Semi-Supervised Enhancement and Suppression of Self-Produced Speech Using Correspondence between Air- and Body-Conducted Signals

Moe Takada[1,†], Shogo Seki[1,‡], Patrick Lumban Tobing[1], Tomoki Toda[2]
[1]*Graduate School of Informatics* / [2]*Information Technology Center, Nagoya university, Nagoya, Japan*
{[†]takada.moe, [‡]seki.shogo}@g.sp.m.is.nagoya-u.ac.jp

***Abstract*—We propose a semi-supervised method for enhancing and suppressing self-produced speech recorded with wearable air- and body-conductive microphones. Body-conducted signals are robust against external noise and predominantly contain self-produced speech. As a result, these signals provide informative acoustical clues when estimating a linear filter to separate a mixed signal into self-produced speech and background noise. In a previous study, we proposed a blind source separation method for handling air- and body-conducted signals as a multi-channel signal. While our previously proposed method demonstrated the superior performance that can be achieved by using air- and body-conducted signals in comparison to using only air-conducted signals, the enhanced and suppressed air-conducted signals tended to be contaminated with the acoustical characteristics of the body-conducted signals due to the nonlinear relationship between these signals. To address this issue, in this paper, we introduce a new source model which takes into consideration the correspondence between these signals and incorporates them within a semi-supervised framework. Our experimental results reveal that this new method alleviates the negative effects of using the acoustical characteristics of the body-conducted signals, outperforming our previously proposed method, as well as conventional methods, under a semi-supervised condition.**

***Index Terms*—Self-produced speech, Semi-supervised speech enhancement and suppression, Air- and body-conducted signals**

I. Introduction

In recent years, compact, high-performance, wireless audio recording devices, such as Bluetooth microphones, have been developed, and these devices are increasingly being used for acoustic scene classification and event detection. As a result, it is expected that various sound signal processing applications involving the use of wearable audio interfaces will soon be developed. Audio samples recorded with wearable devices usually contain a mixture of sound source signals, such as the user's own speech and ambient environmental sounds. Depending on the focus of the application, different sound sources can be designated as the target signals. For example, wearable microphones should be ideal for obtaining only the user's speech for applications involving speech recognition. Similarly, only ambient environmental sounds could be obtained for applications involving acoustic scene classification or acoustic event detection, by suppressing the user's speech. Therefore, sound source separation techniques which can extract the desired target source signals from the observed mixture of signals are essential.

Blind source separation (BSS) [1] is a well-known and long-used technique for separating the underlying source signals present in an observed mixture of signals received by a microphone array. BSS methods based on independent component analysis (ICA) [2], [3] separate the observed signal by estimating a linear separation filter, under the assumption that source signals present in the observed signal are statistically independent of each other. Independent vector analysis (IVA) [4], [5], a natural extension of ICA, solves the permutation problems which occur in frequency-domain ICA (FDICA) [6], [7], resulting in better performance. Recently, independent low-rank matrix analysis (ILRMA) [8] has been proposed, which incorporates a source model based on non-negative matrix factorization (NMF) [9]–[11] into IVA, achieving impressive performance. While microphone arrays placed at a fixed point are generally used in BSS studies, in order to develop applications for wearable audio devices it will be necessary to develop BSS techniques for wearable, roving microphone arrays.

As one possible source separation technique for wearable audio devices, in this paper we propose a system which extracts the wearer's own speech (i.e., the user's speech) from the mixed signals recorded in noisy environments, which can be used for speech applications, or to suppress the user's speech in order to extract only the background sounds for environmental sound applications. In addition to a conventional wearable microphone, skin-attached microphones are also used as they primarily record the user's speech while suppressing external noise [12], [13]. We use one of the promising skin-attached microphones, a non-audible murmur (NAM) microphone [14] developed to detect very soft, whispered speech. These microphones are expected to be useful in the development of next-generation audio interfaces because they can record a wide variety of self-produced speech, not only NAM but also normal speech.

