Online Noisy Single-Channel Source Separation Using Adaptive Spectrum Amplitude Estimator and Masking

N. Tengtrairat, W. L. Woo, *Senior Member, IEEE*, S. S. Dlay, and Bin Gao, *Senior Member, IEEE*

***Abstract*—A novel single-channel source separation method is presented to recover the original signals given only a single observed mixture in noisy environment. The proposed separation method is an online adaptive process and independent of parameters initialization. In this paper, a noisy pseudo-stereo mixing model is developed by formulating an artificial mixture from the observed mixture where the signals are modeled by the autoregressive process. The proposed demixing process composes of two steps: First, the noisy mixing model is enhanced by selecting the time-frequency (TF) units of signal presence and computing the mixture spectral amplitude, and second, an adaptive estimation of the parameters associated with each source is computed frame-by-frame, which is then used to construct a TF mask for the separation process. To assess the performance of the proposed method, noisy mixtures of real-audio sources with nonstationary noise have been conducted under various SNRs. Experiments show that the proposed algorithm has yielded superior separation performance especially in low input SNR compared with existing methods.**

***Index Terms*—Blind source separation, masking, noise reduction, single-channel separation, underdetermined mixture.**

I. Introduction

SINGLE-CHANNEL blind source separation (SCBSS) is the process of recovering underlying source signals from an unknown mixing given only a single sensor without any prior information of source signals. SCBSS has interested many researchers during the last decade. In the field of biomedical signal processing, SCBSS is used in several different areas. Applications of ECG/EEG recordings given by the electromyography (EMG) signal have been developed to distinguish heartbeat signal from an observed recording based on diverse approaches, i.e., independent component analysis (ICA), nonnegative matrix factorization (NMF), singular spectrum analysis (SSA) [1]–[3]. Conventional ICA approach cannot be directly applied to a single-channel source separation. Thus, modified ICA methods were proposed. Single-channel independent component analysis (SCICA) approach in [4] 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 algorithm has certain limitations. For example, signals are assumed to be statistically independent. Secondly mixtures compose of non-overlapping spectrum-density signals. SCBSS of EEG recoding based on singular spectrum analysis (SSA) was proposed in [5]. SSA decomposes a time series into a number of interpretable components with distinct subspaces and selects the subgroup of eigenvalues to reconstruct the original source. Another recent application of the SCBSS is image separation in the field of non-destructive test and evaluation (NDT&E) [6]–[8]. In NDT&E, researchers are interested with the study of defects. Imaging technique is used usually to image the target object when excited by an external signal. The captured image is a result of a superposition of several independent events where each event is associated with a particular physics phenomenon. The aim is to estimate these independent events and monitor the associated physical features in order to detect and monitor defects.

In general, SCBSS can be categorized into two groups, i.e., model-based and data-driven methodologies. In this study, we focus on data-driven SCBSS. A popular method is the computational auditory scene analysis (CASA). CASA has been proposed for the isolation of speech from noise by using the ideal binary masking (IBM) in time-frequency domain. A binary masking approach has been introduced to suppress noise from the noisy input and also maintain speech intelligibility. In [9], this method consists of two phases: Firstly, training phase evaluates an ideal binary masking (IBM) by using a Gaussian mixture model (GMM) to label each TF unit whether speech-dominant or noise-dominant. Secondly, an enhancement phase is to construct a binary masking by using the IBM. Later in [10], a new binary-masking algorithm trained using deep neural networks (DNNs) with unsupervised restricted Boltzmann machines (RBMs) is proposed to improve the intelligibility of hearing-impaired listeners by separation of speech from noise through IBM estimation. Extension of GMM with user-generated exemplar source is proposed in [11]. This work uses an exemplar source provided from an external user to estimate the sources. Data-driven methods such as the sparse non-negative matrix factorization (SNMF) [12], [13] determine

Manuscript received August 12, 2013; revised August 11, 2014 and March 26, 2015; accepted May 13, 2015. Date of publication September 07, 2015; date of current version February 22, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Raviv Raich.

N. Tengtrairat is with Department of Software Engineering, Payap University, Chiang Mai, Thailand.

W. L. Woo and S. S. Dlay are with School of Electrical and Electronic Engineering, Newcastle University, England NE1 7RU, U.K. (e-mail: w.l.woo@ncl.ac.uk).

B. Gao is with School of Automation, University of Electronic Science and Technology of China, Chengdu 610054, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2015.2477059

a set of basis for each speaker and a mixture is mapped onto the joint bases of the speakers. It requires no assumption on sources such as statistical independence or grammatical model. However, the SNMF method does not model the temporal structure [14] and it requires large amount of computation to determine the speaker independent basis. The SNMF2D [15] was proposed which used a double convolution to model both spreading of spectral basis and variation of temporal structure inherent in the sources. Some successes have already been reported in recent literature [16]–[19] to show the validity of SNMF2D in separating single channel mixture. The SNMF has regained interest recently where the domain of interest lies in the complex spectrogram which gives rise to the complex NMF (CNMF). Some promising results have recently been reported in [20] with adaptive sparseness. On the other hand, binaural source separation method generally delivers better separation performance than a single recorder in the underdetermined scenario. The Degenerate Unmixing Estimation Technique (DUET) [21] and its variants [22], [23] 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 part of the clustering process. The DUET algorithm has been demonstrated to recover the underlying sparse sources given two anechoic mixtures in the TF domain. Recently, DUET has been extended to the single-channel mixture and the algorithm was termed as the Single Observation Likelihood estimatiOn (SOLO) [24], [25]. The SOLO constructs an artificial stereo mixture which is then used to form a binary mask for separation.

All of the above SCBSS algorithms are derived for noise-free condition which lacks the potential and robust to solve the problem in noisy environments. Since the presence of noise seriously degrades the performance, many algorithms for handling background noise have been developed. In a realistic situation of audio applications, desired signals will be corrupted by an additive background noise. Mathematically, noisy single-channel blind source separation (NSCBSS) can be expressed as:

$$x(t) = s_1(t) + s_2(t) + \cdots + s_N(t) + n(t) \tag{1}$$

where $t = 1, 2, \ldots, T$ denotes time index, $n(t)$ is unknown noise signal and the goal is to estimate the sources $s_n(t)$, $\forall n \in N$ of length T when only the observation signal $x(t)$ is available. A well-known approach to improve intelligibility and perceptual quality of degraded speech is a speech enhancement approach. The speech enhancement approach is to remove background noise in a noisy speech. Most of the common enhancement techniques operate in the frequency domain which can generally be expressed as

$$X(\tau, \omega) = S(\tau, \omega) + N(\tau, \omega) \tag{2}$$

where $X(\tau, \omega)$ is an observed noisy mixture at the ωth frequency bin of the τth frame, $S(\tau, \omega) = \sum_{i=1}^{N} S_i(\tau, \omega)$ is a sum of the source signals (i.e., mixture signal without noise), and $N(\tau, \omega)$ denotes the noise. An enhanced spectrum of mixture signal $\tilde{S}(\tau, \omega)$ is given as $\tilde{S}(\tau, \omega) = G(\tau, \omega)X(\tau, \omega)$ where $G(\tau, \omega)$ is a spectral gain. Hence, speech-enhancement performance depends solely on the spectral gain by applying a frequency-dependent gain function to the spectral components of the noisy speech, in an effort to suppress the noise components to higher quality of speech components. Many approaches have been established in recent decades, for example the spectral subtraction method, minimum-mean square error (MMSE) estimation, and a maximum *a posteriori* (MAP) estimation. The spectral subtraction method [26] achieves noise reduction by subtracting estimated noise spectral amplitude from the observed spectral amplitude without concern of speech spectral components. Secondly, the MMSE estimator [27] and its more recent versions [28] apply a frequency dependent gain function to the spectral components of the noisy speech. Its solution is featured by the noise variance, *a priori* SNR, and *a posteriori* SNR where the noise variance is known or can be estimated. Lastly, the speech enhancement method using a maximum *a posteriori* (MAP) estimation [29], [30] modeled the speech probability density function (PDF) by a parametric super-Gaussian function developed from a histogram. This method has an effective noise reduction capability especially in low SNR environments which is superior among the three methods.

In the paper, we consider the NSCBSS problem as one noisy mixture of N unknown sources signals. The contributions of the paper are summarized below: 1) It is an online adaptive separation method where the observed mixture is segmented into small frames. The separation process is executed adaptively frame-by-frame. Hence, the robustness of the proposed algorithm can benefit for 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 computational simplicity and does not exploit high-order statistic. Hence this yields the benefit of ease of implementation. To achieve the above, the proposed method requires the following assumptions: the source signals are characterized as AR processes, the sources satisfy the windowed-disjoint orthogonality (WDO) and the local stationary of the time-frequency representation.

The overview of the proposed method is illustrated in Fig. 1 which is organized as follows: Section II introduces the noisy pseudo-stereo mixture model. Section III proposes an online demixing method, i.e., the mixture enhancement and the separation process. Section IV presents the separability of the pseudo-stereo model. Experimental results with a series of performance comparison with other SCBSS methods are conducted and discussed in Section V. Finally, Section VI concludes the paper.

II. Proposed Single—Channel Noisy Mixing Model

A. Proposed Pseudo-Stereo Noisy Mixture Model

In this paper, for simplicity we consider the case of a single-channel noisy mixture of two sources and a noise in time domain as

$$x_1(t) = s_1(t) + s_2(t) + n_1(t) \tag{3}$$

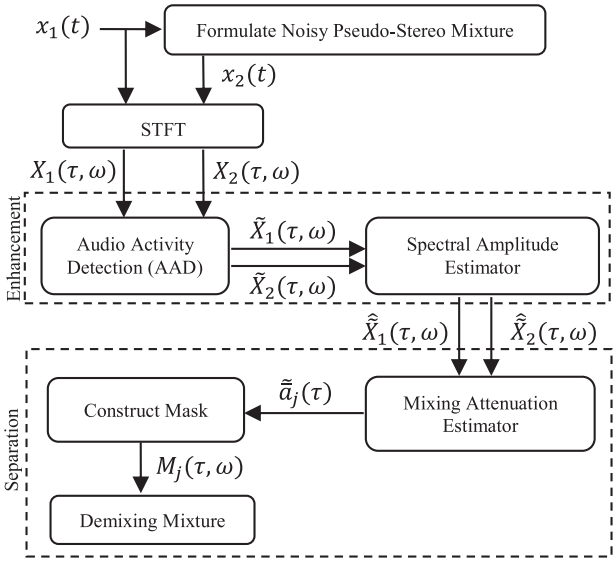


Fig. 1. Overview of the proposed algorithm.

