Adaptive Processing and Learning for Audio Source Separation

Jen-Tzung Chien*, Hiroshi Sawada† and Shoji Makino‡
* Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan
† NTT Service Evolution Laboratories, NTT Corporation, Yokosuka-shi, Kanagawa, Japan
‡ Graduate School of Systems and Information Engineering, University of Tsukuba, Ibaraki, Japan
jtchien@nctu.edu.tw, sawada.hiroshi@lab.ntt.co.jp, maki@tara.tsukuba.ac.jp

***Abstract*— This paper overviews a series of recent advances in adaptive processing and learning for audio source separation. In real world, speech and audio signal mixtures are observed in reverberant environments. Sources are usually more than mixtures. The mixing condition is occasionally changed due to the moving sources or when the sources are changed or abruptly present or absent. In this survey article, we investigate different issues in audio source separation including overdetermined/underdetermined problems, permutation alignment, convolutive mixtures, contrast functions, nonstationary conditions and system robustness. We provide a systematic and comprehensive view for these issues and address new approaches to overdetermined/underdetermined convolutive separation, sparse learning, nonnegative matrix factorization, information-theoretic learning, online learning and Bayesian approaches.**

I. INTRODUCTION

We are surrounded by sounds and noises in presence of room reverberation. The observed mixtures are usually less than source signals. The mixing condition is prone to be varied by the moving sources or in case of source replacement. It becomes challenging to estimate the desired audio and speech signals and develop a comfortable acoustic communication channel between humans and machines. Audio source separation in realistic conditions has been a fascinating avenue for research which is crucial for broad extensions and applications ranging from speech enhancement, speech recognition, music retrieval, sound classification, human-machine communication and many others. How to extract and separate a target audio or speech signal from noisy and nonstationary observations is now impacting the communities of signal processing and machine learning.

The traditional blind source separation (BSS) approaches based on independent component analysis (ICA) were designed to resolve the instantaneous mixtures by optimizing a contrast function or an independence measure. In previous BSS methods, the frequency characteristics and location of each sources and how these sources were mixed were not sophisticatedly investigated. Solving the instantaneous mixtures did not truly reflect the real reverberant environment which structurally mixed the sources as the convolutive mixtures [25][31]. The underdetermined problem in presence of more sources than sensors may not have been carefully treated [30]. The contrast functions may not flexibly and honestly measure the independence for an optimization with convergence [7]. The static mixing system could not catch the underlying dynamics in source signals and sensor networks. The uncertainty of system parameters may not be precisely characterized so that the robustness against adverse conditions was not guaranteed [5][9][15].

Generally, signal processing and machine learning provide fundamental knowledge and algorithm to resolve different issues in audio source separation. In the past years, there have been a remarkable progress on development of cutting-edge adaptive processing and learning algorithms for source separation and its applications. Machine learning has been one of the most rapidly growing areas in international conferences and journals. The goal of this paper is to overview a series of recent advances in adaptive processing and learning algorithms for BSS in presence of speech and music signals. In this paper, we address general issues in an audio source separation system including permutation problem, overdetermined/underdetermined convolutive mixtures, optimization of contrast function, nonstationary mixing condition and model regularization. We provide a comprehensive and unified view for these issues and present a systematic survey over the recent important solutions to overdetermined/underdetermined convolutive separation [14][24][26], sparse source separation [3], reverberant source separation [31], nonnegative matrix factorization (NMF) [8][12][27], information-theoretic learning [6][9], online learning [9] and Bayesian inference [4][18][19][28]. We address how these algorithms are connected and why they work for source separation particularly in speech and music applications.

The remaining of this paper is organized into three sections. Section 2 addresses key issues in BSS system and discusses some challenging issues in real-world audio source separation. The competing solutions to these components and issues are generally categorized into two parts: the *front-end processing* and the *back-end learning*, which are detailed in Second 3 and Section 4, respectively. Section 5 shall summarize this overview study and point out future directions on adaptive processing and learning for audio source separation and applications.

II. CHALLENGES IN AUDIO SOURCE SEPARATION

Audio source separation is known as a challenging research topic and has had a big progress in recent years.

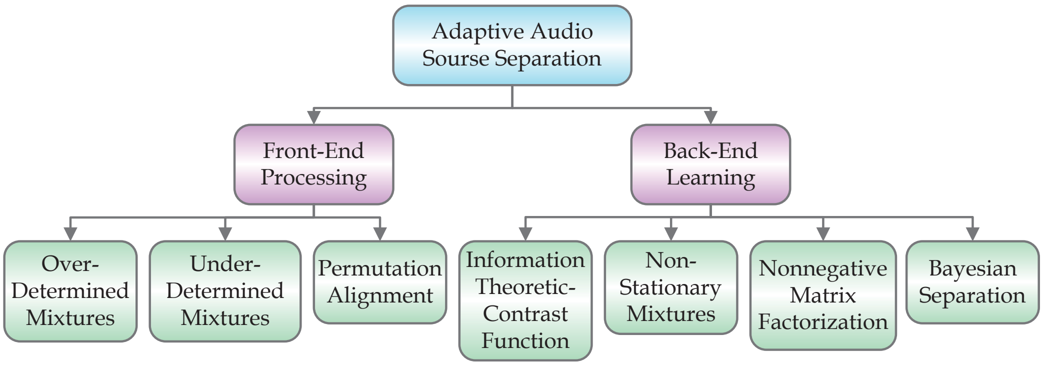
Adaptive Audio Sourse Separation
Front-End Processing
Back-End Learning
Over-Determined Mixtures
Under-Determined Mixtures
Permutation Alignment
Information Theoretic-Contrast Function
Non-Stationary Mixtures
Nonnegative Matrix Factorization
Bayesian Separation

