Semi-supervised Self-produced Speech Enhancement and Suppression Based on Joint Source Modeling of Air- and Body-conducted Signals Using Variational Autoencoder

Shogo Seki[1], *Moe Takada*[1], *Tomoki Toda*[2]

[1]Graduate School of Informatics / [2]Information Technology Center, Nagoya University

{seki.shogo,takada.moe}@g.sp.m.is.nagoya-u.ac.jp, tomoki@itcs.nagoya-u.ac.jp

Abstract

This paper proposes a semi-supervised method for enhancing and suppressing self-produced speech, using a variational autoencoder (VAE) to jointly model self-produced speech recorded with air- and body-conductive microphones. In speech enhancement and suppression for self-produced speech, body-conducted signals can be used as an acoustical clue since they are robust against external noise and include self-produced speech predominantly. We have previously developed a semi-supervised method taking an improved source modeling approach called the joint source modeling, which can capture a nonlinear correspondence of air- and body-conducted signals using non-negative matrix factorization (NMF). This allows enhanced and suppressed air-conducted self-produced speech to be prevented from contaminating by the characteristics of body-conducted signals. However, our previous method employs a rank-1 spatial model, which is effective but difficult to consider in more practical situations. Furthermore, joint source modeling depends on the representation capability of NMF. As a result, enhancement and suppression performances are limited. To overcome these limitations, this paper employs a full-rank spatial model and proposes a joint source modeling of air- and body-conducted signals using a VAE, which has shown to represent source signals more accurately than NMF. Experimental results revealed that the proposed method outperformed baseline methods.

Index Terms: Semi-supervised speech enhancement and suppression, Air- and body-conducted signal, Joint source modeling, Variational autoencoder (VAE)

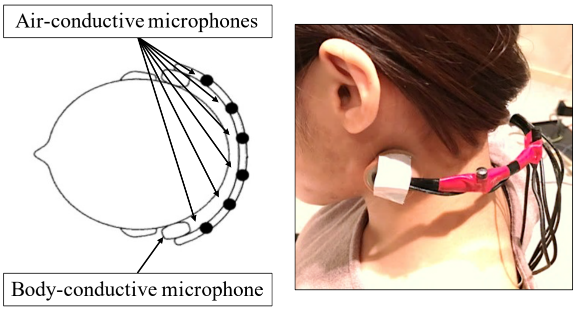
1. Introduction

Speech enhancement refers to the problem of extracting speech signals present in observed signals recorded under noisy conditions and improving their intelligibility [1]. With the recent developments of small and powerful recording devices and auditory scene analysis technologies [2, 3], speech enhancement is expected to build practical sound signal processing applications involving the use of wearable audio interfaces. Audio signals received by wearable devices contain several source signals such as the user's own speech and ambient environmental sounds, each of which can be applicable for various applications. For example, using user's utterances, i.e., self-produced speech, it is possible to develop systems that assist human activities such as speech retrieval [4, 5] and speech operation [6, 7]. It is also possible to design systems that monitor surrounding acoustic scenes and events [8, 9] by using ambient environmental sounds. Hence, to extract the target signals for desired applications, speech enhancement and suppression technologies for self-produced speech are essential.

In self-produced speech enhancement and suppression for wearable audio devices, one of the promising skin-attached body-conductive microphones, a non-audible murmur (NAM) microphone [10] has attracted attention. A NAM microphone is developed to detect very softly whispered speech, i.e., NAM, and it can record a wide variety of body-conducted signals, not only NAM but also normal speech. Although body-conducted signals recorded by a NAM microphone are suffered from strong attenuation of high-frequency components, they are robust against external noise, and they have larger power than air-conducted signals, and include self-produced speech predominantly [11]. These characteristics can be used as acoustical clues for the enhancement and suppression of self-produced speech. Moreover, NAM microphones can be easily installed to neckband-type wearable recording devices since they are used to be set at the back of the speaker‘s ear.

