A METHOD FOR SOLVING THE PERMUTATION PROBLEM OF FREQUENCY-DOMAIN BLIND SOURCE SEPARATION 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. The experimental results of double talk recognition with 20K 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. BSS is the problem of separating independent original sources from a mixed signal where the mixing process is unknown.

The difficulty of separating a mixture 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. 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. If the permutation problem is not solved, performance of sound segregation is insufficient. That is because a separated signal in the time domain contains frequency components from other source signals. 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 interfrequency 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]. This is a precise and robust method, however there is no experiments where the number of microphones is larger than that of the sources.

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 segregated 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 5.

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 input vector $\mathbf{x}(\omega, t)$ as the STFT coefficient of the input signal.

$$\mathbf{x}(\omega, t) = [X_1(\omega, t), \ldots, X_M(\omega, t)]^T \quad (1)$$

$x_m(\omega, t)$ denotes the STFT coefficient at microphone m. ω and t denote the discrete frequency and frame index respectively. Using the transfer function, $\mathbf{x}$ is written as below.

$$\mathbf{x}(\omega, t) = \mathbf{A}(\omega)\mathbf{s}(\omega, t) + \mathbf{n}(\omega, t) \quad (2)$$

where,

$$\mathbf{A}(\omega) = [\boldsymbol{a}_1(\omega), \cdots, \boldsymbol{a}_S(\omega)]^T \quad (3)$$
$$\mathbf{s}(\omega, t) = [S_1(\omega, t), \cdots, S_S(\omega, t)]^T \quad (4)$$
$$\mathbf{n}(\omega, t) = [N_1(\omega, t), \cdots, N_M(\omega, t)]^T \quad (5)$$

The symbol $\boldsymbol{a}_s(k)$ denotes the impulse response converted into the time-frequency domain from the s-th source to the microphones at discrete frequency ω. $\mathbf{s}$ is the time-frequency representation of the source signals. $s_s(\omega, t)$ denotes the spectrum of s-th source. $[\cdot]^T$ denotes the transposition. 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 [3]. Especially, when the number of sources S is different than that of microphones M, it is indispensable to preprocess the inputs.

We employ the subspace method [5] in preprocessing. At each frequency, the spatial correlation matrix $\mathbf{R}$ is defined as $\mathbf{R} = E[\mathbf{x}\mathbf{x}^H]$ where $[\cdot]^H$ denotes the Hermitian transpose. In the subspace method, we assume that $\mathbf{s}(t)$ and $\mathbf{n}(t)$ are uncorrelated. This assumption holds to some extent in a practical sense when the the length of impulse response is shorter than the STFT window length. Furthermore, assuming that $\mathbf{n}(t)$ is spatially white, we apply generalized eigenvalue decomposition $\mathbf{R} = \mathbf{K}\mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^{-1}$ on the assumption of $\mathbf{K} = \mathbf{I}$. The subspace filter $\mathbf{W}$ is defined as below. Using this filter $\mathbf{W}$, the input signal is preprocessed.

$$\mathbf{W} = \sqrt{\boldsymbol{\Lambda}_s}^{-1}\mathbf{E}_s^H \quad (6)$$
$$\mathbf{y} = \mathbf{W}\mathbf{x} \quad (7)$$

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

We separate the signals by processing with the solution of ICA at each frequency after preprocessing. A solution of complexed-value ICA, $\mathbf{U}$ is obtained so that the components of the reconstructed signals

$$\mathbf{z} = \mathbf{U}\mathbf{y} \quad (8)$$

are mutually independent. In this paper, We use JADE (joint approximate diagonalization of eigen matrices) extended to complex values, which is based on the 4th order cumulant [6]. For the sake of convenience, the product of subspace filter $\mathbf{W}$ and $\mathbf{U}$ is defined as separation filter $\mathbf{B}$.

$$\mathbf{B} = \mathbf{U}\mathbf{W} \quad (9)$$

2.4. Scaling problem

Ideally we expect $\mathbf{B}$ to be the inverse of $\mathbf{A}$, but we lack amplitude information of the source signals and their order. So there remains indefiniteness of permutation and scaling factors as below,

$$\mathbf{D}(\omega)\mathbf{P}(\omega)\mathbf{B}(\omega)\mathbf{A}(\omega) = \mathbf{I} \quad (10)$$

where $\mathbf{P}$ is a permutation matrix, i.e. all the elements of each column and row are 0 expect for one element with value 1, and $\mathbf{D}$ is a diagonal matrix. 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

$$\mathbf{D}_m = \mathrm{diag}[B^+_{m1}, \ldots, B^+_{mS}]. \quad (11)$$

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 [5].