where $x_1(t)$ is the single channel mixture, $n_1(t)$ is an additive uncorrelated noise that can be stationary or nonstationary, and $s_1(t)$ and $s_2(t)$ are the original source signals which are assumed to be modeled by the autoregressive (AR) process [31]:

$$s_j(t) = -\sum_{m=1}^{D_j} a_{s_j}(m;t)s_j(t-m) + e_j(t) \tag{4}$$

where $a_{s_j}(m;t)$ denotes the mth order AR coefficient of the jth source 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 $\sigma^2_{e_j}$. This model enables us to formulate a virtual mixture by weighting and time-shifting the single channel mixture $x_1(t)$ as

$$x_2(t) = \frac{x_1(t) + \gamma x_1(t-\delta)}{1+|\gamma|} \tag{5}$$

where $\gamma \in \Re$ is the weight parameter, and $\delta \in \mathbb{Z}$ is the time-delay. The noisy mixture in (3) and (5) 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 (5) in terms of the source signals, AR coefficient and time-delay as

$$\begin{aligned} x_2(t) &= \frac{x_1(t) + \gamma x_1(t-\delta)}{1+|\gamma|} \\ &= \frac{(-a_{s_1}(\delta;t)+\gamma)}{1+|\gamma|}s_1(t-\delta) + \frac{(-a_{s_2}(\delta;t)+\gamma)}{1+|\gamma|}s_2(t-\delta) \\ &\quad + \frac{e_1(t) - \sum_{\substack{m=1\\ m\neq\delta}}^{D_1} a_{s_1}(m;t)s_1(t-m)}{1+|\gamma|} \\ &\quad + \frac{e_2(t) - \sum_{\substack{m=1\\ m\neq\delta}}^{D_2} a_{s_2}(m;t)s_2(t-m)}{1+|\gamma|} \\ &\quad + \frac{n_1(t) + \gamma n_1(t-\delta)}{1+|\gamma|},\ \delta \in \mathbb{Z} \end{aligned} \tag{6}$$

Defining the followings:

$$a_j(t;\delta,\gamma) = \frac{-a_{s_j}(\delta;t)+\gamma}{1+|\gamma|} \tag{7}$$

$$r_j(t;\delta,\gamma) = \frac{e_j(t) - \sum_{\substack{m=1\\ m\neq\delta}}^{D_j} a_{s_j}(m;t)s_j(t-m)}{1+|\gamma|} \tag{8}$$

$$n_2(t;\delta,\gamma) = \frac{n_1(t) + \gamma n_1(t-\delta)}{1+|\gamma|} \tag{9}$$

where $a_j(t;\delta,\gamma)$ and $r_j(t;\delta,\gamma)$ represent the mixing attenuation and the residue of the jth source, respectively, and $n_2(t;\delta,\gamma)$ denotes noise obtained by weighting and time-shifting of the additive noise $n_1(t)$. Using (7)–(9), the overall proposed noisy mixing model can now be formulated in terms of the sources and the noise as

$$\begin{aligned} x_1(t) &= s_1(t) + s_2(t) + n_1(t) \\ x_2(t) &= a_1(t;\delta,\gamma)s_1(t-\delta) + a_2(t;\delta,\gamma)s_2(t-\delta) \\ &\quad + r_1(t;\delta,\gamma) + r_2(t;\delta,\gamma) + n_2(t;\delta,\gamma) \end{aligned} \tag{10}$$

B. Time-Frequency Representation

The TF representation of the noisy mixing model is obtained using the Short-Time Fourier Transform (STFT) of $x_j(t)$, $j = 1, 2$ as

$$\begin{aligned} X_1(\tau,\omega) &= S_1(\tau,\omega) + S_2(\tau,\omega) + N_1(\tau,\omega) \\ X_2(\tau,\omega) &\approx a_1(\tau)e^{-i\omega\delta}S_1(\tau-\delta,\omega) \\ &\quad + a_2(\tau)e^{-i\omega\delta}S_2(\tau-\delta,\omega) - \\ &\quad \cdot \left(\sum_{\substack{m=1\\ m\neq\delta}}^{D_1} \frac{a_{s_1}(m;\tau)}{1+|\gamma|} e^{-i\omega m} S_1(\tau-m,\omega) \right. \\ &\quad \left. + \sum_{\substack{m=1\\ m\neq\delta}}^{D_2} \frac{a_{s_2}(m;\tau)}{1+|\gamma|} e^{-i\omega m} S_2(\tau-m,\omega) \right) \\ &\quad + N_2(\tau,\omega) \end{aligned} \tag{11}$$

for $\forall\tau,\omega$. In (11), we have used the fact that $|e_j(t)| \ll |s_j(t)|$, thus the TF of $r_j(t)$ in (13) can be simplified to

$$R_j(\tau,\omega) = -\sum_{\substack{m=1\\ m\neq\delta}}^{D_j} \frac{a_{s_j}(m;\tau)}{1+|\gamma|} e^{-i\omega m} S_j(\tau-m,\omega) \tag{12}$$

To facilitate further analysis, we also define

$$C_j(\tau,\omega) = \frac{1}{1+|\gamma|} \sum_{\substack{m=1\\ m\neq\delta}}^{D_j} a_{s_j}(m;\tau)e^{-i\omega(m-\delta)} \tag{13}$$

which forms a part of $R_j(\tau,\omega)$ without the contribution of the source $S_j(\tau,\omega)$. Notice that factor $e^{-i\omega\delta}$ is only uniquely specified if $|\omega\delta| < \pi$, otherwise this would cause phase-wrap [32]. Selecting improper time-delay δ will lead to phase-wrap if the maximum frequency of the source is exceeded. In order to avoid phase ambiguity, we must satisfy

$$|\omega_{max}\delta_{max}| < \pi \tag{14}$$

where $\omega_{max} = 2\pi f_{max}/f_s$, δ_{max} is the maximum time delay, f_{max} is the maximum frequency present in the sources and f_s is the sampling frequency. Hence, δ_{max} can be determined from (14) according to

$$\delta_{max} < \frac{f_s}{2f_{max}} \tag{15}$$

As long as the delay parameter is less than δ_{max}, there will not be any phase ambiguity. This condition will be used to determine the range of δ in formulating the pseudo-stereo mixture.

III. Proposed Online Single—Channel Noisy Demixing Method

The proposed online single-channel noisy demixing method mainly comprises of two steps: The first step is mixture enhancement which aims to reduce the additive noise and extracts the source information. 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 source-signature estimator.

A. Proposed Single-Channel Mixture Enhancement

1) Audio Activity Detection: The audio activity detection (AAD) method enhances the noisy mixture by selecting the TF units that contain source signals and removing TF units without source signals. To begin, the two statistical hypotheses are set, i.e., $H_0(\tau,\omega)$ and $H_1(\tau,\omega)$ which denote the source absence and presence at ωth frequency bin of the τth frame, respectively.

$$\begin{aligned} &H_0(\tau,\omega): \text{Source absence}: \quad X(\tau,\omega) = N(\tau,\omega) \\ &H_1(\tau,\omega): \text{Source presence}: \; X(\tau,\omega) = S(\tau,\omega) + N(\tau,\omega) \end{aligned} \tag{16}$$

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 source signals, i.e., $S(\tau,\omega) = S_1(\tau,\omega) + S_2(\tau,\omega)$, and $N(\tau,\omega)$ is the additive noise. The term $S(\tau,\omega)$ and $N(\tau,\omega)$ are assumed to be complex Gaussian distributed. Source presence at a particular (τ,ω) unit is detected by computing a local source absence probability (LSAP) and selecting the (τ,ω) 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

$$\begin{aligned} &p\left(H_0(\tau,\omega)|X(\tau,\omega)\right) \\ &= \frac{p(X(\tau,\omega)|H_0(\tau,\omega))p(H_0)}{p\left(X(\tau,\omega)|H_0(\tau,\omega)\right)p(H_0) + p\left(X(\tau,\omega)|H_1(\tau,\omega)\right)p(H_1)} \\ &= \frac{1}{1+q_\omega \Lambda(\tau,\omega)} \end{aligned} \tag{17}$$

where $p(\cdot)$ denotes a probability density function (PDF), q_ω is the ratio defined by $q_\omega = \frac{p(H_1)}{p(H_0)}$, and $p(H_0)$ and $p(H_1)$ are the prior probabilities of the respective hypotheses. The term $\Lambda(\tau,\omega) = p(X(\tau,\omega)|H_1(\tau,\omega))/p(X(\tau,\omega)|H_0(\tau,\omega))$ is the likelihood ratio of the source presence and source absence at (τ,ω) units defined as: $p(X(\tau,\omega)|H_1(\tau,\omega)) = \frac{1}{\pi(\sigma_S^2(\tau,\omega)+\sigma_N^2(\tau,\omega))}\exp\left(-\frac{|X(\tau,\omega)|^2}{\sigma_S^2(\tau,\omega)+\sigma_N^2(\tau,\omega)}\right)$ and $p(X(\tau,\omega)|H_0(\tau,\omega)) = \frac{1}{\pi\sigma_N^2(\tau,\omega)}\exp\left(-\frac{|X(\tau,\omega)|^2}{\sigma_N^2(\tau,\omega)}\right)$, respectively. In the case of $LSAP \geq T_L$, this particular (τ,ω) unit constitutes as noise. In order to update the noise power, a global source 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 τth frame can be expressed as (18), shown at the bottom of the page. When the GSAP exceeds a global threshold T_G, a noise power estimate is updated. Otherwise, the noise power estimate of the τth frame remains the same as in the previous frame. The noise power estimate can be computed as

$$\hat{\sigma}_N^2(\tau,\omega) = \zeta_N \hat{\sigma}_N^2(\tau-1,\omega) + (1-\zeta_N)\left|N(\tau,\omega)\right|^2 \tag{19}$$

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

In traditional voice activity detection (VAD) method [33], [34], the likelihood of the presence of the sources requires the source 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_ω ratio. This case causes LSAP and GSAP to be independent of the mixture. Therefore, LSAP and GSAP cannot correctly identify (τ,ω) 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 ξ_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 ξ_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 $p(X(\tau,\omega)|H_1(\tau,\omega))$ differs from $p(X(\tau,\omega)|H_0(\tau,\omega))$, 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

$$p\left(H_0(\tau)|X(\tau)\right) = \frac{p(H_0)\Pi_{\omega=1}^{W} p\left(X(\tau,\omega)|H_0(\tau)\right)}{p(H_0)\Pi_{\omega=1}^{W} p\left(X(\tau,\omega)|H_0(\tau)\right) + p(H_1)\Pi_{\omega=1}^{W} p\left(X(\tau,\omega)|H_1(\tau)\right)} = \frac{1}{1+q_\omega \Pi_{\omega=1}^{W}\Lambda(\tau,\omega)} \tag{18}$$

original signals. Additionally, using $\xi_f \sigma_N^2(\tau_N)$ instead $\sigma_S^2(\tau_N)$ 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 source presence can be expressed as

$$p\left(X(\tau,\omega)|H_1(\tau,\omega)\right) = \frac{1}{\pi\sigma_N^2(\tau,\omega)(1+\xi_f)} \exp\left\{-\frac{|X(\tau,\omega)|^2}{\sigma_N^2(\tau,\omega)(1+\xi_f)}\right\} \quad (20)$$

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

$$\begin{aligned} p_e(\xi,\xi_f) &= p(decide\ H_1|H_0)p(H_0) + p(decide\ H_0|H_1)p(H_1) \\ &= \int_0^{X_{T_L}(\tau,\omega)} \int_0^{2\pi} \frac{1}{\pi\sigma_N^2(\tau,\omega)} \\ &\quad \times \exp\left\{-\frac{|X_{T_L}(\tau,\omega)|^2}{\sigma_N^2(\tau,\omega)}\right\} (re^{j\theta}) r dr d\theta p(H_0) \\ &\quad + \int_{X_{T_L}(\tau,\omega)}^{\infty} \int_0^{2\pi} \frac{1}{\pi\sigma_N^2(\tau,\omega)(1+\xi)} \\ &\quad \times \exp\left\{-\frac{|X_{T_L}(\tau,\omega)|^2}{\sigma_N^2(\tau,\omega)(1+\xi)}\right\} (re^{j\theta}) r dr d\theta p(H_1) \\ &= \left(1 - \left(\frac{p(H_0)}{p(H_1)}(1+\xi_f)\right)^{\frac{1+\xi_f}{\xi_f(1+\xi)}}\right) p(H_0) \\ &\quad + \left(\frac{p(H_0)}{p(H_1)}(1+\xi_f)\right)^{\frac{1+\xi_f}{\xi_f}} p(H_1) \end{aligned} \quad (21)$$

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

$$\hat{\xi}_f = \underset{\xi_f}{\operatorname{argmin}} \int_{\xi_{down}}^{\xi_{top}} p_e(\xi,\xi_f) d\xi \quad (22)$$

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

The AAD method enables us to obtain the TF plane of the source-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 source spectral amplitude. In Section III-A-2, we will show how the spectral amplitude of sources can be extracted from the mixing model.

