Sparse source separation based on simultaneous clustering of source locational and spectral features

Shoko Araki*, Tomohiro Nakatani and Hiroshi Sawada

NTT Communication Science Laboratories, NTT Corporation

(Received 28 January 2011, Accepted for publication 24 March 2011)

Keywords: Blind source separation, Sparseness, Time-frequency mask, Phase difference, Common amplitude modulation
PACS number: 43.60.-c [doi:10.1250/ast.32.161]

1. Introduction

Blind source separation (BSS) methods for speech signals have been widely studied, including independent component analysis (ICA) based approaches (e.g., [1]) and methods that rely on sparseness of sources (e.g., [2]), and it has been reported that such approaches perform well even in reverberant conditions. This paper focuses on the latter approach, more specifically, the time-frequency (t-f) mask approach [2–5].

Many BSS algorithms employ a time-frequency (t-f) domain approach, and most of them employ a two-stage approach: the first stage solves BSS at each frequency bin, and then solves the permutation problem between frequencies. As one approach to the permutation problem, a method has been proposed [5,6] based on the common amplitude modulation (AM) property of speech signals [7], which means the temporal envelope of the frequency components tend to be synchronized for one speech signal. It has been confirmed that the common AM based permutation solvers are robust against room reverberation [5]. However, the two-stage approach has inherent difficulties, for example, BSS performance depends on the separation performance of the first stage, and the second stage requires a large calculation cost when the number of sources is large. Moreover, most of the existing BSS methods require the number of sources. If the number is unavailable, we have to estimate it to perform BSS.

This paper proposes a BSS method for sparse sources that has an inherent ability to align the permutation of frequency components, and the approach can be applied even if the number of sources is unknown. The proposed method simultaneously classifies both the source locational feature (phase difference between microphones) and the spectral feature (spectral temporal envelope). With the proposed method, source separation in each frequency bin is performed by clustering the source locational features, while synchronizing the spectral envelope. Thanks to the common AM property, the proposed method inherently aligns the permutation of the frequency components.

The abovementioned simultaneous clustering mechanism is formulated with the maximization of a likelihood function, which can easily be realized with the expectation-maximiza-

tion (EM) algorithm. The authors of [8,9] also formulated a sparse source separation method with the EM algorithm. However, they still needed the number of sources N_s in advance.

The proposed method also has the ability to separate signals by estimating the number of sources when the actual number is unavailable. This capability is realized by defining a cost function that has a high likelihood when the frequency components are synchronized, and when we can model the microphone observations with as small a number of sources as possible. This property enables us to estimate the number of sources by counting the number of dominant estimated source signals.

The organization of this paper is as follows. Section 2 describes our problem. Then, in Section 3 we explain our proposed method with the simultaneous clustering of source locational and spectral features. Section 4 reports source number estimation and source separation results.

2. Problem description

Let us formulate the task. Suppose that $n(= 1, \dots, N_n)$ and $f(= 1, \dots, N_f)$ are time and frequency indices of a t-f slot, and that $N_s \geq 2$ speech signals $s_{n,f,1}, \dots, s_{n,f,N_s}$ are mixed and observed at N_m microphones,

$$x_{n,f,l} = \sum_{m=1}^{N_s} h_{f,l,m} s_{n,f,m}, \quad (1)$$

where $x_{n,f,l}$ is the observation at the l -th microphone ($l = 1, \dots, N_m$), and $h_{f,l,m}$ represents the transfer function from source m to microphone l . In this paper, without loss of generality, we consider a stereo case $N_m = 2$. Our goal is to obtain estimates $y_{n,f,m}$ of the sources $s_{n,f,m}$ from observations $x_{n,f,l}$ without information on the speech sources $s_{n,f,m}$, or the mixing process $h_{f,l,m}$. The number of sources N_s could be unavailable.

