The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Semi-blind suppression of internal noise for hands-free robot spoken dialog system

Jani Even*, Hiroshi Sawada*, Hiroshi Saruwatari*, Kiyohiro Shikano*, Tomoya Takatani**

Abstract— **The speech enhancement architecture presented in this paper is specifically developed for hands-free robot spoken dialog systems. It is designed to take advantage of additional sensors installed inside the robot to record the internal noises. First a modified frequency domain blind signal separation (FD-BSS) gives estimates of the noises generated outside and inside of the robot. Then these noises are canceled from the acquired speech by a multichannel Wiener post-filter. Some experimental results show the recognition improvement for a dictation task in presence of both diffuse background noise and internal noises.**

I. Introduction

In a hands-free dialog system, the user's voice is picked at a distance with a microphone array resulting in a more natural and convenient interface for humans. But an adverse consequence is that speech recognition is more difficult because the user's speech is contaminated by noise (generated by surrounding sound sources) and reverberation. Thus microphone array techniques are used to reduce the effect of the noise and the reverberation before speech recognition is performed [1], [2]. Some of these techniques [3], [4], [5] are based on frequency domain blind signal separation (FD-BSS). FD-BSS is an efficient approach for recovering the speech by separating the unknown observed convolutive mixture of speech and noise in its different components (see review paper [6]).

For hands-free dialog systems mounted on a robot, the situation is even more difficult as the robot itself has several internal noise sources: fans, servo motors, actuators and several mechanical parts. Moreover these internal noise sources are relatively close to the microphone array and thus highly contaminate the acquired user's speech. But contrary to the noise created by the sources that are outside of the robot (referred to as *external noise*), it is possible to install some sensors inside of the robot that collect additional information on the noise from inside of the robot (referred to as *internal noise*).

The speech enhancement architecture presented in this paper is designed to enhance the user's speech by canceling both the external and the internal noises. For the internal noise, in addition to the usual signals obtained by a microphone array, this method exploits additional sensors installed inside of the robot (see Fig. 1). These additional sensors record an unknown convolutive mixture of the signals from the internal noise sources. We assume that this mixture is unknown because the transfer between the internal noise sources and the internal sensors changes when the robot moves or when different internal noise sources are active. Because the mixture is unknown, we propose to process both the signals from the internal sensors and the signals from the microphone array with a modified FD-BSS based approach. In this sense the proposed method is an extension of the method proposed in [3], [4]. But the method is different from [7], [8], where an unfiltered version of a music signal or a speech signal contaminating the user's speech are known.

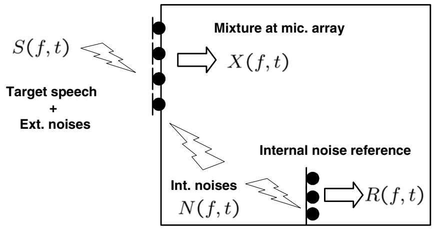

Fig. 1. Sensors and microphone array configuration.

First a modified FD-BSS is applied to the signals from the microphone array and the internal sensors. Since this FD-BSS method incorporates the additional knowledge given by the internal sensors, it is rather referred to as frequency domain semi-blind signal separation (FD-SBSS) (see [9] for SBSS of instantaneous mixtures, [10] for SBSS of convolutive mixture in time and [11] for SBSS of convolutive mixtures in frequency).

The FD-SBSS method gives good estimates of the internal noise and the external noise but it cannot obtain a clean speech estimate. From these estimates, we can obtain the contribution of these noises in the microphone array signals.

Then these noises are canceled by a post processing filter that suppresses the contribution of the noises from the signals observed at the microphone array. Here, we use a Wiener filter on each of the microphone array signals [4]. Finally the output of these Wiener filters are merged together with a delay and sum (DS) beamformer to obtain the enhanced speech fed to the speech recognizer.

Some experimental results show the application of the proposed method for improving the word accuracy in a dictation task in presence of both diffuse background noise and robot internal noise.