2) Mixture Spectral Amplitude Estimator: Let $\tilde{X}(\tau,\omega)$ denotes the mixture with source present at (τ,ω) units from the AAD method. This consists of the sum of the source signals and the residual noise $\tilde{N}(\tau,\omega)$, i.e.,

$$\tilde{X}(\tau,\omega) = S(\tau,\omega) + \tilde{N}(\tau,\omega) \quad (23)$$

where $\tilde{X}(\tau,\omega) = |\tilde{X}(\tau,\omega)|e^{i\theta\omega}$, $S(\tau\omega) = A(\tau,\omega)e^{i\alpha\omega}$ is the sum of the sources (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 source phase, respectively. The residual noise $\tilde{N}(\tau,\omega)$ refers to the remaining noise in the source-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., $\theta\omega = \alpha\omega$ [29]. The conventional MMSE-STSA estimator [29] 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:

$$\begin{aligned} \tilde{A}(\tau,\omega) &= E\left\{A(\tau,\omega)|\tilde{X}_1(\tau,\omega)\right\} \\ &= \frac{q_\omega \Lambda(\tau,\omega)}{1+q_\omega \Lambda(\tau,\omega)} \Gamma(1.5) \frac{\sqrt{v(\tau,\omega)}}{\gamma_{SNR}(\tau,\omega)} \exp\left(-\frac{v(\tau,\omega)}{2}\right) \\ &\quad \times \left[(1+v(\tau,\omega))\, I_0\left(\frac{v(\tau,\omega)}{2}\right)\right. \\ &\quad \left. + v(\tau,\omega) I_1\left(\frac{v(\tau,\omega)}{2}\right)\right] \left|\tilde{X}_1(\tau,\omega)\right| \end{aligned} \quad (24)$$

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 zeroth and first order, respectively. $v(\tau,\omega)$ is defined by $v(\tau,\omega) = \frac{\xi(\tau,\omega)}{1+\xi(\tau,\omega)}(\gamma_{SNR}(\tau,\omega))$, $\gamma_{SNR}(\tau,\omega) \triangleq \frac{|\tilde{X}(\tau,\omega)|^2}{\sigma_N^2(\tau,\omega)}$ 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 the estimates of $\gamma_{SNR}(\tau,\omega)$ and $\xi(\tau,\omega)$, i.e., $\hat{\gamma}(\tau,\omega)$ and $\hat{\xi}(\tau,\omega)$, respectively. These two parameters significantly influence the accuracy of the spectrum amplitude function (24). However, under the case of weak source components and low input SNR, the conventional $\gamma_{SNR}(\tau,\omega)$ estimator causes deterioration of the weak source components. We can analyze this case as follows:

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

Using the subadditivity properties of the absolute value, we obtain

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

In the case of weak source components and low inputs SNR, i.e., $\sigma_S^2(\tau,\omega) \approx 0$, we then have

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

The estimation of $\xi(\tau,\omega)$ can be shown to be given by $\hat{\xi}(\tau,\omega) = \zeta_\xi \hat{A}^2(\tau-1,\omega)/\sigma_N^2(\tau-1,\omega)+(1-\zeta_\xi)max\{\gamma_{SNR}(\tau,\omega)-1,0\}$ 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 $\hat{\gamma}_{SNR}$ based entirely on the current frame. The term ζ_ξ, $0 < \zeta_\xi < 1$, is a weighing factor that controls the trade-off between the noise reduction and the transient distortion brought into the signal. At a particular (τ,ω) unit of weak source activity and low input SNR where $\gamma_{SNR}(\tau,\omega) \leq 1$, 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 $(1-\zeta_\xi)max\{\gamma_{SNR}(\tau,\omega)-1,0\}$. Thus, $\hat{\xi}(\tau,\omega)$ depends only on the scaling of its previous frame without taking the scaled *a posteriori* SNR estimator into account $(1-\zeta_\xi)max\{\gamma_{SNR}(\tau,\omega)-1,0\}$. 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, we can improve the estimation of $\xi(\tau,\omega)$ by computing the *a posteriori* SNR parameter $\hat{\gamma}_{SNR}(\tau,\omega)$ from the source presence probability (SPP) with fixed *a priori* to guarantee that $\hat{\gamma}(\tau,\omega) > 1$. The term $p(H_1(\tau,\omega)|\tilde{X}(\tau,\omega))$ denotes a SPP given by the Bayes' theorem:

$$\begin{aligned} & p\left(H_1(\tau,\omega)|\tilde{X}(\tau,\omega)\right) \\ &= \frac{p\left(\tilde{X}(\tau,\omega)|H_1(\tau,\omega)\right)p(H_1)}{p\Big(\tilde{X}(\tau,\omega)|H_0(\tau,\omega)p(H_0)+p\Big(\tilde{X}(\tau,\omega)|H_1(\tau,\omega)\Big)\Big)p(H_1)} \\ &= \left(\frac{1+\xi_f}{q_\omega}\exp\left\{-E\frac{\left|\tilde{X}(\tau,\omega)\right|^2}{\hat{\sigma}_N^2(\tau,\omega)}\right\}+1\right)^{-1} \end{aligned} \tag{25}$$

where $E = \xi_f/1+\xi_f$. Eqn. (31) is solved for the *a posteriori* SNR based on ξ_f and $p(H_1(\tau,\omega)|\tilde{X}(\tau,\omega))$ as

$$\hat{\gamma}_{SNR}(\tau,\omega) = \frac{1}{E}\log\left(\frac{1}{q_\omega}\left(\frac{1+\xi_f}{p\left(H_1(\tau,\omega)|\tilde{X}(\tau,\omega)\right)^{-1}-1}\right)\right) \tag{26}$$

Using the ξ_f and $p(H_1(\tau,\omega)|\tilde{X}(\tau,\omega)) > 0.08$, the *a posteriori* SNR 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 source information even when source components are weak in low input SNR, the proposed iMMSE-STSA firstly estimate the *a posteriori* SNR using (26) and then using this estimate for computing the spectral amplitude. Finally, the estimated spectra of the mixture can be formulated as

$$\hat{S}(\tau,\omega) = \hat{A}(\tau,\omega)e^{i\theta\omega} \tag{27}$$

In conclusion, the proposed mixture enhancement method will benefit the source separation by providing the greater degree of source information by attempting to select the TF units of source presence and reject the TF units of solely noise. The noise-reduced mixture can now be modeled as $\tilde{\tilde{X}}(\tau,\omega) = \hat{A}(\tau,\omega)e^{i\theta\omega} + \tilde{\tilde{N}}(\tau,\omega)$ which will then be separated by a binary TF mask.

B. Proposed Single—Channel Source Separation

1) Adaptive Mixing Parameter Estimator: The sources are assumed to 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 ϕ is the maximum time-delay (shift) associated with the Short-Time Fourier Transform (STFT) $F^W(\cdot)$ with an appropriate window function $W(\cdot)$. If ϕ is small compared with the length of $W(\cdot)$ then $W(\cdot-\phi) \approx W(\cdot)$. Hence, the Fourier transform of a windowed function with shift ϕ yields approximately the same Fourier transform without ϕ. For the proposed method, the pseudo-stereo mixture is shifted by δ and by invoking the local stationarity this leads to

$$\begin{aligned} s_j(t-\delta) &\xrightarrow{STFT} e^{-i\omega\delta}S_j(\tau-\delta,\omega) \\ &\approx e^{-i\omega\delta}S_j(\tau,\omega),\ \forall\delta, |\delta| \leq \phi \end{aligned} \tag{28}$$

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. Secondly, assuming that the sources satisfy the windowed-disjoint orthogonality (WDO) condition:

$$S_i(\tau,\omega)S_j(\tau,\omega) \approx 0, \quad \forall i \neq j,\ \forall\tau,\omega \tag{29}$$

where $S_i(\tau,\omega)$ and $S_j(\tau,\omega)$ are the STFT of $s_i(t)$ and $s_j(t)$. Hence, the jth source is dominant at a particular (τ,ω) unit, the noise-reduced mixture can be more specifically expressed as:

$$\begin{aligned} \hat{\tilde{X}}_1(\tau,\omega) &= \hat{S}_j(\tau,\omega) + \tilde{N}_1(\tau,\omega) \\ \hat{\tilde{X}}_2(\tau,\omega) &= a_j(\tau)e^{-i\omega\delta}\hat{S}_j(\tau-\delta,\omega) \\ &\quad - \sum_{\substack{m=1 \\ m\neq\delta}}^{D_j}\frac{a_{s_j}(m;\tau)}{1+|\gamma|}e^{-i\omega m}\hat{S}_j(\tau-m,\omega)+\tilde{N}_2(\tau,\omega) \\ &\approx \left[a_j(\tau)-C_j(\tau,\omega)\right]e^{-i\omega\delta}\hat{S}_j(\tau,\omega)+\tilde{N}_2(\tau,\omega), \\ &\quad (\tau,\omega) \in \Omega_j \end{aligned} \tag{30}$$

for δ and $m \leq \phi$. The term $C_j(\tau,\omega) = \frac{1}{1+|\gamma|}\sum_{m=1}^{D_j} a_{s_j}(m;\tau)e^{-i\omega(m-\delta)}$ is given by (13) and Ω_j is the jth source presence area defined as $\Omega := \{(\tau,\omega) : \hat{S}_j(\tau,\omega) \neq 0,\ \forall k \neq j\}$. The estimate of $\bar{a}_j(\tau,\omega) = a_j(\tau) - C_j(\tau,\omega)$ associated with the jth source can be determined as