Fig. 1. Some issues in adaptive audio source separation.

In particular, the challenges in speech separation and recognition have been impacting the development of practical automatic speech recognition systems and attracting the attention of many researchers in different communities, e.g. machine learning, audio signal processing and spoken language processing as illustrated by the recent CHiME challenge (http://spandh.dcs.shef.ac.uk/chime_challenge/).

In this study, we review some advances in adaptive audio source separation which are generally classified into front-end processing and back-end learning as shown in Figure 1. A number of adaptive processing and learning algorithms will be introduced to deal with different issues in audio source separation. In front-end processing, we highlight on the adaptive signal processing to analyse the information on each source, such as its frequency characteristics and location, or identifying how the sources are mixed. The input signals are processed through several processing components to obtain the separated signals. In this section, we review a number of high-impacting works on frequency-domain audio source separation which could align the permutation ambiguities [26], separate the convolutive mixtures [28], identify the number of sources [3], resolve the overdetermined/underdetermined problem [2][30] and compensate for the room reverberation [31].

The back-end learning is devoted to recover the source signals by using only the information about their mixtures observed in each microphone without possessing frequency and location information on each source. We build a statistical model for the whole system and infer the model by using the mixtures. Machine learning algorithms are developed for audio source separation. The model-based speech separation and recognition could be established [20]. In this section, we present the estimation of demixing parameters through construction and optimization of information-theoretic contrast function [6][7]. The solutions to music source separation based on NMF [12][27] and sparse learning [8] are introduced. Next, we focus on the uncertainty modeling for the regularized signal separation in accordance with Bayesian perspective [18]. The nonstationary and temporally-correlated audio source separation [9] is presented.

III. Front-End Processing

Considering the issue of unknown number of sources, a Gaussian mixture model with Dirichlet prior for mixture weight parameter was proposed to identify the direction-of-arrival (DOA) of source speech signal from individual time-frequency units. This model was applied to estimate the number of sources and deal with the sparse source separation [3][4]. Moreover, it is popular to apply linear filtering and spectrum enhancement methods for reverberant speech processing which is applied for speech recognition in reverberant environment [31]. In this section, we mainly focus on the solutions to frequency-domain blind source separation which are applied to resolve the overdetermined/underdetermined problems and the permutation alignment.

A. *Overdetermined System*

When the number N of microphones is enough for the number M of sources (in a determined $M = N$ or overdetermined $M < N$ case), we employ the complex-valued ICA to separate the frequency bin-wise mixtures. Let $\mathbf{x}_{ft} = [x_{ft1}, \ldots, x_{ftN}]^{\mathrm{T}} \in \mathbb{C}^N$ be N dimensional complex vector that represents N microphones' observations at frequency f and time t. ICA obtains the separated signals $\hat{\mathbf{s}}_{ft} = [\hat{s}_{ft1}, \ldots, \hat{s}_{ftM}]^{\mathrm{T}} \in \mathbb{C}^M$ by a linear transformation

$$\hat{\mathbf{s}}_{ft} = \mathbf{W}_f \mathbf{x}_{ft}\,, \quad t = 1, \ldots, T \tag{1}$$

with a frequency-dependent separation matrix $\mathbf{W}_f$ of size $M \times N$. The matrix $\mathbf{W}_f$ is optimized so that the distribution of the vector elements $\hat{s}_{ftm}$, $t = 1, \ldots, T$ is far from a Gaussian distribution. Various optimization learning rules have been proposed, for example FastICA [17] or one based on natural gradient [1][10]. In the learning rule, a non-linear function is utilized to evaluate how similar or different the distribution of each element $\hat{s}_{ftm}$ is compared with the Gaussian distribution. For a complex-valued variable obtained as a result of short-time Fourier transform (STFT), a polar coordinate based non-linear function [21] is effective.

