INTRODUCING NEW MECHANISM IN THE LEARNING PROCESS OF FDICA-BASED SPEECH SEPARATION

Masahiro FURUKAWA, Yusuke HIOKA, Takuro EMA and Nozomu HAMADA

Signal Processing Lab., School of Integrated Design Engineering, Keio University
Hiyoshi 3-14-1, Kohoku-ku, Yokohama, Kanagawa, 223-8522 Japan
{furukawa,hioka,ema,hamada}@hamada.sd.keio.ac.jp

ABSTRACT

The blind source separation for speech using frequency-domain independent component analysis(FDICA) is considered. As a source separation system, Saruwatari *et al.*[1] proposed a method by integrating the independent component analysis (ICA) and array signal processing. In this paper, we introduce the following two techniques into the learning process of the method[1]. (1)Classification of acquired array signals with respect to the number of speakers. (2)Direction-of-Arrival(DOA) estimation for each speaker using the intervals(frames) which are classified into single-speaker frame. Through some experiments, we can confirm that these techniques are effective to guarantee the convergence to the global optimal solution in the learning process.

1. INTRODUCTION

The speech-based human-machine interface is required in various actual environments. The technology, which actualizes the hearing of human and the ability to extract target sound from the midst of many sound sources is expected as a pre-processing of various audio applications.

A recent interesting technique is the blind source separation scheme[2]. The blind source separation for speech using FDICA[3] is considered in this paper. The FDICA system consists of series connection of short term Fourier transform(STFT), source separation system (or unmixing system) and inverse STFT. At the windowed STFT, acquired microphone array signals are separated into those with fixed frame(time) length. As the source separation system, some interesting methods by integrating the ICA and array signal processing are proposed. Saruwatari *et al.* used the coefficients of null-beamforming[4] filter as the initial weight parameters in the unmixing system[1]. Parra *et al.* proposed to constrain the ICA learning system using the array geomatric information[5]. These ideas are effective to guarantee the convergence to the global optimal solution in the learning process. In this paper, we consider the following two subjects in the FDICA system.

1. The learning process to determine the parameters of the unmixing system is solely applied to the frames at which two source signals exist. This may bring fast convergence in the learning process.
2. Initial parameters in unmixing system are necessary.

Two novel techniques are introduced in this paper. First of all, we perform a simple power based voice activity detection to separate non-speech frames. The next step is to classify the acquired signal frames into two-speakers signal frame or one-speaker signal frame. The harmonic structure of power spectrum in speech is used in this discrimination. Because the harmonic structure of speech signal is distinctly observed only in the single-speaker frames, we select the frames whose spectral structure is non-harmonic as the two-speakers frames. Then, the learning process in ICA, or parameter updating ICA process, is applied to the frame at which two source signals exist. For the second subject, single-speaker frames are used for estimating the DOA of each speaker. The coefficients of the null-beamforming filters for the obtained directions are used as the initial unmixing matrix parameters. In the FDICA, the post-processing of permutation and scaling resolution is essential to ensure the separation[3]. Some studies for this subject achieve sufficient performance by taking the geometric information of the microphone array[6][7].

We show the effectiveness of the proposed system by applying it to speech acquired in a meeting room. The proposed system has marked better separation results compared to the conventional methods.

2. PROPOSED METHOD

2.1. Problem Settings

In this study, we assume the 2-speakers and 2-microphones model as shown in Fig.1. For FDICA problem, the short-time Fourier transforms of the input signals are given by the complex valued instantaneous mixture denoted as

$$X_1^{(m)}(\omega) = A_{11}(\omega)S_1^{(m)}(\omega) + A_{12}(\omega)S_2^{(m)}(\omega) \quad (1)$$

$$X_2^{(m)}(\omega) = A_{21}(\omega)S_1^{(m)}(\omega) + A_{22}(\omega)S_2^{(m)}(\omega), \quad (2)$$

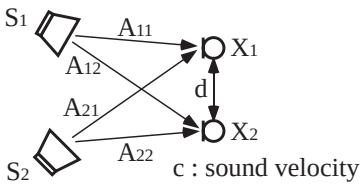

Fig. 1. 2-speakers and 2-microphones model

where $A_{ij}(\omega)$ is time-invariant transfer function between j-th speaker to i-th microphone and $S_j^{(m)}(\omega)$ is the speech of j-th speaker in the Fourier domain, respectively, and m is the frame index. The frame length is taken around 30ms as it is taken in usual speech spectrogram analysis. Our goal is to separate original speech $S_i(i = 1, 2)$ from the observed mixed signals.