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

$\widetilde{N}_1(\tau,\omega)$ and $\widetilde{N}_2(\tau,\omega)$ can be assumed to be small after the mixture enhancement step (as shown in Section V-B). In this case, we can expressed $\bar{a}_j(\tau,\omega)$ as

$$\begin{aligned}\bar{a}_j(\tau,\omega) &= \frac{[a_j(\tau)-C_j(\tau,\omega)]\,e^{-i\omega\delta}\hat{S}_j(\tau,\omega)}{S_j(\tau,\omega)}e^{i\omega\delta}\\ &= a_j(\tau)-C_j(\tau,\omega)\\ &= \bar{a}_j^{(r)}(\tau,\omega)+i\bar{a}_j^{(i)}(\tau,\omega),\ \forall(\tau,\omega)\in\Omega_j \end{aligned} \quad (31)$$

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 $i=\sqrt{-1}$. We propose to adaptively estimate $\bar{a}_j(\tau,\omega)$ frame-by-frame. Firstly, a power weighted TF histogram will be used to estimate $\bar{a}_j(\tau,\omega)$ for each frame and the TF units are then clustered into a number of groups corresponding to the number of sources in the mixture. The power weighted histogram is a function of (τ,ω) with the weight $\sum|\hat{\tilde{X}}_1(\tau,\omega)\hat{\tilde{X}}_2(\tau,\omega)|$ therefore the real and imaginary parts of $\bar{a}_j(\tau,\omega)$ for each frame basis can be estimated as

$$\begin{aligned}\hat{\bar{a}}_j^{(r)}(\tau) &= \frac{\sum_\omega\left|\hat{\tilde{X}}_1(\tau,\omega)\hat{\tilde{X}}_2(\tau,\omega)\right| Re\left[\frac{\hat{\tilde{X}}_2(\tau,\omega)}{\hat{\tilde{X}}_1(\tau,\omega)}e^{i\omega\delta}\right]}{\sum_\omega\left|\hat{\tilde{X}}_1(\tau,\omega)\hat{\tilde{X}}_2(\tau,,\omega)\right|}\\ \hat{\bar{a}}_j^{(i)}(\tau) &= \frac{\sum_\omega\left|\hat{\tilde{X}}_1(\tau,\omega)\hat{\tilde{X}}_2(\tau,\omega)\right| Im\left[\frac{\hat{\tilde{X}}_2(\tau,\omega)}{\hat{\tilde{X}}_1(\tau,\omega)}e^{i\omega\delta}\right]}{\sum_\omega\left|\hat{\tilde{X}}_1(\tau,\omega)\hat{\tilde{X}}_2(\tau,\omega)\right|}\end{aligned} \quad (32)$$

The above can then be combined to form the estimate of (32) as

$$\hat{\bar{a}}_j(\tau) = \hat{\bar{a}}_j^{(r)}(\tau)+i\hat{\bar{a}}_j^{(i)}(\tau) \quad (33)$$

Relating (33) with (31), we can use similar idea to express $\hat{\bar{a}}(\tau)=\hat{a}_j(\tau)-\hat{C}_j(\tau)$ where $\hat{a}_j(\tau)$ 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 $\tilde{\bar{a}}_j(\tau-1)$ and $\hat{\bar{a}}_j(\tau)$:

$$\tilde{\bar{a}}_j(\tau) = \zeta_M\tilde{\bar{a}}_j(\tau-1)+(1-\zeta_M)\hat{\bar{a}}_j(\tau) \quad (34)$$

where $0<\zeta_M<1$ is a smoothing parameter of the adaptive mixing attenuation estimator.

2) Construction of Masks: The binary TF masks can be constructed by labeling each TF unit with the k argument through maximizing the instantaneous likelihood function. The instantaneous likelihood function is derived from the maximum likelihood (ML) method by first formulating the Gaussian likelihood function $p(\hat{\tilde{X}}_1(\tau,\omega),\hat{\tilde{X}}_2(\tau,\omega)|S_j(\tau,\omega),\tilde{\bar{a}}_j(\tau),\sigma^2_{\widetilde{N}_j})$ using (30), 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 assumes the following form:

$$\begin{aligned}&L_j(\tau,\omega)\\ &:= p\left(\hat{\tilde{X}}_1(\tau,\omega),\hat{\tilde{X}}_2(\tau,\omega)\middle|\, S_j(\tau,\omega),\tilde{\bar{a}}_j(\tau),\sigma^2_{\widetilde{N}_j}\right)\\ &= \frac{1}{2\pi}\exp\left(-\frac{1}{2}\frac{\left|\tilde{\bar{a}}_j(\tau)e^{-i\omega\delta}\hat{\tilde{X}}_1(\tau,\omega)-\hat{\tilde{X}}_2(\tau,\omega)\right|^2}{\hat{\sigma}^2_{\widetilde{N}_2}(\tau,\omega)+\hat{\sigma}^2_{\widetilde{N}_1}(\tau,\omega)\tilde{\bar{a}}_j^2(\tau)}\right)\end{aligned} \quad (35)$$

The function $L_j(\tau,\omega)$ clusters every (τ,ω) unit to the jth dominating source for $L_j(\tau,\omega)\geq L_k(\tau,\omega),\ \forall l\neq j$. This process is equivalent to the following minimization problem:

$$F(\tau,\omega)=\underset{k}{\operatorname{argmin}}\frac{\left|\tilde{\bar{a}}_k(\tau)e^{-i\omega\delta}\hat{\tilde{X}}_1(\tau,\omega)-\hat{\tilde{X}}_2(\tau,\omega)\right|^2}{\hat{\sigma}^2_{\widetilde{N}_2}(\tau,\omega)+\hat{\sigma}^2_{\widetilde{N}_1}(\tau,\omega)\tilde{\bar{a}}_k^2(\tau)} \quad (36)$$

Using (30), the term $\hat{\tilde{X}}_2(\tau,\omega)$ can be expressed as:

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

By invoking the local stationarity, we then obtain

$$\hat{\tilde{X}}_2(\tau,\omega)=\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\left(\hat{S}_j(\tau,\omega)+\tilde{N}_1(\tau,\omega)\right)-\frac{E_j(\tau,\omega)}{1+|\gamma|} \quad (37)$$

for $\delta\leq\phi$. The derivation of $\hat{\tilde{X}}_2(\tau,\omega)$ in the source domain in (37) allows us to express $\hat{\tilde{X}}_2(\tau,\omega)$ in the mixture domain as:

$$\hat{\tilde{X}}_2(\tau,\omega)\approx\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\hat{\tilde{X}}_1(\tau,\omega)-\frac{E_j(\tau,\omega)}{1+|\gamma|} \quad (38)$$

In this light, the proposed cost function $G_k(\tau,\omega)$ can be formulated based on the single mixture $\hat{\tilde{X}}_1(\tau,\omega)$ by substituting this expression into (36) which leads to

$$J(\tau,\omega)=\arg\min_k G_k(\tau,\omega) \quad (39)$$

$$G_k(\tau,\omega)=\left|\frac{\tilde{\bar{a}}_k(\tau)e^{-i\omega\delta}\hat{\tilde{X}}_1(\tau,\omega)-\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\hat{\tilde{X}}_1(\tau,\omega)}{\hat{\sigma}^2_{\widetilde{N}_2}(\tau,\omega)+\hat{\sigma}^2_{\widetilde{N}_1}(\tau,\omega)\tilde{\bar{a}}_k^2(\tau)}\right|^2 \quad (40)$$

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 (39) and (40), in the instance when the jth source dominates at $(\tau,\omega)\in\Omega_j$, the function $J(\tau,\omega)$ will correctly identify the source if and only if $G_{k=j}(\tau,\omega)<G_{k\neq j}(\tau,\omega)$. To elucidate

TABLE I
OVERVIEW PROPOSED ALGORITHM

1. **Pseudo-Stereo Mixture step:** Formulate the pseudo-stereo mixture $x_2(t)$ using (5).
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 τ^{th} frame bin and the ω^{th} frequency of two mixtures using (17) and the global SAP for the τ^{th} frame using (18). If the global SAP $> T_G$ then updates $\hat{\sigma}^2_{\widetilde{N}_j}(\tau,\omega)$ using (19).
   2) iMMSE-STSA Estimator: Compute the iMMSE estimator of the source spectral amplitude using (24) and formulate the estimated spectra of the j^{th} sources $\hat{S}(\tau,\omega)$ using (27) for both mixtures.
   B. Separation step:
   1) Compute the mixing attenuation estimators $(\tilde{\hat{a}}_j^{(r)}(\tau), \tilde{\hat{a}}_j^{(i)}(\tau))$ at the τ^{th} frame using (32) and (34).
   2) Label (τ,ω) units using (39) and (40), and form the binary TF mask $M_j(\tau,\omega)$. Recover the original sources by
   $$\hat{\tilde{S}}_j(\tau,\omega) = M_j(\tau,\omega)\hat{\tilde{X}}_1(\tau,\omega) \quad (47)$$
   Finally, convert the estimated sources from TF domain into time domain i.e. $\hat{\tilde{s}}_j(t)$.

this condition, firstly, the case when $k = j$ is considered by setting $\zeta_M = 0$:

$$\begin{aligned}
&G_{k=j}(\tau,\omega)\\
&= \left|\tilde{\hat{a}}_j(\tau)e^{-i\omega\delta}\left(\hat{S}_j(\tau,\omega)+\widetilde{N}_1(\tau,\omega)\right)\right.\\
&\quad \left.-\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\left(\hat{S}_j(\tau,\omega)+\widetilde{N}_1(\tau,\omega)\right)\right|^2\\
&= \left|\hat{a}_j(\tau)e^{-i\omega\delta}\hat{S}_j(\tau,\omega)-\hat{C}_j(\tau)e^{-i\omega\delta}\hat{S}_j(\tau,\omega)\right.\\
&\quad +\tilde{\hat{a}}_j(\tau)e^{-i\omega\delta}\widetilde{N}_1(\tau,\omega)\\
&\quad \left.-\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\hat{S}_j(\tau,\omega)-\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\widetilde{N}_1(\tau,\omega)\right|^2\\
&= \left|-\left(\hat{C}_j(\tau)+\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right)e^{-i\omega\delta}\hat{S}_j(\tau,\omega)\right.\\
&\quad +\tilde{\hat{a}}_j(\tau)e^{-i\omega\delta}\widetilde{N}_1(\tau,\omega)\\
&\quad \left.-\frac{\hat{S}_j(\tau,\omega)}{1+|\gamma|}+\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\widetilde{N}_1(\tau,\omega)\right|^2 \qquad (41)
\end{aligned}$$

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

$$\begin{aligned}
G_{k\neq j}(\tau,\omega) = &\left|\left(\tilde{\hat{a}}_k(\tau)-a_j(\tau)-\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right)e^{-i\omega\delta}\hat{S}_j(\tau,\omega)\right.\\
&+\tilde{\hat{a}}_k(\tau)e^{-i\omega\delta}\widetilde{N}_1(\tau,\omega)\\
&\left.-\frac{\hat{S}_j(\tau,\omega)}{1+|\gamma|}+\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\widetilde{N}_1(\tau,\omega)\right|^2 \qquad (42)
\end{aligned}$$

