hat{\tilde{X}}_1(\tau,\omega), \hat{\tilde{X}}_2(\tau,\omega)|\hat{\bar{a}}_j) \\ &= \frac{1}{2\pi}\, exp\left(-\frac{1}{2}\frac{\left|\bar{\tilde{a}}_j(\tau)e^{-i\omega\delta}\hat{\tilde{X}}_1(\tau,\omega) - \hat{\tilde{X}}_2(\tau,\omega)\right|^2}{\hat{\sigma}^2_{\tilde{N}_2}(\tau,\omega) + \hat{\sigma}^2_{\tilde{N}_1}(\tau,\omega)\,\bar{\tilde{a}}_j^2(\tau)}\right) \qquad (5.30) \end{aligned}$$

The function $L_j(\tau,\omega)$ in (5.30) clusters every $(\tau,\omega)$ unit to the $j^{th}$ dominating signal for $L_j(\tau,\omega) \geq L_k(\tau,\omega),\ \forall k \neq j$. This process is equivalent to the following minimization problem:

$$F(\tau,\omega) = \underset{k}{argmin} \frac{\left|\tilde{\bar{a}}_k(\tau)e^{-i\omega\delta}\hat{\tilde{X}}_1(\tau,\omega)-\hat{\tilde{X}}_2(\tau,\omega)\right|^2}{\hat{\sigma}^2_{\tilde{N}_2}(\tau,\omega)+\hat{\sigma}^2_{\tilde{N}_1}(\tau,\omega)\,\tilde{\bar{a}}^2_k(\tau)} \tag{5.31}$$

Using (5.24), the term $\hat{\tilde{X}}_2(\tau,\omega)$ can be expressed as:

$$\begin{aligned}
\hat{\tilde{X}}_2(\tau,\omega) &= a_j(\tau)e^{-i\omega\delta}\hat{S}_j(\tau-\delta,\omega) - \sum_{\substack{m=1\\ m\neq\delta}}^{D_j} \frac{a_{s_j}(m;\tau)}{1+|\gamma|} e^{-i\omega m}\,\hat{S}_j(\tau-m,\omega) + \tilde{N}_2(\tau,\omega) \\
&= \frac{-a_{s_j}(\delta;t)+\gamma}{1+|\gamma|} e^{-i\omega\delta}\hat{S}_j(\tau-\delta,\omega) - \sum_{\substack{m=1\\ m\neq\delta}}^{D_j} \frac{a_{s_j}(m;\tau)}{1+|\gamma|} e^{-i\omega m}\,\hat{S}_j(\tau-m,\omega) + \\
&\tilde{N}_2(\tau,\omega) \\
&= \frac{\gamma}{1+|\gamma|} e^{-i\omega\delta}\hat{S}_j(\tau-\delta,\omega) - \sum_{m=1}^{D_j} \frac{a_{s_j}(m;\tau)}{1+|\gamma|} e^{-i\omega m}\,\hat{S}_j(\tau-m,\omega) + \tilde{N}_2(\tau,\omega) \\
&= \frac{\gamma}{1+|\gamma|} e^{-i\omega\delta}\hat{S}_j(\tau-\delta,\omega) + \frac{\hat{S}_j(\tau,\omega)-E_j(\tau,\omega)}{1+|\gamma|} + \frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\tilde{N}_1(\tau,\omega)
\end{aligned}$$

By invoking the local stationarity, the above is then obtained

$$\hat{\tilde{X}}_2(\tau,\omega) = \frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\left(\hat{S}_j(\tau,\omega)+\tilde{N}_1(\tau,\omega)\right) - \frac{E_j(\tau,\omega)}{1+|\gamma|} \tag{5.32}$$

for $\delta \leq \phi$. The derivation of $\hat{\tilde{X}}_2(\tau,\omega)$ in the signal domain in (5.32) allows $\hat{\tilde{X}}_2(\tau,\omega)$ to be expressed in the mixture domain as:

$$\hat{\tilde{X}}_2(\tau,\omega) \approx \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\hat{\tilde{X}}_1(\tau,\omega) - \frac{E_j(\tau,\omega)}{1+|\gamma|} \tag{5.33}$$

In this light, the proposed cost function can be formulated based on the single mixture $\hat{\tilde{X}}_1(\tau,\omega)$ by substituting this relation into (5.31) which leads to

$$J(\tau,\omega) = \underset{k}{\operatorname{argmin}}\, G_k(\tau,\omega) \tag{5.34}$$

where

$$G_k(\tau,\omega) = \left|\frac{\tilde{\bar{a}}_k e^{-i\omega\delta}\hat{\tilde{X}}_1(\tau,\omega)-\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\hat{\tilde{X}}_1(\tau,\omega)}{\hat{\sigma}^2_{\tilde{\tilde{N}}_2}(\tau,\omega)+\hat{\sigma}^2_{\tilde{\tilde{N}}_1}(\tau,\omega)\,\tilde{\bar{a}}^2_k(\tau)}\right|^2 \tag{5.35}$$

Since $e_j(t) \ll s_j(t)$, the term $E_j(\tau,\omega)/(1+|\gamma|)$ is negligible. Hence, $\hat{\tilde{X}}_2(\tau,\omega) \approx \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\hat{\tilde{X}}_1(\tau,\omega)$. Using (5.34) and (5.35), in the instance when the $j^{th}$ signal dominates at $(\tau,\omega) \in \Omega_j$ the cost function will correctly identify the signal if and only if $G_{k=j}(\tau,\omega) < G_{k\neq j}(\tau,\omega)$. To elucidate this condition, firstly, the case when $k = j$ is considered by setting $\zeta_M = 0$:

$$G_{k=j}(\tau,\omega) = \left|\tilde{\tilde{a}}_j(\tau)e^{-i\omega\delta}\left(\hat{S}_j(\tau,\omega) + \tilde{N}_1(\tau,\omega)\right) - \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\left(\hat{S}_j(\tau,\omega) + \tilde{N}_1(\tau,\omega)\right)\right|^2$$

$$= \left|\hat{a}_j(\tau)e^{-i\omega\delta}\hat{S}_j(\tau,\omega) - \hat{C}_j(\tau)e^{-i\omega\delta}S_j(\tau,\omega) + \tilde{\tilde{a}}_j(\tau)e^{-i\omega\delta}\tilde{N}_1(\tau,\omega) - \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\hat{S}_j(\tau,\omega) - \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\tilde{N}_1(\tau,\omega)\right|^2$$

$$= \left|-\hat{C}_j(\tau)e^{-i\omega\delta}\hat{S}_j(\tau,\omega) + \tilde{\tilde{a}}_j(\tau)e^{-i\omega\delta}\tilde{N}_1(\tau,\omega) + \sum_{\substack{m=1\\ m\neq\delta}}^{D_j}\left(\frac{a_{S_j}(m;\tau)e^{-i\omega m}}{1+|\gamma|}\right)\hat{S}_j(\tau-m,\omega) - \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\tilde{N}_1(\tau,\omega)\right|^2$$

$$= \left|-\hat{C}_j(\tau)e^{-i\omega\delta}\hat{S}_j(\tau,\omega) + \tilde{\tilde{a}}_j(\tau)e^{-i\omega\delta}\tilde{N}_1(\tau,\omega) - \frac{\hat{S}_j(\tau,\omega)}{1+|\gamma|} - \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}e^{-i\omega\delta}\hat{S}_j(\tau-\delta,\omega) - \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\tilde{N}_1(\tau,\omega)\right|^2$$

$$= \left|-\left(\hat{C}_j(\tau) + \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right)e^{-i\omega\delta}\hat{S}_j(\tau,\omega) + \tilde{\tilde{a}}_j(\tau)e^{-i\omega\delta}\tilde{N}_1(\tau,\omega) - \frac{\hat{S}_j(\tau,\omega)}{1+|\gamma|} + \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\tilde{N}_1(\tau,\omega)\right|^2 \quad (5.36)$$

When $k \neq j$, following the above step leads to

$$G_{k\neq j}(\tau,\omega) = \left|\left(\tilde{\tilde{a}}_k(\tau) - a_j(\tau) - \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right)e^{-i\omega\delta}\hat{S}_j(\tau,\omega) + \tilde{\tilde{a}}_k(\tau)e^{-i\omega\delta}\tilde{N}_1(\tau,\omega) - \frac{\hat{S}_j(\tau,\omega)}{1+|\gamma|} + \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\tilde{N}_1(\tau,\omega)\right|^2 \quad (5.37)$$

To guarantee that $G_{k=j}(\tau,\omega) < G_{k\neq j}(\tau,\omega)$ is always satisfied, a condition for $\hat{C}_j$ must then be specified. Starting with (5.36) and (5.37), the condition $\hat{C}_j$ can be expressed

$$\left|-\left(\hat{C}_j(\tau) + \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right)e^{-i\omega\delta}\hat{S}_j(\tau,\omega) + \tilde{\tilde{a}}_j(\tau)e^{-i\omega\delta}\tilde{N}_1(\tau,\omega) - \left(\frac{\hat{S}_j(\tau,\omega)}{1+|\gamma|} + \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\tilde{N}_1(\tau,\omega)\right)\right|^2 <$$
$$\left|\left(\tilde{\tilde{a}}_k(\tau) - a_j(\tau) - \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right)e^{-i\omega\delta}\hat{S}_j(\tau,\omega) + \tilde{\tilde{a}}_k(\tau)e^{-i\omega\delta}\tilde{N}_1(\tau,\omega) - \left(\frac{\hat{S}_j(\tau,\omega)}{1+|\gamma|} + \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\tilde{N}_1(\tau,\omega)\right)\right|^2 \quad (5.38)$$

Eq. (5.38) is bounded by

$$\left|\hat{C}_j(\tau)\hat{S}_j(\tau,\omega)\right| - \left|\frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\hat{S}_j(\tau,\omega) - \tilde{\tilde{a}}_j(\tau)\tilde{N}_1(\tau,\omega)\right| - \left|\frac{\hat{S}_j(\tau,\omega)}{1+|\gamma|} + \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\tilde{N}_1(\tau,\omega)\right| <$$

$$\left|\left(\tilde{\tilde{a}}_k(\tau) - a_j(\tau) - \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right)\hat{S}_j(\tau,\omega) + \tilde{\tilde{a}}_k(\tau)\tilde{N}_1(\tau,\omega)\right| + \left|\frac{\hat{S}_j(\tau,\omega)}{1+|\gamma|} + \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\tilde{N}_1(\tau,\omega)\right|$$

and therefore the $|\hat{C}_j(\tau)|$ condition then obtains

$$|\hat{C}_j(\tau)| < \left|\left(\tilde{\tilde{a}}_k(\tau) - a_j(\tau) - \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right) + \tilde{\tilde{a}}_k(\tau)\frac{\tilde{N}_1(\tau,\omega)}{\hat{S}_j(\tau,\omega)}\right| + \left|\frac{a_{S_j}(\delta;\tau)}{1+|\gamma|} - \tilde{\tilde{a}}_j(\tau)\frac{\tilde{N}_1(\tau,\omega)}{\hat{S}_j(\tau,\omega)}\right| + \frac{2}{1+|\gamma|}\left|1 + \left(1+\gamma e^{-i\omega\delta}\right)\frac{\tilde{N}_1(\tau,\omega)}{\hat{S}_j(\tau,\omega)}\right| \tag{5.39}$$

for $\forall j \neq k$. As $\tilde{N}_1(\tau,\omega)$ has small energy compared with signal energy it can be treated as negligible. Hence, Eq. (5.39) can be simplified to

$$|\hat{C}_j(\tau)| < \left|\left(\tilde{\tilde{a}}_k(\tau) - a_j(\tau) - \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right)\right| + \left|\frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right| + \frac{2}{1+|\gamma|} \tag{5.40}$$

If the condition in (5.40) is satisfied across $\Omega_j$, the cost function (5.34) - (5.35) will then correctly assign the $(\tau,\omega)$ unit to the $j^{th}$ signal. On the contrary, the respectively $(\tau,\omega)$ unit will be wrongly assigned if $|\hat{C}_j(\tau)|$ is larger than the right-hand side of (5.40). Once the cost function is calculated, the binary TF mask for the $j^{th}$ signal can then be constructed as

$$M_j(\tau,\omega) := \begin{cases} 1 & J(\tau,\omega) = j \\ 0 & otherwise \end{cases}. \tag{5.41}$$

The proposed algorithm is summarized in Table 5.1.

Table 5.1: Overview proposed algorithm

1. **Pseudo-Stereo Mixture step:** Formulate the pseudo-stereo mixture $x_2(t)$ using (5.10).
2. **Transform step:** Transform two mixtures $x_1(t)$ and $x_2(t)$ into TF domain by using STFT.

**3. Online Single-Channel Demixing:**

*A. Single-Channel Source Enhancement step:*

1) Audio Activity Detection: Compute the local SAP at the $\tau^{th}$ frame bin and the $\omega^{th}$ frequency of two mixtures using (5.13) and the global SAP for the $\tau^{th}$ frame using (5.14). If the global SAP $> T_G$ then update $\hat{\sigma}^2_{\tilde{N}_j}(\tau,\omega)$ using (5.15).

2) iMMSE-STSA Estimator: Compute the iMMSE estimator of the signal spectral amplitude using (5.20) and formulate the estimated spectra of the $j^{th}$ signals $\tilde{S}(\tau,\omega)$ using (5.23) for both mixtures.

*B. Separation step:*

1) Compute the mixing attenuation estimators $(\hat{\hat{a}}_j^{(r)}(\tau), \hat{\hat{a}}_j^{(i)}(\tau))$ at the $\tau^{th}$ frame using (5.27).

2) Evaluate the cost function $J(\tau,\omega)$ using (5.34, 5.35), and form the binary TF mask $M_j(\tau,\omega)$ using (5.41). Recover the original signals by

$$\hat{\hat{S}}_j(\tau,\omega) = M_j(\tau,\omega)\hat{\tilde{X}}_1(\tau,\omega) \tag{5.42}$$

Finally, convert the estimated signals from TF domain into time domain.

## 5.4 Results and Analysis

The mixture enhancement and the separation performance of the proposed method have been evaluated on real-audio signals in nonstationary noise. A noisy mixture is generated by adding two audio signals and an uncorrelated nonstationary noise with various input SNRs. 20 speech, 20 music signals and noise signals have been randomly selected from TIMIT, RWC, and Noisex [57] databases, respectively. The Noisex database contains 15 various noise signals which can be classified into stationary noise group (i.e. HF radio channel and white noises) and nonstationary ones (i.e. voice babble, factory babble, and various military noises). Additionally, experiments have been conducted to determine the optimal $\xi_f$ and the choice of $\zeta_M$. All experiments have been conducted under the same conditions as follows: The signals are mixed with normalized power over the duration of the signals. All mixed signals are sampled at 16 kHz sampling rate. The TF representation is computed by using the STFT of 1024-point Hamming window with 50% overlap. The parameters are set as follows: for the pseudo-stereo noisy mixture $\delta = 2$ and $\gamma = 4$ for the smoothing parameter of the noise power and the *a priori* SNR estimates $\zeta_N = 0.95$ and $\zeta_\xi = 0.98$, respectively, and $p(H_0) = p(H_1) = 0.5$. These parameters are selected after conducting the Monte-Carlo experiment over 100 independent realizations of 100 mixtures. The separation performance is evaluated by measuring the distortion between the original signal and the estimated one according to the signal-to-distortion (SDR) ratio defined as $\mathrm{SDR} = 10\,log_{10}\left(\left\|s_{target}\right\|^2/\left\|e_{interf}+e_{noise}+e_{artif}\right\|^2\right)$ where $e_{interf}$, $e_{noise}$, and $e_{artif}$ represent the interference from other signals, noise and artifact signals. MATLAB is used as the programming platform. All simulations and analyses are performed using a PC with Intel Core 2 CPU 3GHz and 3GB RAM.

### *5.4.1 Determination of Optimal $\xi_f$ for Mixture Enhancement*

The optimal $\xi_f$ is determined by minimizing the proposed integrated probability of error in (5.17) and (5.18) in *Section 5.2.1)*. The term $\xi$ varies from $0dB$ to $30dB$ by $5dB$ increment. The candidate $\xi_f$ is converted from linear scale to dB (i.e. $10\log_{10}\xi_f = \xi_f^{dB}dB$) with various $\xi_f^{dB}$ from $0dB$ to $50dB$ by $5dB$ increment.

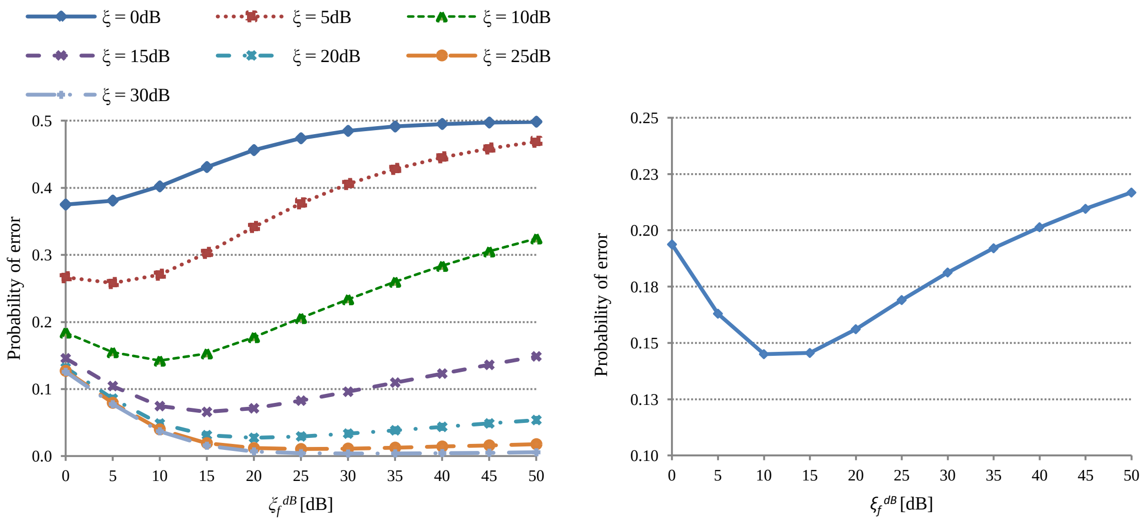


Figure 5.2: Probability of error $p_e\left(\xi, \xi_f\right)$ of individual $\xi$ value (left) and integrated probability of error for various $\xi_f$ (right).

Fig.5.2 on the left-hand side shows the plot of $p_e(\xi, \xi_f)$ for various $\xi$ values. As a result of individual $\xi$, the minimum $p_e(\xi, \xi_f)$ is obtained at $\xi_f = \hat{\xi}_f = \xi$. Therefore, the optimal $\xi_f$ is then set by $\xi$. However in realistic scenario, the term $\xi$ is unknown. Thus, the optimal $\xi_f$ in (5.18) is determined by approximating the above integral in (5.18) by discretely evaluating the term at various $\xi$ values and taking the average. The result is shown on the right-hand side of Fig.5.2. It can be seen that the range of $\tilde{\xi}_f$ that yields the minimum error is between $10\text{dB}$ and $15\text{dB}$. Based on this result, the optimal $\xi_f$ can be set at $10\log_{10}\xi_f = 12.5\text{ dB}$ for all experiments.

### *5.4.2 Mixture Enhancement Performance*

To verify the proposed mixture enhancement method, a test has been conducted on the proposed method by using the mean-square error (MSE, $MSE = \frac{1}{N}\sum_{n=1}^{N}|s(n) - \hat{s}(n)|^2$) and the perceptual evaluation of speech quality (PESQ) measures. The PESQ has been found to correlate highly with both the intelligibility and the quality of speech [58].

Higher PESQ values signify better speech quality with the possible range between -0.5 (worst) to +4.5 (best) defined by ITU Recommendation P.862. The MATLAB implementation provided by [59] has been used to measure PESQ. The experiments have been assessed on three types of mixtures i.e. music + music, speech + music, and speech + speech. Fifty noisy mixtures have been conducted for each mixing type. Each noisy mixture is manually mixed by adding a clean mixture of speech and music signals with an additive nonstationary noise. The noisy mixture has 7 levels of input SNR from 0dB to 30dB increased by 5dB.

In Fig.3 on the left-hand side; from 10dB and below, our proposed enhancement method gains an MSE improvement of twice over the observed noisy mixture. In the case with above 15dB, the MSE of the enhanced mixture is less than 0.1 and approaches 0 from 20dB onwards. This implies that the enhanced mixture progressively resembles the noise-free mixture. Hence, this allows the residual noise to be neglected in (5.26) and (5.40) in *Section 5.3*. In case of PESQ measurement in Fig.5.3 on the right-hand side, the average PESQ improvement of the enhanced mixture over the noisy mixture are 0.7, 0.3, and 0.2 for below 15dB, 20dB and 25dB input SNR, respectively. This translates into 28%, 8%, 4%, respectively. The enhanced mixture has significantly improved the noisy.