II. PRELEMINARIES

A. Frequency domain blind signal separation

For a hands-free speech interface, the propagation of the sounds from their locations of emission to the microphone

*Graduate school of information science Nara Institute of science and technology, Ikoma, Nara, Japan even@is.naist.jp
**Toyota Motor Corporation, Toyota, Aichi, Japan.

978-1-4244-3804-4/09/$25.00 ©2009 IEEE

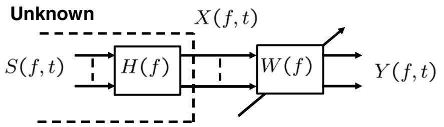

Fig. 2. Mixture matrix $H(f)$ and separation matrix $W(f)$ at the fth frequency bin.

array is modeled by an unknown convolutive mixture. FD-BSS aims at recovering the emitted signals by processing the observed signals in the frequency domain. The frequency domain processing is interesting because after applying a F points short time Fourier transform (STFT) to the observed signals, the convolutive mixture is equivalent to F instantaneous mixtures. Then in each frequency bin, the blind estimation of the emitted signal components is possible using BSS [12].

At the fth frequency bin, the observed signal $X(f,t) = [x_1(f,t),\ldots,x_n(f,t)]^T$ (size $n \times T$) is

$$X(f,t) = H(f)S(f,t),$$

where the $n \times n$ complex valued matrix $H(f)$ represents the instantaneous mixture received by the n microphone array and $S(f,t) = [s_1(f,t),\ldots,s_n(f,t)]^T$ is the emitted signal at the fth frequency bin (size $n \times T$). f denotes the frequency bin, t the frame index and T the number of frames.

A theorem [13] states that if the components of $S(f,t)$ are statistically independent (and at most one is Gaussian) then it is possible to recover them up to scale and permutation indeterminacy by finding the separation matrix $W(f)$ such that the components $y_1(f,t),\ldots,y_n(f,t)$ of

$$Y(f,t) = W(f)X(f,t)$$

are statistically independent (see Fig. 2). These scale and permutation indeterminacy are represented by

$$Y(f,t) = P(f)\Lambda(f)S(f,t),$$

where $P(f)$ is a $n \times n$ permutation matrix and $\Lambda(f)$ is a diagonal $n \times n$ matrix. Namely $W(f)$ is such that

$$W(f)H(f) = P(f)\Lambda(f).$$

Consequently several FD-BSS methods adapt the matrices $W(f)$ in order to minimize a cost function measuring the statistical independence of the components of $Y(f,t)$ (see [6]). Because of the permutation indeterminacy, in order to achieve separation in the time domain , it is necessary to match the components from the same signal in all the frequency bin before transforming back the signals in time. This is referred to as *permutation resolution*. After permutation resolution, the estimated signals are still filtered by an indeterminate filter because of the scaling indeterminacy $\Lambda(f)$. A solution is to *project back* the estimated signals to the microphone array [14].

B. FD-BSS in presence of diffuse noise

For the cocktail party problem where the goal is to separate several speech signals, the author in [15] showed that FD-BSS achieves the separation of a source by placing directional nulls in the direction of the interfering sources. Thus FD-BSS can be seen as a set of adaptive null beamformers.

At this point, it is important to notice that for a hands-free robot spoken dialog system, the user is close to microphone array and can be considered a point source. On the contrary, the external noise sources are far from the array and create a diffuse background noise. Because the user is a point source, FD-BSS is able to obtain a good estimate of the background noise by placing a directional null in the user's direction. But with a limited number of microphones, FD-BSS cannot cancel the diffuse background noise. Thus FD-BSS does not give a good estimate of the speech. In such situation, the speech estimate given by the FD-BSS is equivalent to a delay and sum (DS) beamformer set in the direction of the user [16]. FD-BSS has to be combined with some nonlinear post-filtering techniques in order to improve the quality of the captured speech [3], [4], [5], [16]. An efficient approach uses Wiener filtering to suppress the diffuse background noise estimated by FD-BSS [4].

III. Main results

A. Additional sensors

Let us consider a hands-free robot spoken dialog system that has a n microphone array and r internal sensors (see Fig. 1). Moreover, let use assume that the signals observed at these additional sensors are from an unknown convolutive mixture of the signals from the internal noise sources but contain no contribution of the external noise or user's speech (The type of these sensors and their positions are instrumental in realizing this condition).