There is scaling ambiguity in an ICA solution. For an audio source separation task, the scaling ambiguity is resolved by trying to represent the observed signals at microphones with scaled separated signals. For this purpose, we calculate the inverse matrix $\mathbf{A}_f = [\mathbf{a}_{f1}, \ldots, \mathbf{a}_{fM}] = \mathbf{W}_f^{-1}$ or the Moore-Penrose pseudo inverse matrix $\mathbf{A}_f = [\mathbf{a}_{f1}, \ldots, \mathbf{a}_{fM}] = \mathbf{W}_f^{+}$ of the separated matrix. By multiplying $\mathbf{A}_f$ on both sides of (1), we have

$$\mathbf{A}_f \hat{\mathbf{s}}_{ft} = \sum_{m=1}^{M} \mathbf{a}_{fm} \hat{s}_{ftm} = \mathbf{x}_{ft} \,. \tag{2}$$

Then we obtain a vector $\breve{\mathbf{s}}_{ft}^{(m)}$ of the scaled (ambiguity-resolved) separated signals as

$$\breve{\mathbf{s}}_{ft}^{(m)} = \mathbf{a}_{fm} \hat{s}_{ftm} \,. \tag{3}$$

B. Underdetermined System

When the number N of microphones is insufficient for the number M of sources (in an underdetermined $M > N$ case), we typically employ the method based on time-frequency masking, where we need to estimate which source has the largest amplitude for each time frequency slot (f, t). For that purpose, we apply a clustering method to observation vectors $\mathbf{x}_{ft}$ and to calculate the posterior probability $p(C_m|\mathbf{x}_{ft})$ that a vector $\mathbf{x}_{ft}$ belongs to a cluster C_m. Then, time frequency masks are made by

$$\mathcal{M}_{ftm} = \begin{cases} 1 & p(C_m|\mathbf{x}_{ft}) \geq p(C_{m'}|\mathbf{x}_{ft}), \, \forall\, m' \neq m \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

and the separated signals are obtained by

$$\hat{\mathbf{s}}_{ft}^{(m)} = \mathcal{M}_{ftm} \mathbf{x}_{ft} \,. \tag{5}$$

In terms of clustering methods, one based on an anechoic propagation model [2][24] is easy and simple, and works well under low reverberant conditions. However, to cope with more complicated real-room sound propagation, frequency bin-wise clustering has been proposed [26]. Specifically, a Gaussian mixture model (GMM)

$$p(\mathbf{x}_{ft}|\boldsymbol{\Theta}) = \sum_{m=1}^{M} \pi_{fm}\, p(\mathbf{x}_{ft}|\mathbf{a}_{fm}, \sigma_{fm}^2) \tag{6}$$

with a complex Gaussian density function of the form [26]

$$\begin{aligned} p(\mathbf{x}|\mathbf{a}_{fm}, \sigma_{fm}^2) &= \frac{1}{(2\pi\sigma_{fm}^2)^{N-1}} \\ &\times \exp\left(-\frac{||\mathbf{x} - (\mathbf{a}_{fm}^{\mathrm{H}}\mathbf{x}) \cdot \mathbf{a}_{fm}||^2}{\sigma_{fm}^2} \right) \end{aligned} \tag{7}$$

is assumed for each frequency bin f, and we estimate a parameter set $\boldsymbol{\Theta} = \{\pi_{f1}, \mathbf{a}_{f1}, \sigma_{f1}^2, \ldots, \pi_{fM}, \mathbf{a}_{fM}, \sigma_{fM}^2\}$ that maximizes the likelihood $p(\mathbf{X}_f|\boldsymbol{\Theta}) = \prod_{t=1}^{T} p(\mathbf{x}_{ft}|\boldsymbol{\Theta})$. In (6), $\mathbf{a}_{fm}$ is the mean vector, σ_{fm}^2 is the variance, and π_{fm} is the mixture ratio of the m-th cluster. In (7), $(\mathbf{a}_{fm}^{\mathrm{H}}\mathbf{x}) \cdot \mathbf{a}_{fm}$ denotes the orthogonal projection of $\mathbf{x}$ onto the subspace spanned by $\mathbf{a}_{fm}$. After the parameter set is estimated, the posterior probabilities used in (4) is given by

$$p(C_m|\mathbf{x}_{ft}) = \frac{\pi_{fm}\, p(\mathbf{x}_{ft}|\mathbf{a}_{fm}, \sigma_{fm}^2)}{\sum_{m=1}^{M} \pi_{fm}\, p(\mathbf{x}_{ft}|\mathbf{a}_{fm}, \sigma_{fm}^2)} \,. \tag{8}$$

C. Permutation Alignment

The method based on ICA or GMM, described in previous Subsections, performs a source separation task in a frequency bin-wise manner. Therefore, we need to align the permutation ambiguity of the ICA or GMM results in each frequency bin so that a separated signal in the time domain contains frequency components from the same source signal. This problem is well known as the permutation problem of frequency-domain BSS [22]. Although various approaches to the permutation problem have been proposed [25], the following approach based on the dominance measures [23][26] performs very well.

