IEEJ TRANSACTIONS ON ELECTRICAL AND ELECTRONIC ENGINEERING
IEEJ Trans 2015; **10**: 114–115
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI:10.1002/tee.22072

Letter

Semi-Blind Source Separation using Binary Masking and Independent Vector Analysis

Yuuki Tachioka[a], Member
Tomohiro Narita, Non-Member
Jun Ishii, Non-member

Recent prevalence of speech recognition system increases the opportunity of simultaneous recognition of multiple speakers' utterances. There are two types of source separation methods: physical and statistical. The former is based on the physical information such as a direction of arrival of sound sources. The latter only uses statistical independence. The advantage of the former is fast computation and effectiveness with precise information; and that of the latter is no need for physical information, which leads to the robustness of measurement errors. In this paper, we propose to combine these approaches effectively. Experiments on a speech recognition task show that the proposed method can achieve the upper limit performance of the two approaches. © 2014 Institute of Electrical Engineers of Japan. Published by John Wiley & Sons, Inc.

Keywords: binary masking, independent vector analysis, automatic speech recognition

Received 4 April 2014; Revised XXXX; Accepted XXXX

1. Introduction

Recent progress in speech recognition widens the target user base. In this scenario, simultaneous recognition of multiple speakers' utterances for a real-time use of a single system is required. Before recognition, some source separation approaches are applied. The most general one is based on physical information, such as the direction of arrival of the sound [1]. This method is fast and effective but susceptible to errors in physical information. On the other hand, blind source separation approach based on statistical independence [2] is more time consuming and may be inferior to the physical method with precise information but can be robust for measurement errors. In this paper, we propose to combine these physical and statistical approaches effectively to improve the robustness of source separations.

2. Binary masking in the time–frequency domain

From now on, the number of microphones is assumed to be 2. When x_1 and x_2 are the short-time Fourier transforms of the observed signals for the first and second microphone, respectively, a cross-spectrum of them at the time frame t $(1 \leq t \leq T)$ and frequency bin ω is represented as

$$x_2(\omega,t)/x_1(\omega,t) = A\mathrm{e}^{J\omega\tau(\omega,t)}, \tag{1}$$

where J is an imaginary unit, A is a positive amplitude ratio, and $\tau(\omega,t)$ is a time difference between them. The masking matrix W is composed of two vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$:

$$W(\omega,t) = (\boldsymbol{w}_1(\omega,t), \boldsymbol{w}_2(\omega,t))^h, \tag{2}$$

where h is an Hermite transpose. If the direction of the sound source θ is known, binary masking (BM) on time–frequency domain constructs the masks W as [1]

$$\boldsymbol{w}_k(\omega,t) = \begin{cases} \epsilon \boldsymbol{e}_k & : |\frac{c}{l_m}\sin^{-1}\tau_{\omega,t} - \theta| > \theta_c, \\ \boldsymbol{e}_k & : \text{otherwise}, \end{cases} \tag{3}$$

where k is the microphone ID, $\boldsymbol{e}_k$ is a unit vector whose kth element is 1, ϵ is a small number for smoothing, and θ_c is a tolerance error. c is a sound velocity and l_m is the distance between microphones. Separated signal $\boldsymbol{y}$ is obtained as

$$\boldsymbol{y}(\omega,t) = W(\omega,t)\boldsymbol{x}(\omega,t), \tag{4}$$

where $\boldsymbol{x}(\omega,t)$ and $\boldsymbol{y}(\omega,t)$ are vector forms of $(x_1(\omega,t), x_2(\omega,t))^\top$ and $(y_1(\omega,t), y_2(\omega,t))^\top$. $\top$ denotes a transpose. Separation is effective when the physical variables above are all reliable.

