ENVIRONMENTAL SOUND EXTRACTION USING ONOMATOPOEIC WORDS

Yuki Okamoto[1]* *Shota Horiguchi*[2] *Masaaki Yamamoto*[2] *Keisuke Imoto*[3] *Yohei Kawaguchi*[2]

[1] Ritsumeikan University, Japan
[2] Hitachi, Ltd., Japan
[3]Doshisha University, Japan

ABSTRACT

An onomatopoeic word, which is a character sequence that phonetically imitates a sound, is effective in expressing characteristics of sound such as duration, pitch, and timbre. We propose an environmental-sound-extraction method using onomatopoeic words to specify the target sound to be extracted. By this method, we estimate a time-frequency mask from an input mixture spectrogram and an onomatopoeic word using a U-Net architecture, then extract the corresponding target sound by masking the spectrogram. Experimental results indicate that the proposed method can extract only the target sound corresponding to the onomatopoeic word and performs better than conventional methods that use sound-event classes to specify the target sound.

***Index Terms*—** Sound extraction, deep learning, environmental sound, onomatopoeic word, onomatopoeia, sound event detection

1. INTRODUCTION

Environmental sounds are essential for expressive media content, e.g., movies, video games, and animation, to make them immersive and realistic. One way to prepare a desired sound is to obtain it from an environmental sound database. However, the number of databases currently available is very limited [1], so the desired sound is not always in the database. On the other hand, there is a large amount of unlabeled environmental sounds on the Internet, but it is not easy to expand the database because it requires rich domain knowledge and taxonomy.

Even if the database became large, its usability might decrease because it would also require users to have domain knowledge. Intuitive methods for sound retrieval have been proposed. For example, vocal imitation [2, 3, 4, 5] and onomatopoeic words [6] were used as search queries in some sound retrieval systems. It has also been reported that user satisfaction is high when using an intuitive sound-retrieval technique [2, 3]. Therefore, it would also be useful for content creators if they can extract a desired sound intuitively.

We propose an environmental-sound-extraction method using an onomatopoeic word, which is a character sequence that phonetically imitates a sound. It has been shown that onomatopoeic words are effective in expressing the characteristics of sound [7, 8] such as sound duration, pitch, and timbre. Onomatopoeic words are also advantageous in terms of low labeling cost since they do not require domain knowledge and taxonomy for labeling. In our proposed method, therefore, uses an onomatopoeic word is used to specify the sound to extract from a mixture sound, as shown in Fig. 1. We used a U-Net architecture [9], which has been used in various source-separation and sound-extraction studies [10, 11, 12, 13], to estimate the time-frequency mask of the target sound. To the best of our knowledge, there has been no study on extracting only specific sound by using an onomatopoeic word.

*This work has been done during internship at Hitachi, Ltd.

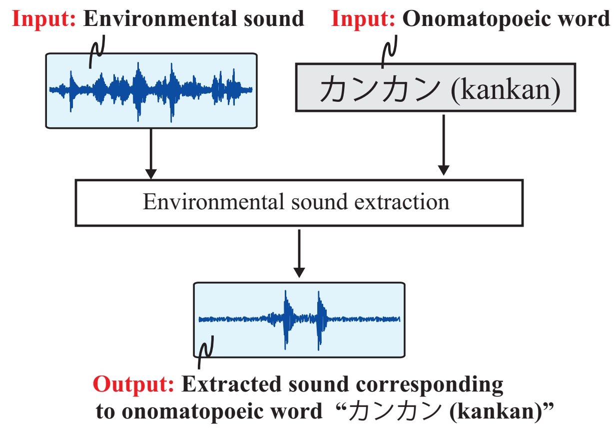

Fig. 1. Overview of environmental sound extraction using an onomatopoeic word. The word “kankan” is often used in Japanese to represent hitting sounds.

The rest of the paper is organized as follows. In Sec. 2, we describe related work on environmental sound extraction. In Sec. 3, we present our proposed method for extracting environmental sounds using an onomatopoeic word from an input mixture. In Sec. 4, we discuss experiments we conducted on the effectiveness of our proposed method compared with baseline methods that use class labels to specify the target sound. Finally, we summarize and conclude this paper in Sec. 5.