To obtain the separated signals $y_{n,f,m}$, this paper utilizes t-f masks $M_{n,f,m}$, which extract the t-f points of the m -th source. I.e., the separated signals $y_{n,f,m}$ are obtained by

$$y_{n,f,m} = M_{n,f,m} x_{n,f,1}. \quad (2)$$

To estimate the t-f masks $M_{n,f,m}$, this paper assumes the sparseness of the sources [2], that is, at most one source is dominant at each t-f slot. With this assumption, we rewrite (1) as follows:

*e-mail: araki.shoko@lab.ntt.co.jp

Fig. 1 Graphical model.

$$s_{n,f,m} = \begin{cases} x_{n,f,1} & \text{for } m = z_{n,f} \\ 0 & \text{for } m \neq z_{n,f} \end{cases} \quad (3)$$

where $z_{n,f}$ is the dominant source index at the t-f slot (n, f) . In (3), we assume that $|h_{f,1,m}| = 1$ and $\arg(h_{f,1,m}) = 0$, without loss of generality.

3. Proposed method

Our proposed method is based on the simultaneous clustering of source locational and spectral features. The clustering of the locational features, which in this paper are the phase differences between two channels $A_{n,f} = \arg[\frac{x_{n,f,2}}{x_{n,f,1}}]$, separates signals in each frequency bin. At the same time, the clustering of the complex spectrum $X_{n,f} = x_{n,f,1}$ maximizes the synchrony of the AM structure across frequencies.

3.1. Probabilistic model

Letting θ be a set of model parameters that are detailed in the following subsections, we model the features $X_{n,f}$ and $A_{n,f}$ with a mixture model:

$$p_{n,f}(X_{n,f}, A_{n,f}; \theta) = \sum_m^M p_{n,f}(X_{n,f}, A_{n,f}|m; \theta) p(m; \theta), \quad (4)$$

where we denote " $z_{n,f} = m$ " by " m ," and $p(m; \theta) = \alpha_m$ is the mixture weight ($\sum_m \alpha_m = 1$). To handle the unknown number of sources N_s , we prepare a sufficient number M of classes (i.e., $m = 1, \dots, M$).

Assuming that $X_{n,f}$ and $A_{n,f}$ are mutually independent for a given m , we write $p_{n,f}(X_{n,f}, A_{n,f}|m; \theta)$ as

$$p_{n,f}(X_{n,f}, A_{n,f}|m; \theta) = p_{n,f}(X_{n,f}|m; \theta_X) \cdot p_f(A_{n,f}|m; \theta_A) \quad (5)$$

where $p_{n,f}(X_{n,f}|m; \theta_X)$ and $p_f(A_{n,f}|m; \theta_A)$ are the sparse observation model (Sect. 3.2) and the phase difference model (Sect. 3.3), respectively. The maximization of $p_f(A_{n,f}|m; \theta_A)$ corresponds to the frequency bin-wise separation, and that of $p_{n,f}(X_{n,f}|m; \theta_X)$ synchronizes the AM structure across frequencies and solves the permutation problem. The graphical model is shown in Fig. 1.

3.2. Sparse observation model

This section defines the observation model $p_{n,f}(X_{n,f}|m; \theta_X)$. We interpret the sparseness assumption in (3) as follows: two types of observations, $x_{n,f,1}$ and zeros, are always obtained simultaneously in relation to the dominant source and the other non-dominant sources. This interpretation can be represented by letting $s_{n,f} = [s_{n,f,1}, \dots, s_{n,f,N_s}]$, $X_{n,f} = x_{n,f,1}$ and $\delta(s)$ be the Dirac delta, as follows:

$$p_{n,f}(X_{n,f}|s_{n,f}, m) = \delta(X_{n,f} - s_{n,f,m}) \prod_{m' \neq m} \delta(-s_{n,f,m'}) \quad (6)$$

In this paper, we assume that $s_{n,f}$ is independent of m

Fig. 2 Generative model of (9).

$$p_{n,f}(s_{n,f}|m) = p_{n,f}(s_{n,f}) = \prod_m p_{n,f}(s_{n,f,m}), \quad (7)$$

and we model the complex spectrum of each source $s_{n,f,m}$ with a t-f dependent complex Gaussian distribution N_c :

$$p_{n,f}(s_{n,f,m}) = N_c(s_{n,f,m}; 0, \gamma_{n,f,m}^2) \quad (8)$$

where $\gamma_{n,f,m}^2$ is the variance of the source spectrum $E[|s_{n,f,m}|^2]$ at each t-f point.

