社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究資料
JSAI Technical Report
SIG-Challenge-A902-5

Blind signal separation with selective post filtering: application to speech enhancement

Jani Even, Hiroshi Saruwatari, Kiyohiro Shikano, Tomoya Takatani[+]
Nara Institute of Science and Technology, Ikoma, JAPAN
[+]Toyota motor corporation, Toyota, JAPAN
even@is.naist.jp

Abstract

In this paper, we consider the human/machine hands-free speech interface where the user voice is picked at a distance with a microphone array. The proposed method aims at suppressing the diffuse background noise efficiently without distorting the speech estimate. This method is a modification of a method combining frequency domain blind signal separation (FD-BSS) and Wiener filter based post-processing. Contrary to the conventional approach, the Wiener post filter is only applied to a selected number of the components separated by FD-BSS. Simulation results show that the proposed approach can achieve a better speech enhancement, measured in term of word recognition in a speech recognition task, than the conventional Wiener filter based post-processing.

1 Introduction

In hands-free speech recognition, microphone array techniques are used to improve the captured speech by reducing the effect of noise and reverberation ([7, 4]). Among these techniques, in recent years, frequency domain blind signal separation (FD-BSS) has been used with success for recovering the speech by separating the observed signals in their different components (see review paper [13]). FD-BSS is efficient for speech/speech separation [11]. But in the human/machine communication where the user's voice has to be extracted from a diffuse background noise, FD-BSS gives a better estimate of the diffuse background noise than of the target speech. Consequently FD-BSS has to be combined with some nonlinear post-filtering techniques in order to improve the quality of the captured speech [18, 11, 16, 17, 9]. An efficient approach suppresses the diffuse background noise estimated by FD-BSS via Wiener filtering [16].

In this paper, our goal is to improve the speech recognition performance for the human/machine hands-free speech interface. The user is assumed to be close to the microphone array and thus is modeled as a point source whereas the other sources create a diffuse background noise. We use a similar approach as in [16] where the noise estimate is obtained by FD-BSS and noise suppression is performed via Wiener filtering. But we propose a modified noise estimate and we do not apply the Wiener filtering directly to the observations.

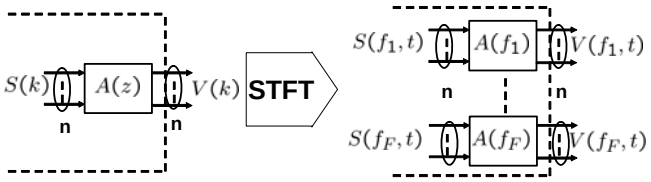

Figure 1: Equivalent mixtures in frequency domain.

The main idea is that if some of the sources from the diffuse background noise are efficiently canceled by the FD-BSS (linear processing), it is better not to include them in the noise estimate used by the post-filter (non linear processing) in order to keep the distortion of the estimated speech low. This is particularly important for speech recognition tasks where the the post-filter should give a good trade-off between high SNR and low distortion [16]. In the proposed approach, after the FD-BSS, we exclude from the noise estimate the estimated noise components that are the least correlated with the speech estimate. Using this modified noise estimate in the Wiener filter based post-filter also requires the modification of the observation before filtering.

Experimental results show the impact of the proposed method on the quality of the speech estimate in a speech recognition task. In particular, the proposed method achieves better performance than the conventional Wiener filter based post-processing.

2 Preliminaries

2.1 Frequency Domain Blind Signal Separation

In blind signal separation of acoustic signals, the propagation of the sounds from their locations of emission

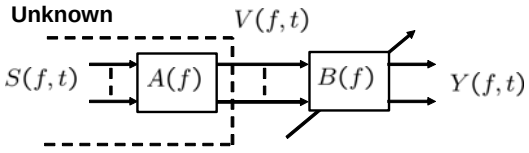

Figure 2: BSS at frequency bin f.

to the microphone array is modeled by a convolutive mixture. After applying a F points short time Fourier transform (STFT) to the observed signals, the convolutive mixture is equivalent to F instantaneous mixtures in the frequency domain (see Fig. 1). At the fth frequency bin, the observed signals are

$$V(f,t) = A(f)S(f,t)$$

where the $n \times n$ complex valued matrix $A(f)$ represents the instantaneous mixture received by the n microphone array and

$$S(f,t) = [s_1(f,t), \ldots, s_n(f,t)]^T$$

are the emitted signal components at the fth frequency bin. t denotes the frame index.