2. RELATED WORK

Methods of environmental sound extraction and separation using deep learning have been developed [11, 14, 15, 16]. Sudo et al. developed an environmental-sound-separation method based on U-Net architecture [11]. A similar method using U-Net was also proposed for source separation [10, 17]. Ochiai et al. used Conv-TasNet [18], which was originally proposed for speech separation, to extract only the sounds of specific sound events [14]. These methods use the sound-event class as the input to specify the desired sound. However, environmental sounds have various characteristics that cannot be described as a sound class, such as sound duration, pitch, and timbre. For example, if the “whistle sound” class is defined regardless of the pitch, it is not possible for a conventional method to extract only the sound of the desired pitch. One possible solution is to define more fine-grained sound-event classes, e.g., “high-pitched whistle sound”

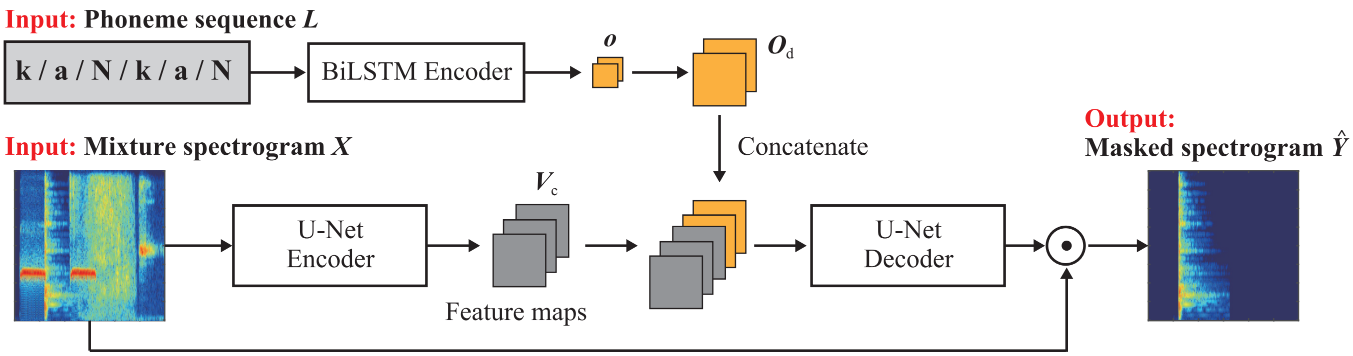

Fig. 2. Detailed architecture of proposed environmental-sound-extraction method using an onomatopoeic word.

and "low-pitched whistle sound." However, this is impractical because the labeling cost will increase. Even if we could define such fine-grained sound-event classes, there would always be intra-class variation, and we have no way to distinguish between them. Therefore, the method of conditioning by sound-event class is not suitable to extract specific sounds.

Another possible method similar to that discussed in this paper is singing voice extraction using humming [19]. In this case, the target sound is always a human voice, so humming is sufficient to represent it. However, in the case of environmental sound extraction, humming is insufficient to determine the target sound because it cannot express timbre, and some kinds of sounds cannot be represented by humming, e.g., plosive sounds.

3. PROPOSED METHOD

3.1. Overview of environmental sound extraction using an onomatopoeic word

Our purpose was to reconstruct a target sound $\boldsymbol{y}$ from a mixture sound $\boldsymbol{x}$, where the target is specified by an onomatopoeic word w. We estimate $\hat{\boldsymbol{y}}$ from $\boldsymbol{x}$ and w using a nonlinear transformation $\mathsf{Extractor}(\cdot,\cdot)$ as follows:

$$\hat{\boldsymbol{y}} = \mathsf{Extractor}(\boldsymbol{x}, w). \tag{1}$$

We explain this $\mathsf{Extractor}(\cdot,\cdot)$ in Sec. 3.2.

3.2. Proposed sound extraction method

Figure 2 shows the detailed architecture of the proposed method. The method involves time-frequency mask estimation using U-Net and feature vector extraction from an onomatopoeic word. We condition the output of the U-Net encoder with an onomatopoeic word to specify the target environmental sound to extract. In previous studies, the target sound to be extracted was conditioned by sound-event class [10], or further conditioned by the estimated interval of the target sound [11]. These studies have shown that conditioning on intermediate features after passing through convolutional neural network layers can be effective. Thus, we also use conditioning on the intermediate features of the U-Net encoder.

