PERMUTATION-FREE CLUSTERING OF RELATIVE TRANSFER FUNCTION FEATURES FOR BLIND SOURCE SEPARATION

Nobutaka Ito, Shoko Araki, and Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, "Keihanna Science City" Kyoto 619-0237 Japan

ABSTRACT

This paper describes an application of relative transfer functions (RTFs) to underdetermined blind source separation (BSS). A clustering-based BSS approach has the advantage that it can even deal with the underdetermined case, where the sources outnumber the microphones. Among others, clustering of a normalized observation vector (NOV) has proven effective for BSS even under reverberation. We here point out that the NOV gives information about RTFs of the dominant source, and hence call it the RTF features. Most of the previous BSS methods are limited in that they undergo significant performance degradation when the number of sources is not known precisely. This paper introduces our recently developed method for joint BSS and source counting based on permutation-free clustering of the RTF features. We demonstrate the effectiveness of the method in experiments with reverberant mixtures of an unknown number of sources with a reverberation time of up to 440 ms.

Index Terms— Blind source separation, source counting, relative transfer functions, clustering, permutation problem

1. INTRODUCTION

Over the last decade, the clustering-based BSS approach has been studied extensively [1–5], primarily because it can even deal with the underdetermined case. In this approach, we assume that each source signal is sparse enough in the time-frequency domain that at most one source signal is dominant in each time-frequency slot (disjointness) [1, 6]. Under this assumption, source location features (*e.g.*, the time difference of arrival (TDOA) or the direction of arrival (DOA)) form N clusters (N: the number of sources). Since each of the clusters corresponds to a source, we can realize source separation by collecting the time-frequency components in each cluster.

Among others, clustering of the normalized observation vector (NOV) based on fitting of a Watson mixture model (WMM) [5, 7] has proven effective for BSS even under reverberation. Since the NOV is differently distributed in different frequency bins, the clustering is performed in each frequency bin separately. After the clustering, there remains permutation ambiguity: it remains unknown which source each bin-wise cluster corresponds to. Therefore, to group together the bin-wise clusters corresponding to the same source, clustering of the activity sequences of the bin-wise clusters is performed. In this paper, we point out that the NOV gives information about relative transfer functions (RTFs) [8–10] of the dominant source, and hence call it the *RTF features*.

Most of the previous BSS methods including those proposed in [5, 7] require that the number of sources should be given. When it is not known precisely, such methods undergo significant performance degradation. Although there exist some methods for counting sources prior to BSS [11], it has been difficult to count sources under reverberation. This has limited the application area of BSS significantly, and therefore BSS for an unknown number of sources remains an important fundamental problem.

This paper introduces our recently developed method for joint BSS and source counting [12], which extends the previous methods [5, 7]. It is known that the frequency components of a speech signal tend to be activated synchronously, which is called the common amplitude modulation property. The property enables us to group together the frequency components activated synchronously as corresponding to the same source. We model the property by introducing time-variant, frequency-invariant mixture weights. By exploiting the property, our method can perform bin-wise clustering and, at the same time, group together the bin-wise clusters corresponding to the same source, which we call *permutation-free clustering*. Furthermore, our method can perform BSS even when N is unknown, since if we set the number of clusters, L, greater than N, the grouping results in N dominant clusters.

The method presented in this paper has been published partly in [12, 13]. Here, we clarify the relationship between the NOV and RTFs, and also present new experiments including evaluation of the source counting performance of our method.

