Journal of Signal Processing, Vol.27, No.6, pp.207-211, Novemtber 2023

RESEARCH NOTE

Amplitude Spectrogram Prediction from Mel-Frequency Cepstrum Coefficients Using Deep Neural Networks

Shoya Kawaguchi and Daichi Kitamura

National Institute of Technology, Kagawa College, Kagawa 761-8058, Japan
E-mail: kitamura-d@t.kagawa-nct.ac.jp

Abstract **Timbre conversion of musical instrument sounds, utilizing deep neural networks (DNNs), has been extensively researched and continues to generate significant interest in the development of more advanced techniques. We propose a novel algorithm for timbre conversion that utilizes a variational autoencoder. However, this system must be capable of predicting the amplitude spectrogram from the mel-frequency cepstrum coefficient (MFCC). This research aims to build a DNN-based decoder that utilizes the MFCC and time-frame-wise total amplitude as inputs to predict the amplitude spectrogram. Experiments conducted using a musical instrument sound dataset show that a decoder incorporating bidirectional long short-term memory yields accurate predictions of amplitude spectrograms.**

Keywords: deep learning, mel-frequency cepstrum coefficient, timbre conversion

1. Introduction

The generation of musical instrument sounds and the conversion of timbre have been the focus of extensive research. Various techniques based on deep neural networks (DNNs) have been proposed to tackle this problem. For example, differentiable digital signal processing [1] synthesizes musical instrument sounds by utilizing multiple sinusoidal waves and filtered noise, whose parameters are generated by pre-trained DNNs. In [2], the waveform of speech signals is predicted from mel-frequency cepstrum coefficients (MFCCs) [3] using generative adversarial networks [4]. WaveNet [5] is also utilized to achieve timbre conversion of musical instrument sounds in an end-to-end manner [6]. These methods directly generate time-domain waveforms and can be interpreted as a DNN model that includes the training of phase information in the time-frequency domain. However, timbre of sounds strongly depends on the amplitude (or power) information in the time-frequency domain. In this paper, we focus on predicting only the amplitude information using DNNs, omitting the generation of the time-domain waveform (estimation of phase and transformation from time-frequency to time domains) from the DNN training. This approach enables us to develop a simple timbre conversion system that does not require a complex DNN architecture and a large training dataset.

Another approach involves using a variational autoencoder (VAE) [7] to generate musical instrument sounds [8, 9, 10]. This method uses the VAE to extract disentangled latent features of pitch (note number) and timbre (musical instrument) labels.

Similar to the aforementioned approach, this paper also focuses on a VAE-based system for the timbre conversion of musical instruments. We are currently developing a new timbre conversion system, as depicted in Fig. 1. Our system aims to incorporate the three fundamental elements of musical instrument sounds: pitch, timbre, and volume. To represent these features, we use traditional, well-established features: note number as a parameter of pitch, an MFCC as a parameter of timbre, and time-frame-wise total amplitude as a parameter of volume. The timbre conversion of an input sound is achieved by modifying only the MFCC on the basis of the pre-trained VAE. We believe that our proposed system will enable the interpolation of timbre across multiple types of musical instruments, and thus contributing to the advancement of new art and music.

The proposed system (Fig. 1) requires a decoding process that involves calculating an amplitude spectrogram from the note number, manipulated MFCC, and time-frame-wise total amplitude. However, this type of decoder cannot be achieved in an analytical