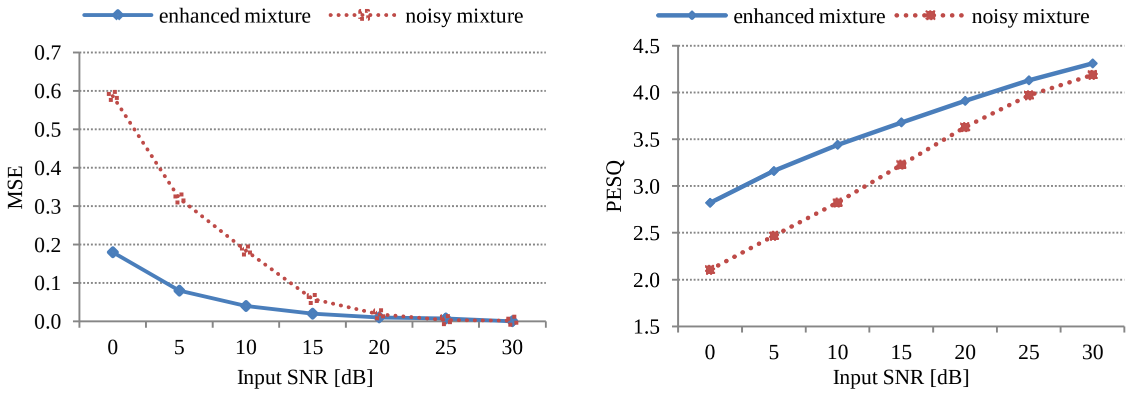
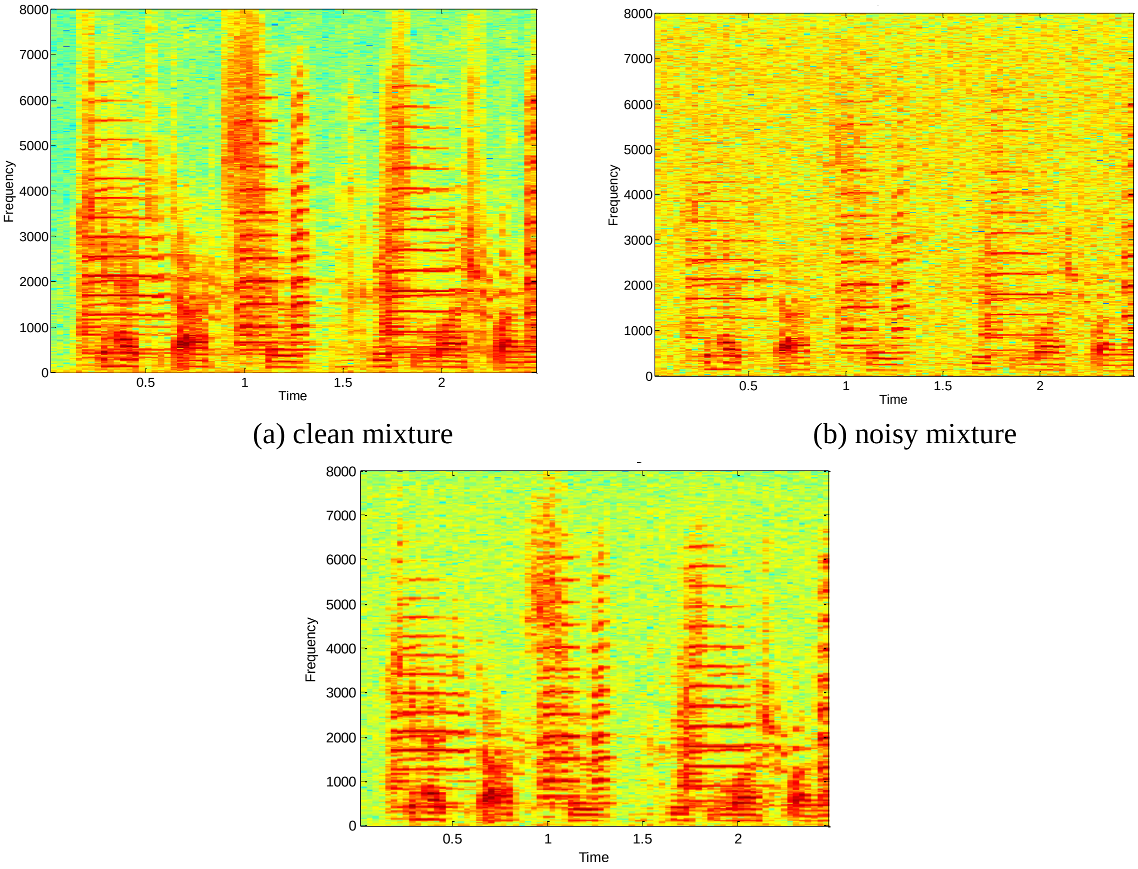


Figure 5.3: MSE (left) and PESQ (right) on mixtures of two signals and additive noises at different input SNRs.

A visual test has also been conducted by using mixing real-audio signals (speech + music) and an uncorrelated additive noise. A clean mixture of speech and musical signals is shown in Fig.5.4(a). A noisy mixture consists of the two audio signals and a white Gaussian noise with 5dB SNR. The enhanced mixture is obtained by applying the proposed enhancement method on the noisy mixture. Visually, an enhanced mixture in Fig.5.4(c) has efficiently extracted the signals spectrum compared with the noisy mixture in Fig.5.4(b).



(a) clean mixture

(b) noisy mixture

(c) enhanced mixture

Figure 5.4: Spectrograms of original clean mixture, clean mixture and additive white noise, noisy mixture enhanced using proposed iMMSE-STSA estimator.

### *5.4.3 Choice of $\zeta_M$ for Estimating $\tilde{\tilde{a}}_j(\tau)$*

The adaptive mixing attenuation estimator in (5.29) i.e. $\tilde{\tilde{a}}_j(\tau) = \zeta_M \tilde{\tilde{a}}_j(\tau - 1) + (1 - \zeta_M)\hat{\tilde{a}}_j(\tau)$ is weighted at every two consecutive frame through $\zeta_M$. To determine $\zeta_M$, 100 experiments have been conducted on 100 noise-free mixtures by implementing the proposed algorithm but excluded the enhancement step. Each noise-free mixture is simulated by adding two synthetic nonstationary AR signals. The nonstationary AR signal is synthesized by using the model (5.3) with $2.56s$ length which divided into five sections i.e. $T_1 = [0\,, 0.51s]$, $T_2 = (0.51s\,, 1.03s]$, $T_3 = (1.03s\,, 1.54s]$, $T_4 = (1.54s\,, 2.05s]$, and $T_5 = (2.05s\,, 2.56s]$, respectively. The term $a_{s_j}$ and $e_j(t)$ of $s_j(t)$ have been changed section by section. The samples of synthetic original signals are shown in Fig.5.5 in the top row.

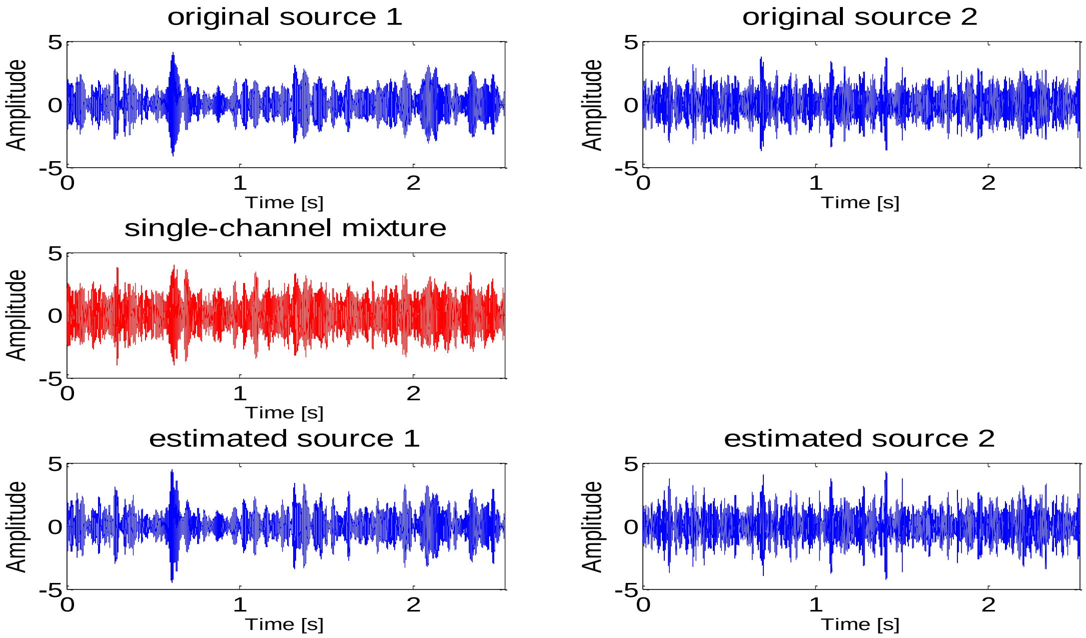


Figure 5.5: Two original signals, noise-free mixture and two estimated signals with $\zeta_M = 0.95$.

Firstly, the term $\zeta_M$ is tested on a range from $0.05$ to $0.95$ by $0.1$ increment. As a result, from $\zeta_M = 0.05$ to $\zeta_M = 0.85$, the average SDR results have increased slightly. Between $0.85 \leq \zeta_M \leq 0.95$ , the average SDR rises sharply with the average improvement of $3dB$ per signal. The term $\zeta_M$ is then further tested on [0.86,0.99] with $0.01$ increments and its results are illustrated in Fig.5.6. The highest average SDR is within the interval of $\zeta_M$ from $0.91$ to $0.98$. Hence the optimal choice of $\zeta_M$ will be within [0.91,0.98]. An example of $\tilde{\bar{a}}_1(\tau)$ against $\bar{a}_1(\tau)$ with different $\zeta_M$ values has been plotted in Fig.5.7. The term $\tilde{\bar{a}}_1(\tau)$ of $\zeta_M = 0.7$ has highly oscillatory values. Conversely, $\tilde{\bar{a}}_j(\tau)$ varies slowly and resembles a straight line when $\zeta_M = 0.99$ because $\tilde{\bar{a}}_j(\tau)$ at the $\tau^{th}$ frame depends 99% on its previous value. When $\zeta_M = 0.95$, $\tilde{\bar{a}}_j(\tau)$ tracks very closely with the true $\bar{a}_j(\tau)$. Hence, $\zeta_M$ has a crucial role in tracking the behavior of $\bar{a}_j(\tau)$.

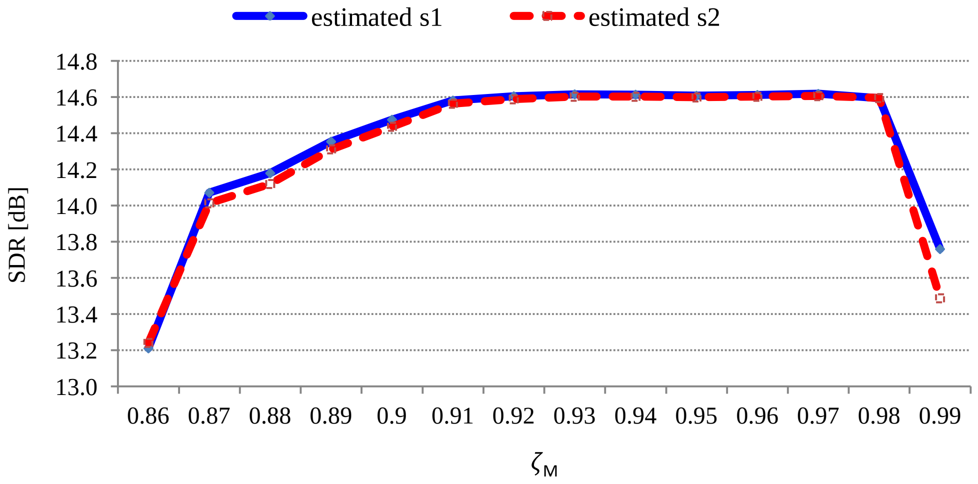


Figure 5.6: Average SDR on the noise-free mixture of two synthetic AR signals with various $\zeta_M$

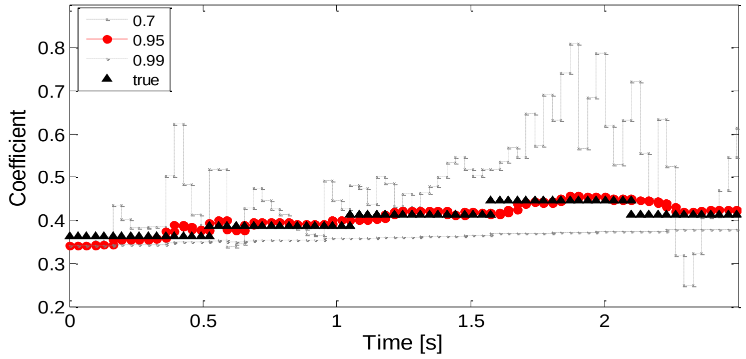


Figure 5.7: Mixing coefficients of $\bar{a}_1(\tau)$ (true) and $\tilde{\bar{a}}_1(\tau)$ for $\zeta_M = 0.7, 0.95\,, 0.99$

Although $\tilde{\bar{a}}_j(\tau)$ is an estimate of $\bar{a}_j(\tau)$, the separating performance of $\tilde{\bar{a}}_j(\tau)$ yields the same SDR as $\bar{a}_j(\tau)$ at $14.7dB$ and $14.9dB$ for $\hat{s}_1(t)$ and $\hat{s}_2(t)$, respectively. This is because the condition $G_{k=j}(\tau,\omega) < G_{k\neq j}(\tau,\omega)$ has been satisfied when $\left|\hat{C}_j(\tau)\right| < \left|\left(\tilde{\bar{a}}_k(\tau) - a_j(\tau) - \frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right)\right| + \left|\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right| + \frac{2}{1+|\gamma|}$ according to (5.40). The $\left|\hat{C}_j(\tau)\right|$ condition for $j = 1$ and $2$ have been computed and shown in Fig.8. For $j = 1\,,$ $\left|\hat{C}_1(\tau)\right| < \left|\left(\tilde{\bar{a}}_2(\tau) - a_1(\tau) - \frac{a_{s_1}(\delta;\tau)}{1+|\gamma|}\right)\right| + \left|\frac{a_{s_1}(\delta;\tau)}{1+|\gamma|}\right| + \frac{2}{1+|\gamma|}$, thus the $\left|\hat{C}_1(\tau)\right|$ condition is satisfied. For $j = 2$, the $\left|\hat{C}_2\right|$ condition is also true. Therefore, the cost function has correctly assigned all $(\tau,\omega)$ units to their respective original signals. This is clearly evident by the same SDR results between the $\tilde{\bar{a}}_j(\tau)$ and the $\bar{a}_j(\tau)$. Therefore, the term $\zeta_M$ has been selected around $0.95$ for all experiments.

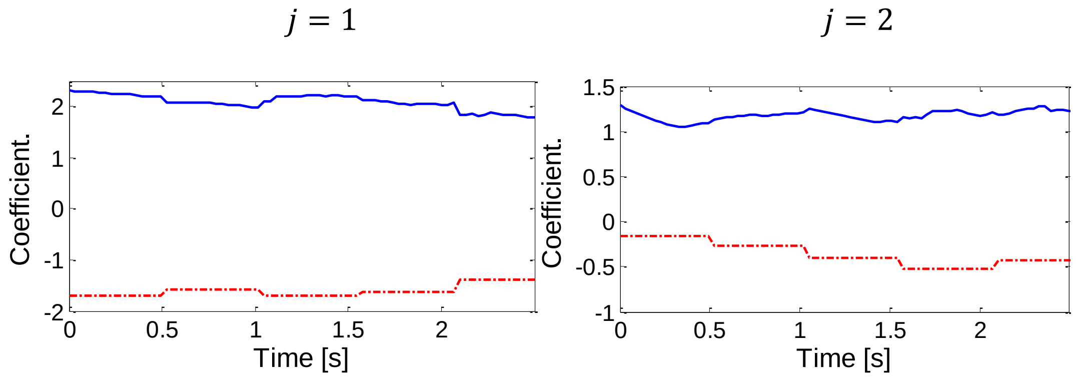


Figure 5.8: $\left|\hat{C}_j(\tau)\right|$ condition of $j = 1$ on the left plot and $j = 2$ on the plot where dot-dash line refers to $\left|\hat{C}_j(\tau)\right|$ and continuous line refers to $\left|\left(\tilde{\bar{a}}_k(\tau) - a_j(\tau) - \frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right)\right| + \left|\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right| + (2/1 + |\gamma|),\ j \neq k.$

### *5.4.4 Separation Performance*

The separation performance of the proposed method has been assessed by using 150 mixtures for three types of mixtures i.e. music + music, speech + music, and speech + speech. Mixtures were conducted in *Section 5.4.2*. The separation performance has been computed from the average of a hundred experiments of each of 150 mixtures for three mixing types. The proposed approach will be compared with the sparse nonnegative matrix 2-dimensional factorization (SNMF2D) and the single-channel independent component analysis (SCICA). The SNMF2D parameters are set as follows: the number of factors is 2, sparsity weight of 1.1, number of phase shift and time shift is 31 and 7, respectively for music. As for speech, both shifts are set to 4. The TF domain used in SNMF2D is based on the log-frequency spectrogram. Cost function of SNMF2D is based on the Kullback-Leibler divergence. As for the SCICA, the number of block is 10 with time delay set to unity.

In Fig.5.9, $\tilde{\tilde{\alpha}}_1(\tau)$ and $\tilde{\tilde{\alpha}}_2(\tau)$ change from frame to frame (this is natural as they correspond to speech and music signals, respectively). Examples of two audio signals with equal power, the additive noise, and the noisy mixture at 0dB SNR are shown in Fig.5.10 at the top and the second row. The SNR has been computed by $10\log_{10}(P_{signal}/P_{noise})$ where $P_{signal}$ and $P_{noise}$ denote a power of signal and a power of noise, respectively. Visually in Fig.5.10, the estimated signals (bottom) have been clearly separated when compared with the original signals (top). On the other hand, the estimated signals from SCICA and SNMF2D have not been well separated as shown in Figs.5.11 and 5.12, respectively.

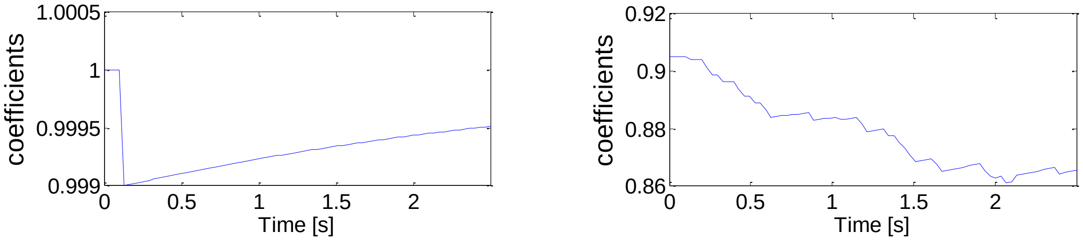


Figure 5.9: Estimated coefficients of $\tilde{\tilde{a}}_1(\tau)$ (left) and $\tilde{\tilde{a}}_2(\tau)$ (right).

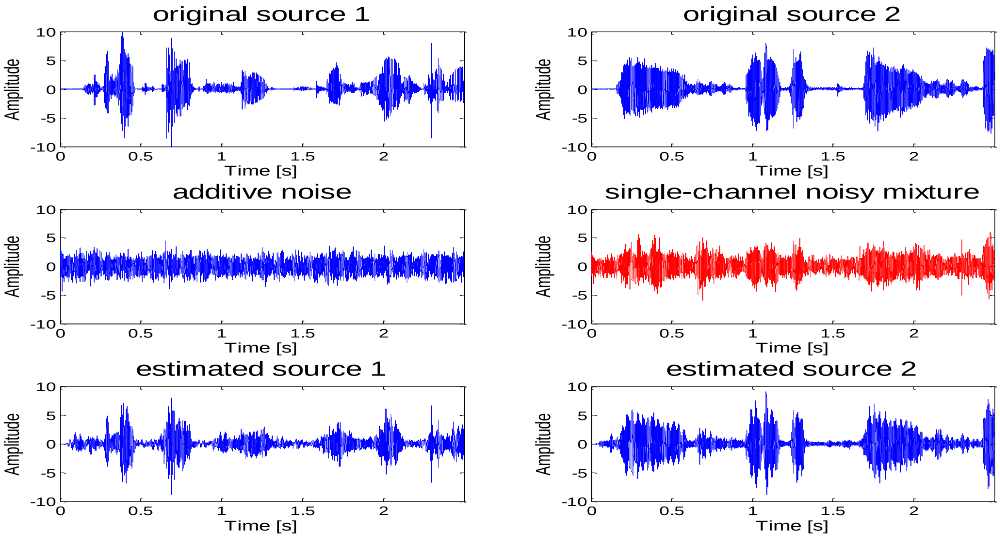


Figure 5.10: Two original signals, observed noisy mixture of 0dB SNR, and two estimated signals using the proposed method.