To guarantee that $G_{k=j}(\tau,\omega) < G_{k\neq j}(\tau,\omega)$ is always satisfied, then we must specified a condition for $\hat{C}_j$. Starting with (41) and (42), we have

$$\begin{aligned}
&\left|-\left(\hat{C}_j(\tau)+\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right)e^{-i\omega\delta}\hat{S}_j(\tau,\omega)+\tilde{\hat{a}}_j(\tau)e^{-i\omega\delta}\widetilde{N}_1(\tau,\omega)\right.\\
&\left.-\left(\frac{\hat{S}_j(\tau,\omega)}{1+|\gamma|}+\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\widetilde{N}_1(\tau,\omega)\right)\right|^2\\
&<\left|\left(\tilde{\hat{a}}_k(\tau)-a_j(\tau)-\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right)e^{-i\omega\delta}\hat{S}_j(\tau,\omega)\right.\\
&\quad +\tilde{\hat{a}}_k(\tau)e^{-i\omega\delta}\widetilde{N}_1(\tau,\omega)\\
&\quad \left.-\left(\frac{\hat{S}_j(\tau,\omega)}{1+|\gamma|}+\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\widetilde{N}_1(\tau,\omega)\right)\right|^2 \qquad (43)
\end{aligned}$$

Eq. (43) is bounded by

$$\begin{aligned}
&\left|\hat{C}_j(\tau)\hat{S}_j(\tau,\omega)\right|-\left|\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\hat{S}_j(\tau,\omega)-\tilde{\hat{a}}_j(\tau)\widetilde{N}_1(\tau,\omega)\right|\\
&-\left|\frac{\hat{S}_j(\tau,\omega)}{1+|\gamma|}+\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\widetilde{N}_1(\tau,\omega)\right|\\
&<\left|\left(\tilde{\hat{a}}_k(\tau)-a_j(\tau)-\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right)\hat{S}_j(\tau,\omega)+\tilde{\hat{a}}_k(\tau)\widetilde{N}_1(\tau,\omega)\right|\\
&\quad +\left|\frac{\hat{S}_j(\tau,\omega)}{1+|\gamma|}+\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)\widetilde{N}_1(\tau,\omega)\right|
\end{aligned}$$

and therefore we obtain

$$\begin{aligned}
\left|\hat{C}_j(\tau)\right| < &\left|\left(\tilde{\hat{a}}_k(\tau)-a_j(\tau)-\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right)+\tilde{\hat{a}}_k(\tau)\frac{\widetilde{N}_1(\tau,\omega)}{\hat{S}_j(\tau,\omega)}\right|\\
&+\left|\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}-\tilde{\hat{a}}_j)(\tau)\frac{\widetilde{N}_1(\tau,\omega)}{\hat{S}_j(\tau,\omega)}\right|\\
&+\frac{2}{1+|\gamma|}\left|1+(1+\gamma e^{-i\omega\delta})\frac{\widetilde{N}_1(\tau,\omega)}{\hat{S}_j(\tau,\omega)}\right| \qquad (44)
\end{aligned}$$

for $\forall j \neq k$. As $\widetilde{N}_1(\tau,\omega)$ has small energy compared with source energy they can be treated as negligible. Hence, (44) can be simplified to

$$\left|\hat{C}_j(\tau)\right| < \left|\left(\tilde{\hat{a}}_k(\tau)-a_j(\tau)-\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right)\right|+\left|\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right|+\frac{2}{1+|\gamma|} \qquad (45)$$

If the condition in (45) is satisfied across Ω_j, the function (39), (40) will then correctly assign the (τ,ω) unit to the jth source. Once the TF plane of the mixtures are assigned into k groups of (τ,ω) units, the binary TF mask for the jth source can then be constructed as

$$M_j(\tau,\omega) := \begin{cases} 1 & J(\tau,\omega) = j \\ 0 & otherwise. \end{cases} \qquad (46)$$

The proposed algorithm is summarized in Table I.

IV. Analysis of Separability of the Proposed Pseudo-Stereo Mixture Model

The separability of the noise-free mixing model can be examined from the noise-free pseudo-stereo mixture by considering $a_j(t;\delta,\gamma)$ and $r_j(t;\delta,\gamma)$ in the following three cases. Case 1 refers to identical sources mixed in the single channel, Case 2 represents different sources but setting γ and δ for the pseudo-stereo mixture such that $a_1(t;\delta,\gamma) = a_2(t;\delta,\gamma)$, and Case 3 corresponds to the most general case where the sources are distinct, and γ and δ are selected arbitrarily such that the mixing attenuations and residues are also different. The above cases are demonstrated by using the functions $J(\tau,\omega)$ and $G_k(\tau,\omega)$ from Section III-B-2). These function are recapped here as:

$$J(\tau,\omega) = \underset{k}{\operatorname{argmin}}\, G_k(\tau,\omega) \tag{48}$$

$$G_k(\tau,\omega) = \left|\bar{a}_k(\tau)e^{-i\omega\delta}X_1(\tau,\omega) - \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)X_1(\tau,\omega)\right|^2 \tag{49}$$

For each TF unit, the kth argument that gives the minimum cost will be assigned to the kth source. We may analyze (49) further by assuming that the jth source dominates at a particular TF unit. In this case, the observed mixture in TF domain reduces to $X_1(\tau,\omega) = S_j(\tau,\omega)$ and therefore, (49) becomes

$$\begin{aligned}
&G_k(\tau,\omega)\\
&= \left|\bar{a}_k(\tau)e^{-i\omega\delta}S_j(\tau,\omega) - \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)S_j(\tau,\omega)\right|^2\\
&= \left|\bar{a}_k(\tau)e^{-i\omega\delta}S_j(\tau,\omega) - \frac{S_j(\tau,\omega)}{1+|\gamma|} - \frac{\gamma e^{-i\omega\delta}}{1+|\gamma|}S_j(\tau,\omega)\right|^2\\
&= \Big|a_k(\tau)e^{-i\omega\delta}S_j(\tau,\omega) - C_k(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega)\\
&\quad + \sum_{\substack{m=1\\ m\neq\delta}} \frac{a_{s_j}(m;\tau)e^{-i\omega m}}{1+|\gamma|}S_j(\tau-m,\omega)\\
&\quad - a_j(\tau)e^{-i\omega\delta}S_j(\tau,\omega)\Big|^2
\end{aligned} \tag{50}$$

We consider the following three cases:

Case 1: 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 $x_2(t;\delta,\gamma) = \left(\frac{-a(\delta;t)+\gamma}{1+|\gamma|}\right)x_1(t-\delta) + 2r(t;\delta,\gamma)$.

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 1 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 1, the function $J(\tau,\omega)$ given by (50) becomes:

$$\begin{aligned}
J(\tau,\omega) = \underset{k}{\operatorname{argmin}}\Big|&a(\tau)e^{-i\omega\delta}S_j(\tau,\omega) - C(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega)\\
&+ \sum_{\substack{m=1\\ m\neq\delta}}^{D} \frac{a_s(m;\tau)e^{-i\omega m}}{1+|\gamma|}S_j(\tau-m,\omega) - a(\tau)e^{-i\omega\delta}S_j(\tau,\omega)\Big|^2
\end{aligned}$$

Invoking the local stationarity of the sources $S_j(\tau - D_j,\omega) = S_j(\tau,\omega)$ for $|D_j| \le \phi$, the above leads to

$$\begin{aligned}
J(\tau,\omega) &= \underset{k}{\operatorname{argmin}}\left|\sum_{\substack{m=1\\ m\neq\delta}}^{D}\frac{\left(a_s(m;\tau)e^{-i\omega m} - a_s(m;\tau)e^{-i\omega m}\right)}{1+|\gamma|}\right|^2\\
&\quad\times |S_j(\tau,\omega)|^2\\
&= 0 \text{ for } \forall k.
\end{aligned}$$

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

Case 2: 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 $x_2(t;\delta,\gamma) = \left(\frac{-a(\delta;t)+\gamma}{1+|\gamma|}\right)x_1(t-\delta) + r_1(t;\delta,\gamma) + r_2(t;\delta,\gamma)$.

This case remains almost similar to the previous case and 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 jth source via $C_j(\tau,\omega)$, the separability of this mixture can be analyzed using $J(\tau,\omega)$ and (50) as

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

It can be deduced from above that the cost function yields a zero value for $k = j$, and nonzero value for $k \neq j$. Despite the mixing attenuation for both sources are identical, the function $J(\tau,\omega)$ is still able to distinguish the k arguments by using only the difference of residues. Therefore, the mixture of Case 2 is separable.

Case 3: $a_1(t;\delta,\gamma) \neq a_2(t;\delta,\gamma)$ and $r_1(t;\delta,\gamma) \neq r_2(t;\delta,\gamma)$ (or $r_1(t;\delta,\gamma) = r_2(t;\delta,\gamma)$) then $x_2(t;\delta,\gamma) = \left(\frac{-a_{s_1}(\delta;t)+\gamma}{1+|\gamma|}\right)s_1(t-\delta) + \left(\frac{-a_{s_2}(\delta;t)+\gamma}{1+|\gamma|}\right)s_2(t-\delta) + r_1(t;\delta,\gamma) + r_2(t;\delta,\gamma)$

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 function $J(\tau,\omega)$ given by (50) can be expressed as

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

This cost function yields a nonzero value only for $k \neq j$. In this case, the function $J(\tau,\omega)$ can separate the k arguments due to

the difference of 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 function $J(\tau,\omega)$ becomes

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

This cost function yields a nonzero value only for $k \neq j$; thus the function $J(\tau,\omega)$ is able to distinguish the k arguments. In summary, by considering $a_j(t;\delta,\gamma)$ and $r_j(t;\delta,\gamma)$ with respect to above three cases, only Case 2 and Case 3 are separable.

V. Results and Analysis

A noisy mixture is generated by adding two sources and an uncorrelated nonstationary noise with various input SNRs. 20 speech, 20 music signals and noise signals are selected from TIMIT, RWC, and Noisex databases, respectively. Additionally, we have conducted experiments to determine the optimal ξ_f and the choice of ζ_M. All experiments are conducted under the same conditions as follows: The sources 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 $p(H_0) = p(H_1) = 0.5$. The separation performance is evaluated by measuring the distortion between the original source and the estimated one according to the signal-to-distortion (SDR) ratio [35] defined as $\text{SDR} = 10log_{10}(\|s_{target}\|^2/\|e_{interf} + e_{noise} + e_{artif}\|^2)$ where e_{interf}, e_{noise}, and e_{artif} represent the interference from other sources, 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 3 GHz and 3GB RAM.

A. Determination of Optimal ξ_f for Mixture Enhancement

The optimal ξ_f is determined by minimizing the proposed integrated probability of error in (21) and (22) in Section III-A-1. The term ξ varies from 0 dB to 30 dB by 5 dB increment. The candidate ξ_f is converted from linear scale to dB (i.e., $10\log_{10}\xi_f = \xi_f^{dB} dB$) with various ξ_f^{dB} from 0 dB to 50 dB by 5 dB increment.