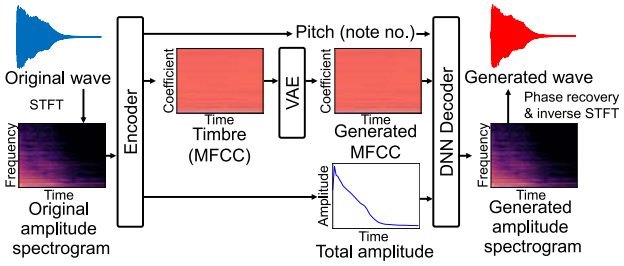

Fig. 1 Process flow of proposed timbre conversion system

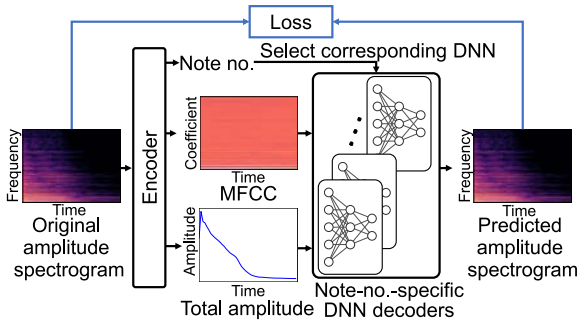

Fig. 2 Training process flow of proposed DNN-based timbre decoder

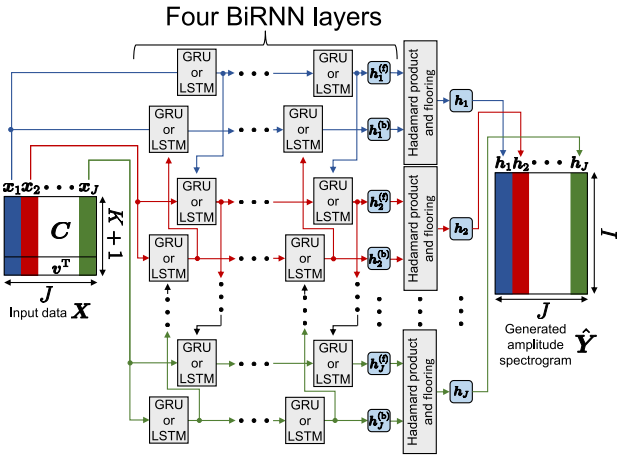

Fig. 3 Architecture of BiRNN used as DNN decoder

manner. To solve this problem, we propose the utilization of three DNN architectures and evaluate their suitability for predicting amplitude spectrograms.

2. Proposed System and Its DNN Decoder

2.1 Overview of proposed system

As depicted in Fig. 1, the proposed system first computes an amplitude spectrogram of the input sound via a short-time Fourier transform (STFT). The note number, MFCC, and time-frame-wise total amplitude are then extracted. While the note number can be extracted using various techniques as f_o, including DNN-based methods such as [11], the MFCC and time-frame-wise total amplitude are obtained through deterministic calculations. In the proposed system, the MFCC is assumed to be a reliable feature that represents the timbre of the sound and is disentangled from the pitch and volume. To train the timbre-embedded latent space, the MFCCs of various multiple musical instruments are input into the VAE. This training enables us to manipulate only the timbre of the sounds after the training of the entire proposed system. Finally, the note number, manipulated MFCC, and time-frame-wise total amplitude are decoded to obtain the amplitude spectrogram, and the waveform of the generated sound is obtained via a phase recovery technique, such as the Griffin–Lim algorithm [12], and the inverse STFT.

2.2 Motivations and contributions of this paper

In the proposed system, the output of the VAE (generated MFCC) must be decoded into an amplitude spectrogram using the note number and time-frame-wise total amplitude. Since the MFCC is a dimensionality-reduced feature, this decoding cannot be achieved through linear operations or analytical manners. To circumvent this issue, we use a DNN as the decoder in the proposed system, namely, the DNN decoder predicts the amplitude spectrogram of the synthesized sound from the inputted note number, MFCC, and time-frame-wise total amplitude. In this paper, we investigate a suitable architecture for this DNN decoder and evaluate its accuracy. The evaluation of the entire system shown in Fig. 1 is our future work.

The training procedure for the DNN decoder is depicted in Fig. 2. For the decoder, multiple DNNs are independently prepared and trained for each note number (from C3 to B5). Thus, f_o estimated by the encoder is used to select the note-number-specific DNN, and the MFCC and time-frame-wise total amplitude are inputted into the selected one. This note-number-wise model training eliminates the need for each DNN to posses generalization capabilities for all, resulting in a high prediction performance.

In this paper, we conduct experimental investigations to determine the suitable DNN architecture for the DNN decoder. The performance of the multilayer perception (MLP), bidirectional gated recurrent unit (BiGRU) [13], and bidirectional long short-term memory (BiLSTM) [14] are compared. BiGRU and BiLSTM are referred to as bidirectional recurrent neural networks (BiRNNs) [15], which can effectively capture latent structures of time-series data.