In each frequency bin, the blind estimation of the emitted signal components is possible using BSS [15]. The estimates

$$Y(f,t) = [y_1(f,t), \ldots, y_n(f,t)]^T$$

are obtained by applying an unmixing matrices $B(f)$ to the observed signals (see Fig.2)

$$Y(f,t) = B(f)X(f,t) = B(f)A(f)S(f,t). \qquad (1)$$

If the components of $S(f,t)$ are statistically independent (and at most one is Gaussian) then it is possible to recover the components of $S(f,t)$ up to scale and permutation indeterminacy by finding the separation matrix $B(f)$ such that the components of $Y(f,t)$ are statistically independent [3]. Namely $B(f)$ is such that

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

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

Consequently several FD-BSS methods adapt the matrices $B(f)$ in order to minimize a cost function measuring the statistical dependence between the components of the estimate $Y(f,t)$ (see [13]).

Because of the unknown order of the estimated components $y_i(f,t)$, in order to achieve separation in the time domain , it is necessary to match the components from the same signal in all the frequency bins before transforming back the signals in time. This is referred to as *permutation resolution*. After resolving the permutation, 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 [12]. The projection back of the ith estimate is a n component signal defined by

$$Z_i(f,t) = B(f)^{-1} D_i Y(f,t)$$

where D_i is a matrix having only one non null entry $d_{ii} = 1$. If we assume perfect separation $B(f)A(f) = P(f)\Lambda(f)$ and the estimated signal is $s_j(f,t)$ then $P(f)$ is such that

$$P(f)^{-1} D_i P(f) = D_1$$

and

$$Z_i(f,t) = A(f,t)^{(:,j)} s_j(f,t)$$

where $A(f,t)^{(:,j)}$ is the j^{th} column of $A(f,t)$. Namely $Z_i(f,t)$ is equal to the contribution of the jth estimated signal at the microphone array because the projection back replaces the indeterminate filtering of the estimated signal by the estimate of the room impulse response between the location of the jth signal and the microphone array (represented by $A(f,t)^{(:,j)}$). Note that the observation is the sum of all the projected back components

$$X(f,t) = \sum_{i=1}^{n} Z_i(f,t).$$

3 Proposed method

The block diagram in Fig 3 shows the proposed processing in the frequency domain. The different blocks are explained in the following sections.

3.1 Speech and Diffuse Background Noise Blind Separation

In [14], the authors showed that for speech/speech separation (cocktail party model) FD-BSS is equivalent to a set of adaptive null beamformers (ANBF) each having its null toward different speakers. Thus the separation is achieved because FD-BSS is able to cancel the speeches that are point sources. In our case, FD-BSS gives a good estimate of the diffuse background noise by placing a null in the direction of the speech. But it is not possible to get a good speech estimate since with a limited number of microphones it is not possible to cancel the diffuse background noise [18].

Another problem of the separation of 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 [5]. Here, in order to find the speech component in each of the frequency bins, 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 use the average of the modulus of the $y_i(f,t)$

$$\alpha_i(f) = \mathcal{E}\{|y_i(f,t)|\}$$

under the constraint

$$\mathcal{E}\{|y_i(f,t)|^2\} = 1$$

where $\mathcal{E}\{\cdot\}$ denotes the expectation operator. The component with the smallest parameter is selected as the target speech (for details see [6]). 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.

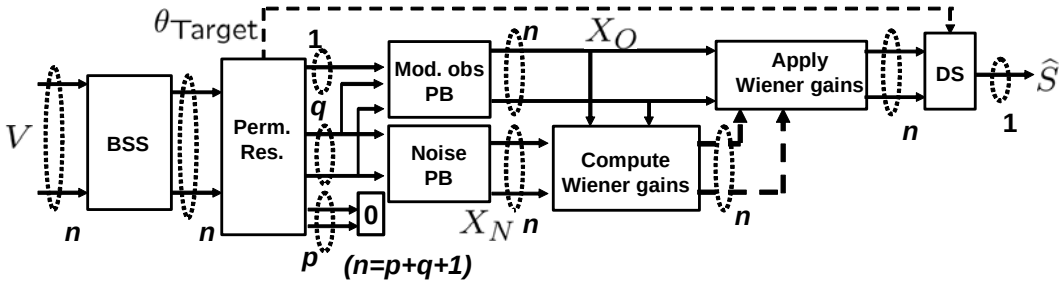

Figure 3: Overview of the proposed architecture.