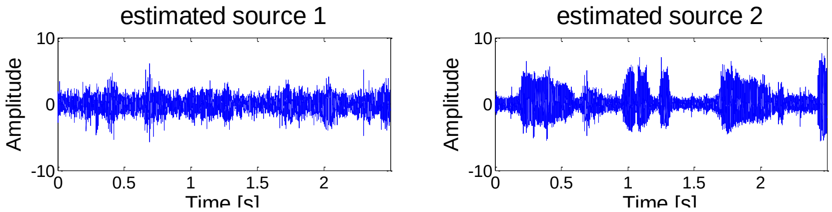


Figure 5.11: Two estimated signals using SCICA method.

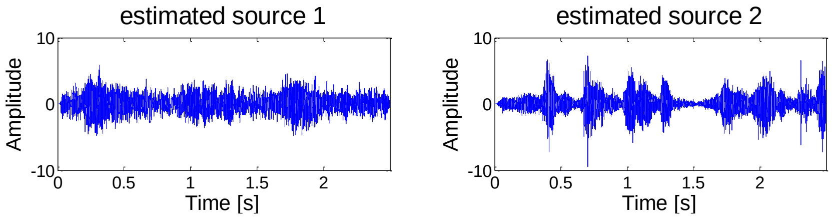


Figure 5.12: Two estimated signals using SNMF2D method.

The average SDR results using the proposed method for three mixing types with various inputs SNR have also been illustrated in Fig.5.13. As expected, the mixture of music +

music obtains the best separation performance followed by speech + music and speech + speech, respectively. The reasons are firstly the difference of AR coefficients between music and music is more distinct than the other two types. Secondly, the speech signals are highly nonstationary thus it is more difficult to separate than music. Additionally, the additive noise signals, i.e. babble and destroyer operations room background noises, have similar frequency components to speech components in which the spectrums of speech signal will be submerged by the noise signal.

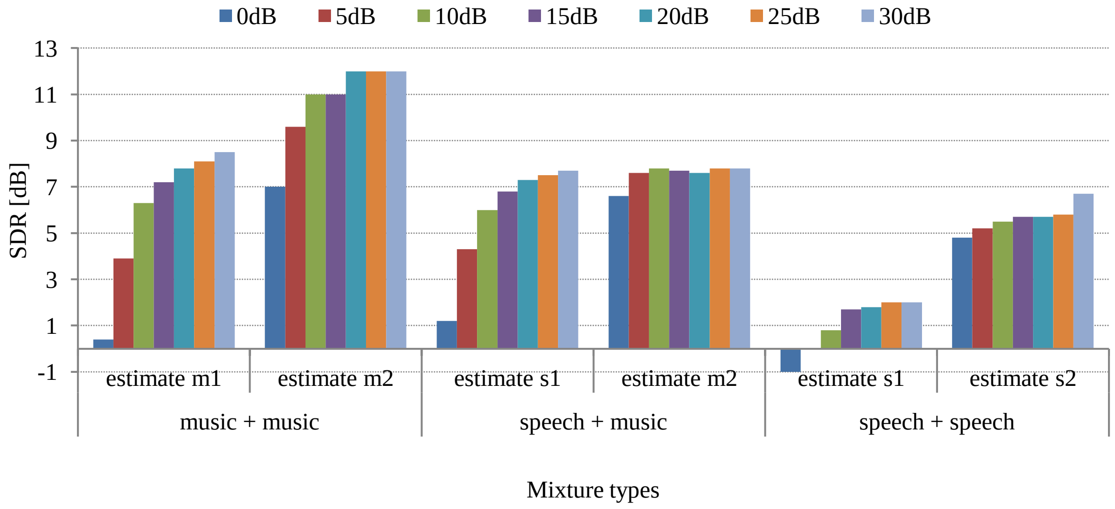


Figure 5.13: Average SDR performance of three mixing types with various input SNR using the proposed method.

Next, the separation performances of the proposed methods are compared with SCICA and SNMF2D shown in Fig.5.14. The proposed method shows better separation performance than SCICA and SNMF2D across input SNRs. The proposed method can well separate the noisy mixture while the SCICA and SNMF2D cannot when, in particular, the input SNR is below 15 dB. This is because the proposed method removes noise components and emphasises the signal components through the mixture enhancement step. For the SCICA and SNMF2D methods, their separation performances depend critically on signal information, given by the highly noisy mixture, thus these two methods are hampered by interference of noise. Fig.5.15 shows a comparison of SCICA, SNMF2D and the proposed method based on the mixing types. The proposed method renders the best separation performance of all mixture types among the three methods. Particularly in low input SNR i.e. below 15dB, the proposed method performs far superior than the SNMF2D and SCICA.

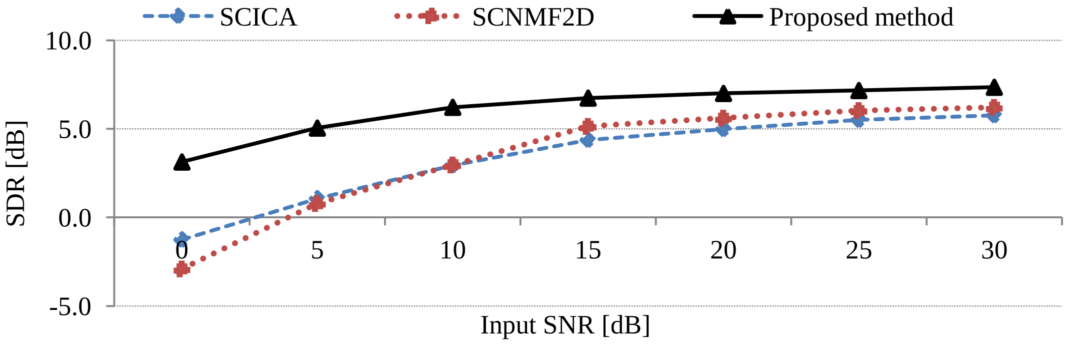
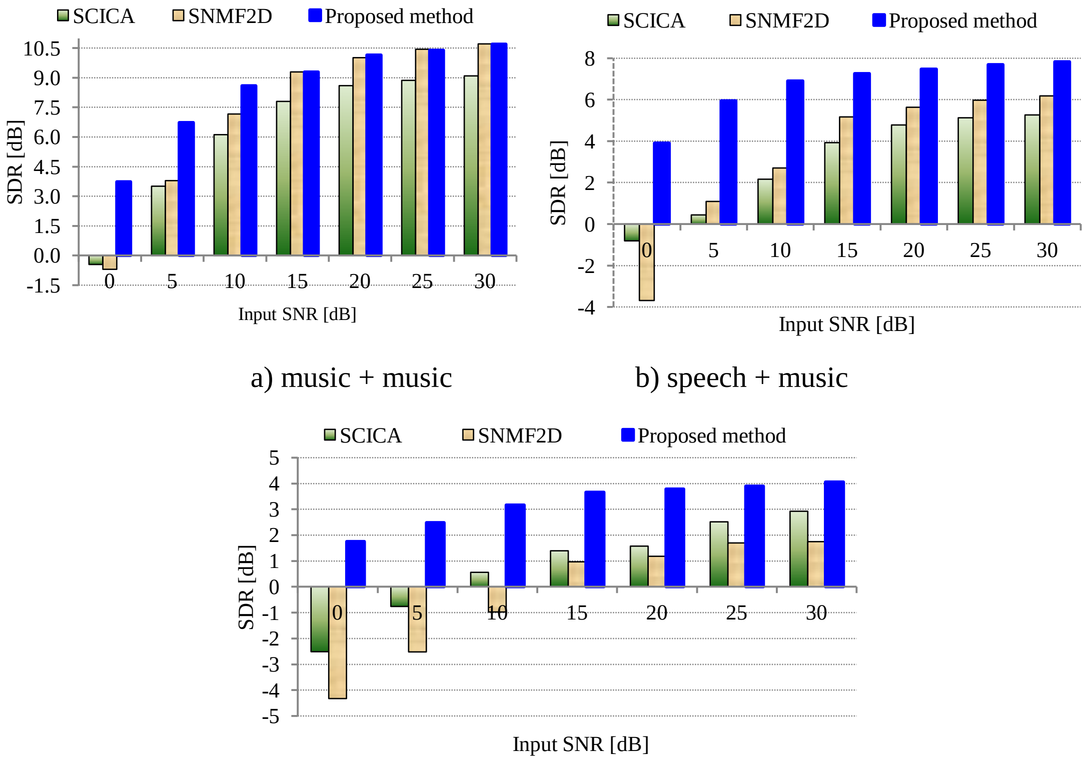


Figure 5.14: Comparison of average SDR performance among SCICA, SNMF2D and the proposed method.



a) music + music

b) speech + music

c) speech + speech

Figure 5.15: Comparison of average SDR performance of three mixing types with various input SNR between SNMF2D, SCICA, and the proposed method.

## 5.4 Summary

In this paper, a novel noisy single channel blind separation algorithm has been presented. The proposed method constructs a noisy pseudo-stereo mixture by time-delaying and weighting the observed mixture. The method assumes that the source signals are characterized as AR processes and the separability analysis of the pseudo-stereo mixture has been derived. The proposed method enhances the signals in the noisy mixing model and then separates the enhanced mixture. Furthermore, the conditions required for a unique mask construction from the maximum likelihood framework have also been identified. The proposed method has demonstrated a high level separation performance for real-audio signals in nonstationary noisy environment. The proposed method gains the desirable properties for the online applications: Firstly, it is able to adapt the parameter estimate frame-by-frame and separates the mixture given by small blocks. Secondly, it can separate the original signals from the high noisy mixture.

# CHAPTER 6

# SINGLE CHANNEL BLIND SOURCE SEPARATION USING MULTIPLE TIME DELAY PSEUDO – STEREO MIXTURE

In this chapter, the pseudo-stereo mixture is further extended by using multiple time delay of the observed mixture. Separability analysis of the proposed multiple times delay mixing model and the analysis of the difference between the new mixing model and the pseudo-stereo model from Chapter 3 Section 3.2 are articulated. As such, the new mixing model improves the pseudo-stereo mixture in term of increasing the difference of the mixing coefficients between signals, and reducing the residues of AR coefficients. As a result, the peaks in the histogram corresponding to each signal will be revealed wider apart from one another which are then be used for constructing a binary mask. Subsequently, the proposed multiple times delay model will improve the separation performance. Finally, experimental testing has been conducted on both syntheticed AR signals and real-audio signals to assess the proposed multiple time delay mixing model compared with the SOLO method.

The chapter is organized as follows: the ‘multi-time delay pseudo-stereo’ mixing model is developed in Section 6.1. Next, the separability of the proposed model is elucidated in Section 6.2. In Section 6.3, the separation method is presented. Then, the proposed multiple time delay mixing model is compared with the pseudo-stereo model in Section 6.4. Experimental results coupled with a series of performance comparison with the proposed multiple times delay mixing model compared with the SOLO method are presented in Section 6.5. Finally, Section 6.6 concludes this chapter.

## 6.1 Proposed Multiple Times Delay Pseudo-Stereo Mixture Model

The case of a mixture of two signals in time domain is considered as

$$x_1(t) = s_1(t) + s_2(t) \tag{6.1}$$

where $x_1(t)$ is the single channel mixture, and $s_1(t)$ and $s_2(t)$ are the original signals which are assumed to be modeled by the autoregressive (AR) process i.e. $s_j(t) = -\sum_{m=1}^{D_j} a_{s_j}(m;t)s_j(t-m) + e_j(t)$ where $a_{s_j}(m;t)$ denotes the $m^{th}$ order AR coefficient of the $j^{th}$ signal at time $t$, $D_j$ is the maximum AR order, and $e_j(t)$ is an independent identically distributed (i.i.d.) random signal with zero mean and variance $\sigma_{e_j}^2$. For simplicity, the pseudo-stereo mixture is formulated by two weighing; $\{\gamma_1, \gamma_2\} \in \Re$, and two time-shifting; $\{\delta_1, \delta_2\} \in \mathcal{I}$, the single channel mixture $x_1(t)$.

$$x_2(t) = \frac{x_1(t) + \gamma_1 x_1(t-\delta_1) + \gamma_2 x_1(t-\delta_2)}{1+|\gamma_1|+|\gamma_2|} \tag{6.2}$$

The mixture in (6.1) and (6.2) is termed as “multiple time delay (MTD) pseudo-stereo”. Eq.(6.2) can be expressed in terms of the original signals, AR coefficients and time-delays as

$$\begin{aligned} x_2(t) &= \frac{x_1(t) + \gamma_1 x_1(t-\delta_1) + \gamma_2 x_1(t-\delta_2)}{1+|\gamma_1|+|\gamma_2|} \\ &= \frac{s_1(t)+s_2(t)+\gamma_1[s_1(t-\delta_1)+s_2(t-\delta_1)\,]+\gamma_2[s_1(t-\delta_2)+s_2(t-\delta_2)\,]}{1+|\gamma_1|+|\gamma_2|} \\ &= \frac{-\sum_{m=1}^{D_1} a_{s_1}(m)s_1(t-m) + e_1(t)}{1+|\gamma_1|+|\gamma_2|} + \frac{\gamma_1 s_1(t-\delta_1)}{1+|\gamma_1|+|\gamma_2|} + \frac{\gamma_2 s_1(t-\delta_2)}{1+|\gamma_1|+|\gamma_2|} + \end{aligned}$$

$$\frac{-\sum_{m=1}^{D_1} a_{s_2}(m)s_2(t-m)+e_2(t)}{1+|\gamma_1|+|\gamma_2|}+\frac{\gamma_1 s_2(t-\delta_1)}{1+|\gamma_1|+|\gamma_2|}+\frac{\gamma_2 s_2(t-\delta_2)}{1+|\gamma_1|+|\gamma_2|}$$

$$=\frac{\left(\gamma_1-a_{s_1}(\delta_1)\right)}{1+|\gamma_1|+|\gamma_2|}s_1(t-\delta_1)+\frac{\left(\gamma_2-a_{s_1}(\delta_2)\right)}{1+|\gamma_1|+|\gamma_2|}s_1(t-\delta_2)-\frac{\sum_{\substack{m=1\\ m\neq\delta_1,\delta_2}}^{D_1} a_{s_1}(m;t)s_1(t-m)+e_1(t)}{1+|\gamma_1|+|\gamma_2|}+$$

$$\frac{\left(\gamma_1-a_{s_2}(\delta_1)\right)}{1+|\gamma_1|+|\gamma_2|}s_2(t-\delta_1)+\frac{\left(\gamma_2-a_{s_2}(\delta_2)\right)}{1+|\gamma_1|+|\gamma_2|}s_2(t-\delta_2)-\frac{\sum_{\substack{m=1\\ m\neq\delta_1,\delta_2}}^{D_2} a_{s_2}(m;t)s_2(t-m)+e_2(t)}{1+|\gamma_1|+|\gamma_2|} \quad (6.3)$$

Define

$$a_{ij}(t;\{\delta_1,\delta_2\},\{\gamma_1,\gamma_2\})=\frac{\gamma_i-a_{s_j}(\delta_i)}{1+\sum_{p=1}^{P}\left|\gamma_p\right|} \quad (6.4)$$

$$r_j(t;\{\delta_1,\delta_2\},\{\gamma_1,\gamma_2\})=\frac{e_j(t)-\sum_{\substack{m=1\\ m\neq\delta_1,\delta_2}}^{D_j} a_{s_j}(m;t)s_j(t-m)}{1+\sum_{p=1}^{P}\left|\gamma_p\right|} \quad (6.5)$$

where $a_{ij}(t;\{\delta_1,\delta_2\},\{\gamma_1,\gamma_2\})$ and $r_j(t;\{\delta_1,\delta_2\},\{\gamma_1,\gamma_2\})$ represent the $i^{th}|_{i=1,2}$ mixing attenuation and residue of the $j^{th}$ signal, respectively, and $p=1,2$ denotes the index of the $\{\gamma_p\}_{p=1,2}$ and $\{\delta_p\}_{p=1,2}$ parameters. Notice that: $a_{ij}(t)$ and $r_j(t)$ will be used for $a_{ij}(t;\{\delta_1,\delta_2\},\{\gamma_1,\gamma_2\})$ and $r_j(t;\{\delta_1,\delta_2\},\{\gamma_1,\gamma_2\})$, respectively, to further facilitate. The parameterization of $a_{ij}(t)$ and $r_j(t)$ depends on $\{\delta_1,\delta_2\}$ and $\{\gamma_1,\gamma_2\}$ although this is not shown explicitly. The proposed MTD mixture contains an extra mixing attenuation which causes less the residue of the MTD mixture compared with the pseudo-stereo mixture. As a result of using (6.4) and (6.5), the MTD mixing model of two signals; $s_1(t)$ and $s_2(t)$, can be expressed in time domain as

$$x_1(t)=s_1(t)+s_2(t)$$

$$x_2(t)=a_{11}(t)s_1(t-\delta_1)+a_{21}(t)s_1(t-\delta_2)+r_1(t)$$

$$+a_{12}(t)s_2(t-\delta_1)+a_{22}(t)s_2(t-\delta_2)+r_2(t) \quad (6.6)$$

The assumptions are required as in the previous chapters which are recapped as follows: **Assumption 1**: The source signals satisfy the local stationarity of the time-frequency representation: $S_j(\tau-\phi,\omega)\approx S_j(\tau,\omega)$ . **Assumption 2**: phase ambiguity. Phase ambiguity can be avoided by satisfying the following condition: $\delta_p^{max}<\frac{f_s}{2f_{max}}$, where $p$ indicates the index of the time-delay. **Assumption 3**: The source signals are modelled as quasi-stationary where the AR parameters in AR process are stationary within a block but can change from block to block. **Assumption 4**: The source signals satisfy the windowed-disjoint orthogonality (WDO) condition; $S_i(\tau,\omega)S_j(\tau,\omega)\approx 0$ , $\forall i\neq j$ , $\forall\tau,\omega$.

Based on the above assumptions, the TF representation of the MTD mixing model is obtained using the STFT of $x_j(t)$, $j=1{,}2$ as

$$X_1(\tau,\omega)=S_1(\tau,\omega)+S_2(\tau,\omega)$$

$$X_2(\tau,\omega)=a_{11}(\tau)e^{-i\omega\delta_1}S_1(\tau-\delta_1,\omega)+a_{21}(\tau)e^{-i\omega\delta_2}S_1(\tau-\delta_2,\omega)+$$

$$a_{12}(\tau)e^{-i\omega\delta_1}S_2(\tau-\delta_1,\omega)+a_{22}(\tau)e^{-i\omega\delta_2}S_2(\tau-\delta_2,\omega)-$$

$$\left(\sum_{\substack{m=1\\m\neq\delta_1,\delta_2}}^{D_1}\frac{a_{s_1}(m;\tau)}{1+|\gamma_1|+|\gamma_2|}e^{-i\omega m}\,S_1(\tau-m,\omega)+\sum_{\substack{m=1\\m\neq\delta_1,\delta_2}}^{D_2}\frac{a_{s_2}(m;\tau)}{1+|\gamma_1|+|\gamma_2|}e^{-i\omega m}\,S_2(\tau-m,\omega)\right)\quad(6.7)$$

for $\forall\tau,\omega$. In (6.7), $e_j(t)$ can be negligible based on the fact that $e_j(t)\ll s_j(t)$, thus the TF of $r_j(t)$ in (5) simplifies to

$$R_j(\tau,\omega)=-\sum_{\substack{m=1\\m\neq\delta_1,\delta_2}}^{D_j}\frac{a_{s_j}(m;\tau)}{1+\sum_{p=1}^{2}|\gamma_p|}e^{-i\omega m}S_j(\tau-m,\omega)\qquad(6.8)$$

To facilitate further analysis, $C_j(\tau,\omega)$ is defined as

$$C_j(\tau,\omega) = \sum_{\substack{m=1\\ m\neq\delta_1,\delta_2}}^{D_j} \frac{a_{s_j}(m;\tau)}{1+\sum_{p=1}^{2}\left|\gamma_p\right|} e^{-i\omega m} \tag{6.9}$$

which forms a part of $R_j(\tau,\omega)$ without the contribution of the signal $S_j(\tau,\omega)$. From (6.7), it can be seen that $a_{ij}e^{-i\omega\delta}$ and $C_j(\tau,\omega)$ represent the signature of the $j^{th}$ signal which can be used for recovering the original signals. Next, the separability of the proposed MTD mixing model will be analyzed in the following section.