Table 1 Experimental conditions

Window and shift lengths in STFT	64/32 ms
Window function in STFT	Hann window
Sampling frequency	16 kHz
Maximum frequency of mel filter bank	8 kHz
Minimum frequency of mel filter bank	0 kHz
Number of mel-filters (K)	64
Flooring value ε	2.0×10^{-7}

2.3 Architecture of DNN decoders

Let $\boldsymbol{Y} \in \mathbb{R}_{\geq 0}^{I \times J}$ and y_{ij} be the amplitude spectrogram of an input sound and its element at frequency i and time j, respectively, as obtained by STFT. The time-frame-wise total amplitude of this signal is computed by

$$v_j = \sum_{i=1}^{I} (y_{ij} + \varepsilon) \tag{1}$$

where ε represents a small value to avoid zero-division. The amplitude spectrogram is normalized using the time-frame-wise total amplitude as

$$\overline{y}_{ij} = \frac{y_{ij}}{v_j} \tag{2}$$

Then, the MFCC is computed with the normalized power spectrogram $\boldsymbol{P} \in \mathbb{R}_{\geq 0}^{I \times J}$, whose element is $\overline{y}_{ij}^2$. This MFCC is denoted as $\boldsymbol{C} \in \mathbb{R}^{K \times J}$, where K is the order of the MFCC, which is equal to the number of filters in the bandpass filter known as the mel filter bank used in the MFCC computation. The input data of the DNN decoder is defined as

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{v}^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{(K+1) \times J} \tag{3}$$

where $\boldsymbol{v} = [v_1, v_2, \cdots, v_J]^{\mathrm{T}} \in \mathbb{R}^J$.

For the training of MLP, we vectorize $\boldsymbol{X}$ and input the vector into the first layer. The label (reference of the DNN prediction) is defined as $\boldsymbol{y} \in \mathbb{R}_{\geq 0}^{IJ}$, a vectorized version of the original amplitude spectrogram $\boldsymbol{Y}$. This MLP consists of three fully connected hidden layers with 1024–512–512 dimensions, and each hidden layer has a rectified linear unit as the activation function.

The network architecture of BiGRU or BiLSTM is illustrated in Fig. 3. The column vectors of $\boldsymbol{X}$, denoted as $\boldsymbol{x}_j \in \mathbb{R}^{K+1}$, are input into the first BiRNN layer. This bidirectional computation is applied four times, acquiring the forward output vectors $\boldsymbol{h}_1^{(\mathrm{f})}, \boldsymbol{h}_2^{(\mathrm{f})}, \cdots, \boldsymbol{h}_J^{(\mathrm{f})} \in \mathbb{R}_{\geq 0}^I$ and the backward output vectors $\boldsymbol{h}_J^{(\mathrm{b})}, \boldsymbol{h}_{J-1}^{(\mathrm{b})}, \cdots, \boldsymbol{h}_1^{(\mathrm{b})} \in \mathbb{R}_{\geq 0}^I$. The dimension of each vector is increased from $K+1$ to I in the first BiRNN layer and is maintained throughout the remaining layers. Finally, the predicted amplitude spectrogram is composed as

$$\hat{\boldsymbol{Y}} = [\boldsymbol{h}_1 \ \boldsymbol{h}_2 \ \cdots \ \boldsymbol{h}_J] \in \mathbb{R}_{\geq \varepsilon}^{I \times J} \tag{4}$$

$$\boldsymbol{h}_j = \max\left(\boldsymbol{h}_j^{(\mathrm{f})} \odot \boldsymbol{h}_j^{(\mathrm{b})}, \varepsilon\right) \ \forall j \tag{5}$$

where $\odot$ denotes the Hadamard product and $\max(\cdot, \cdot)$ returns the maximum value of inputs in each element.

All the DNNs are trained by minimizing the mean squared error $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2$, where $\hat{\boldsymbol{y}} \in \mathbb{R}_{\geq \varepsilon}^{IJ}$ is a vectorized version of the predicted amplitude spectrogram $\hat{\boldsymbol{Y}}$ and $\|\cdot\|_2$ is the L_2 norm.