When using ICA, we employ the power ratio

$$r_f^{(m)}(t) = \frac{||\breve{\mathbf{s}}_{ft}^{(m)}||^2}{\sum_{m=1}^{M} ||\breve{\mathbf{s}}_{ft}^{(m)}||^2} \tag{9}$$

of the scaled separated signals (3) as a dominance measure [23]. On the other hand, when using a GMM for time-frequency masking, we employ the posterior probability (8)

$$r_f^{(m)}(t) = p(C_m|\mathbf{x}_{ft}) \tag{10}$$

as a dominance measure [26]. After calculating the dominance measure, we basically interchange the indices m of the separated signals so that the correlation coefficient $\rho(r_f^{(m)}, r_{f'}^{(m)})$ between the dominance measures at different frequency bins f and f' is maximized for the same source. The optimization procedure is described in details in a reference [26].

IV. Back-End Learning

In this section, we focus on the *machine learning* solutions to audio source separation. We consider blind speech or music separation as a learning problem without special treatment on convolutive mixtures or extraction of frequency features and location information on each source signal. The information-theoretic learning, online learning, dictionary learning, sparse learning and Bayesian learning are investigated.

A. Information-Theoretic Contrast Function

Information-theoretic learning is important to find the demixing solution to audio source separation. Let the observation vector $\mathbf{x}_t = [x_{t1}, \ldots, x_{tN}]^{\mathrm{T}}$ from N microphones at time frame t be mixed by $\mathbf{x}_t = \mathbf{A}\mathbf{s}_t$ where $\mathbf{A}$ is an unknown $N \times M$ mixing matrix and $\mathbf{s}_t = [s_{t1}, \ldots, s_{tM}]^{\mathrm{T}}$ denotes a vector of M mutually-independent source signals. For the case of $N = M$, BSS problem is resolved by ICA method which optimizes a contrast function $\mathcal{D}(\mathbf{X}, \mathbf{W})$ measuring the independence or non-Gaussianity of the demixed signals $\hat{\mathbf{s}}_t$ based on a demixed matrix or separation matrix $\mathbf{W}$, i.e. $\hat{\mathbf{s}}_t = \mathbf{W}\mathbf{x}_t$. The demixing matrix can be estimated in accordance with the gradient descent algorithm or the natural gradient algorithm [1] from a set of audio signals $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$

$$\mathbf{W}^{(n+1)} = \mathbf{W}^{(n)} - \eta \frac{\partial \mathcal{D}(\mathbf{X}, \mathbf{W}^{(n)})}{\partial \mathbf{W}^{(n)}} \mathbf{W}^{(n)\mathrm{T}} \mathbf{W}^{(n)} \tag{11}$$

where n is the iteration index and η is the learning rate. The scaled natural gradient algorithm was proposed to improve learning process by imposing *a posteriori* scalar gradient constraint [13]. The metrics of likelihood function, negentropy and kurtosis are popular to serve as ICA contrast functions. More meaningfully, the information-theoretical contrast function is adopted to measure the independence between the demixed signals.

The *statistical hypothesis test* was recently proposed to carry out an information measure of confidence towards independence by investigating the null hypothesis $\mathcal{H}_0$ where the demixed signals $\hat{\mathbf{S}} = \{\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_T\}$ are independent against the alternative hypothesis $\mathcal{H}_1$ where the demixed signals are dependent [6]. The contrast function was formed as a log likelihood ratio given by

$$\mathcal{D}_{\text{NLR}}(\mathbf{X}, \mathbf{W}) = \log p(\hat{\mathbf{S}}|\mathcal{H}_0) - \log p(\hat{\mathbf{S}}|\mathcal{H}_1). \tag{12}$$

However, the parametric Gaussian distribution is not allowed to represent demixed signals in $p(\hat{\mathbf{S}}|\mathcal{H})$ by using ICA method. The *nonparametric distribution* based on Parzen window density function was applied to develop the nonparametric likelihood ratio (NLR) contrast function for speech separation. This method was extended to unsupervised learning of acoustic hidden Markov models for speech recognition. More generally, the measure of independence is calculated as a divergence between the joint distribution of the demixed signals and the product of marginal distributions of individual demixed signals. This divergence measure equals to zero in case that the condition of independence is met. A general convex divergence measure was derived by substituting a general convex function $f(t) = \frac{4}{1-\alpha^2}\left[\frac{1-\alpha}{2} + \frac{1+\alpha}{2}t - t^{(1+\alpha)/2}\right]$ into Jensen's inequality to construct a contrast function for ICA optimization. This convex divergence $\mathcal{D}_{\text{C}}(\mathbf{X}, \mathbf{W}, \alpha)$ is written by [7]

$$\frac{2}{1-\alpha^2}\sum_{t=1}^{T}\left\{ 2\left[\frac{1}{2}\left(p(\mathbf{W}\mathbf{x}_t) + \prod_{m=1}^{M} p(\mathbf{w}_m\mathbf{x}_t)\right)\right]^{(1+\alpha)/2} - \left[p(\mathbf{W}\mathbf{x}_t)^{(1+\alpha)/2)} + \left(\prod_{m=1}^{M} p(\mathbf{w}_m\mathbf{x}_t)\right)^{(1+\alpha)/2}\right]\right\} \tag{13}$$

