LOW-LATENCY REAL-TIME BLIND SOURCE SEPARATION FOR HEARING AIDS BASED ON TIME-DOMAIN IMPLEMENTATION OF ONLINE INDEPENDENT VECTOR ANALYSIS WITH TRUNCATION OF NON-CAUSAL COMPONENTS

Masahiro Sunohara, Chiho Haruta

RION CO., LTD.
3-20-41 Higashimotomachi, Kokubunji,
Tokyo, 185-8533, Japan
suno@rion.co.jp

Nobutaka Ono

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo, 101-8430, Japan
onono@nii.ac.jp

ABSTRACT

In this paper, we present a low-latency scheme for real-time blind source separation (BSS) based on online auxiliary-function-based independent vector analysis (AuxIVA). In many real-time audio applications, especially hearing aids, low latency is highly desirable. Conventional frequency-domain BSS methods suffer from a delay caused by frame analysis. To reduce the delay, we implement separation filters as multiple FIR filters in the time domain, which are converted from demixing matrices estimated by online AuxIVA in the frequency domain. Also, to further reduce the latency, part of the non-causal components of the FIR filters are truncated on the basis of causality analysis for ideal separation filters using a simple model. By experimental evaluation using a head and torso simulator in a real environment, the proposed algorithm with an algorithmic delay of less than 10 ms exhibited a separation performance of 7.7 dB in terms of the signal-to-interference ratio (SIR), which was less than 1.4 dB degradation from the case of conventional frequency-domain implementation.

***Index Terms*—** blind source separation, independent vector analysis, hearing aids, algorithmic delay, causality

1. INTRODUCTION

Most hearing-impaired people find it difficult to focus on a desired sound in noisy environments such as a party venue or a crowded restaurant. Improving speech communication in such environments is one of the most important issues to be resolved by hearing aids. Many types of single-channel noise reduction techniques are widely used for hearing aids to reduce undesired noise [1]; however, they cannot easily to reduce nonstationary sound sources. Also, many studies on multichannel techniques have been conducted, although they require a perfect voice activity detector [2,3] or the directions of the desired sound sources [4, 5] in advance to design an appropriate beamformer.

As a technique for solving these problems, blind source separation (BSS) may be applicable [6, 7]. BSS is a signal processing method that can extract a desired sound source from a mixture by using multiple microphones without requiring information on the source signals. To apply BSS to hearing aids, which are real-time systems, there are two important issues. One is to reduce the computational time. In the frequency-domain approach for convolutive BSS, independent vector analysis (IVA) has been proposed as a technique that does not require the solution of a permutation ambiguity problem [8–10], and auxiliary-function-based IVA (AuxIVA) is a state-of-the-art approach with fast convergence and a low calculation cost [11–13].

Another important issue is to reduce the latency from the input to the output. In addition to its computational complexity, an algorithm may require an inherent delay, which is referred to as an algorithmic delay. In the case of frequency-domain BSS, a delay of at least one frame length is necessary for frame analysis [14]. Although several real-time implementations of IVA have been proposed [15, 16], this delay is not avoidable, similarly to in other frequency-domain BSS approaches, and can be relatively large. For instance, the algorithmic delay becomes 256 ms when the frame length is 4096 samples at a sampling frequency of 16 kHz. Such a large delay causes various problems in a hearing aid system such as a difficulty in speaking owing to the delayed auditory feedback effect or a sense of discomfort due to the loss of lip synchronization [17]. Several studies have been conducted on tolerable delays for hearing aids [18, 19]. One study indicated that a tolerable delay is about 6 ms at 1 kHz [18].

To solve the problem of an inherent delay, a low-latency scheme for real-time BSS is proposed in this paper. Focusing on reducing the latency, we convert demixing matrices estimated by online AuxIVA into multiple FIR filters in the time domain. Also, to further reduce the latency, we truncate part of the non-causal components of the FIR filters. To evaluate the proposed system, a PC simulation was carried out using real mixtures of two speeches recorded by two microphones installed in binaural hearing aids attached to each artificial ear of a head and torso simulator.