2.2. Ideas in the proposed method

As far as the number of speakers is assumed less than two, we roughly classify each frame of the received signal into three types; (i)double talk segment, (ii)single talk segment and (iii)silent segment. From the discussions in Sec.1, only the type(i) segment should be utilized in the ICA learning. The type(ii) segment is suitable for the DOA estimation to determine the initial parameter of unmixing system. For these purposes, we set a frame analyzing stage in the conventional FDICA method as shown in Fig.2. Furthermore, an initial parameter is derived by the DOA estimated in the type(ii) segment.

Fig.3 shows the flow diagram of the procedure and the basic ideas for each stage are summarized below.

1. **Voice activity detection using the power localization**
   Because speech signal component is temporally isolated at particular frames, such tupe(iii) segments can be extracted by applying the voice activity detection based on the signal's power. These segments are deleted.

2. **Extraction of the type(i) segment based on the harmonic structure**
   It is well-known[8] that harmonic structure distinctly appears in the spectrum of voiced sound. Due to the frame length, voiced sound from each speaker has a harmonic structure. But in particular cases that a frame includes the boundary of two voiced phonemes, they are mixed and the harmonic structure is not distinctly found eventually. So we decide the frames without distinct harmonic structure as the type(i) segment.

3. **Determination of the type(ii) segment based on DOA**
   The rest of the segments after the previous process are still the mixture of type(i) and type(ii), because some double talk frames may contain distinct harmonic structure in their spectra. We measure the difference between type(i) and type(ii) using the DOA information.

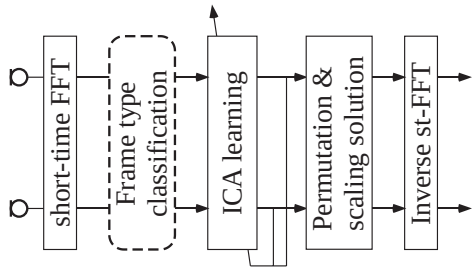

Fig. 2. FDICA with frame analysis

After the segment classification, the ICA learning is applied only to type(i) frame using the initial values determined by the DOA estimated from type(ii) frame. In the following Sec.2.3, we explain the procedures mentioned above.

2.3. Procedures

2.3.1. Voice activity detection

For the detection of voice activity frame number set $\mathbf{m}$, we take a simple power based discrimination given by

$$\mathbf{m} = \{m : P(m) \geq Th\} \quad (3)$$

$$P(m) = 10 \log_{10} \sum_{\omega} |X_1^{(m)}(\omega)|^2, \quad (4)$$

where Th is a threshold determined by the rule of thumb. In the following process, we deal with only the frames in $\mathbf{m}$.

2.3.2. Extraction of double talk segments

To extract double talk segments using the harmonic structure, we first estimate the pitch frequency ω_0 by the logarithmic harmonic product spectrum(LHPS) defined as

$$\omega_0^{(m)} = \arg\max_{\omega} \{C^{(m)}(\omega)\} \quad (5)$$

$$C^{(m)}(\omega) = 10 \sum_{k=1}^{K} \log_{10} |X_1^{(m)}(k\omega)|^2. \quad (6)$$

If the harmonic structure is distinct, a prominent peak appears in the LHPS. For signal with more than one voiced sound, in contrast, the LHPS shows some major peaks. From this fact, we extract the double talk segments by the following manner.

$$\mathbf{m}_D = \left\{ m : \left(C_1^{(m)}(\omega_0) - C_2^{(m)}(\omega_0') \right) < 3 \right\} \cap \mathbf{m} \quad (7)$$

