Single-Channel Blind Separation Using Pseudo-Stereo Mixture and Complex 2-D Histogram

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

***Abstract*— A novel single-channel blind source separation (SCBSS) algorithm is presented. The proposed algorithm yields at least three benefits of the SCBSS solution: 1) resemblance of a stereo signal concept given by one microphone; 2) independent of initialization and *a priori* knowledge of the sources; and 3) it does not require iterative optimization. The separation process consists of two steps: 1) estimation of source characteristics, where the source signals are modeled by the autoregressive process and 2) construction of masks using only the single-channel mixture. A new pseudo-stereo mixture is formulated by weighting and time-shifting the original single-channel mixture. This creates an artificial mixing system whose parameters will be estimated through our proposed weighted complex 2-D histogram. In this paper, we derive the separability of the proposed mixture model. Conditions required for unique mask construction based on maximum likelihood are also identified. Finally, experimental testing on both synthetic and real-audio sources is conducted to verify that the proposed algorithm yields superior performance and is computationally very fast compared with existing methods.**

***Index Terms*— Blind source separation, single-channel, underdetermined mixture, unsupervised signal processing.**

I. Introduction

With only the sensor signals, blind source separation (BSS) is the process of recovering underlying source signals from an unknown mixing [1]–[3]. BSS has interested many researchers during the last decade because of its potential to solve problems in an ubiquitous range of disciplines. In the early BSS era, independent component analysis (ICA) was first proposed as a solution [4]. The ICA approach aims to seek the unknown mixing matrices for extracting a number of sources from a number of observed mixtures based on the critical assumption that the source signals are non-Gaussian [5] and mutually independent. BSS using ICA approaches is straightforward and is used in many applications with great success. For more information, see [6] and [7]. In practice, it may not be able to provide a sensor for individual source because of limited spaces, high cost of sensors, violation of assumptions, and so on [8]. Thus, the number of sensors is mostly less than the number of source signals. In addition, there is a case where only one sensor is available and this corresponds to the extreme case of the underdetermined BSS problem. Under this circumstance, most conventional BSS methods fail to recover the signal source from the single-channel observation. This leads to a research avenue of single-channel BSS (SCBSS) problem. SCBSS represents the separation of mixed signal from a single sensor. Mathematically, it can be treated as one mixture of unknown source signals as follows:

$$x(t) = s_1(t) + s_2(t) + \cdots + s_N(t) \qquad (1)$$

where $t = 1, 2, \ldots, T$ is the time index 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. Many SCBSS approaches are proposed to solve the problem. In general, it can be categorized into two groups, i.e., model-based and data-driven methodologies.

The term model-based separation approach requires prior knowledge from the training data sets to estimate the unknown sources. Model-based SCBSS methods are dominantly illuminated by computational auditory scene analysis (CASA) [9]–[11], and hidden Markov models (HMMs) methods [12]–[14]. The goal of CASA is to replicate the process of human auditory system using signal processing approaches and grouping them into auditory streams using psychoacoustical cues. It exploits an appropriate transform hence the observed mixture is segmented into time-frequency (TF) cells that are then used to characterize note objects by harmonicity, common onset, correlated modulation and duration of sinusoidal partials, and finally to build note streams based on pitch proximity. Nevertheless, CASA methods cannot efficiently segregate instruments playing in the different pitch range into different streams. They also cannot replicate the entire process performed in the auditory system as the process beyond the auditory nerve is not well studied. In addition, it is difficult to group the sources if one of them is assumed to be fully voiced. In HMM-based methods, one recent technique [14] proposes an eigenvoice speech models that define the space of speaker variation and an iterative algorithm to infer the parameters for each source. The separation is based on Viterbi algorithm to find a path through a factorial HMM. This method shows good results. However, the drawbacks of the system are the computational time consumption not only for the training, but also for the separation process. In addition, it needs access to the source signals for the training purposes that renders the method nonblind. A related technique to model-based SCBSS is the underdetermined-ICA method [15]. This technique models the sources as sparse combination of a set of time-domain basis functions that are initially derived using the ICA methods. The sources are subsequently estimated by

Manuscript received July 4, 2012; revised April 3, 2013; accepted April 3, 2013. Date of publication August 1, 2013; date of current version October 15, 2013.

The authors are with the School of Electrical and Electronic Engineering, Newcastle University, Newcastle NE1 7RU, U.K. (e-mail: naruephorn.tengtrairat@ncl.ac.uk; bin.gao@ncl.ac.uk; w.l.woo@ncl.ac.uk; s.s.dlay@ncl.ac.uk).

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/TNNLS.2013.2258680

maximizing the loglikelihood with the ICA-derived basis functions. This method renders optimal separation when the ICA basis functions corresponding to each source have minimal time-domain overlap. When the basis functions have significant overlap with each other, e.g., mixture of two speech sources, this method performs very poorly.

For data-driven SCBSS, these methods perform source separation without any recourse to the training information. A popular method in this category is the sparse nonnegative matrix factorization (SNMF). The SNMF method [16] determines 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 that has the advantage of simplicity. It requires no assumption on sources 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. In addition, 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 [17]. To integrate these features into the SNMF, a 2-D model leading to the sparse nonnegative matrix 2-D factorization (SNMF2-D) was thus developed in [18]. The SNMF2-D uses a double convolution to model both spreading of spectral basis and variation of temporal structure inherent in the sources. Some success was already reported in [19] and [20] to show the validity of SNMF2-D in separating single-channel mixture.

In binaural BSS method, the degenerate unmixing estimation technique (DUET) [21] and its variants [22], [23] are proposed as a separating method using binary TF masks. A major advantage of DUET is that the estimates from two channels are combined inherently as part of the clustering process. DUET algorithm is demonstrated to recover the underlying sparse sources given two anechoic mixtures in the TF domain. However, the DUET algorithm is practically handicapped to separate signals when only one recording channel is available. In addition, 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.

In this paper, a new framework for solving the above problem is presented by reformulating the binaural BSS problem using monaural method. This paper contributes a novel method whose strength is summarized as follows: 1) 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 SNMF2-D and underdetermined-ICA SCBSS that require many iterative optimization of the solution; 2) It is independent of initialization condition, i.e., no need for random initial inputs or any predetermined structure on the sensors. This renders robustness to the proposed method; and 3) it has low-computational complexity and does not exploit high-order statistic. Hence this yields the benefit of implementation. We term the proposed method as single observation likelihood estimation (SOLO) algorithm.

This paper is organized as follows. Section II introduces a pseudo-stereo mixture model, the assumptions of SOLO method, and the separability of the proposed mixture model. Parameter estimation and construction of masks are described in Section III. Section IV describes a method for selecting the parameters of the pseudo-stereo mixture. Next, experimental results with a series of performance comparison with other unsupervised SCBSS methods are described in Section V. Finally, conclusion is summarized in Section VI.

II. Single-Channel Mixing Model

A. Pseudo-Stereo Mixture Model

In this paper, for simplicity we consider the case of a mixture of two sources in time domain as follows:

$$x(t) = s_1(t) + s_2(t) \tag{2}$$

where $x_1(t)$ is the single-channel mixture, and $s_1(t)$ and $s_2(t)$ are the original source signals that are assumed to be modeled by the autoregressive (AR) process [24] as follows:

$$s_j(t) = -\Sigma_{m=1}^{D_j} a_{s_j}(m;t) s_j(t-m) + e_j(t) \tag{3}$$

where $a_{s_j}(m;t)$ is 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 random signal with zero mean and variance $\sigma_{e_j}^2$. This model is particularly interesting in source separation: 1) many audio signals satisfy this process and 2) it enables us to formulate a virtual mixture by weighting and time-shifting the single-channel mixture $x_1(t)$ as follows:

$$x_2(t) = \frac{x_1(t) + \gamma x_1(t-\delta)}{1+|\gamma|} \tag{4}$$

where $\gamma \in \Re$ is the weight parameter, and δ is the time-delay. The mixture in (2) and (4) is termed as pseudo-stereo because it has an artificial resemblance of a stereo signal except that it is given by one location that results in the same time-delay but different attenuation of the source signals. To show this, we can express (4) in terms of the source signals, AR coefficient, and time-delay as follows:

$$\begin{aligned}
x_2(t) &= \frac{x_1(t) + \gamma x_1(t-\delta)}{1+|\gamma|} \\
&= \frac{s_1(t) + s_2(t) + \gamma\,[s_1(t-\delta) + s_2(t-\delta)]}{1+|\gamma|} \\
&= \frac{-\Sigma_{m=1}^{D_1} a_{s_1}(m) s_1(t-m) + e_1(t)}{1+|\gamma|} + \frac{\gamma s_1(t-\delta)}{1+|\gamma|} \\
&\quad \frac{-\Sigma_{m=1}^{D_2} a_{s_2}(m) s_2(t-m) + e_2(t)}{1+|\gamma|} + \frac{\gamma s_2(t-\delta)}{1+|\gamma|} \\
&= \frac{(-a_{s_1}(\delta) + \gamma)}{1+|\gamma|} s_1(t-\delta) + \frac{(-a_{s_2}(\delta) + \gamma)}{1+|\gamma|} s_2(t-\delta) \\
&\quad + \frac{e_1(t) - \sum_{\substack{m=1 \\ m \neq \delta}}^{D_1} a_{s_1}(m) s_1(t-m)}{1+|\gamma|} \\
&\quad + \frac{e_2(t) - \sum_{\substack{m=1 \\ m \neq \delta}}^{D_1} a_{s_2}(m) s_2(t-m)}{1+|\gamma|}.
\end{aligned} \tag{5}$$

Define

$$a_j(t;\delta,\gamma) = \frac{-a_{s_j}(\delta;t)+\gamma}{1+|\gamma|} \tag{6}$$

$$r_j(t;\delta,\gamma) = \frac{e_j(t) - \sum_{\substack{m=1\\ m\neq\delta}}^{D_j} a_{s_j}(m;t)s_j(t-m)}{1+|\gamma|} \tag{7}$$

where $a_j(t;\delta,\gamma)$ and $r_j(t;\delta,\gamma)$ are the mixing attenuation and residue of the jth source, respectively. The parameterization of $a_j(t;\delta,\gamma)$ and $r_j(t;\delta,\gamma)$ depends on δ and γ 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., $a_j(t), \delta, r_j(t)$ that are used to construct the complex 2-D histogram for estimating the sources. Using (6) and (7), the overall proposed mixing model of the SOLO can now be formulated in terms of the sources as follows:

$$\begin{aligned} x_1(t) &= s_1(t) + s_2(t) \\ x_2(t) &= a_1(t;\delta,\gamma)s_1(t-\delta) + a_2(t;\delta,\gamma)s_2(t-\delta) \\ &\quad + r_1(t;\delta,\gamma) + r_2(t;\delta,\gamma). \end{aligned} \tag{8}$$

B. Method Assumptions

The proposed SOLO method focuses on separating sources from one mixture using a binary TF mask. To achieve this, the following assumptions will be used.