## 6.2 Separability of Multiple Times Delay Pseudo–Stereo Mixing Model

The separability of the proposed mixing model in (6.6) can be examined from the MTD mixture by considering $a_{ij}(t)$ and $r_j(t)$ and evaluating the following MTD cost function:

$$J(\tau,\omega) = \underset{k}{\operatorname{argmin}}\left| \bar{a}_k(\tau,\omega)X_1(\tau,\omega) - \left(\frac{1+\gamma_1 e^{-i\omega\delta_1}+\gamma_2 e^{-i\omega\delta_2}}{1+\sum_{p=1}^{2}\left|\gamma_p\right|}\right)X_1(\tau,\omega)\right|^2 \tag{6.10}$$

where

$$\bar{a}_k(\tau,\omega) = a_{1k}(\tau)e^{-i\omega\delta_1} + a_{2k}(\tau)e^{-i\omega\delta_2} - C_k(\tau,\omega) \tag{6.11}$$

with $a_k(\tau)$ and $C_k(\tau,\omega)$ are defined in (6.5) and (6.9). The MTD cost function (6.10) is derived by following similar steps applied to the cost function in Chapter 3 Section 3.3. Here, (6.10) can be further derived in term of the $j^{th}$ source signal by using the observed mixture where $X_1(\tau,\omega) = S_j(\tau,\omega)$ as follow:

$$J(\tau,\omega) = \underset{k}{\operatorname{argmin}}\left| \bar{a}_k(\tau,\omega)S_j(\tau,\omega) - \left(\frac{1+\gamma_1 e^{-i\omega\delta_1}+\gamma_2 e^{-i\omega\delta_2}}{1+\sum_{p=1}^{2}\left|\gamma_p\right|}\right)S_j(\tau,\omega)\right|^2$$

$$= \underset{k}{\operatorname{argmin}}\left| \left[a_{1k}(\tau)e^{-i\omega\delta_1} + a_{2k}(\tau)e^{-i\omega\delta_2} - C_k(\tau,\omega)\right]S_j(\tau,\omega)\right.$$

$$-\left(\frac{S_j(\tau,\omega)+\gamma_1 e^{-i\omega\delta_1}S_j(\tau,\omega)+\gamma_2 e^{-i\omega\delta_2}S_j(\tau,\omega)}{1+\sum_{p=1}^{2}\left|\gamma_p\right|}\right)\Bigg|^2$$

$$= \underset{k}{\operatorname{argmin}}\Bigg| \left[a_{1k}(\tau)e^{-i\omega\delta_1} + a_{2k}(\tau)e^{-i\omega\delta_2} - C_k(\tau,\omega)\right]S_j(\tau,\omega)$$

$$-\frac{S_j(\tau,\omega)}{1+\sum_{p=1}^{2}\left|\gamma_p\right|} - \frac{\gamma_1 e^{-i\omega\delta_1}S_j(\tau,\omega)}{1+\sum_{p=1}^{2}\left|\gamma_p\right|} - \frac{\gamma_2 e^{-i\omega\delta_2}S_j(\tau,\omega)}{1+\sum_{p=1}^{2}\left|\gamma_p\right|}\Bigg|^2$$

$$= \underset{k}{\operatorname{argmin}}\Bigg| \left[a_{1k}(\tau)e^{-i\omega\delta_1} + a_{2k}(\tau)e^{-i\omega\delta_2} - C_k(\tau,\omega)\right]S_j(\tau,\omega)$$

$$+\frac{\sum_{m=1}^{D_j} a_{s_j}(m;\tau)e^{-iwm}S_j(\tau-m,\omega)}{1+\sum_{p=1}^{2}\left|\gamma_p\right|} - \frac{\gamma_1 e^{-i\omega\delta_1}S_j(\tau,\omega)}{1+\sum_{p=1}^{2}\left|\gamma_p\right|} - \frac{\gamma_2 e^{-i\omega\delta_2}S_j(\tau,\omega)}{1+\sum_{p=1}^{2}\left|\gamma_p\right|}\Bigg|^2$$

$$= \underset{k}{\operatorname{argmin}}\Bigg| \left[a_{1k}(\tau)e^{-i\omega\delta_1} + a_{2k}(\tau)e^{-i\omega\delta_2} - C_k(\tau,\omega)\right]S_j(\tau,\omega)$$

$$+\frac{\sum_{\substack{m=1\\ m\neq\delta_1,\delta_2}}^{D_j} a_{s_j}(m;\tau)e^{-iwm}S_j(\tau-m,\omega)}{1+\sum_{p=1}^{2}\left|\gamma_p\right|} - \frac{\gamma_1 e^{-i\omega\delta_1}S_j(\tau,\omega)+a_{s_j}(\delta_1;\tau)e^{-iw\delta_1}S_j(\tau-\delta_1,\omega)}{1+\sum_{p=1}^{2}\left|\gamma_p\right|} -$$

$$\frac{\gamma_2 e^{-i\omega\delta_2}S_j(\tau,\omega)+a_{s_j}(\delta_2;\tau)e^{-iw\delta_2}S_j(\tau-\delta_2,\omega)}{1+\sum_{p=1}^{2}\left|\gamma_p\right|}\Bigg|^2$$

$$= \underset{k}{\operatorname{argmin}}\Bigg| \left[a_{1k}(\tau)e^{-i\omega\delta_1} + a_{2k}(\tau)e^{-i\omega\delta_2} - C_k(\tau,\omega)\right]S_j(\tau,\omega)$$

$$+\sum_{\substack{m=1\\ m\neq\delta_1,\delta_2}}^{D_j} \frac{a_{s_j}(m;\tau)e^{-iwm}}{1+\sum_{p=1}^{2}\left|\gamma_p\right|}S_j(\tau,\omega) - \frac{(\gamma_1+a_{s_j}(\delta_1;\tau))}{1+\sum_{p=1}^{2}\left|\gamma_p\right|}e^{-i\omega\delta_1}S_j(\tau,\omega) - \frac{(\gamma_2+a_{s_j}(\delta_2;\tau))}{1+\sum_{p=1}^{2}\left|\gamma_p\right|}e^{-i\omega\delta_2}S_j(\tau,\omega)\Bigg|^2$$

$$= \underset{k}{\operatorname{argmin}}\Bigg| \left[a_{1k}(\tau)e^{-i\omega\delta_1} + a_{2k}(\tau)e^{-i\omega\delta_2} - C_k(\tau,\omega)\right]S_j(\tau,\omega)$$

$$+\left[\sum_{\substack{m=1\\ m\neq\delta_1,\delta_2}}^{D_j} \frac{a_{s_j}(m;\tau)e^{-iwm}}{1+\sum_{p=1}^{2}\left|\gamma_p\right|} - \frac{(\gamma_1+a_{s_j}(\delta_1;\tau))}{1+\sum_{p=1}^{2}\left|\gamma_p\right|}e^{-i\omega\delta_1} - \frac{(\gamma_2+a_{s_j}(\delta_2;\tau))}{1+\sum_{p=1}^{2}\left|\gamma_p\right|}e^{-i\omega\delta_2}\right]S_j(\tau,\omega)\Bigg|^2$$

$$= \underset{k}{\operatorname{argmin}}\Big| \left[a_{1k}(\tau)e^{-i\omega\delta_1} + a_{2k}(\tau)e^{-i\omega\delta_2} - C_k(\tau,\omega)\right]$$

$$-\left[a_{1j}(\tau)e^{-i\omega\delta_1} + a_{2j}(\tau)e^{-i\omega\delta_2} - C_j(\tau,\omega)\right]\Big|^2\left|S_j(\tau,\omega)\right|^2 \qquad (6.12)$$

The MTD cost function in (6.12) will be able to distinguish the $k$ arguments by yielding a zero value for $k=j$ and nonzero value for $k\neq j$. The cost function can separate the $k$ arguments due to the difference of $a_{1k}$ and $a_{1j}$, or $a_{2k}$ and $a_{2j}$, or $C_k(\tau,\omega)$ and $C_j(\tau,\omega)$. Based on this fact, the proposed MTD mixing model can be separated by using at least one pair of the different coefficient parameters, for example $a_{1k}\neq a_{1j}$, which can be considered through $a_{ij}(t)$ and $r_j(t)$ in the following cases:

Case I: If $a_{11}(t) \neq a_{12}(t)$, $a_{21}(t) = a_{22}(t) = a_2(t)$ and $r_1(t) = r_2(t) = r(t)$, then

$$x_2(t) = \left(\frac{\gamma_1 - a_{s_j}(\delta_1;t)}{1+\sum_{p=1}^{2}|\gamma_p|}\right) x_1(t-\delta_1) + \left(\frac{\gamma_2 - a_s(\delta_2;t)}{1+\sum_{p=1}^{2}|\gamma_p|}\right) x_1(t-\delta_2) + 2r(t)$$

Case II: If $a_{11}(t) = a_{12}(t) = a_1(t)$, $a_{21}(t) \neq a_{22}(t)$ and $r_1(t) = r_2(t) = r(t)$, then

$$x_2(t) = \left(\frac{\gamma_1 - a_s(\delta_1;t)}{1+\sum_{p=1}^{2}|\gamma_p|}\right) x_1(t-\delta_1) + \left(\frac{\gamma_2 - a_{s_j}(\delta_2;t)}{1+\sum_{p=1}^{2}|\gamma_p|}\right) x_1(t-\delta_2) + 2r(t).$$

Case III: If $a_{11}(t) = a_{12}(t) = a_1(t)$, $a_{21}(t) = a_{22}(t) = a_2(t)$ and $r_1(t) \neq r_2(t)$, then

$$x_2(t) = \left(\frac{\gamma_1 - a_s(\delta_1;t)}{1+\sum_{p=1}^{2}|\gamma_p|}\right) x_1(t-\delta_1) + \left(\frac{\gamma_2 - a_s(\delta_2;t)}{1+\sum_{p=1}^{2}|\gamma_p|}\right) x_1(t-\delta_2) + r_1(t;\delta,\gamma) + r_2(t;\delta,\gamma).$$

Since the term $a_{ij}(t)$ and $r_j(t)$ is related to the $j^{th}$ signal via $\frac{\gamma_i - a_{s_j}(\delta_i;t)}{1+\sum_{p=1}^{2}|\gamma_p|}$ and $C_j(\tau,\omega) = \sum_{\substack{m=1 \\ m\neq\delta_1,\delta_2}}^{D_j} \frac{a_{s_j}(m;\tau)}{1+\sum_{p=1}^{2}|\gamma_p|} e^{-i\omega m}$, respectively, the separability of the mixture of the above three cases can be sequentially analyzed using the MTD cost function in (6.12) as

Case I:
$$\begin{aligned} J(\tau,\omega) &= \underset{k}{\operatorname{argmin}} \Big| \left[a_{1k}(\tau)e^{-i\omega\delta_1} + a_2(\tau)e^{-i\omega\delta_2} - C(\tau,\omega)\right] \\ &\qquad - \left[a_{1j}(\tau)e^{-i\omega\delta_1} + a_2(\tau)e^{-i\omega\delta_2} - C(\tau,\omega)\right]\Big|^2 \left|S_j(\tau,\omega)\right|^2 \\ &= \underset{k}{\operatorname{argmin}} \left|a_{1k}(\tau)e^{-i\omega\delta_1} - a_{1j}(\tau)e^{-i\omega\delta_1}\right|^2 \left|S_j(\tau,\omega)\right|^2 \end{aligned} \tag{6.13}$$

Case II:
$$\begin{aligned} J(\tau,\omega) &= \underset{k}{\operatorname{argmin}} \Big| \left[a_1(\tau)e^{-i\omega\delta_1} + a_{2k}(\tau)e^{-i\omega\delta_2} - C(\tau,\omega)\right] \\ &\qquad - \left[a_1(\tau)e^{-i\omega\delta_1} + a_{2j}(\tau)e^{-i\omega\delta_2} - C(\tau,\omega)\right]\Big|^2 \left|S_j(\tau,\omega)\right|^2 \\ &= \underset{k}{\operatorname{argmin}} \left|a_{2k}(\tau)e^{-i\omega\delta_2} - a_{2j}(\tau)e^{-i\omega\delta_2}\right|^2 \left|S_j(\tau,\omega)\right|^2 \end{aligned} \tag{6.14}$$

Case III:
$$J(\tau,\omega) = \underset{k}{\operatorname{argmin}} \Big| \left[a_1(\tau)e^{-i\omega\delta_1} + a_2(\tau)e^{-i\omega\delta_2} - C_k(\tau,\omega)\right]$$

$$-\left[a_1(\tau)e^{-i\omega\delta_1}+a_2(\tau)e^{-i\omega\delta_2}-C_j(\tau,\omega)\right]\Big|^2\left|S_j(\tau,\omega)\right|^2$$

$$=\underset{k}{\operatorname{argmin}}\left|C_k(\tau,\omega)-C_j(\tau,\omega)\right|^2\left|S_j(\tau,\omega)\right|^2 \qquad (6.15)$$

Eqs.(6.13 – 6.15) yield nonzero value for $k \neq j$ thus these cost functions are able to distinguish the $k$ arguments. Therefore, the mixtures of Case I - III are separable. Case I - III denote different signals but setting $\{\gamma_1,\gamma_2\}$ and $\{\delta_1,\delta_2\}$ for the MTD mixture such that at least one pair of coefficient parameters differs.

On the other hands, if the coefficient parameters in (6.12) are not different, this can then be expressed as

Case IV: $a_{11}(t)=a_{12}(t)=a_1(t)$ , $a_{21}(t)=a_{22}(t)=a_2(t)$ and $r_1(t)=r_2(t)=r(t)$, then $x_2(t)=\left(\frac{\gamma_1-a_s(\delta_1;t)}{1+\sum_{p=1}^{P}|\gamma_p|}\right)x_1(t-\delta_1)+\left(\frac{\gamma_2-a_s(\delta_2;t)}{1+\sum_{p=1}^{P}|\gamma_p|}\right)x_1(t-\delta_2)+2r((t))$.

Case IV refers to identical signals mixed in the single channel. In this case, there is no benefit achieved at all. The second mixture is simply formulated as time-delayed of the first mixture multiply by a scalar plus the redundant residue. The separability of this case is presented by substituting the MTD mixture of Case IV into the cost function (6.12). Since both residues are equal, then $C_1(\tau,\omega)=C_2(\tau,\omega)=C(\tau,\omega)=\sum_{\substack{m=1\\ m\neq\delta_1,\delta_2}}^{D}\frac{a_s(m;\tau)e^{-iwm}}{1+\sum_{p=1}^{2}|\gamma_p|}$. For Case IV, the cost function becomes:

$$J(\tau,\omega)=\underset{k}{\operatorname{argmin}}\Big|\left[a_1(\tau)e^{-i\omega\delta_1}+a_2(\tau)e^{-i\omega\delta_2}-C(\tau,\omega)\right]$$

$$-\left[a_1(\tau)e^{-i\omega\delta_1}+a_2(\tau)e^{-i\omega\delta_2}-C(\tau,\omega)\right]\Big|^2\left|S_j(\tau,\omega)\right|^2$$

$$=0 \qquad \text{for } \forall k. \qquad (6.16)$$

As a result, the cost function $J(\tau,\omega)$ is zero for all $k$ arguments i.e. $J_1=J_2=0$ thus, the MTD cost function cannot distinguish the $k$ arguments, the mixture is not separable.

According to the above four cases, the proposed MTD pseudo-stereo mixing model is separable when two signals are different and $\{\gamma_1, \gamma_2, \delta_1, \delta_2\}$ are approximately set for the MTD mixture such that at least one pair of coefficient parameters i.e. $a_{1j}(\tau), a_{2j}(\tau)$, and $C_j(\tau, \omega)$ differs.

## 6.3 Proposed Separation Method

### *6.3.1 Parameter Estimation using Complex 2D Histogram*

The TF representation of the MTD mixing model in (6.7) is assumed that the $j^{th}$ signal is dominant at a particular TF unit which can be expressed as

$$\begin{aligned} X_1(\tau,\omega) &= S_j(\tau,\omega) \\ X_2(\tau,\omega) &= a_{1j}(\tau)e^{-i\omega\delta_1}S_j(\tau-\delta_1,\omega) + a_{2j}(\tau)e^{-i\omega\delta_2}S_j(\tau-\delta_2,\omega) - \\ &\qquad \left(\sum_{\substack{m=1 \\ m\neq\delta_1,\delta_2}}^{D_j} \frac{a_{s_j}(m;\tau)}{1+|\gamma_1|+|\gamma_2|} e^{-i\omega m}\, S_j(\tau-m,\omega)\right) \\ &= \left[a_{1j}(\tau)e^{-i\omega\delta_1} + a_{2j}(\tau)e^{-i\omega\delta_2} - \sum_{\substack{m=1 \\ m\neq\delta_1,\delta_2}}^{D_j} \frac{a_{s_j}(m;\tau)}{1+|\gamma_1|+|\gamma_2|} e^{-i\omega m}\right] S_j(\tau,\omega) \end{aligned} \tag{6.17}$$

for $\{\delta_1, \delta_2, m\} \leq \phi$, $(\tau,\omega) \in \Omega_j$ and $\Omega_j$ is the active area of $S_j(\tau,\omega)$ defined as

$$\Omega_j := \{(\tau,\omega): S_j(\tau,\omega) \neq 0, \ \forall k \neq j\} \tag{6.18}$$

The estimate of $\bar{a}_j(\tau,\omega) = a_{1j}(\tau)e^{-i\omega\delta_1} + a_{2j}(\tau)e^{-i\omega\delta_2} - C_j(\tau,\omega)$ associated to the $j^{th}$ signal can be determined as

$$\bar{a}_j^{MTD}(\tau,\omega) = \frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}$$

$$= a_{1j}(\tau)e^{-i\omega\delta_1} + a_{2j}(\tau)e^{-i\omega\delta_2} - C_j(\tau,\omega)$$

$$= \bar{a}_j^{(r)}(\tau,\omega) + i\bar{a}_j^{(i)}(\tau,\omega) \ , \quad \forall(\tau,\omega) \in \Omega_j \qquad (6.19)$$

where $\bar{a}_j^{(r)}(\tau,\omega) = Re\left[\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}\right]$, $\bar{a}_j^{(i)}(\tau,\omega) = Im\left[\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}\right]$ are the real and imaginary parts of $\bar{a}_j(\tau,\omega)$, respectively, and $i = \sqrt{-1}$. Notice that the mixing coefficient of the MTD mixing model is without the factor $e^{i\omega\delta}$ comapared with the mixing coefficient of the pseudo-stereo mixing model in (4.7) (Chapter 4 Section 4.2.1) i.e. $\bar{a}_j(\tau,\omega)$ $=\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}e^{i\omega\delta}$. The proposed weighted complex 2-dimentional (2D) histogram in Chapter 4 is reformulated for estimating $\bar{a}_j^{MTD}(\tau,\omega)$. The proposed complex 2D histogram of the MTD mixing model can be expressed as

$$\hat{\bar{a}}_j^{(r)} = \frac{\sum_{\tau,\omega}|X_1(\tau,\omega)X_2(\tau,\omega)|Re\left[\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}\right]}{\sum_{\tau,\omega}|X_1(\tau,\omega)X_2(\tau,\omega)|}$$

$$\hat{\bar{a}}_j^{(i)} = \frac{\sum_{\tau,\omega}|X_1(\tau,\omega)X_2(\tau,\omega)|Im\left[\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}\right]}{\sum_{\tau,\omega}|X_1(\tau,\omega)X_2(\tau,\omega)|} \qquad (6.20)$$

The above can then be combined to form the estimate of (6.19) as $\hat{a}_j = \hat{\bar{a}}_j^{(r)} + i\hat{\bar{a}}_j^{(i)}$. This expression can be expressed by using the similar idea to that expressed in (6.19) as;

$$\hat{\bar{a}}_j = \hat{a}_{1j} + \hat{a}_{2j} - \hat{C}_j \qquad (6.21)$$

where $\hat{a}_{ij}$ and $\hat{C}_j$ are the complex 2D histogram estimates of $a_{ij}(\tau)$ and $C_j(\tau,\omega)$, respectively.

