14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

A METHOD FOR SOLVING THE PERMUTATION PROBLEM OF FREQUENCY-DOMAIN BSS USING REFERENCE SIGNAL

Takashi Isa, Toshiyuki Sekiya, Tetsuji Ogawa and Tetsunori Kobayashi

Department of Computer Science, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

ABSTRACT

This paper presents a method for solving the permutation problem. This is a problem specific to frequency domain blind source separation within the framework of independent component analysis. Towards this problem, we propose a method which uses reference signals. For each frequency bin, the permutation alignment is fixed by calculating correlation coefficients between the reference signal and the separated signal. Reference signals are obtained as signals corresponding to each individual original sources. The reference signals are chosen or obtained subjectively, and do not need to be separated well. For example, the conventional beamforming technique gives suitable reference signals. To show the effectiveness of this method, we conducted a experiment of continuous speech recognition in a real room. The experimental results of double talk recognition with 20 K vocabulary show that the proposed method is effective to achieve 20 % error reduction rate compared with the established DOA-based approach.

1. INTRODUCTION

Multi-talk recognition is indispensable to realize various applications of hands free speech recognition, for example, conversation systems such as a humanoid robot, dictation systems of a meeting, interfaces of car-navigation systems.

Recently, blind source separation (BSS) within the framework of independent component analysis (ICA) has been studied actively as one of the approaches for speech segregation or enhancement. The difficulty of separating a mixed speech signal is due to the delays and reflections of the ambient environment. The recorded signals are no longer instantaneous mixtures but convoluted mixtures. An approach toward convoluted mixture is to transform time signals into time-frequency signals using windowed Fourier Transform in order to employ the complex-valued instantaneous ICA algorithm. The merit of this approach is that the ICA algorithm becomes simple and can be separated for each frequency. Also, any complex-valued instantaneous ICA algorithm can be employed with this approach.

However, BSS in the frequency domain includes the so called permutation problem, caused by the permutation ambiguity of the ICA solution. This problem affects the speech segregation performance seriously, and it is necessary to align the permutation precisely for each frequency.

Various methods has been proposed for solving the permutation problem. One approach is using the property of inter-frequency correlations of output signal envelopes [1]. In this approach, it is known that misalignment is collected consecutively after failing to align precise permutation for one frequency. Another is based on direction of arrival (DOA) estimation from the ICA solution [2]. DOAs are estimated at each subband and clustered so as to correspond to each original signal. This is thought not to use inter-frequency continuity of speech signal. Therefore, preciseness of DOA estimation at a frequency influences the permutation alignment. To improve the permutation alignment, the method using both DOA and inter-frequency correlations has been proposed [3]. However, there is no experiment in the case that the number of microphones is larger than that of speakers.

We propose a new method by taking advantage of the correlation between reference signals and estimated original sources for each frequency. The reference signals are obtained corresponding to original components. The permutation is chosen so that its alignment gives a correlation as large as possible. The reference signal is obtained suitably for the problem and not needed to be separated completely. We apply beamforming and time-frequency masking to acquire reference signals. In this way, the permutation ambiguity is removed from the separated speech.

In the following section 2, formulation of the BSS is described. In section 3, the definition of the reference signal and the algorithm of the proposed method is described in detail. In section 4, conditions and results of a continuous speech recognition experiment are described. We give the conclusion in section 6.

2. BLIND SOURCE SEPARATION IN THE FREQUENCY DOMAIN

2.1 Formulation of the sound

We assume the environment where S sound sources exist and the sound field is observed by M microphones. We define the vector $\mathbf{x}(\omega, t)$ as the STFT coefficient of the input signal.

$$\mathbf{x}(\omega,t) = [X_1(\omega,t),\ldots,X_M(\omega,t)]^T$$

$X_i(\omega,t)$ denotes the STFT coefficient at i-th microphone. ω and t denote the discrete frequency and frame index respectively. The operator $[\cdot]^T$ the transposition.