3. IVA using auxiliary function

Statistical method uses only the independence between sources and needs no physical information above. The most major statistical method, namely independent component analysis (ICA), causes the permutation problem about separated speakers because this method separates sources at each frequency bin. To address this problem, independent vector analysis (IVA) minimizes the objective function (5) across frequency bins and determines time-invariant separation matrices $W(\omega)$.

$$J(\boldsymbol{W}) = \sum_k E[r_{k,t}] - \sum_\omega \log|\mathrm{det}\, W(\omega)|. \tag{5}$$

where $\boldsymbol{W}$ is a set of $W(\omega)$, and $r_{k,t}$ is an auxiliary variable in (6). This can be optimized using an auxiliary function as an upper limit of J [2]. This method outperforms gradient-decent-based conventional methods. After the update of auxiliary variables (6), the separation matrices are updated in two steps: direction update rule (7) and norm normalization rule (8).

$$\begin{aligned} r_{k,t} &= \sqrt{\sum_\omega |\boldsymbol{w}_k^h(\omega)\boldsymbol{x}(\omega,t)|^2}, \\ V_k(\omega) &= \sum_{t=1}^{T}\left[\frac{\boldsymbol{x}(\omega,t)\boldsymbol{x}^h(\omega,t)}{Tr_{k,t}}\right]. \end{aligned} \tag{6}$$

[a] Correspondence to: Yuuki Tachioka.
E-mail: Tachioka.Yuki@eb.MitsubishiElectric.co.jp

Information Technology R & D Center, Mitsubishi Electric Corporation, 5–1–1, Ofuna, Kamakura, Kanagawa 247-8501, Japan

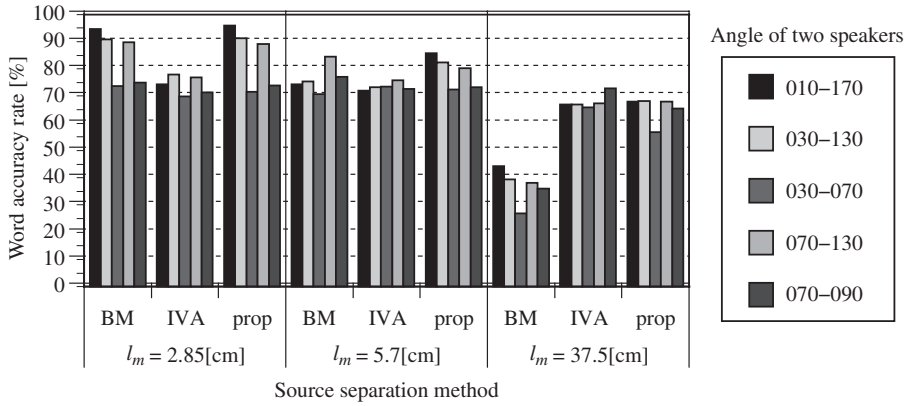

Fig. 1. Word accuracy rate [%] in terms of methods (BM, binary masking; IVA, independent vector analysis; prop, proposed method that combines BM and IVA) and angle of speaker to the microphone array. The iteration number of IVA and prop was 20.

Table I. Word accuracy rate [%] in terms of methods and the number of iterations

	$l_m = 2.85$[cm]			$l_m = 5.7$[cm]			$l_m = 37.5$[cm]		
iter	BM	IVA	prop	BM	IVA	prop	BM	IVA	prop
5	**8**4.8	61.1	84.2	76.4	60.9	78.2	37.0	57.6	56.8
10	-	69.1	84.3	-	69.3	79.1	-	64.0	61.8
15	-	72.6	84.3	-	72.5	79.0	-	66.8	64.4
20	-	74.1	84.4	-	73.5	**7**8.9	-	**6**8.0	65.3

$$\boldsymbol{w}_k(\omega) \leftarrow (W(\omega)V_k(\omega))^{-1}\boldsymbol{e}_k, \tag{7}$$

$$\boldsymbol{w}_k(\omega) \leftarrow \boldsymbol{w}_k(\omega)/\sqrt{\boldsymbol{w}_k^h(\omega)V_k(\omega)\boldsymbol{w}_k(\omega)}. \tag{8}$$

Finally, projection back [3] is applied to the separated matrix.