We have previously proposed self-produced speech enhancement and suppression methods using body-conducted signals captured with a NAM microphone and air-conducted signals captured by conventional microphones [12, 13]. In [12], a blind enhancement and suppression method was proposed, which uses the audio signals recorded with the air- and body-conductive microphones and estimates self-produced speech and ambient environmental sounds from the separated signals obtained by independent low-lank matrix analysis (ILRMA) [14]. Thanks to the use of body-conducted signals, this method can estimate self-produced speech effectively, outperforming that with air-conducted signals only. However, the sound quality of the estimated signals is suffered from the contamination caused by the characteristics of body-conducted signals. To address this issue, we developed an improved source model called the “joint source model”, which captures a nonlinear relationship between air- and body-conducted signals, and proposed a semi-supervised method for enhancing and suppressing self-produced speech [13]. This allows air- conducted signals to be prevented from contaminating by characteristics of body-conducted signals. However, our previous method employs a rank-1 spatial model based on ILRMA, which is useful but difficult to consider in wearable audio devices. Furthermore, joint source modeling depends on the representation capability of non-negative matrix factorization (NMF) [15], resulting in the performance limitation.

To overcome these limitations, this paper proposes a self-produced speech enhancement and suppression method using a variational autoencoder (VAE) [16] in the joint source modeling of air- and body-conducted signals. The proposed method employs a full-rank spatial model and integrates with the joint source modeling. Furthermore, a VAE-based semi-supervised speech enhancement methodology [17, 18, 19] is applied.



(a) Microphone positions (b) Recording environments

Figure 1: *An air- and body-conductive microphone array, where multiple air-conductive microphones and a body-conductive microphone (NAM microphone) are allocated on neckband.*

2. Related Works

2.1. Self-produced Speech Enhancement and Suppression Using Body-conducted Signals [12]

In [12], a multichannel air-conducted signal and a single channel body-conducted signal recorded with an air- and body-conductive microphone array in Figure 1 are treated as a multichannel signal. ILRMA is directly applied to this multichannel signal, and the same number of separated signals are estimated. Post-processing to classify the separated signals into self-produced speech and ambient environmental sounds is required, this can be solved by utilizing the characteristics of body-conducted signals. Since body-conducted signals have larger powers than air-conducted signals and include self-produced speech predominantly, 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. While this method shows superior performance to that uses air-conducted signals only, the separated signals can be contaminated by the acoustic characteristics of the body-conducted signals such as the degradation of the sound quality.

2.2. Self-produced Speech Enhancement and Suppression Considering Correspondence between Air- and Body-conducted Signals [13]

To address the contamination issue, in [13], we proposed a new source model taking the correspondence of air- and body-conducted signals into account, which capable of applying the linear separation to only the air-conducted signals. We modified ILRMA so that the spectral patterns of air- and body-conducted signals are modeled independently while sharing the same temporal activations. This allows us to model the nonlinear relationship between air- and body-conducted signals. We call this source modeling approach the joint source modeling and refer the modified variant of ILRMA to as “basis-coupled ILRMA” (BCILRMA). Moreover, a multi-step method was developed by incorporating ILRMA and BCILMA within a semi-supervised framework [13]. Our previous method enables us to avoid contamination and improve the performances by using a semi-supervised speech enhancement framework. However, our previous method employs a rank-1 spatial model based on ILRMA, which is difficult to consider in the use of wearable audio devices, such as longer reverberant conditions and obstacles, i.e., speaker‘s head. Another shortcoming is that our previous method uses NMF-based source models, which remains room to be improved.

3. Proposed Method

3.1. Overview

The proposed method starts by employing a full-rank spatial model to generalize conventional methods [12, 13]. Self-produced speech enhancement and suppression methods based on the joint source modeling using NMF and VAE are then derived.

3.2. General Formulation

Let us denote the short-time Fourier transform (STFT) coefficients of self-produced speech and ambient sounds recorded by a I-channel air- and body-conductive microphone array as $\{\mathbf{s}(\mathrm{A},f,t) \in \mathbb{C}^{I-1}, s(\mathrm{B},f,t) \in \mathbb{C}\}$, $\{\mathbf{n}(\mathrm{A},f,t) \in \mathbb{C}^{I-1}, n(\mathrm{B},f,t) \in \mathbb{C}\}$, where A and B represent the air- and body-conducted signals, and t and f are the time and frequency indices, respectively. The mixture signals are given as:

$$\mathbf{x}(\mathrm{A},f,t) = \mathbf{s}(\mathrm{A},f,t) + \mathbf{n}(\mathrm{A},f,t), \quad (1)$$

$$x(\mathrm{B},f,t) = s(\mathrm{B},f,t) + n(\mathrm{B},f,t). \quad (2)$$

We introduce the local Gaussian model (LGM) [20, 21, 22] so that each signal independently follows a zero-mean complex Gaussian distribution. The mixture signals can be written as:

$$\mathbf{x}(\mathrm{A},f,t) \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{V}_{\mathrm{X}}(\mathrm{A},f,t)\right), \quad (3)$$

$$x(\mathrm{B},f,t) \sim \mathcal{N}_{\mathbb{C}}\left(0, v_{\mathrm{X}}(\mathrm{B},f,t)\right). \quad (4)$$

$\mathbf{V}_{\mathrm{X}}(\mathrm{A},f,t)$ and $v_{\mathrm{X}}(\mathrm{B},f,t)$ are given as:

$$\mathbf{V}_{\mathrm{X}}(\mathrm{A},f,t) = v_{\mathrm{S}}(\mathrm{A},f,t)\mathbf{R}_{\mathrm{S}}(\mathrm{A},f) + v_{\mathrm{N}}(\mathrm{A},f,t)\mathbf{R}_{\mathrm{N}}(\mathrm{A},f), \quad (5)$$

$$v_{\mathrm{X}}(\mathrm{B},f,t) = v_{\mathrm{S}}(\mathrm{B},f,t) + v_{\mathrm{N}}(\mathrm{B},f,t), \quad (6)$$

where $(\cdot)_{\mathrm{S}}$ and $(\cdot)_{\mathrm{N}}$ represent the components of self-produced speech and ambient sounds, and $v(\cdot)$ and $\mathbf{R}(\cdot)$ denote the source variance and spatial covariance, respectively. Hence, given an observed mixture signal $\mathcal{X} = \{\mathbf{x}(\mathrm{A},f,t), x(\mathrm{B},f,t)\}_{f,t}$, using the source variance $\mathcal{V} = \{v_{\mathrm{S}}(\mathrm{A},f,t), v_{\mathrm{S}}(\mathrm{B},f,t), v_{\mathrm{N}}(\mathrm{A},f,t), v_{\mathrm{N}}(\mathrm{B},f,t)\}_{f,t}$ and the spatial covariance $\mathcal{R} = \{\mathbf{R}_{\mathrm{S}}(\mathrm{A},f), \mathbf{R}_{\mathrm{N}}(\mathrm{A},f)\}_f$ in source signals, the negative log-likelihood is written by:

$$\begin{aligned} &-\log p(\mathcal{X}|\mathcal{V},\mathcal{R}) \overset{c}{=} \\ &\sum\nolimits_{f,t}\left[\mathrm{tr}\left(\mathbf{X}(\mathrm{A},f,t)\mathbf{V}_{\mathrm{X}}^{-1}(\mathrm{A},f,t)\right) + \mathrm{logdet}\mathbf{V}_{\mathrm{X}}(\mathrm{A},f,t)\right] \\ &+\sum\nolimits_{f,t}\left[|x(\mathrm{B},f,t)|^2 v_{\mathrm{X}}^{-1}(\mathrm{B},f,t) + \log v_{\mathrm{X}}(\mathrm{B},f,t)\right], \quad (7) \end{aligned}$$

where $\overset{c}{=}$ denotes the equality up to constant terms and

$$\mathbf{X}(\mathrm{A},f,t) = \mathbf{x}(\mathrm{A},f,t)\mathbf{x}^{\mathsf{H}}(\mathrm{A},f,t). \quad (8)$$

3.3. Joint Source Modeling Using NMF

As with the previous study [13], we apply the joint source modeling, which assumes that the air- and body-conducted signals of source $j \in \{\mathrm{S},\mathrm{N}\}$ take the following structure by NMF:

$$\begin{bmatrix} v_j(\mathrm{A},f,t) \\ v_j(\mathrm{B},f,t) \end{bmatrix} = \sum\nolimits_k \begin{bmatrix} h_{j,k}(\mathrm{A},f) \\ h_{j,k}(\mathrm{B},f) \end{bmatrix} u_{j,k}(t), \quad (9)$$

where k is the index of the number of bases. Since these air- and body-conducted source models share the temporal activation $u_{j,k}(t)$, while having individual spectral patterns $h_{j,k}(\mathrm{A},f)$

and $h_{j,k}(\mathrm{B}, f)$, the nonlinear relationship between these signals is taken to be account.