In our previous work [15], we proposed a self-produced speech enhancement and suppression method using body-conducted signals captured with a NAM microphone, in addition to air-conducted signals captured by a conventional microphone. Although the sound quality of the body-conducted signal was significantly degraded due to the effects of body-conduction, such as strong attenuation of its high-frequency components, it was robust against external noise, therefore this

can be used effectively for the enhancement and suppression of self-produced speech. Our previously proposed method [15] directly applied ILRMA to air- and body-conducted signals by handling them as an observed multichannel signal, resulting in superior performance to ILRMA using air-conducted signals only. However, we also found that the enhanced, self-produced speech signal tended to be contaminated with the acoustic characteristics of the body-conducted signals, due to the nonlinear relationship between the air- and body-conducted signals.

To address this issue, in this study we propose a semi-supervised method of speech enhancement and suppression for self-produced speech which is capable of handling the nonlinear relationship between air- and body-conducted signals. While the conventional method uses a regular source model, in which air- and body-conducted signals are represented with shared spectral patterns, the proposed method adopts an improved source model which can represent the individual spectral patterns of air- and body-conducted signals while sharing the temporal activation structure of the self-produced speech. We develop a semi-supervised framework for self-produced speech enhancement and suppression based on the proposed method, and experimentally demonstrate that it significantly outperforms the conventional method.

II. General formulation

Suppose that there are N source signals and that the mixed signal is recorded with M microphones. Let us denote the short-time Fourier transform (STFT) coefficients of source signal $\mathbf{s}_{ij}$, mixture signal $\mathbf{x}_{ij}$, and separated signal $\mathbf{y}_{ij}$ as:

$$\mathbf{s}_{ij} = [s_{ij,1}, \ldots, s_{ij,N}]^{\mathsf{T}} \in \mathbb{C}^N, \tag{1}$$

$$\mathbf{x}_{ij} = [x_{ij,1}, \ldots, x_{ij,M}]^{\mathsf{T}} \in \mathbb{C}^M, \tag{2}$$

$$\mathbf{y}_{ij} = [y_{ij,1}, \ldots, y_{ij,M}]^{\mathsf{T}} \in \mathbb{C}^N, \tag{3}$$

where $(\cdot)^{\mathsf{T}}$ represents the transpose and $\mathbb{C}$ denotes complex numbers. The frequency and time indices are represented as i and j, respectively. We assume that the mixing process can be modeled using the time-invariant mixing matrix $\mathbf{A}_i = [\mathbf{a}_{i,1}, ..., \mathbf{a}_{i,N}] \in \mathbb{C}^{M\times N}$ as follos:

$$\mathbf{x}_{ij} = \mathbf{A}_i \mathbf{s}_{ij}. \tag{4}$$

If $N = M$ and $\mathbf{A}_i$ is regular, there exists a separation matrix $\mathbf{W}_i = [\mathbf{w}_{i,1}, ..., \mathbf{w}_{i,M}]^{\mathsf{H}} \in \mathbb{C}^{M\times M}$ that satisfies $\mathbf{A}_i^{-1} = \mathbf{W}_i$. Thus, the separated signal is given by:

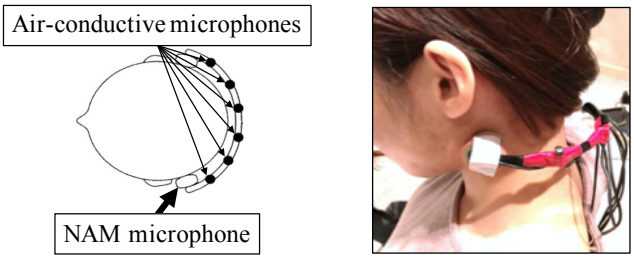
$$\mathbf{y}_{ij} = \mathbf{W}_i \mathbf{x}_{ij}. \tag{5}$$

In ICA-based source separation methods, including IVA and ILRMA, the separation matrix $\mathbf{W}_i$ is estimated assuming that the source signals are statistically independent of each other.

III. Conventional method

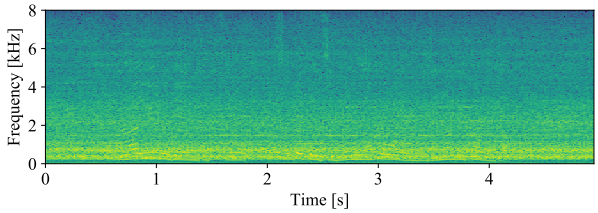
A. Multichannel recording with air- and body-conductive microphones

Several conventional speech enhancement methods using air- and body-conductive microphones have been proposed.