2. LOW-LATENCY SCHEME FOR REAL-TIME BLIND SOURCE SEPARATION

2.1. Frequency-domain BSS

We assume that K sources are observed by K microphones and that their short-time Fourier transform (STFT) representations are known. Let $\boldsymbol{s}(\omega,\tau) = [s_1(\omega,\tau)\cdots s_K(\omega,\tau)]^t$ and $\boldsymbol{x}(\omega,\tau) = [x_1(\omega,\tau)\cdots x_K(\omega,\tau)]^t$ be the vector representations of the source and the observation signal in the (ω,τ)th time-frequency bin, respectively, where ^{t} denotes the vector transpose. In the frequency-domain approach for a convolutive mixture, the following linear mixing model is assumed:

$$\boldsymbol{x}(\omega,\tau) = A(\omega;\tau)\boldsymbol{s}(\omega,\tau). \tag{1}$$

Here, $A(\omega;\tau)$ is a $K\times K$ mixing matrix. The sources are estimated by the following linear demixing process:

$$\boldsymbol{y}(\omega,\tau) = W(\omega;\tau)\boldsymbol{x}(\omega,\tau), \tag{2}$$

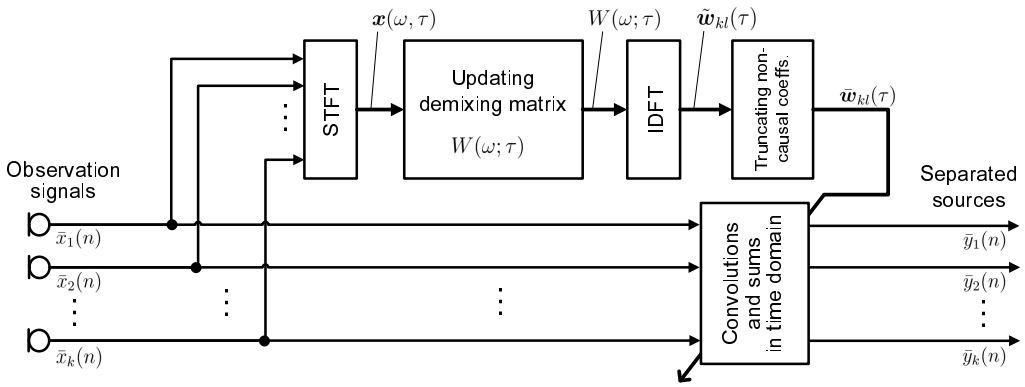


Fig. 1. Signal block diagram of the proposed algorithm.

where

$$W(\omega;\tau) = (\boldsymbol{w}_1(\omega;\tau)\cdots\boldsymbol{w}_K(\omega;\tau))^h \tag{3}$$

is a demixing matrix, ^{h} denotes the Hermitian transpose, and

$$\boldsymbol{y}(\omega,\tau) = [y_1(\omega,\tau)\cdots y_K(\omega,\tau)]^t \tag{4}$$

represents the estimated sources.

2.2. Online AuxIVA

As an effective algorithm to estimate the demixing matrices $W(\omega;\tau)$, an online AuxIVA algorithm has been proposed [16]. The algorithm consists of the following alternate update rules.

Weighted covariance matrix update

$$r_k(\tau) = \sqrt{\sum_{\omega=1}^{N_\omega} \left|\boldsymbol{w}_k^h(\omega;\tau)\boldsymbol{x}(\omega,\tau)\right|^2}, \tag{5}$$

$$\begin{aligned} V_k(\omega;\tau) &= \alpha V_k(\omega;\tau-1) \\ &\quad + (1-\alpha)\frac{G'(r_k(\tau))}{r_k(\tau)}\boldsymbol{x}(\omega,\tau)\boldsymbol{x}^h(\omega,\tau), \end{aligned} \tag{6}$$

where N_ω is the number of frequency bins equal to the frame length, $0 \le \alpha < 1$ is the forgetting factor, and $G(r)$ is called the spherical contrast function. In AuxIVA, $G(r)$ is selected such that $G'(r)/r$ is monotonically decreasing for $r > 0$ [11], where $'$ denotes the derivative. A typical choice is $G(r) = r$, which corresponds to the Laplace distribution.

