A NOISE SUPPRESSION METHOD FOR BODY-CONDUCTED SOFT SPEECH BASED ON NON-NEGATIVE TENSOR FACTORIZATION OF AIR- AND BODY-CONDUCTED SIGNALS

Yusuke Tajiri† *Hirokazu Kameoka*‡ *Tomoki Toda*†

† Graduate School of Information Science, Nagoya University, Japan
‡ Media Information Laboratory, NTT Communication Science Laboratories, Japan
tajiri.yusuke@g.m.sp.is.nagoya-u.ac.jp, kameoka.hirokazu@lab.ntt.co.jp, tomoki@icts.nagoya-u.ac.jp

ABSTRACT

This paper presents a novel noise suppression method to enhance soft speech recorded with a special body-conductive microphone called nonaudible murmur (NAM) microphone. NAM microphone is capable of detecting extremely soft speech, but the recorded soft speech easily suffers from external noise due to its faint volume. To effectively suppress noise on the body-conducted signals, an external noise monitoring framework using an air-conducive microphone has been proposed. In this study, we propose a noise suppression method for this framework based on a probabilistic observation model robust against phase variations. In the proposed method, noise suppression process is formulated as a special case of non-negative tensor factorization of the observed air- and body-conducted signals. Experimental results demonstrate that 1) the proposed method consistently outperforms the conventional method under real noisy environments and 2) the proposed method effectively deals with speech acoustic changes caused by the Lombard reflex.

Index Terms— Silent speech communication, nonaudible murmur, noise suppression, external noise monitoring, non-negative matrix factorization

1. INTRODUCTION

Speaking is the most efficient way of human communication. In recent decades, the style of speech communication has been changed as a result of advancement of information and communication technologies, such as mobile phones or smartphones. These technologies allow us to talk to each other beyond geographic distances and also make such a speech communication style common today. This newly developed speech communication style has reminded us that there are some situations where we hesitate to talk; e.g., we have difficulty in talking about private information in a crowd; or speaking by yourself would sometimes annoy others in quiet environments. To address this issue, *silent speech interfaces* [1] have recently attracted attention as a technology to make it possible for us to talk without the necessity of emitting an audible signal. There are several ways to detect silent speech, such as body-conductive microphones [2, 3], electromyography [4], ultrasound imaging [5], and so on.

In this paper, we focus on the nonaudible murmur (NAM) microphone [3], which is one of the body-conductive microphones. The NAM microphone was designed to detect an extremely soft whispered voice called NAM, which is so quiet that people around the speaker barely hear its emitted sound. It can also detect various types of speech, such as a whispered voice, a soft voice, and normal speech. Although severe degradation of speech quality is caused by body-conductive recording [6], the recorded body-conducted speech is still comprehended if people get used to this special kind of sound. Moreover, some attempts have been made to apply statistical voice conversion (VC) techniques with the aim of further improving the quality and intelligibility of the recorded body-conducted speech [7, 8]. Thus, the NAM microphone has a great potential to be used as one of the silent speech interfaces.

To practically use the NAM microphone for silent speech communication, there still remain some issues. Robustness against external noise is one of the advantages of NAM microphone, but the detected unvoiced soft speech (such as NAM) is still significantly deteriorated as the surrounding noise level increases. Such a noisy NAM signal also causes severe failure in the conversion process of the VC-based enhancement system [9]. To address this issue, we have proposed a noise suppression method based on external noise monitoring using an air-conductive microphone [10]. This method utilizes the air-conductive microphone to detect only the external noise signal leveraging a property of NAM (i.e., its faint volume). It has been reported in [10] that the air-conducted signal is effectively used as a reference signal to suppress the external noise on the body-conducted signal in over 60 dBA noise condition (corresponding to about -20 dB SNR at the air-conductive microphone).