Using the transfer function, $\mathbf{x}(\omega,t)$ is written as below.

$$\begin{aligned}\mathbf{x}(\omega,t) &= \mathbf{A}(\omega)\,\mathbf{s}(\omega,t)+\mathbf{n}(\omega,t) \quad (1)\\ \mathbf{A}(\omega) &= [\boldsymbol{a}_1(\omega),\cdots,\boldsymbol{a}_S(\omega)]\\ \mathbf{s}(\omega,t) &= [s_1(\omega,t),\cdots,s_S(\omega,t)]^T\\ \mathbf{n}(\omega,t) &= [N_1(\omega,t),\cdots,N_M(\omega,t)]^T\end{aligned}$$

The symbol $\boldsymbol{a}_j(\omega)$ denotes the impulse response converted into the time-frequency domain from the j-th source to the microphones at discrete frequency ω. $\mathbf{s}$ is the time-frequency representation of the source signals. $s_j(\omega,t)$ denotes the spectrum of j-th source. To simplify the expression, we omit the symbol ω and t. This shows that a convoluted mixture is transformed into a simple instantaneous mixture for a fixed ω.

2.2 Preprocessing

It is essential for good performance of ICA that the signal is preprocessed [4]. Especially, when the number of sources S is different than that of microphones M, it is indispensable to preprocess the inputs to use all information of the observed signal. Subspace method is applied to preprocess in this study [6]. The spatial correlation matrix $\mathbf{R}=E\left[\mathbf{x}\mathbf{x}^H\right]$ is decomposed into the signal subspace and the noise subspace. The subspace filter $\mathbf{W}$ is defined as below. Using this filter $\mathbf{W}$, the input signal is preprocessed.

$$\begin{aligned}\mathbf{W} &= \sqrt{\boldsymbol{\Lambda}_s}^{-1}\mathbf{E}_s^H \quad (2)\\ \mathbf{y} &= \mathbf{W}\mathbf{x} \quad (3)\end{aligned}$$

The symbol $\mathbf{E}_s$ and $\boldsymbol{\Lambda}_s$ is eigenvector and eigenvalue respectively. They represent signal subspace, and correspond to the S largest eigenvalues chosen. The input $\mathbf{x}$ is preprocessed by subspace filter in this way. In the case with the number of microphones equivalent with that of sound sources $M=S$, the subspace filter is replaced by the PCA filter $\mathbf{W}=\sqrt{\boldsymbol{\Lambda}}^{-1}\mathbf{E}^H$.

2.3 ICA

A solution of complexed-value ICA, $\mathbf{U}$ is obtained so that the components of the reconstructed signals

$$\mathbf{z}=\mathbf{U}\mathbf{y} \quad (4)$$

are mutually independent. In this paper, We use JADE (joint approximate diagonalization of eigen matrices) extended to complex values [7]. For the sake of convenience, the product of subspace filter $\mathbf{W}$ and $\mathbf{U}$ is defined as separation filter $\mathbf{B}$.

$$\mathbf{B}=\mathbf{U}\mathbf{W} \quad (5)$$

2.4 Scaling problem

We expect $\mathbf{B}$ to be the inverse of $\mathbf{A}$, but we lack the information of amplitude and order of source signals. So there remains indefiniteness of permutation and scaling factors. The output of the separation filter must be processed with the permutation matrix $\mathbf{P}$ and the scaling matrix $\mathbf{D}$. The scaling matrix $\mathbf{D}_m$ is a $S\times S$ diagonal matrix represented as follows

$$\mathbf{D}_m=\mathrm{diag}[B^+_{m1},\ldots,B^+_{mS}]. \quad (6)$$

B^+_{ms} is the (m,s)-th element of the Moore-Penrose pseudoinverse of $\mathbf{B}$. The signal processed by $\mathbf{S}_m$ is S estimates of sources observed at the m-th microphone [6].

However, the problem of resolving the permutation matrix $\mathbf{P}$ is still open. We describe how to solve the permutation problem in the section that follows.