### *6.3.2 Construction of Masks*

In this section, the binary TF masks will be established by using $X_1(\tau,\omega)$ alone. The binary TF masks can be constructed by labeling each TF unit with the $k$ argument through maximizing the following instantaneous likelihood function given by (4.13) in Chapter 4 Section 4.2.2:

$$G(\tau,\omega) = \underset{k}{\operatorname{argmin}} |\hat{\hat{a}}_k X_1(\tau,\omega) - X_2(\tau,\omega)|^2 \qquad (6.22)$$

To substitute $X_2(\tau,\omega)$ with $X_1(\tau,\omega)$, the proposed MTD mixture in (6.2) is derived in term of the $j^{th}$ signal as

$$x_2(t) = \frac{s_j(t)+\gamma_1 s_j(t-\delta_1)+\gamma_2 s_j(t-\delta_2)}{1+|\gamma_1|+|\gamma_2|} \qquad (6.23)$$

The TF representation of (6.23) can be expressed by using STFT as

$$X_2(\tau,\omega) = \frac{S_j(\tau,\omega)+\gamma_1 e^{-i\omega\delta_1} S_j(\tau-\delta_1,\omega)+\gamma_2 e^{-i\omega\delta_2} S_j(\tau-\delta_1,\omega)}{1+|\gamma_1|+|\gamma_2|},\ (\tau,\omega) \in \Omega_j \qquad (6.24)$$

for $\delta_1$, $\delta_2$, and $m \le \phi$. Using local stationary assumption in (6.24), it can then be obtained

$$\begin{aligned} X_2(\tau,\omega) &\approx \frac{S_j(\tau,\omega)+\gamma_1 e^{-i\omega\delta_1} S_j(\tau,\omega)+\gamma_2 e^{-i\omega\delta_2} S_j(\tau,\omega)}{1+|\gamma_1|+|\gamma_2|},\ (\tau,\omega) \in \Omega_j \\ &= \frac{1+\gamma_1 e^{-i\omega\delta_1} + \gamma_2 e^{-i\omega\delta_2}}{1+|\gamma_1|+|\gamma_2|} S_j(\tau,\omega) \\ &= \frac{1+\gamma_1 e^{-i\omega\delta_1} + \gamma_2 e^{-i\omega\delta_2}}{1+|\gamma_1|+|\gamma_2|} X_1(\tau,\omega) \end{aligned} \qquad (6.25)$$

Hence, the proposed cost function can be formulated based on the single mixture $X_1(\tau,\omega)$ by substituting (6.25) into (6.22) which leads to

$$J(\tau,\omega) = \underset{k}{\operatorname{argmin}}\ H_k(\tau,\omega) \qquad (6.26)$$

where

$$H_k(\tau,\omega) = \left| \hat{\hat{a}}_k X_1(\tau,\omega) - \frac{1+\gamma_1 e^{-i\omega\delta_1} + \gamma_2 e^{-i\omega\delta_2}}{1+|\gamma_1|+|\gamma_2|} X_1(\tau,\omega) \right|^2 \qquad (6.27)$$

Technically, the TF plane of the mixed signal $X_1(\tau,\omega)$ is partitioned into $k$ groups of $(\tau,\omega)$ units by evaluating the proposed cost function. For each $(\tau,\omega)$ unit, the $k^{th}$ argument that gives the minimum cost will be assigned to the $k^{th}$ signal. Once the cost function is evaluated, the binary TF mask for the $j^{th}$ signal can be constructed as

$$M_j(\tau,\omega) := \begin{cases} 1 & J(\tau,\omega) = j \\ 0 & otherwise \end{cases}. \qquad (6.28)$$

The original signals will be recovered by

$$\hat{S}_j(\tau,\omega) = M_j(\tau,\omega)X_1(\tau,\omega). \qquad (6.29)$$

Finally, the estimated signals are converted back into time domain by using the inverse STFT.

The proposed MTD pseudo-stereo algorithm is summarized and expressed in a general term as:

**In Time Domain, Multiple Times Delay Pseudo – Stereo Mixing Model:**

$$x_1(t) = s_1(t) + s_2(t)$$

$$x_2(t) = \frac{x_1(t) + \sum_{p=1}^{P} \gamma_p x_1(t-\delta_p)}{1+\sum_{p=1}^{P}\left|\gamma_p\right|}$$

$$= \sum_{i=1}^{I} a_{ij}\big(t;\{\delta_p,\gamma_p\}_{p=1,\ldots,P}\big)s_j(t-\delta_i) \; + r_j\big(t;\{\delta_p,\gamma_p\}_{p=1,\ldots,P}\big)$$

where $p$ denotes the index of the MTD parameter i.e. $\{\delta_p,\gamma_p\}_{p=1,\ldots,P}$ , $a_{ij}(t;\{\delta_p,\gamma_p\}_{p=1,\ldots,P}) = \frac{\gamma_i - a_{s_j}(\delta_i)}{1+\sum_{p=1}^{P}\left|\gamma_p\right|}$ , and

$$r_j\big(t;\{\delta_p,\gamma_p\}_{p=1,\ldots,P}\big) = \frac{e_j(t) - \sum_{\substack{m=1 \\ m\neq\{\delta_p\}_{p=1,\ldots,P}}}^{D_j} a_{s_j}(m;t)s_j(t-m)}{1+\sum_{p=1}^{P}\left|\gamma_p\right|}$$

**In TF representation:**

$$X_1(\tau,\omega) = S_j(\tau,\omega)$$

$$X_2(\tau,\omega) = \sum_{i=1}^{I} a_{ij}(\tau) e^{-i\omega\delta_i} S_j(\tau-\delta_i,\omega) - \sum_{\substack{m=1 \\ m\neq\{\delta_p\}_{p=1,2,\ldots,P}}}^{D_j} \frac{a_{s_j}(m;\tau)}{1+\sum_{p=1}^{P}|\,\gamma_p|} e^{-i\omega m} S_j(\tau-m,\omega)$$

**The estimate of** $\bar{a}_j^{MTD}(\tau,\omega)$

$$\bar{a}_j^{MTD}(\tau,\omega) = \frac{X_2(\tau,\omega)}{X_1(\tau,\omega)} = \sum_{i=1}^{I} a_{ij}(\tau) e^{-i\omega\delta_i} - \sum_{\substack{m=1 \\ m\neq\{\delta_p\}_{p=1,2,\ldots,P}}}^{D_j} \frac{a_{s_j}(m;\tau)}{1+\sum_{p=1}^{P}|\,\gamma_p|} e^{-i\omega m}$$

**The MTD cost function:**

$$J(\tau,\omega) = \underset{k}{\operatorname{argmin}} \left| \hat{\bar{a}}_k X_1(\tau,\omega) - \left(\frac{1+\sum_{p=1}^{P} \gamma_p e^{-i\omega\delta_p}}{1+\sum_{p=1}^{P}|\,\gamma_p|}\right) X_1(\tau,\omega)\right|^2$$

## 6.4 Difference of Pseudo-Stereo Mixture and Multiple Times Delay Pseudo-Stereo Mixture

This section presents the difference between the proposed pseudo-stereo mixtures (in Chapter 3) and the newly proposed MTD pseudo-stereo mixtures. The pseudo-stereo mixtures and the MTD pseudo-stereo mixtures are regarded to be different through its mixing coefficient $\bar{a}_j(\tau,\omega)$. The mixing coefficient of the pseudo-stereo mixtures and the MTD mixtures are expressed in Table 6.1

Table 6.1: Mixing coefficient of the pseudo-stereo mixture and the MTD mixture

| Pseudo-Stereo Mixture | MTD Mixture $(p = 2)$ |
|---|---|
| $\bar{a}_j(\tau,\omega) = \frac{X_2(\tau,\omega)}{X_1(\tau,\omega)} e^{i\omega\delta}$<br>$= a_j(\tau) - \sum_{\substack{m=1 \\ m\neq\delta}}^{D_j} \frac{a_{s_j}(m;\tau)}{1+\lvert\gamma\rvert} e^{-i\omega(m-\delta)}$<br>$= a_j(\tau) - C_j(\tau,\omega)$ (6.30) | $\bar{a}_j^{MTD}(\tau,\omega)(\tau,\omega) = \frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}$<br>$= a_{1j}(\tau)e^{-i\omega\delta_1} + a_{2j}(\tau)e^{-i\omega\delta_2} - \sum_{\substack{m=1 \\ m\neq\delta_1,\delta_2}}^{D_j} \frac{a_{s_j}(m;\tau)}{1+\sum_{p=1}^{2}\lvert\gamma_p\rvert} e^{-i\omega m}$<br>$= a_{1j}(\tau)e^{-i\omega\delta_1} + a_{2j}(\tau)e^{-i\omega\delta_2} - C_j^{p=2}(\tau,\omega)$ (6.31) |

Eq. (6.30) and (6.31) show that the mixing coefficients of both methods differ in term of the mixing attenuation i.e. $a_j(\tau)$ and $a_{1j}(\tau)e^{-i\omega\delta_1} + a_{2j}(\tau)e^{-i\omega\delta_2}$ and the residue of the AR coefficients i.e. $C_j(\tau,\omega)$ and $C_j^{p=2}(\tau,\omega)$. Therefore, the mixing attenuations of both methods are firstly analysed by measuring theirs distinguishabilities. Next, the AR coefficients of $C_j(\tau,\omega)$ and $C_j^{p=2}(\tau,\omega)$ are compared by using Schwarz's inequality.

### *6.4.1 Comparison of Mixing Attenuation Distinguishability*

Mixing Attenuation Distinguishability (MAD) of the pseudo-stereo mixture introduced in Chapter 3 is now recapped here in (6.32). MAD of the MTD mixture can be derived by following similar line as (3.32) and expressed in (6.33).

Table 6.2: MAD of the pseudo-stereo mixture and the MTD mixture

| Pseudo-Stereo Mixture | MTD Mixture $(p = 2)$ |
|---|---|
| $\theta = \lvert a_k(\tau) - a_j(\tau)\rvert^2$<br>(6.32) | $\theta_{p=2} = \frac{\left\lvert\left(a_{1k}(\tau)e^{-i\omega\delta_1}+a_{2k}(\tau)e^{-i\omega\delta_2}\right)X_1(\tau,\omega)-X_2(\tau,\omega)\right\rvert^2}{\lvert X_1(\tau,\omega)\rvert^2}$<br>$= \frac{\left\lvert\left(a_{1k}(\tau)e^{-i\omega\delta_1}+a_{2k}(\tau)e^{-i\omega\delta_2}\right)S_j(\tau,\omega)-\left(a_{1j}(\tau)e^{-i\omega\delta_1}+a_{2j}(\tau)e^{-i\omega\delta_2}\right)S_j(\tau,\omega)\right\rvert^2}{\left\lvert S_j(\tau,\omega)\right\rvert^2}$<br>$= \left\lvert a_{1k}(\tau)e^{-i\omega\delta_1} + a_{2k}(\tau)e^{-i\omega\delta_2} - a_{1j}(\tau)e^{-i\omega\delta_1} - a_{2j}(\tau)e^{-i\omega\delta_2}\right\rvert^2$<br>$= \left\lvert\left(a_{1k}(\tau) - a_{1j}(\tau)\right)e^{-i\omega\delta_1} + \left(a_{2k}(\tau) - a_{2j}(\tau)\right)e^{-i\omega\delta_2}\right\rvert^2$<br>$\leq \left(\lvert a_{1k}(\tau) - a_{1j}(\tau)\rvert + \lvert a_{2k}(\tau) - a_{2j}(\tau)\rvert\right)^2$ (6.33) |

where $k \neq j$, $a_j(\tau;\delta,\gamma) = \frac{-a_{s_j}(\delta;\tau)+\gamma}{1+|\gamma|}$, and $a_{ij}(\tau;\{\delta_1,\delta_2\},\{\gamma_1,\gamma_2\}) = \frac{\gamma_i - a_{s_j}(\delta_i;\tau)}{1+|\gamma_1|+|\gamma_2|}$. Mixing Attenuation Distinguishabilities of $\theta$ and $\theta_{p=2}$ are assumed as $\theta_{p=2} > \theta$. The condition $\theta_{p=2} > \theta$ can be further derived as:

$$\theta_{p=2} > \theta$$

$$\left(\left|a_{1k}(\tau) - a_{1j}(\tau)\right| + \left|a_{2k}(\tau) - a_{2j}(\tau)\right|\right)^2 > \left|a_k(\tau) - a_j(\tau)\right|^2$$

$$\left|a_{1k}(\tau) - a_{1j}(\tau)\right| + \left|a_{2k}(\tau) - a_{2j}(\tau)\right| > \left|a_k(\tau) - a_j(\tau)\right|$$

$$\left|\frac{\gamma_1 - a_{s_k}(\delta_1;\tau)}{1+|\gamma_1|+|\gamma_2|} - \frac{\gamma_1 - a_{s_j}(\delta_1;\tau)}{1+|\gamma_1|+|\gamma_2|}\right| + \left|\frac{\gamma_2 - a_{s_k}(\delta_2;\tau)}{1+|\gamma_1|+|\gamma_2|} - \frac{\gamma_2 - a_{s_j}(\delta_2;\tau)}{1+|\gamma_1|+|\gamma_2|}\right| > \left|\frac{-a_{s_k}(\delta;\tau)+\gamma}{1+|\gamma|} - \frac{-a_{s_j}(\delta;\tau)+\gamma}{1+|\gamma|}\right|$$

$$\left|\frac{\gamma_1 - a_{s_k}(\delta_1;\tau) - \gamma_1 + a_{s_j}(\delta_1;\tau)}{1+|\gamma_1|+|\gamma_2|}\right| + \left|\frac{\gamma_2 - a_{s_k}(\delta_2;\tau) - \gamma_2 + a_{s_j}(\delta_2;\tau)}{1+|\gamma_1|+|\gamma_2|}\right| > \left|\frac{-a_{s_k}(\delta;\tau) + \gamma + a_{s_j}(\delta;\tau) - \gamma}{1+|\gamma|}\right|$$

$$\left|\frac{a_{s_j}(\delta_1;\tau) - a_{s_k}(\delta_1;\tau)}{1+|\gamma_1|+|\gamma_2|}\right| + \left|\frac{a_{s_j}(\delta_2;\tau) - a_{s_k}(\delta_2;\tau)}{1+|\gamma_1|+|\gamma_2|}\right| > \left|\frac{a_{s_j}(\delta;\tau) - a_{s_k}(\delta;\tau)}{1+|\gamma|}\right| \quad (6.34)$$

An Analysis of (6.34) will reveal that $\theta_{p=2} > \theta$ is true when the denominators: $|\gamma| \leq |\gamma_1| + |\gamma_2|$. To show that, two AR signals are synthesized i.e. $s_k(t)$ and $s_j(t)$ using the AR model with the following the coefficients:

$$a_{s_k} = [-3.8893, 5.7219, -3.7449, 0.9224]$$

$$a_{s_j} = [-2.7836, 3.8383, -2.6461, 0.9037]$$

The MAD results of $\theta$ and $\theta_{p=2}$ are tabulated in Table 6.3.

Table 6.3: Model parameters and MAD values of the pseudo-stereo mixture and the MTD mixture

| Pseudo-Stereo Mixture:<br>Set $\delta = 2$, and $\gamma = 1$ | MTD Mixture $(p = 2)$<br>Set $\delta_1 = 1, \delta_2 = 2,\ \gamma_1 = 1, \gamma_2 = 1$ |
|---|---|
| $\theta = \left|a_k(\tau) - a_j(\tau)\right|^2$<br>$= \left|\frac{-3.8383+5.7219}{2}\right|^2 = 0.8870$ | $\theta_{p=2} \leq \left(\left|a_{1k}(\tau) - a_{1j}(\tau)\right| + \left|a_{2k}(\tau) - a_{2j}(\tau)\right|\right)^2$<br>$= \left(\left|\frac{-2.7836-(-3.8893)}{1+2}\right| + \left|\frac{3.8383-5.7219}{1+2}\right|\right)^2 = 0.9930$ |

In Table 6.3, the MAD results show that $\theta_{p=2} > \theta$ is satisfied. The higher the MAD score is, the more isolated are the mixing coefficients associated with $s_k$ and $s_j$ signals. In this case, the MAD improvement of the MTD mixture is at 1.5% over the pseudo-stereo mixture. Additionally, MAD of the both mixtures have been computed with various weight parameters and plotted in Fig.6.1.

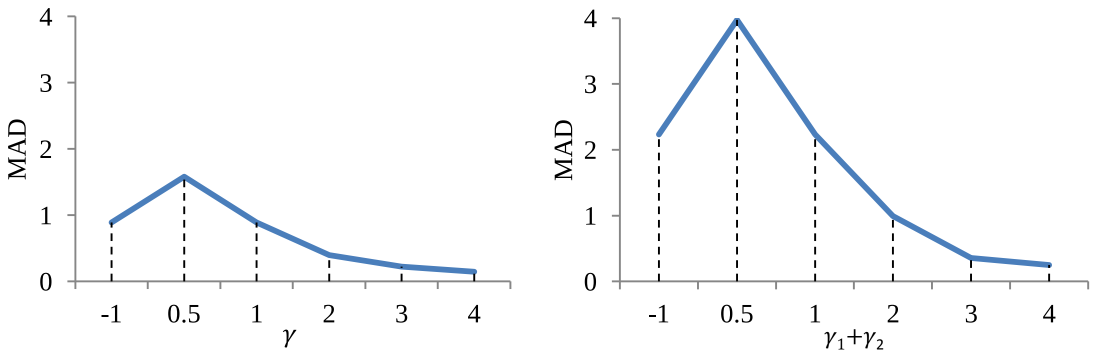


Figure 6.1: Mixing Attenuation Distinguishability of the pseudo-stereo mixture (*left*) and the MTD mixture (*right*)

In Fig.6.1, the MTD mixture delivers the better distinguishability than the pseudo-stereo mixture across the range. Mixing Attenuation Distinguishability of the MTD mixtures decreases exponentially when the weighted parameters are increased.

### *6.4.2 Comparison of AR Coefficients of Residue*

Secondly, the two mixtures are compared via their AR coefficients residue $C_j(\tau,\omega)$ and $C_j^{p=2}(\tau,\omega)$ by applying with Schwarz's inequality. The supremum of $C_j(\tau,\omega)$ and $C_j^{p=2}(\tau,\omega)$ are tabulated in Table 6.4.

