Some Advances in Adaptive 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

I. Introduction

We are surrounded by sounds and noises in presence of room reverberation [15]. The observed mixed signals 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 based on the measure of independence or non-Gaussianity. In previous BSS methods, the frequency characteristics and location of individual 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 [11]. The underdetermined problem in presence of more sources than sensors may not have been carefully treated [14]. The contrast functions may not flexibly and honestly measure the independence for an optimization with convergence [3]. 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].

Generally, signal processing and machine learning provide fundamental knowledge and algorithm to resolve different issues in audio source separation. The goal of this article is to overview a series of recent advances in adaptive processing and learning algorithms for BSS in presence of speech and music signals. We survey the recent solutions to overdetermined/underdetermined convolutive separation [12], sparse source separation [1], nonnegative matrix factorization (NMF) [4][13], information-theoretic learning [2][5], online learning [5] and Bayesian inference.

In general, these algorithms are classified into front-end processing and back-end learning as shown in Figure 1 [6]. 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. We review the works on frequency-domain audio source separation which could align the permutation ambiguities [12], separate the convolutive mixtures, identify the number of sources [1], resolve the overdetermined/underdetermined problem [14]. The back-end learning is devoted to recover the source signals by using only the information about their mixtures observed in each microphone without frequency and location information on each source. We build a statistical model and infer the model by using the mixtures. We introduce the estimation of demixing parameters through construction and optimization of information-theoretic contrast function [2][3]. The solutions to music source separation based on NMF [13] and sparse learning [4] are addressed. Next, we focus on the uncertainty modeling for the regularized signal separation in accordance with Bayesian perspective. The nonstationary and temporally-correlated source separation [5] is presented.

II. 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 [1].

For the determined or the overdetermined problem, the number of microphones is enough for the number of sources. The complex-valued ICA was proposed to separate the frequency bin-wise mixtures. For each frequency bin, the ICA demixing matrix is optimized so that the distribution of the demixed elements is far from a Gaussian [11].

There is scaling ambiguity in an ICA solution. For an audio source separation task, the scaling ambiguity is resolved by representing the observed signals at microphones with the scaled separated signals [11].

In an underdetermined system, the number of microphones N is insufficient for the number of sources M, we typically employ the method based on time-frequency masking, where we need to estimate which source has the largest amplitude for

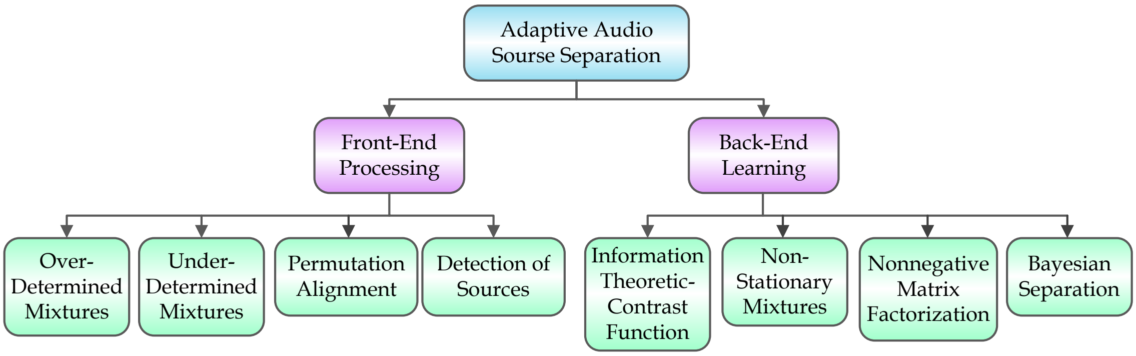

Fig. 1. Some issues in adaptive audio source separation.

each time frequency slot (f, t). For this purpose, we apply a clustering method to M-dimensional 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 or a source C_m. Then, the time frequency masks $\mathcal{M}_{ftm}$ are made and used to find the separated signals $\hat{\mathbf{s}}_{ft}^{(m)} = \mathcal{M}_{ftm}\mathbf{x}_{ft}$. The posterior probability $p(C_m|\mathbf{x}_{ft})$ is calculated by using a likelihood function $p(\mathbf{x}_{ft}|\mathbf{\Theta})$ based on a Gaussian mixture model (GMM) with parameters $\mathbf{\Theta}$ [12].

The method based on ICA or GMM 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 [9]. The *dominance measures* [10][12] performs very well for this problem. When using ICA, we employ the power ratio of the scaled separated signals as a dominance measure $r_f^{(m)}(t)$ [10]. On the other hand, when using a GMM for time-frequency masking, we employ the posterior probability $r_f^{(m)}(t) = p(C_m|\mathbf{x}_{ft})$ as a dominance measure [12]. 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.

III. 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. 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 from a set of audio signals $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$. 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 [2]. The contrast function was formed as a log likelihood ratio given by $\mathcal{D}(\mathbf{X}, \mathbf{W}) = \log p(\hat{\mathbf{S}}|\mathcal{H}_0) - \log p(\hat{\mathbf{S}}|\mathcal{H}_1)$. 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 $\mathcal{D}(\mathbf{X}, \mathbf{W}, \alpha)$ is developed with an adjustable convexity parameter. 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 dictionary learning based on the nonnegative matrix factorization (NMF) is recently hot issue in audio source separation [7]. 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 criterion with sparsity constraint.

More recently, the convex divergence [3] and Itakura-Saito (IS) divergence [13] were treated as the 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 $\frac{X_{nt}}{[\mathbf{AS}]_{nt}}$. In [8][13], 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.

In [4], 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 are Laplacian distributed, *Bayesian group sparse learning* for NMF was performed to resolve the underdetermined source separation through a Gibbs sampling procedure.

Further, 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 [5]. A meaningful approach to deal with the robustness issue in audio source separation is constructed from *Bayesian perspective*. Some prior information is introduced for uncertainty modeling and knowledge integration. 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. We also 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* are merged into a separation system which compensates for the nonstationary and temporally correlated mixing environments and source signals, respectively.

IV. Conclusions

We have presented a series of adaptive methods which were developed for different issues in BSS. 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 NMF 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.

References

[1] 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 ICASSP*, 2009, pp. 33–36.

[2] 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.

[3] 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.

[4] J.-T. Chien and H.-L. Hsieh, "Bayesian group sparse learning for music source separation," *EURASIP Journal on Audio, Speech, and Music Processing*, 5 July, 2013. (doi: 10.1186/1687-4722-2013-18)

[5] 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.

[6] J.-T. Chien, H. Sawada and S. Makino, "Adaptive processing and learning for audio source separation," in *Proc. of APSIPA ASC*, 2013.

[7] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation", in *Proc. of ICASSP*, 2006, pp. 621-624.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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, pp. 114-126, 2012.