In semi-supervised scenarios, using fixed spectral patterns of self-produced speech $\{h_{\mathrm{S},k}(\mathrm{A}, f)\}_{k,f}$ and $\{h_{\mathrm{S},k}(\mathrm{B}, f)\}_{k,f}$ trained by dataset, the estimation algorithm iteratively updates the other source model parameters $\{h_{\mathrm{N},k}(\mathrm{A}, f)\}_{k,f}$, $\{h_{\mathrm{N},k}(\mathrm{B}, f)\}_{k,f}$, $\{u_{\mathrm{S},k}(t)\}_{k,t}$, and $\{u_{\mathrm{N},k}(t)\}_{k,t}$, and spatial covariances $\{\mathbf{R}_{\mathrm{S}}(\mathrm{A}, f)\}_f$ and $\{\mathbf{R}_{\mathrm{N}}(\mathrm{A}, f)\}_f$. By using the Majorization-Minimization (MM) algorithm [23, 24], the optimal updates of each parameter can be derived as:

$$h_{\mathrm{N},k}(\mathrm{A}, f) \leftarrow h_{\mathrm{N},k}(\mathrm{A}, f)\sqrt{\frac{\sum_t u_{\mathrm{N},k}(t) l_{\mathrm{N}}(\mathrm{A}, f, t)}{\sum_t u_{\mathrm{N},k}(t) m_{\mathrm{N}}(\mathrm{A}, f, t)}}, \quad (10)$$

$$h_{\mathrm{N},k}(\mathrm{B}, f) \leftarrow h_{\mathrm{N},k}(\mathrm{B}, f)\sqrt{\frac{\sum_t u_{\mathrm{N},k}(t) l_{\mathrm{N}}(\mathrm{B}, f, t)}{\sum_t u_{\mathrm{N},k}(t) m_{\mathrm{N}}(\mathrm{B}, f, t)}}, \quad (11)$$

$$u_{j,k}(t) \leftarrow u_{j,k}(t)\sqrt{\frac{\sum_{c,f} h_{j,k}(c, f) l_j(c, f, t)}{\sum_{c,f} h_{j,k}(c, f) m_j(c, f, t)}}, \quad (12)$$

$$\mathbf{R}_j(\mathrm{A}, f) \leftarrow \mathbf{\Lambda}_j^{-1}(\mathrm{A}, f) \# \left(\mathbf{R}_j(\mathrm{A}, f)\mathbf{\Omega}_j(\mathrm{A}, f)\mathbf{R}_j(\mathrm{A}, f)\right), \quad (13)$$

where $c \in \{\mathrm{A}, \mathrm{B}\}$ denotes conducted signal and $\#$ denotes the geometric mean between two positive definite matrices [25], and

$$l_j(\mathrm{A}, f, t) = \mathrm{tr}(\mathbf{V}_{\mathrm{X}}^{-1}(\mathrm{A}, f, t)\mathbf{X}(\mathrm{A}, f, t)\mathbf{V}_{\mathrm{X}}^{-1}(\mathrm{A}, f, t)\mathbf{R}_j(\mathrm{A}, f)), \quad (14)$$

$$m_j(\mathrm{A}, f, t) = \mathrm{tr}\left(\mathbf{V}_{\mathrm{X}}^{-1}(\mathrm{A}, f, t)\mathbf{R}_j(\mathrm{A}, f)\right), \quad (15)$$

$$l_j(\mathrm{B}, f, t) = |x(\mathrm{B}, f, t)|^2\, v_{\mathrm{X}}^{-2}(\mathrm{B}, f, t), \quad (16)$$

$$m_j(\mathrm{B}, f, t) = v_{\mathrm{X}}^{-1}(\mathrm{B}, f, t), \quad (17)$$

$$\mathbf{\Lambda}_j(\mathrm{A}, f) = \sum\nolimits_t v_j(\mathrm{A}, f, t)\mathbf{V}_{\mathrm{X}}^{-1}(\mathrm{A}, f, t), \quad (18)$$

$$\mathbf{\Omega}_j(\mathrm{A}, f) = \sum\nolimits_t v_j(\mathrm{A}, f, t)\mathbf{V}_{\mathrm{X}}^{-1}(\mathrm{A}, f, t)\mathbf{X}(\mathrm{A}, f, t)\mathbf{V}_{\mathrm{X}}^{-1}(\mathrm{A}, f, t). \quad (19)$$