As the conventional noise suppression method based on the external noise monitoring, we proposed the semi-blind source separation (semi-BSS) techniques to estimate a time-invariant linear filter and reported that it worked well under a particular situation where only one fixed sound source exists in a sound-proof room. As a result of further investigation, we also found that its performance tended to degrade in more realistic experimental conditions, including environ-

This work was supported in part of JSPS KAKENHI Grant Numbers: 15K12064, 26280060, and 16J08977.

ments with many moving sound sources. In such situations, the consistency of the interchannel phase differences of the external noise can be easily destroyed, making it difficult to deal with noise suppression using a time-invariant system. To develop a noise suppression algorithm that is robust against the phase variations of noise, one possible approach is to treat the interchannel phase differences between the two channel inputs as latent variables to be marginalized out so that the algorithm becomes less sensitive to that factor. Based on this idea, we formulate a probabilistic model of the two channel observations within the external noise monitoring framework and develop a parameter inference algorithm. It turns out that the proposed noise suppression algorithm is a special case of non-negative tensor factorization of the observed air- and body-conducted signals. Experimental results demonstrated that 1) the proposed method consistently outperformed the conventional method under real noisy environments and 2) the proposed method effectively deals with speech acoustic changes caused by the Lombard reflex [11].

2. NOISE SUPPRESSION BASED ON EXTERNAL NOISE MONITORING

2.1. External noise monitoring [10]

NAM is an extremely soft whispered voice and it is practically difficult to be detected with an usual air-conductive microphone in noisy environments. By leveraging this property, only external noise signal can be detected with the air-conductive microphone placed away from mouth. Although the NAM signal is actually leaked into the air-conductive microphone from mouth, the signals detected with the air-conductive microphone can be well approximated with only the external noise signals if the sound pressure level of the external noise is higher than 60 dBA. It is also expected that this setting position of the air-conductive microphone close to the NAM microphone is helpful to detect the external noise signals well corresponding to noise signals detected with the NAM microphone. Consequently, the mixing process of the observed body- and air-conducted signals in noisy environments is assumed as follows:

$$y_1(t) = s_1(t) + \sum_{u=0}^{U} f_t(u) s_2(t-u) \quad (1)$$

$$y_2(t) \approx s_2(t) \quad (2)$$

where $s_1(t)$ is a clean body-conducted NAM signal, $s_2(t)$ is an air-conducted external noise signal, and $\{f_t(0), \cdots, f_t(U)\}$ is an acoustic transfer function to transfer the air-conducted external noise signal into the body-conducted external noise signal.

2.2. Noise suppression based on semi-BSS

Let us denote the frequency components of the source signals by $\boldsymbol{s}_{\omega,\tau} = [s_{1,\omega,\tau}, s_{2,\omega,\tau}]^\top$ and those of the observed signals by $\boldsymbol{y}_{\omega,\tau} = [y_{1,\omega,\tau}, y_{2,\omega,\tau}]^\top$, where ω and τ are frequency and time indices, respectively. By assuming that the acoustic transfer function is time-invariant, the mixing process given by Eqs. (1) and (2) is modeled as instantaneous mixture in the frequency domain as follows:

$$\boldsymbol{y}_{\omega,\tau} = \boldsymbol{F}_\omega \boldsymbol{s}_{\omega,\tau} \quad (3)$$

where $\boldsymbol{F}_\omega$ is a (2×2) time-invariant mixing matrix. In a standard BSS problem, the (2×2) demixing matrix $\boldsymbol{D}_\omega$ is the parameter to be estimated, for example using independent component analysis. By contrast, our noise monitoring problem assumes that one of the two source signals (i.e., $s_{2,\omega,\tau}$) is known, and some elements of the demixing matrix $\boldsymbol{D}_\omega$ can be fixed as follows:

$$\boldsymbol{D}_\omega = \begin{bmatrix} 1 & d_{1,2,\omega} \\ 0 & 1 \end{bmatrix}. \quad (4)$$

Therefore, only the component $d_{1,2,\omega}$ needs to be estimated by maximizing independence between the separated NAM signal and the observed air-conducted signal. This estimation can be done using the natural gradient algorithm [12].

3. PROPOSED NOISE SUPPRESSION METHOD

3.1. Probabilistic observation model

Although the effectiveness of the semi-BSS-based noise suppression method was confirmed on synthetic data in [10], the method had an essential problem when used in real environments. In real environments, there can be many moving sound sources including the user itself. In such situations, the impulse response between the air- and body-conducted external noise signals varies over time, which is not easy to handle with a time-invariant filter. Thus, we must consider a time-variant system. One possible approach for this problem is to treat a time-variant factor as a latent variable to be marginalized out. This idea has been adopted in some challenging task, e.g., a speech dereverberation robust against speaker's movement [13], a blind source separation robust against sampling rate mismatch of an ad-hoc microphone array [14] and a reverberation-robust blind source separation [15]. Inspired by these studies, the proposed method takes a similar approach to develop a robust noise suppression algorithm that is designed to be less sensitive to the phase variation caused by many sound sources and those movements.