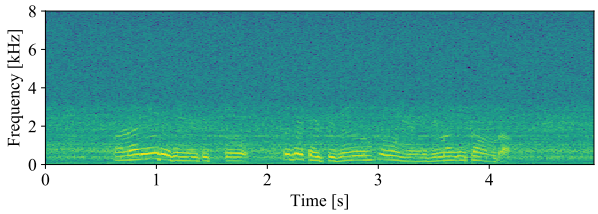


(a) Microphone positions (b) Recording environment

Fig. 1. A neckband-type recording device, where multiple air-conductive microphones and a NAM (body-conductive) microphone are positioned along the neckband.



(a) Air-conducted signal



(b) Body-conducted signal

Fig. 2. Spectrograms of (a) air-conducted signal and (b) body-conducted signal recorded in a noisy environment.

Figure 1 shows a neckband-type wearable recording device which contains multiple air-conductive microphones and a NAM microphone. The air-conductive microphones are installed on the device at equal intervals, and are set at the back of the speaker's neck. The NAM microphone is also installed on the neckband, and is attached at the position shown in Fig. 1 (a).

Figure 2 shows spectrograms of air- and body-conducted speech signals recorded under a 70 dBA noisy condition. We can see in Fig. 2 that the body-conducted signal clearly captures the components of self-produced speech, however the high-frequency components are suppressed due to the mechanical properties of body conduction [16]. Therefore, the sound quality of the body-conducted signal is significantly degraded.

B. Enhancement and suppression of self-produced speech using air- and body-conducted signals

The conventional method of speech enhancement and suppression using air- and body-conducted signals treats an M-channel air-conducted signal and a single channel body-conducted signal as a multichannel signal. ILRMA is directly applied to this $(M+1)$-channel signal, consisting of air- and body-conducted signals $\mathbf{x}_{ij}^{(\mathrm{A})} (= [x_{ij,1}^{(\mathrm{A})}, \ldots, x_{ij,M}^{(\mathrm{A})}])$ and $x_{ij}^{(\mathrm{B})}$, in order to estimate the $(M+1)\times(M+1)$ separation filter and the $M+1$ separated signals $y_{ij,1}, \ldots, y_{ij,N}$. Here, $\cdot^{(\mathrm{A})}$ and

$.^{(\mathrm{B})}$ represent air- and body-conducted signals, respectively. The conventional method models the power spectrograms of the source signals $s_{ij,1}, \ldots, s_{ij,N}$ as follows:

$$|s_{ij,n}|^2 = \sum\nolimits_k z_{nk} t_{ik} v_{kj}, \tag{6}$$

where k is the basis index, and $t_{ik} \geq 0$ and $v_{kj} \geq 0$ are components of spectral patterns and time-varying activations, respectively. Note that $z_{nk} \in [0, 1]$ satisfies $\sum_k z_{nk} = 1$ and represents the contribution of the k-th spectral pattern to the n-th source signal.

The separation algorithm iteratively updates the separation matrix and source model parameters. The update rules for the source model parameters can be derived by using the Majorization-Minimization (MM) algorithm [17] and the separation matrix can be updated using the Iterative-Projection algorithm [18].

After estimating a separation filter, the n-th multichannel separated signals $\hat{\mathbf{y}}_{ij,n} = [y^{(\mathrm{A})}_{ij,n1}, \ldots, y^{(\mathrm{A})}_{ij,nM}, y^{(\mathrm{B})}_{ij,n}]^{\mathsf{T}} \in \mathbb{C}^{M+1}$ are generated by using the projection back method [19]:

$$\hat{\mathbf{y}}_{ij,n} = \mathbf{W}_i^{-1} \mathbf{M}_n \mathbf{y}_{ij}, \tag{7}$$

where $\mathbf{M}_n$ is a diagonal matrix to extract only the n-th component of $\mathbf{y}_{ij}$ by masking the other components. The separated signal that has the largest power at the channel corresponding to the body-conducted signal is then identified as the self-produced speech. Once the self-produced speech is estimated, a Wiener filter [20] is applied to the observed air-conducted signals to enhance and suppress the speech.