3.4. Joint Source Modeling Using VAE

In the proposed method, we employ a VAE instead of NMF to represent self-produced speech $v_{\mathrm{S}}(\mathrm{A}, f, t)$ and $v_{\mathrm{S}}(\mathrm{B}, f, t)$, while modeling ambient sounds by NMF as well as the previous study [13].

Let $\tilde{\mathbf{S}}(\mathrm{A}) \in \mathbb{C}^{F \times T}$ and $\tilde{\mathbf{S}}(\mathrm{B}) \in \mathbb{C}^{F \times T}$ denote the complex spectrograms of self-produced speech recorded by air- and body-conductive microphones. In the joint source modeling using VAE, the spectrograms of air- and body-conducted signals are stacked along with frequency bin and treated as the network input. Given an concatenated spectrogram $\tilde{\mathbf{S}} = [\tilde{\mathbf{S}}^{\mathsf{T}}(\mathrm{A}), \tilde{\mathbf{S}}^{\mathsf{T}}(\mathrm{B})]^{\mathsf{T}} \in \mathbb{C}^{2F \times T}$, the encoder distribution $q_\phi(\mathbf{Z}|\tilde{\mathbf{S}})$ and the decoder distribution $p_\theta(\tilde{\mathbf{S}}|\mathbf{Z}, g)$ are expressed as a Gaussian distribution and a zero-mean complex Gaussian distribution:

$$q_\phi(\mathbf{Z}|\tilde{\mathbf{S}}) = \prod\nolimits_d \mathcal{N}_{\mathbb{C}}\left(z(d)|\mu_\phi(d; \tilde{\mathbf{S}}), \sigma_\phi^2(d; \tilde{\mathbf{S}})\right), \quad (20)$$

$$p_\theta(\tilde{\mathbf{S}}|\mathbf{Z}, g) = \prod\nolimits_{c,f,t} \mathcal{N}_{\mathbb{C}}\left(\tilde{s}(\mathrm{c}, f, t)|0, v(\mathrm{c}, f, t)\right), \quad (21)$$

$$v(\mathrm{c}, f, t) = g\sigma_\theta^2(c, f, t; \mathbf{Z}) \quad (22)$$

where $z(d)$, $\mu_\phi(d; \tilde{\mathbf{S}})$, and $\sigma_\phi^2(d; \tilde{\mathbf{S}})$ represent the d–th elements of a latent variable $\mathbf{Z}$ and encoder outputs $\mu_\phi(\tilde{\mathbf{S}})$ and $\sigma_\phi^2(\tilde{\mathbf{S}})$, $\sigma_\theta^2(c, f, t; \mathbf{Z})$ represent the (f, t)–th elements of the decoder output $\sigma_\theta^2(\mathbf{Z})$, and g is the global scale of generated spectrogram. During VAE training, both the encoder network and decoder network parameters ϕ and θ are trained using the following objective function:

$$\mathcal{J}(\phi, \theta) = \mathbb{E}_{\tilde{\mathbf{S}} \sim p(\tilde{\mathbf{S}})}[\mathbb{E}_{\mathbf{Z} \sim q(\mathbf{Z})}[\log p(\tilde{\mathbf{S}}|\mathbf{Z})] - \mathrm{KL}(q(\mathbf{Z}|\tilde{\mathbf{S}})||p(\mathbf{Z}))], \quad (23)$$