When the acoustic transfer function is time-variant, the mixing process in Eqs. (3) is rewritten as

$$\boldsymbol{y}_{\omega,\tau} = \sum_i \boldsymbol{a}_{i,\omega,\tau} s_{i,\omega,\tau} \quad (5)$$

where $\boldsymbol{a}_{i,\omega,\tau}$ is the steering vector of the ith source signal depending on time τ. We assume that the time-frequency component of each source signal independently follows a complex Gaussian distribution, i.e., $s_{i,\omega,\tau} \sim \mathcal{N}_{\mathbb{C}}(0, p_{i,\omega,\tau})$ where

$p_{i,\omega,\tau}$ is the power spectrogram of the ith source signal.

Here, we assume that the magnitude part of the steering vector is less sensitive to a slight movement of the sound sources, and we decompose the steering vector into its magnitude and phase parts as follows:

$$\boldsymbol{a}_{i,\omega,\tau} = \underbrace{\begin{bmatrix} |a_{1,i,\omega}| & 0 \\ 0 & |a_{2,i,\omega}| \end{bmatrix}}_{\boldsymbol{A}_{i,\omega}} \underbrace{\begin{bmatrix} e^{j\phi_{1,i,\omega,\tau}} \\ e^{j\phi_{2,i,\omega,\tau}} \end{bmatrix}}_{\boldsymbol{\psi}_{i,\omega,\tau}}, \quad (6)$$

where the number of sources $I = 2$ and $\boldsymbol{a}_{1,\omega,\tau} = [1, 0]^\top$ in the external noise monitoring framework. Thus, the time-frequency components of the observed signals follow a complex Gaussian distribution as follows:

$$\boldsymbol{y}_{\omega,\tau} \sim \mathcal{N}_{\mathbb{C}}\Big(\boldsymbol{0}, \sum_i p_{i,\omega,\tau} \boldsymbol{A}_{i,\omega} \boldsymbol{\psi}_{i,\omega,\tau} \boldsymbol{\psi}_{i,\omega,\tau}^{\mathsf{H}} \boldsymbol{A}_{i,\omega}^{\mathsf{H}}\Big). \quad (7)$$

To treat the time-variant factor $\phi_{c,i,\omega,\tau}$ as a latent variable to be marginalized out, we make the following assumptions:

- $\phi_{c,i,\omega,\tau}$ and $\phi_{c',i,\omega,\tau}$ ($c \neq c'$ or $\tau \neq \tau'$) are statistically independent of each other.
- $\phi_{c,i,\omega,\tau}$ follows a uniform distribution in $[0, 2\pi)$.

Finally, the following probabilistic observation model is obtained: $y_{c,\omega,\tau} \sim \mathcal{N}_{\mathbb{C}}\Big(0, \sum_i p_{i,\omega,\tau} |a_{c,i,\omega}|^2\Big)$.

3.2. Non-negative tensor factorization (NTF) of air- and body-conducted signals

From Eqs. (7), the log-likelihood function of the modeled power spectrogram $x_{c,\omega,\tau} = \sum_i p_{i,\omega,\tau} |a_{c,i,\omega}|^2$ is given by

$$L(x_{c,\omega,\tau}) = -\log \pi x_{c,\omega,\tau} - \frac{|y_{c,\omega,\tau}|^2}{x_{c,\omega,\tau}}. \quad (8)$$

Eqs. (8) is maximized when $x_{c,\omega,\tau} = |y_{c,\omega,\tau}|^2$, and we can replace the maximization problem of this function with a minimization problem of the following difference function

$$L(|y_{c,\omega,\tau}|^2) - L(x_{c,\omega,\tau}) = \frac{|y_{c,\omega,\tau}|^2}{x_{c,\omega,\tau}} - \log \frac{|y_{c,\omega,\tau}|^2}{x_{c,\omega,\tau}} - 1, \quad (9)$$

i.e., the Itakura-Saito divergence $\mathcal{D}_{\mathrm{IS}}(|y_{c,\omega,\tau}|^2 \,|\, x_{c,\omega,\tau})$ [16].