Demixing matrix update

$$W(\omega;\tau) = W(\omega;\tau-1). \tag{7}$$

$$\boldsymbol{w}_k(\omega;\tau) \leftarrow (W(\omega;\tau)V_k(\omega;\tau))^{-1}\boldsymbol{e}_k, \tag{8}$$

$$\boldsymbol{w}_k(\omega;\tau) \leftarrow \boldsymbol{w}_k(\omega;\tau)/\sqrt{\boldsymbol{w}_k^h(\omega;\tau)V_k(\omega;\tau)\boldsymbol{w}_k(\omega;\tau)}, \tag{9}$$

where $\boldsymbol{e}_k$ is the column vector whose kth element is one and all the other elements are zeros.

2.3. Time-domain implementation

In frequency-domain BSS, an algorithmic delay corresponding to the frame length is unavoidable. A means of shortening this delay is to form two paths for updating the demixing matrices in the frequency domain and separating the sources using FIR filters in the time domain as shown in Fig. 1. After applying back-projection [20], the frequency-domain demixing matrix $W(\omega;\tau)$ is converted to multiple time-domain FIR filters using the inverse DFT as follows:

$$\tilde{w}_{kl}(n;\tau) = \frac{1}{N_\omega}\sum_{\omega=1}^{N_\omega} w_{kl}(\omega;\tau)e^{j2\pi(\omega-1)n/N_\omega}, \tag{10}$$

for $n = -N_\omega/2,\ldots,N_\omega/2-1$, where $w_{kl}(\omega;\tau)$ indicates the (k,l) element of the matrix $W(\omega;\tau)$. The vector $\tilde{\boldsymbol{w}}_{kl}(\tau) = (\tilde{w}_{kl}(-N_\omega/2,\tau)\cdots\tilde{w}_{kl}(N_\omega/2-1,\tau))$ is considered as a time-domain FIR filter of length N_ω. The elements of the vector $\tilde{\boldsymbol{w}}_{kl}(\tau)$ with positive and negative time indexes correspond to causal and non-causal components, respectively. The non-causal components must be shifted to the causal domain for implementation in a real system, and this shift determines the algorithmic delay of FIR filtering. To shift all the non-causal components, an algorithmic delay of $N_\omega/2$ samples is necessary.

To further shorten the algorithmic delay, we here consider shifting only N_d non-causal components and truncating the others. This process can be written as

$$\bar{w}_{kl}(n;\tau) = \tilde{w}_{kl}(n-N_d;\tau)\ (n = 0,\ldots,\frac{N_\omega}{2}+N_d-1). \tag{11}$$

Then, the separated time-domain signal $\bar{y}_k(n)(k = 1,\ldots,K)$ can be obtained as

$$\bar{y}_k(n) = \sum_{l=1}^{K}\sum_{m=0}^{\frac{N_\omega}{2}+N_d-1}\bar{w}_{kl}(m;\tau)\bar{x}_k(n-m), \tag{12}$$

where $\bar{x}_k(n)$ is the time-domain observation signal of the kth microphone at discrete time index n. Note that the algorithmic delay is determined by only N_d, although the total filter length is $N_\omega/2+N_d$.

Generally, the truncation of the non-causal components should degrade the separation performance. However, if all the non-causal components of $\tilde{\boldsymbol{w}}_{kl}(\tau)$ are originally zero, these components can be truncated and the algorithmic delay of the system theoretically can be zero without performance degradation. Therefore, the analysis of the causality of ideal separation filters is important for justifying this truncation.