where $\mathbf{W} = [\mathbf{w}_1^{\text{T}}, \ldots, \mathbf{w}_M^{\text{T}}]^{\text{T}}$ and α is an adjustable convexity parameter. By substituting the nonparametric distribution $p(\cdot)$ and adjusting the convexity parameter α, we estimated different nonparametric solutions to the demixing matrices $\hat{\mathbf{W}}$ for audio source separation. In cases of $\alpha = 1$ and $\alpha = -1$, the general convex divergence is realized to the *convex-Shannon* divergence and the *convex-logarithm* divergence where the convex functions based on Shannon's entropy and negative logarithm are adopted, respectively. The convergence in optimizing convex divergence for ICA was improved by choosing the convexity parameter with sharp learning curve [7]. Information-theoretic learning is illustrated to work for adaptive source separation.

B. Nonnegative Matrix Factorization

The information-theoretic learning is further extended to dictionary learning based on the nonnegative matrix factorization (NMF) which is recently hot issue in audio source separation [11]. NMF attempts to decompose the nonnegative mixed samples $\mathbf{X} \in \mathbb{R}^{N \times T}$ into a product of nonnegative mixing matrix $\mathbf{A} \in \mathbb{R}^{N \times M}$ and nonnegative source signals $\mathbf{S} \in \mathbb{R}^{M \times T}$ by minimizing a divergence measure $\mathcal{D}(\mathbf{X}, \mathbf{A}, \mathbf{S})$ between $\mathbf{X}$ and $\mathbf{AS}$. NMF is a parts-based representation which only allows additive combination and can be directly applied to decompose the nonnegative mixed audio signals. The absolute values of short-time Fourier transform (STFT) are calculated to form $\mathbf{X}$. The standard NMF is fulfilled according to a regularized least square (RLS) criterion $\mathcal{D}_{\text{RLS}}(\mathbf{X}, \mathbf{A}, \mathbf{S})$ with sparsity constraint

$$\sum_{n,t} |X_{nt} - [\mathbf{AS}]_{nt}|^2 + \gamma_a \sum_{n,m} f(A_{nm}) + \gamma_s \sum_{m,t} f(S_{mt}) \tag{14}$$

where $\gamma_a \geq 0$ and $\gamma_s \geq 0$ denote the regularization parameters. The sparseness measure $f(A_{nm}) = |A_{nm}|$, $f(A_{nm}) = A_{nm}$ or $f(A_{nm}) = A_{nm}\log(A_{nm})$ could be imposed in (14).

More recently, the α divergence [12], convex divergence [7] and Itakura-Saito (IS) divergence [27] were treated as objective function to derive solution to NMF. For example, IS divergence is written by $\mathcal{D}_{\text{IS}}(\mathbf{X}, \mathbf{A}, \mathbf{S}) = \sum_{n,t}(\frac{X_{nt}}{[\mathbf{AS}]_{nt}} - \log\frac{X_{nt}}{[\mathbf{AS}]_{nt}} - 1)$ which depends only on the ratio $\frac{X_{nt}}{[\mathbf{AS}]_{nt}}$. This property is favorable for analyzing music and speech signals where low frequency components have much higher energy than high frequency components. In [16][27], minimizing IS divergence was shown to be equivalent to maximizing the log-likelihood $\log p(\tilde{\mathbf{X}}|\mathbf{A}, \mathbf{S})$ based on the multivariate complex-valued Gaussian distributions where $\tilde{\mathbf{X}}$ denotes a matrix of STFT complex-valued coefficients. The multiplicative updating rule for NMF was obtained by

$$A_{nm} \leftarrow A_{nm}\sqrt{\frac{\sum_t \frac{X_{nt}S_{mt}}{([\mathbf{AS}]_{nt})^2}}{\sum_t \frac{S_{mt}}{[\mathbf{AS}]_{nt}}}},\ S_{mt} \leftarrow S_{mt}\sqrt{\frac{\sum_n \frac{X_{nt}A_{nm}}{([\mathbf{AS}]_{nt})^2}}{\sum_n \frac{A_{nm}}{[\mathbf{AS}]_{nt}}}}. \tag{15}$$

However, this solution is only designed for the case of an M-dimensional observation frame $\mathbf{x}_t$ at a single frequency bin f. The multichannel time-frequency NMF was developed for music source separation where an M-dimensional mixed signals $\mathbf{x}_{ft}$ was extended to different frequency bins $\{(f, t), 1 \leq f \leq F\}$. A clustering procedure of NMF bases was performed. The separate bases were used to recover individual source signals.