C. Drawback

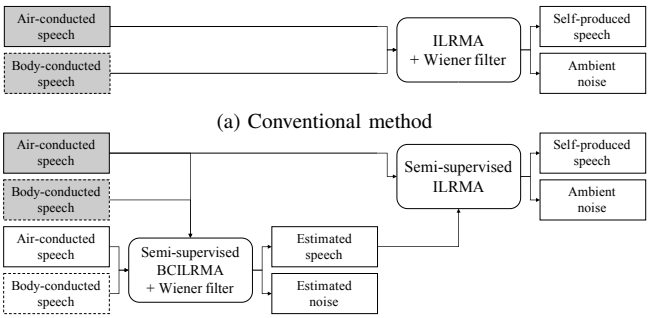
While the conventional method has been shown to improve enhancement and suppression performances for self-produced speech, one drawback is that the separated signals can be contaminated by the acoustic characteristics of the body-conducted signals, such as the degradation of the sound quality. Because there is a nonlinear relationship between air- and body-conducted signals, the mixing process assumed in (4) is not valid between these signals. Consequently, the linear separation given in (5) causes this contamination issue.

IV. Proposed Method Using Correspondence Between Air- and Body-Conducted Signals

A. Overview

To address this issue, in this study we propose a new source model which takes the nonlinear relationship between air- and body-conducted signals into account, which capable of applying the linear separation to only the air-conducted signals. The proposed method modifies ILRMA so that it can use coupled spectral patterns, making it possible to model the nonlinear relationship between air- and body-conducted signals. In this paper, we call the proposed source modeling approach "joint source modeling" and the modified method of ILRMA is referred to as "basis-coupled ILRMA" (BCILRMA).

Figure 3 shows flowcharts of the conventional and the proposed methods. The proposed method consists of a sequential process of semi-supervised BCILRMA for air- and body-conducted signals, followed by semi-supervised ILRMA for the air-conducted signals. First, using clean, self-produced speech data, the semi-supervised BCILRMA is applied to air- and body-conducted signals to estimate the air-conducted self-produced speech signal. Then, the estimated speech is modeled using non-negative tensor factorization (NTF) [21] to initialize the source model parameters of the semi-supervised ILRMA in the second step. Finally, the semi-supervised ILRMA is applied to only the air-conducted signals, to separate them into a self-produced speech signal and an ambient environmental sound signal.



(a) Conventional method

(b) Proposed method

Fig. 3. Flowcharts of (a) conventional method and (b) proposed method, where air- and body-conducted speech are denoted with solid and dotted frames, respectively, and noisy components are highlightened in gray.

B. Basis-coupled ILRMA (BCILRMA)

To model the nonlinear relationship between the air- and body-conducted signals, we employ the following new, joint-source model:

$$|s^{(\mathrm{A})}_{ij,n}|^2 = \sum\nolimits_k z^{(\mathrm{A})}_{nk} t^{(\mathrm{A})}_{ik} v_{kj}, \tag{8}$$

$$|x^{(\mathrm{B})}_{ij}|^2 = \sum\nolimits_k t^{(\mathrm{B})}_{ik} v_{kj}. \tag{9}$$

These source models share v_{kj} which represents time-varying components of air- and body-conducted signals, while having individual spectral patterns $t^{(\mathrm{A})}_{ik}$ and $t^{(\mathrm{B})}_{ik}$. Note that $t^{(\mathrm{B})}_{ik}$ captures the nonlinear relationship between these signals. Therefore, different acoustic characteristics of air- and body-conducted signals are modeled using the coupled spectral patterns while sharing the same temporal activation structure between those two signals. The objective function of BCILRMA is given by:

$$\mathcal{L} = \mathcal{L}^{(\mathrm{A})}_{\mathrm{ILRMA}} + \mathcal{L}^{(\mathrm{B})}_{\mathrm{ISNMF}}, \tag{10}$$

$$\mathcal{L}^{(\mathrm{A})}_{\mathrm{ILRMA}} = \sum\nolimits_{i,j,n} \left[\frac{|(\mathbf{w}^{(\mathrm{A})}_{i,n})^{\mathsf{H}} \mathbf{x}^{(\mathrm{A})}_{ij}|^2}{\sum_k z^{(\mathrm{A})}_{nk} t^{(\mathrm{A})}_{ik} v_{kj}} + \log \sum\nolimits_k z^{(\mathrm{A})}_{nk} t^{(\mathrm{A})}_{ik} v_{kj} \right] - 2J \log |\det \mathbf{W}^{(\mathrm{A})}_i|, \tag{11}$$