4. Proposed method

The main reason for the degradation of physical methods is spatial aliasing, which occurs in the frequency bands more than $f_c = \frac{c}{2l_m}$. For these bands, the performance of physical methods is significantly degraded; on the other hand, statistical method is robust. To address this problem, in the bands less than f_c, BM is used; otherwise, IVA is used. However, this simple combination causes a permutation problem similar to the ICA, thus we insert BM into the framework of IVA optimization. After BM is applied to the bands less than f_c, in the other bands IVA separates sources where, for all ωs, auxiliary variables and separation matrices are updated to guarantee the identity of separated speakers. Instead of the update rule (7), the following update rule is used:

$$\boldsymbol{w}_k(\omega) \leftarrow \begin{cases} (W(\omega)V_k(\omega))^{-1}\boldsymbol{e}_k & : \omega > 2\pi f_c, \\ \boldsymbol{e}_k & : \text{otherwise.} \end{cases} \tag{9}$$

5. Experiments

5.1. Setup Experiments on automatic speech recognition were performed. The impulse responses were measured in a variable reverberant room whose reverberation time was 300 ms. This was included in the Real World Computing Partnership (RWCP)-SSD database (E2A). Two microphones were picked up from the line array. The microphone intervals l_m were 2.85, 5.7, and 37.5 cm. Direction of arrival was given in this experiment, because that can be estimated with high accuracy [4]. Impulse responses were provided with the direction of arrival from 10 to 170° by 20°. This experiment used five combinations of them: (10,170), (30,130), (30,70), (70,130), and (70,90)°. The distance between the center of microphone array and the sound source was 2 m. Utterances were taken from Japan Electronic Industry Development Association (JEIDA)-JCSD (B-set), which was composed of 100 area names. Although the dictionary of the automatic speech recognition system was 100 area names, mixed speech was made from 30 area names with different area names. For speaker variety, 20 speaker sets were prepared from five male and five female speakers. The window length and window shift of short-time Fourier transform were 60 and 30 ms, respectively, and Mel-Frequency Cepstrum Coefficients (MFCC) features were used.

6. Results and Discussion

Table I shows the relationship between the word accuracy rate and the number of iterations for BM, IVA, and the proposed method (prop). Note that BM needs no iterations. For IVA and prop, 20 iterations were enough. For the $l_m = 2.85$[cm] case, BM achieved the highest performance, but increasing l_m degraded the performance. IVA was less susceptible for l_m, but for the $l_m = 2.85$[cm] case the performance was lower than that of BM. The proposed method achieved performance equal to that of BM for the $l_m = 2.85$[cm] case and to that of IVA for the $l_m = 37.5$[cm] case and achieved the best performance for the $l_m = 5.7$[cm] case.

Figure 1 shows the influence of speaker positions. When two speakers are positioned with more than 40° intervals, word accuracies were high for BM and prop; for the $l_m = 2.85$[cm] case, BM and prop achieved word accuracies more than 90%. IVA was less susceptible for speaker positions.

7. Conclusion

We proposed an effective combination of the physical and statistical methods. This can combine the advantages of two methods and improve the robustness of source separations. Speech recognition experiments showed that the proposed method achieved the upper limit of performance of the two methods.

References

(1) Sawada H, Araki, S, Makino S. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Transactions on Audio, Speech, and Language Process* 2011; **19**:516–527.
(2) Ono N. Stable and fast update rules for independent vector analysis based on auxiliary function technique. Proceedings of Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011, 189–192.
(3) Murata N, Ikeda S, Ziehe A. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing* 2001; **41**:1–24.
(4) Tachioka Y, Narita T, Iwasaki T. Direction of arrival estimation by cross-power spectrum phase analysis using prior distributions and voice activity detection information. *Acoustical Science & Technology* 2012; **33**:68–71