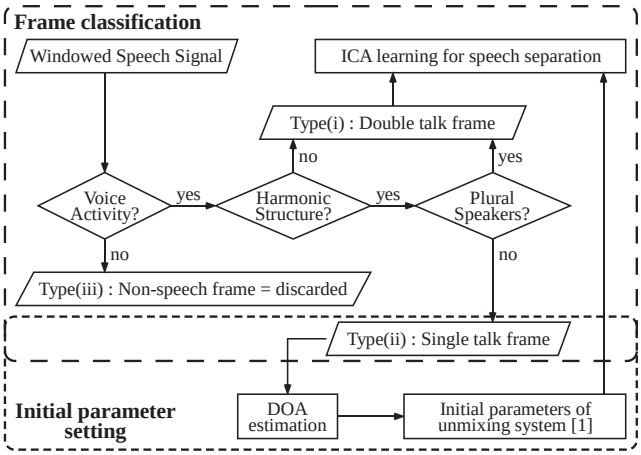

Fig. 3. Data flow in the frame analysis

Here $C_i^{(m)}$ is the i-th largest local maximum in the LHPS of the m-th frame.

2.3.3. Determination of single talk segments

Whether a segment is classified into single talk or not is determined by applying DOA estimation to the complement of $\mathbf{m}_D$, i.e. $(\mathbf{m}'_D \cap \mathbf{m})$. First of all, we calculate the average of estimated DOA $\bar{\theta}_A^{(m)}$ (group A) and $\bar{\theta}_B^{(m)}$ (group B), on the harmonic elements of ω_0 and the other peaks, respectively. For the DOA estimation, the beamforming method is adopted due to its simplicity and robustness to the noise. The estimated angle $\bar{\theta}$ is given by

$$\bar{\theta}^{(m)} = \arg\max_{\theta} \{P_{BF}^{(m)}(\theta)\}, \tag{8}$$

where,

$$\begin{aligned} P_{BF}^{(m)}(\theta,\omega) &= \frac{\mathbf{a}_\omega^H(\theta) R_{xx}^{(m)}(\omega) \mathbf{a}_\omega(\theta)}{\mathbf{a}_\omega^H(\theta)\mathbf{a}_\omega(\theta)} \\ \mathbf{a}_\omega(\theta) &= \left[\begin{array}{cc} \exp(-j\omega\tau_1) & \exp(-j\omega\tau_2) \end{array}\right]^T \\ R_{xx}^{(m)}(\omega) &= \left[\begin{array}{cc} |X_1^{(m)}(\omega)|^2 & X_1^{(m)}(\omega) X_2^{(m)*}(\omega) \\ X_1(\omega)^{(m)*} X_2(\omega) & |X_2^{(m)}(\omega)|^2 \end{array}\right], \end{aligned}$$

$\tau_1 = 0$ and $\tau_2 = \frac{d \sin\theta}{c}$ are the factors for the delay compensation. Comparing the difference between these average DOAs, we determine the single talk segments.

$$\mathbf{m}_S = \{m : |\bar{\theta}_A^{(m)} - \bar{\theta}_B^{(m)}| < 10\} \cap \mathbf{m} \tag{9}$$

2.3.4. Initial value setting based on DOA

The initial value of the ICA learning is obtained by the estimated DOA. We calculate the initial unmixing system parameters by the null-beamforming method[1] with setting its null direction to the mode of the estimated direction in group A, i.e. $\bar{\theta}_A^{(m)} (m \in \mathbf{m}_S)$.

Table 1. Parameter settings for the experiments

Sampling Frequency	8 kHz
Frame length	256 samples
Frame overlap	128 samples
DFT points	512
Microphone distance d	0.04 m
Stepsize μ	0.004
No. of iteration	500
Power ratio of two speeches	0 dB

Table 2. DOAs for computing initial values in conventional method

	Source 1	Source 2
Conventional(1)	$60°$	$-60°$
Conventional(2)	$45°$	$-45°$
Conventional(3)	$30°$	$-30°$

3. EXPERIMENTAL EVALUATION

We performed some experiments in a real acoustic environment for performance evaluation. The parameter settings are summarized in Tab.1. The conventional methods for comparison are the method[1] with different DOAs for computing initial values as shown in Tab.2. For the objective evaluation criteria, we use the average of SIR defined in [9]. Fig.4 and Fig.5 show the results for different combinations of speakers and DOAs, respectively. For the cases A to F in Fig.4, we use two speeches of two males and females, and note that the sources adopted in case G are same data used in [1]. The DOA combinations are given in Tab.3. In each case, we performed 4 trials and took the average of the results. Through the experiments, the proposed method is shown to give higher performance than that of the conventional method. From Fig.4, we can find that the proposed method is robust to the speech difference among individuals. On the other hand, Fig.5 shows that the performance of the conventional method is highly deteriorated in the case of $3, 6, 7, 9, 10$ and 11, even though the proposed method keeps its better performance. In these cases, speakers' locations are asymmetric with respect to the broadside and therefore the initial values in Tab.2 are not appropriate for the conventional method. In contrast, the proposed method is not affected by such factors owing to the DOA based initial value.