In [8], Bayesian NMF was proposed for monaural music source separation which decomposed a single-channel mixed signal $\mathbf{X}$ into a rhythmic signal $\mathbf{X}_r$ and a harmonic signal $\mathbf{X}_h$. Let the nonnegative monaural matrix $\mathbf{X} \in \mathbb{R}^{F \times T}$ in time-frequency domain be chunked into L segments $\{\mathbf{X}^{(l)}\}$. Each segment is represented by $\mathbf{X}^{(l)} = \mathbf{X}_r^{(l)} + \mathbf{X}_h^{(l)} = \mathbf{A}_r\mathbf{S}_r^{(l)} + \mathbf{A}_h^{(l)}\mathbf{S}_h^{(l)}$ where $\{\mathbf{S}_r^{(l)}, \mathbf{S}_h^{(l)}\}$ are two groups of segment-dependent encoding coefficients, $\mathbf{A}_h^{(l)}$ denotes the bases for harmonic source which are individual for different segments

l, and $\mathbf{A}_r$ denotes the bases for rhythmic source which are shared across segments. Assuming the basis components are Gamma distributed and the encoding coefficients $\mathbf{S}^{(l)} = \{\mathbf{S}_r^{(l)}, \mathbf{S}_h^{(l)}\}$ are Laplacian distributed by $p([\mathbf{S}^{(l)}]|\lambda^{(l)}) = \frac{1}{2}\lambda^{(l)} \exp\{-\lambda^{(l)}[\mathbf{S}^{(l)}]\}$ with hyperparameter $\lambda^{(l)}$, *Bayesian group sparse learning* for NMF was performed to resolve the underdetermined problem in audio source separation through a Gibbs sampling procedure over different parameters and hyperparameters of Gamma distributions and Laplacian distributions. BSS was viewed as a probabilistic framework with latent variables including source signals, mixing coefficients, modeling errors and their associated parameters.

C. Bayesian Learning for Nonstationary Source Separation

In real-world audio source separation, we face the challenges of changing sources or moving speakers, namely the source signals may abruptly appear or disappear, the speakers may be replaced by new ones or even moving from one location to the other. The mixing conditions and source signals are accordingly nonstationary and should be traced to assure robustness in nonstationary source separation [9]. A meaningful approach to deal with the robustness issue in audio source separation is constructed from *Bayesian perspective* [5][18]. The Bayesian approaches to blind speech separation [9][18] are examined. Some prior information is introduced for uncertainty modeling and knowledge integration. Model regularization and system robustness could be improved. Let $\mathbf{X}^{(l)} = \{\mathbf{x}_t^{(l)}\}$ denote a set of mixed signals at segment l. The signals are mixed by a linear combination of M unknown source signals $\mathbf{S}^{(l)} = \{\mathbf{s}_t^{(l)}\}$ using a mixing matrix $\mathbf{A}^{(l)}$, i.e. considering a noisy ICA model $\mathbf{x}_t^{(l)} = \mathbf{A}^{(l)}\mathbf{s}_t^{(l)} + \boldsymbol{\varepsilon}_t^{(l)}$ where $\mathbf{E}^{(l)} = \{\boldsymbol{\varepsilon}_t^{(l)}\}$ denotes the noise signals. We assume that $\mathbf{A}^{(l)}$ and $\mathbf{S}^{(l)}$ are unchanged within a segment l but varied across segments. To tackle the nonstationary source separation, we attempt to incrementally characterize the variations of $\mathbf{A}^{(l)}$ and $\mathbf{S}^{(l)}$ from the observed segments $\mathcal{X}^{(l)} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(l)}\}$. *Online learning* is conducted to compensate for nonstationary conditions of mixing coefficients and source signals segment by segment. At each learning epoch l, we first accumulate to calculate the predictive distribution of current ICA parameters $\boldsymbol{\Theta}^{(l)} = \{\mathbf{A}^{(l)}, \mathbf{S}^{(l)}, \mathbf{E}^{(l)}\}$ based on previous segment data $\mathcal{X}^{(l-1)}$. When new data $\mathbf{X}^{(l)} = \{\mathbf{x}_t^{(l)}\}$ of segment l is enrolled, the posterior distribution is corrected by $p(\boldsymbol{\Theta}^{(l)}|\mathcal{X}^{(l)}) \propto p(\mathbf{X}^{(l)}|\boldsymbol{\Theta}^{(l)})p(\boldsymbol{\Theta}^{(l)}|\mathcal{X}^{(l-1)})$ which is proportional to the product of a likelihood function of current data segment $\mathbf{X}$ and *a posteriori* distribution given the previous data segments $\mathcal{X}^{(l-1)}$. In this sense, the modes of posterior distributions are updated in an online manner $\boldsymbol{\Theta}^{(0)} \rightarrow \boldsymbol{\Theta}^{(1)} \rightarrow \cdots \rightarrow \boldsymbol{\Theta}^{(l)}$.