The rest of this paper is organized as follows. In Section 2, we formulate the BSS problem we deal with in this paper. In Section 3, we introduce our method for the permutation-free clustering of the RTF features. In Section 4, we present experimental results to verify our method, and in Section 5 we conclude.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

2. PROBLEM FORMULATION

Suppose that we observe N concurrent speech signals using M microphones, where N is unknown. In the short-time Fourier transform (STFT) domain, the observed signals

$$\boldsymbol{y}_{tf} \triangleq \begin{bmatrix} y_{tf}^{(1)} & \cdots & y_{tf}^{(M)} \end{bmatrix}^{\mathsf{T}} \tag{1}$$

are modeled as follows:

$$\boldsymbol{y}_{tf} = \sum_{n=1}^{N} \tilde{s}_{tf}^{(n)} \tilde{\boldsymbol{h}}_{f}^{(n)}. \tag{2}$$

Here, $t \in \{1, \ldots, T\}$ and $f \in \{1, \ldots, F\}$ denote the frame and the frequency-bin indices, $\tilde{s}_{tf}^{(n)}$ the STFT of the nth source signal,

$$\tilde{\boldsymbol{h}}_{f}^{(n)} \triangleq \begin{bmatrix} \tilde{h}_{f}^{(1,n)} & \cdots & \tilde{h}_{f}^{(M,n)} \end{bmatrix}^{\mathsf{T}} \tag{3}$$

the time-invariant transfer function from the nth source to the microphones, and $^{\mathsf{T}}$ transposition.

(2) can be rewritten as

$$\boldsymbol{y}_{tf} = \sum_{n=1}^{N} s_{tf}^{(n)} \boldsymbol{h}_{f}^{(n)}, \tag{4}$$

where

$$s_{tf}^{(n)} \triangleq \tilde{s}_{tf}^{(n)} \tilde{h}_{f}^{(1,n)}, \tag{5}$$

$$\boldsymbol{h}_{f}^{(n)} \triangleq \frac{\tilde{\boldsymbol{h}}_{f}^{(n)}}{\tilde{h}_{f}^{(1,n)}}. \tag{6}$$

$s_{tf}^{(n)}$ is the nth source signal observed at the first microphone. $\boldsymbol{h}_{f}^{(n)}$ consists of relative transfer functions (RTFs) $\tilde{h}_{f}^{(m,n)}/\tilde{h}_{f}^{(1,n)}$ $(m = 1, \ldots, M)$, and is hence called the RTF vector in this paper.

Under the disjointness assumption (see Section 1), the observation model (4) is simplified as

$$\boldsymbol{y}_{tf} = s_{tf}^{(\nu)} \boldsymbol{h}_{f}^{(\nu)} \ \text{ where } \ \nu = d_{tf}, \tag{7}$$

where d_{tf} denotes the index of the dominant source in the time-frequency slot (t, f).

In (7), we assume no noise for simplicity. The readers are referred to [14] for integration with noise suppression.

Our goal is to estimate $s_{tf}^{(n)}$ from $\boldsymbol{y}_{tf}$ without knowing the RTF vector $\boldsymbol{h}_{f}^{(n)}$ or the number of sources, N. Once we have estimated d_{tf}, we can estimate $s_{tf}^{(n)}$ by, *e.g.*,

$$\hat{s}_{tf}^{(n)} \triangleq \mathcal{M}_{tf}^{(n)} y_{tf}^{(1)}, \tag{8}$$

where

$$\mathcal{M}_{tf}^{(n)} \triangleq \begin{cases} 1, & \text{if } \hat{d}_{tf} = n, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Here, $\hat{}$ denotes the estimate, and $\mathcal{M}_{tf}^{(n)}$ is the time-frequency mask for extracting the nth source signal.

Therefore, our problem has boiled down to the estimation of d_{tf}, which is dealt with in Section 3.

3. OUR METHOD

3.1. Relative Transfer Function (RTF) Features

Our observation model (7) implies that $\boldsymbol{y}_{tf}$ is parallel to the RTF vector $\boldsymbol{h}_{f}^{(\nu)}$, while its Euclidean norm $\|\boldsymbol{y}_{tf}\|$ is affected by the source spectrum $s_{tf}^{(\nu)}$. Therefore, a normalized observation vector [5, 7]

$$\boldsymbol{x}_{tf} \triangleq \frac{\boldsymbol{y}_{tf}}{\|\boldsymbol{y}_{tf}\|} \tag{10}$$

contains information on the RTF vector. Indeed, under (7), (10) equals an unit directional vector of the RTF vector $\boldsymbol{h}_{f}^{(\nu)}$, because substitution of (7) into (10) gives

$$\boldsymbol{x}_{tf} = \frac{s_{tf}^{(\nu)}}{\left|s_{tf}^{(\nu)}\right|} \cdot \frac{\boldsymbol{h}_{f}^{(\nu)}}{\left\|\boldsymbol{h}_{f}^{(\nu)}\right\|} \ \text{ where } \ \nu = d_{tf}. \tag{11}$$