Fig. 2 on the left-hand side shows the plot of $p_e(\xi,\xi_f)$ for various ξ values. As a result of individual ξ, the minimum $p_e(\xi,\xi_f)$ is obtained at $\xi_f = \hat{\xi}_f = \xi$. Therefore, the optimal ξ_f is then set by ξ. However in realistic scenario, the term ξ is unknown. Thus, the optimal ξ_f in (22) is determined by approximating the above integral in (22) by discretely evaluating the term at various ξ values and taking the average. The result is shown on the right-hand side of Fig. 2. It can be seen that the range of $\hat{\xi}_f$ that yields the minimum error is between 10 dB and 15 dB. Based on this result, the optimal ξ_f can be set at $10\log_{10}\xi_f = 12.5$ dB for all experiments.

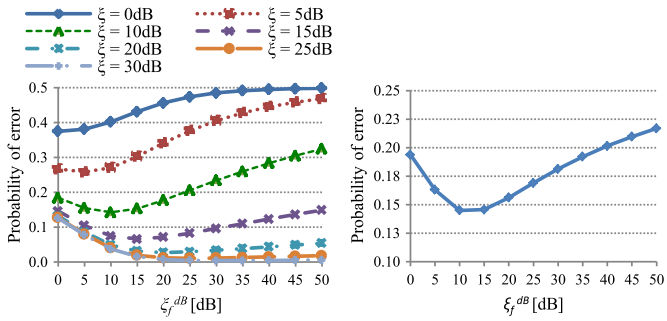


Fig. 2. Probability of error $p_e(\xi,\xi_f)$ of individual ξ value (left) and integrated probability of error for various ξ_f (right).

B. Mixture Enhancement Performance

To verify the proposed mixture enhancement method, a test has been conducted and compared the mixture enhancement method with the original MMSE and the recent modified MMSE [36] by using segmental SNR (SegSNR, in dB) and the perceptual evaluation of speech quality (PESQ) measures [38]. The experiments have been assessed on three types of mixtures, i.e., music + music, speech + music, and speech + speech.

For the standard MMSE, the smoothing parameter ζ_ξ was set at 0.98 according to [27] which shows a strong correlation of $\hat{\xi}(\tau,\omega)$ and corresponding previous enhanced spectral amplitudes. As such, the term $\hat{\xi}(\tau,\omega)$ will be smoothness across time where this property suits for stationary signals. Thus, the current frame estimation $\hat{\xi}(\tau,\omega)$ inclines to be smaller than its previous estimation $\hat{\xi}(\tau-1,\omega)$. Consequently, the smooth $\hat{\xi}(\tau,\omega)$ will be underestimated. This leads to over-suppression: not only noise components but also the source signals; and the sensitive spectral amplitude estimator $\tilde{A}(\tau,\omega)$. The modified MMSE gives better noise suppression and the quality of reconstructed signals than the standard MMSE method where $\zeta_\xi = 0.5$ for low input SNR and $\zeta_\xi = 0.8$ for high input SNR as shown in Fig. 3. However, the modified MMSE demands higher computational time consuming for the training step but still removes more source components compared with the proposed mixture enhancement method. In Fig. 3, the modified MMSE gives lower perceptual intelligibility and quality of the estimated signals than the proposed mixture enhancement method even though the modified MMSE yields better SegSNR. Therefore, our proposed mixture enhancement method retains the perceptual quality of the sources and maintain a comparably high SegSNR while being able to reduce noise. The proposed mixture enhancement method yields the best PESQ performance where the average PESQ improvement are 27% and 19% over the standard and modified MMSE methods, respectively. In the interval of [0, 20] dB input SNR, the proposed mixture-enhancement method is able to significantly remove noise from the noisy mixture and also retain intelligible perception of the noise-reduced mixture. As evidenced in Fig. 3, the proposed enhancement method gains the average improvement over the noisy mixture at 3.0 dB (76%) for SegSNR and 0.4 (12%) for PESQ.

The subjective testing of signal quality and intelligibility has been conducted based on ITU-T standard (P.835). The signal distortion (SIG) [43] has been used as the opinion test of intelligibility. A five-category rating scale is used for each aspect of

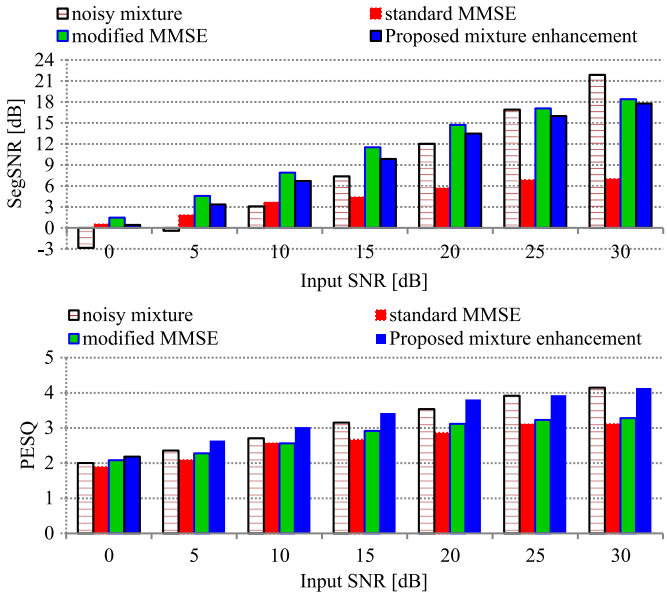


Fig. 3. SegSNR (top) and PESQ (bottom) on mixtures of two sources and additive noises at different input SNRs.

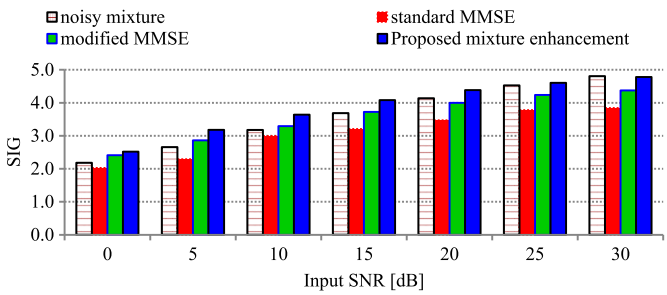


Fig. 4. Comparison of average SIG testing for the noisy mixture, standard MMSE, modified MMSE, and Proposed mixture enhancement.

the evaluation. A five-category rating scale is used for each aspect of the evaluation. For SIG, the corresponding scales are: 1) Very unnatural, very degraded, 2) Fairly unnatural, fairly degraded, 3) Somewhat natural, somewhat degraded, 4) Fairly natural, little degradation, 5) Very natural, no degradation. The SIG results are shown in Fig. 4.

The proposed mixture enhancement method renders the best quality and intelligibility of the enhanced mixture among the three MMSE methods for across the range of input SNR. A visual test has also been conducted by using mixed real-audio sources (speech + music) and an uncorrelated additive noise. A clean mixture of speech and musical sources is shown in Fig. 5(a). A noisy mixture consists of the two audio sources and a white Gaussian noise with 5 dB SNR. The enhanced mixture is obtained by applying the proposed enhancement method on the noisy mixture. Visually, an enhanced mixture in Fig. 5(c) has efficiently extracted the sources spectrum compared with the noisy mixture in Fig. 5(b).

C. Choice of ζ_M for estimating $\tilde{\tilde{a}}_j(\tau)$

The adaptive mixing attenuation estimator in (34), 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 of $\hat{\tilde{a}}_j$ through ζ_M. To determine ζ_M,

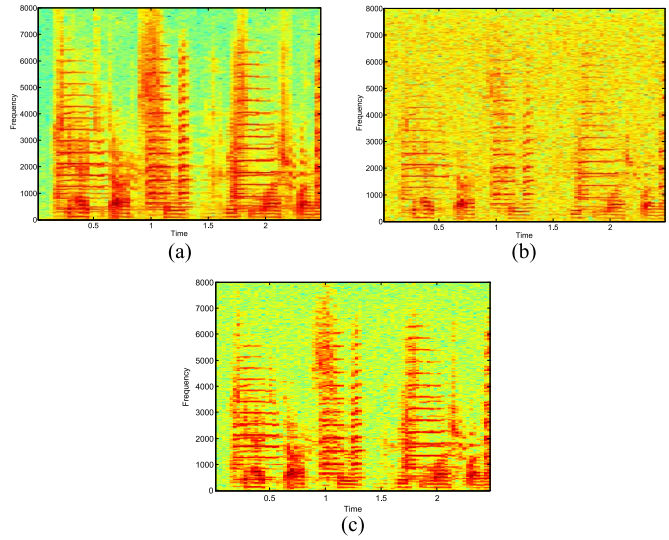


Fig. 5. Spectrograms of original clean mixture, clean mixture and additive white noise, noisy mixture enhanced using proposed iMMSE-STSA estimator. (a) clean mixture; (b) noisy mixture; (c) enhanced mixture.

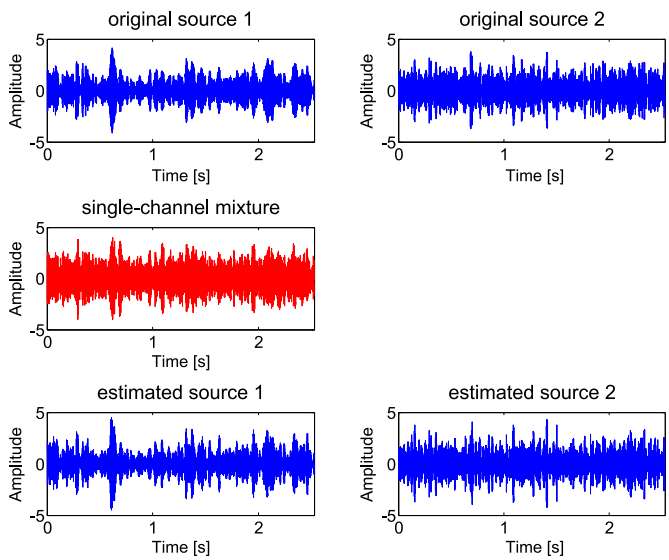


Fig. 6. Two original sources, noise-free mixture and two estimated sources with $\zeta_M = 0.95$.

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 sources. The nonstationary AR source is synthesized by using the model (3) with 256 s length which divided into five sections, i.e., $T_1 = [0, 0.51\text{ s}]$, $T_2 = (0.51\text{ s}, 1.03\text{ s}]$, $T_3 = (1.03\text{ s}, 1.54\text{ s}]$, $T_4 = (1.54\text{ s}, 2.05\text{ s}]$, and $T_5(2.05\text{ s}, 2.56\text{ s}]$, respectively. The term a_{s_j} and $e_j(t)$ of $s_j(t)$ have been changed section by section. The samples of synthetic source signals are shown in Fig. 6 in the top row.

Firstly, the term ζ_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 3 dB per source. The term ζ_M is then further tested on [0.86,0.99] with 0.01 increments and its results are

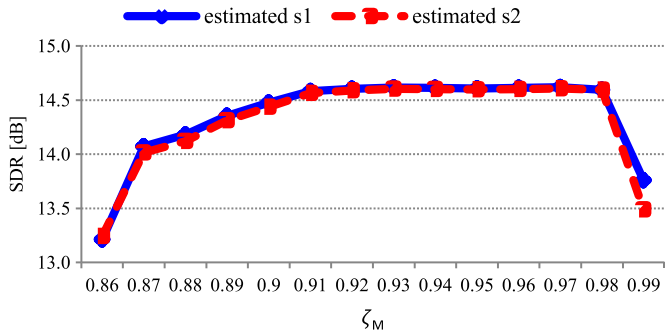


Fig. 7. Average SDR on the noise-free mixture of two synthetic AR sources with various ζ_M.