Here, we assume that the power spectrogram $p_{i,\omega,\tau}$ is modeled as a product of two non-negative matrices in the same manner as the conventional non-negative matrix factorization (NMF) [17], $p_{i,\omega,\tau} = \sum_k w_{i,\omega,k} h_{i,k,\tau}$ where k shows an index of spectral templates. $w_{i,\omega,k}$ and $h_{i,k,\tau}$ represent the ith spectral template and the corresponding temporal activation function. From this assumption, the maximum likelihood estimation of the parameters ($\boldsymbol{A'} = (|a_{c,i,\omega}|^2)_{2\times 2\times \Omega}$, $\boldsymbol{W} = (w_{i,\omega,k})_{2\times\Omega\times K}$ and $\boldsymbol{H} = (h_{i,k,\tau})_{2\times K\times T}$), where $|a_{c,1,\omega}|^2$ is fixed, is formulated as a special case of non-negative tensor factorization of the observed power spectrogram tensor $\boldsymbol{Y'} = (|y_{c,\omega,\tau}|^2)_{2\times\Omega\times T}$.

Now, we can define the objective function as

$$D_{\mathrm{IS}}(\Theta) \stackrel{c}{=} \sum_{c,\omega,\tau} \left(\frac{|y_{c,\omega,\tau}|^2}{x_{c,\omega,\tau}} + \log x_{c,\omega,\tau} \right) \quad (10)$$

where Θ is the set of the parameters, and $\stackrel{c}{=}$ denotes equality up to constant terms. By using an auxiliary function method [18], the update rules for the parameters can be obtained as

$$|a_{c,i,\omega}|^2 \leftarrow |a_{c,i,\omega}|^2 \left(\frac{\sum_{k,\tau} |y_{c,\omega,\tau}|^2 w_{i,\omega,k} h_{i,k,\tau} / x_{c,\omega,\tau}^2}{\sum_{k,\tau} w_{i,\omega,k} h_{i,k,\tau} / x_{c,\omega,\tau}} \right)^{\frac{1}{2}}, \quad (11)$$

$$w_{i,\omega,k} \leftarrow w_{i,\omega,k} \left(\frac{\sum_{c,\tau} |y_{c,\omega,\tau}|^2 |a_{c,i,\omega}|^2 h_{i,k,\tau} / x_{c,\omega,\tau}^2}{\sum_{c,\tau} |a_{c,i,\omega}|^2 h_{i,k,\tau} / x_{c,\omega,\tau}} \right)^{\frac{1}{2}}, \quad (12)$$

$$h_{i,k,\tau} \leftarrow h_{i,k,\tau} \left(\frac{\sum_{c,\omega} |y_{c,\omega,\tau}|^2 |a_{c,i,\omega}|^2 w_{i,\omega,k} / x_{c,\omega,\tau}^2}{\sum_{c,\omega} |a_{c,i,\omega}|^2 w_{i,\omega,k} / x_{c,\omega,\tau}} \right)^{\frac{1}{2}}, \quad (13)$$

where $|a_{1,1,\omega}|^2$ and $|a_{2,1,\omega}|^2$ are fixed to 1 and 0, respectively, and the spectral template $w_{1,\omega,k}$ is pre-trained and fixed during the parameter updating process.

4. EXPERIMENTAL EVALUATION

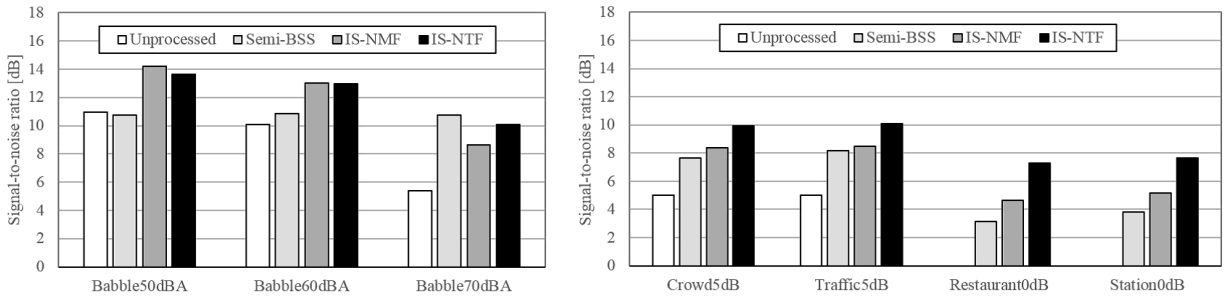
4.1. Experimental Conditions

We simultaneously recorded clean body- and air-conducted NAM signals uttered by one Japanese male speaker using the NAM microphone and the air-conductive microphone, respectively in a sound-proof room. We also recorded the following 7 kinds of noise.