The proposed method takes the following as inputs, as shown in Fig. 2. One is a T-length F-dimensional mixture spectrogram $\boldsymbol{X} \in \mathbb{R}^{F\times T}$ extracted from the input mixture sound $\boldsymbol{x}$. The other is a one-hot encoded phoneme sequence $\boldsymbol{L} = (\boldsymbol{l}_1,\dots,\boldsymbol{l}_N)$ extracted from w. The extracted acoustic feature $\boldsymbol{X}$ is fed to the U-Net encoder, which consists of K-stacked convolutional layers. In each layer of the U-Net encoder, the time-frequency dimension decreases by half and the number of channels doubles. As a result, $C\left(=2^K\right)$ feature maps are calculated as

$$[\boldsymbol{V}_1,\dots,\boldsymbol{V}_C] = \mathsf{UNetEncoder}\left(\boldsymbol{X}\right) \in \mathbb{R}^{F'\times T'\times C}, \tag{2}$$

where $\boldsymbol{V}_c \in \mathbb{R}^{F'\times T'} (c = 1,\dots,C)$ denotes the feature map of the c-th channel.

At the same time, the phoneme sequence $\boldsymbol{L}$ is fed to the bidirectional long short-term memory (BiLSTM) encoder. As a result, a D-dimensional word-level embedding $\boldsymbol{o} = [o_1,\dots,o_D]^\mathsf{T} \in \mathbb{R}^D$ that captures the entire onomatopoeic word is extracted as follows:

$$\boldsymbol{o} = \mathsf{BiLSTMEncoder}\left(\boldsymbol{L}\right) \in \mathbb{R}^D. \tag{3}$$

The extracted embedding $\boldsymbol{o}$ is stretched in the time and frequency directions to form D feature maps $[\boldsymbol{O}_1,\dots,\boldsymbol{O}_D]$, where $\boldsymbol{O}_d \in \mathbb{R}^{F'\times T'}$ for $d \in \{1,\dots,D\}$ is the matrix the whose elements are all o_d.

Finally, a time-frequency soft mask $M \in (0,1)^{F\times T}$ is estimated using the U-Net decoder, which consists of K-stacked deconvolutional layers. The feature maps from the U-Net encoder and BiLSTM encoder are concatenated to be $C + D$ channels and fed to the U-Net decoder followed by the element-wise sigmoid function $\sigma\left(\cdot\right)$ as

$$\boldsymbol{Z} = \mathsf{UNetDecoder}\left([\boldsymbol{V}_1,\dots,\boldsymbol{V}_C,\boldsymbol{O}_1,\dots,\boldsymbol{O}_D]\right) \in \mathbb{R}^{F\times T}, \tag{4}$$

$$\boldsymbol{M} = \sigma\left(\boldsymbol{Z}\right) \in (0,1)^{F\times T}. \tag{5}$$

The target signal in time-frequency domain $\hat{\boldsymbol{Y}}$ is then recovered by masking the input $\boldsymbol{Y}$ as

$$\hat{\boldsymbol{Y}} = \boldsymbol{M} \odot \boldsymbol{X} \in \mathbb{R}^{F\times T}, \tag{6}$$

where $\odot$ is the Hadamard product.

During training, the loss function defined as root mean square error between $\hat{\boldsymbol{Y}}$ and target features $\boldsymbol{Y} \in \mathbb{R}^{F\times T}$, which is extracted from $\boldsymbol{x}$, is used:

$$L\left(\boldsymbol{Y},\hat{\boldsymbol{Y}}\right) = \sqrt{\frac{1}{TF}\left\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right\|_F^2}, \tag{7}$$

where $\|\cdot\|_F$ is the Frobenius norm.

In the inference phase, we reconstruct an environmental sound wave from the masked acoustic features $\hat{\boldsymbol{Y}}$ using the Griffin–Lim algorithm [20].