In such case, the mixing at the fth frequency bin has the following block structure (see Fig. 3)

$$\begin{bmatrix} X(f,t) \\ R(f,t) \end{bmatrix} = \begin{bmatrix} H_1(f) & H_2(f) \\ 0 & H_3(f) \end{bmatrix} \begin{bmatrix} S(f,t) \\ N(f,t) \end{bmatrix}, \qquad (1)$$

where $R(f,t)$ ($r \times T$) is the signal observed at the additional sensors and $N(f,t)$ ($r \times T$) represents the internal noise sources. The signal $R(f,t)$, that is only a function of $N(f,t)$, is referred to as *reference* in the remainder. This structure corresponds to the situation described in Fig. 1, with n external signals and r internal noises.

B. Frequency domain semi-blind signal separation

The FD-SBSS is a modification of the FD-BSS that takes advantage of the block structure of the mixing process by using a demixer that has a block structure of compatible dimensions with the matrices $H_1(f)$, $H_2(f)$ and $H_3(f)$.

In the fth frequency bin the demixing process is

$$\begin{bmatrix} Y(f,t) \\ Q(f,t) \end{bmatrix} = \begin{bmatrix} W_1(f) & W_2(f) \\ 0 & W_3(f) \end{bmatrix} \begin{bmatrix} X(f,t) \\ R(f,t) \end{bmatrix}.$$

The components of $Q(f,t)$ (size $r \times T$) are only function of the internal noises.

Using the results in [13] presented in Sect. II-A, the components of $Y(f,t)$ and $Q(f,t)$ are all statistically independent if and only if the matrices $W_1(f)$, $W_2(f)$ and

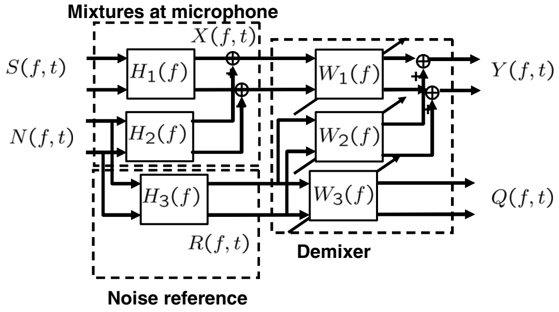

Fig. 3. Block structure of mixing and demixing at the fth frequency bin.

$W_3(f)$ are such that

$$\begin{bmatrix} W_1(f) & W_2(f) \\ 0 & W_3(f) \end{bmatrix} \begin{bmatrix} H_1(f) & H_2(f) \\ 0 & H_3(f) \end{bmatrix} = \begin{bmatrix} P_1(f)\Lambda_1(f) & 0 \\ 0 & P_2(f)\Lambda_2(f) \end{bmatrix},$$

where $P_1(f)(n \times n)$ and $P_2(f)$ $(r \times r)$ are permutation matrices and $\Lambda_1(f)(n \times n)$ and $\Lambda_2(f)$ $(r \times r)$ are diagonal matrices.

Consequently, at the fth frequency bin, it is possible to perform the semi-blind separation by updating $W_1(f)$, $W_2(f)$ and $W_3(f)$ until the components of $Y(f,t)$ and $Q(f,t)$ are all statistically independent. The semi-blind separation method uses the mutual information of $Y(f,t)$ and $Q(f,t)$ to measure the statistical independence of their components. The criterion is optimized by an iterative gradient descent on the matrices $W_1(f)$, $W_2(f)$ and $W_3(f)$. At iteration k, we have the following unmixing system

$$\begin{bmatrix} Y^{(k)}(f,t) \\ Q^{(k)}(f,t) \end{bmatrix} = \begin{bmatrix} W_1^{(k)}(f) & W_2^{(k)}(f) \\ 0 & W_3^{(k)}(f) \end{bmatrix} \begin{bmatrix} X(f,t) \\ R(f,t) \end{bmatrix}.$$