3. CAUSALITY OF DEMIXING IMPULSE RESPONSES

We here investigate the causality of impulse responses of ideal separation filters. To consider a hearing aid application, we focus

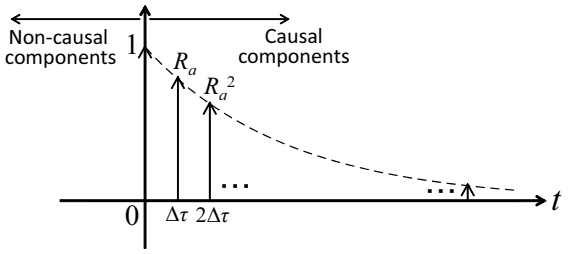


Fig. 2. Demixing impulse response $\tilde{w}_{21}(t)$ with $R_a < 1$ and $\Delta\tau > 0$.

on a simple model consisting of two sound sources, $s_1(\omega,\tau)$ and $s_2(\omega,\tau)$, and two observations, $x_1(\omega,\tau)$ and $x_2(\omega,\tau)$, in an anechoic environment. Let $a(\theta)$ and $\tau(\theta)$ be the amplitude ratio and the time difference of the second channel relative to the first channel for a source with direction θ, respectively. The incident directions of s_1 and s_2 have angles θ_1 and θ_2, respectively, and we write $a_k = a(\theta_k)$ and $\tau_k = \tau(\theta_k)$ $(k = 1, 2)$ for simplicity. Then, the microphone signals are given by the following frequency-domain expression:

$$\begin{pmatrix} x_1(\omega,\tau) \\ x_2(\omega,\tau) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ a_1 e^{-j\omega\tau_1} & a_2 e^{-j\omega\tau_2} \end{pmatrix} \begin{pmatrix} s_1(\omega,\tau) \\ s_2(\omega,\tau) \end{pmatrix}. \quad (13)$$

Then, the source signals can be expressed as

$$\begin{pmatrix} s_1(\omega,\tau) \\ s_2(\omega,\tau) \end{pmatrix} = \frac{1}{D} \begin{pmatrix} a_2 e^{-j\omega\tau_2} & -1 \\ -a_1 e^{-j\omega\tau_1} & 1 \end{pmatrix} \begin{pmatrix} x_1(\omega,\tau) \\ x_2(\omega,\tau) \end{pmatrix}, \quad (14)$$

$$D = a_2 e^{-j\omega\tau_2} - a_1 e^{-j\omega\tau_1} = -a_1 e^{-j\omega\tau_1}(1 - R_a e^{-j\omega\Delta\tau}), \quad (15)$$

where $R_a = a_2/a_1$ and $\Delta\tau = \tau_2 - \tau_1$. When $R_a < 1$, the elements of the demixing matrix can be represented by applying the formula for the sum to infinity of a geometric progression to Eq. (14) as follows:

$$w_{21}(\omega) = \sum_{m=0}^{\infty} (R_a e^{-j\omega\Delta\tau})^m, \quad (16)$$

where $w_{kl}(\omega)$ $(k, l = 1, 2)$ represents the klth element of the ideal demixing matrix. Here we show only $w_{21}(\omega)$ owing to limited space but $w_{11}(\omega), w_{12}(\omega)$, and $w_{22}(\omega)$ are also represented in a similar way.

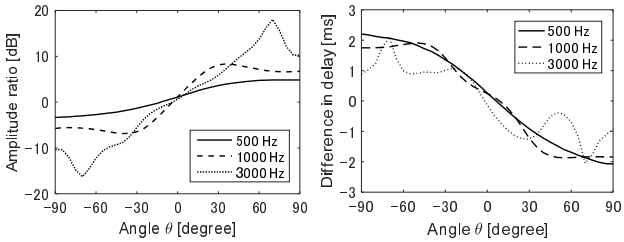
By applying the inverse Fourier transform to Eq. (16), the time-domain impulse response $\tilde{w}_{21}(t)$ can be obtained as

$$\tilde{w}_{21}(t) = \sum_{m=0}^{\infty} R_a^m \delta(t - m\Delta\tau). \quad (17)$$