$$\mathcal{L}^{(\mathrm{B})}_{\mathrm{ISNMF}} = \sum\nolimits_{i,j} \left[\frac{|x^{(\mathrm{B})}_{ij}|^2}{\sum_k t^{(\mathrm{B})}_{ik} v_{kj}} + \log \sum\nolimits_k t^{(\mathrm{B})}_{ik} v_{kj} \right], \tag{12}$$

where $\mathcal{L}^{(\mathrm{A})}_{\mathrm{ILRMA}}$ represents the objective function of ILRMA for the air-conducted signals. Note that the separation filter $\mathbf{W}^{(\mathrm{A})}_i$ is applied only to the observed air-conducted signals. On the other hand, $\mathcal{L}^{(\mathrm{B})}_{\mathrm{ISNMF}}$ is the objective function of the NMF with

Itakura-Saito divergence for the body-conducted signal [22]. The update rules for these parameters can be derived in the same manner as those for ILRMA and ISNMF (we have omitted these update rules here, due to space limitations).

C. Semi-supervised BCILRMA for air- and body-conducted signals

Using the clean, self-produced speech data from the air- and body-conducted self-produced speech signals, the coupled spectral dictionaries for the self-produced speech are trained in advance. In semi-supervised BCILRMA, these coupled spectral dictionaries are fixed and the other parameters, such as the other dictionaries and all temporal activations are optimized. Then, the separated self-produced speech signal and ambient environmental signal are determined. Finally, a single channel Wiener filter is applied separately to the observed air-conducted signals in each channel, in order to extract the self-produced speech signals. The Wiener gain $G_{ij,m}$ for the air-conducted, self-produced speech can be expressed as follows:

$$G_{ij,m} = \|\hat{\mathbf{y}}_{ij,m}^{(\mathrm{S})}\|_2^2 / (\|\hat{\mathbf{y}}_{ij,m}^{(\mathrm{S})}\|_2^2 + \|\hat{\mathbf{y}}_{ij,m}^{(\mathrm{N})}\|_2^2), \qquad (13)$$

where $\hat{\mathbf{y}}_{ij,m}^{(\mathrm{S})}$ and $\hat{\mathbf{y}}_{ij,m}^{(\mathrm{N})}$ are the separated self-produced speech and the ambient environmental sounds, respectively.

D. Semi-supervised ILRMA for air-conducted signals

The estimated, air-conducted speech signals are less affected by ambient sound, however they still tend to suffer from some artifacts caused by the Wiener filter. Therefore, in our proposed method, these signals are used to initialize some of the source model parameters used in the second stage of processing (semi-supervised ILRMA), but only for the air-conducted signals. As mentioned earlier, NTF is applied to the estimated air-conducted self-produced speech signals to determine the corresponding source model parameters. After fixing some or all of the parameters, the second stage of semi-supervised ILRMA is performed, where the hidden variable corresponding to the self-produced speech $z_{nk}^{(\mathrm{A})}$ is set to 1. Finally, the separated self-produced speech and the ambient environmental sounds are determined by applying the obtained separation filter to the observed air-conducted signals. Consequently, our proposed method uses a first stage of separation, based on BCILRMA, to obtain prior information about the targeted speech signals for a second stage of separation using semi-supervised ILRMA.

V. Experimental evaluation

A. Experimental settings

The proposed method was experimentally evaluated by enhancing and suppressing self-produced speech recorded under noisy conditions. The experimental conditions are shown in Table I. Self-produced speech and environmental ambient sound were recorded separately and superimposed to generate noisy speech. We used the neckband-type wearable recording device shown in Fig. 1, which contains multiple air-conductive microphones and a NAM microphone. Our self-produced speech consisted of 50 sentences uttered by one female, Japanese speaker, i.e., this evaluation was performed in a speaker-dependent setting. Crowd noise with a sound pressure level of 70 dBA was used for the ambient environmental sound. Six noise sources were arranged at intervals of 60 degrees around the speaker, at a distance of 2 meters from the speaker, with the location directly in front of the speaker designated as zero degrees.