The update rules for the matrices have the following form [11], [17]

$$W_j^{(k+1)}(f) = W_j^{(k)}(f) - \mu \Delta W_j^{(k)}(f),$$

where (dropping the frequency and frame indexes for $Y(f,t)$ and $Q(f,t)$)

$$\begin{aligned} \Delta W_1^{(k)}(f) &= \left(I- < \Phi(Y^{(k)}) Y^{(k)H} >_t\right) W_1^{(k)}(f), \\ \Delta W_2^{(k)}(f) &= \left(I- < \Phi(Y^{(k)}) Y^{(k)H} >_t\right) W_2^{(k)}(f) \\ &\quad - \left(< \Phi(Y^{(k)}) Q^{(k)H} >_t\right) W_3^{(k)}(f), \\ \Delta W_3^{(k)}(f) &= \left(I- < \Phi(Q^{(k)}) Q^{(k)H} >_t\right) W_3^{(k)}(f). \end{aligned}$$

The nonlinear functions $\Phi(\cdot)$ are estimated from the data using a kernel based estimate of the score function of the components of $Y(f,t)$ and $Q(f,t)$ [18] these estimates assume circularity [19].

C. Permutation resolution

A problem that arises when separating speech and diffuse background noise is the permutation resolution. The methods developed for the speech/speech separation are often not efficient for the case of speech in diffuse background noise [20].

Because of the block structure (1), the estimated speech is one of the components of $Y(f,t)$. The other components of $Y(f,t)$ are related to the external noise only (assuming separation is good) and the components of $Q(f,t)$ to the internal noise only.

Here, in order to find the speech component in each of the frequency bin, we rely on the fact that the speech distribution is spikier than that of the diffuse background noise. To measure the 'spikedness' of the distribution, we determine the scale parameter $\alpha_i(f)$ of the Laplacian distribution that fits the distribution of the modulus of $y_i(f,t)$. The maximum likelihood estimate of this parameter is $\alpha_i(f) = \left(\mathcal{E}\left\{|y_i(f,t)|\right\}\right)^{-1}$. The component with the largest parameter is selected as the target speech (for details see [20]). After this first step of permutation resolution, we assume that the components are permuted such that $y_1(f,t)$ is the speech component in the fth bin.

An advantage of the FD-SBSS approach is that, if well separated, the internal noises does not interfere with the permutation resolution. This is of particular interest as some of the robot internal noises have spiky distributions. Thus if these components appear in the permutation resolution, finding the speech components in each frequency bin is more difficult and it would force us to significantly modify the permutation resolution based on Laplacian distribution.

D. Noise estimation

After finding the speech component in each of the frequency bins, we can obtain the estimates of both the external and the internal noises.

The projection back of the external noise is a n component signal defined by

$$X_E(f,t) \quad = \quad W_1(f)^{-1} D Y(f,t), \qquad (2)$$

where D is a $n \times n$ diagonal matrix with ones on its diagonal except for the first entry that is null.

If we assume perfect separation then, after permutation resolution, $y_1(f,t)$ is the speech estimate and we have

$$W_1(f) H_1(f) = \begin{bmatrix} 1 & 0 \\ 0 & P(f) \end{bmatrix} \Lambda_1(f),$$

where $P(f)$ is a $(n-1) \times (n-1)$ permutation matrix.

Let us define $\hat{P}(f) = \begin{bmatrix} 1 & 0 \\ 0 & P(f) \end{bmatrix}$.

Then we have $W_1(f) = \hat{P}(f)\Lambda_1(f)H_1(f)^{-1}$ and $W_1(f)^{-1} = H_1(f)\Lambda_1(f)^{-1}\hat{P}(f)^{-1}$.