When $R_a > 1$, we have

$$\tilde{w}_{21}(t) = -\sum_{m=0}^{\infty} R_a^{-(m+1)} \delta(t + (m+1)\Delta\tau). \quad (18)$$

Therefore, when $R_a < 1$ and $\Delta\tau > 0$ or $R_a > 1$ and $\Delta\tau < 0$, which indicates that an earlier channel is larger, the demixing impulse response $\tilde{w}_{21}(t)$ exhibits a periodic impulse train with exponential decay in only the causal-component part as shown in Fig. 2. On the other hand, when $R_a > 1$ and $\Delta\tau > 0$ or $R_a < 1$ and $\Delta\tau < 0$, the demixing impulse response $\tilde{w}_{21}(t)$ consists of only the non-causal components.



(a) Amplitude ratio $a(\theta)$ (b) Time difference $\tau(\theta)$

Fig. 3. Amplitude ratio and time difference of the right channel relative to the left channel measured using KEMAR.

Table 1. Experimental conditions

microphone spacing	18 cm
reverberation time	650 ms at 500 Hz
signal length	30 s × 10
sampling frequency	16 kHz
frame length	4096
frame shift	1024
window function	Hanning
forgetting factor	0.98

A sufficient condition that the ideal separation filters are causal is that $a(\theta)$ and $\tau(\theta)$ are monotonically increasing and monotonically decreasing functions of θ, respectively. Figure 3 shows some examples of the amplitude ratio and time difference of the right channel relative to the left one, measured using a KEMAR dummy-head microphone [21]. Although the amplitude ratio and time difference have frequency dependence and the monotonicity is not perfect, the condition is roughly satisfied. Therefore, we can expect that the performance degradation due to the truncation of non-causal components will not be large in the case of hearing aids.

4. EVALUATION

To evaluate the performance of the proposed algorithm with hearing aids, a PC simulation was carried out using real mixtures of two speeches recorded by a head and torso simulator (G.R.A.S.: KEMAR type 45BB) in a meeting room with a volume of 135 m^3. Figure 4 shows the setup of the loudspeakers and microphones in the evaluation. Two electret condenser microphones were installed into behind-the-ear (BTE)-type hearing aids attached to each ear of the head and torso simulator. The direction of one of the two sources was fixed at 0° and that of the other source was varied from 30° to 180° in steps of 30°. We selected ten speech sources for each direction from the RWCP Japanese News Speech Corpus [22]. The other experimental conditions are summarized in Table 1.

The number of remaining non-causal components N_d was varied from 0 to 160 samples, corresponding to an algorithmic delay from 0 to 10 ms. As a conventional system for comparison, we used online AuxIVA implemented in the STFT domain with a frame length of 4096 or 160 samples. The experiments were performed on the recorded mixtures using MATLAB R2016a on a laptop PC with an Intel Core i7-3770 3.40 GHz. We confirmed that this algorithm acted as a real-time system by another C-language-based implementation with real audio I/O. The performance was evaluated