where $\mathbb{E}_{\tilde{\mathbf{S}} \sim p(\tilde{\mathbf{S}})}[\cdot]$ denotes the sample mean over the dataset and $\mathrm{KL}(\cdot||\cdot)$ denotes Kullback-Leivler (KL) divergence. Since the decoder distribution is designed to be the same form as the LGM, the decoder network $\sigma_\theta^2(\mathbf{Z})$ leads the same (negative) log-likelihood as (7), which can be seen as a generative model of self-produced speech. Thus, the air- and body-conducted variances of self-produced speech are given by:

$$\begin{bmatrix} v_{\mathrm{S}}(\mathrm{A}, f, t) \\ v_{\mathrm{S}}(\mathrm{B}, f, t) \end{bmatrix} = g \begin{bmatrix} \sigma_\theta^2(\mathrm{A}, f, t; \mathbf{Z}) \\ \sigma_\theta^2(\mathrm{B}, f, t; \mathbf{Z}) \end{bmatrix} \quad (24)$$

The iterative algorithm consists of updating the joint source model parameters for self-produced speech g and $\mathbf{Z}$, those for ambient sounds $\{h_{\mathrm{N},k}(\mathrm{A}, f)\}_{k,f}$, $\{h_{\mathrm{N},k}(\mathrm{B}, f)\}_{k,f}$, and $\{u_{\mathrm{N},k}(t)\}_{k,t}$, and spatial covariances $\{\mathbf{R}_{\mathrm{S}}(\mathrm{A}, f)\}_f$ and $\{\mathbf{R}_{\mathrm{N}}(\mathrm{A}, f)\}_f$, where g and $\mathbf{Z}$ are new parameters. From the MM algorithm, the optimal update for g is obtained as:

$$g \leftarrow g\sqrt{\frac{\sum_{c,f,t} \sigma_\theta^2(c, f, t; \mathbf{Z}) l_{\mathrm{S}}(c, f, t)}{\sum_{c,f,t} \sigma_\theta^2(c, f, t; \mathbf{Z}) m_{\mathrm{S}}(c, f, t)}}. \quad (25)$$

As shown in [26], it is possible to build a majorizer for the negative log-likelihood defined in (7). Hence, $\mathbf{Z}$ can be updated using backpropagation [27, 28], where the majorizer is used as the objective function (we have omitted the derivation, due to space limitations).

4. Experimental Evaluation

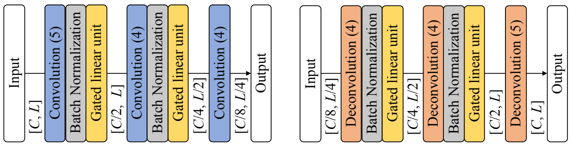
4.1. Experimental Settings

The proposed method was experimentally evaluated under a semi-supervised self-produced speech enhancement and suppression scenario. We used the neckband-type wearable recording device shown in Figure 1, where three-channel air-conducted signals and one-channel body-conducted signals were used. Self-produced speech and environmental ambient sound were recorded separately and superimposed to generate noisy speech. Our self-produced speech consisted of 50 Japanese sentences uttered by one Japanese female. 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. We used 32 utterances for training, and 18 utterances were used for evaluation. All the signals were sampled at 24 kHz. STFT analysis was conducted with a 21.3 ms window length and a 10.7 ms shift length.

We tested semi-supervised BCILRMA, that with post-processing [13], semi-supervised multichannel NMF (MNMF) [29], and VAE-NMF [18, 19] as baseline methods (dAB-NMF, dAB-NMF+, uA-NMF, and uA-VAE). These methods were compared with the proposed NMF-based and VAE-based methods (uAB-NMF and uAB-VAE). The categorization of each method is shown in Table 1. The number of NMF bases was set to 32 per each source signal, and the

Table 1: *Methods for comparison.*

Method	Conducted signal	Spatial model	Speech model
dAB-NMF	Air & Body	Rank-1	NMF
dAB-NMF+ [13]	Air & Body	Rank-1	NMF
uA-NMF [29]	Air	Full-rank	NMF
uA-VAE [18, 19]	Air	Full-rank	VAE
uAB-NMF	Air & Body	Full-rank	NMF
uAB-VAE	Air & Body	Full-rank	VAE



(a) Encoder network (b) Decoder network

Figure 2: *Architectures of (a) encoder and (b) decoder networks, where $[C, L]$ denotes the input channel and input length. Both convolution and deconvolution represent 1-dimensional operation.* (k) *represents kernel size.*

Itakura-Saito NMF (IS-NMF) [30] with 1000 iterations was used for training. We used the encoder and decoder networks shown in Figure 2 for our VAE, where the Adam [31] algorithm with a learning rate of 0.0001 was used for training. All the methods were initialized with 100 iterations and then ran for 100 iterations.