3. Experimental Evaluation of DNN Decoders

3.1 Dataset and conditions

To evaluate the performance of the DNN decoder, we conducted an experiment on amplitude spectrogram prediction using the neural audio synthesis (Nsynth) dataset [6]. Nsynth is an audio dataset comprising four-second-long signals of various musical instrument sounds and consists of 305,979 signals. These signals were split into 289,205 training, 12,678 validation, and 4,096 test data. The other experimental conditions are shown in Table 1. As evaluation criteria, we used amplitude relative squared error (ARSE) and MFCC relative squared error (MRSE), defined as

$$\mathrm{ARSE} = 10 \log_{10} \frac{\sum_{j=1}^{J} \sum_{i=1}^{I} (y_{ij} - \hat{y}_{ij})^2}{\sum_{j=1}^{J} \sum_{i=1}^{I} y_{ij}^2} \ [\mathrm{dB}] \tag{6}$$

$$\mathrm{MRSE} = 10 \log_{10} \frac{\sum_{j=1}^{J} \sum_{k=2}^{14} (c_{kj} - \hat{c}_{kj})^2}{\sum_{j=1}^{J} \sum_{k=2}^{14} c_{kj}^2} \ [\mathrm{dB}] \tag{7}$$

respectively, where $\hat{y}_{ij}$ and c_{kj} are the elements of $\hat{\boldsymbol{Y}}$ and $\boldsymbol{C}$, respectively, and $\hat{c}_{kj}$ is the MFCC calculated from the predicted amplitude spectrogram $\hat{\boldsymbol{Y}}$. Small values of ARSE and MRSE indicate accurate prediction performance. Note that when calculating MRSE, we only used the MFCCs from $k = 2$ to 14, as these dimensions specifically encompass the timbre characteristics.

3.2 Results

Figs. 4, 5, and 6 show examples of original and predicted amplitude spectrograms. These results confirm that the MLP consistently fails to predict the amplitude spectrogram: the predicted results contain numerous spectral holes, resulting in artificial distortions. In contrast, BiGRU and BiLSTM accurately predict the harmonic structures and temporal transitions within the original amplitude spectrogram. This

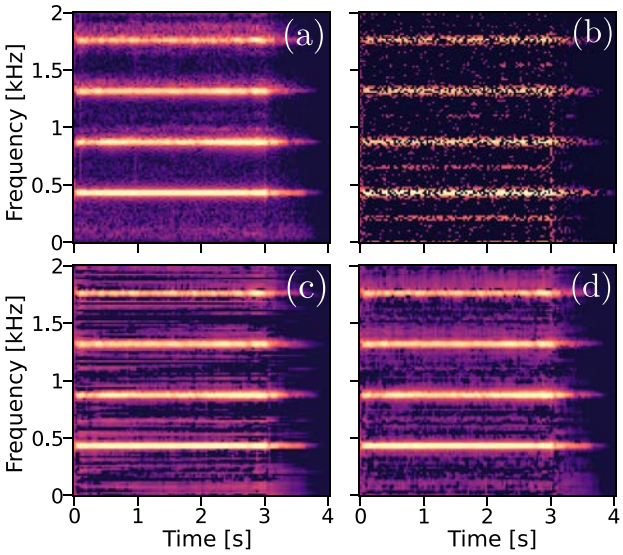

Fig. 4 Example of spectrograms for test data (acoustic flute sound): (a) Original, (b) Predicted by MLP, (c) Predicted by BiGRU, and (d) Predicted by BiLSTM

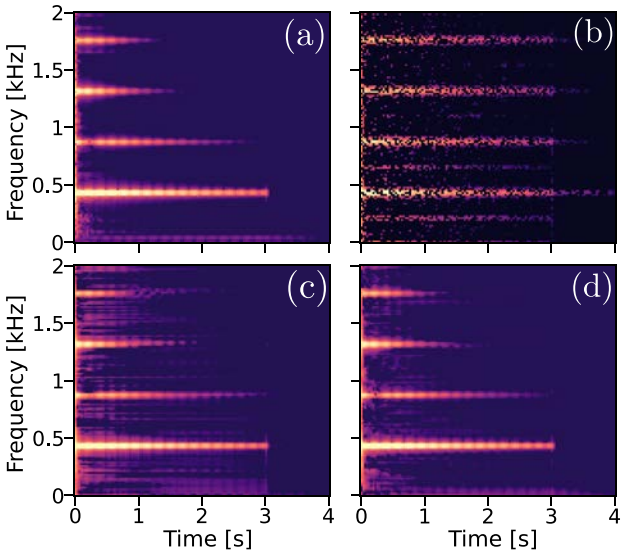

Fig. 5 Example of spectrograms for test data (synth. keyboard sound): (a) Original, (b) Predicted by MLP, (c) Predicted by BiGRU, and (d) Predicted by BiLSTM