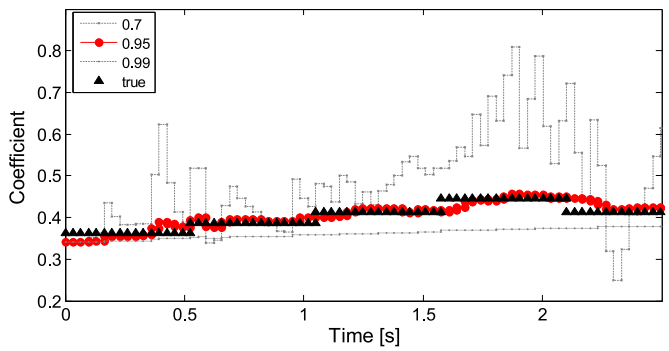


Fig. 8. Mixing coefficients of $\bar{a}_1(\tau)$ (true) and $\tilde{\bar{a}}(\tau)$ for $\zeta_M = 0.7, 0.95, 0.99$.

illustrated in Fig. 7. The highest average SDR is within the interval of ζ_M from 0.91 to 0.98. Hence the optimal choice of ζ_M will be within [0.91,0.98].

We have plotted an example of $\tilde{\bar{a}}_1(\tau)$ against $\bar{1}(\tau)$ with different ζ_M values in Fig. 8. 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 τth frame depends 99% on its previous value. When $\zeta_M = 0.95$, $\tilde{\bar{a}}(\tau)$ tracks very closely with the true $\bar{a}_j(\tau)$. Hence, ζ_M has a crucial role in tracking the behavior of $\bar{a}_j(\tau)$. Although $\tilde{\bar{a}}(\tau)$ is an estimate of $\bar{a}_j(\tau)$, the separating performance of $\tilde{\bar{a}}(\tau)$ yields the same SDR as $\bar{a}_j(\tau)$ at 14.7 dB and 14.9 dB for $\hat{\bar{s}}_1(t)$ and $\hat{\bar{s}}_2(t)$, respectively. This is because the condition $G_{k=j}(\tau,\omega) < G_{k\neq j}(\tau,\omega)$ has been satisfied when $|\hat{C}_j(\tau)| < \left|\left(\tilde{\bar{a}}(\tau) - a_j(\tau) - \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right)\right| + \left|\frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right| + \frac{2}{1+|\gamma|}$ according to (45). We have computed the $|\hat{C}_j(\tau)|$ condition for $j = 1$ and 2 as shown in Fig. 9. For $j = 1$, $|\hat{C}_j(\tau)| < \left|\left(\tilde{\bar{a}}_2(\tau) - a_1(\tau) - \frac{a_{S_1}(\delta;\tau)}{1+|\gamma|}\right)\right| + \left|\frac{a_{S_1}(\delta;\tau)}{1+|\gamma|}\right| + \frac{2}{1+|\gamma|}$, thus the $|\hat{C}_1(\tau)|$ condition is satisfied. For $j = 2$, the $|\hat{C}_2|$ condition is also true. Therefore, the cost function has correctly assigned all (τ,ω) units to their respective original sources. This is clearly evident by the same SDR results between the $\tilde{\bar{a}}_j(\tau)$ and the $\bar{a}_j(\tau)$. Therefore, we selected ζ_M around 0.95 for all experiments.

D. Separation Performance

The separation performance of the proposed method has been assessed by using 150 mixtures. The noises have been randomly selected from the NOISEX database which are: pink.wav, destroyerops.wav and factory2.wav. These noises

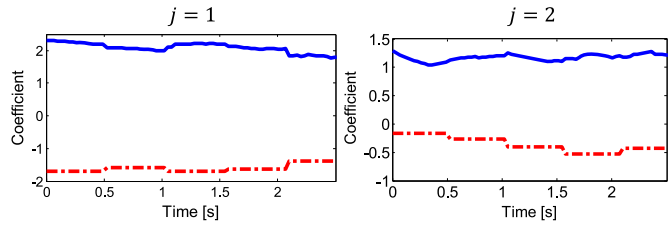


Fig. 9. $|\hat{C}_j(\tau)|$ condition of $j = 1$ on the left plot and $j = 2$ on the plot where the dot-dash line refers to $|\hat{C}_j(\tau)|$ and the continuous line refers to $\left|\left(\tilde{\bar{a}}_k(\tau) - a_j(\tau) - \frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right)\right| + \left|\frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right| + \frac{2}{1+|\gamma|}, j \neq k$.

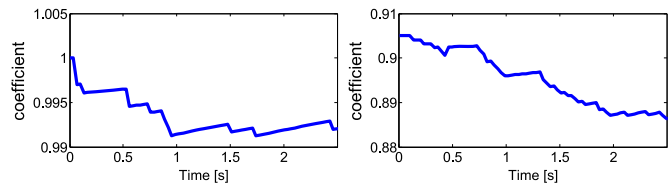


Fig. 10. Estimated coefficients of $\tilde{\bar{a}}_1(\tau)$ (left) and $\tilde{\bar{a}}_2(\tau)$ (right).

represent stationary, nonstationary and highly nonstationary noises, respectively. The proposed approach will be compared with the single-channel nonnegative matrix 2-dimensional factorization (SNMF2D) and the single-channel independent component analysis (SCICA) [4]. The SNMF2D parameters are set as follows [4]: 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. Cost function of SNMF2D is based on the Kullback-Leibler divergence. As for the SCICA, the number of block is 10 with unity time delay.

In Fig. 10, $\tilde{\bar{a}}_1(\tau)$ and $\tilde{\bar{a}}_2(\tau)$ change from frame to frame (this is natural as they correspond to speech and music signals, respectively). Examples of two audio sources with equal power, the additive noise, and the noisy mixture at 0 dB SNR are shown in Fig. 11 at the top and the second row. Visually in Fig. 10, the estimated sources (bottom) have been clearly separated when compared with the original sources (top). On the other hand, the estimated sources from SCICA and SNMF2D have not been well separated as shown in Figs. 12 and 13, respectively. We have also illustrated the average SDR results using the proposed method for three mixing types with various inputs SNR in Fig. 14. As expected, the mixture of music + music yields 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 have similar frequency components to speech components in which the spectrums of speech signal will be submerged by the noise signal.

Fig. 15 illustrates the separation performance of SCICA, SNMF2D, and the proposed method based on three different noises with various input SNRs. An error bar denotes a standard deviation of each method across input SNRs. From the above it can be seen that the proposed method yields superior separation performance with the average SDR at 6.14, 6.47, and 6.24 dB per source for stationary, nonstationary, and highly nonstationary noises, respectively. The average improvement

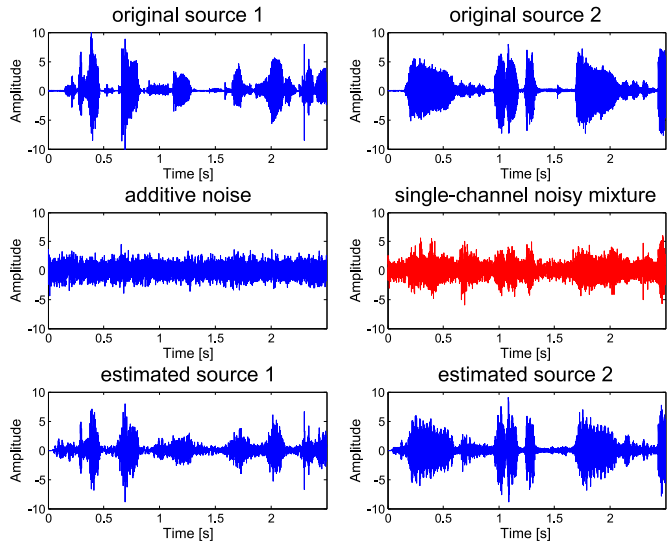


Fig. 11. Two original sources, observed noisy mixture of 0 dB SNR, and two estimated sources using the proposed method.

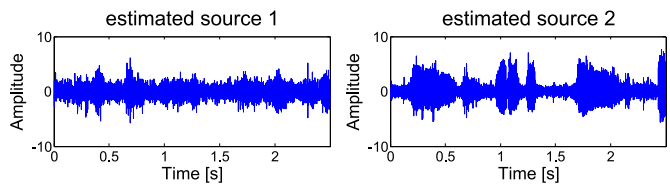


Fig. 12. Two estimated sources using SCICA method.

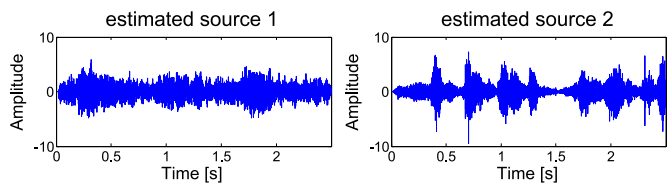


Fig. 13. Two estimated sources using SNMF2D method.

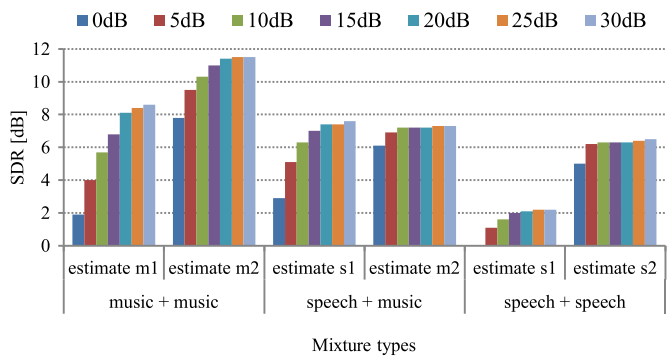


Fig. 14. Average SDR performance of three mixing types with various input SNR using the proposed method.