Rewriting (2) (dropping frequency index) we have

$$\begin{aligned} X_E(t) &= W_1^{-1} D \begin{bmatrix} W_1 \, W_2 \end{bmatrix} \begin{bmatrix} H_1 \, H_2 \\ 0 \;\; H_3 \end{bmatrix} \begin{bmatrix} S(t) \\ N(t) \end{bmatrix} \\ &= W_1^{-1} D W_1 H_1 S(t) \\ &= H_1 \Lambda_1^{-1} \hat{P}^{-1} D \hat{P} \Lambda_1 S(t) \\ &= H_1 D S(t). \end{aligned}$$

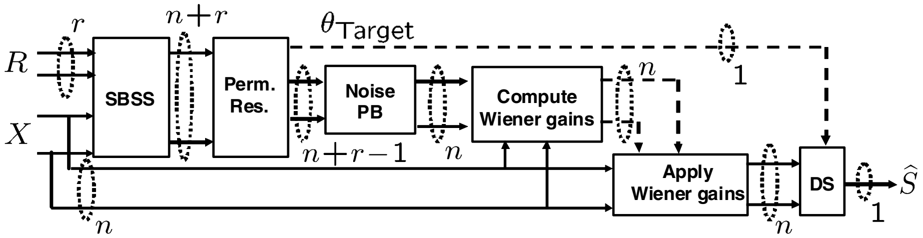

Fig. 4. Overview of the proposed method.

Because D cancels the first component of $S(f,t)$ that is the speech, $X_E(f,t)$ is equal to the contribution of the external noise at the microphone array.

The internal noise projection back is also a n component signal defined by

$$X_I(f,t) \quad = \quad -W_1(f)^{-1}W_2(f)W_3(f)^{-1}Q(f,t). \quad (3)$$

If the separation is perfect we have

$$\begin{aligned} W_2(f) &= -W_1(f)H_2(f)H_3(f)^{-1}, \\ W_3(f)^{-1} &= H_3(f)\Lambda_2(f)^{-1}P_2(f)^{-1}. \end{aligned}$$

We can rewrite (3) (dropping frequency index) as

$$\begin{aligned} X_I(t) &= H_2\Lambda_2^{-1}P_2^{-1}\begin{bmatrix} 0 \, W_3 \end{bmatrix}\begin{bmatrix} H_1 \, H_2 \\ 0 \;\; H_3 \end{bmatrix}\begin{bmatrix} S(t) \\ N(t) \end{bmatrix} \\ &= H_2\Lambda_2^{-1}P_2^{-1}W_3H_3N(t) \\ &= H_2N(t). \end{aligned}$$

Consequently $X_I(f,t)$ is the contribution of the internal noises to the microphone array.

These noise estimates are good because FD-SBSS can efficiently cancel the contribution of the user's speech at the microphone array by steering a directional null in the direction of the user.

E. Wiener post-filter

The noise cancellation is performed by applying a Wiener filter on each of the microphone array signals. The noise estimates used in the Wiener filter are obtained by adding the contributions of both internal and external noises at the microphone

$$X_N(f,t) = X_E(f,t) + X_I(f,t).$$

The Wiener filter gain for the ith component is

$$G^{(i)}(f,t) \quad = \quad \frac{|\widehat{X}^{(i)}(f,t)|^2}{|\widehat{X}^{(i)}(f,t)|^2 + \gamma|\widehat{X}_N^{(i)}(f,t)|^2},$$

where the subscript (i) denotes the ith component and γ is a parameter controlling the noise reduction. If we have additional information concerning what movement the robot is doing at a given time, we can set γ to a specific value that is well adapted for a given condition of the robot otherwise a fixed parameter γ is used. The ith component of the filtered target speech is

$$\widehat{S}^{(i)}(f,t) \quad = \quad \sqrt{G^{(i)}(f,t)|\widehat{X}^{(i)}(f,t)|^2}\,\frac{\widehat{X}^{(i)}(f,t)}{|\widehat{X}^{(i)}(f,t)|}.$$

Finally the n components of the Wiener filtered speech estimates are merged into one by applying a delay and sum (DS) beamformer in the direction θ_{target} of the target speech

$$\widehat{S}(f,t) = \sum_{i=1}^{n} G_{\text{DS}\theta}^{(i)}(f,t)\widehat{S}^{(i)}(f,t),$$

where $G_{\text{DS}\theta}^{(i)}(f,t)$ the gain of the DS beamformer at the ith microphone. The target DOA is estimated during the permutation resolution step after the speech components are found with a method similar to the one in [19].