is due to the recurrent architecture in BiGRU and BiLSTM, which effectively captures the time-series structures of the input data. In addition, Fig. 7 shows ARSE and MRSE results averaged over all test data for each musical instrument. In both scores, BiLSTM consistently provides the best result for all the instruments. The results for mallets are inferior as compared with other musical instruments. This is likely because mallets are typically classified as percussion instruments, and their sounds are characterized by complex harmonic structures and time transitions, as shown in Fig. 6. From these results, we found that BiLSTM provides satisfactory performance in predicting the amplitude spectrogram, making it the preferred choice for the DNN decoder in the proposed system.

4. Conclusion

We examined the efficacy of predicting amplitude spectrograms from the MFCC and time-frame-wise total amplitude using DNNs. Experiments using the Nsynth dataset showed that BiLSTM consistently outperforms MLP and BiGRU models. From the results, the proposed method achieved accurate spectrogram prediction for several musical instruments, e.g., string, flute, guitar, and keyboard. We expect that a VAE-based timbre conversion system can provide better performance, particularly for these musical instruments. In future work, we intend to construct the proposed sound generation system (Fig. 1) utilizing the BiLSTM-based DNN decoder.

Acknowledgement

This work was partly supported by Ono Charitable Trust for Acoustics and JSPS KAKENHI Grant Number 22H03652.

References

[1] J. Engel, L. Hantrakul, C. Gu and A. Roberts: Differentiable digital signal processing, Proc. Int. Conf. Learning Representations, 2020.

[2] L. Juvela, B. Bollepalli, X. Wang, H. Kameoka, M. Airaksinen, J. Yamagishi and P. Alku: Speech waveform synthesis from MFCC sequences with generative adversarial networks, Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, pp. 5679–5683, 2018.

[3] F. Zheng, G. Zhang and Z. Song: Comparison of different implementations of MFCC, J. Computer Science and Technology, Vol. 16, pp. 582–589, 2001.

[4] I. Goodfellow, J. P.-Abadie, M. Mirza, B. Xu, D. W.-Farley, S. Ozair, A. Courville and Y. Bengio: Generative adversarial networks, Proc. Advances in Neural Information Processing Systems, Vol. 27, pp. 2672–2680, 2014.

[5] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu: WaveNet: A generative model for raw audio, arXiv:1609.03499v2, 2016.

[6] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan and M. Norouzi: Neural audio synthesis of musical notes with WaveNet autoencoders, Proc. Int. Conf. Machine Learning, pp. 1068–1077, 2017.

[7] D. P. Kingma and M. Welling: Auto-encoder variational Bayes, Proc. Int. Conf. Learning Representations, 2014.

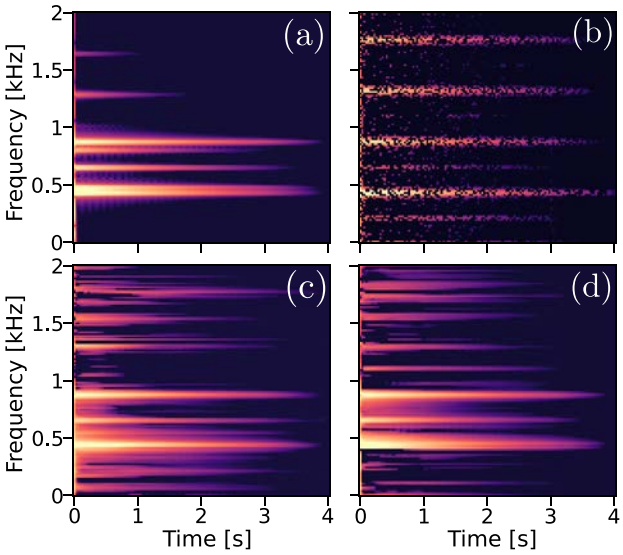

Fig. 6 Example of spectrograms for test data (acoustic mallet sound): (a) Original, (b) Predicted by MLP, (c) Predicted by BiGRU, and (d) Predicted by BiLSTM