Hence, we call $\boldsymbol{x}_{tf}$ the RTF features. Note that $s_{tf}^{(\nu)}/|s_{tf}^{(\nu)}|$ is a mere phase factor, which affects neither the direction nor the norm of $\boldsymbol{x}_{tf}$.

3.2. Watson Mixture Model for RTF Feature Clustering

We perform clustering of the RTF features by fitting a Watson mixture model (WMM) [5, 7]

$$p(\boldsymbol{x}_{tf}|\Theta) = \sum_{n=1}^{N} P(d_{tf} = n|\Theta) \mathcal{W}\big(\boldsymbol{x}_{tf}; \boldsymbol{a}_{f}^{(n)}, \kappa_{f}^{(n)}\big), \tag{12}$$

where Θ denotes the set of all model parameters. Here, the Watson distribution

$$\mathcal{W}\big(\boldsymbol{x}_{tf}; \boldsymbol{a}_{f}^{(n)}, \kappa_{f}^{(n)}\big) \propto \exp\big(\kappa_{f}^{(n)} \big|\boldsymbol{a}_{f}^{(n)\mathsf{H}} \boldsymbol{x}_{tf}\big|^2\big) \tag{13}$$

is defined on the unit hypersphere in $\mathbb{C}^M$, and models the histogram of $\boldsymbol{x}_{tf}$ for the nth source. The mean orientation

$$\boldsymbol{a}_{f}^{(n)} \triangleq \frac{\boldsymbol{h}_{f}^{(n)}}{\left\|\boldsymbol{h}_{f}^{(n)}\right\|} \tag{14}$$

equals a nominal value of $\boldsymbol{x}_{tf}$ for the nth source, which can be seen from (11). The concentration parameter $\kappa_{f}^{(n)} \geq 0$ controls the degree of deviation of $\boldsymbol{x}_{tf}$ from $\boldsymbol{a}_{f}^{(n)}$ due to modeling errors of (7).

Note that (13) is an increasing function of $\big|\boldsymbol{a}_{f}^{(n)\mathsf{H}} \boldsymbol{x}_{tf}\big| = \cos\theta$, where θ $(0 \leq \theta \leq \pi/2)$ denotes the intersection angle between $\boldsymbol{x}_{tf}$ and $\boldsymbol{a}_{f}^{(n)}$. This has two important implications.

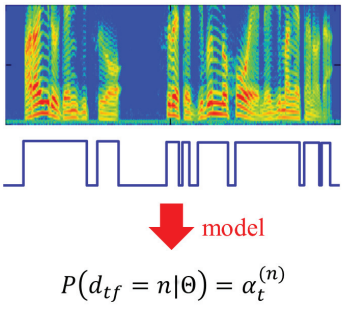

Fig. 1. The common amplitude modulation of a speech signal is modeled by the time-variant, frequency-invariant prior probability of the signal being dominant.

First, (13) is a decreasing function of θ, and takes its maximum (minimum) value at $\boldsymbol{x}_{tf}$ parallel (perpendicular) to $\boldsymbol{a}_f^{(n)}$. Second, (13) is insensitive to the phase factor $s_{tf}^{(\nu)}/|s_{tf}^{(\nu)}|$ in (11), since the factor does not affect the direction of the vector.

Experiments have shown that the clustering of the RTF features is effective for BSS even under reverberation [5]. This is partly attributed to the distance measure defined by the WMM. Another reason is that this approach does not assume planewave propagation of the source signals unlike fullband clustering of, *e.g.*, the DOA or the TDOAs [1–3].