As the evaluation metrics, we used the signal-to-distortion ratio (SDR), the source image-to-spatial distortion ratio (ISR), the signal-to-inference ratio (SIR), the signal-to-artifact ratio (SAR) [32], the perceptual evaluation of speech quality (PESQ) [33], and the short-time objective intelligibility (STOI) [34].

4.2. Experimental Results

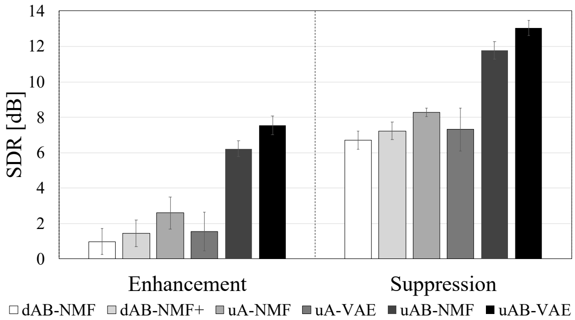


Figure 3: *SDR comparison of self-produced speech enhancement and suppression, where error bars show the 95 % confidence intervals.*

Table 2: *Averaged performances.*

Method	SDR	ISR	SIR	SAR	PESQ	STOI
dAB-NNF	3.8	8.6	7.9	11.2	1.2	0.6
dAB-NMF+	4.4	9.2	8.2	11.0	N/A	N/A
uA-NMF	5.4	11.1	10.2	10.4	1.2	0.7
uA-VAE	4.4	12.0	10.1	12.9	1.1	0.7
uAB-NMF	9.0	14.2	14.5	13.0	1.3	**0.8**
uAB-VAE	**10.3**	**15.6**	**16.9**	**14.2**	**1.7**	**0.8**

Figure 3 shows an SDR comparison of the self-produced speech enhancement and suppression performances of each method. First, we show the conventional performances of our previously proposed method. When comparing the SDR of the baseline methods which employ the rank-1 spatial models (dAB-NMF and dAB-NMF+), the previously proposed method slightly performs better than that without post-processing. We can see from these results that the post-processing works reasonably well, however, these performances are still limited. We next focus on the effectiveness of the full-rank spatial model. The comparison of dAB-NMF and uAB-NMF directly reflects the ability between the rank-1 and full-rank spatial models. We can confirm that the use of a full-rank spatial model provides significant performance improvements. Another finding is that the methods employing full-rank spatial models (uA-NMF and uA-VAE) can get comparable or better performances than those employing rank-1 spatial models, using air-conducted signals only. These results imply that the full-rank spatial model is more appropriate for practical microphone settings shown in Figure 1. Furthermore, focusing on the comparison of the NMF-based joint source modeling and the VAE-based joint source modeling (uAB-NMF and uAB-VAE), uAB-VAE got further improvements and achieved the best performances. On the other hand, the conventional VAE-based source modeling (uA-VAE) failed to improve the enhancement and suppression performances. This might be because the power of air-conducted signals was much smaller than that of body-conducted signals, which made the parameter estimation to the global scale g more sensitive and caused computational instability. These results indicate that the VAE is capable of representing self-produced speech more precisely, and the body-conducted signal helps stabilize parameter estimations in the VAE-based source modeling.

A comparison of the averaged performances of each method is shown in Table 2. The proposed VAE-based joint source modeling (uAB-VAE) yielded the best performances at each metric. Specifically, uAB-VAE gained 0.4 points larger than uAB-NMF at the PESQ evaluation, achieving significant improvement. We can confirm from this result that the VAE-based joint source modeling contributes to improving the sound quality of self-produced speech[1].