- **Babble(50, 60, 70)dBA**: (50, 60, 70) dBA babble noise
- **Crowd5dB**: Real crowd noise (5 dB SNR)
- **Traffic5dB**: Real traffic noise (5 dB SNR)
- **Restaurant0dB**: Real restaurant noise (0 dB SNR)
- **Station0dB**: Real station noise (0 dB SNR)

The first three babble noise signals were recorded in the sound-proof room by using a loud speaker to present them, and they were directly superimposed on the clean air- and body-conducted NAM signals. The sound pressure levels of the individual noise signals were measured by a sound level meter placed at around the speaker's head. The human speech-like noise [19] generated by superimposing 20 different speaker's speech signals was used as the babble noise. The other four noise signals were recorded in real noisy environments. These signals include unspecified number of sound sources which were not controlled by us. The sound volumes of real noise signals were adjusted before the superimposition in order to minimize gain mismatches of the recording condition between real noisy environments and the sound-proof room.

To investigate the adverse effect of the Lombard reflex [11], which was naturally caused in also speaking NAM under noisy conditions [20], on the semi-supervised frameworks (e.g., the semi-supervised NMF [21] and the proposed NTF), we additionally recorded clean Lombard NAM signals uttered



(a) Only one fixed sound source exists. (b) Unspecified number of moving sound sources exist.

Fig. 1. SNR of the enhanced NAM signals.

by the same speaker by presenting the noise signals to him using headphones. Three sound levels of babble noise (**Babble50, 60, 70dBA**) were used in this recording. Fifty sentences in a phoneme balanced sentence set [22] were uttered in NAM. The sampling frequency was set to 16 kHz. The window length of STFT was set to 64 ms and the shift length was set to 32 ms.

The following 4 methods were evaluated in this paper.

- **Unprocessed**: no noise suppression
- **Semi-BSS**: semi-BSS (conventional method)
- **IS-NMF**: semi-supervised NMF with the IS div.
- **IS-NTF**: proposed semi-supervised NTF with the IS div.

For the **Semi-BSS**, the step-size parameter η and the number of iteration times were set to 0.01 and 200, respectively. The element of the demixing matrix $d_{2,2,\omega}$ was additionally updated to obtain better performance. For the **IS-NMF** and **IS-NTF**, the number of spectral templates for each sound source signal was set to 20. The spectral templates of target NAM signal were pre-trained by the conventional NMF using clean NAM signals, and they were fixed during the parameter updating process. The leave-one-out cross-validation was used in the evaluation. Noise suppression performance was measured by averaged SNR of the estimated body-conducted NAM signals.

4.2. Robustness against the change of environments

To verify the effectiveness of the proposed method, we investigated performance differences caused by changes of a surrounding situation. **Figure 1** shows the results. If only one fixed source exists, the conventional method (**Semi-BSS**) consistently yields about 10 dB SNR. Especially, **Semi-BSS** outperforms the others when the noise level is 70 dBA. However, the performance of **Semi-BSS** deteriorates in case of using real environment noise signals. This tendency can be confirmed by comparing the result of 70 dBA noise condition in **Figure 1 (a)** and all noise conditions in **Figure 1 (b)**. By contrast, for the proposed method (**IS-NTF**), there is no performance degradation caused by the change of the surrounding environments, and it is found that **IS-NTF** always outperforms **IS-NMF** thanks to the external noise monitoring.