TABLE I
Experimental conditions

Training data	32 sentences
Evaluation data	18 sentences
Sampling frequency	24 kHz
Frame size	21.3 ms (512 pt)
Shift size	10.7 ms (256 pt)
# basis spectra	16

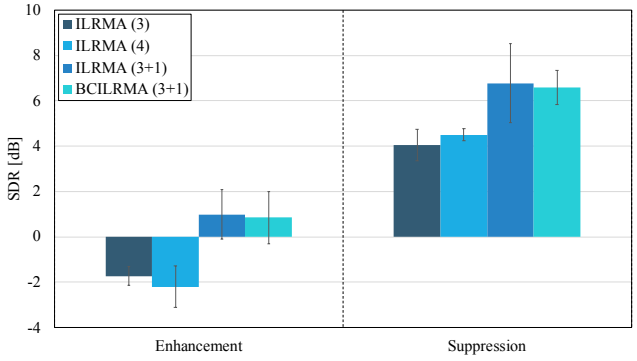
TABLE II
Experimental configurations (proposed method)

	Self-produced speech		Ambient noise	
	Spectral basis	Temporal activation	Spectral basis	Temporal activation
Fixed-A	Estimate	Fixed	Estimate	
Fixed-B	Fixed	Estimate	Estimate	
Fixed-AB	Fixed	Fixed	Estimate	

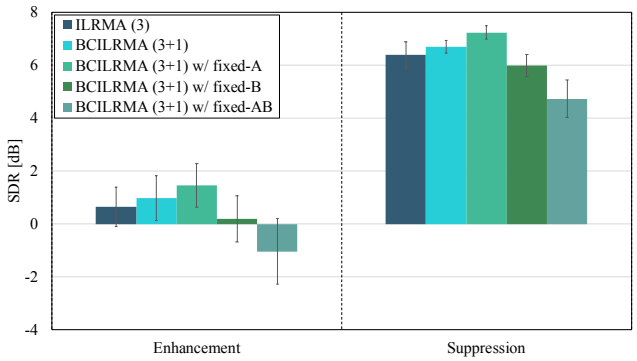
We conducted performance evaluations under both blind and semi-supervised conditions. Under the blind condition, regular ILRMAs with three- and four-channel air-conducted signals were tested as baseline methods (ILRMA (3) and ILRMA (4)), and these methods were compared with the conventional method (ILRMA (3+1)) and the proposed BCILRMA method (BCILRMA (3+1)). Under the semi-supervised condition, we evaluated ILRMA (3), BCILRMA (3+1), and BCILRMA (3+1) with a post-processing based on semi-supervised ILRMA (3), where three different source model parameter settings for the post-processing were investigated by fixing either or both of the temporal activations and the spectral bases of the self-produced speech (BCILRMA (3+1) w/ fixed-A, BCILRMA (3+1) w/ fixed-B and BCILRMA (3+1) w/ fixed-AB, as summarized in Table II). We used the signal-to-distortion ratio (SDR) [23] as the performance measurement criterion.

B. Experimental results

Figure 4 shows a comparison of the performance of each method. We can see from these results that in a blind setting the conventional and the proposed BCILRMA methods, which both jointly use air- and body-conducted signals, outperformed the baseline methods, which only use air-conducted signals. Table III shows the mel-cepstral distances (Mel-CDs) for each method, where Mel-CD is a well-known metric to capture sound quality often used in speech syntehsis research. Thus, we can confirm that the proposed BCILRMA method significantly improves the quality of self-produced speech, while achieving comparable performances as the conventional method. In a semi-supervised setting, although the conventional ILRMA (3+1) also tended to improve performance, the proposed method in configuration "BCILRMA (3+1) w/ fixed-A" yielded the best performance. These results suggest that the use of the body-conducted signal is helpful for improving speech enhancement and suppression performance for



(a) Blind condition.



(b) Semi-supervised condition.

Fig. 4. Self-produced speech enhancement and suppresion performances under (a) blind and (b) semi-supervised conditions, where the error bars show the 95% confidence intervals.

TABLE III
MEL-CEPSTRAL DISTANCES FOR SELF-PRODUCED SPEECH

Method	ILRMA (3)	ILRMA (3+1)	BCILRMA (3+1)
Mel-CD [dB]	8.75	7.71	6.25

self-produced speech, and that the proposed semi-supervised method is the most effective approach.