However, the problem of resolving the permutation matrix $\mathbf{P}$ is still open.

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 has two properties 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.

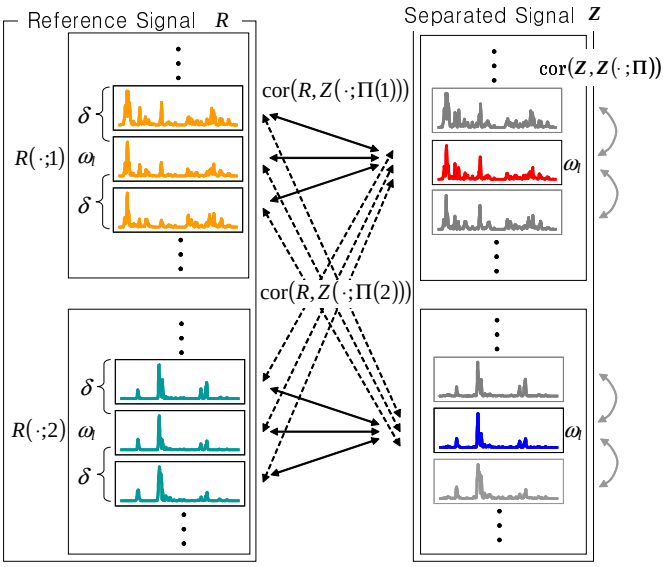

Fig. 1. Conceptual diagram to fix permutation alignment (Permutation is addressed with respect to each frequency. A Two-headed arrow means a correlation pair calculated. Solid line and dotted line mean different permutation alignment.)

Basically, we can make S (the number of sources) reference signals. However, we do not always have to prepare S reference signals. At least, we can get the separated signals solved the permutation problem by an increment of the reference signals.

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

In this subsection, we show the proposed method to fix permutation alignment in detail, that is, obtaining the permutation matrix $\mathbf{P}$ based on the reference signal. We assume that the reference signals have been obtained by another process which differ from ICA. We described the way of producing the reference signals in the next subsection.

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

$$\mathcal{E}Z(\omega; s) = \frac{1}{M}\sum_{j=1}^{M}|Z_j(\omega; s)|. \tag{12}$$

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

$$\mathrm{cor}(a, b) = \frac{\mu_{a\cdot b} - \mu_a \cdot \mu_b}{\sigma_a \cdot \sigma_b} \tag{13}$$

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

The practical step is described below.

1. We define the similarity $\mathrm{sim}(\omega)$ using the correlation. We consider that this measures how well the signal is separated from the observed mixture.

$$\mathrm{sim}(\omega) = \sum_{i\neq j}\mathrm{cor}\big(\mathcal{E}Z(\omega; i), \mathcal{E}Z(\omega; j)\big) \tag{14}$$

   Sort ω in order of weakness of correlation between independent components in ω

$$\mathrm{sim}(\omega_1) \leq \mathrm{sim}(\omega_2) \leq, \ldots, \leq \mathrm{sim}(\omega_{\max}) \tag{15}$$

   We process this order $\omega_1, \ldots, \omega_{\max}$ to fix the permutation alignment.

2. For ω_l, we resolve the permutation $\Pi_{\omega_l}(i)$ which maximizes the summation of correlation between the separated envelope of ω_l and the reference envelope at frequencies from $\omega_{l-\delta}$ to $\omega_{l+\delta}$. Additionally, when any permutation Π from $\omega_{l-\delta}$ to $\omega_{l+\delta}$ is already fixed, we consider the correlation between the envelope of ω_l and $\omega_{\mathcal{L}}$, where $\mathcal{L}$ denotes the set of frequencies whose permutation has already been fixed. This is achieved by maximizing the formula within all the possible permutations Π as described below.

$$\sum_{i=1}^{S}\Big\{\alpha\sum_{|l'-l|\leq\delta}\mathrm{cor}\Big(\mathcal{E}R\big(\omega_{l'}; i\big), \mathcal{E}Z\big(\omega_l; \Pi(i)\big)\Big) + \beta\sum_{\substack{|l'-l|\leq\delta \\ l'\in\mathcal{L}}}\mathrm{cor}\Big(\mathcal{E}Z\big(\omega_{l'}; i\big), \mathcal{E}Z\big(\omega_l; \Pi(i)\big)\Big)\Big\} \tag{16}$$

   Figure 1 shows the conceptual diagram of proposed method.

3. We assign the permutation to $Z_j(\omega_l; i)$ for all j to get the separated spectrogram.

$$Z_j(\omega_l; i) = Z_j(\omega_l; \Pi(i)) \tag{17}$$

Figure 2 shows the difference between a minimum and a maximum of formula (16) when the number of the separated signal and reference signal is two. We consider that it expresses the correlation between separated signal and reference signal can be a criterion to determine the permutation. 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 signal. It is sufficient to use the conventional method to make the reference signal.

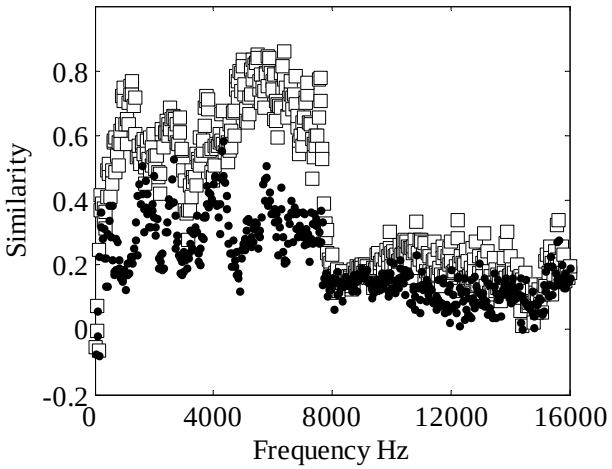

Fig. 2. Similarity between separated signal and reference signal. (Dots and squares are a minimum and a maximum of equation (16) respectively at each discrete frequency.)

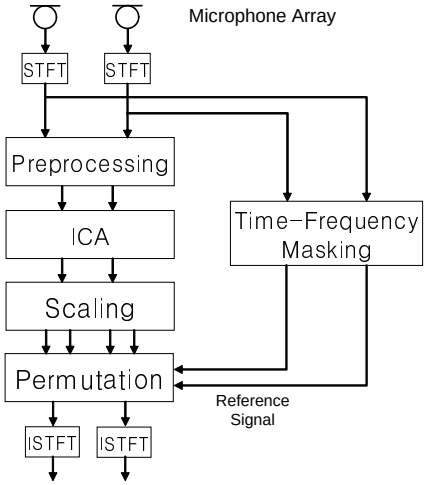

Fig. 3. Diagram of the proposed method using time-frequency masking to synthesize the reference signal.

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

When there are two microphones, we utilize a basic binary masking technique such as that described in [7]. The time-frequency mask is estimated using the sparseness of sources. We estimate DOA using phase difference between two observations.

$$\theta = \arccos \frac{\arg\left(\frac{X_i(\omega,t)}{X_j(\omega,t)}\right)}{\omega c^{-1} r} \quad (i \neq j) \tag{18}$$

We express microphone space as r and speed of sound as c in the equation above. The DOA histogram has S clusters and each cluster corresponds to one source. Using the average DOAs of these clusters $\theta_1, \ldots, \theta_S$, we define a time-frequency mask BM_k. We obtain the reference signal utilizing BM_k.

$$BM_k(\omega,t) = \begin{cases} 1 & (\theta_k - \xi \leq \theta \leq \theta_k + \xi\ ) \quad (19) \\ 0 & \text{otherwise} \quad (20) \end{cases}$$

$$R_k(\omega,t) = BM_k(\omega,t)\ X_j(\omega,t) \quad (j \in 1, \ldots, M)$$

In the equation (19), the symbol ξ is an extraction range parameter.

Figure 3 depicts the process using the time-frequency masking output as the reference signal.

3.3.2. Beamforming

When there are two microphones, we apply the beamforming technique. Especially, the modified minimum variance beamforming (modified MVBF) filter $\mathbf{F}_{s\ \mathrm{MV}}$ [4] was used.

$$\mathbf{F}_{s\ \mathrm{MV}} = \frac{\mathbf{R}^{-1}\boldsymbol{a}_s}{\boldsymbol{a}_s^H \mathbf{R}^{-1} \boldsymbol{a}_s} \tag{21}$$

$$\mathbf{R} = \mathbf{Q} + \gamma \mathbf{I} \tag{22}$$

$$\mathbf{Q} = \hat{\mathbf{A}}\mathbf{I}\hat{\mathbf{A}}^{\mathbf{H}} \tag{23}$$

In the equtation described above, matrix $\mathbf{I}$ denotes an unit matrix. The matrix $\hat{\mathbf{A}}$ is constructed by the location vectors. The vector $\boldsymbol{a}_s$ is the location vector corresponding to the s-th source signal.

However, it is required to estimate the source localization to utilize modified MVBF. The direction of arrival (DOA) is estimated by using the information of separating matrix $\mathbf{B}$[2].

$$\theta = \arccos \frac{\arg\left(\frac{\mathbf{B}^{+}_{m\Pi(i)}}{\mathbf{B}^{+}_{m'\Pi(i)}}\right)}{\omega c^{-1}(r_m - r_{m'})} \tag{24}$$

The symbol r_m is the location vector corresponding to m-th microphone. $[\cdot]^+$ denotes the Moore-Penrose pseudoinverse. We use mean value of θ as the DOA estimator.

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 shows that the speech recognition performance does not depend on the assumption of distance from microphone to sound source.

As a result of using the modified MVBF, we acquire

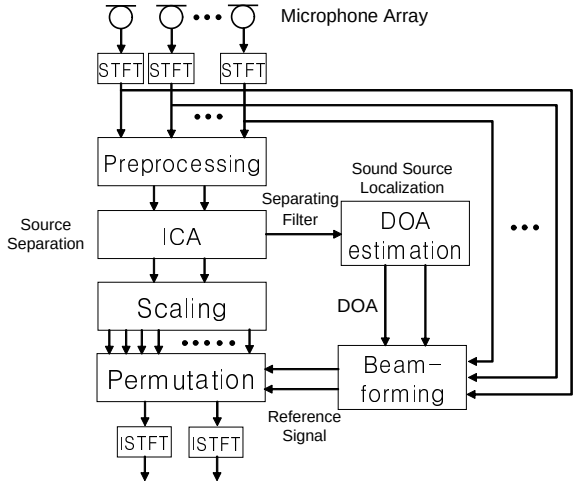

Fig. 4. Diagram of proposed method using beamforming to make the reference signal.

reference the signal $\mathbf{r}$ using $\mathbf{F}_{s\ \mathrm{MV}}$.

$$R_s(\omega,t) = \mathbf{F}^H_{s\ \mathrm{MV}}(\omega)\mathbf{x}(\omega,t) \quad (25)$$

$$\mathbf{r}(\omega,t) = [R_1(\omega,t),\ldots,R_S(\omega,t)]^T \quad (26)$$

Figure 4 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.

4.1. Experimental Setup

We recorded speech data to enable continuous speech recognition. Sampling Frequency is 32 kHz. Quantization is 16 bits. The microphone array consisted of eight omni directional microphones. Array form was linear with consistent spacing of 3cm. Figure 5 shows the recording condition. The reverberation time (RT) can be changed to 240ms and 320ms 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[8] 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.

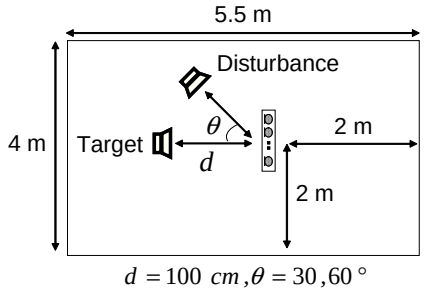

Fig. 5. Recording condition.(We recorded evaluation data in a real room. The reverberation time can be changed by drawing curtains.)

4.2. Speech Processing

4.2.1. Analysis condition

The STFT window is a Hamming window. Frame size and frame shift is 64ms and 8ms respectively. The separated signal in equation (8) and the reference signal are made up under this condition.

4.2.2. Methods for permutation

We evaluate two methods for solving the permutation problem; proposed method, method based on estimating DOA [2].

The reference signal is obtained by two techniques; time-frequency masking (when there are two microphones) and beamforming (when there are eight microphones). In this masking method, extraction range parameter in equation (19) is set to ten degree. In the beamforming method, the location vectors in equation (23) are calculated at intervals of five degree in the range of -90 to 90 degree. We assume that the distance from microphone to sound source is 150cm. The noise power parameter in equation (23) γ is set $\gamma = \|\mathbf{Q}\| \times 0.03$. The assumption of distance being unrelated to speech recognition performance was already stated. We do not use the recorded transfer function to prevent dependence on room acoustics. The confident parameter α and β in equation (16) is 0.9 and 0.1 respectively. The preliminary experiments shows that these parameters do not affect speech recognition rate.

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

The parameters of the acoustic features are as follows. Acoustic features are 12-dimensional MFCC and ΔMFCC and Δpower. Pre-emphasis is done with $1 - 0.97z^{-1}$. Frame

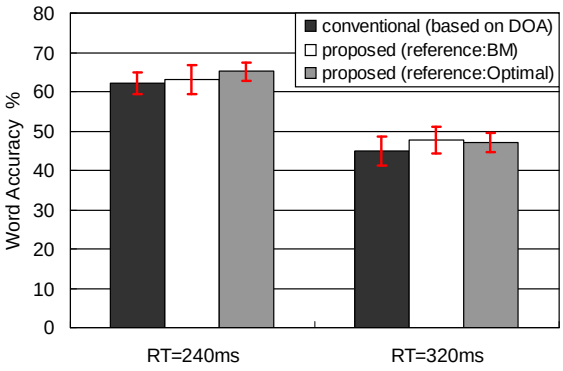

Fig. 6. Results of continuous speech recognition in the case of using two microphones ($M = 2$). (BM is time-frequency masking output and optimal is the signal recorded with only target source. Each thick bar represents the average recognition performance in two experimental conditions. Line on the bar represents the maximum and minimum performance)

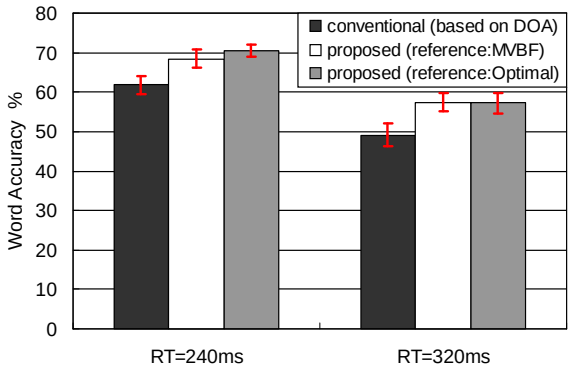

Fig. 7. Results of continuous speech recognition in the case of using eight microphones ($M = 8$). (MVBF is modified minimum variance beam former output and optimal is the signal recorded with only target source. Each thick bar represents the average recognition performance in two experimental conditions. Line on the bar represents the maximum and minimum performance)

length is 25ms and frame shift is 10ms. Window function is Hamming. 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 32kHz, while the acoustic models are trained with speech data sampled at 16kHz. Segregated speech is downsampled to 16kHz and converted to acoustic features.

4.4. Results

Figure 6 shows the results of the experiment under the condition that there are 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 7 shows the results of the experiment under the condition that there are eight microphones. The conventional method based on DOA performed insufficiently. The proposed method performed over 68% and 57% recognition rate in the reverberation time 240ms and 320ms respectively. This was a 20% error reduction compared with the established DOA-based approach.

5. CONCLUSION

We proposed the method of solving the permutation problem, which is based on the reference signal. 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. Our proposal achieved 20% error reduction rate compared with the DOA-based approach.

6. REFERENCES

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

[2] H. Sawada, R. Mukai, S. Araki, and S. Makino, “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.

[3] A. Hyvärinen, J. Karhunen, E. Oja, “Independent Component Analysis,” John Wiley, 2001.

[4] F. Asano, H. Asou, and T. Matsui, “Sound Source Localization and Separation in Near Field,” IEICE Trans. Fundamentals., vol. E83, pp. 2286–2294, 2000.

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

[6] Jean,F, C., and Souloumiac, A. “Blind beamforming for non Gaussian signals,” IEE Proc., vol. F140, pp. 362–370, 1993.

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

[8] K. Itou, K. Takeda, T. Takezawa, T. Matsuoka, K. Shikano, T. Kobayashi, M. Yamamoto, S. Itahashi,“The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus,” in Proc. ISCA Int Conf. Spoken, Language Processing, pp.3261-3264, Nov. 1998.