Table 6.4: Schwarz's inequality of $C_j(\tau,\omega)$ and $C_j^{p=2}(\tau,\omega)$

| Pseudo-Stereo Mixture | MTD Mixture $(p = 2)$ |
|---|---|
| $C_j(\tau,\omega) \leq \sqrt{\left(D_j-1\right)\sum_{\substack{m=1\\m\neq\delta}}^{D_j}\left\lvert\frac{a_{S_j}(m)}{1+\lvert\gamma\rvert}\right\rvert^2}$ (6.35) | $C_j^{p=2}(\tau,\omega) = \sqrt{\left(D_j-1\right)\sum_{\substack{m=1\\m\neq\delta_1,\delta_2}}^{D_j}\left\lvert\frac{a_{s_j}(m)}{1+\lvert\gamma_1\rvert+\lvert\gamma_2\rvert}\right\rvert^2}$ (6.36) |

To begin, assume that the AR coefficients residue of $C_j(\tau,\omega)$ and $C_j^{p=2}(\tau,\omega)$ are such that $C_j^{p=2}(\tau,\omega) > C_j(\tau,\omega)$. This condition can be further derived as:

$$\begin{gathered}
\sqrt{\left(D_j-1\right)\sum_{\substack{m=1\\m\neq\delta_1,\delta_2}}^{D_j}\left\lvert\frac{a_{s_j}(m)}{1+\lvert\gamma_1\rvert+\lvert\gamma_2\rvert}\right\rvert^2} < \sqrt{\left(D_j-1\right)\sum_{\substack{m=1\\m\neq\delta}}^{D_j}\left\lvert\frac{a_{s_j}(m)}{1+\lvert\gamma\rvert}\right\rvert^2} \\
\left(D_j-1\right)\sum_{\substack{m=1\\m\neq\delta_1,\delta_2}}^{D_j}\left\lvert\frac{a_{s_j}(m)}{1+\lvert\gamma_1\rvert+\lvert\gamma_2\rvert}\right\rvert^2 < \left(D_j-1\right)\sum_{\substack{m=1\\m\neq\delta}}^{D_j}\left\lvert\frac{a_{s_j}(m)}{1+\lvert\gamma\rvert}\right\rvert^2 \\
\sum_{\substack{m=1\\m\neq\delta_1,\delta_2}}^{D_j}\left\lvert\frac{a_{s_j}(m)}{1+\lvert\gamma_1\rvert+\lvert\gamma_2\rvert}\right\rvert^2 < \sum_{\substack{m=1\\m\neq\delta}}^{D_j}\left\lvert\frac{a_{s_j}(m)}{1+\lvert\gamma\rvert}\right\rvert^2 \qquad (6.37)
\end{gathered}$$

From (6.37), the total number of AR order coefficients of the MTD mixture is less than the pseudo-stereo mixture by one order. This is significant as it enforces the AR coefficients residue of the MTD mixture to be less than the value of the pseudo-stereo mixture. Hence, $C_j^{p=2}(\tau,\omega) < C_j(\tau,\omega)$ is true when $\lvert\gamma\rvert \leq \lvert\gamma_1\rvert + \lvert\gamma_2\rvert$. To validate this assumption, a synthesized AR signal is used. The AR coefficients of $s_j$ are as follow:

$$a_{s_j} = [-2.7836, 3.8383, -2.6461, 0.9037]$$

Table 6.5: Model parameters and MAD values of the pseudo-stereo mixture and the MTD mixture

| Pseudo-Stereo Mixture: $\sum_{\substack{m=1\\m\neq\delta}}^{D_j}\left\lvert\frac{a_{s_j}(m)}{1+\lvert\gamma\rvert}\right\rvert^2$ | MTD Mixture: $\sum_{\substack{m=1\\m\neq\delta_1,\delta_2}}^{D_j}\left\lvert\frac{a_{s_j}(m)}{1+\lvert\gamma_1\rvert+\lvert\gamma_2\rvert}\right\rvert^2$ |
|---|---|
| Set $\delta = 1$, and $\gamma = 1$ | Set $\delta_1 = 1, \delta_2 = 2,\ \gamma_1 = 1, \gamma_2 = 1$ |
| $\sum_{\substack{m=1\\m\neq\delta}}^{D_j}\left\lvert\frac{a_{s_j}(m)}{1+\lvert\gamma\rvert}\right\rvert^2 = \left\lvert\frac{3.8383-2.6461+0.9037}{2}\right\rvert^2$<br>$= 1.0981$ | $\sum_{\substack{m=1\\m\neq\delta_1,\delta_2}}^{D_j}\left\lvert\frac{a_{s_j}(m)}{1+\lvert\gamma_1\rvert+\lvert\gamma_2\rvert}\right\rvert^2 = \left\lvert\frac{-2.6461,0.9037}{3}\right\rvert^2$<br>$= 0.3373$ |

For $|\gamma| \leq |\gamma_1| + |\gamma_2|$, the condition in (6.37) is satisfied where $0.3373 < 1.0981$. Therefore, $C_j^{p=2}(\tau, \omega) < C_j(\tau, \omega)$ is satisfied, the MTD mixture contains smaller amount of residue than the pseudo-stereo mixture.

According to the above analysis, MTD mixture yields better distinguishability of the mixing attenuation between two source signals and lesser AR coefficients residue than the pseudo-stereo mixture. This will lead to better distinction between the peak regions in the complex 2D histogram. Subsequently, the peaks of the MTD mixing model can then accurately be identified to render the more accurate mask construction. Therefore, the MTD algorithm will perform better separation performance than the pseudo-stereo mixture.

## 6.5 Results and Analysis

The separation performance of the proposed MTD method is demonstrated by separating synthetic and real-audio signals. The synthetic signals represent stationary AR signals. The real-audio signals which are inherently non-stationary include voice and music signals. Additionally, the proposed MTD method is evaluated by setting the multi-time delay parameters $\{\delta_p, \gamma_p\}$ with $p = 2$. All experiments have been conducted under the same conditions as follows: The signals are mixed with normalized power over the duration of the signals. The proposed MTD algorithm (MTD-SOLO) will be compared with the SOLO algorithm as proposed in Chapters 3 and 4. All mixed signals are sampled at 16 kHz sampling rate. The TF representation is computed by using the STFT of 1024-point Hamming window with 50% overlap. The STFT setting performs the high degree of the approximate window-disjoint orthogonality as proposed in [12]. The separation performance is evaluated by measuring the distortion between original signal and the estimated one according to the signal-to-distortion (SDR) ratio defined as $SDR = 10\, log_{10}\left(\left\|s_{target}\right\|^2/\left\|e_{interf} + e_{artif}\right\|^2\right)$ where $e_{interf}$ represent the interference from other signals and $e_{artif}$ is the artefact. MATLAB is used as the programming platform. All simulations and analyses are performed using a PC with Intel Core 2 CPU 3GHz and 3GB RAM.

### *6.5.1 Synthetic AR Signals*

### 6.5.1.1 Stationary AR Signals

Two stationary AR signals are synthesized for $s_1(t)$ and $s_2(t)$ using the AR process with following the coefficients: $a_{s_1} = [-3.8893, 5.7219, -3.7449, 0.9224]$ and $a_{s_2} = [-2.7836, 3.8383, -2.6461, 0.9037]$ and $e_1(t)$ and $e_2(t)$ are zero mean white Gaussian signal with average variances of $2.6 \times 10^{-8}$ and $1.1 \times 10^{-4}$, respectively. The coefficients and the variances are randomly selected. Two AR signals model two audio signals from a concert flute at notes C4 (262Hz) and G7 (1,976Hz), respectively. The original signals are shown in Fig. 6.3. For all methods, the histogram-resolution parameters are set at $\Delta_{\alpha^{(r)}}= 5$ , $\Delta_{\alpha^{(i)}}= 200$ , $\zeta^{(r)} = 101$ and $\zeta^{(i)} = 3$ . The pseudo-stereo parameters are selected to be $\gamma = 4$ and $\delta = 2$ for the SOLO. The MTD pseudo-stereo parameters are determined as $\delta_{p=1,2} = \{1,2\}$ and $\gamma_{p=1,2} = \{1,1\}$.

Fig.6.2 illustrates the clustering of the signals into two peaks of SOLO (left) and MTD-SOLO (right). According to the analysis in Section 6.4, the histogram of the MTD mixture obviously reveals two peaks of more distant than that obtained from the SOLO.

Fig.6.3 also shows the mixed signal and the separated signals based on the SOLO method. Visually, it can be seen that the mixture has been very well separated comparing with the original signals. The separation performance is tabulated in Table 6.6 which shows the comparative results of SOLO, MTD-SOLO and IBM.

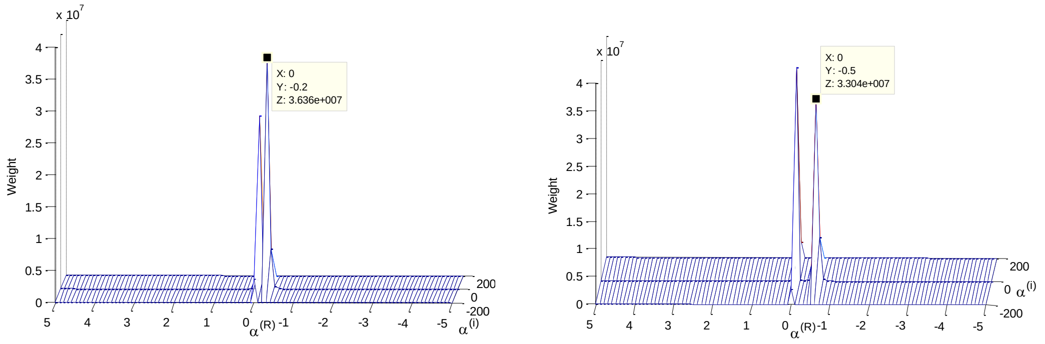


Figure 6.2: A complex 2D histogram corresponding to two signals of SOLO (*left*) and MTD-SOLO (*right*).

.

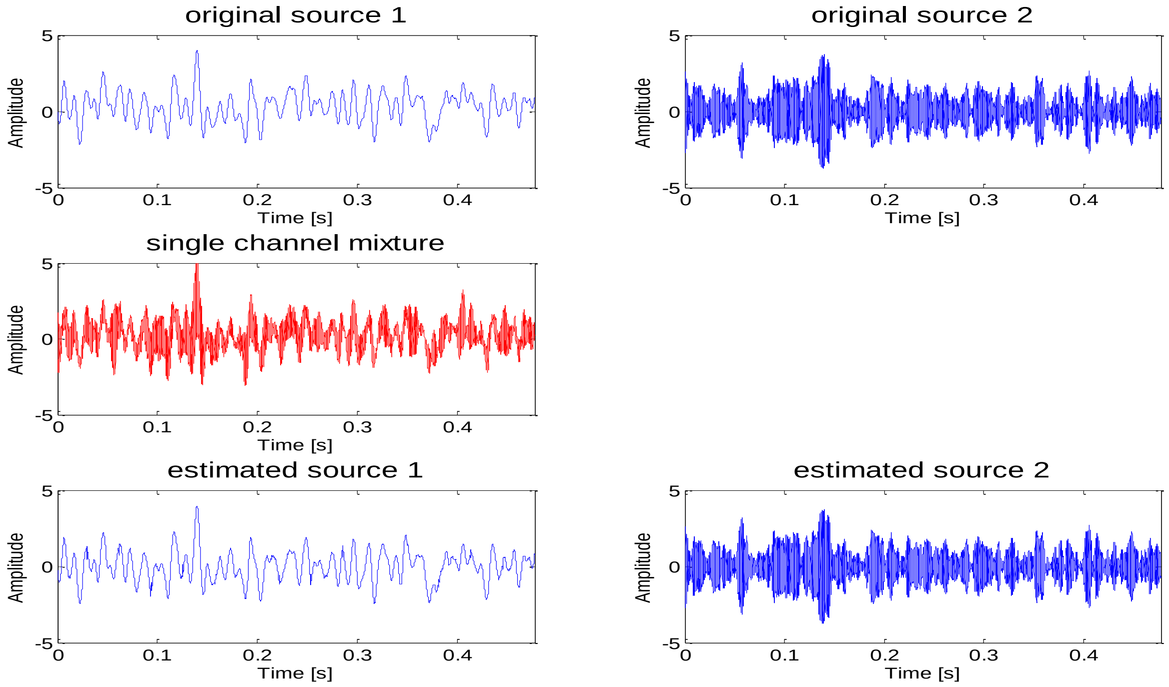


Figure 6.3: Two original signals, observed mixture and two estimated signals using SOLO.

Table 6.6: Comparison of average SDR performance on mixture of two AR signals with SOLO, MTD-SOLO and IBM

| Methods | SDR $s_1$ | SDR $s_2$ | Average |
|---|---|---|---|
| SOLO | 12.4 | 20.0 | 16.20 |
| MTD-SOLO | 12.4 | 20.1 | 16.25 |
| IBM | 12.5 | 20.5 | 16.50 |

All proposed methods have successfully separated the mixture with high accuracy comparing with the IBM. The separation performance of the MTD-SOLO method is slightly better than the SOLO methods. The clear distinction of the peaks of the MTD-SOLO method has conducted a more accurate mask and subsequently resulted in better separation performance. However, both SOLO and MTD-SOLO methods estimate $\bar{a}_j(\tau, \omega)$ by using the whole TF units that befits a stationary signal since their AR coefficients are constant for all times. This reason causes similar separation performance of the both methods.

### 6.5.1.2 Separation of more than two stationary AR Signals

In this evaluation, the proposed method is tested by increasing the number of signals from $j = 3, 4, 5$. Each mixture of 3 to 5 signals is executed 100 times. Five stationary AR signals are synthesized using the AR process with the following centre frequencies and the coefficients:

| | |
|---|---|
| $s_1$: French Horn (A2), 110Hz, | $a_{s_1} = [-3.9163,\ 5.7552, -3.7613, 0.9224]$ |
| $s_2$: Trumpet (E5), 988Hz, | $a_{s_2} = [\ -3.6286, 5.2126, -3.4849, 0.9224]$ |
| $s_3$: Concert flute (B6), 1,976Hz, | $a_{s_3} = [-2.7836, 3.8383, -2.6461, 0.9037]$ |

$s_4$: Guita bass (C4, 10 Harmonics), 2,620Hz, $a_{s_4} = [-2.0322, 2.9729, -1.9717, 0.9413]$

$s_5$: Violin (G7), 3,136Hz, $a_{s_5} = [-1.3113, 2.3703, -1.2723, 0.9413]$

and $e_1(t)$ to $e_5(t)$ are zero mean white Gaussian signals with variances $7.44 \times 10^{-10}$, $4.93 \times 10^{-6}$, $1.13 \times 10^{-4}$, $5.41 \times 10^{-5}$ and $8.36 \times 10^{-5}$, respectively. The coefficients and the variances are randomly selected. All experiments are conducted under the same conditions: $\Delta_{\alpha^{(r)}} = 5$, $\Delta_{\alpha^{(i)}} = 200$, $\zeta^{(r)} = 101$ and $\zeta^{(i)} = 3$. SOLO parameters are set as follows: $\delta = 2$, $\gamma = 4$, and $\delta_{p=2} = \{1,2\}$, $\gamma_{p=2} = \{1,1\}$ for the MTD-SOLO.

The separation performances for mixtures of 3 to 5 are tabulated in Table 6.7. By comparing with the IBM, the SDR results of both methods nearly reach the same as the results of the IBM method. For the mixture of 3 and 4 signals, the MTD-SOLO method slightly surpasses the SOLO methods for the estimated $s_1$ at the average SDR improvement of 0.1 dB per signal while the other signals have the same SDR. In the case of 5 mixing signals, the MTD-SOLO method significantly achieves the better separation performance for four estimated signals at the average SDR improvement of 0.1dB per signal. Because the advantage of more MAD and less residual AR coefficients benefits the MTD-SOLO method for estimating and identifying five peaks in the histogram. The average SDRs of 3 to 5 signals are plotted in Fig.6.4 for comparing SOLO, MTD-SOLO, and IBM.

Table 6.7: Average SDR results for mixture of 3 to 5 signals

| Methods | SDR $s_1$ | SDR $s_2$ | SDR $s_3$ |
|---|---|---|---|
| SOLO | 13.9 | 16.6 | 19.7 |
| MTD-SOLO | **14.0** | 16.6 | 19.7 |
| IBM | 14.4 | 17.2 | 19.9 |

| Methods | SDR $s_1$ | SDR $s_2$ | SDR $s_3$ | SDR $s_4$ |
|---|---|---|---|---|
| SOLO | 13.9 | 16.5 | 18.6 | 15.4 |
| MTD-SOLO | **14.0** | 16.5 | 18.6 | 15.4 |
| IBM | 14.4 | 17.2 | 19.0 | 15.6 |

| Methods | SDR $s_1$ | SDR $s_2$ | SDR $s_3$ | SDR $s_4$ | SDR $s_5$ |
|---|---|---|---|---|---|
| SOLO | 13.9 | 16.5 | 18.5 | 14.9 | 19.8 |
| MTD-SOLO | **14.0** | 16.5 | **18.6** | **15.0** | **19.9** |
| IBM | 14.4 | 17.1 | 18.9 | 15.2 | 20.3 |

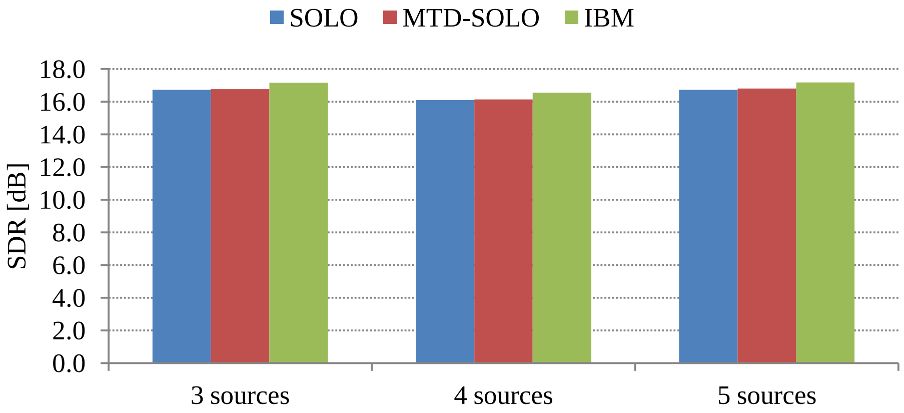


Figure 6.4: Average SDR results for mixture of 3 to 5 signals.

In Fig. 6.5 based on MTD-SOLO, the mixture of 5 signals presents and their separated signals are illustrated in Fig. 6.6 against the original signals. Visually, it is seen that all estimated signals are splendidly separated from the mixture comparing with its original one.

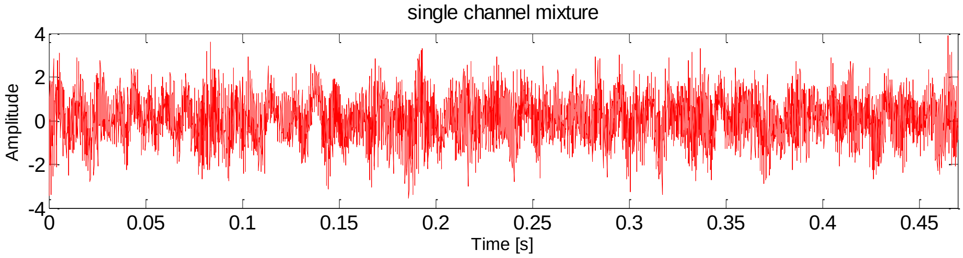


Figure 6.5: Single channel mixture of 5 AR signals using MTD-SOLO.

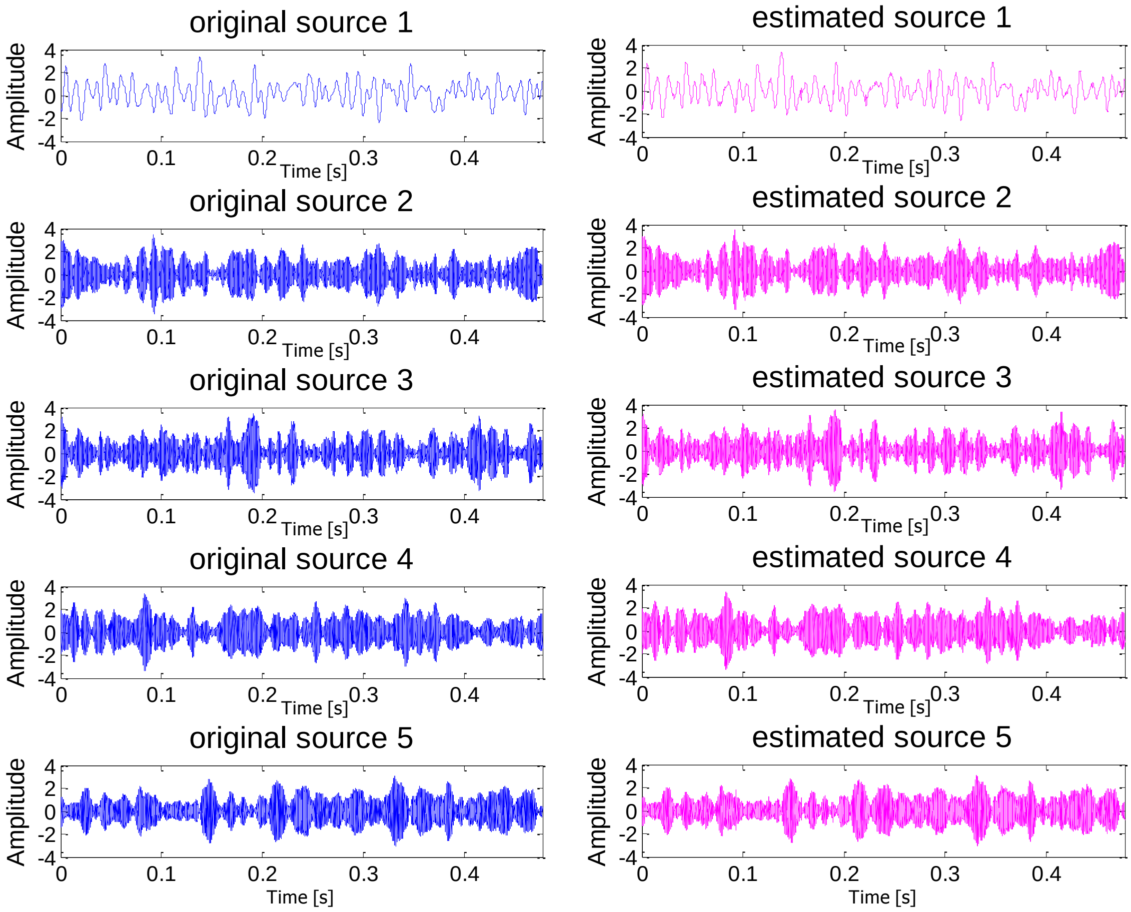


Figure 6.6: Original signals (*left*) and estimated signals (*right*) using MTD-SOLO.