3.3. Time-variant Source Dominance Prior for Modeling Common Amplitude Modulation of a Speech Signal

It is known that the frequency components of a speech signal tend to be activated synchronously, which is called the *common amplitude modulation* property (see Fig. 1). The property enables us to group together the frequency components activated synchronously as corresponding to the same source.

We model the common amplitude modulation property by assuming that the prior probability of the nth source being dominant, $P(d_{tf} = n|\Theta)$, is dependent on t but independent of f:

$$P(d_{tf} = n|\Theta) = \alpha_t^{(n)}. \tag{15}$$

The parameter $\alpha_t^{(n)}$ is called the *source dominance prior*, and represents the amplitude modulation of the nth source. Note that $\alpha_t^{(n)}$ should satisfy

$$\sum_{n=1}^{N} \alpha_t^{(n)} = 1. \tag{16}$$

3.4. BSS Based on Permutation-free Clustering

Noting that the source dominance priors (15) are also the mixture weights in the WMM (12), we obtain

$$p(\boldsymbol{x}_{tf}|\Theta) = \sum_{n=1}^{N} \alpha_t^{(n)} \mathcal{W}\big(\boldsymbol{x}_{tf}; \boldsymbol{a}_f^{(n)}, \kappa_f^{(n)}\big). \tag{17}$$

By using (17), which models the common amplitude modulation property explicitly, our method can perform bin-wise clustering and, at the same time, group together the bin-wise clusters corresponding to the same source. We call it *permutation-free clustering*. This contrasts with the conventional approach [5, 7], which performs bin-wise clustering and permutation alignment separately.

The permutation-free clustering is realized by jointly estimating $\alpha_t^{(n)}$, $\boldsymbol{a}_f^{(n)}$, and $\kappa_f^{(n)}$ based on maximum *a posteriori* (MAP) estimation. Specifically, we maximize the logarithm of the posterior probability defined by

$$\ln p\big(\Theta\big|\{\boldsymbol{x}_{tf}\}_{tf}\big) \stackrel{\mathrm{c}}{=} \sum_{t=1}^{T}\sum_{f=1}^{F} \ln p(\boldsymbol{x}_{tf}|\Theta) + \ln p(\Theta), \tag{18}$$

with respect to the parameter set Θ, where $\stackrel{\mathrm{c}}{=}$ denotes equality up to a constant. The prior $p(\Theta)$ is the following isotropic Dirichlet prior on $\{\alpha_t^{(n)}\}_{n=1}^N$:

$$p(\Theta) \propto \prod_{n=1}^{N} \big(\alpha_t^{(n)}\big)^{\phi-1}, \tag{19}$$

where ϕ is a hyperparameter. The optimization can be performed efficiently based on an expectation-maximization (EM) based algorithm detailed in [13].

Using the estimated parameters, the posterior probability of the nth source being active, $\gamma_{tf}^{(n)} \triangleq P(d_{tf} = n|\boldsymbol{x}_{tf}, \Theta)$, can be computed as follows:

$$\gamma_{tf}^{(n)} = \frac{\alpha_t^{(n)} \mathcal{W}\big(\boldsymbol{x}_{tf}; \boldsymbol{a}_f^{(n)}, \kappa_f^{(n)}\big)}{\sum_{\nu=1}^{N} \alpha_t^{(\nu)} \mathcal{W}\big(\boldsymbol{x}_{tf}; \boldsymbol{a}_f^{(\nu)}, \kappa_f^{(\nu)}\big)}. \tag{20}$$

Source separation can be performed by (8), where the dominant source index d_{tf} is estimated as

$$\hat{d}_{tf} = \arg\max_n \gamma_{tf}^{(n)}. \tag{21}$$

Note that we can design not only time-frequency masks but also beamformers, because we can also identify the RTF vector using the estimated parameters. Noting that (14) implies $\boldsymbol{h}_f^{(n)} \parallel \boldsymbol{a}_f^{(n)}$ and that $h_f^{(1,n)} = 1$ by definition, we obtain

$$\boldsymbol{h}_f^{(n)} = \frac{\boldsymbol{a}_f^{(n)}}{a_f^{(1,n)}}. \tag{22}$$

Here, $a_f^{(1,n)}$ denotes the first entry of $\boldsymbol{a}_f^{(n)}$.