A block diagram of the processing done in each frequency bin is given in Fig. 4 (the plain arrows are signals and the dashed arrows are parameters).

IV. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of the proposed method, a database of robot internal noises was created. Three additional sensors were installed inside of a robot equipped with a four microphone array. Then for different actions of the robot, the contributions of the internal noises at the additional sensors and at the microphone array were recorded. We also recorded the contributions of an external diffuse background noise at the microphone array and at the additional sensors (which are insignificant compared to the previous ones). The impulse response from one meter in front of the microphone array was also estimated (it also include the impulse response at the additional sensors).

Since our goal is to perform speech recognition, a 20K-word Japanese dictation task from JNAS [21] is used as performance measure. The recognizer is JULIUS [22] using Phonetically Tied Mixture (PTM) model [23]. The open test set is composed of 100 utterances (female speakers). The conditions used in recognition are given in Table II. The acoustic model is a clean model with super-imposed noise (office noise 25dB SNR).

To obtain the test set, the speech signals from the database are first convolved to the estimated impulse response giving the speech contribution at the microphone array. These signals are mixed with the recorded diffuse background noise at different SNR values. Then these mixtures are mixed with the contributions of the internal noises at the microphone array. In the remainder, the SNR between the speech and the diffuse background noise is referred to as external SNR whereas the SNR of the second mixture is referred to as internal SNR. All SNR values are computed at the microphone array when the speech is active. The corresponding mixtures at the additional sensors are also obtained during this process.

TABLE I

Recorded gestures.

gesture number	SNR	$\{\gamma_S, \gamma_{NS}\}$
0	20dB	$\{5, 5\}$
1	16.66dB	$\{5, 5\}$
2	4.54dB	$\{5, 25\}$
3	0.39dB	$\{5, 10\}$
4	5.37dB	$\{5, 100\}$

TABLE II

System specifications.

Sampling frequency	16 kHz
Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCC, 12-order ΔMFCCs 1-order ΔE
HMM	PTM , 2000 states
Training data	Adult and Senior (JNAS)
Test data	Adult and Senior female (JNAS)

For the frequency domain processing, the short time Fourier transform uses a 512 point hamming window with 50% overlap. The semi-blind signal separation is performed by 300 iterations of the SBSS method with adaptation step of 0.1 divided by two every 100 iterations (the method uses adaptive nonlinear functions adapted from [18]). With the notations of Fig. 4, $X(f, t)$ is obtained from the microphone array whereas $R(f, t)$ comes from the internal sensors.

The robot internal noises were recorded in five situations (see Table I): gesture 0 while the robot is not moving and the noise is created by fans and servo motors, gestures 1 to 4 while performing different movements that create additional mechanical and motor noises for short periods. The gestures 1 to 4 are highly non stationary, when the robot is moving the internal SNR is changing from the value for gesture 0 to the values in Table I.

The coefficient γ of the Wiener post-filter is fixed to $\gamma = 5$ or changed according to the noise type while the robot is moving. The values for stationary (γ_S) and non stationary (γ_{NS}) parts are given in Table I. For each noise, γ_{NS} was determined by taking the value from the set $\{5, 7, 9, 10, 12, 15, 20, 25, 30, 50, 100\}$ that results in the best word accuracy .

The word accuracy for the 20k-word dictation task are given in Fig. 5. The results for the proposed method (SBSS-post) are compared with the ones without processing using only second microphone (OBS), a delay and sum beamformer in the direction of the user (DS) and the semi-blind signal separation without post-processing (SBSS).

First we can see that for the gestures 0 and 1, that do not present severe non stationarity, the word accuracy scores are significantly higher. The strong adverse effect of internal noise is especially present for the gestures 2 to 4 that are highly non stationary. The proposed method performs better except for the gestures 2 and 3 when the external SNR is the lowest where the semi-blind signal separation without post-processing is slightly better (0.38% for gesture 2 and 2.98% for gesture 3). We can also note that the improvement between the unprocessed signals (OBS) and the delay and sum beamformer (DS) is very small meaning that the internal noise effect prevents the delay and sum beamformer to perform well. The semi-blind signal separation without post-processing also performs poorly because in presence of diffuse background noise the speech estimate given by the blind signal separation approach is close to a delay and sum beamformer. As a consequence, for the suppression of both the external and the internal noises, it is necessary to replace the linear speech estimation by a nonlinear approach like the one proposed here.