3.2 Modified Noise Estimate and Modified Observation

Assuming that the FD-BSS method achieved the best possible separation, the estimated noise components $y_2(f,t),\cdots,y_n(f,t)$ contain no speech however the speech estimate $y_1(f,t)$ is still contaminated by the noise.

The noise estimate for the conventional Wiener filter based post-processing is obtained by projecting back the $n-1$ components $y_2(f,t),\cdots,y_n(f,t)$ to the microphone array [16]. But in our approach we do not project back the $n-1$ noise components.

The noise estimate is composed of several components and these components may contribute at different levels in the noise still present in the speech estimate. In particular some of these estimated noise components may have a very small contribution in the noise contaminating the speech estimate. Meaning that FD-BSS suppressed some part of the diffuse background noise (the diffuse background noise may contain contributions of point sources for example). In such case, we propose to exclude these components from the noise estimate used by the Wiener filter post-processing. The reason is that it is better in term of speech distortion to suppress these components with the FD-BSS filter that is linear than with the nonlinear post-processing.

To determine the noise components that have few contribution in the noise contaminating the speech estimate, we compute the correlation between the speech estimate $y_1(f,t)$ and the estimated noise components $y_2(f,t),\cdots,y_n(f,t)$. This correlation is denoted by

$$C_i = \mathcal{E}\{y_1(f,t)y_i(f,t)^*\}$$

where * denotes the complex conjugation.

The noise components are sorted according to the absolute value of these correlations. In the remainder, the components are permuted such that $C_2 > \cdots > C_n$. The p components with smallest correlation are not projected back (see the p components set to 0 in Fig. 3).

Thus the noise estimate $X_N(f,t)$ is only composed of the projection back of $y_2(f,t),\cdots,y_{n-p}(f,t)$

$$X_N(f,t) = B(f)^{-1}D_N Y(f,t) = \sum_{i=2}^{n-p} Z_i(f,t)$$

where D_N is a matrix selecting $y_2(f,t),\cdots,y_{n-p}(f,t)$ (the *Noise PB* block in Fig. 3).

Since the last p components are not projected back the Wiener filtering has to be applied to the modified observation $X_O(f,t)$ obtained by projecting back all the components except these p last ones

$$X_O(f,t) = B(f)^{-1}D_O Y(f,t) = \sum_{i=1}^{n-p} Z_i(f,t)$$

where D_O is a matrix selecting $y_1(f,t),\cdots,y_{n-p}(f,t)$ (the *Mod. obs PB* block in Fig. 3).

3.3 Wiener post-filter and delay and sum beamformer

The modified noise estimate $X_N(f,t)$ and the modified observation $X_O(f,t)$ both have n components. The Wiener filtering is applied component wise and the Wiener gain for the ith component is

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

where the subscript (i) denotes the ith component and γ is a parameter controlling the noise reduction. The ith component of the filtered target speech is

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

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

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

where $G_{\mathrm{DS}\theta}^{(i)}(f,t)$ the gain of the DS beamformer at the ith microphone (the target DOA is estimated during the permutation resolution step. It is an average over all bins of the estimated DOA of the separated speech component).

4 Experimental Results

To demonstrate the effectiveness of the proposed post-processing based on selective projection back, we compare it to the conventional Wiener filter based post-processing, to the FD-BSS with no post-processing and to a delay and sum beamformer (DS).

A four ($n = 4$) microphone array (inter microphone spacing of 2.15cm) was used to record a diffuse background noise (a vacuum cleaner at two meters from the array and -60^o), the impulse responses at one meter from the array in front of the array and at an angle of 60^o (see Fig. 4). The recorded noise is mixed with the convolution of the impulse response at an angle of 60^o with a recorded fan noise. The SNR of this mixture is 0dB. Then this mixture of noises is mixed with the convolution of the impulse responses and a clean speech (100 signals from the JNAS database of Japanese sentences [8]). A second set of data is obtain by mixing only the diffuse background noise with the filtered speeches. The first data set is referred to by 'fan' whereas the second is referred to by 'no fan'. The SNR values between noise and speech are adjusted to be the same for both datasets.

For the frequency domain processing, the short time Fourier transform uses a 512 point hamming window with 50% overlap. The separation is performed by 300 iterations of a BSS method with adaptation step of 0.1 divided by two every 100 iterations (the method is adapted from [2, 19]).