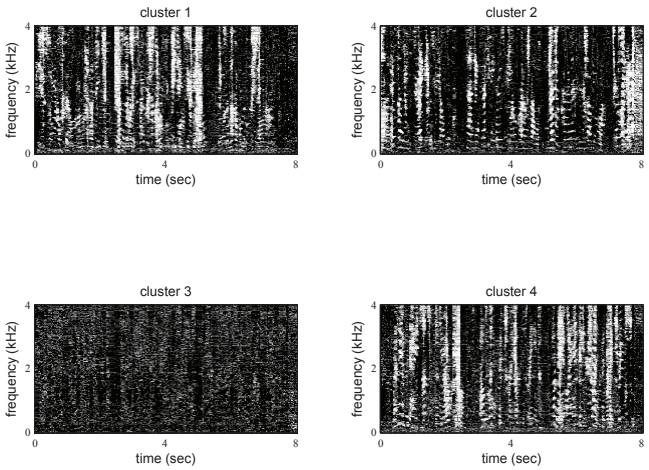

Fig. 2. Posterior probabilities $\gamma_{tf}^{(n)}$ for $N = 3, L = 4$. A lighter color means a larger $\gamma_{tf}^{(n)}$. Clusters 1, 2, and 4 are judged as the dominant clusters.

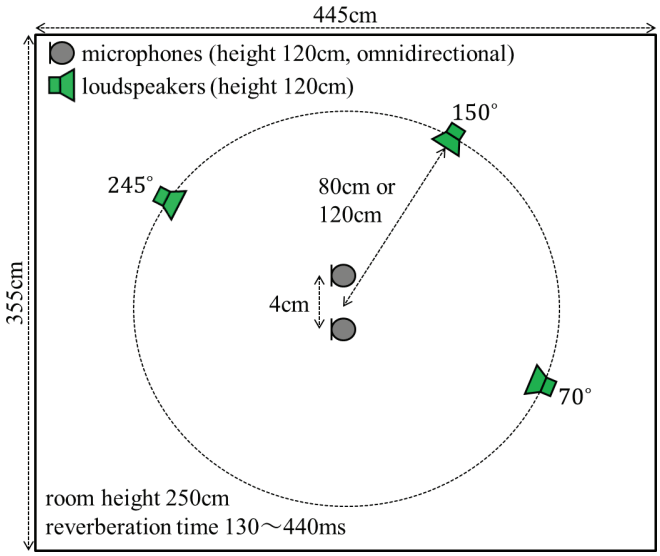

Fig. 3. Experimental conditions.

3.5. Source Counting

Our method can perform BSS even when N is unknown. Indeed, if we set the number of clusters, L, greater than N, the common amplitude modulation property makes clusters with synchronous activation grouped together, which results in N dominant clusters and $L - N$ almost empty clusters (see Fig. 2). Therefore, N can be estimated by counting the dominant clusters.

To this end, we first calculate the following total activity for each cluster:

$$\rho^{(l)} \triangleq \sum_{t=1}^{T} \sum_{f=1}^{F} \gamma_{tf}^{(l)}. \tag{23}$$

We then apply k-means clustering with two clusters to $\{\rho^{(l)}\}_{l=1}^{L}$. Since only N elements of $\{\rho^{(l)}\}_{l=1}^{L}$ have significant values, N can be obtained as the number of elements of the cluster with the larger centroid.

4. PERFORMANCE EVALUATION

To evaluate the source separation and the source counting performances of our method (called the "proposed method" hereafter), we conducted experiments under the conditions shown in Fig. 3. To generate mixtures, we convolved 8s-long English speech signals with room impulse responses measured in the environment shown in Fig. 3. The sampling frequency was 8 kHz, the frame length was 1024 points (128 ms), and the frame shift was 256 points (32 ms). The hyperparameter of the Dirichlet prior was set at $\phi = 600$. The number of iterations in the EM-based algorithm was 100. As a baseline, we also evaluated the source separation performance of the conventional method [5], which assumes that N is given.