3. PROPOSED METHOD

3.1 What is the reference signal?

The reference signals correspond to each of the individual original sources. The reference signals are roughly separated from observed mixture with a different process than ICA. The reference signal does not need to be separated thoroughly. We expect that the reference signal meets two conditions as below. First, the reference signal correlates with the original source properly. Second, if the original source is the same, the envelope of source signal and reference signal correlates even for different frequencies. If the signal has these two properties, it can be the reference signal.

3.2 Detail algorithm for solving the permutation problem based on the reference signal

We assume that the reference signals have been obtained by another process which differs from ICA. We described the way of producing the reference signals in the next subsection.

We define the separated signal corresponding to the j-th source and observed at the i-th microphone in discrete frequency ω as $Z_m(\omega;j)$. Similarly, We define the reference signal as $R(\omega;j)$. We use the envelope estimator $\mathcal{E}$ as

$$\mathcal{E}Z(\omega;j)=\frac{1}{M}\sum_{i=1}^{M}|Z_i(\omega;j)|. \quad (7)$$

We define the correlation of the two signals $a(t)$ and $b(t)$ as below

$$\mathrm{cor}(a,b)=\frac{\mu_{a\cdot b}-\mu_a\cdot\mu_b}{\sigma_a\cdot\sigma_b} \quad (8)$$

where μ_a and σ_a denotes the mean and standard deviation of a respectively.

At a frequency ω_l, the permutation Π_{ω_l} is determined so that the summation of correlation between the envelope of the separated signals and that of the reference signals. It is realized by the equation following.

$$\Pi_{\omega_l}=\underset{\Pi}{\mathrm{argmax}}\sum_{i}^{M}\mathrm{cor}\Big(\mathcal{E}R(\omega_l;i),\mathcal{E}Z(\omega_l;\Pi(i))\Big) \quad (9)$$

where, the symbol $\Pi(i)$ is conversion as permutation such as $\Pi(1)=2,\ \Pi(2)=1$ or $\Pi(1)=1,\ \Pi(2)=2$.

Furthermore, we modify the equation (9) on the basis of the condition of the reference signals, that is, if the original source is the same, the envelope of source signal and reference signal correlates even for different frequencies. The equation (9) become bellow.

$$\Pi_{\omega_l}=\underset{\Pi}{\mathrm{argmax}}\sum_{i}^{M}\sum_{|\omega_l-\omega'|\le\delta}\mathrm{cor}\Big(\mathcal{E}R(\omega';i),\mathcal{E}Z(\omega_l;\Pi(i))\Big) \quad (10)$$

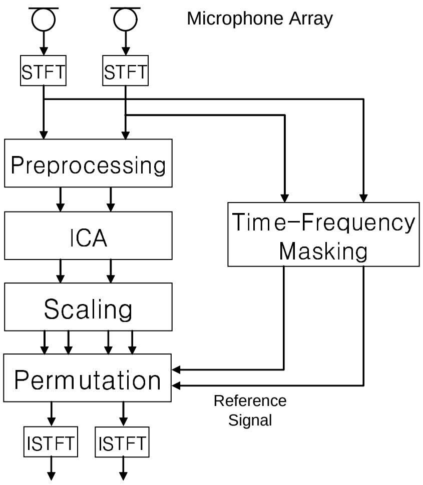

Figure 1: Diagram of the proposed method using time-frequency masking to synthesize the reference signal.

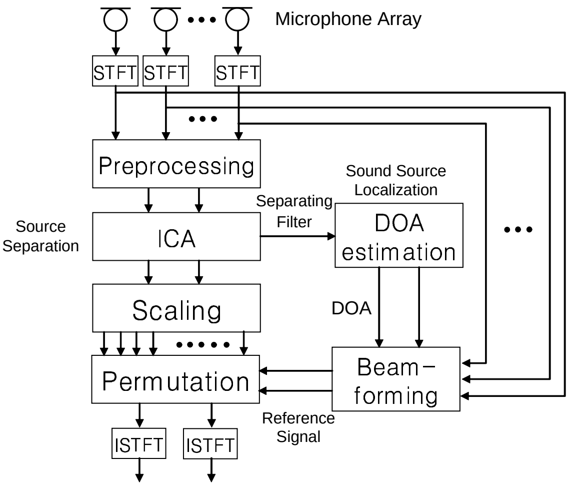

Figure 2: Diagram of proposed method using beamforming to make the reference signal.