Assumption 1: The sources are modeled as quasi-stationary. This refers to the condition where the AR parameters in (3) 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 $\forall t \in T_l = \{t_l, \ldots, t_{l+1}-1\}$ such that

$$s_j(t) = -\Sigma_{m=1}^{D_j} a_{s_j}(m;T_l)s_j(t-m) + e_j(t)\ \forall t \in T_l. \tag{9}$$

Stationary AR sources are special case of the above where the AR parameters do not vary with time [25] and this is equivalent to setting $L = 1$ in (9).

Assumption 2: The sources satisfy the windowed-disjoint orthogonality [26] condition as follows:

$$S_i(\tau,\omega)S_j(\tau,\omega) \approx 0\ \forall i \neq j\ \forall \tau,\ \omega \tag{10}$$

where $S_i(\tau,\omega)$ is the short-time Fourier transform (STFT) of $S_j(t)$ defined as follows:

$$\begin{aligned} S_j(\tau,\omega) &= F^W[s_j(t)](\tau,\omega) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} W(t-\tau)s_j(t)e^{i\omega t}\mathrm{dt} \end{aligned} \tag{11}$$

and $W(t)$ is the window function. The STFT is performed on the signal frame-by-frame and thus, τ is the window shift. Let us denote the shift size as Δ_τ. To ensure the quasi-stationarity of the source is maintained in the TF domain, it is required $\Delta_\tau \le B_l$ for all l. This is practically justified by choosing the appropriate length of $W(t)$. Hence, we can write $a_{s_j}(m;T_l) = a_{s_j}(m;\tau)$ provided that $\Delta_\tau \le B_l$ and $\tau \in T_l$.

Assumption 3: The sources satisfy the local stationarity of the TF 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 $F^W(\cdot)$ with an appropriate window function $W(\cdot)$. If ϕ is small compared with the length of $W(\cdot)$ then [27]. 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

$$\begin{aligned} s_j(t-\delta) &\overset{\text{STFT}}{\to} e^{-i\omega\delta}S_j(\tau-\delta,\omega) \\ &\approx e^{i\omega\delta}S_j(\tau,\omega)\ \ \forall\delta, |\delta| \le \phi. \end{aligned} \tag{12}$$

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

Assumption 4: Phase ambiguity. The factor $e^{-i\omega\delta}$ is only uniquely specified if $|\omega\delta| < \pi$, otherwise this would cause phase-wrap [28]. Selecting improper time-delay δ will lead to phase-wrap if the maximum frequency of the source is exceeded. To avoid phase ambiguity, we must satisfy the following:

$$|\omega_{\max}\delta_{\max}| < \pi \tag{13}$$

where $\omega_{\max} = 2\pi f_{\max}/f_s$, $\delta_{\max}$ is the maximum time delay, $f_{\max}$ is the maximum frequency present in the sources and f_s is the sampling frequency. Hence, $\delta_{\max}$ can be determined from (13) according to the following:

$$\delta_{\max} < \frac{f_s}{2f_{\max}}. \tag{14}$$

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, we obtain $\delta_{\max} < 2.28$ using (14). Therefore, phase ambiguity can be avoided if δ is selected to be either one or two. In addition, 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 δ in formulating the pseudo-stereo mixture.

C. TF Representation

With the above assumptions, the TF representation of the mixing model is obtained using the STFT of $x_j(t)$, $j = 1, 2$, as follows:

$$\begin{aligned} X_1(\tau,\omega) &= S_1(\tau,\omega) + S_2(\tau,\omega) \\ X_2(\tau,\omega) &= a_1(\tau)e^{-i\omega\delta}S_1(\tau-\delta,\omega) \\ &\quad + a_2(\tau)e^{-i\omega\delta}S_2(\tau-\delta,\omega) \\ &\quad - \left(\sum_{\substack{m=1\\ m\neq\delta}}^{D_1} \frac{a_{s_1}(m;\tau)}{1+|\gamma|}e^{-i\omega m}S_1(\tau-m,\omega)\right. \\ &\quad \left. + \sum_{\substack{m=1\\ m\neq\delta}}^{D_1} \frac{a_{s_2}(m;\tau)}{1+|\gamma|}e^{-i\omega m}S_2(\tau-m,\omega)\right) \end{aligned} \tag{15}$$

for $\forall\tau,\omega$. In (15), we use the fact that $e_j(t) \ll s_j(t)$, thus the TF of $r_j(t)$ in (7) simplifies to the following:

$$R_j(\tau,\omega) = -\sum_{\substack{m=1\\ m\neq\delta}}^{D_1} \frac{a_{s_j}(m;\tau)}{1+|\gamma|}e^{-i\omega m}S_1(\tau-m,\omega). \tag{16}$$

To facilitate further analysis, we also define as follows:

$$C_j(\tau,\omega) = \frac{1}{1+|\gamma|}\sum\nolimits_{\substack{m=1\\ m\neq\delta}}^{D_j} a_{s_j}(m;\tau)e^{-i\omega(m-\delta)} \quad (17)$$

which forms part of $R_j(\tau,\omega)$ without the contribution of the source $S_j(\tau,\omega)$. From (15), 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 (15) will reveal that even if $S_j(\tau,\omega)$ is unknown, the signature of each source can be extracted directly from $X_1(\tau,\omega)$ using only information of $a_j e^{-i\omega\delta}$ and $C_j(\tau,\omega)$. Thus, this constitutes the separability of the SOLO model that will be analyzed in the following section.

D. Analysis of the Proposed Mixture Model

The separability of SOLO can be examined from the pseudo-stereo mixture by considering $a_j(t;\delta,\gamma)$ and $r_j(t;\delta,\gamma)$ and evaluating in the light of the following cost function:

$$\begin{aligned} J(\tau,\omega) = \arg\min_k \Bigg|&\bar{a}_k(\tau,\omega)e^{-i\omega\delta}X_1(\tau,\omega) \\ &-\left(\frac{1+\gamma\, e^{i\omega\delta}}{1+|\gamma|}X_1(\tau,\omega)\right)\Bigg|^2 \end{aligned} \quad (18)$$

where

$$\bar{a}_k(\tau,\omega) = a_k(\tau) - C_k(\tau,\omega) \quad (19)$$

with $a_k(\tau)$ and $C_k(\tau,\omega)$ are defined in (6) and (17). The full development of the above cost function will be described in Section III-B. Technically, this function partitions the TF plane of the mixed signal into k groups of (τ,ω) units by evaluating the cost function. For each TF unit, the kth argument that gives the minimum cost will be assigned to the kth source. We may analyze (18) further by assuming the jth source dominates at a particular TF unit. Here, the first line of (15) reduces to $X_1(\tau,\omega) = S_j(\tau,\omega)$ and therefore, (18) becomes as follows:

$$\begin{aligned} J(\tau,\omega) &= \arg\min_k \Bigg|\bar{a}_k(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega) \\ &\qquad -\left(\frac{1+\gamma\, e^{-i\omega\delta}}{1+|\gamma|}\right)S_j(\tau,\omega)\Bigg|^2 \\ &= \arg\min_k \Bigg|a_k(\tau)e^{-i\omega\delta}S_j(\tau,\omega) - C_k(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega) \\ &\qquad +\sum\nolimits_{\substack{m=1\\ m\neq\delta}}^{D_1} = \frac{a_{s_j}(m;\tau)e^{-i\omega m}}{1+|\gamma|}S_j(\tau-,m,\omega) \\ &\qquad -a_j(\tau)e^{-i\omega\delta}S_j(\tau,\omega)\Bigg|^2 \end{aligned} \quad (20)$$

We consider the following three cases.

Case 1: Identical sources mixed in the single channel that can be expressed as follows: If $a_1(t;\delta,\gamma) = a_2(t;\delta,\gamma) = a(t;\delta,\gamma)$ and $r_1(t;\delta,\gamma) = r(t;\delta,\gamma)$, then $x_2(t) = ((-a(\delta;t)+\gamma)/1+|\gamma|)x_1(t-\delta) + 2r(t;\delta,\gamma)$.

The separability of this case is presented by substituting the pseudo-stereo mixture of Case 1 into the cost function. As both residues are equal, then $C_1(\tau,\omega) = C_2(\tau,\omega) = C(\tau,\omega) = (1/1+|\gamma|)\sum\nolimits_{\substack{m=1\\ m\neq\delta}}^{D} a_s(m;\tau)e^{-i\omega(m-\delta)}$. For Case 1, the cost function becomes as follows:

$$\begin{aligned} J(\tau,\omega) = \arg\min_k &|a(\tau)\,e^{-i\omega\delta}s_j(\tau,\omega) - C(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega) \\ &+\Sigma_{\substack{m=1\\ m\neq\delta}}^{D}\frac{a_s(m;\tau)e^{-i\omega m}}{1+|\gamma|}S_j(\tau-m,\omega) \\ &-a(\tau)e^{-i\omega\delta}S_j(\tau,\omega)|^2. \end{aligned} \quad (21)$$

Invoking the local stationarity of the sources $S_j(\tau-D_j,\omega) = S_j(\tau,\omega)$ for $|D_j| \leq \phi$, (21) leads to the following:

$$\begin{aligned} J(\tau,\omega) &= \arg\min_k \left|\sum\nolimits_{\substack{m=1\\ m\neq\delta}}^{D_j}\frac{\left(a_{s_j}(m;\tau)-a_{s_k}(m;\tau)\right)}{1+|\gamma|}\ e^{-i\omega m}\right|^2 \\ &\qquad \times\left|S_j(\tau,\omega)\right|^2 \\ &= 0 \ \text{ for } \forall\, k. \end{aligned} \quad (22)$$

Therefore, the cost function $J(\tau,\omega)$ is zero for all k arguments, i.e., $J_1 = J_2 = 0$ thus there is no benefit achieved at all. Here, the cost function cannot distinguish the k arguments, the mixture is not separable.