By multiplying (6) and (7) considering (8) and marginalizing it on $s_{n,f}$, we obtain the sparse observation model:

$$p_{n,f}(X_{n,f}|m; \gamma_{n,f,m}^2) = N_c(X_{n,f}; 0, \gamma_{n,f,m}^2) \cdot \prod_{m' \neq m}^M N_c(0; 0, \gamma_{n,f,m'}^2). \quad (9)$$

The generative model of (9) is illustrated in Fig. 2. It should be noted that we may be able to employ a different type of sparse observation model (e.g., [10]), however, this paper utilizes (9).

3.2.1. Spectral model with common AM

We also assume that each source $s_{n,f,m}$ has a common AM structure [7] across frequencies. That is, the onset and offset of the frequency components are synchronized. Figure 3 shows examples of speech spectra. We can see the synchrony of the frequency components. This synchrony is termed "source activity" in [5].

To model the common AM structure of source signals, we parameterize the spectral variance $\gamma_{n,f,m}^2$ with the time-variant spectral envelope $\rho_{n,m}$ and the t-f dependent spectral fine structure $a_{n,f,m}$:

$$\gamma_{n,f,m} = a_{n,f,m} \rho_{n,m}. \quad (10)$$

The time-variant spectral envelope $\rho_{n,m}$ models the onset synchrony across frequencies. For the spectral fine structure $a_{n,f,m}$ in (10), we use the observation spectrum $a_{n,f,m} = |X_{n,f}|$ in this paper. The spectral parameter set we have to estimate is $\theta_X = \{\rho_{n,m}\}$.

3.3. Phase difference model

This section defines the phase difference model $p_f(A_{n,f}|m; \theta_A)$. Assuming that all sources are located at different locations, and thus have different directions of arrival (DOA), we model the phase difference $A_{n,f}$ with a Gaussian

$$p_f(A_{n,f}|m; \mu_{f,m}, \sigma_{f,m}^2) = N(A_{n,f}; \mu_{f,m}, \sigma_{f,m}^2), \quad (11)$$

where $\mu_{f,m}$ and $\sigma_{f,m}^2$ are the frequency dependent mean and variance of phase difference, and $\theta_A = \{\mu_{f,m}, \sigma_{f,m}^2\}$.

Fig. 3 Example spectrum of a speech signal. We can see the common AM structure.

3.4. EM algorithm for likelihood maximization

This section provides the optimization algorithm. Let $\theta = \{\mu_{f,m}, \sigma_{f,m}^2, \rho_{n,m}^2, \alpha_m\}$ be a model parameter set. In the following, the dominant source index m is dealt with as a hidden variable.

The likelihood function is

$$L(\theta) = \sum_n \sum_f \log p_{n,f}(X_{n,f}, A_{n,f}; \theta), \quad (12)$$

where $p_{n,f}(X_{n,f}, A_{n,f}; \theta)$ is given by (4) and (5). The likelihood function (12) can be maximized by using the EM algorithm.