The proposed approach is tested with two modified noise estimates corresponding to $p = 1$ and $p = 2$. The result are compared to the delay and sum beamformer in front of the array (DS), the FD-BSS with no post processing (BSS) and the conventional Wiener filter $p = 0$ (note: the FD-BSS with no post processing can be seen as discarding all the noise components $p = 3$). Several values of the coefficient γ of the Wiener filter were tested for each method: $\gamma \in \{1, 5, 10, 15, 20, 25\}$.

Since our goal is speech recognition, a 20K-word Japanese dictation task from JNAS is used as performance measure. The word accuracy achieved by the recognizer is function of both the SNR and the amount of distortion of the speech estimate. The recognizer is JULIUS [1] using Phonetically Tied Mixture (PTM) model [10]. The open test set is composed of 100 utterances (female speakers). The conditions used in recognition are given in Table 1. The acoustic model is a clean model with super-imposed noise (office noise 25dB SNR).

Figure 5 shows the word accuracy achieved by the different methods on the two data sets ('fan' and 'nofan') for the different SNR values. For each case the result is the one obtain with the parameter γ giving the best

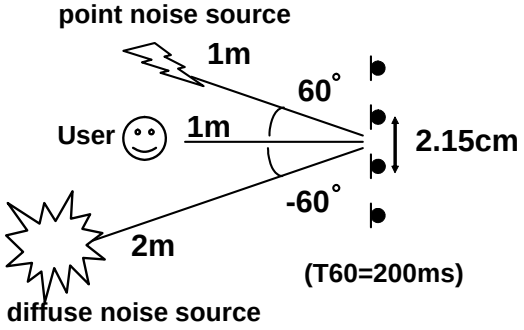

Figure 4: Experimental setup.

Table 1: *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)

word accuracy (also see first row of Table 2).

We can see that, at the same SNR, the performances are better for the 'fan' dataset that contains a point source in addition to the diffuse background noise. In particular for the lower SNRs (5dB and 10dB), the improvement of the word accuracy with the proposed method over the conventional method is better for the 'fan' dataset. This shows that if some components of the noise are canceled by the FD-BSS (the point source fan noise), modifying the noise estimate improves the performance.

There is also a performance gain on the 'nofan' dataset showing that some of the noise components of the diffuse background noise contributed less to the noise contaminating the speech estimate given by FD-BSS. We can also notice that for $p = 2$ the performance is better than for $p = 1$. Meaning that discarding more noise components lead to better results on these datasets.

These results also show the necessity of the nonlinear post-processing as in all cases there is an improvement over the FD-BSS.

The effect of the coefficient γ is depicted in Fig. 6 (the three plots share same color scale). For the proposed post-processing, like for the conventional Wiener filter there is a trade-off between SNR and distortion, the word accuracy is better with a larger γ at low SNR and a smaller γ at high SNR.

Table 2 shows the difference of word accuracy between the proposed method with $p = 2$ and the conventional method $p = 0$ for different choice of γ ()A positive value indicates that the proposed method is superior to the conventional method). The row 'best γ' is obtained by selecting for each method at each SNR the parameter γ from the list $\{1, 5, 10, 15, 20, 25\}$ that gives the best word accuracy. This row shows the improvement for the proposed method for $p = 2$ over conventional method ($p = 0$) in Fig. 5. The other rows show the improvement for fixed values of γ. Note that for larger γ there is no performance improvement on the 'fan' dataset as the proposed method perform best with small γ (see bottom of Fig. 6). This shows that at high SNR it is important to choose a smaller γ for the proposed method. We can also notice that for the 'nofan' dataset the performance difference is larger than for the 'fan' dataset at high SNR.

	'fan' dataset				'nofan' dataset			
	5dB	10dB	15dB	20dB	5dB	10dB	15dB	20dB
best γ	4.92	5.1	2.74	0.91	2.18	2.28	3.53	1.36
$\gamma = 1$	13.3	15.03	6.79	0.91	9.3	12.47	7.13	4
$\gamma = 15$	6.19	5.08	0.77	−0.63	0.72	4.12	3.71	2.18
$\gamma = 25$	1.81	4.26	−1.3	−1.77	3.58	3.98	1.74	2.44