### *6.5.2 Real-Audio Sources*

Audio signals can be characterized as non-stationary AR processes since their AR coefficients vary with time. Three type of mixtures are generated i.e. music + music (MM), music+speech (MS), and speech+speech (SS). The male and female speeches are randomly selected from TIMIT and music signals from the RWC database. Both signals are mixed with equal power to generate the mixture. All experiments are conducted under the same conditions: $\Delta_{\alpha^{(r)}} = 2$, $\Delta_{\alpha^{(i)}} = 500$, $\zeta^{(r)} = 101$ and $\zeta^{(i)} = 3$. SOLO parameters are set as follows: $\delta = 2$, $\gamma = 4$, and $\delta_{p=2} = \{1,2\}$, $\gamma_{p=2} = \{1,1\}$ for the MTD-SOLO.

Firstly, in the case of the musical mixture, the MTD-SOLO method yields the highest SDRs for most of the estimated musical signals as shown in Table 6.8. The average SDR of SOLO and MTD-SOLO are 9.8 and 9.9 dB per signal. The MTD-SOLO improves the separation performance over the SOLO method at the average SDR 0.1 dB per signal.

Table 6.8: Comparison of SDR performance among three mixtures of two musical signals by SOLO and MTD-SOLO

| Method | drum + jazz 1 | | piano + jazz 1 | | piano + jazz 2 | |
|---|---|---|---|---|---|---|
| | SDR $s_1$ | SDR $s_2$ | SDR $s_1$ | SDR $s_2$ | SDR $s_1$ | SDR $s_2$ |
| SOLO | 9.74 | 13.10 | 9.80 | 11.63 | 4.17 | 10.09 |
| MTD-SOLO | 9.83 | 13.11 | 9.88 | 11.64 | 4.62 | 10.58 |

An example of the separated music signals based on MTD-SOLO is illustrated in Fig 6.4. Visually, the MTD-SOLO method has successfully separated the original drum from the mixture. The estimated jazz 1 is well also separated comparing with its original signals.

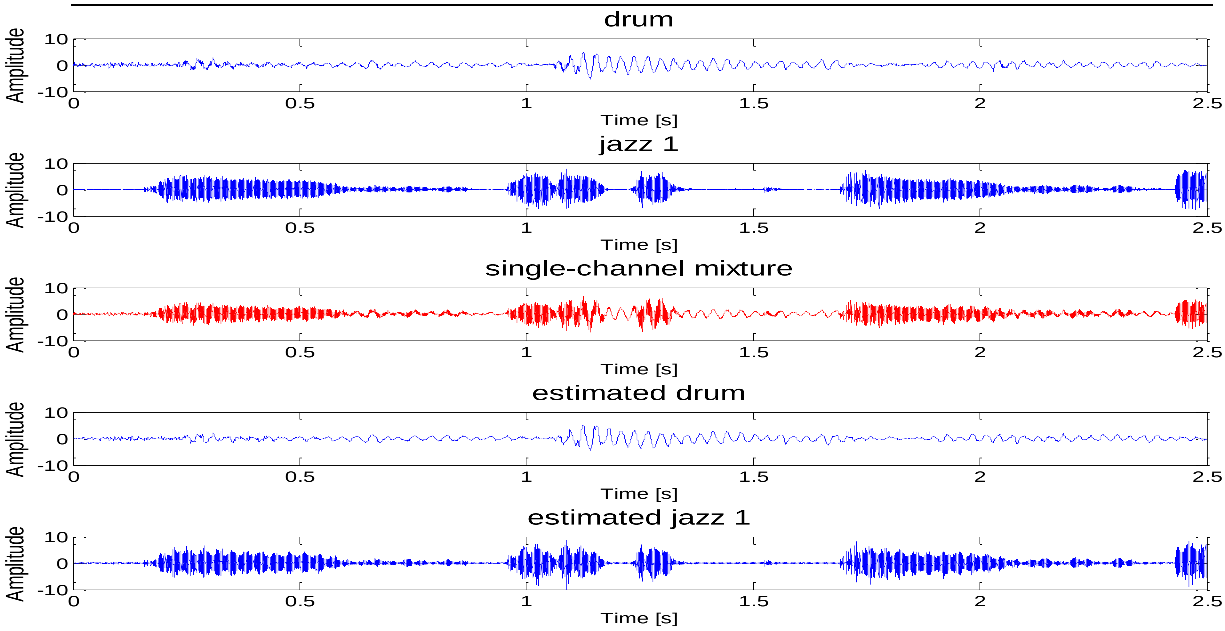


Figure 6.7: Original musical signals, single channel mixture, and estimated signals in time domain using the MTD-SOLO method.

Secondly, for the music and speech mixtures, the MTD-SOLO method performs the best across the three mixtures as tabulated in Table 6.8. The improvement of MTD-SOLO is 0.7dB per signal (8%) higher than the SOLO method where the average SDR of both methods are 9.2 and 8.5 dB per signal, respectively. Fig. 6.9 shows the original woman1 and piano signals, the mixture and the separated signals from top to bottom using the MTD-SOLO method. Visually, the separated signals are similar to the original signals.

Table 6.9: Comparison of SDR performance on three mixtures of music and speech signals for SOLO and MTD-SOLO

| Method | man 1+ jazz 1 | | man 1+ drum | | woman 1+ piano | |
|---|---|---|---|---|---|---|
| | SDR $s_1$ | SDR $s_2$ | SDR $s_1$ | SDR $s_2$ | SDR $s_1$ | SDR $s_2$ |
| SOLO | 7.38 | 6.84 | 9.94 | 6.51 | 10.23 | 10.04 |
| MTD-SOLO | **7.90** | **7.65** | **10.73** | **7.28** | **11.17** | **10.27** |

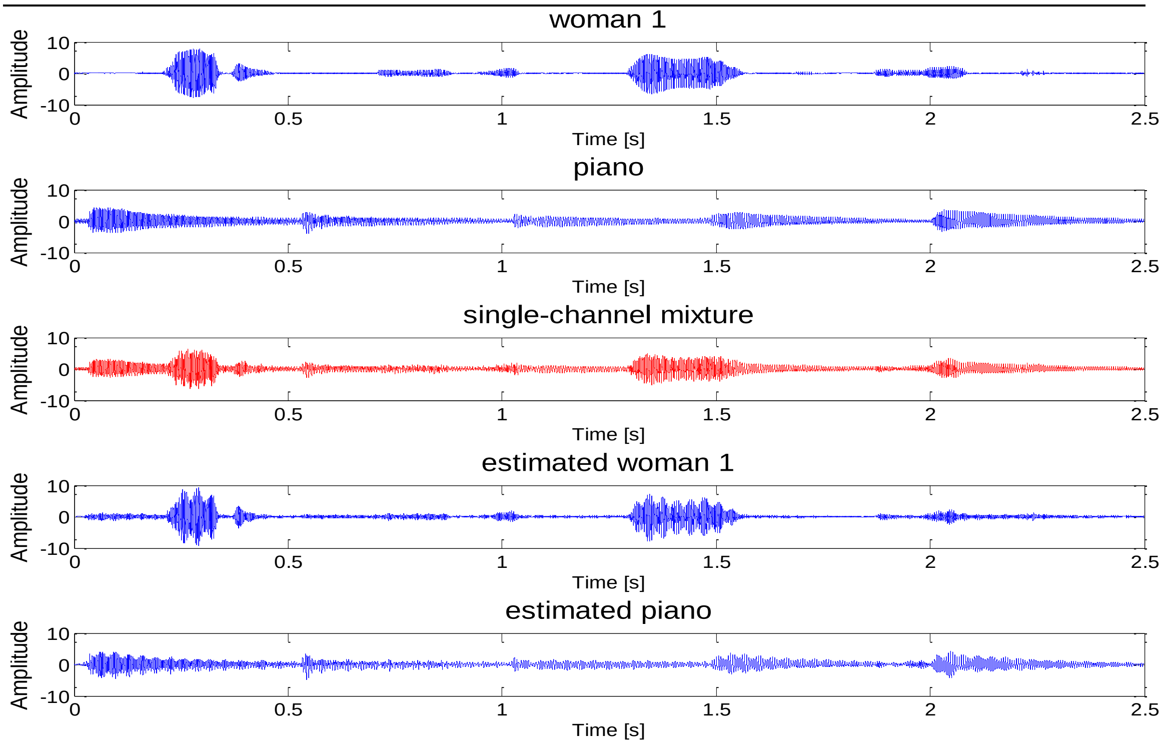


Figure 6.8: Original speech and music, single channel mixture, and estimated signals in time domain using the MTD-SOLO method.

In the extreme case of the speech mixture, the separation performances of the SOLO and MTD-SOLO methods decrease when compared with the two previous types of mixtures. In Table 6.10, the MTD-SOLO method still yields the best separation performance across all speech mixtures and the average SDR improvement over the SOLO method is 0.6 dB per signal (39%). The average SDR are 1.5 and 2.2 dB per signal for SOLO and MTD-SOLO methods, respectively.

Table 6.10: Comparison of SDR performance on three mixtures of two speech for SOLO and MTD-SOLO

| Method | man 2 + woman 1 | | man 1+ woman 2 | | man 1+ woman 3 | |
|---|---|---|---|---|---|---|
| | SDR $s_1$ | SDR $s_2$ | SDR $s_1$ | SDR $s_2$ | SDR $s_1$ | SDR $s_2$ |
| SOLO | 1.18 | 0.92 | 1.49 | 4.25 | 1.17 | 0.27 |
| MTD-SOLO | **1.94** | **1.82** | **1.75** | **4.89** | **1.37** | **1.16** |

Fig. 6.9 illustrates the separated signals of the man1 and woman mixture compared with the original signals. Visually, the speech mixture can be separated where the estimated man 1 and woman2 have the main signature of its original signals. However, the recovered signals still contain some interfered signals.

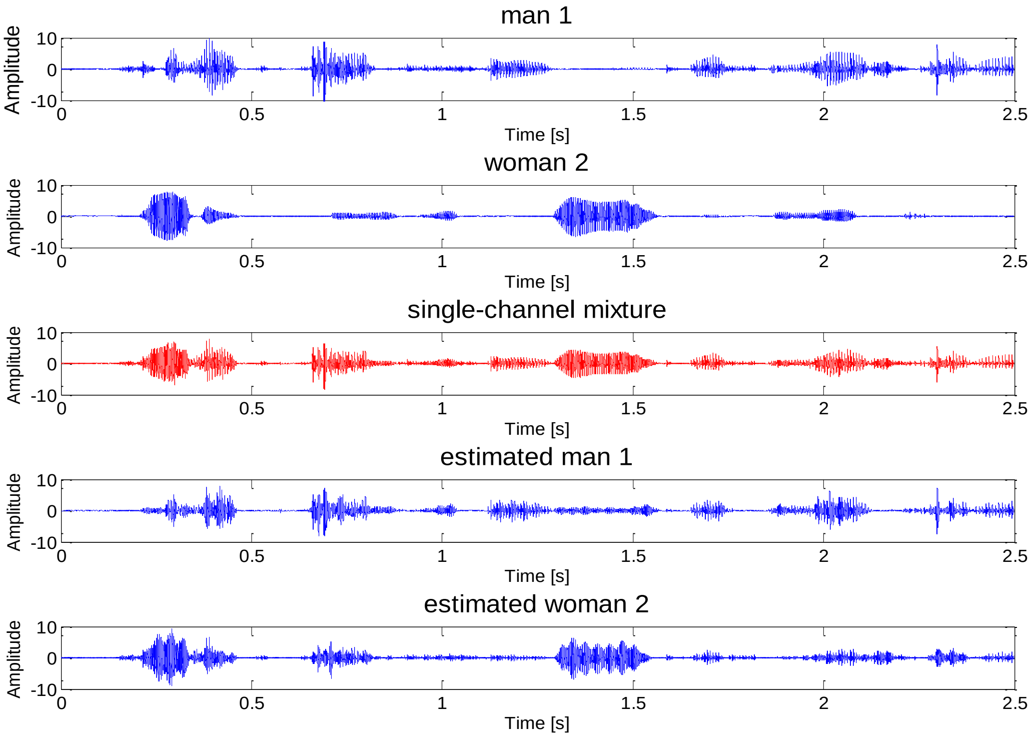


Figure 6.9: Original speech signals, single channel mixture, and estimated signals in time domain using the MTD-SOLO method.

In conclusion, the average SDR of all types of mixtures for SOLO and MTD-SOLO methods presents in Fig.6.10. The MTD-SOLO method produces the best separation performance. The MTD-SOLO method enhances the separation performance over the SOLO method at the average SDR 0.5dB per signal (7%).

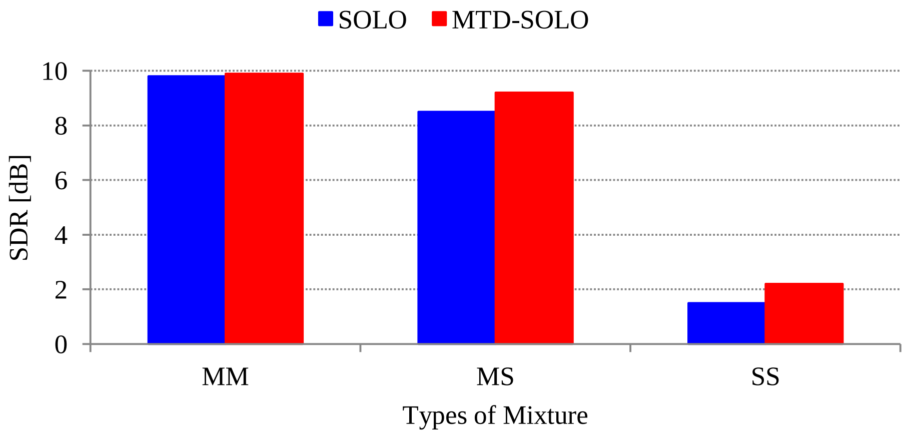


Figure 6.10: Average SDR performance of three types of mixture for SOLO and MTD-SOLO.

The characteristic of each signal in the mixtures of MM and MS differs significantly from one signal to anothers, thus the AR coefficients of two signals are also different. On the other hand, the AR coefficients of speech signals are closer to each other and highly nonstationary. Thus it becomes more difficult to separate than the MM and MS mixtures.

In the case of the MTD-SOLO method, the average SDR results have higher SDR values than the SOLO method. This can be explained that the mixing coefficient of the MTD-SOLO method takes advantage of more distinguishability $\theta_{p=2} > \theta$ and the residual coefficients $C_j^{p=2}(\tau,\omega) < C_j(\tau,\omega)$ which is less than the SOLO method as the analysis in Section 6.4. These properties befit for complex mixtures as audio mixing signals. Therefore, the MTD-SOLO is able to recover better signals than the SOLO method.

## 6.6 Summary

In this chapter, a novel family of the SOLO method has been proposed. The chapter presents the multi-time delay mixing model to solve the single-channel signal separation problem. The MTD mixture creates an extra mixing attenuation $a_{2j}(t)$ compared with the observed mixture. Hence, the separability analysis of the MTD mixing model has more flexibility than the pseudo-stereo mixtures. Additionally, the proposed MTD mixing model contributes the desirable properties to estimate signal characteristic: (*i*) increase the distinguishability of the mixing attenuation between signals (*ii*) reduce AR coefficients residues. Therefore, combining the MTD mixing model has improved the separation process of the SOLO algorithm.

# CHAPTER 7

# CONCLUSION OF THE THESIS

The work in this thesis has fulfilled all the aims and objectives set out in Chapter 1. The novel artificial mixture created from the observed mixture has been proposed. This paves the way for the development of the three new single-channel blind signal separation (SCBSS) methods as follows: Firstly, the SOLO algorithm has been demonstrated to distinguish both stationary and nonstationary mixtures in one go. In particular, the separation performance is closed to the ideal binary mask for a stationary mixture in which case its result is based on the averaged AR coefficient of each source. For the nonstationary mixture, it can be segmented into arbitrary blocks and then proceed with the separation process to achieve better signal separation performance. Secondly, the noisy pseudo-stereo mixing model using the mixture enhancement and online parameter estimation giving rise to the SOLO-APE method. This method is able to recover the original signal from stationary and nonstationary noisy environment. Based on frame-by-frame estimation, this method naturally befits nonstationary signals. Finally, the multi-time delay (MTD) pseudo-stereo mixture using the SOLO separation algorithm; the MTD-SOLO method; enhances separation performance of SOLO especially for the complex mixture which contains more than two sources or similar sources where theirs frequencies are in the same range.

In Chapter 2, an overview of the SCBSS methods was presented. Both *model-based* SCBSS and data-driven SCBSS methods aim to increase the accuracy of the separated signals. The various algorithms suit for different type of signals in different situations based on their limitations and constraints. Therefore, these approaches have not solved the SCBSS problem. These problems have been concluded in Chapter 2. Therefore, it is essential to develop an efficient solution for the separation of single channel mixtures to improve the degree of separation performance at both theoretical and practical issues. This fact is the motivation of this thesis, which commits to develop new algorithms for alleviating the SCBSS problem.

## 7.1 Proposed SCBSS Methods

In Chapter 3, a new pseudo-stereo mixing model presents the extendsion of binaural signal separation approaches to solve a monaural signal separation problem. A novel “pseudo-stereo” mixing model is proposed to create a synthetic stereo signal by weighting and time-shifting the original single-channel mixture. Separability analysis of the proposed model has also been derived to verify that the artificial stereo mixture is separable. This work overcomes the under-determined ill-conditions associated with monaural signal separation and path the way forward for binaural signal separation approaches to solve monaural mixture. For practical application, the recommend ranges of the $(\gamma, \delta)$ pairs have been provided by measuring the proportion of distinguishability between the mixing attenuations and AR coefficients residue.

In Chapter 4, a novel ‘SOLO’ framework for solving the unsupervised SCBSS problem is presented. The proposed method takes an advantage of the relationship between the readily available single-channel mixed signal and the ‘pseudo-stereo’ mixture to estimate the signature of the signals. For separation, a binary maximum likelihood time-frequency mask is construced and the conditions required for unique mask construction from the maximum likelihood framework have also been identified. The proposed algorithm yields superior performance and is computationaly very fast compared with the existing SCBSS methods.

To this end, the proposed method enjoys at least three advantages: Firstly, it does not require *a priori* knowledge of the signals. Secondly, the proposed approach is able to capture the music and speech characteristics and hence, renders robustness to the separation method. Finally, the proposed technique holds a desirable property — neither iterative optimization nor parameter initialization is required and this enables the separation process to be fast and executable in “one-go”.

In Chapter 5, a novel framework to solving SCBSS in noisy environment is proposed. The proposed method enhances the signals in the noisy mixing model and then separates the enhanced mixture. The proposed framework contributes to the desirable properties which are summarized below: 1) It is an online adaptive separation where the observed mixture is segmented into small frames. The separation process is then executed adaptively frame-by-frame. This online separation reduces the computational complexity of the whole observed mixture. Thus, it needs low computational cost. Batch methods usually suffer from large storage requirement and high computational complexity when the observed mixture is of large scale. Hence, the robustness of the proposed algorithm benefits the real-time signal processing applications. 2) It is an adaptive parameters estimation method. The parameters are adaptively estimated from two consecutive frames. The self-adaptive property is preferred for time-varying signals especially speech and highly nonstationary noise. 3) It is independent of parameters initialization, i.e. no need for random initial inputs or any predetermined structure on the sensors. This renders robustness to the proposed method. 4) It has the computational simplicity and does not exploit high-order statistics. Hence this results in the ease of implementation. The proposed method has demonstrated high level separation performance for real-audio signals in nonstationary noisy environment.

In Chapter 6, the pseudo-stereo mixture in Chapter 3 is further extended by using multiple times delay the observed mixture. A novel family of the SOLO method is presented. The new proposed mixing model increases the distinguishability of the mixing attenuation between signals and reduces residual AR coefficients. Therefore, reformulating the SOLO algorithm using the MTD mixing model improves the separation performance over the SOLO method.

In conclusion, the proposed methods are summarized for each critical issue and presented in the following tables. Firstly, SOLO, SOLO-APE, and MTD-SOLO are compared in the mixing model as shown in Table 7.1. SOLO and MTD-SOLO have been modeled in anechoic environment while SOLO-APE is in echoic environment.