The symbol ω' expresses a neighboring frequency to ω. Our preliminary experiment showed the results that equation (10) gave better permutation than equation (9). For all ω, the permutation is aligned by realizing above. Output signals by ICA is permuted such as follows.

$$Z_j(\omega_l; i) = Z_j(\omega_l; \Pi_{\omega_l}(i)). \tag{11}$$

As a result, the permutation ambiguity can be solved. The separated spectrogram is converted into the time domain signal by applying the inverse Discrete Fourier Transform (IDFT) to $Z_j(\omega; i)$

3.3 How to make the reference signal

In this subsection, we discuss the method of producing the reference signals. It is sufficient to use the conventional method in order to make the reference signals.

We assume two cases, in which there are enough microphones to apply beamforming technique and not. Concretely speaking, we process in the case of the number of microphones of two and eight ($M = 2, 8$) when the number of sources is two ($S = 2$). In this work, we implement two techniques to get the reference signal for each different case.

3.3.1 Time-frequency masking

Figure 1 depicts the process using the time-frequency masking output as the reference signal. When there are two microphones, we utilize a basic binary mask such as that described in [8].

3.3.2 Beamforming

When there are eight microphones, we apply the beamforming technique. Especially, the modified minimum variance beamformer (modified MVBF) [5] is carried out. However, it is required to estimate the source localization to utilize modified MVBF. The direction of arrival (DOA) can be estimated by using the information of separating matrix $\mathbf{B}$ [3].

The distance from microphone to sound source is not given yet, but we obtain the modified MVBF filter on the assumption that the sound field is the near field. That is because that the preliminary experiment showed that the speech recognition performance does not depend on the distance from microphone to sound source. Additionally, the spatial correlation matrix is calculated by steering vectors.

Figure 2 shows the diagram of the proposed method using beamformer output as the reference signal.

4. EXPERIMENT

We applied the proposed method to the double-talk recognition and evaluated under the condition where the number of sources is given and room acoustics is unknon.

4.1 Experimental Setup

We recorded speech data to enable continuous speech recognition. The speech was sampled at 32 kHz. The microphone array consisted of eight omni directional microphones. Array form was linear with consistent spacing of 3 cm. We used center two microphones in the condition of using only two microphones. Figure 3 shows the recording condition. The reverberation time (RT) could be changed to 240 ms and 320 ms by drawing heavy curtains or not. The loudspeaker arranged in front of the microphone array was the target source. Another loudspeaker was the disturbance source and was moved to vary experimental conditions. Evaluation data was recorded for a total of four different conditions. As for the target utterances, we selected total 100 sentences spoken by 23 male speakers from ASJ-JNAS [9] continuous speech corpus. As for the disturbance utterances, we selected speech data spoken by different male speakers from ASJ-JNAS. Each utterance was adjusted to almost the same duration and energy. The SNR was almost 0 dB.

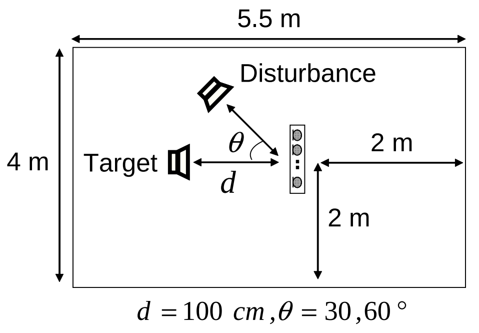

Figure 3: Recording condition.(We recorded evaluation data in a real room. The reverberation time could be changed by drawing curtains.)

4.2 Speech Processing

4.2.1 Methods for permutation

We evaluate two methods for solving the permutation problem; proposed method, a method based on estimating DOA [3].