The results obtained while using the noise type and time information to change the coefficient γ are compared to the fixed γ case in Fig. 6.

V. CONCLUSIONS

This paper proposes a speech enhancement architecture developed for hands-free robot spoken dialog system. The specificity of this architecture is the addition of sensors inside the robot that measure the internal noises and the use of semi-blind signal separation (SBSS) with multi-channel Wiener post-filter. Experiments were conducted to show the improvement of word accuracy for a dictation task. The proposed architecture outperformed other approaches demonstrating the usefulness of the additional sensors and the effectiveness of the proposed SBSS based processing. The future research aims at taking into account the non stationarity of the internal noises by using an adaptive SBSS method.

REFERENCES

[1] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Prapagation*, vol. AP-30, pp. 27–34, 1982.

[2] S. Doclo, A. Spriet, and M. Moonen, "Efficient frequency-domain implementation of speech distortion weighted multi-channel wiener filtering for noise reduction," *in Proc. EUSIPCO, Vienna, Austria*, pp. 2007–2010, 2004.

[3] Y. Mori, T. Takatani, H. Saruwatari, K. Shikano, T. Hiekata, and T Morita, "Blind source separation combining simo-ica and simo-model-based binary masking," *ICASSP 2006, Toulouse,France*, pp. 81–84, 2006.

[4] Y. Takahashi, K. Osako, H. Saruwatari, and K. Shikano, "Blind source extraction for hands-free speech recognition based on wiener filtering and ica-based noise estimation," *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSMCA)*, pp. 164–167, 2008.

[5] J. Kocinski, "Speech intelligibility improvement using convolutive blind source separation assisted by denoising algorithms," *Speech Communication*, vol. 50, pp. 29–37, 2008.

[6] M.S. Pedersen, J. Larsen, U. Kjems, and L.C. Parra, *A Survey of Convolutive Blind Source Separation Methods*, Springer, 2007.

[7] R. Takeda, K. Nakadai, K. Komatani, T. Ogata, and H.G. Okuno, "Exploiting known sound sources to improve ica-based robot audition in speech separation and recognition," *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2007)*, pp. 1757–1762, 2007.

[8] R. Takeda, K. Nakadai, K. Komatani, T. Ogata, and H.G. Okuno, "Barge-in-able robot audition based on ica and missing feature theory under semi-blind situation," *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2008)*, pp. 1718–1723, 2008.

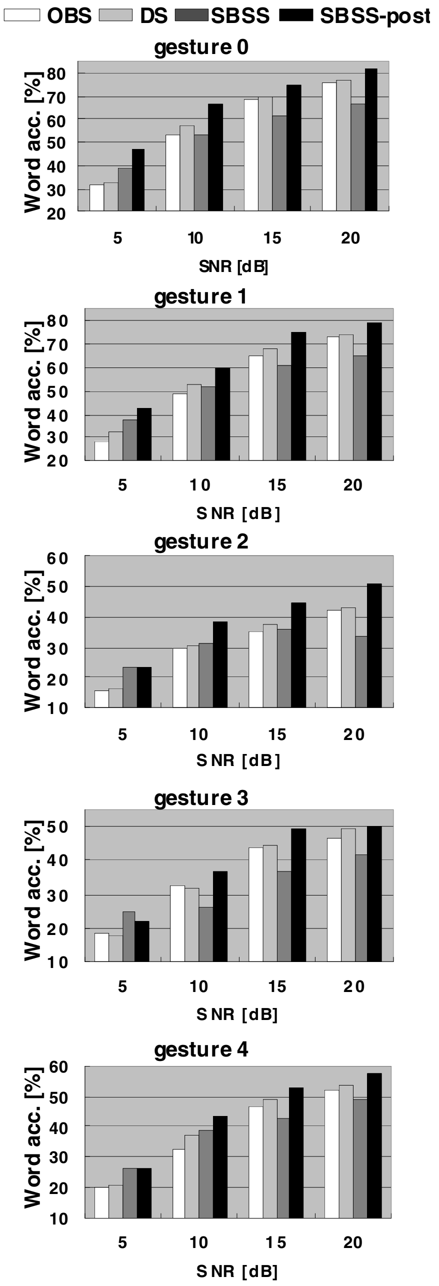

Fig. 5. Word accuracy for all gestures at different SNR values.