5. Conclusion

In this paper, we proposed a semi-supervised method for enhancing and suppressing self-produced speech, using a VAE to jointly model self-produced speech recorded with air- and body-conductive microphones. The proposed method generalizes our previous methods so that it can employ a full-rank spatial model to consider more practical recording settings. The proposed method was integrated with the joint source modeling, which can capture the nonlinear correspondence of air- and body-conducted signals. Furthermore, a VAE-based semi-supervised speech enhancement framework was then incorporated. From experimental evaluations using a neckband-type air- and body-conductive microphone array, our proposed method outperformed baseline methods, including our previous method, demonstrating the effectiveness of a full-rank spatial model and a VAE-based joint source modeling.

6. Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 17H01763, Japan and JST CREST Grant Number JPMJCR19A3, Japan.

[1] The audio samples can be found at `https://drive.google.com/drive/folders/1XVrRiiSav1EEf7zOp_iJIFXscpSiGWbU?usp=sharing`

7. References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[2] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An ieee aasp challenge," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4, 2013.

[3] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2018.

[4] J. . Van Thong, P. J. Moreno, B. Logan, B. Fidler, K. Maffey, and M. Moores, "Speechbot: an experimental speech-based search engine for multimedia content on the web," *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 88–96, 2002.

[5] E. Barnard, J. Schalkwyk, C. van Heerden, and P. J. Moreno, "Voice search for development," in *Interspeech 2010 – 10th Annual Conference of the International Speech Communication Association*, pp. 282–285, 2010.

[6] D. Valtchev and I. Frankov, "Service gateway architecture for a smart home," *IEEE Communications Magazine*, vol. 40, no. 4, pp. 126–132, 2002.

[7] S. Krstulović, "Audio event recognition in the smart home," in *Computational Analysis of Sound Scenes and Events*. Springer, 2018, pp. 335–371.

[8] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 158–161, 2005.

[9] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *IEEE International Conference on Multimedia and Expo*, pp. 1306–1309, 2005.

[10] Y. Nakajima, H. Kashioka, N. Campbell, and K. Shikano, "Non-audible murmur (nam) recognition," *IEICE Transactions on Information and Systems*, vol. 89, no. 1, pp. 1–8, 2006.

[11] 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 Communication*, vol. 52, no. 4, pp. 301–313, 2010.

[12] M. Takada, S. Seki, and T. Toda, "Self-produced speech enhancement and suppression method using air- and body-conductive microphones," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1240–1245, 2018.

[13] S. Seki, M. Takada, K. Takeda, and T. Toda, "Semi-supervised enhancement and suppression of self-produced speech using correspondence between air- and body-conducted signals," in *28th European Signal Processing Conference*, 2020, (Submitted).

[14] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1622–1637, 2016.

[15] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE International Workshop on Applications of Signal Processing to Audio and Acoustics*, no. 3, pp. 177–180, 2003.

[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014.

[17] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 716–720, 2018.

[18] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 101–105, 2019.

[19] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Semi-supervised multichannel speech enhancement with a deep speech prior," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2197–2212, 2019.

[20] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency gaussian source models," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, pp. 78–81, 2005.

[21] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local gaussian modeling," in *International Conference on Independent Component Analysis and Signal Separation*, pp. 775–782, 2009.

[22] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*. IGI global, 2011, pp. 162–185.

[23] J. De Leeuw and W. J. Heiser, "Convergence of correction matrix algorithms for multidimensional scaling," *Geometric representations of relational data*, pp. 735–752, 1977.

[24] D. R. Hunter and K. Lange, "A tutorial on mm algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[25] K. Yoshii, "Correlated tensor factorization for audio source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 731–735, 2018.

[26] H. Kameoka, H. Sawada, and T. Higuchi, "General formulation of multichannel extensions of nmf variants," in *Audio Source Separation*. Springer, 2018, pp. 95–124.

[27] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural computation*, vol. 31, no. 9, pp. 1891–1914, 2019.

[28] S. Seki, H. Kameoka, L. Li, T. Toda, and K. Takeda, "Underdetermined source separation based on generalized multichannel variational autoencoder," *IEEE Access*, vol. 7, no. 1, pp. 168 104–168 115, 2019.

[29] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[30] 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.

[31] D. P. Kingma and J. L. Ba, "Adam: Amethod for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[32] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[33] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 749–752, 2001.

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time $-$ frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.