E-step: The auxiliary function Q is given as

$$\begin{aligned} Q(\theta|\theta^t) &= \sum_n \sum_f E[\log p_{n,f}(X_{n,f}, A_{n,f}, m; \theta) | \theta^t] \\ &= \sum_n \sum_f \sum_m g_{n,f,m} \log p_{n,f}(X_{n,f} | m; \theta) \\ &\quad + \sum_n \sum_f \sum_m g_{n,f,m} \log p_f(A_{n,f} | m; \theta) \\ &\quad + \sum_n \sum_f \sum_m g_{n,f,m} \log p(m; \theta), \end{aligned} \quad (13)$$

where θ^t is an estimate of the parameters after the t -th iteration and

$$\begin{aligned} g_{n,f,m} &= p_{n,f}(m | X_{n,f}, A_{n,f}; \theta^t) \\ &= \frac{p_{n,f}(X_{n,f} | m; \theta^t) p_f(A_{n,f} | m; \theta^t) p(m; \theta^t)}{\sum_m p_{n,f}(X_{n,f} | m; \theta^t) p_f(A_{n,f} | m; \theta^t) p(m; \theta^t)}. \end{aligned} \quad (14)$$

This posterior $g_{n,f,m}$ corresponds to the probability that source m is active at a t-f slot (n, f) .

M-step: The update rules for the phase difference parameters are obtained by setting $\frac{\partial Q(\theta|\theta^t)}{\partial \mu_{f,m}} = 0$ and $\frac{\partial Q(\theta|\theta^t)}{\partial \sigma_{f,m}^2} = 0$:

$$\mu_{f,m}^{t+1} = \frac{\sum_n g_{n,f,m} A_{n,f}}{\sum_n g_{n,f,m}}, \quad (15)$$

$$(\sigma_{f,m}^2)^{t+1} = \frac{\sum_n g_{n,f,m} (A_{n,f} - \mu_{f,m})^2}{\sum_n g_{n,f,m}}. \quad (16)$$

The update rule for time-variant spectral envelope $\rho_{n,m}$ is derived by $\frac{\partial Q(\theta|\theta^t)}{\partial \rho_{n,m}^2} = 0$ as:

$$(\rho_{n,m}^2)^{t+1} = \frac{\sum_f g_{n,f,m}}{N_f} \quad (17)$$

where N_f is the number of frequency bins. Note that Eq. (17) has the same form as the calculation of the centroid of the source activity in [5].

Moreover, by using the Lagrange multiplier method and $\sum_m^M \alpha_m = 1$, the mixture weight is obtained as follows:

$$\alpha_m^{t+1} = \frac{\sum_n \sum_f g_{n,f,m}}{N_n N_f}, \quad (18)$$

where N_n is the number of time frames.

3.5. Source number estimation

The spectral model (9) has a high likelihood value when only one source has a large variance and the others are zero. That is, the model (9) explains the observed signal $X_{n,f}$ with as small a number of sources as possible. From this fact and (18), just a few mixture weights α_m for dominant sources have a sufficiently large values, and other $\alpha_{m'}$ values become almost zero. Therefore, we can determine the number of sources N_s by counting the number of mixture weights whose values are sufficiently large, i.e., $\alpha_m \geq th$ ($th = 0.1$ in this paper).

3.6. Source separation

The t-f mask $M_{n,f,m}$ in (2) for the m -th separated source is obtained by the posterior (14), that is, $M_{n,f,m} = g_{n,f,m}$. The separated signal is obtained by

$$y_{n,f,m} = g_{n,f,m} x_{n,f,1}. \quad (19)$$

4. Experiments

4.1. Experimental setup

The mixed signals were made by convolving the measured room impulse responses 5-second English speech signals sampled at 8 kHz. The number of sources was $N_s = 2$. We utilized two microphones with spaces of 4 cm (no spatial aliasing) and 20 cm (spatial aliasing occurs above 850 Hz). The room reverberation time was 130 ms. The frame size N_f for STFT was 512 (64 ms), and the frame shift was 128 (16 ms).

In our implementation, to balance the range of the phase and spectrum likelihoods in (5), we employ weight parameters w_a and w_x for these likelihood values, i.e., $p_{n,f}(X_{n,f}, A_{n,f} | m; \theta) = p_{n,f}^{w_x}(X_{n,f} | m; \theta_x) \cdot p_f^{w_a}(A_{n,f} | m; \theta_a)$. In the EM algorithm, we have to consider these weights for $g_{n,f,m}$ calculation (14). The weights in this paper were $w_a = 1$ and $w_x = 0.2$.