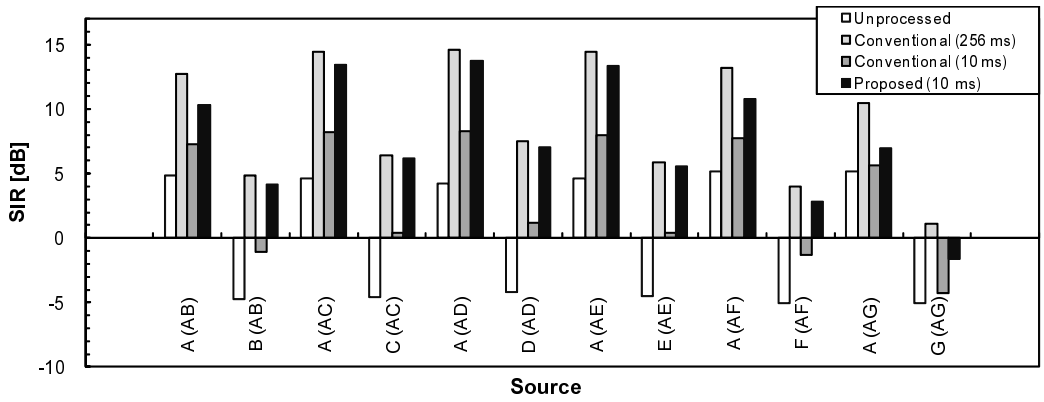


Fig. 5. Separation performance of the proposed algorithm compared with that of the conventional algorithm with algorithmic delays of 10 and 256 ms.

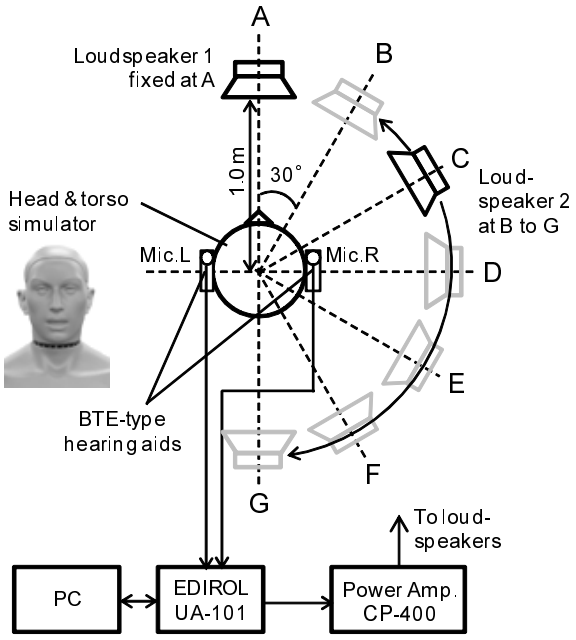


Fig. 4. Setup of loudspeakers and microphones in the evaluation.

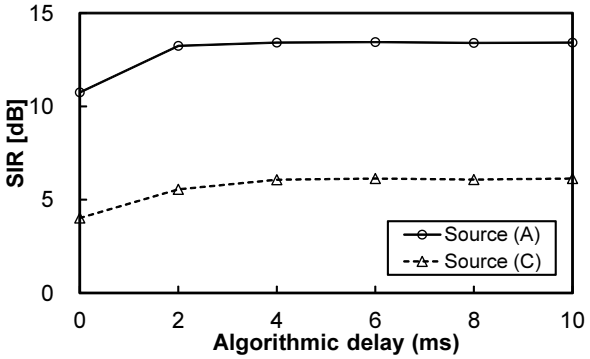


Fig. 6. Separation performance vs algorithmic delay of the proposed algorithm.

by the average of the signal-to-interference ratio (SIR) over all trials with exception of the first three seconds, which was calculated by bss_eval_images.m in the BSS toolbox [23].

Figure 5 shows the separation performance for the proposed algorithm with an algorithmic delay of 10 ms compared with that for the conventional frequency-domain implementation with an algorithmic delay of 10 or 256 ms. On the horizontal axis, A(AB) denotes source A in a mixture of source A and source B. The large difference in the SIR between the center (A) and non-center (B-G) sources under the unprocessed condition was caused by the head-related transfer function (HRTF) of the torso simulator. In the conventional frequency-domain implementation, shortening the algorithmic delay by using a short window length (10 ms) results in unsatisfactory separation performance. On the other hand, in spite of the short algorithmic delay of 10 ms, the proposed algorithm shows better separation performance, which was on average only 1.4 dB less than that of the conventional algorithm with an algorithmic delay of 256 ms.

Figure 6 shows the resultant SIRs of the proposed algorithm for algorithmic delays from 0 to 10 ms (N_d was set from 0 to 160 samples) when the two sources were located at A(0°) and C(60°). From the figure, it is observed that algorithmic delays of 2 ms and above resulted in better performance.