In particular, source signals are represented by a mixture of Gaussian distributions $p(\mathbf{s}_t^{(l)}|\boldsymbol{\Theta}^{(l)})$ given by

$$\prod_{m=1}^{M} \left[\sum_{k=1}^{K} \pi_{mk}^{(l)} \mathcal{N}(s_{m,t}^{(l)}|\mu_{mk}^{(l)}, (\gamma_{mk}^{(l)})^{-1}) \right] \tag{16}$$

with parameters $\boldsymbol{\Theta}^{(l)}$ consisting of mixture weights $\boldsymbol{\Pi}^{(l)} = \{\pi_{mk}^{(l)}\}$, means $\mathbf{M}^{(l)} = \{\mu_{mk}^{(l)}\}$ and precisions $\mathbf{R}^{(l)} = \{\gamma_{mk}^{(l)}\}$. The noise vector $\boldsymbol{\varepsilon}_t^{(l)}$ is assumed to be Gaussian $\mathcal{N}(\varepsilon_t^{(l)}|0, (\mathbf{B}^{(l)})^{-1})$ with zero mean and diagonal precision matrix $\mathbf{B}^{(l)} = \text{diag}\{\beta_n^{(l)}\}$. The prior density of $N \times M$ mixing matrix $\mathbf{A}^{(l)} = \{a_{nm}^{(l)}\}$ is distributed by $p(\mathbf{A}^{(l)}|\boldsymbol{\alpha}^{(l)}) = \prod_{n=1}^{N} \left[\prod_{m=1}^{M} \mathcal{N}(a_{nm}^{(l)}|0, (\alpha_m^{(l)})^{-1}) \right]$. Importantly, the hyperparameter $\alpha_m^{(l)}$ is known as an automatic relevance determination (ARD) [29], which reveals the activity of a source signal $s_{m,t}^{(l)}$. The matrix $\mathbf{A}^{(l)}$ is prone to be sparse with near zero entries at the mth column of $\mathbf{A}^{(l)}$ if the estimated ARD $\alpha_m^{(l)}$ is large. The mth source is likely inactive at segment l. The redundant sources are disregarded automatically.

To fulfill *full Bayesian* approach to audio source separation, we consider the uncertainties of parameters and hyperparameters $\boldsymbol{\Theta}^{(l)} = \{\mathbf{A}^{(l)}, \mathbf{S}^{(l)}, \mathbf{E}^{(l)}, \boldsymbol{\Pi}^{(l)}, \mathbf{M}^{(l)}, \mathbf{R}^{(l)}, \mathbf{B}^{(l)}, \boldsymbol{\alpha}^{(l)}\}$ which are latent variables and compensate these uncertainties by using the solution obtained by optimizing the marginal likelihood over all latent variables. *Conjugate priors* $p(\boldsymbol{\Theta}^{(l)}|\boldsymbol{\Phi}^{(l-1)})$ are introduced to characterize the uncertainties of $\{\boldsymbol{\Pi}^{(l)}, \mathbf{M}^{(l)}, \mathbf{R}^{(l)}, \mathbf{B}^{(l)}, \boldsymbol{\alpha}^{(l)}\}$. The hyperparameters $\boldsymbol{\Phi}^{(l-1)}$ are continuously updated from history data $\mathcal{X}^{(l-1)}$. With these conjugate priors, the integral in marginal likelihood has closed-form solution and is calculated efficiently. This nonstationary Bayesian method was formulated based on the variational Bayesian expectation maximization (VB-EM) algorithm [9] where the variational inference was solved by maximizing the lower bound of marginal likelihood.

More recently, we present the solution to nonstationary and *temporally correlated* source separation where the mixing condition is changed continuously and the temporal correlation in time-series signals, e.g. mixing coefficients and source signals, is taken into account. Online learning and *Gaussian process* (GP) [19] are merged into a separation system which compensates for the nonstationary and temporally correlated mixing environments and source signals, respectively. Considering a noisy ICA model with time-varying mixing matrix $\mathbf{A}_t^{(l)}$ at segment l and time bin t, the temporally correlated mixing coefficients are generated by the distributions of nonparametric latent functions. GP flexibly explores the unknown temporal structure of $a_{nm,t}^{(l)}$. A latent function $f(\cdot)$ is employed to connect the relation between current coefficient $a_{nm,t}^{(l)}$ and its past p coefficients $\vec{\mathbf{a}}_{nm,t-1}^{(l)} = [a_{nm,t-1}^{(l)}, \ldots, a_{nm,t-p}^{(l)}]^T$ by

$$a_{nm,t}^{(l)} = f(\vec{\mathbf{a}}_{nm,t-1}^{(l)}) + \varepsilon_{nm,t}^{(l)} \tag{17}$$