When comparing the performance of the various configurations of the proposed method, we can see that the use of fixed activation (BCILRMA (3+1) w/ fixed-A) is the most effective approach. In this implementation, the temporal activation structure of the self-produced speech signal is determined during first-stage separation using BCILRMA. Therefore, the proposed method can also improve separation performance when using ILRMA by effectively using the body-conducted signal, in particular, to determine the activation patterns of the speech signal.

VI. CONCLUSION

In this paper, we have proposed a semi-supervised speech enhancement and suppression method for self-produced speech using the correspondence between air- and body-conducted signals recorded with a wearable device containing multiple air-conductive microphones and a NAM microphone. Our proposed method uses BCILRMA to model the differing acoustic characteristics of the air- and body-conducted signals, and incorporates this model within a semi-supervised speech enhancement framework. Our experimental results demonstrated that the proposed BCILRMA-based method significantly improves enhancement and suppression performance for self-produced speech, and that performance is further improved within a semi-supervised framework.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number JP17H01763 and JST CREST Grant Number JPMJCR19A3.

REFERENCES

[1] P. Comon and C. Jutten, "Handbook of Blind Source Separation, Independent Component Analysis and Applications," *Academic Press*, 2010.
[2] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, pp. 287–314, 1994.
[3] S. Makino, S. Araki, R. Mukai and H. Sawada, "Audio Source Separation based on Independent Component Analysis," in *Proc. IEEE ISCAS*, pp. 668–671, 2004.
[4] T. Kim, I. Lee and T.-W. Lee, "Independent vector analysis: definition and algorithms," in *Proc. IEEE ACSSC*, pp. 1393–1396, 2006.
[5] A. Hiroe, "Solution of Permutation Problem in Frequency Domain ICA, Using Multivariate Probability Density Functions," in *Proc. ICA*, pp. 601–608, 2006.
[6] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1–3, pp. 21–34, 1998.
[7] H. Saruwatari, T. Kawamura and K. Shikano, "Blind sourceseparation for speech based on fast-convergence algorithmwith ICA and beamforming," in *Proc. Eurospeech 2001*, pp. 2603–2606, 2001.
[8] D. Kitamura, N. Ono, H. Sawada, H. Kameoka and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," *Audio Source Separation*, pp. 125–155, 2018.
[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788, 1999.
[10] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, pp. 556–562, 2001.
[11] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE WASPAA*, pp. 177–180, 2003.
[12] A. Subramanya, Z. Zhang, Z. Liu, and A. Acero, "Multisensory processing for speech enhancement and magnitude-normalized spectra for speech modeling," *Speech Communication*, vol. 50, no. 3, pp. 228–243, 2008.
[13] Z. Liu, Z. Zhang, A. Acero, J. Droppo and X. Huang, "Direct filtering for air-and bone-conductive microphones," in *Proc. IEEE MMSP*, pp. 363–366, 2004.
[14] Y. Nakajima, H. Kashioka, N. Campbell and K. Shikano, "Non-audible murmur (NAM) recognition," *IEICE Trans. Information and Systems*, vol. E89-D, no. 1, pp. 1–8, 2006.
[15] M. Takada, S. Seki and T. Toda, "Self-Produced Speech Enhancement and Suppression Method using Air-and Body-Conductive Microphones," in *Proc. APSIPA ASC*, pp. 1240–1245, 2018.
[16] T. Hirahara, M. Otani, S. Shimizu, T. Toda, K. Nakamura, Y. Nakajima and K. Shikano, "Silent-speech enhancement using body-conducted vocal-tract resonance signals," *Speech Communitcation*, vol. 52, no. 4, pp. 301–313, 2010.
[17] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.
[18] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE WASPAA*, pp. 189–192, 2011.
[19] 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.
[20] Liang Hong, J. Rosca, and R. Balan, "Independent Component analysis based single channel speech enhancement," in *Proc. IEEE ISSPIT*, pp. 522–525, 2003.
[21] A. Cichocki, R. Zdunek, A. H. Phan and S. Amari, "Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation," *John Wiley & Sons*, 2009.
[22] C. Févotte, N. Bertin and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.
[23] E. Vincent, R. Gribonval and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.