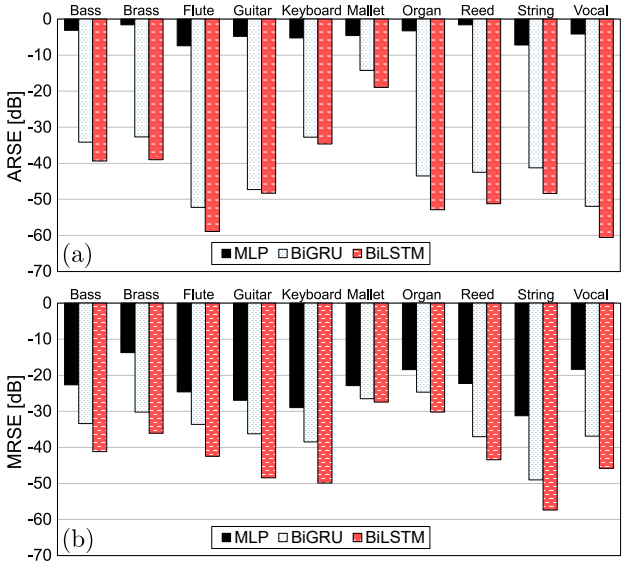

Fig. 7 (a) ARSE and (b) MRSE results averaged over test data for each musical instrument

[8] Y. J. Luo, K. Agres and D. Herremans: Learning disentangled representations of timbre and pitch for musical instrument sounds using Gaussian mixture variational autoencoders, Proc. Int. Society for Music Information Retrieval, pp. 746–753, 2019.

[9] Y. J. Luo, K. W. Cheuk, T. Nakano, M. Goto and D. Herremans: Unsupervised disentanglement of pitch and timbre for isolated musical instrument sounds, Proc. Int. Society for Music Information Retrieval, pp. 700–707, 2020.

[10] K. Tanaka, R. Nishikimi, Y. Bando, K. Yoshii and S. Morishima: Pitch-timbre disentanglement of musical instrument sounds based on VAE-based metric learning, Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, pp. 111–115, 2021.

[11] J. W. Kim, J. Salamon, P. Li and J. P. Bello: CREPE: A convolutional representation for pitch estimation, Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, pp. 161–165, 2018.

[12] D. Griffin and J. Lim: Signal estimation from modified short-time Fourier transform, IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. 32, No. 2, pp. 236–243, 1984.

[13] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio: Learning phrase representations using RNN encoder-decoder for statistical machine translation, arXiv:1406.1078, 2014.

[14] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink and J. Schmidhuber: LSTM: A search space odyssey, IEEE Trans. Neural Networks and Learning Systems, Vol. 28, No. 10, pp. 2222–2232, 2016.

[15] M. Schuster and K. K. Paliwal: Bidirectional recurrent neural networks, IEEE Trans. Signal Processing, Vol. 45, No. 11, pp. 2673–2681, 1997.

Shoya Kawaguchi is a bachelor student of Advanced Course in Electronics, Information and Communication Engineering, National Institute of Technology, Kagawa College. His research interests include timbre conversion and instrumental sound generation using deep learning.

Daichi Kitamura received the Ph.D. degree from SOKENDAI, Hayama, Japan. He joined the University of Tokyo in 2017 as a Research Associate, and he moved to National Institute of Technology, Kagawa Collage in 2018. His research interests include audio source separation, statistical signal processing, and machine learning. He was the recipient of the Awaya Prize Young Researcher Award from the Acoustical Society of Japan (ASJ) in 2015, Ikushi Prize from Japan Society for the Promotion of Science in 2017, Best Paper Award from IEEE Signal Processing Society Japan in 2017, Itakura Prize Innovative Young Researcher Award from ASJ in 2018, and IEEE Signal Processing Society 2019 Young Author Best Paper Award.

(Received May 2, 2023; revised June 13, 2023)