Table 2. Source counting accuracy (%) of the proposed method. We conducted 16 trials, and calculated the ratio of the trials with correct source counting to the total number of trials.

N	reverberation time (ms)					
	130	200	250	300	370	440
2	100%	100%	100%	100%	100%	100%
3	100%	100%	93%	100%	100%	93%

Table 1 shows the source separation performance in terms of the signal-to-distortion ratio (SDR) [15], when N was given. We see that the proposed method achieved SDRs comparable to the conventional method when N was given.

Table 2 shows the source counting accuracy by the proposed method, when N was unknown. We set the number of clusters $L = 4$, which exceeded the number of sources $N = 2, 3$. We see that the proposed method counted sources almost perfectly under all conditions including underdetermined and reverberant conditions, where source counting has conventionally been difficult [11, 16]. Note that the conventional method [5] cannot be applied in this case, because N was unknown.

5. CONCLUSION

In this paper, we have introduced our method for joint BSS and source counting based on permutation-free clustering of the RTF features.

The future work includes extension of the application area of the method by taking advantage of its probabilistic formulation. Firstly, we are working on integration of the method

Table 1. Source separation performance in terms of signal-to-distortion ratio (SDR). We averaged the SDRs of 16 trials with different combinations of speech signals and different distances between sources and the array centroid.

N	method	reverberation time (ms)					
		130	200	250	300	370	440
2	conventional [5]	16.6	14.9	13.0	12.1	10.9	10.0
	proposed	16.6	14.9	13.2	12.3	11.0	9.9
3	conventional [5]	9.7	8.8	8.1	7.0	6.2	5.8
	proposed	9.2	8.2	7.4	6.6	5.7	5.4

with dereverberation and denoising techniques. Initial such attempts are found in [14, 17]. Secondly, we plan to develop the method to an online method, so that it can deal with a time-variant number of moving sources.

REFERENCES

[1] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. SP*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[2] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, Aug. 2007.

[3] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *Proc. WASPAA*, Oct. 2007, pp. 147–150.

[4] M.I. Mandel, R.J. Weiss, and D.P.W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. ASLP*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[5] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. ASLP*, vol. 19, no. 3, pp. 516–527, Mar. 2011.

[6] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoust. Sci. Tech.*, vol. 22, no. 2, pp. 149–157, May 2001.

[7] D.H. Tran Vu and R. Häb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. ICASSP*, Mar. 2010, pp. 241–244.

[8] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. SP*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[9] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. SAP*, vol. 12, no. 5, pp. 451–459, Sept. 2004.

[10] Z. Koldovsky, J. Malek, and S. Gannot, "Spatial source subtraction based on incomplete measurements of relative transfer function," 2014, arXiv:1411.2744v2.

[11] K. Yamamoto, F. Asano, W.F.G. van Rooijen, E.Y.L. Ling, T. Yamada, and N. Kitawaki, "Estimation of the number of sound sources using support vector machines and its application to sound source separation," in *Proc. ICASSP*, Apr. 2003, vol. V, pp. 485–488.

[12] N. Ito, S. Araki, K. Kinoshita, and T. Nakatani, "Permutation-free clustering method for underdetermined blind source separation based on source location information," *IEICE Trans.*, vol. J97-A, no. 4, pp. 234–246, Apr. 2014 (in Japanese).

[13] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *Proc. ICASSP*, May 2013, pp. 3238–3242.

[14] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Trans. ASLP*, vol. 21, no. 9, pp. 1913–1928, Sept. 2013.

[15] E. Vincent, R. Griboval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[16] L. Drude, A. Chinaev, D.H. Tran Vu, and R. Häb-Umbach, "Source counting in speech mixtures using a variational EM approach for complex Watson mixture models," in *Proc. ICASSP*, May 2014, pp. 6834–6838.

[17] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept. 2014, pp. 269–273.