Table 7.1: Summary of pseudo-stereo mixing model for SOLO, SOLO-APE, and MTD-SOLO

| Methods | Pseudo-Stereo Mixing Model |
|---|---|
| SOLO | $x_1(t) = \sum_{j=1}^{2} s_j(t)$<br>$x_2(t) = \sum_{j=1}^{2} a_j(t;\delta,\gamma)s_j(t-\delta) + \sum_{j=1}^{2} r_j(t;\delta,\gamma)$ |
| SOLO-APE | $x_1(t) = \sum_{j=1}^{2} s_j(t) + n_1(t)$<br>$x_2(t) = \sum_{j=1}^{2} a_j(t;\delta,\gamma)s_j(t-\delta) + \sum_{j=1}^{2} r_j(t;\delta,\gamma) + n_2(t;\delta,\gamma)$ |
| MTD-SOLO | $x_1(t) = \sum_{j=1}^{2} s_j(t)$<br>$x_2(t) =$<br>$\sum_{j=1}^{2}\sum_{i=1}^{P} a_{ij}\big(t;\{\delta_p,\gamma_p\}_{p=1,\ldots,P}\big)s_j(t-\delta_i) + \sum_{j=1}^{2} r_j\big(t;\{\delta_p,\gamma_p\}_{p=1,\ldots,P}\big)$ |

Secondly, the mixing coefficient of the $j^{th}$ source represents the $j^{th}$ sources' signature to be estimated. Determining the mixing coefficient is based on the pseudo-stereo mixing model. Comparison of the mixing coefficients for SOLO, SOLO-APE, and MTD-SOLO is presented in Table 7.2.

Table 7.2: Summary of mixing coefficient for SOLO, SOLO-APE, and MTD-SOLO

| Methods | Mixing coefficient $\bar{a}_j(\tau,\omega)$ |
|---|---|
| SOLO | $\bar{a}_j^{SOLO}(\tau,\omega) = \frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}e^{i\omega\delta} = a_j(\tau) - C_j(\tau,\omega)$ |
| SOLO-APE | $\bar{a}_j^{APE}(\tau,\omega) =$<br>$\frac{\hat{\hat{X}}_2(\tau,\omega)}{\hat{\hat{X}}_1(\tau,\omega)}e^{i\omega\delta} = \frac{[a_j(\tau)-C_j(\tau,\omega)]e^{-i\omega\delta}\hat{S}_j(\tau,\omega)+\tilde{N}_2(\tau,\omega)}{\hat{S}_j(\tau,\omega)+\tilde{N}_1(\tau,\omega)}e^{i\omega\delta}$ |
| MTD-SOLO | $\bar{a}_j^{MTD}(\tau,\omega) = \frac{X_2(\tau,\omega)}{X_1(\tau,\omega)} = \sum_{i=1}^{I} a_{ij}(\tau)e^{-i\omega\delta_i} - C_j^{\{\gamma_p\delta_p\}}(\tau,\omega)$ |

Thirdly, a TF mask can be constructed by evaluating the cost function $J(\tau,\omega)$. The cost function partitions the TF plane of the mixture into $k$ groups of $(\tau,\omega)$ units by assigning each $(\tau, \omega)$ unit with the $k^{th}$ argument that gives the minimum cost. The cost functions of SOLO, SOLO-APE, and MTD-SOLO are presented in Table 7.3

Table 7.3: Summary of cost function for SOLO, SOLO-APE, and MTD-SOLO

| **Methods** | **Cost function** $J(\tau, \omega) = \underset{k}{argmin}\, H_k(\tau, \omega)$ |
|---|---|
| SOLO | $H_k(\tau, \omega) = \left\lvert \hat{\hat{a}}_k e^{-i\omega\delta} X_1(\tau, \omega) - \left(\frac{1+\gamma e^{-i\omega\delta}}{1+\lvert\gamma\rvert}\right) X_1(\tau, \omega) \right\rvert^2$ |
| SOLO-APE | $H_k(\tau, \omega) = \left\lvert \frac{\tilde{\tilde{a}}_k e^{-i\omega\delta} \hat{\tilde{X}}_1(\tau,\omega) - \left(\frac{1+\gamma e^{-i\omega\delta}}{1+\lvert\gamma\rvert}\right) \hat{\tilde{X}}_1(\tau,\omega)}{\hat{\sigma}^2_{\tilde{N}_2}(\tau,\omega) + \hat{\sigma}^2_{\tilde{N}_1}(\tau,\omega)\, \tilde{\tilde{a}}^2_k(\tau)} \right\rvert^2$ |
| MTD-SOLO | $H_k(\tau, \omega) = \left\lvert \hat{\hat{a}}_k X_1(\tau, \omega) - \left(\frac{1+\sum_{p=1}^{P} \gamma_p e^{-i\omega\delta_p}}{1+\sum_{p=1}^{P}\lvert \gamma_p\rvert}\right) X_1(\tau, \omega) \right\rvert^2$ |

Finally, prerequisite the separation process for SOLO, SOLO-APE, and MTD-SOLO is presented in Table 7.4.

Table 7.4: Summary of preprocess for SOLO, SOLO-APE, and MTD-SOLO

| **Methods** | **Preprocess** |
|---|---|
| SOLO | - |
| SOLO-APE | Mixture enhancement |
| MTD-SOLO | - |

## 7.2 Future Work

### *7.2.1 Development of Adaptive Mixing Coefficient Estimator for Multi-Time Delay Pseudo-Stereo Mixing Model*

As described in Chapter 5, audio signal is nonstationary and correlated between neighbouring frequencies bins of consecutive frames. Thus, the adaptive mixing coefficient estimator is proposed to compute the mixing coefficient of the signal frame-by-frame. The SOLO-APE method yields the best separation performance. The MTD mixing model as shown in Chapter 6 has enhanced the sepration performance of the SOLO method. Therefore, the development of multi-time delay mixing model using adaptive mixing coefficient estimator will improve the separation performance for isolated the nonstationary signals from the mixture of more than two sources.

### *7.2.2 Development of Multi-Time Delay Mixing Model based Cochleagram*

The time-frequency (TF) analysis is the core technique of characterizing and manipulating audio signals. The study of three TF representations i.e. classic spectrogram, log-frequency spectrogram and cochleagram is presented in [60]. According to [60], the cochlear suits to be the TF representative of the time-varying signals due to the following reasons: The cochlear model based on the gammatone filters bank is approximately logarithmically spaced with Q constant for frequencies that range between $f_s/10$ to $f_s/2$ and approximately linearly spaced for frequencies below $f_s/10$. Hence, the cochlear has a non-uniform TF resolution while it is balanced between high and low frequency zones. A cochleargram is inspired by the auditory nerve. The cochleargram is modelled by using the gammatone filterbank which decomposes the time-domain input into the frequency domain. The impulse response of a gammatone filter centered at frequency $f$ is give by:

$$g(f,t)=\begin{cases} t^{l-1}e^{-2\pi vt}, \ t \geq 0 \\ \qquad\quad 0, \ else \end{cases} \tag{7.1}$$

where $l$ is the order of filter, $v$ denotes the rectangular bandwidth which increases with the center frequency $f$. The filter output response $x(c,t)$ can be expressed with regards to a particular filter channel $c$ as:

$$x(c,t) = x(t) * g(f_c, t) \tag{7.2}$$

where $f_c$ denotes the center frequency, and '*' indicates a convolution operator.

Therefore, the development of multi-time delay mixing model using Cochleagram will improve the accuracy of the separation performance.

### *7.2.3 Development of Component Regeneration for SCBSS*

The quality improvement of the speech signals has caught the attention of entusiastic researchers in broad disciplines. In noise reducetion approaches, the speech signal is generally corrupted with background noise. Based on noise reduction methods, the output singnal is always distorted as a result of the over-attenuation of speech components. Low energy components are usually regarded as noise in noise reduction processing and are then highly suppressed. To enhance the deteriorated speech signal, postprocessingn methods have been proposed in [61, 62]. Harmonic component regeneration is a method to reduce the speech distortion which can noticeably improve the voiced as proposed in [62]. On the other hand, the recovering of both the voiced and unvoiced speech is proposed by synthesizing the missing components in [62].

In the SCBSS problem, the separated outputs are always distorted to some degree caused by the imperfection of the SCBSS methods. Signal components are arbitrarily assigned for a particular signal. Thus each component has a chance of error in determining to the original signals that will lead the distortion of the estimated signals. Therefore, a postprocessing is proposed to improve the quality of the estimated signals by regenerating the missing signal components in TF representation.

# REFERENCE

[1] J. H. McDermott, “The cocktail party problem,” Current Biology, vol. 19, no. 22, pp. 1024 -1027, 2009

[2] E. C. Cherry, “Some experiments on the recoginition of speech, with one and two ears,” *Journal of the Acoustic Society of America*, vol. 25, pp. 975 – 979, 1953.

[3] A. Deleforge, and R. Horaud, “The cocktail party robot: Sound source separation and localisation with an active binaural head,” in *Proc. of 7th ACM/IEEE International Conference Human-Robot Interaction*, pp. 31 - 438 , Mar. 2012.

[4] J. Zhang, W.L. Woo, and S.S. Dlay, “Blind Source Separation of Post-Nonlinear Convolutive Mixture,” *IEEE Trans. Audio, Speech, and Lang. Process*., vol. 15, no. 8, pp. 2311-2330, 2007.

[5] J.-T. Chien and H.-L. Hsieh, “Nonstationary source separation using sequential and variational Bayesian learning,” *IEEE Trans. Neural Netw. and Learning Sys*., vol. 24, no. 5, pp. 681–694, May. 2013.

[6] M. Anderson, T. Adali, and X.-L. Li , “Joint Blind Source Separation With Multivariate Gaussian Model: Algorithms and Performance Analysis,” *IEEE Trans. Signal Process.*, vol.60, no.4, pp. 1672 - 1683, Apr. 2012.

[7] N. Das, A. Routray, and P. K.e Dash, “ICA Methods for Blind Source Separation of Instantaneous Mixtures: A Case Study,” *Neural Information Processing*, vol. 11, no. 11, Nov. 2007.

[8] A. Hyvärinen, “Fast and robust fixed-point algorithms for independent component analysis,” *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, 1999.

[9] S. Moon, and H. Qi, “Hybrid dimensionality reduction method based on support vector machine and independent component analysis,” *IEEE Trans. Neural Netw. and Learning Sys*., vol. 23, no. 5, pp.749 -761, 2012.

[10] S. Javidi, C. C. Took, and D. P. Mandic, “Fast independent component analysis algorithm for quaternion valued signals,” *IEEE Trans. Neural Netw*., vol. 22, no.12, pp. 1967–1978, Dec 2011.

[11] P. Gao, W. L. Woo, and S. S. Dlay, “Nonlinear signal separation for multinonlinearity constrained mixing model,” *IEEE Trans. Neural Netw*., vol. 17, no. 3, pp. 796–802, May. 2006.

[12] Ö. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. Signal Process*., vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[13] T. May, S. V. D. Par, and A. Kohlrausch, “A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation,” *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 7, pp. 2016–2030, Sep. 2012.

[14] J. Woodruff and D.L. Wang, “Binaural localization of multiple sources in reverberant and noisy environments,” *IEEE Trans. Audio, Speech, and Lang. Process*., vol. 20, no. 5, pp. 1503–1512, Jul. 2012.

[15] W.L. Woo and S.S. Dlay, “Neural network approach to blind signal separation of mono-nonlinearly mixed signals,” *IEEE Trans. Circuits and System*, vol. 52, no. 2, pp. 1236-1247, Jun. 2005.

[16] C. J. James, and S. Wang, “Blind Source Separation in single-channel EEG analysis: An application to BCI, in *Proc. of 28th IEEE EMBS annual International conference New York City USA*, pp. 6544 – 6547, Aug.2006.

[17] D. A. Reynolds, and R. C. Rose, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models,” *IEEE Trans. on Speech and Audio Process*., vol. 3, no. 1, pp. 72-83, Jan. 1995.

[18] L.R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceeding of the IEEE*, Vol. 88, No. 2, February 1989.

[19] M. P. Cooke, Modelling Auditory Processing and Organisation. Cambridge University Press, Cambridge, UK, 1993.

[20] http://en.wikipedia.org/wiki/Pitch_(music)

[21] M. Weintraub, A theory and computational model of monaural auditory sound separation. Ph. D. Thesis, Standford University, 1985.

[22] D. L. Wang and G. J. Brown, Computational Auditory Scene Analysis. *The Institute of Electrical and Electronics Engineers*, Inc., pp. 1-43. 2006

[23] C.-L. Hsu and J.-S. R. Jang, “On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset,” *IEEE Trans. Audio, Speech, and Lang. Proc*., vol. 18, no. 2, pp. 310-319, Feb. 2010.

[24] Z. Jin and D.L. Wang, “Reverberant Speech Segregation Based on Multipitch Tracking and Classification,” *IEEE Trans. Audio, Speech, and Lang. Proc*., vol. 19, no. 8, pp. 2328 - 2337, Nov. 2011.

[25] Y. Shao and D. L. Wang, “Sequential organization of speech in computational auditory scene analysis,” *Speech Commun*., vol. 51, pp. 657–667, Aug. 2009

[26] L.Benaroya and F. Bimbot, “Wiener Based Source Separation With HMM/GMM Using A Single Sensor”, in *Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation*, Nara, Japan, Apr. 2003.

[27] M. E. Davies and C. J. James, “Source separation using single channel ICA,” *Signal Process.*, vol. 87, no. 8, pp. 1819–1832, 2007.

[28] B. Mijovic, M. D. Vos, I. Gligorijevic, J. Taelman, and S. V. Haffel, “Source separation from single-channel recordings by combining empirical-mode decomposition and independent component analysis,” *IEEE Trans. Biomedical Eng*., vol. 57, no. 9, Sep. 2010.

[29] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[30] B. Gao, W. L. Woo, and S. S. Dlay, “Variational regularized 2-D nonnegative matrix factorization,” *IEEE Trans. Neural Netw. and Learning Sys*., vol. 23, no. 5, pp. 703–716, May. 2012.

[31] P. O. Hoyer, “Non-negative matrix factorization with sparseness constraints”, *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[32] K. E. Hild II, H. T. Attias, and S. S. Nagarajan, “An expectation–maximization method for spatio–temporal blind source separation using an AR-MOG source model,” *IEEE Trans. Neural Netw.*, vol. 19, no. 3, pp. 508-519, Mar. 2008.

[33] M. N. Schmidt and M. Morup, “Nonnegative matrix factor 2-D deconvolution for blind single channel source separation,” in *Proc. ICABSS 2006*, vol. 3889, pp. 700–707, Mar. 2006.

[34] G. Zhou, S. Xie, Z. Yang, J.-M. Yang, and Z. He, “Minimum-volume-constrained nonnegative matrix factorization: enhanced ability of learning parts,” *IEEE Trans. Neural Netw*., vol. 22, no. 10, pp. 1626-1637, Oct. 2011.

[35] B. Gao, W. L. Woo, and S. S. Dlay, “Single-channel source separation using EMD-subband variable regularized sparse features,” *IEEE Trans. Audio, Speech, and Lang. Process*., vol. 19, no. 4, pp. 961-976, May 2011.

[36] R. Balan, A. Jourjine, and J. Rosca, “Ar process and sources can be reconstructed from degenerate mixtures,” in Independent Component Analysis and Blind Signal Separation, International Conference on (ICA), pp. 467–472, Jan. 1999.

[37] R. Balan and J. Rosca, “A spectral power factorization,” Siemens Corporate Research. Princeton, NJ, Tech. Rep. SCR-01-TR-703, Sep 2001.

[38] http://documentation.apple.com/en/finalcutpro/usermanual/index.html#chapter=52%26section=6%26tasks=true

[39] http://www.dpamicrophones.com/en/Mic-University/Stereo-Techniques/Background.aspx

[40] R. Balan, J. Rosca, S. Rickard, and J. O'Ruanaidh, "The influence of windowing on time delay estimates," in *Proc. Conf. Inform. Sci. Syst.*, vol. 1, Princeton, NJ, Mar. 2000, pp. WP1-15 –17.

[41] R. de Frein and S. Rickard, "The synchronized short-time-Fourier-transform: properties and definitions for multichannel source separation," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 91-103, Jan. 2011.

[42] R.G. McKilliam, B.G. Quinn, I.V.L. Clarkson, and B. Moran, "Frequency estimation by phase unwrapping," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 2953-2963, Jun. 2010.

[43] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 4, pp. 1462-1469, Jul. 2006.

[44] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: music genre database and musical instrument sound database," in *Proc. ISMIR 2003,* pp. 229-230, Oct. 2003.

[45] K. Achan, S.T. Roweis and B.J. Frey, "Probabilistic inference of speech signals from phaseless spectrograms," in *Advances in Neural Information Processing Systems* 16, MIT Press, Cambridge, MA, pp.1393-1400. 2004.

[46] Y. Song and X. Peng, "Spectra analysis of sampling and reconstructing continuous signal using hamming window function," in *Proc. 4th IEEE Intl. Conf. Natural Comp.*, pp. 48-52, Nov. 2008

[47] W.-K. Ma, T.-H. Hsieh, and C.-Y. Chi, “DOA estimation of quasi-stationary signals with less sensors than sources and unknown spatial noise covariance: a khatri–rao subspace approach,” *IEEE Trans. Signal Process*., vol. 58, no. 4, pp. 2168–2180, Apr. 2010

[48] K. E. Hild II, H. T. Attias, and S. S. Nagarajan, “An expectation–maximization method for spatio–temporal blind source separation using an AR-MOG source model,” *IEEE Trans. Neural Netw*., vol. 19, no. 3, pp. 508-519, Mar. 2008.

[49] B. Mijovic, M. D. Vos, I. Gligorijevic, J. Taelman, and S. V. Haffel, “Source separation from single-channel recordings by combining empirical-mode decomposition and independent component analysis,” *IEEE Trans. Biomedical Eng*., vol. 57, no. 9, Sep. 2010.

[50] Y. Li and D. Wang, “On the optimality of ideal binary time-frequency masks,” in Proc. *IEEE Int. Conf. Acoust., Speech, Signal Process*., pp. 3501-3504, May 2008.

[51] B. Gao, W. L. Woo, and S. S. Dlay, “Unsupervised single-channel separation of nonstationary signals using Gammatone filterbank and Itakura–Saito nonnegative matrix two-dimensional factorizations,” *IEEE Trans. Circuits and Sys*. I, vol. 60, no. 3, pp. 662-675, 2013.

[52] B. Gao, W. L. Woo, and S. S. Dlay, “Adaptive sparsity nonnegative matrix factorization for single channel source separation,” *IEEE Journal of Selected Topics in Signal Process*., vol. 5, no. 5, pp. 989-1001, 2011.

[53] Y. Xiang, S. K. Ng, and V. K. Nguyen, “Blind Separation of Mutually Correlated Sources Using Precoders,” *IEEE Trans. Neural Netw*., vol. 21, no. 1, pp. 82–90, Jan. 2010.

[54] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process*. Lett., vol. 6, no. 1, pp. 1–3, Jan 1999.

[55] S. Mousazadeh and I. Cohen, “AR-GARCH in presence of noise: parameter estimation and its application to voice activity detection,” *IEEE Trans. Audio, Speech, and Lang. Process*., vol. 19, no. 4, pp. 916-926, May 2011.

[56] Y. Ephraim, and D. Malah, “Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator,” *IEEE Trans. Audio, Speech, and Lang. Process*., vol. 32 , no. 6, pp. 1109–1121, Dec. 1984.

[57] Signal Processing Information Base: Noise Data, Rice Univ., Houston, TX [Online]. Available: http://spib.rice.edu/spib/select_noise.html

[58] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs,” in *Proc. ICASSP 2001*, pp. 749–752, 2001.

[59] J. Ma, Y. Hu, and P. C. Loizou, “Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions,” *J. Acoust. Soc. Amer*., vol. 125, pp. 3387–3405, 2009.

[60] B. Gao, Single channel blind source separation, Ph.D. thesis, Newcastle University, Newcastle upon Tyne, United Kingdom, 2011.

[61] C. Plapous, C. Marro, and P. Scalart, “Improved signal-to-noise ratio estimation for speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process*., vol. 14, no. 6, pp. 2098–2108, Nov. 2006.

[62] M. Parvaix, L. Girin, and J-M. Brossier, “A watermarking-based method for informaed source separation of audio signals with a single sensor”, *IEEE Trans. Audio Speech and Lang. Process.*, vol. 18, no.6, pp. 1464-1475, Aug. 2010.