Case 2: Different sources but setting γ and δ for the pseudo-stereo mixture such that $a_1(t;\delta,\gamma) = a_2(t;\delta\gamma)$ that 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 $x_2(t) = ((-a(\delta;t)+\gamma)/1+|\gamma|)x_1(t-\delta) + r_1(t;\delta,\gamma) + r_2(t;\delta,\gamma)$.

This case differs from the previous case 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 through $C_j(\tau,\omega)$, the separability of this mixture can be analyzed using (20) as follows:

$$\begin{aligned} J(\tau,\omega) &= \arg\min_k |a(\tau)e^{-i\omega\delta}S_j(\tau,\omega) \\ &\qquad -C_k(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega) \\ &\qquad +\Sigma_{\substack{m=1\\ m\neq\delta}}^{D_j}\frac{a_{s_j}(m;\tau)e^{-i\omega m}}{1+|\gamma|}S_j(\tau-m,\omega) \\ &\qquad -a(\tau)e^{-i\omega\delta}S_j(\tau,\omega)|^2 \\ &= \arg\min_k \left|\Sigma_{\substack{m=1\\ m\neq\delta}}^{D_j}\frac{\left(a_{s_j}(m;\tau)-a_{s_k}(m;\tau)\right)}{1+|\gamma|}e^{-i\omega m}\right|^2 \\ &\qquad \times\left|S_j(\tau,\omega)\right|^2. \end{aligned} \quad (23)$$

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

Case 3: General case where the sources are distinct, and γ and δ are selected arbitrarily such that the mixing attenuations and residues are also different. Case 3 can be expressed as follows: If $a_1(t;\delta,\gamma) \neq a_2(t;\delta,\gamma)$ and $r_1(t;\delta,\gamma) = r_2(t;\delta,\gamma)$ (or $r_1(t;\delta,\gamma) = r_2(t;\delta,\gamma)$) then $x_2(t) = ((-a_1(\delta;t)+\gamma)/1+|\gamma|)s_1(t-\delta)((-a_2(\delta;t)+\gamma)/1+|\gamma|)s_2(t-\delta) + r_1(t;\delta,\gamma) + r_2(t;\delta,\gamma)$.

We first treat the situation of $r_1(t;\delta,\gamma) = r_2(t;\delta,\gamma)$. As the mixing attenuations $a_1(\tau)$ and $a_2(\tau)$ correspond, respectively,

TABLE I
SUMMARY OF THE SEPARABILITY OF THE PSEUDO-STEREO MIXTURE MODEL THROUGH $a_j(t;\delta,\gamma)$ AND $r_j(t;\delta,\gamma)$ PARAMETERS

Case	I	II	III
Condition	$a_1 = a_2$& $r_1 = r_2$	$a_1 = a_2$& $r_1 \neq r_2$	$a_1 \neq a_2$& $(r_1 = r_2 \parallel r_1 \neq r_2)$
Separability	No	Yes	Yes

a_j for $a_j(t;\delta,\gamma)$ and r_j for $r_j(t;\delta,\gamma)$.

to $s_1(t)$ and $s_2(t)$ then the cost function can be expressed as follows:

$$\begin{aligned} J(\tau,\omega) = \arg\min_k & |a_k(\tau)e^{-i\omega\delta}S_j(\tau,\omega) \\ & -C(\tau,\omega)e^{-i\omega\delta}S_j(\tau,\omega) \\ & +\sum_{\substack{m=1\\ m\neq\delta}}^{D_j}\frac{a_s(m;\tau)e^{-i\omega m}}{1+|\gamma|}S_j(\tau-m,\omega) \\ & -a_j(\tau)e^{-i\omega\delta}S_j(\tau,\omega)|^2 \\ = \arg\min_k & \left|\left(a_k(\tau)-a_j(\tau)\right)e^{-i\omega\delta}\right|^2\left|S_j(\tau,\omega)\right|^2. \end{aligned} \tag{24}$$

This cost function yields a nonzero value only for $k \neq j$. Here, the cost function can separate the k arguments because of 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 cost function becomes as follows:

$$\begin{aligned} J(\tau,\omega) = \arg\min_k & \Bigg[|(a_k(\tau)-a_j(\tau))e^{-i\omega\delta}+\Sigma_{\substack{m=1\\ m\neq\delta}}^{D_j} \\ & \times\frac{(a_{s_j}m;\tau)-a_{s_k}(m;\tau)}{1+|\gamma|}e^{-i\omega m}|^2\,|S_j(\tau,\omega)|^2\Bigg]. \end{aligned} \tag{25}$$

This cost function yields a nonzero value only for $k \neq j$; thus the cost function is able to distinguish the k arguments. By considering $a_j(t;\delta,\gamma)$ and $r_j(t;\delta,\gamma)$ with respect to above three cases, the summarized results are shown in Table I.

III. SINGLE-CHANNEL DEMIXING METHOD

A. Parameter Estimation Using Complex 2-D Histogram

The core concept of developing a separating process is to construct a cost function that usually requires the knowledge of the relative AR coefficients and time-delay. In direct contrast to the DUET algorithm, the proposed SOLO does not require explicit estimation of the time-delay parameter δ. Rather, this parameter is selected by the user to form the pseudo-stereo mixture in (4) along with the weight parameter γ. Detailed information on this selection is deferred to Section IV. By formulation, the conventional DUET method considers the residues in (7) as noise because it does not represent the variable to be optimized. Therefore, this leads to biased estimate of the mask. In the proposed SOLO, we view the residues as information-carrying signal because it still retains information on other AR coefficients that are beneficial to the separability of the mixture (Section II-D). Hence, the proposed method does not need exact estimation of the AR coefficients. Rather, it uses $\bar{a}_k(\tau,\omega)$ in (19) and as $C_j(\tau,\omega)$ is spread over the frequency, $\bar{a}_k(\tau,\omega)$ does not necessarily have to be precisely estimated. To begin, let us assume the jth source is dominant at a particular TF unit as follows:

$$\begin{aligned} X_1(\tau,\omega) &= S_j(\tau,\omega) \\ X_2(\tau,\omega) &= a_j(\tau)e^{-i\omega\delta}S_j(\tau-\delta,\omega) \\ &\quad -\Sigma_{\substack{m=1\\ m\neq\delta}}^{D_j}\frac{a_{s_j}(m;\tau)}{1+|\gamma|}e^{-i\omega m}S_j(\tau-m,\omega) \\ &\approx \left[a_j(\tau)-C_j(\tau,\omega)\right]e^{-i\omega\delta}S_j(\tau,\omega), \quad (\tau,\omega)\in\mathbf{\Omega}_j \end{aligned} \tag{26}$$

for δ and $m \leq \phi, C_j(\tau,\omega)$ is given by (17) and $\mathbf{\Omega}_j$ is the active area of $S_j(\tau,\omega)$ defined as follows:

$$\mathbf{\Omega}_j := \left\{(\tau,\omega)\,;\, S_j(\tau,\omega)\neq 0\ \forall k\neq j\right\}. \tag{27}$$

The estimate of $\bar{a}_j(\tau,\omega) = a_j(\tau) - C_j(\tau,\omega)$ associated with the source can be determined as follows:

$$\begin{aligned} \bar{a}_j(\tau,\omega) &= \frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}e^{i\omega\delta} \\ &= a_j(\tau) - C_j(\tau,\omega) \\ &= \bar{a}_j^{(r)}(\tau,\omega) + i\bar{a}_j^{(i)}(\tau,\omega)\ \forall(\tau,\omega)\in\mathbf{\Omega}_j \end{aligned} \tag{28}$$

where $\bar{a}_j^{(r)}(\tau,\omega) = \mathrm{Re}[X_2(\tau,\omega)e^{i\omega\delta}/X_1(\tau,\omega)]$, $\bar{a}_j^{(i)}(\tau,\omega) = \mathrm{Im}[X_2(\tau,\omega)e^{i\omega\delta}/X_1(\tau,\omega)]$ 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. To overcome this problem, we propose a weighted complex 2-D histogram estimation method as a function of (τ,ω) with the weight $\Sigma_{\tau,\omega}|X_1(\tau,\omega)X_2(\tau,\omega)|$ to estimate $\bar{a}_j(\tau,\omega)$ and cluster them into N groups (where N is the total number of sources in the mixture). In particular, the real and imaginary parts of $\bar{a}_j(\tau,\omega)$ can be estimated as follows:

$$\begin{aligned} \hat{\bar{a}}_j^{(r)} &= \frac{\Sigma_{\tau,\omega}|X_1(\tau,\omega)X_2(\tau,\omega)|\,Re\left[\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}e^{i\omega\delta}\right]}{\Sigma\tau,\omega\,|X_1(\tau,\omega)X_2(\tau,\omega)|} \\ \hat{\bar{a}}_j^{(i)} &= \frac{\Sigma_{\tau,\omega}|X_1(\tau,\omega)X_2(\tau,\omega)|\,Im\left[\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)}e^{i\omega\delta}\right]}{\Sigma\tau,\omega\,|X_1(\tau,\omega)X_2(\tau,\omega)|}. \end{aligned} \tag{29}$$

The above can then be combined to form the estimate of (28) as follows:

$$\hat{\bar{a}}_j = \hat{\bar{a}}_j^{(r)} + \hat{\bar{a}}_j^{(i)}. \tag{30}$$

Relating (30) with (28), we can use similar idea to express $\widehat{\bar{a}}_j = \widehat{a}_j - \hat{C}_j$ where $\hat{a}_j$ and $\hat{C}_j$ are the complex 2-D histogram estimates of $a_j(\tau)$ and $C_j(\tau,\omega)$, respectively.