5. CONCLUSION

In this paper, we presented a real-time BSS algorithm with low latency based on online IVA for hearing aids. The proposed algorithm can significantly shorten the algorithmic delay by the time-domain implementation of demixing matrices as FIR filters and the truncation of part of their non-causal components. This was justified by an analysis of the causality of ideal separation filters. From the result of the evaluation, the algorithmic delay in the proposed system was within 10 ms and the average SIR was 7.7 dB, which is a performance degradation of less than 1.4 dB compared with the average performance of a conventional method with an algorithmic delay of 256 ms. These results suggest that the proposed algorithm can be used for real-time audio devices such as hearing aids.

6. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. ASSP*, vol. 32, no. 6, pp. 1109–1121, 1984.

[2] T. J. Klasen, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural multichannel Wiener filtering for hearing aids: preserving interaural time and level differences," in *Proc. ICASSP*, vol. 5, pp. 145–148, 2006.

[3] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids," *J. Acoust. Soc. Am.*, vol. 125, no. 1, pp. 360–371, 2009.

[4] J. E. Greenberg and P. M. Zurek, "Evaluation of an adaptive beamforming method for hearing aids," *J. Acoust. Soc. Am.*, vol. 91, pp. 1662–1676, 1992.

[5] M. E. Lockwood, D. L. Jones, R. C. Bilger, C. R. Lansing, W. D. O'Brien, B. C. Wheeler, and A. S. Feng, "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms," *J. Acoust. Soc. Am.*, vol. 115, no. 1, pp. 379–391, 2004.

[6] Y. Mori, T. Takatani, H. Saruwatari, K. Shikano, T. Hiekata, and T. Morita, "High-presence hearing-aid system using DSP-based real-time blind source separation module," in *Proc. ICASSP*, pp. 609–612, 2007.

[7] K. Reindl, Y. Zheng, and W. Kellermann, "Speech enhancement for binaural hearing aids based on blind source separation," in *Proc. ISCCSP*, pp. 609–612, 2010.

[8] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. ICA*, pp. 601–608, 2006.

[9] T. Kim, T. Eltoft, and T. W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. ICA*, pp. 165–172, 2006.

[10] T. Kim, H. T. Attias, S. Y. Lee, and T. W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2007.

[11] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, pp. 189–192, 2011.

[12] N. Ono, "Auxiliary-function based independent vector analysis with power of vector-norm type weighting functions," in *Proc. APSIPA ASC*, 2012.

[13] N. Ono, "Fast stereo independent vector analysis and its implementation on mobile phone," in *Proc. IWAENC*, 2012.

[14] D. Mauler and R. Martin, "A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement," in *Proc. EUSIPCO*, pp. 222–226, 2007.

[15] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Trans. CAS I*, vol. 57, no. 7, pp. 1431–1438, 2010.

[16] T. Taniguchi, N. Ono, A. Kawamata, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proc. HSCMA*, pp. 107–111, 2014.

[17] J. Agnew and J. M. Thornton, "Just noticeable and objectionable group delays in digital hearing aids," *J. Am. Acad. Audiol.*, vol. 11, no. 6, pp. 330–336, 2000.

[18] M. A. Stone, B. C. J. Moore, K. Meisenbacher, and R. P. Derleth, "Tolerable hearing aid delays. V. Estimation of limits for open canal fittings," *Ear Hear.*, vol. 29, pp. 601–617, 2008.

[19] R. Heurig and J. Chalupper, "Acceptable processing delay in digital hearing aids," *Hear. Rev.*, vol. 17, no. 1, pp. 28–31, 2010.

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

[21] B. Gardner and K. Martin, "HRTF measurement of a KEMAR dummy-head microphone," 1994 [Online], Available: http://sound.media.mit.edu/resources/KEMAR.html. [Accessed: 10 Sep. 2016].

[22] "Real World Computing Project News Speech Corpus," http://research.nii.ac.jp/src/en/RWCP-SP99.html. [Accessed: 11 Sep. 2016].

[23] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.