The separation performance was evaluated with the signal-to-interference ratio (SIR) as a measure of separation performance and the signal-to-distortion ratio (SDR) as a measure of sound quality. Their definitions can be found in [3]. We calculated SIR and SDR values for the separated sources that are counted as sources by the method in Sect. 3.5. We performed 20 trials with different speech source and location combinations and then averaged the results.

4.2. Results

Figure 4 plots the example t-f masks $M_{n,f,m} = g_{n,f,m}$ obtained with the proposed method by assuming $M = 8$. The figure shows that just two ($= N_s$) masks have large values. This result and (18) prove that we can estimate the number of sources by using the method described in Section 3.5. The accuracies for source number estimation were 100% and

Fig. 4 Example estimated masks $M_{n,f,m} = g_{n,f,m}$. The number of classes $M = 8$ ($m = 1, \dots, 8$).

Fig. 5 Example estimated parameters for $m = 4$ and $m = 5$ in Fig. 4. (a) Mean parameter $\mu_{f,m}$ of the phase difference and (b) spectral envelope parameter $\rho_{n,m}$.

75.0% when the microphone spacings were 4 cm and 20 cm, respectively. The reason for the poor performance with a microphone spacing of 20 cm case is because source separation becomes difficult at some frequencies due to spatial aliasing.

Figures 5(a) and (b) show the phase $\mu_{f,m}$ and spectral parameters $\rho_{n,m}$ for $m = 4$ and $m = 5$ in Fig. 4, respectively. It can be seen that we obtain the phase parameters $\mu_{f,m}$ that have a linear phase property, and that the spectrum parameter $\rho_{n,m}$ provides the spectral temporal envelope.

Figure 6 summarizes the separation performance for $M = 8$. In the figure, “k-means” and “proposed” denote the results obtained with the conventional method [3] and the

Fig. 6 Experimental results. Input SIR was 0 [dB].

proposed method, respectively. We can see that the proposed method achieves high performance even when the number of sources is unknown.

5. Conclusion

This paper proposed BSS method based on the simultaneous clustering of source locational and spectrum features. We confirmed that our proposed method can separate signals successfully even when the number of sources N_s is unknown. Future study will include an evaluation where there are more sources, i.e., $N_s \geq 3$.

References

- [1] A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis* (John Wiley & Sons, New York, 2001).
- [2] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. Signal Process.*, **52**, 1830–1847 (2004).
- [3] S. Araki, H. Sawada, R. Mukai and S. Makino, “Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors,” *Signal Process.*, **77**, 1833–1847 (2007).
- [4] Y. Izumi, N. Ono and S. Sagayama, “Sparseness-based 2ch BSS using the EM algorithm in reverberant environment,” *Proc. WASPAA2007*, pp. 147–150 (2007).
- [5] H. Sawada, S. Araki and S. Makino, “Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment,” *IEEE Trans. Audio Speech Lang Process.*, **19**, 516–527 (2011).
- [6] N. Murata, S. Ikeda and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals,” *Neurocomputing*, **41**(1–4), pp. 1–24 (2001).
- [7] G. J. Brown, *Computational Auditory Scene Analysis: A Representational Approach*, Ph.D. thesis, University of Sheffield (1992).
- [8] M. Mandel, D. Ellis and T. Jebara, “An EM algorithm for localizing multiple sound sources in reverberant environments,” *Proc. Neural Inf. Process. Syst.* (2006).
- [9] P. O’Grady and B. Pearlmutter, “Soft-LOST: EM on a mixture of oriented lines,” *Proc. ICA 2004 (LNCS 3195)*, pp. 430–436 (2004).
- [10] T. Nakatani, S. Araki, T. Yoshioka and M. Fujimoto, “Multi-channel source separation based on source location cue with log-spectral shaping by hidden Markov source model,” *Interspeech 2010*, pp. 2766–2769 (2010).