Table 2: Word Accuracy differences for $p = 2$ and $p = 0$ versus γ

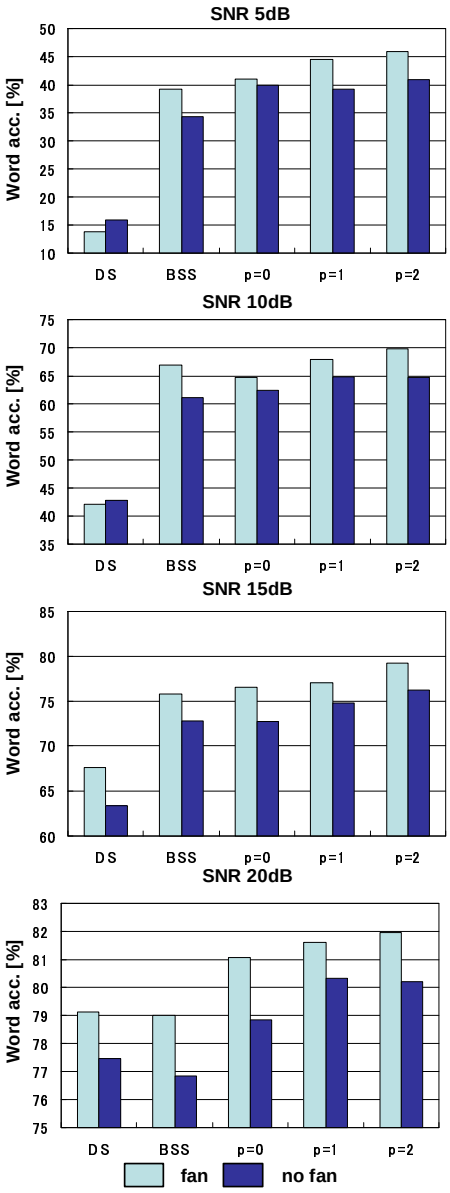

Figure 5: Word accuracy for different SNR values with the different methods for both datasets.

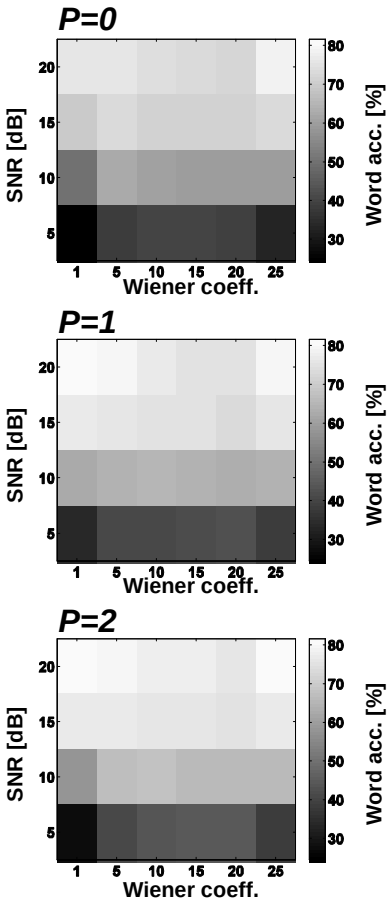

Figure 6: Effect of Wiener coefficient on word accuracy for the different methods ('fan' dataset only).

5 Conclusion

In this paper, we consider the suppression of the diffuse background noise in the human/machine communication scenario. We proposed a modification of the noise estimation given by FD-BSS. This modification leads to a more efficient Wiener filter based post-processing of the speech estimate. Some experimental results showed that this approach increases the word accuracy in a dictation task.

References

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

[2] A. J. Bell and T. J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[3] P. Comon. Independent component analysis, a new concept ? *Signal Processing*, 36:287–314, 1994.

[4] 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*, pages 2007–2010, 2004.

[5] 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*, pages 2172–2177, 2008.

[6] J. Even, H. Saruwatari, and K. Shikano. Blind signal extraction based speech enhancement in presence of diffuse background noise. *2009 IEEE Workshop on Statistical Signal Processing (SSP2009), Cardiff, Wales, UK*, pages 513–516, 2009.

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

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

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

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

[11] 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*, pages 81–84, 2006.

[12] N. Murata, S. Ikeda, and A. Zieh. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1-4):1–24, 2001.

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

[14] H. Saruwatari et al. Blind source separation combining independent component analysis and beamforming. *EURASIP Jour. on Appl. Sig. Proc.*, 2003(11):1135–1146, 2003.

[15] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22(1-3):21–34, 1998.

[16] 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)*, pages 164–167, 2008.

[17] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano. Blind spatial subtraction array for speech enhancement in noisy environment. *IEEE Transaction on Audio, Speech and Language Processing*, 17(4):650–664, 2009.

[18] 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.

[19] N. Vlassis and Y. Motomura. Efficient source adaptivity in independent component analysis. *IEEE Trans. Neural Networks*, 12(3):559–566, 2001.