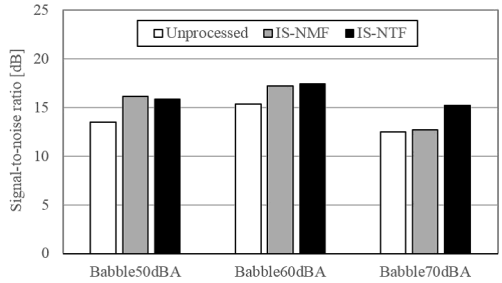


Fig. 2. SNR of the enhanced Lombard NAM signals.

4.3. Robustness against the change of speaking style

To investigate the adverse effect of the Lombard reflex on the proposed method, We used the simulated noisy Lombard NAM, which were generated by superimposing the noise signals on the clean Lombard NAM signals, as the input of each noise suppression method. There are two possible factors causing performance degradation under real noisy situations. One is an acoustic mismatch between the spectral feature of the input Lombard NAM and that of the spectral templates trained using normal NAM. The other is deterioration in the approximation accuracy of the external noise monitoring due to increase in sound intensity of NAM emitted from mouth.

Figure 2 shows the result. It is found that the conventional semi-supervised NMF (**IS-NMF**) is ineffective when the noise level raised to 70 dBA. On the other hand, the proposed method (**IS-NTF**) still work in such case. These results reveal that the proposed method is robust against the acoustic mismatches caused by the Lombard reflex and the external noise monitoring is still effective even for the Lombard NAM.

5. CONCLUSION

This paper presented a novel noise suppression method for the body-conducted soft speech. To deal with the real environment noise, we applied a probabilistic observation model into the external noise monitoring framework, and formulated its process as a special case of non-negative tensor factorization of the observed signals. Experimental results demonstrated that the proposed method consistently outperformed the conventional method under real noisy environments.

6. REFERENCES

[1] B. Denby, T. Schultz, K. Honda, J.M. Gilbert, and J.S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[2] S.C. Jou, T. Schultz, and A. Waibel, "Adaptation for soft whisper recognition using a throat microphone," in *Proc. INTERSPEECH*, 2004, pp. 1493–1496.

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

[4] T. Schultz and M. Wand, "Modeling coarticulation in EMG-based continuous speech recognition," *Speech Communication*, vol. 52, no. 4, pp. 341–353, 2010.

[5] T. Hueber, E.L. Benaroya, G. Ghollet, B. Denby, G.Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[6] 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.

[7] T. Toda, K. Nakamura, H. Sekimoto, and K. Shikano, "Voice conversion for various types of body transmitted speech," in *Proc. ICASSP*, 2009, pp. 3601–3604.

[8] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, 2012.

[9] Y. Tajiri, K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Non-audible murmur enhancement based on statistical conversion using air- and body-conductive microphones in noisy environments," in *Proc. INTERSPEECH*, 2015, pp. 2769–2773.

[10] Y. Tajiri, T. Toda, and S. Nakamura, "Noise suppression method for body-conducted soft speech enhancement based on external noise monitoring," in *Proc. ICASSP*, 2016, pp. 5935–5939.

[11] J.C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[12] S. Amari, "Natural gradient works efficiently in learning," *Neural computation*, vol. 10, no. 2, pp. 251–276, 1998.

[13] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. ICASSP*, 2009, pp. 45–48.

[14] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," in *Proc. IWAENC*, 2014, pp. 203–207.

[15] N. Murata, H. Kameoka, K. Kinoshita, S. Araki, T. Nakatani, S. Koyama, and H. Saruwatari, "Reverberation-robust underdetermined source separation with non-negative tensor double deconvolution," in *Proc. EUSIPCO*, 2016, pp. 1648–1652.

[16] C. Févotte, N. Bertin, and J. L. Durrieu, "Non-negative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[17] D. D. Lee and S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2001, pp. 556–562.

[18] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, vol. 30, Siam, 1970.

[19] S. Kajita, D. Koyabashi, K. Takeda, and F. Itakura, "Analysis of speech features included in human speech-like noise," *Journal of ASJ*, vol. 53, no. 5, pp. 337–345, 1997.

[20] P. Heracleous, T. Kaino, H. Saruwatari, and K. Shikano, "Investigating the role of the Lombard reflex in non-audible murmur (NAM) recognition," in *Proc. INTERSPEECH*, 2005, pp. 252–255.

[21] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *ICA*, 2007, pp. 414–421.

[22] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuwabara, "A large-scale japanese speech database," in *First International Conference on Spoken Language Processing*, 1990.