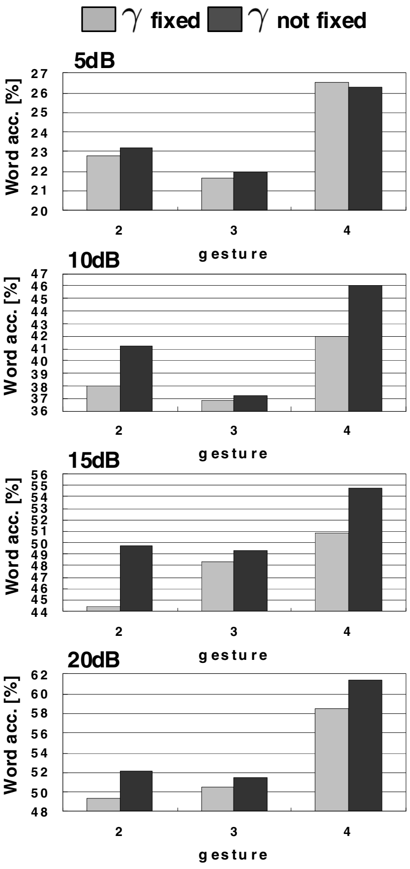

Fig. 6. Effect of variable γ on word accuracy.

[9] M. Joho et al., "Combined blind/nonblind source separation based on the natural gradient," *IEEE Sig. Proc. Letters*, vol. 8, no. 8, pp. 236–238, 2001.

[10] J. Even and K. Sugimoto, "Ica based adaptive disturbance cancellation for mimo system with unknown dynamics," *SICE-ICASE International Joint Conference (SICE-ICCAS 2006), Busan, Korea*, pp. 4473–4478, 2006.

[11] S. Miyabe, T. Takatani, H. Saruwatari, K. Shikano, and Y. Tatekura, "Barge-in- and noise-free spoken dialogue interface based on sound field control and semi-blind source separation," *Proc. 15th European Signal Processing Conference (EUSIPCO 2007)*, pp. 232–236, 2007.

[12] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

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

[14] N. Murata, S. Ikeda, and A. Zieh, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

[15] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1135–1146, 2003.

[16] Y. Takahashi, T. Takatani, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array with independent component analysis for hands-free speech recognition," *International Work Shop on Acoustic Echo and Noise Control (IWAENC) (CD-ROM)*, 2006.

[17] J. Even, H. Saruwatari, and K. Shikano, "Frequency domain semi-blind signal separation: Application to the rejection of internal noises," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, USA*, pp. 4581–4584, 2008.

[18] N. Vlassis and Y. Motomura, "Efficient source adaptivity in independent component analysis.," *IEEE Trans. Neural Networks*, vol. 12, no. 3, pp. 559–566, 2001.

[19] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," *IEICE Trans. Fundamentals*, vol. E86-A, no. 3, pp. 590–596, 2003.

[20] J. Even, H. Saruwatari, and K. Shikano, "An improved permutation solver for blind signal separation based front-ends in robot audition," *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, Nice, France*, pp. 2172–2177, 2008.

[21] K. Ito et al., "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," *The Journal of Acoust. Soc. of Japan*, vol. 20, pp. 196–206, 1999.

[22] "Julius, an open-source large vocabulary csr engine - http://julius.sourceforge.jp," .

[23] A. Lee, T. Kawahara, K. Takeda, and Shikano K., "A new phonetic tied-mixture model for efficient decoding," *In Proceedings of ICASSP*, pp. 1269–1272, 2000.