where $\varepsilon_{nm,t}^{(l)}$ denotes the white noise. This function is generated from a zero-mean Gaussian with a variance $\kappa(\vec{\mathbf{a}}_{nm,t-1}^{(l)}, \vec{\mathbf{a}}_{nm,\tau-1}^{(l)})$ given by $\xi_{a_{nm}}^{(l-1)} \exp\left[-\frac{\lambda_{a_{nm}}^{(l-1)}}{2} \left\| \vec{\mathbf{a}}_{nm,t-1}^{(l)} - \vec{\mathbf{a}}_{nm,\tau-1}^{(l)} \right\|^2 \right]$ which is an exponential-quadratic kernel function with hyperparameters $\{\lambda_{a_{nm}}^{(l-1)}, \xi_{a_{nm}}^{(l-1)}\}$. The GP prior $p(\mathbf{a}_{nm}^{(l)}|\boldsymbol{\mu}_{a_{nm}}^{(l-1)}, \mathbf{R}_{a_{nm}}^{(l-1)})$ for the mixing coefficients $\mathbf{a}_{nm}^{(l)} = [a_{nm,1}^{(l)}, \ldots, a_{nm,L}^{(l)}]^T$ written by $\mathcal{N}(\mathbf{a}_{nm}^{(l)}|0, \mathbf{K}_{a_{nm}}^{(l-1)})$

where $[\mathbf{K}_{a_{nm}}^{(l-1)}]_{t\tau} = \kappa(\vec{\mathbf{a}}_{nm,t-1}^{(l)}, \vec{\mathbf{a}}_{nm,\tau-1}^{(l)}) + \delta_{t\tau}$. With GP prior and its hyperparameters, VB-EM algorithm is again applied for model inference of the proposed online GP. The variational and sequential Bayesian inference is presented to implement a dynamic audio source separation system [9].

V. Conclusions

We have presented a series of adaptive methods which were developed for different issues in audio source separation. These methods were systematically categories into front-end processing and back-end learning. In front-end processing, we addressed high-performance solutions to overdetermined and underdetermined problems which are based on the processing of complex-valued time-frequency signals and the noise-masking method using Gaussian mixture model. The permutation problem was solved according to the correlation coefficient between dominance measures at different frequency bins. In back-end learning, we addressed the importance of information-theoretical learning for ICA optimization. The recent methods of sparse learning and dictionary learning based on nonnegative matrix factorization were presented for speech/music source separation. The online learning and Bayesian learning designed for nonstationary source separation were also presented for improving the robustness for audio source separation. In the future, the directions of single-channel and multi-channel dereverberation are still impacting the communities of signal processing and machine learning. The application of audio source separation in robust speech recognition is continuing to be attractive for future studies.

References

[1] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, pp. 251-276, 1998.

[2] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, pp. 1833–1847, 2007.

[3] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 33–36.

[4] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem," in *Proc. of Independent Component Analysis and Blind Signal Separation*, 2009, pp. 742–750.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[6] J.-T. Chien and B.-C. Chen, "A new independent component analysis for speech recognition and separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1245–1254, 2006.

[7] J.-T. Chien and H.-L. Hsieh, "Convex divergence ICA for blind source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 290–301, 2012.

[8] J.-T. Chien and H.-L. Hsieh, "Bayesian group sparse learning for nonnegative matrix factorization," in *Proc. of INTERSPEECH*, 2012, pp. 1552–1555.

[9] J.-T. Chien and H.-L. Hsieh, "Nonstationary source separation using sequential and variational Bayesian learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 5, pp. 681–694, 2013.

[10] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*, John Wiley & Sons, 2002.

[11] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation", in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006, pp. 621-624.

[12] A. Cichocki, H. Lee, Y.-D. Kim, and S. Choi, "Non-negative matrix factorization with α-divergence," *Pattern Recognition Letters*, vol. 29, pp. 1433–1440, 2008.

[13] S. C. Douglas and M. Gupta, "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation", in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007, pp. 637-640.

[14] S. Douglas, M. Gupta, H. Sawada, and S. Makino, "Spatio-temporal FastICA algorithms for the blind separation of convolutive mixtures," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1511–1520, 2007.

[15] C. Fevotte, "Bayesian audio source separation," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. Springer, 2007, pp. 305–335.

[16] C. Fevotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[17] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[18] A. Mohammad-Djafari and K. Knuth, "Bayesian approaches," in *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, P. Common and C. Jutten, Eds. Elsevier, 2010, pp. 467–513.

[19] S. Park and S. Choi, "Gaussian processes for source separation," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 1909–1912.

[20] S. J. Rennie, J. R. Hershey, and P. A. Olsen,"Single-channel multitalker speech recognition," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 66–80, 2010.

[21] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency domain blind source separation," *IEICE Trans. Fundamentals*, vol. E86-A, pp. 590–596, 2003.

[22] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 530–538, 2004.

[23] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *Proc. ISCAS*, pp. 3247–3250, 2007.

[24] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

[25] H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. Springer, 2007, pp. 47–78.

[26] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[27] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[28] H. K. Solvang, Y. Nagahara, S. Araki, H. Sawada, and S. Makino, "Frequency-domain Pearson distribution approach for independent component analysis (FD-Pearson-ICA) in blind source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 639–649, 2009.

[29] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine", *Journal of Machine Learning Research*, vol. 1, pp. 211-244, 2001.

[30] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–12, 2007.

[31] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms - robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.