The reference signal is obtained by two techniques; time-frequency masking (when there are two microphones) and beamforming (when there are eight microphones). In the beamforming method, the steering vectors were calculated considering direct sound not to rely on the room acoustics. We do not use the recorded transfer function to prevent dependence on room acoustics. Frame size and frame shift is 64 ms and 8 ms respectively.

Additionally, we use a signal recorded with only the target source as the reference signal. It is a cheating experiment to determine a true permutation as possible. We consider that this signal gives the ideal permutation alignment.

4.3 Speech recognition

Acoustic features are 12-dimensional MFCC and ΔMFCC and Δpower. The acoustic models are trained with 20 K sentences spoken by about 100 male speakers from ASJ-JNAS corpus. The training data is recorded with close-talk microphones. The language models are trigram language models using lexicon of 20 K vocabulary size. In this experiment, the speech data is sampled at 32 kHz, while the acoustic models are trained with speech data sampled at 16 kHz. Segregated speech is downsampled to 16kHz and converted to acoustic features.

4.4 Evaluation

We apply the frequency domain BSS to double-talk recognition under the condition that the number of sound sources is known. We compare two methods for solving the permutation problem; our proposal, method based on DOA and correlation [3].

With the proposed method, two kinds of reference signal is evaluated, one is the signal separated by modified MVBF described section 3.3. The other is target signal, which is recorded by the same environment without disturbance signal.

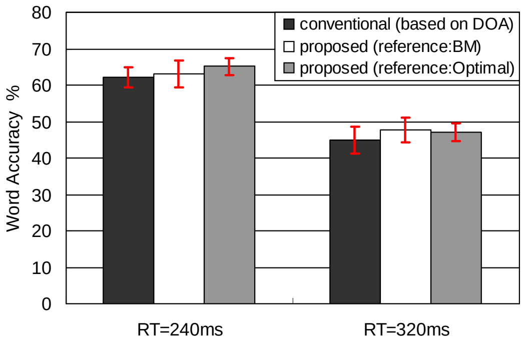

Figure 4: Results of continuous speech recognition in the case of using two microphones ($M = 2$). (BM is time-frequency masking output and optimal is the signal recorded with only target source. Each thick bar represents the average recognition performance in two experimental conditions. Line on the bar represents the maximum and minimum performance)

4.5 Results

Figure 4 shows the results of the experiment under the condition that there were two microphones. Through an approach based on DOA and our proposal, almost a precise permutation was obtained.

Our proposed method had about a three point advantage over conventional methods based on DOA where the reverberation time was long. Figure 5 shows the results of the experiment under the condition that there were eight microphones. The conventional method based on DOA performed insufficiently. The proposed method performed over 68 % and 57 % recognition rate in the reverberation time 240 ms and 320 ms respectively.

5. DISCUSSION

In this section, we discuss the advantages of proposed method over the method based on DOAs estimated with respect to each frequency.

Figure 6 and Figure 7 illustrate the results of DOA estimation respectively in the case that the source was placed at 0 degree and 60 degree. The *with confidence* and *without confidence* means whether estimate is reliable or not. These figures shows that preciseness of DOA estimation was not raised by eight microphones and subspace method. The standard deviation did not become smaller and the number of frequencies where DOA was reliable did not increase by eight microphones.

We consider the misalignment has two causes. First is the miss-clustering of DOAs. This was seriously effected by accuracy of the DOA estimation. To deal with this problem, we can tune the parameters and criteria which represents *confidence* through a trial and error process. Second is continuing to estimate the unreliable DOAs. These DOAs is useless to determine the permu-

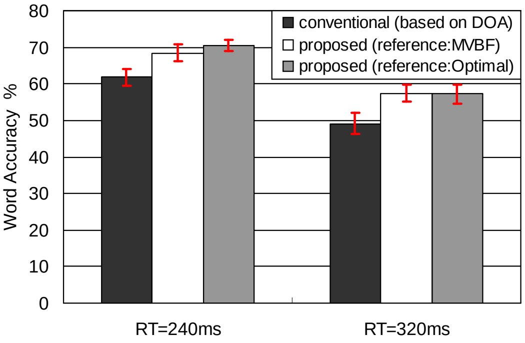

Figure 5: Results of continuous speech recognition in the case of using eight microphones ($M = 8$). (MVBF is modified minimum variance beam former output and optimal is the signal recorded with only target source. Each thick bar represents the average recognition performance in two experimental conditions. Line on the bar represents the maximum and minimum performance)

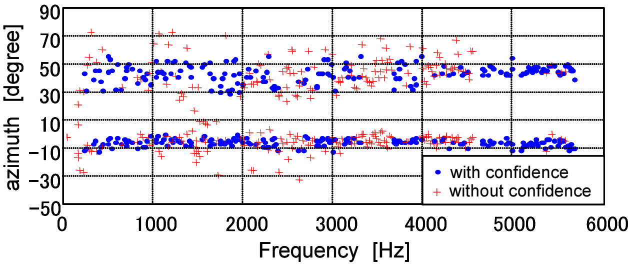

Figure 6: DOA estimates in the case of using two microphones.

tations. In the frequency which does not give permutation, unavoidly, conventional method must calculate the correlation between a frequency and neighboring frequencies in which the permutation is decided. However, if the reliable permutation is not fixed at neighboring bands, it is difficult to determine the precise permutation.

In our method, we used the mean of DOAs to obtain the reference signal. The mean of DOAs was more reliable than DOA at a band. The proposed method consider the correlation between the reference signal, which is obtained by using reliable DOA, and the separated signal at the same frequency. This process treats inter-frequency correlation effectively. Thus, the proposed method has advantages over the conventional method.

6. CONCLUSION

We proposed a method of solving the permutation problem, which is based on the reference signal. The reference signal can be obtained by incorporating the conventional technique. To show the effectiveness of this method, we conducted a experiment in a real room. The proposed method achieved about 70 % word accuracy in double-talk recognition of 20 K vocabulary. From the comparison of the DOA-based method, the advantage of the proposed method was shown.

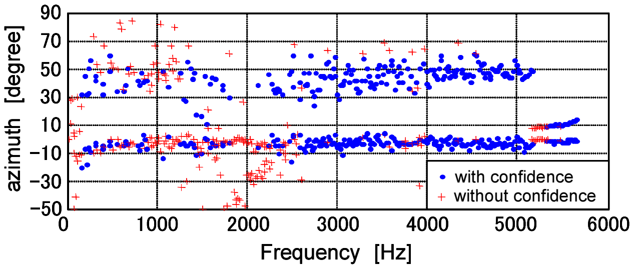

Figure 7: DOA estimates in the case of using eight microphones.

REFERENCES

[1] N. Murata et al., “An approach to blind source separation based on temporal structure of speech signals,” Neurocomputing, vol. 41, pp. 1–24, 2001.

[2] S. Kurita et al., “Evaluation of blind source separation method using directivity pattern under reververant conditions,” IEEE Proc. ICASSP2000, vol. 5, pp. 3140–3143, 2000.

[3] H. Sawada et al., “A Robust and precise method for solving the permutation problem of frequency-domain blind source separation,” IEEE Trans. ASSP, vol. ASSP-12, pp. 530–538, 2004.

[4] A. Hyvärinen et al., “Independent Component Analysis,” John Wiley, 2001.

[5] F. Asano et al., “Sound Source Localization and Separation in Near Field,” IEICE Trans. Fundamentals., vol. E83, pp. 2286–2294, 2000.

[6] F. Asano et al. “Speech enhancement using array signal processing based on the coherent-subspace method,” IEICE Trans. Fundamentals., vol. E80, pp. 2276–2285, 1997.

[7] J. F. Cardoso et al., “Blind beamforming for non Gaussian signals,” IEE Proc., vol. F140, pp. 362–370, 1993.

[8] Ö. Yilmaz et al., “Blind separation of speech mixture via time-frequency masking,” IEEE Trans. SP. vol. SP-52, pp. 1830–1847, 2004.

[9] K. Itou et al., “The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus,” in Proc. ISCA98 pp. 3261–3264.