B. Construction of Masks

Here, we will establish the construction of the binary TF masks 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 following instantaneous likelihood function as follows:

$$\begin{aligned} &F_j(\tau,\omega) \\ &:= p\left(X_1(\tau,\omega),\ X_2(\tau,\omega)|\hat{\hat{a}}_j\right) \\ &= \frac{1}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}\frac{\left|\hat{\hat{a}}_j e^{-i\omega\delta}X_1(\tau,\omega)-X_2(\tau,\omega)\right|^2}{1+\hat{\hat{a}}_j^2}\right). \quad (31) \end{aligned}$$

We can show that the above is derived from the maximum likelihood (ML) framework by first formulating the Gaussian likelihood function $p(X_1(\tau,\omega), X_2(\tau,\omega)|S_j(\tau,\omega), \widehat{\widehat{a}}_j)$ using (26), maximizing the likelihood function with respect to $S_j(\tau,\omega)$ and then substituting the obtained result into the Gaussian likelihood function. The instantaneous likelihood function $F_j(\tau,\omega)$ in (31) clusters every (τ,ω) unit to the jth dominating source for $F_j(\tau,\omega) \geq F_k(\tau,\omega), \forall k \neq j$. This process is equivalent in minimizing the following:

$$G(\tau,\omega) = \arg\min_k \left|\hat{\hat{a}}_k e^{-i\omega\delta}X_1(\tau,\omega)-X_2(\tau,\omega)\right|^2. \quad (32)$$

Using (15), the third term of $X_2(\tau,\omega)$ can be expressed as follows:

$$\begin{aligned} &\frac{1}{1+|\gamma|}\Sigma^{D_j}_{\substack{m=1\\ m\neq\delta}} a_{S_j}(m;\tau)e^{-i\omega m}S_j(\tau-m,\ \omega) \\ &= \frac{1}{1+|\gamma|}\big(-S_j(\tau,\omega)+E_j(\tau,\omega) \\ &\quad -a_{S_j}(\delta;\tau)e^{-i\omega\delta}S_j(\tau-\delta,\ \omega)\big) \\ &\approx -\frac{1}{1+|\gamma|}\left(1+a_{S_j}(\delta;\tau)e^{-i\omega\delta}\right)S_j(\tau,\omega)+\frac{E_j(\tau,\omega)}{1+|\gamma|} \\ &= -\frac{1}{1+|\gamma|}\left(1+(\gamma-a_j(\tau)(1+|\gamma|))e^{-i\omega\delta}\right)S_j(\tau,\omega) \\ &\quad +\frac{E_j(\tau,\omega)}{1+|\gamma|} \\ &= -\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)S_j(\tau,\omega)+a_j(\tau)e^{-i\omega\delta}S_j(\tau,\omega) \\ &\quad +\frac{E_j(\tau,\omega)}{1+|\gamma|} \quad (33) \end{aligned}$$

for $\delta \leq \phi$ and by invoking the local stationarity at the fourth line of (33). Using (33), $X_2(\tau,\omega)$ can be now expressed as follows:

$$X_2(\tau,\omega) \approx \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)X_1(\tau,\omega)-\frac{E_j(\tau,\omega)}{1+|\gamma|}. \quad (34)$$

In this light, the proposed cost function can be formulated based on the single mixture $X_1(\tau,\omega)$ by substituting this relation into (32) which leads to the following:

$$J(\tau,\omega) = \arg\min_k H_k(\tau,\omega) \quad (35)$$

where

$$H_k(\tau,\omega)=\left|\hat{\hat{a}}_k e^{-i\omega\delta}X_1(\tau,\omega)-\left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)X_1(\tau,\omega)\right|^2. \quad (36)$$

As $e_j(t) \ll s_j(t)$, the term $E_j(\tau,\omega)/(1+|\gamma|)$ is negligible. Hence, $X_2(\tau,\omega) = \left(\frac{1+\gamma e^{-i\omega\delta}}{1+|\gamma|}\right)$. In below, we elucidate how the above cost function works. Initially, we assume the jth source is dominant at $(\tau,\omega) \in \mathbf{\Omega}_j$ and then consider the case when $k = j$

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

$$\begin{aligned} H_{k\neq j}(\tau,\omega) = &\left|\left(\hat{\hat{a}}_k - a_j(\tau) - \frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right)e^{-i\omega\delta}-\frac{1}{1+|\gamma|}\right|^2 \\ &|S_j(\tau,\omega)|^2. \quad (38) \end{aligned}$$

Using (37), as shown at the bottom of the page, and (38), we may thus state that when the jth source dominates at $(\tau,\omega) \in \mathbf{\Omega}_j$ the cost function will correctly identify the source 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 (37) and (38), we have the following:

$$\begin{aligned} &\left|\left(\hat{C}_j+\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right)e^{-i\omega\delta}+\frac{1}{1+|\gamma|}\right|^2 \\ &< \left|\left(\hat{\hat{a}}_k-a_j(\tau)-\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right)e^{-i\omega\delta}-\frac{1}{1+|\gamma|}\right|^2. \quad (39) \end{aligned}$$

Let $\beta_j = \hat{C}_j + a_{s_j}(\delta;\tau)/1+|\gamma|$ and $\beta_l = \widehat{\widehat{a}}_k - a_j(\tau) - a_{s_j}(\delta;\tau)/1+|\gamma|$, then the above becomes as follows:

$$\left|\beta_j e^{-i\omega\delta}+\frac{1}{1+|\gamma|}\right|^2 < \left|\beta_l e^{-i\omega\delta}-\frac{1}{1+|\gamma|}\right|^2. \quad (40)$$

$$\begin{aligned} H_{k=j}(\tau,\omega) &= \left|\hat{\hat{a}}_j e^{-i\omega\delta}S_j(\tau,\omega)-\left(\frac{1+\gamma e^{-\omega\delta}}{1+|\gamma|}\right)S_j(\tau,\omega)\right|^2 \\ &= \left|\hat{\hat{a}}_j e^{-i\omega\delta}S_j(\tau,\omega)-\widehat{C}_j e^{-i\omega\delta}S_j(\tau,\omega)-a_j(\tau)e^{-i\omega\delta}S_j(\tau,\omega)+\ \Sigma^{D_j}_{\substack{m=1\\ m\neq\delta}}\left(\frac{a_{S_j(m;\tau)}e^{-i\omega m}}{1+|\gamma|}\right)S_j(\tau-m,\omega)\right|^2 \\ &= \left|-\widehat{C}_j e^{-\omega\delta}S_j(\tau,\omega)+\Sigma^{D_j}_{m=1}\frac{a_{S_j}(m;\tau)e^{-i\omega\delta}}{1+|\gamma|}S_j(\tau-m,\omega)-\frac{a_{S_j}(\delta;\tau)e^{-i\omega\delta}}{1+|\gamma|}S_j(\tau-\delta,\omega)\right|^2 \\ &= \left|-\widehat{C}_j e^{-i\omega\delta}S_j(\tau,\omega)-\frac{S_j(\tau,\omega)}{1+|\gamma|}-\frac{a_{S_j}(\delta;\tau)e^{-i\omega\delta}}{1+|\gamma|}S_j(\tau-\delta,\omega)\right|^2 \\ &= \left|\left(\widehat{C}_j+\frac{a_{S_j}(\delta;\tau)}{1+|\gamma|}\right)e^{-i\omega\delta}+\frac{1}{1+|\gamma|}\right|^2 |S_j(\tau,\omega)|^2 \quad (37) \end{aligned}$$

The left-hand side of (40) is bounded below by

$$\begin{aligned}\left|\beta_j e^{-i\omega\delta} + \frac{1}{1+|\gamma|}\right| &\geq |\beta_j| - \frac{1}{1+|\gamma|}\\ &= \left|\hat{C}_j + \frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right| - \frac{1}{1+|\gamma|}\\ &\geq \left|\hat{C}_j\right| - \left|\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right| - \frac{1}{1+|\gamma|} \quad (41)\end{aligned}$$

and the right-hand side of (40) is bounded above by

$$\begin{aligned}\left|\beta_l e^{-i\omega\delta} + \frac{1}{1+|\gamma|}\right| &\leq \left|\beta_l e^{-i\omega\delta}\right| + \frac{1}{1+|\gamma|}\\ &= |\beta_l| + \frac{1}{1+|\gamma|}\\ &= \left|\hat{\hat{a}}_k - a_j - \frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right| + \frac{1}{1+|\gamma|}. \quad (42)\end{aligned}$$

Substituting (41) and (42) into (40) and replugging the terms for β_j and β_l, (40) results in the following:

$$|\hat{C}_j| < \widehat{a}_k - a_j(\tau) - \frac{a_{s_j}(\delta;\tau)}{1+|\gamma|} + \frac{2}{1+|\gamma|} + \left|\frac{a_{s_j}(\delta;\tau)}{1+|\gamma|}\right| \quad (43)$$

for $\forall_j \neq k$.The cost function (35) and (36) will correctly assign the (τ,ω) unit to the jth source if the $|\hat{C}_j|$ condition in (43) is satisfied across $\mathbf{\Omega}_j$. Conversely, if $|\hat{C}_j|$ is larger than the right-hand side of (43) then this will lead to wrong assignment of the TF units. Once the cost function is evaluated, the binary TF mask for the jth source can be constructed as follows:

$$M_j(\tau,\omega) := \begin{cases} 1, & J(\tau,\omega) = j\\ 0, & \text{otherwise.}\end{cases} \quad (44)$$

The proposed SOLO algorithm is shown in Algorithm 1.

IV. Determination of the Values for γ and δ

The pseudo-stereo mixture is formulated through determining the weight γ and the time-delay δ parameters. The separability of the proposed method depends on the sources' AR coefficients estimated from the relation of $X_1(\tau,\omega)$, $X_2(\tau,\omega)$, and their residues. The weight γ parameter acts as a controlling factor to maintain the difference of the sources' AR coefficients and to control the amount of the residues $r_j(t;\delta,\gamma)$. On the one hand, if $\gamma \gg a_{s_j}(\delta;t)$ then the distinguishing ability of the mixing attenuations in (6) will tend to be small such that $a_1(\tau) = a_2(\tau)$ and thereby, we lose the benefit of the pseudomixture signal. In additional, it reduces the residues in (7) that subsequently diminishes the contribution of $C_j(\tau,\omega)$ in $\bar{a}_j(\tau,\omega)$. On the other hand, if $\gamma \ll a_{s_j}(\delta;t)$ 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.

1) Mixing Attenuation Distinguishability: We define the distinguishability function of the mixing attenuation as follows:

$$\begin{aligned}\theta &= \frac{|a_k e^{i\omega\delta} X_1(\tau,\omega) - X_2(\tau,\omega)|^2}{|X_1(\tau,\omega)|^2}\\ &= \frac{|a_k e^{i\omega\delta} S_j(\tau,\omega) - a_j e^{-i\omega\delta} S_j(\tau,\omega)|^2}{|S_j(\tau,\omega)|^2} = |a_k - a_j|^2 \quad (46)\end{aligned}$$

Algorithm 1 Pseudocode of SOLO Algorithm

```
Determine values of $\gamma$ and $\delta$ parameters. (Algorithm 1: for selecting an appropriate $\gamma$ and $\delta$.
Transform $x_1(t)$ and $x_2(t)$ into TF domain $X_1(\tau,\omega)$ and $X_2(\tau,\omega)$ using STFT.
Construct the weighted complex 2-D histogram in terms of $(\widehat{\bar{a}}_j^{(r)}, \widehat{\bar{a}}_j^{(r)})$ according to (29).
Identify $N$ peaks and combine $(\widehat{\bar{a}}_j^{(r)}, \widehat{\bar{a}}_j^{(i)})$ to form $\hat{\bar{a}}_j$ in (30).
For $k = 1{:}N$
    Evaluate the cost function $J(\tau,\omega)$ using (35) and (36).
For $j = 1$: number of sources
    Formulate the binary TF mask $M_j(\tau,\omega)$ using (44).
    Separate the observed mixture using
        $\tilde{S}_j(\tau,\omega) = M_j(\tau,\omega) X_1(\tau,\omega).$    (45)
    Convert $\tilde{S}_j(\tau,\omega)$ from TF domain into time domain $\tilde{S}_j(t)$.
```

for $k \neq 1$. The second line of (46) is obtained using (26). Larger value of θ implies that the mixing attenuations between the two sources are distant further from each other. This will yield two distinct peaks in the complex 2-D histogram. Alternatively, we can use the concept of symmetric mixing attenuation α_j which that is defined as $\alpha_j := a_j - 1/a_j$. Here, the distinguishability in terms of α_k and α_j takes the form as follows:

$$\theta = |\alpha_k - \alpha_j|^2. \quad (47)$$

2) Attenuation-to-Residue Ratio (ARR):

$$\text{ARR} = \arg\max_{\gamma,\delta} \frac{\theta}{\left|\frac{\kappa_1(\tau,\omega)+\kappa_2(\tau,\omega)}{\kappa_1(\tau,\omega)-\kappa_2(\tau,\omega)}\right|^2} \quad (48)$$

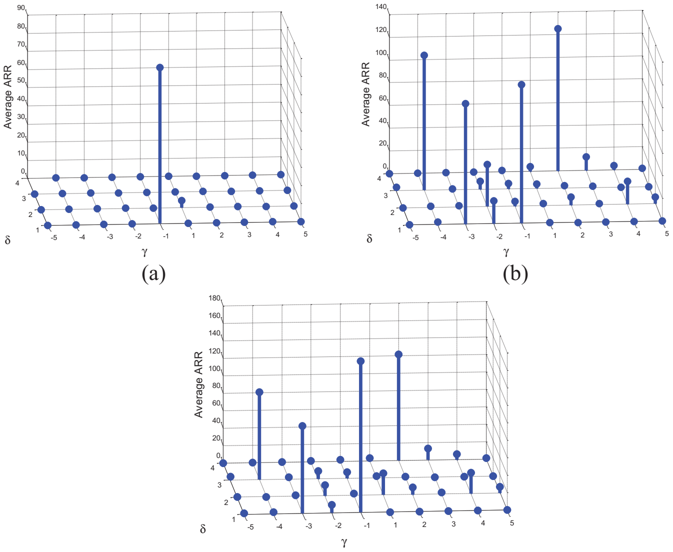
where

$$\kappa_j(\tau,\omega) = \sqrt{(D_j - 1)\Sigma_{\substack{m \neq 1\\ m \neq \delta}}^{D_j} \left|\frac{a_{s_j}(m)}{1+|\gamma|}\right|}^{\,2}$$

is the supremum of $C_j(\tau,\omega)$ that is obtained by applying the Schwarz's inequality to $C_j(\tau,\omega)$. The term $\kappa_1(\tau,\omega) - \kappa_2(\tau,\omega)|^2$ is the maximum difference of residues inspired by the cost function of Case 2 in Section II-C. On the other hand, the term $|\kappa_1(\tau,\omega) + \kappa_2(\tau,\omega)|^2$ is 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 $\widehat{a}_j$ of the two sources 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 $\widehat{\hat{a}}_j$ between the sources. Thus, choosing the appropriate γ and δ such that the two peak regions are clearly distinct in the complex 2-D 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 signal-to-distortion

Algorithm 2 Determination of γ and δ

```
1. Select an arbitrary range of $\delta$ that satisfies (14).
2. Calculate the ARR matrices on the range of $\gamma$ and $\delta$ using (47) and (48).
3. Choose pairs of $\gamma$ and $\delta$ such that the ARR is greater than a threshold, i.e.,
   $\Lambda = \{(\gamma, \delta) | ARR > \psi\}$   (49)
   where $\Lambda$ is the set of the selected pairs, and $\psi$ is a threshold.[1]
```



(a) (b) (c)

Fig. 1. Set of γ and δ for mixture of (a) SS, (b) MM, and (c) SM.

(SDR). Algorithm 2 shows the steps for determining the range of values for γ and δ.

A set of experiments is conducted to determine the γ and δ pairs using real-audio sources from TIMIT and RWC [29] databases. Seventy-five types of mixtures are constructed from these databases that are divided into three categories: 1) speech and speech (SS); 2) music and music (MM); and 3) speech and music (SM). Each type contains two sources and each source has unit power. All experiments are performed under the following conditions: STFT of 1024-point with 50% overlap and sampling frequency of 16 kHz. Source's AR coefficients are calculated using Yule–Walker method. A finite range of γ and δ is selected to be $[-5, 5]$ (excluding $\gamma = 0$) and [1, 4], respectively. Following the steps in Algorithm 1, the set of the pairs Λ with the threshold $\psi = 20$ is found by calculating the average ARR for each category and this is plotted is shown in Fig. 1. The results in Fig. 1(a) show at least a pair is found for the SS category, i.e., $\Lambda = \{(\gamma, \delta)|\text{ARR} > 20\} = (-1, 1)$. The results indicate that only the low-order AR coefficients, i.e., $\delta = 1$ are beneficial for separation. This is not surprising as 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 γ on the ARR. As γ increases in magnitude for both positive and negative directions, θ and $C_j(\tau, \omega)$ become progressively smaller such that the ARR is almost zero. Fig. 1(b) shows the results for the MM category with nine 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. 1(b). Finally, Fig. 1(c) shows the results of the SM category with six pairs identified as $\Lambda = \{(-4, 3), (-3, 1), (-1, 1), (1, 2), (2, 4), (4, 2)\}$. We may note that both MM and SM categories have broader range of γ and δ than the SS group because of 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 are shown in Table II.

TABLE II

COMMON PAIRS OF (γ, δ) IN THE MM AND SM CATEGORIES

γ	-1	4	2
δ	1	2	4

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

A. Number of Bins for the Complex 2-D Histogram

The weighted complex 2-D histogram is used to reveal the signature of each source by clustering the mixing parameter pairs $(\bar{a}_j^{(r)}(\tau, \omega), \bar{a}_j^{(i)}(\tau, \omega))$ within the histogram widths $(\Delta_{\alpha^{(r)}}, \Delta_{\alpha^{(i)}})$. Here, $\Delta_{\alpha^{(r)}}$ and $\Delta_{\alpha^{(i)}}$ are the maximum value of $\alpha^{(r)}$ and $\alpha^{(i)}$, respectively. The selected peak positions are the estimated $\widehat{\bar{a}}_j = \widehat{\bar{a}}_j^{(r)} + i\widehat{\bar{a}}_j^{(i)}$. This therefore means that the resolution of the complex 2-D histogram depends on the number of bins. If the number of bins is too small, then the histogram loses resolution. Conversely, if it is too large, then the histogram will appear scattered. Hence, the determination of the number of bins required to construct an appropriately resolved complex 2-D histogram is vital. In this paper, we propose that the number of resolution bins $\zeta^{(r)}$ and $\zeta^{(i)}$ can be calculated, respectively, as follows:

$$\zeta^{(r)} = \psi \Delta_{\alpha^{(r)}} + 1$$
$$\zeta^{(i)} = \frac{\Delta_{\alpha^{(i)}}}{\psi} - 1 \tag{50}$$

where $\zeta^{(r)}$ and $\zeta^{(i)}$ are the number of bins for $\alpha^{(r)}$ and $\alpha^{(i)}$, respectively, and ψ is the threshold selected in (49).

[1] By means of Monte-Carlo experiments [30], $\psi = 20$ has been experimentally verified to yield satisfactory performance.

V. Results and Analysis

The performance of SOLO is demonstrated by separating synthetic and real-audio sources. The synthetic sources are divided into stationary and nonstationary types that are, respectively, given by the stationary AR signals and chirp signals.[2] The real-audio sources that are inherently nonstationary include voice and music signals. 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 using the STFT of 1024-point Hamming window with 50% overlap [20], [31]. The separation performance is evaluated by measuring the distortion between original source and the estimated one according to the SDR ratio and signal-to-interference (SIR) ratio [32] defined as SDR$= 10\log_{10}(\| s_{\text{target}} \|^2 / \| e_{\text{interf}} + e_{\text{artif}} \|^2)$ and SIR $= 10\log_{10}(\| s^2_{\text{target}}\|/ \| e_{\text{interf}} \|^2)$ where e_{interf} is the interference from other sources and e_{artif} is the artifacts. The proposed approach will be compared with the (SNMF2-D) [17], the single-channel ICA (SCICA) [33], and the ideal binary mask (IBM) [34] that represents the ideal separation performance. The SNMF2-D parameters are set as follows [35], [36]: number of factors is two, sparsity weight of 1.1, number of phase shift and time shift is 31 and seven, respectively, for music. As for speech, both shifts are set to four. The TF domain used in SNMF2-D is based on the log-frequency spectrogram. Cost function of SNMF2-D is based on the Kullback–Leibler divergence. As for the SCICA, the number of block is ten 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 3-GHz and 3-GB RAM.

A. Stationary Sources

1) AR Sources: Two stationary AR sources are synthesized for $s_1(t)$ and $s_2(t)$ using the model (3) with 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 6.2×10^{-6} and 7.6×10^{-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 is not included in the above mentioned to avoid cluttering the notation. The source signals are shown in Fig. 3. 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 $\zeta^{(i)} = 3$.

Fig. 2 shows the clustering of the sources into two peaks that accord with the number of sources in the mixture. Fig. 3 also shows the mixed signal and the separated sources based on the SOLO method. It can be seen that the mixture is very well-separated comparing with the original sources.

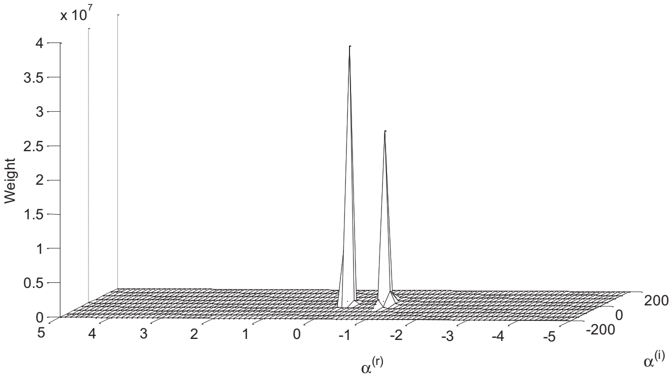


Fig. 2. Complex 2-D histogram corresponding to two sources.

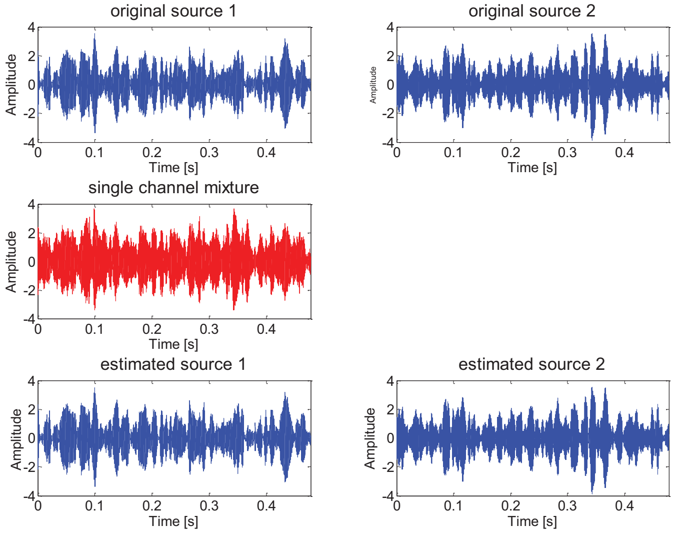


Fig. 3. Two original sources, observed mixture, and two estimated sources.

TABLE III

Comparison of Average SDR and SIR Performances on Mixture of Two AR Sources With SNMF2-D, SCICA, SOLO, and IBM

Methods	SDR s_1	SDRs_2	SIRs_1	SIRs_2
SNMF2-D	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

The separation performance is shown in Table III that shows the comparison results of SNMF2-D, 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 sources with high accuracy. In particular, the SOLO method renders an average SDR improvement of 13.4 dB/source over the SNMF2-D and 14.6 dB/source over the SCICA and an average SIR improvement of 56.1 dB per source and 53 dB/source over the SNMF2-D and SCICA, respectively. Because of the stationarity of the sources, the AR coefficients do not change with τ and thus $H_{k=j}(\tau, \omega) < H_{k\neq j}(\tau, \omega)$ can be satisfied only when $\left|\hat{C}_j\right| < \left|\widehat{\overline{a}}_k - a_j - a_{s_j}(\delta)/1 + |\gamma|\right| + 2/1 + |\gamma| + \left|a_{s_j}(\delta)/1 + |\gamma|\right|$ according to (43). For $j = 1$, we compute $|\hat{C}_1| = 14.3$ and $\left|\widehat{\overline{a_2}} - a_1 - a_{s_1}(2)/6\right| + 2/6 + \left|a_{s_2}(2)/6\right| = 2.24$ in which case we have $1.43 < 2.24$ thus, the $\left|\hat{C}_1\right|$ condition is satisfied. For $j = 2$, we have $\left|\hat{C}_2\right|$ and 1.01

[2] Chirp signal is classified as nonstationary because of its time-varying instantaneous frequency.

TABLE IV
AVERAGE SDR AND SIR RESULTS FOR MIXTURE OF 2–5 SOURCES

Mixture	γ	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

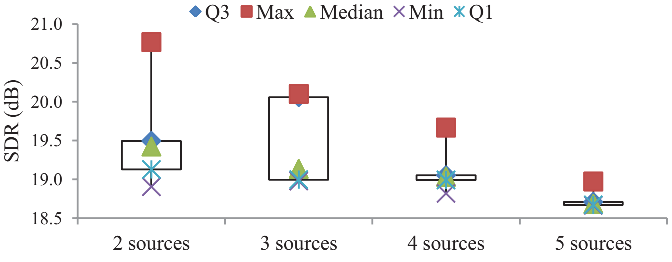


Fig. 4. Box plot of average SDR results.

$\left|\widehat{\overline{a_1}} - a_2 - a_{s_2}(2)/6\right| + 2/6 + \left|a_{s_2}(2)/6\right| = 1.67$ 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 (γ, ω) units to their respective original sources. This is clearly evident by the same SDR results between the SOLO and the IBM.

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

$$\begin{aligned} a_{s_1} &= [-3.8604, 5.6466, -3.7076, 0.9224] \\ a_{s_2} &= [-2.6189, 3.5578, -2.4136, 0.8493] \\ a_{s_3} &= [0.8773, 2.0937, 0.8340, 0.9037] \\ a_{s_4} &= [2.9132, 3.9841, 2.7128, 0.8672] \\ a_{s_5} &= [3.8148, 5.5394, 3.6264, 0.9037] \end{aligned}$$

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 8.54×10^{-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 is shown in Table IV and Fig. 4 shows the corresponding box plot. Separation performance progressively deteriorates as the number of sources increases. When the sources are not perfectly estimated and become slightly mutually correlated [24], the projection of these sources to the original source 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 sources increases, this inadvertently leads to larger values of e_{interf} and e_{artif}, and subsequently decreased the SDR and SIR performances. This explains the result for five sources that shows a drop in performance. Although this is the case, the SDR and SIR results are still maintained at a high level. The complex 2-D histogram, shown in Fig. 5, distinctively enumerates five peaks that correspond to the number of sources in the mixture. Figs. 6 and 7 show the original sources, the mixture, and the separated sources. We can visually inspect that the separated sources are very similar to the original sources. In this experiment, the sources satisfy the assumptions in Section II-B 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 successfully identifies and partitions the mixed signal TF plane into the correct group of sources.

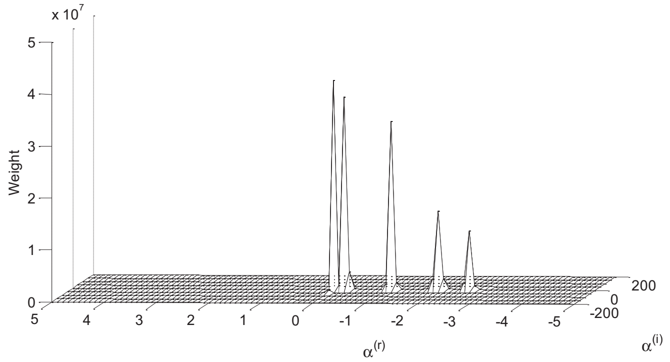


Fig. 5. Complex 2-D histogram of a mixture of five sources.

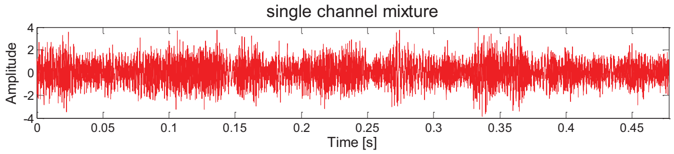


Fig. 6. Single-channel mixture.

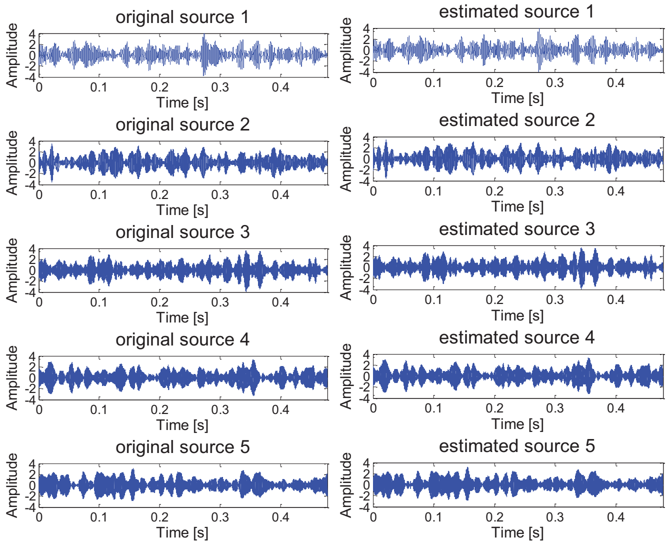


Fig. 7. Original sources (left) and estimated sources (right).

B. Nonstationary Sources

As the proposed method estimates the parameter $\widehat{\overline{a}}_j$ from the complex 2-D histogram, its result is based on the averaged AR coefficient of each source. As such, the estimated $\widehat{\overline{a}}_j$ befits very well the purpose of separating stationary AR sources. In nonstationary sources, we may readily adapt this approach and

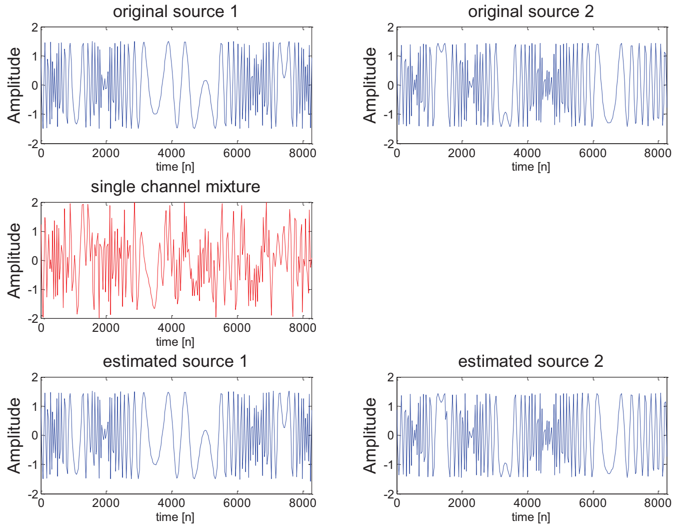


Fig. 8. Original sources, single-channel mixture, and estimated sources using SOLO with $L = 5$.

invoke the assumption of quasi-stationary from Section II-B. In effect this enables us to work under the condition where the AR parameters are stationary within a block but variable 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 $\widehat{\bar{a}}_j = [\widehat{\bar{a}}_{j1}\widehat{\bar{a}}_{j2}\cdots\widehat{\bar{a}}_{jL}]$ where $\widehat{\bar{a}}_{jl}$ is the estimate of $\widehat{\bar{a}}_j$ from the lth block. A mask will subsequently be constructed in exactly the same manner in (35) but using the aggregated $\widehat{\bar{a}}_j$ obtained from each block.

1) Chirp Sources: In this example, chirp signals are used to demonstrate the effectiveness of the SOLO method in dealing with nonstationary sources. s_1 is a down-chirp whose center frequency varies from 3.3 to 2 kHz. s_2 is a quadratic-chirp signal whose center frequency varies from 0.5 to 1.8 kHz. Both sources are mixed with equal average power over the duration of the signals. The single-channel mixture is first divided into L nonoverlapping 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. 8 shows the two-synthesized chirp sources, the single-channel mixture and the separated sources using the SOLO with $L = 5$. From the plots, it is visually evident that the mixture is clearly separated comparing with the original sources.

In Table V, we show the comparison results of SNMF2-D, SCICA, SOLO with $L = 1, 3, 5$, and IBM. In general, the SOLO yields far superior separating results compared with the SNMF2-D and the SCICA with an average SDR improvement of 9.0 and 8.3 dB/source, and with an average SIR improvement of 15.3 and 16.0 dB, respectively. With the use of $\widehat{\bar{a}}_j = [\widehat{\bar{a}}_{j1}\widehat{\bar{a}}_{j2}\cdots\widehat{\bar{a}}_{jL}]$ partition, SOLO with $L > 1$ leads to substantially better separation performance than the SOLO with $L = 1$. From Table V, the average SDR and SIR performances increases by Warning: 4 and 6 dB/source, respectively, when $L = 5$.

Because the sources have time-varying instantaneous frequencies, $R_j(\tau, \omega)$ in (16) will change accordingly with ω and τ. As $\bar{a}_j(\tau, \omega)$ composes $a_j(\tau)$ and $R_j(\tau, \omega)$, it follows that $\bar{a}_j(\tau, \omega)$ will also vary with ω and τ. Unfortunately, setting $L = 1$ will mean that $\widehat{\bar{a}}_j = \widehat{\bar{a}}_{j1}$, which only estimates the global average of $\bar{a}_j(\tau, \omega)$ for all (τ, ω). Thus, the obtained result of $\widehat{\bar{a}}_{j1}$ can yield significant deviation from the true $\bar{a}_j(\tau, \omega)$. Therefore, the SDR and SIR performances of SOLO with $L = 1$ are not as high as in the previous case of stationary sources. On the other hand, when the mixture signal is divided into L blocks such that each block resembles a mixture of frequency-invariant sources similar to the AR sources, then $\bar{a}_j(\tau, \omega)$ in each block can be treated as constant. As such, the cost function renders by $\widehat{\bar{a}}_j = [\widehat{\bar{a}}_{j1}\widehat{\bar{a}}_{j2}\cdots\widehat{\bar{a}}_{jL}]$ will enable all the TF units in each block to be specifically labeled using the estimated $\widehat{\bar{a}}_{jl}$ derived from that block. Therefore, better separation performance can be obtained as shown in Table V.

TABLE V

COMPARISON OF SDR AND SIR PERFORMANCES ON MIXTURE OF CHIRP SOURCES WITH SNMF2-D, SCICA, SOLO, AND IBM

Methods	SDR S_1	SDR S_2	SIR S_1	SIR S_2
SNMF2-D	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

2) Real Audio Sources: Audio sources can be characterized as nonstationary AR processes as their AR coefficients vary with time. As an example, three type of mixtures are generated, i.e., male speech + jazz, female speech + jazz, and male speech + piano. The male and female speeches are selected from TIMIT and music sources from the RWC [29] database. Both sources are mixed with equal power to generate the mixture. This is shown in the first three panels of Fig. 9. To perform separation, we first divide the mixture into L nonoverlapping partitions. Two possible choices are available. The first choice is to partition the mixture into equal-length L blocks. We investigate the separation performance 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 shown in Table VI along with SNMF2-D, SCICA, and IBM. In general the SOLO with increasing the number of blocks shows better separation performance than the SNMF2-D and SCICA. From Table VI, the performance remains high when using $L = 15$ where the average SDR and SIR results are 7.7 dB per source and 19.7 dB/source, respectively. When L increases, each block becomes progressively narrower and contains less samples. The condition (43) may not be satisfied in some of these blocks particularly those of small amplitudes. Here, the obtained mask may wrongly assign some of the TF units to the incorrect source. Therefore, the SIR value is slightly decreased. The proposed SOLO method renders an average SDR improvement of 1.2 and 2.1 dB/source over SNMF2-D and SCICA, respectively. Fig. 10 shows the box plot corresponding to the above results.

The second choice is to examine the characteristics and identify the transition behavior in the mixture signal. In this

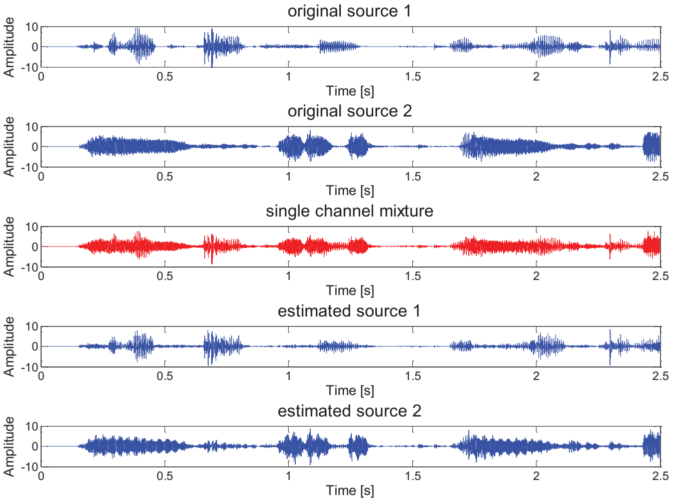


Fig. 9. Original sources, single-channel mixture, and estimated sources in time domain using the SOLO with $L = 64$ nonuniform blocks.

TABLE VI

COMPARISON OF AVERAGE SDR AND SIR PERFORMANCES ON MIXTURE OF TWO AUDIO SOURCES BETWEEN SNMF2-D, SCICA, SOLO, AND IBM

Methods	SDR S_1	SDR S_2	SIR S_1	SIR S_2
SNMF2-D	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 SM, respectively.

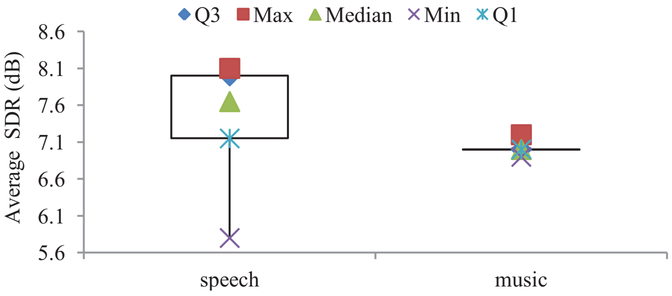


Fig. 10. Box plot of average SDR results on mixture of two audio sources versus the number of blocks.

TABLE VII

COMPARISON OF SDR PERFORMANCE ON MIXTURE OF TWO AUDIO SOURCES USING SOLO WITH NONUNIFORM LENGTH

Methods	SDR S_1	SDR S_2	SIR S_1	SIR S_2
SOLO ($L = 3$ with nonuniform blocks)	7.9	7.1	20.8	17.3
SOLO ($L = 6$ with nonuniform blocks)	8.1	7.3	21.7	17.5

Note that s_1 and s_2 refer to SM, respectively.

TABLE VIII

COMPUTATION COMPLEXITY OF SNMF2-D, SCICA, AND SOLO

Methods	Number of Operations
SNMF2-D	$2N \log_2 L + CN_s[3\tau \frac{L}{2} + 2N_\phi N_\tau N + N_\phi(4\frac{N}{L} + 2N_\tau N)$ $+2N_\phi N_\tau(N + \frac{L}{2} + N_\tau(N + \frac{L}{2} + N_s \frac{L}{2}))]$
SCICA	$[2K(K+1)(N-K+1)IK + K^3 + 2(K(N-K+1))+$ $(K^2 + K(K-1))(N-K+1)]N_s$
SOLO	$5N + L + 4NN_s + 2N \log_2 L$

case, the window size for each block is not required to be identical. We will consider two examples. In the first example, we set $L = 3$ where it can be observed that the mixture of a male speech and Jazz music shows a transition at time $t = 0.85\,\text{s}$ and in the interval around $t = 1.5\,\text{s}$. Thus, this enables us to partition the mixture into the following blocks, i.e., $T_1 = [0, 0.85\,s]$, $T_2 = (0.85\,s, 1.5\,s]$, and $T_3 = (1.5\,s, 2.5\,s]$. In the second example, the mixture signal is partitioned into $L = 6$ blocks, i.e., $T_1 = [0, 0.64\,s]$, $T_2 = (0.64\,s, 0.86\,s]$, $T_3 = (0.86\,s, 1.06\,s]$, $T_4 = (1.06\,s, 1.38\,s]$, $T_5 = (1.38\,s, 2.18\,s]$, and $T_6 = (2.18\,s, 2.5\,s]$. The SDR results are shown in Table VII. With $L = 3$ nonuniform blocks, the SDR performance gives 7.5 dB/source that matches the case of $L = 9$, and $L = 12$ equal-length blocks. On the other hand, with $L = 6$ nonuniform blocks the SDR performance gives 7.7 dB/source that matches the equal-length partition scheme of $L = 15$. The separated sources are plotted in the last panels of Fig. 9. The separated sources resemble closely to the original sources. The IBM results are also 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.

We also calculate the computational complexity of SNMF2-D, SCICA, and the proposed SOLO on a function of N sample size of a signal (N), number of sources (N_s), length of the STFT window (L), number of frequency-shifts (N_ϕ) and time-shift (N_τ) for the SNMF2-D, number of iterations for SNMF2-D (C) and SCICA (I), and number of SCICA blocks (K). This is shown in Table VIII.

We plot the computation complexity of the above algorithms and this is shown in Fig. 11 with the following parameters: $N_s = 2$, $L = 1024$, $N_\phi = 31$, $N_\tau = 7$, $C = 100$, $I = 100$, $K = 10$, and N varies from 1×10^4 to 8×10^4. We note that SOLO is computationally less demanding than SNMF2-D and SCICA. The reason is SOLO does not require any iteration for updating parameters. On the other hand, SNMF2-D requires updating the spectral basis and the mixing of the sources. 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 source of interest; and 3) the steps are repeated until all sources are extracted. Fig. 10 shows that the complexity of SCICA is almost identical to SNMF2-D in the region of 10^{10} operations. Thus, the overall computational complexity associated with

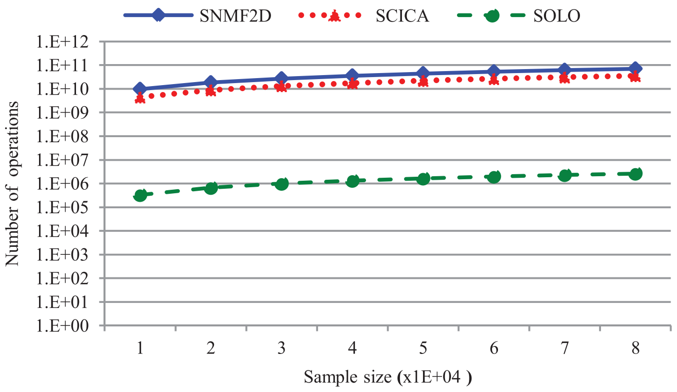


Fig. 11. Comparison of computational complexity on mixture of two audio sources between SNMF2-D, SCICA, and SOLO.

both algorithms is significantly high. On the other hand, the proposed SOLO consumes the least computation that renders it very fast and yet yields the best separation performance among the three methods.

VI. Conclusion

In this paper, a novel single-channel blind separation algorithm was presented. The proposed method constructed a pseudo-stereo mixture by time-delaying and weighting the observed single-channel mixture. The method assumed that the source signals were characterized as AR processes. Experiments were conducted successfully to separate stationary as well as time-varying AR sources. In this paper, the separability analysis of the pseudo-stereo mixture was derived and the conditions required for unique mask construction from the ML framework were also identified. The proposed method demonstrated high-level separation performance for both synthetic and real-audio sources. The proposed method enjoys at least three advantages: 1) it does not require *a priori* knowledge of the sources; 2) the proposed approach is able to capture the music and speech characteristics and hence, renders robustness to the separation method; and 3) 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.

References

[1] J. Zhang, W. L. Woo, and S. S. Dlay, “Blind source separation of post-nonlinear convolutive mixture,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2311–2330, Nov. 2007.

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

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

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

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

[6] 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.

[7] 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.

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

[9] G. J. Brown and M. P. Cooke, “Computational auditory scene analysis,” *Comput. Speech Lang.*, vol. 8, no. 4, pp. 297–336, 1994.

[10] Y. Jiang and H. Zhou, “An algorithm combined with spectral subtraction and binary masking for monaural speech segregation,” in *Proc. IEEE Int. Conf. Signal Process., Commun. Compon.*, Sep. 2011, pp. 1–4.

[11] K. Hu and D. L. Wang, “An unsupervised approach to cochannel speech separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 122–131, Jan. 2013.

[12] B. Gao, W. L. Woo, and S. S. Dlay, “Single channel blind source separation using the best characteristic basis,” in *Proc. 3rd Int. Conf. Inf. Commun. Technol., Theory Appl.*, Apr. 2008, pp. 1–5.

[13] J.-I. Hirayama, S.-I. Maeda, and S. Ishii, “Markov and semi-Markov switching of source appearances for nonstationary independent component analysis,” *IEEE Trans. Neural Netw.*, vol. 18, no. 5, pp. 1326–1342, Sep. 2007.

[14] R. J. Weiss and D. P. W. Ellis, “Speech separation using speaker-adapted eigenvoice speech models,” *Comput. Speech Lang.*, vol. 24, no. 1, pp. 16–29, 2010.

[15] G. J. Jang and T. W. Lee, “A maximum likelihood approach to single-channel source separation,” *J. Mach. Learn. Res.*, vol. 4, nos. 7–8, pp. 1365–1392, 2004.

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

[17] K. E. Hild, 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.

[18] M. N. Schmidt and M. Morup, “Nonnegative matrix factor 2-D deconvolution for blind single channel source separation,” in *Proc. 6th Int. Conf. Independ. Compon. Anal. Blind Signal Separat.*, vol. 3889. Mar. 2006, pp. 700–707.

[19] 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.

[20] 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, May 2011.

[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] 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.

[25] 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.

[26] 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.

[27] R. Balan, J. Rosca, S. Rickard, and J. O’Ruanaidh, “The influence of windowing on time delay estimates,” in *Proc. Conf. Inf. Sci. Syst.*, vol. 1. Mar. 2000, pp. 1–3.

[28] 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.

[29] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. Int. Conf. Music Inf. Retr.*, Oct. 2003, pp. 229–230.

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

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

[32] 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.

[33] 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.

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

[35] 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 Syst. I, Reg. Papers*, vol. 60, no. 3, pp. 662–675, Mar. 2013.

[36] 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, Sep. 2011.

N. Tengtrairat received the B.Eng. degree in computer engineering from Chiang Mai University, Chiang Mai, Thailand, and the M.Sc. degree in management information systems from Chulalongkorn University, Bangkok, Thailand. She is currently pursuing the Ph.D. degree with Newcastle University, Newcastle upon Tyne, U.K., in statistical single-channel blind source separation.

She has been a Lecturer with the Department of Computer Science, Payap University, Thailand. Her current research interests include speech and audio signal processing, speech enhancement, noise cancelling, and machine learning.

Bin Gao received the B.S. degree in communications and signal processing from Southwest Jiao Tong University, Chengdu, China, in 2005, the M.Sc. degree (Distinction) in communications and signal processing in 2007, and the Ph.D. degree from Newcastle University, Newcastle upon Tyne, U.K., in 2011.

He is currently a Research Associate with Newcastle University. His current research interests include audio and image processing, machine learning, structured probabilistic modeling on audio applications such as audio source separation, feature extraction and denoising, and single channel blind source separation.

W. L. Woo (M'11–SM'12) was born in Malaysia. He received the B.Eng. degree (First Class Hons.) in electrical and electronics engineering and the Ph.D. degree from Newcastle University, Newcastle upon Tyne, U.K.

He is currently a Senior Lecturer with the School of Electrical, Electronics and Computer Engineering, Newcastle University. His current research interests include the mathematical theory and algorithms for nonlinear signal and image processing. This includes areas of machine learning for signal processing, blind source separation, multidimensional signal processing, and signal/image deconvolution and restoration. He has an extensive portfolio of relevant research supported by a variety of funding agencies. He has authored over 250 papers on these topics on various journals and international conference proceedings.

Dr. Woo currently serves on the editorial board of several international signal processing journals. He actively participate in international conferences and workshops, and serves on their organizing and technical committees. In addition, he acts as a Consultant to a number of industrial companies that involve the use of statistical signal and image processing techniques. He is a member of the Institution Engineering Technology. He was the recipient of the IEE Prize and the British Scholarship in 1998 to continue his research work.

S. S. Dlay received the B.Sc. (Hons.) degree in electrical and electronic engineering and the Ph.D. degree in VLSI design from Newcastle University, Newcastle upon Tyne, U.K.

He was a Post Doctoral Research Associate at Newcastle University in 1984 and helped to establish an Integrated Circuit Design Centre, funded by the EPSRC. In 1984, he was a Lecturer with the Department of Electronic Systems Engineering, University of Essex, Essex, U.K. In 1986, he re-joined the Newcastle University as a Lecturer with the School of Electrical, Electronic and Computer Engineering, then he was promoted to Senior Lecturer in 2001. In recognition of his major achievements, he has been appointed to a Personal Chair in Signal Processing Analysis. He has authored over 250 research papers ranging from biometrics and security, biomedical signal processing, and implementation of signal processing architectures.

Prof. Dlay serves on many editorial boards and has played an active role in numerous international conferences in terms of serving on technical and advisory committees as well as organizing special sessions. He is a College Member of the Engineering and Physical Science Research Council (EPSRC). He was the recipient of a Scholarship from the EPSRC and the Charles Hertzmann Award.