SDR of the proposed method over the SCICA and SNMF2D methods are 3.2 and 3.1 dB per source, respectively. 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 emphasizes the source components through the mixture enhancement step. For the SCICA

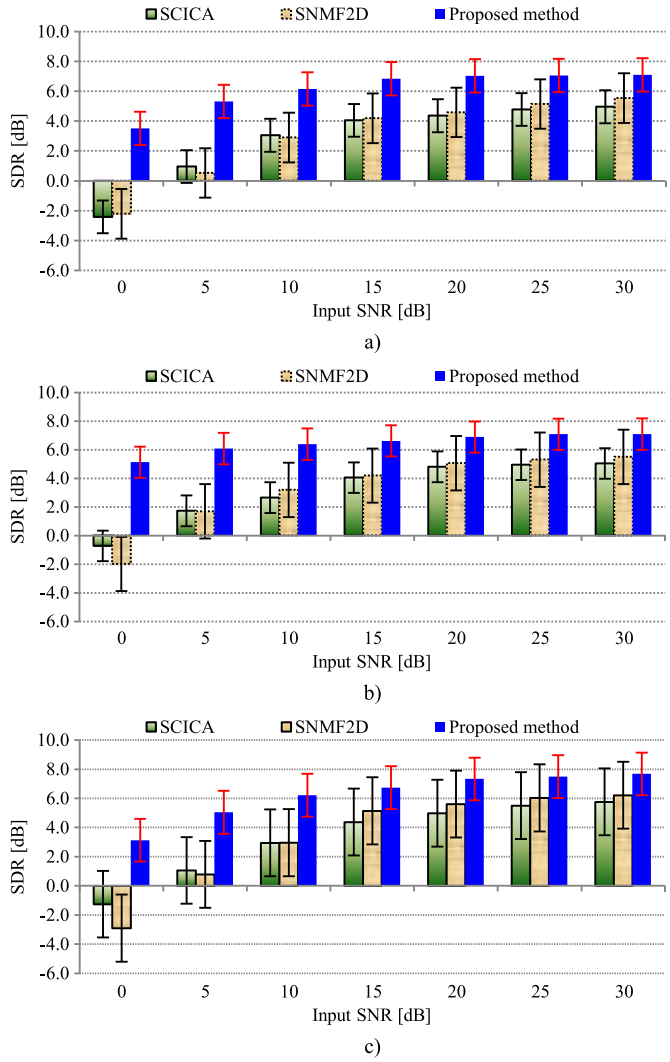


Fig. 15. Comparison of average SDR performance among SCICA, SNMF2D and the proposed method. (a) Stationary; (b) nonstationary noise; (c) highly nonstationary noise.

and SNMF2D methods, their separation performances depend critically on source information, given by the highly noisy mixture, thus these two methods are hampered by interference of noise.

Fig. 16 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 15 dB, the proposed method performs far superior than the SNMF2D and SCICA.

VI. Conclusion

In this paper, a novel noisy single channel source 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 sources in the noisy mixing model

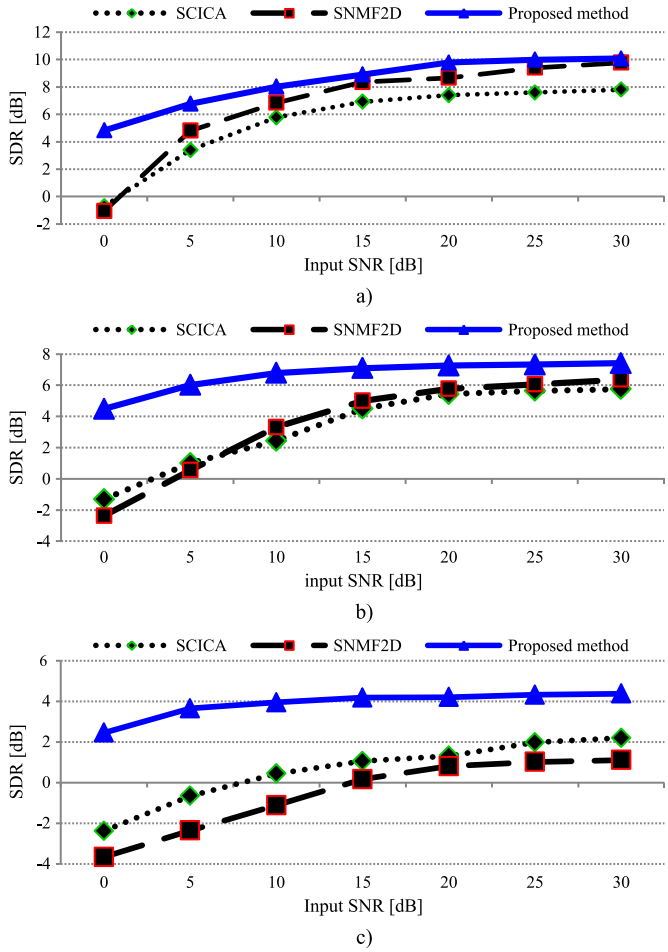


Fig. 16. Comparison of average SDR performance of the three mixing types with various input SNR between SNMF2D, SCICA, and the proposed method. (a) music + music; (b) speech + music; (c) speech + speech.

and then separates the enhanced mixture. Furthermore, the conditions required for unique mask construction from the maximum likelihood method have also been identified. The proposed method has demonstrated a high level separation performance for sources in nonstationary noisy environment. The proposed method gains at least three advantages: Firstly, the proposed approach is able to adapt the parameter estimated frame-by-frame and separates the mixture given by small blocks. Secondly, it does not require *a priori* knowledge of the sources. Finally, neither iterative optimization nor parameter initialization is required. Hence, these render the robustness to the proposed method for implementation in practical scenarios.

REFERENCES

[1] P. Tichavský and A. Yeredor, "Fast approximate joint diagonalization incorporating weight matrices," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 878–891, Mar. 2009.

[2] C. Damon, A. Liutkus, A. Gramfort, and S. Essid, "Non-negative Tensor factorization for single-channel EEG artifact rejection," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, pp. 1–6.

[3] S. Sanei, T. K. M. Lee, and V. Abolghasemi, "A new adaptive line enhancer based on singular spectrum analysis," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 2, pp. 428–432, 2012.

[4] 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. Biomed. Eng.*, vol. 57, no. 9, pp. 2188–2196, Sep. 2010.

[5] S. Kouchaki and S. Sanei, "Supervised single channel source separation of EEG signals," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Sep. 2013, pp. 1–5.

[6] B. Gao, L. Bai, W. L. Woo, G. Y. Tian, and Y. Cheng, "Automatic defect identification of eddy current pulsed thermography using single channel blind source separation," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 4, pp. 913–922, 2014.

[7] B. Gao, L. Bai, W. L. Woo, and G. Y. Tian, "Thermography pattern analysis and separation," *Appl. Phys. Lett.*, vol. 104, no. 25, p. 251902, 2014.

[8] B. Gao, H. Zhang, W. L. Woo, G. Y. Tian, L. Bai, and A. Yin, "Smooth nonnegative matrix factorization for defect detection using microwave nondestructive testing and evaluation," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 4, pp. 923–934, 2014.

[9] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1486–1494, 2009.

[10] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Amer.*, vol. 134, no. 4, pp. 3029–3038, 2013.

[11] Q. Wang, W. L. Woo, and S. S. Dlay, "Informed single channel speech separation using HMM-GMM user-generated exemplar source," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2087–2100, 2014.

[12] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.

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

[14] 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.

[15] B. Gao, W. L. Woo, and S. S. Dlay, "Variational regularized two-dimensional nonnegative matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 5, pp. 703–716, 2012.

[16] B. Gao, W. L. Woo, and S. S. Dlay, "Adaptive sparsity nonnegative matrix factorization for single channel source separation," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 989–1001, 2011.

[17] B. Gao, W. L. Woo, and S. S. Dlay, "Unsupervised single channel separation of non-stationary signals using Gammatone Filterbank and Itakura-Saito nonnegative matrix two-dimensional factorizations," *IEEE Trans. Circuits Syst. I*, vol. 60, no. 3, pp. 662–675, 2013.

[18] B. Gao, W. L. Woo, and L. C. Khor, "Cochleagram-based audio pattern separation using two-dimensional non-negative matrix factorization with automatic Sparsity adaptation," *J. Acoust. Soc. Amer.*, vol. 135, pp. 1171–1185, 2014.

[19] B. Gao, W. L. Woo, and B. W.-K. Ling, "Machine learning source separation using maximum a posteriori nonnegative matrix factorization," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1169–1179, 2014.

[20] P. Parathai, W. L. Woo, S. S. Dlay, and B. Gao, "Single-channel blind separation using L1-sparse complex nonnegative matrix factorization for acoustic signals," *J. Acoust. Soc. Amer.*, vol. 137, p. EL124, 2015.

[21] 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.

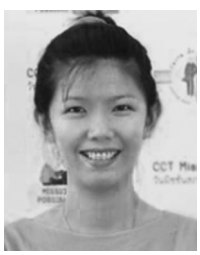
[22] T. May, S. V. D. Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 2016–2030, Sep. 2012.

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

[24] N. Tengtrairat, B. Gao, W. L. Woo, and S. S. Dlay, "Single-channel blind separation using pseudo-stereo mixture and complex 2-D histogram," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 11, pp. 1722–1735, 2013.

[25] N. Tengtrairat and W. L. Woo, "Single-channel separation using underdetermined blind method and least absolute deviation," *Neurocomput.*, vol. 147, pp. 412–425, 2015.

[26] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Process. Lett.*, vol. 17, no. 6, pp. 599–602, 2010.
[27] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.
[28] E. Plourde and B. Champagne, "Multidimensional STSA estimators for speech enhancement with correlated spectral components," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3013–3024, Jul. 2011.
[29] N. Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2012, pp. 4561–4564.
[30] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, 1992.
[31] M. Athineos and D. P. W. Ellis, "Autoregressive modeling of temporal envelopes," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5237–5245, 2007.
[32] 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.
[33] 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.
[34] S. Mousazadeh and I. Cohen, "AR-GARCH in presence of noise: Parameter estimation and its application to voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 916–926, May 2011.
[35] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[36] J. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria," *Speech Commun.*, vol. 49, pp. 530–541, 2007.
[37] *Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for end-to-end Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, ITU-T P.862, International Telecommunication Union, Geneva, Switzerland, 2001.
[38] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Speech, Audio Process.*, vol. 16, no. 1, pp. 229–238, 2008.



N. Tengtrairat received the B.Eng. degree in computer engineering from the Chiang Mai University, Chiang Mai, Thailand, M.Sc. degree in management information system from the Chulalongkorn University, and Ph.D. from Newcastle University, U.K. She has been a lecturer with the Department of Software Engineering at Payap University, Thailand. Her research interests include statistical single-channel blind source separation, speech and audio signal processing, speech enhancement, noise cancelling, and machine learning.

W. L. Woo (M'09–SM'11) received the B.Eng. degree (1st Class Hons.) in Electrical and Electronics Engineering and the Ph.D. degree from the Newcastle University, U.K. He was awarded the IEE Prize and the British Scholarship in 1998 to continue his research work. He is currently Director of Operations for the international branch of the University. His major research is in the mathematical theory and algorithms for nonlinear signal and image processing. He has published over 250 papers on these topics on various journals and international conference proceedings. Dr. Woo is a member of the Institution Engineering Technology (IET).

S. S. Dlay received his B.Sc. (Hons.) degree in Electrical and Electronic Engineering and his Ph.D. in VLSI Design from the Newcastle University. In 1986 he rejoined the Newcastle University as a Lecturer in the School of Electrical, Electronic and Computer Engineering and was later appointed to a Personal Chair in Signal Processing Analysis. He has published over 250 research papers ranging from biometrics and security, biomedical signal processing, and implementation of signal processing architectures. Professor Dlay is a College Member of the EPSRC.

Bin Gao (M'12–SM'14) received the B.S. degree in communications and signal processing from Southwest Jiao Tong University, Chengdu, China, in 2005, the M.Sc. (Hons.) degree in communications and signal processing, and the Ph.D. degree from Newcastle University, Newcastle, U.K., in 2011. He is currently an Associate Professor with the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu.