OVERDETERMINED BLIND SEPARATION FOR CONVOLUTIVE MIXTURES OF SPEECH BASED ON MULTISTAGE ICA USING SUBARRAY PROCESSING

Tsuyoki NISHIKAWA Hiroshi ABE Hiroshi SARUWATARI Kiyohiro SHIKANO

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192, JAPAN
E-mail: {tsuyo-ni, hiro-abe, sawatari, shikano}@is.aist-nara.ac.jp

ABSTRACT

We propose a new algorithm for overdetermined blind source separation based on multistage independent component analysis (MS ICA). To improve the separation performance, we have proposed MSICA in which frequency-domain ICA and time-domain ICA are cascaded. In the original MSICA, the specific mixing model, where the number of microphones is equal to that of sources, was assumed. However, additional microphones are required to achieve an improved separation performance under reverberant environments. This leads to alternative problems, e.g., a complication of the permutation problem. In order to solve them, we propose a new extended MSICA using subarray processing, where the number of microphones and that of sources are set to be the same in every subarray. The experimental results obtained under the real environment reveal that the separation performance of the proposed MSICA is improved as the number of microphones is increased.

1. INTRODUCTION

Blind source separation (BSS) is an approach for estimating original source signals only from the information of the mixed signals observed in each input channel. This technique is applicable to high-quality hands-free speech recognition systems. Many BSS methods based on independent component analysis (ICA) [1] have been proposed [2, 3] for the acoustic signal separation. However, the performances of these methods degrade particularly seriously under heavily reverberant conditions. In order to improve the separation performance, we have proposed multistage ICA (MSICA) [4], in which frequency-domain ICA (FDICA) [3, 5] and time-domain ICA (TDICA) [2] are cascaded (see Fig. 1). In this method, first, FDICA finds an approximate solution to separate the sources to a certain extent, and finally TDICA removes the residual crosstalk components arising in FDICA.

In the conventional ICA research, the specific mixing model is often assumed where the number of microphones is equal to that of sources. In the original MSICA, we also assumed this model and performed the source separation. However, additional microphones are required to achieve an improved separation performance because of the existence of the reflection and the reverberation component. In this paper, we set the number of microphones to be larger than that of sources and we extend the conventional MSICA into a new MSICA method using a large microphones. We point out that the following problems arise in the simple extension of MSICA: (1) the permutation problem [3] in FDICA part becomes very complicated, and (2) the solution of FDICA is likely to be trapped within a trivial solution. In this paper, as a robust method against these problems, we propose a new MSICA method using subarray processing, where the number of each subarray's microphones is set to be equal to that of the sources, and the outputs of FDICA performed in every subarray are weighted to be inserted into TDICA. The experimental results obtained under real acoustic conditions reveal that the separation performance of the proposed MSICA is improved over that of an original MSICA as the number of microphones is increased.

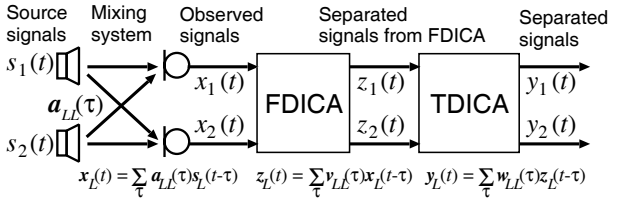

Fig. 1. Blind source separation procedure performed in original MSICA which has been previously proposed by the authors [4].

2. CONVENTIONAL MSICA AND PROBLEMS

2.1. Sound Mixing Model of Microphone Array

In general, the observed signals $\boldsymbol{x}_K(t) = [x_1(t), \cdots, x_K(t)]^{\mathrm{T}}$ in which multiple source signals $\boldsymbol{s}_L(t) = [s_1(t), \cdots, s_L(t)]^{\mathrm{T}}$ are convolved with room impulse responses (see Fig. 1) are obtained as $\boldsymbol{x}_K(t) = \sum_{\tau=0}^{P-1} \boldsymbol{a}_{KL}(\tau)\, \boldsymbol{s}_L(t-\tau)$, where K is the number of array elements (microphones) and L is the number of sound sources. Here, $\boldsymbol{a}_{KL}(\tau) = [a_{ij}(\tau)]_{ij}$ ($[\cdot]_{ij}$ denotes the matrix in which the ij-th element is $[\cdot]$) is the $K \times L$ mixing filter matrix, and P is the length of the impulse response.

2.2. BSS Algorithm Based on MSICA[4]

Figure 1 shows the procedure of the original MSICA. In the case of $K = L$, MSICA is conducted in the following steps. First, we perform FDICA [3, 5] to separate the source signals to some extent with the advantage of high stability. The output signals $\boldsymbol{z}_L(t) = [z_1(t), \cdots, z_L(t)]^{\mathrm{T}}$ from FDICA can be given as $\boldsymbol{z}_L(t) = \sum_{\tau=0}^{Q-1} \boldsymbol{v}_{LL}(\tau)\, \boldsymbol{x}_L(t-\tau)$, where $\boldsymbol{v}_{LL}(\tau) = [v_{ij}(\tau)]_{ij}$ is the separation filter matrix for FDICA, and Q is the length of the separation filter of FDICA. In FDICA, we optimize $\boldsymbol{v}_{LL}(\tau)$ so that the narrowband output signals are mutually independent at each frequency.

Second, we regard the output signals $\boldsymbol{z}_L(t)$ from FDICA as the input signals for TDICA, and we can remove the residual crosstalk components of FDICA by using TDICA. Finally, we regard the output signals from TDICA as the resultant separated signals. The separated signals $\boldsymbol{y}_L(t) = [y_1(t), \cdots, y_L(t)]^{\mathrm{T}}$ of MSICA can be given as $\boldsymbol{y}_L(t) = \sum_{\tau=0}^{R-1} \boldsymbol{w}_{LL}(\tau)\, \boldsymbol{z}_L(t-\tau)$, where $\boldsymbol{w}_{LL}(\tau)$ is the separation filter matrix for TDICA, and R is the length of the separation filter of TDICA. In TDICA, we optimize $\boldsymbol{w}_{LL}(\tau)$ so that the fullband separated signals are mutually independent.

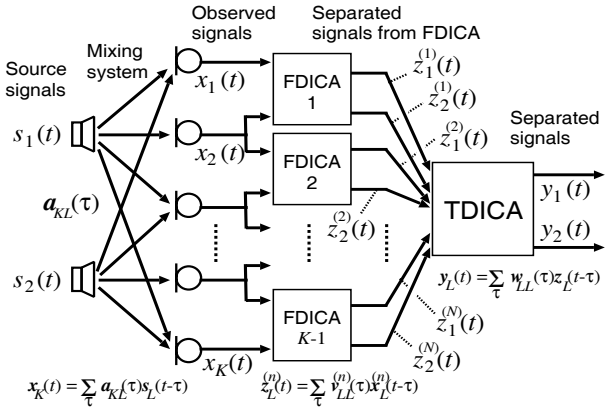

Fig. 2. Blind source separation procedure performed in the proposed MSICA using subarray processing.

2.3. Simple Extension of Conventional MSICA

In the conventional MSICA, the specific mixing model is assumed, where the number of microphones is equal to that of sources. However, additional microphones are required to achieve an improved separation performance because of the reflection and the reverberation component. Thus, we should set the number of microphones to be larger than that of sources (i.e., $K > L$), and we extend the conventional MSICA into a new MSICA method by using a large number of microphones. First, as the simple extension of MSICA, we consider the following two methods in the specific case of $K > L$.

[Method 1] The K output signals are obtained from FDICA and L separated signals are obtained from TDICA: $\boldsymbol{z}_K(t) = \sum_{\tau=0}^{Q-1} \boldsymbol{v}_{KK}(\tau)\, \boldsymbol{x}_K(t-\tau)$, $\boldsymbol{y}_L(t) = \sum_{\tau=0}^{R-1} \boldsymbol{w}_{LK}(\tau)\, \boldsymbol{z}_K(t-\tau)$. There is a permutation problem [3] of sources in every frequency bin in FDICA. By using recently proposed techniques [6, 7, 8], we can easily solve the problem only in the case of $K = L$. However, **(P1)** the permutation problem in FDICA becomes very complicated as the number of microphones is increased. Also, **(P2)** the discrimination of the output signals corresponding to the true sources is needed because there exist $K - L$ imaginary outputs. Therefore Method 1 is not applicable to separating sources in the real environment.

[Method 2] The L output signals are obtained from FDICA and the L separated signals are obtained from TDICA: $\boldsymbol{z}_L(t) = \sum_{\tau=0}^{Q-1} \boldsymbol{v}_{LK}(\tau)\, \boldsymbol{x}_K(t-\tau)$, $\boldsymbol{y}_L(t) = \sum_{\tau=0}^{R-1} \boldsymbol{w}_{LL}(\tau)\, \boldsymbol{z}_L(t-\tau)$. There still exist some problems as follows. **(P3)** In the iterative learning of FDICA, the solution is likely to be trapped within a trivial solution as described in Sect. 4.2. **(P4)** We cannot utilize all the information of the observed signals at K microphones in TDICA because the number of the input signals for TDICA is decreased to L by FDICA.

Due to these problems, a new extension algorithm of MSICA which is not affected by **(P1)–(P4)** is desired to achieve a superior separation performance. Therefore, in the next section we propose a new BSS algorithm based on the extended MSICA using subarray processing.

3. PROPOSED MSICA USING SUBARRAY PROCESSING

In the proposed extended MSICA, we regard the K observed signals as combinations of the $L(< K)$ observed signals, and we regard this combination as a subarray (see Fig. 2). First, we divide the whole inputs into $K - 1$ subarrays, and we perform FDICA in every subarray. The output signals $\boldsymbol{z}_L^{(n)}(t) = [z_1^{(n)}(t), \cdots, z_L^{(n)}(t)]^{\mathrm{T}}$ from FDICA in the n-th subarray can be given as $\boldsymbol{z}_L^{(n)}(t) = \sum_{\tau=0}^{Q-1} \boldsymbol{v}_{LL}^{(n)}(\tau)\, \boldsymbol{x}_L^{(n)}(t-\tau)$, where $\boldsymbol{v}_{LL}^{(n)}(\tau)$ is the separation filter matrix of FDICA in the n-th subarray and $\boldsymbol{x}_L^{(n)}(t) = [x_n(t), x_{n+1}(t), \cdots, x_{n+L-1}(t)]^{\mathrm{T}}$. As the FDICA algorithm for optimization of the separation filter $\boldsymbol{v}_{LL}^{(n)}(\tau)$, we introduce the fast-convergence FDICA proposed by one of the authors [5]. In the FDICA, the optimal $\boldsymbol{v}_{LL}^{(n)}(\tau)$ is obtained by the following iterative equation [3]:

$$\begin{aligned}\boldsymbol{V}_{LL}^{(n)}(f)_{i+1} \;=\; \alpha\Big[&\mathrm{diag}\Big(\big\langle \boldsymbol{\Phi}(\boldsymbol{Z}_L^{(n)}(f,m))\boldsymbol{Z}_L^{(n)}(f,m)^{\mathrm{H}}\big\rangle_m\Big)\\ &-\big\langle \boldsymbol{\Phi}(\boldsymbol{Z}_L^{(n)}(f,m))\boldsymbol{Z}_L^{(n)}(f,m)^{\mathrm{H}}\big\rangle_m\Big]\boldsymbol{V}_{LL}^{(n)}(f)_i\\ &+\boldsymbol{V}_{LL}^{(n)}(f)_i, \qquad (1)\end{aligned}$$

where $\boldsymbol{V}_{LL}^{(n)}(f)$ is a Fourier transform result of $\boldsymbol{v}_{LL}^{(n)}(\tau)$, $\boldsymbol{Z}_L^{(n)}(f,m)$ is the narrow-band output signal in the time-frequency domain and $\mathrm{diag}(\cdot)$ is the operation for setting every off-diagonal element of matrix as zero. Also, f is frequency, m is the analysis frame of short-time DFT, $\langle\cdot\rangle_m$ denotes the frame-averaging operator, i is used to express the value of the i-th step in the iterations, and α is the step-size parameter. We define the nonlinear vector function $\boldsymbol{\Phi}(\cdot)$ as

$$\boldsymbol{\Phi}(\boldsymbol{Z}_L(f,m)) \;\equiv\; \big[\Phi(Z_1(f,m)),\cdots,\Phi(Z_L(f,m))\big]^{\mathrm{T}}, \qquad (2)$$

$$\Phi(Z_l(f,m)) \;\equiv\; \tanh(Z_l^{(\mathrm{R})}(f,m)) + j\cdot\tanh(Z_l^{(\mathrm{I})}(f,m)), \qquad (3)$$

where $Z_l^{(\mathrm{R})}(f,m)$ and $Z_l^{(\mathrm{I})}(f,m)$ are the real and imaginary parts of $Z_l(f,m)$, respectively.

Next, we regard all output signals from FDICA in $K-1$ subarrays as the input signals for TDICA, and we remove the residual crosstalk components from FDICAs. The resultant separated signals $\boldsymbol{y}_L^{(n)}(t)$ can be given as $\boldsymbol{y}_L(t) = \sum_{\tau=0}^{R-1} \boldsymbol{w}_{LL\cdot(K-1)}(\tau)\, \boldsymbol{z}_{L\cdot(K-1)}(t-\tau)$, where $\boldsymbol{w}_{LL\cdot(K-1)}(\tau)$ is the $L \times L\cdot(K-1)$ separation filter matrix and

$$\begin{aligned}\boldsymbol{z}_{L\cdot(K-1)}(t) \;=\; [&z_1^{(1)}(t),\cdots,z_1^{(K-1)}(t),z_2^{(1)}(t),\cdots,\\ &z_2^{(K-1)}(t),\cdots,z_L^{(1)}(t),\cdots,z_L^{(K-1)}(t)]^{\mathrm{T}}. \qquad (4)\end{aligned}$$

In the TDICA, the optimal $\boldsymbol{w}_{LL\cdot(K-1)}(\tau)$ is obtained by the following iterative equation [9]:

$$\begin{aligned}\boldsymbol{w}_{LL\cdot(K-1)}(\tau)_{i+1} = \beta\sum_{d=0}^{R-1}\Big\{&\mathrm{diag}\Big(\langle\boldsymbol{\phi}(\boldsymbol{y}_L(t))\boldsymbol{y}_L(t-\tau+d)^{\mathrm{T}}\rangle_t\Big)\\ &-\langle\boldsymbol{\phi}(\boldsymbol{y}_L(t))\boldsymbol{y}_L(t-\tau+d)^{\mathrm{T}}\rangle_t\Big\}\boldsymbol{w}_{LL\cdot(K-1)}(d)_i\\ &+\boldsymbol{w}_{LL\cdot(K-1)}(\tau)_i, \qquad (5)\end{aligned}$$

where β is the step-size parameter, $\langle\cdot\rangle_t$ denotes the time-aver- aging operator, and $\boldsymbol{\phi}(\boldsymbol{y}_L(t)) \equiv [\tanh(y_1(t)),\cdots,\tanh(y_L(t))]^{\mathrm{T}}$.

We can easily solve the permutation problem by using the conventional methods [6, 7, 8] because the number of microphones is equal to that of sources in every subarray. Also, the discrimination of the output signals corresponding to the true sources is not required because the number of output signals from FDICA is equal to that of sources, i.e., there are no imaginary outputs. The separation filter of FDICA is likely to converge on the optimal point, particularly in the case of $K = L$ (see Sect. 4.2). Therefore, in the proposed MSICA, the problems **(P1)–(P3)** described in Sect. 2.3

do not arise. In addition, we can utilize the information of all the element of the microphone array in the TDICA because we use the output signals from FDICA in all subarrays with the information from all microphones. Therefore, **(P4)** is also solved by the proposed MSICA.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

A 14-element linear array with the interelement spacing of 2.83 cm is assumed. The speech signals are assumed to arrive from two directions, -40° and 20°. The distance between the microphone array and the loudspeakers is 2.0 m. Two sentences spoken by two male and two female speakers are used as the original speech samples and the sampling frequency is 8 kHz. Using these sentences, we obtain 12 combinations with respect to speakers and source directions. In these experiments, we use the following signals as the source signals: the original speech convolved with the impulse responses specified by the reverberation times of 300 ms. We use the impulse responses recorded in a real room selected from the Real World Computing Partnership (RWCP) sound scene database [10]. In order to evaluate the performance, we used the *noise reduction rate* (NRR), which is defined as the output signal-to-noise ratio (SNR) in dB minus input SNR in dB. Also, the filter length of FDICA is 1024 taps.

4.2. Problems in Simply Extended MSICA Based on Method 2

In order to visually evaluate the convergence by FDICA of Method 2, we plot the directivity pattern of the separation filter $\boldsymbol{v}_{LK}(\tau)$ provided by FDICA of Method 2. Figure 3 shows the directivity pattern for a different number of microphones ($K = 2$ or 12), where "Filter 1" is extracting source 1, and "Filter 2" is extracting source 2. In Fig. 3 (a), the directional nulls of the separation filters given by FDICA steer in the direction of interference when two microphones are used. However, in Fig. 3 (b) where 12 microphones are used, the nulls of separation filter 2 steer not only in the direction of interference but also in the target speech direction. Therefore, the output signal from separation filter 2 becomes a zero signal.

In FDICA, the separation filters are updated so that the output signals are mutually independent and the separated signal from FDICA can be generally given as $Z_l(f,m) = a_l(f)\ S_l(f,m)$, where $S_l(f,m)$ is the source signal in the time-frequency domain and $a_l(f)$ is the arbitrary complex-valued coefficient. The coefficient $a_l(f)$ is not determined because we evaluate only the independence between the output signals in FDICA. The coefficient $a_1(f)$ in Fig. 3(b) becomes approximately zero and the output signal from filter 1 becomes the zero signal. The speech signal and the zero signal are mutually independent and consequently, the independence assumption holds. However, needless to say, this solution is trivial with respect to the separation of source signals. This phenomenon occurs due to the fact that the degree of freedom of the separation filter becomes high when we use many microphones. We can conclude that the separation filter with a low degree of freedom is desirable in FDICA. This is the motivation behind proposing the extended MSICA using subarray processing in which the number of each subarray's microphones is equal to that of sources.

4.3. Separation Performance in Every Subarray

Figure 4 shows the NRR results of FDICA and the conventional MSICA for different subarrays. The separation filter length of the TDICA part in MSICA is 2048 taps. These separation performances are averaged for 12 combinations of speakers. From Fig. 4, we can confirm that the source-separation performances in each subarray are disperse. We speculate the reason as being that there are differences in the standing wave condition, the reflection component, and reverberant component at each microphone. The blind determination of the subarray which can achieve a superior separation performance is a difficult problem. Also, we must perform the conventional MSICA in all subarrays and huge amounts of calculations are required. Therefore, it is unreasonable to perform the original MSICA in each subarray.

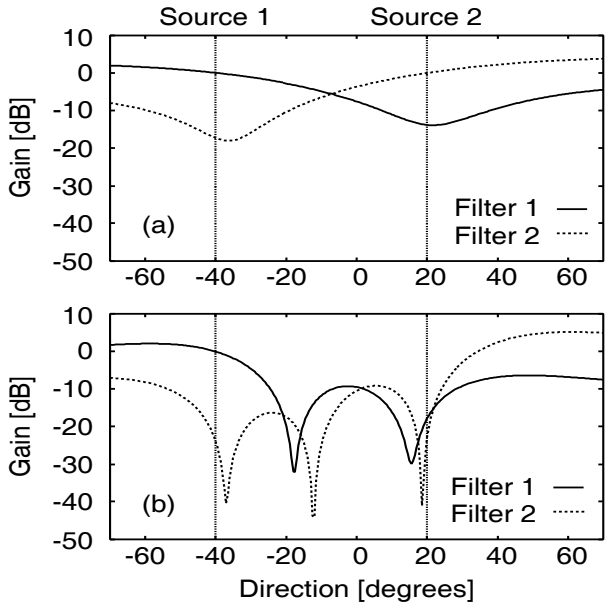

Fig. 3. Directivity patterns in 1812.5 Hz of the separation filters provided by FDICAs of Method 2 by using (a) two microphones are (b) 12 microphones. The number of sources is two.

4.4. Separation Results of Proposed MSICA Using Subarray Processing

In the proposed MSICA using subarray processing, the microphones which are selected symmetrically with respect to the array center are used. For example, the "four-element array" consists of microphones #6, #7, #8, and #9.

As the initial value of the TDICA part in the proposed MSICA, we introduce the following coefficient:

$$\boldsymbol{w}_{LL\cdot(K-1)}(\tau) = \begin{cases} \left[\dfrac{c_{k-(l-1)\times(K-1)}^{-\gamma}}{\sum_{n=1}^{K-1} c_n^{-\gamma}} \cdot \mathrm{IDFT}[\exp(j\omega d_{lk})] \right]_{lk} \\ \quad \text{if } (l-1)\times(K-1) < k \leq l\times(K-1), \\ [0]_{lk} \quad \text{otherwise}, \end{cases} \tag{6}$$

$$c_n = \sum_{\tau=-T}^{T} \Big\{ |\langle \phi(z_i^{(n)}(t)) z_j^{(n)}(t-\tau)\rangle_t| + |\langle \phi(z_j^{(n)}(t)) z_i^{(n)}(t-\tau)\rangle_t| \Big\}, \tag{7}$$

where IDFT$[\cdot]$ denotes an inverse DFT of $\cdot$, T is the length of the output signals from FDICA, ω is an angular frequency, and d_{lk} is the phase delay of input signals for TDICA so that the correlation

between the input signal $z_l{}^{(i)}$ and $z_l{}^{(j)}$ is maximum. Also, γ is the enhancement parameter to weight with the correlation c_n. c_n corresponds to the Frobenius norm of the update term $\{\cdot\}$ in the TDICA algorithm given by Eq. (5), and we estimate the degree of the separation performance by using this value. We introduce this filter (Eq. (6)) as the initial value of the TDICA part in MSICA. If $\gamma = 0$ in Eq. (6), this filter corresponds to a conventional delay-and-sum beamformer. On the other hand, highly separated output signals from specific FDICAs are strongly weighted as the γ is increased. We compare the separation performances of the initial value and the proposed MSICA by changing γ and the number of microphones.

Figures 5 and 6 show the NRR results of the initial value and the proposed MSICA for different γ and numbers of microphones. From Fig. 5, the separation performances of the initial value for the proposed MSICA are improved as γ is increased in all microphones. Therefore, the weighting equation (Eq. (6)) with the input signals for TDICA works effectively. The final separation performance is improved as the number of microphones is increased (see Fig. 6). However, the separation performances of the proposed MSICA which are improvements from the initial values using different γ are not very different in all microphones. We can conclude that the proposed MSICA does not depend on the initial value in the TDICA part and we can achieve a superior separation performance by using the information from many microphones.

5. CONCLUSION

In this paper, we proposed a MSICA, by setting the number of microphones to be larger than that of sources to achieve an improved separation performance. In the FDICA part in the simple extension of MSICA, the use of additional microphones led to alternative problems: the solution is likely to be trapped within a trivial solution and the permutation problem in FDICA becomes very complicated. In order to solve these problems, we proposed a new extended MSICA using subarray processing, where the number of microphones and that of sources are set to be the same in every subarray. The experimental results obtained under real acoustic environmental conditions reveal that the separation performance of the proposed MSICA is improved as the number of microphones is increased.

6. ACKNOWLEDGEMENT

This work was partly supported by NISSAN MOTOR CO., LTD. in Japan and Core Research for Evolutional Science and Technology (CREST) Program "Advanced Media Technology for Everyday Living" of Japan Science and Technology Agency (JST).

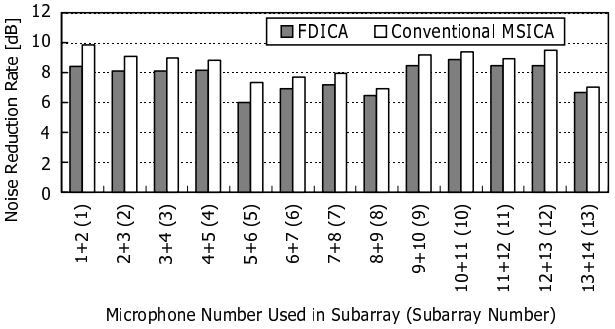

Fig. 4. Comparison of the source-separation performance by FDICA and conventional MSICA in every subarray.

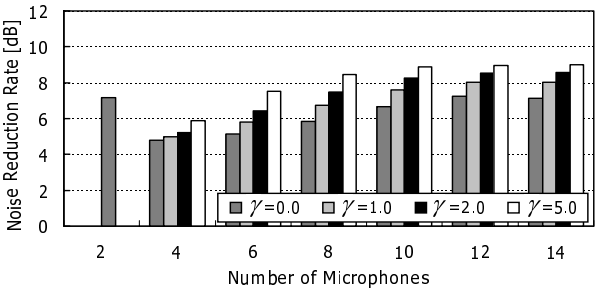

Fig. 5. Comparison of the initial value in TDICA part of the proposed MSICA for different γ and the number of microphones.

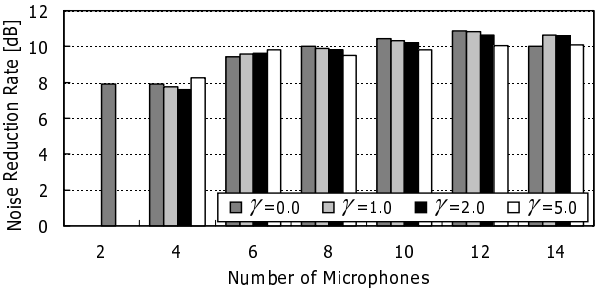

Fig. 6. Comparison of the proposed MSICA for different γ and the number of microphones.

7. REFERENCES

[1] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol.36, pp.287–314, 1994.

[2] S. Amari, S. C. Douglas, A. Cichocki, and H. H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," *Proc. SPAWC97*, pp.101–104, April 1997.

[3] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," *Proc. of Int. Symp. on Nonlinear Theory and Its Application*, pp.923–926, Sept. 1998.

[4] T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA," *IEICE Trans. Fundamentals*, Vol.E86-A, No.4, pp.846–858, April, 2003.

[5] 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, March 2003.

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

[7] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech and Audio Processing*, vol.8, no.3, pp.320–327, May 2000.

[8] 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," *Proc. Int. Symp. on ICA and BSS*, pp.505–510, April 2003.

[9] S. Choi, S. Amari, A. Cichocki, and R. Liu, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," *Proc. Int. Symp. on ICA and BSS*, pp.371–376, January 1999.

[10] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," *Proc. Int. Conf. on Language Resources and Evaluation*, pp.965–968, June 2000.