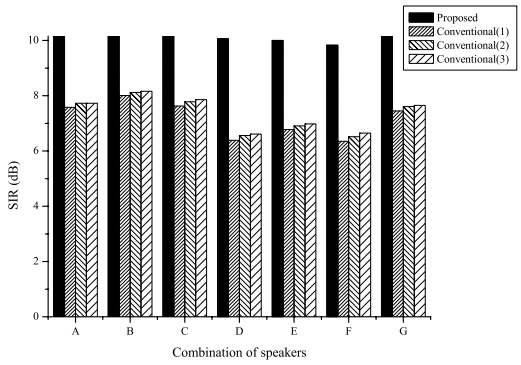

Fig. 4. Results for different speaker combination

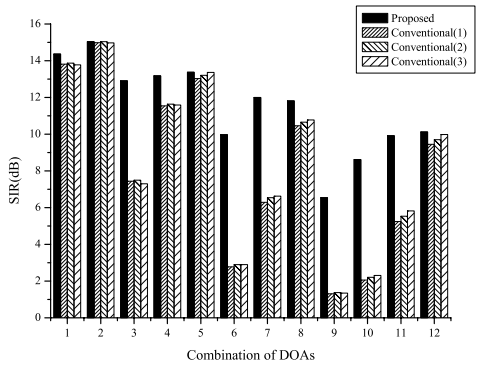

Fig. 5. Results for different DOA combination

4. CONCLUDING REMARKS

In this contribution, we proposed a new mechanism based on the speech signal features to the learning process of the FDICA-based speech separation. At the first step of the proposed method, each frame of the received signal is classified based on the number of speakers. Then the double talk segments are adopted to the ICA learning, on the other hand, the DOA estimation is performed on the single talk segments. We confirmed the efficiency through the experiments in a real acoustic environment.

5. ACKNOWLEDGEMENT

This work is supported in part by a Grant in Aid for the 21st century Center Of Excellence for Optical and Electronic Device Technology for Access Network from the Ministry of Education, Culture, Sport, Science, and Technology in Japan.

6. REFERENCES

[1] H. Saruwatari, T. Kawamura, T. Nishikawa and K. Shikano, "Fast-Convergence Algorithm for Blind Source Separation Based on Array Signal Processing," IEICE Trans. Fundamentals, Vol.E86-A, No.3, pp.286–291, Mar. 2003.

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

[3] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Neurocomputing, vol.22, pp.21–34, 1998.

[4] D. H. Johnson and D. E. Dudgeon, "Array Signal Processing," PRENTICE HALL, 1993.

[5] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," IEEE Trans. on Sppech and Audio Processing, vol.10, no.6, pp.352–362, Sep. 2002.

[6] S. Kurita, H. Saruwatari, S. Kajita, K.Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pettern under reverberant conditions," Proceedings of ICASSP2000, vol.5, pp.3140–3143, 2000.

[7] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and pricise method for solving the permutation problem of frequency-domain blind source separation," Proceedings of ICA2003, pp.505–510, 2003.

[8] S. Furui,"Digital Speech Processing, Synthesis and Recognition" Marcel Dekker Inc., 1989.

[9] R. Mukai, S. Araki, H. Sawada and S. Makino, "Removal of Residual Cross-talk Components in Blind Source Separation using LMS Filters," IEEE NNSP2002, pp.435-444, Sep. 2002.

Table 3. DOA combination

Case	1	2	3	4	5	6
θ_1[deg]	60	45	60	45	30	60
θ_2[deg]	-30	-45	0	-15	-30	15
Case	7	8	9	10	11	12
θ_1[deg]	45	30	60	45	30	15
θ_2[deg]	0	-15	30	15	0	-15