4. EXPERIMENTS

4.1. Dataset construction

To construct the datasets for this task, we used environmental sounds extracted from RealWorld Computing Partnership-Sound Scene Database (RWCP-SSD) [21]. Some sound events in RWCP-SSD are labeled in the “event entry + ID” format, e.g., *whistle1* and *whistle2*. We created hierarchical sound-event classes by grouping labels with the same event entry, e.g., *whistle*. We first selected 44 sound events from RWCP-SSD, which we call subclasses, and grouped them into 16 superclasses. The superclasses and subclasses used in this study are listed in Table 2. We selected 16 types of sound events in superclass and 44 types of sound events in subclass from RWCP-SSD to construct the dataset. The sounds in each subclass were divided as 7:2:1, used for training, validation, and evaluation, respectively. The onomatopoeic words corresponding to each environmental sound were extracted from RWCP-SSD-Onomatopoeia [22]. Each sound was annotated with more than 15 onomatopoeic words in RWCP-SSD-Onomatopoeia, and we used randomly selected three onomatopoeic words for each sound for our experiments.

We constructed the following three evaluation datasets using the selected sound events:

- **Inter-superclass dataset**: Each mixture sound in this dataset is composed of a target sound and interference sounds, the superclass of each is different from that of the target sound.
- **Intra-superclass dataset**: Each mixture sound in this dataset is composed of a target sound and interference sounds, the superclass of each is the same as that of the target sound, but the subclass is different.
- **Intra-subclass dataset**: Each mixture sound in this dataset is composed of a target sound and interference sounds, the subclass of each is the same as that of the target sound, but the onomatopoeic words are different.

The mixture sounds in each dataset were created by varying the signal-to-noise ratio (SNR) by $\{-10, -5, 0, 5, 10\}$ dB. The SNR between a target signal $\boldsymbol{s}_{\text{target}}$ and an interference signal $\boldsymbol{s}_{\text{interference}}$ is defined as

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\|\boldsymbol{s}_{\text{target}}\|^2}{\|\boldsymbol{s}_{\text{interference}}\|^2}\right). \tag{8}$$

The training and validation sets consisted of 7,563 and 2,160 mixture sounds, respectively. Each evaluation set consisted of 1,107 mixture sounds for each SNR. The audio clips for these sets were randomly selected from RWCP-SSD.

4.2. Training and evaluation setup

Table 1 shows the experimental conditions and parameters used for the proposed method (onomatopoeia-conditioned method). As baselines, we also evaluated the methods with which the target sound is conditioned by the superclass or subclass sound-event class. We used the one-hot representation of the label for $\boldsymbol{o}$ in (3) instead of the word embeddings.

To evaluate each method, we used signal-to-distortion ratio improvement (SDRi) [24] as an evaluation metric. SDRi is defined as the difference between the SDR of the target sound to the mixture and that of the target sound to the extracted sound as follows:

$$\mathrm{SDRi} = 10\log_{10}\left(\frac{\|\boldsymbol{y}\|^2}{\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|^2}\right) - 10\log_{10}\left(\frac{\|\boldsymbol{y}\|^2}{\|\boldsymbol{y}-\boldsymbol{x}\|^2}\right). \tag{9}$$

Table 1. Experimental conditions

Mixture-sound length	5 s
Sampling rate	16 kHz
Waveform encoding	16-bit linear PCM
# of U-Net encoder blocks	4
# of U-Net decoder blocks	4
# of BiLSTM encoders	1
# of units in BiLSTM encoder	512
Batch size	8
Optimizer	RAdam [23]
Acoustic feature	Amplitude spectrogram
Window length for FFT	128 ms (2,048 samples)
Window shift for FFT	32 ms (512 samples)

Table 2. Superclass and subclass sound events used in this study

Superclass	Subclass	Superclass	Subclass
metal	metal05, metal10, metal15	**bells**	bells1, bells2, bells3, bells4, bells5
dice	dice1, dice2, dice3	**coin**	coin1, coin2, coin3
bottle	bottle1, bottle2	**coins**	coins1, coins2, coins3, coins4, coins5
cup	cup1, cup2		
particl	particl1, particl2	**whistle**	whistle1, whistle2, whistle3
cap	cap1, cap2		
clap	clap1, clap2	**phone**	phone1, phone2, phone3, phone4
claps	claps1, claps2		
clip	clip1, clip2	**toy**	toy1, toy2
bell	bell1, bell2		

We conducted evaluations regarding SDRi on each of the three evaluation datasets introduced in Sec. 4.1.

4.3. Experimental results

Table 3 shows the SDRi on each evaluation dataset. We observed that the superclass-conditioned method performed well on the inter-superclass dataset but performed poorly on the intra-superclass and intra-subclass datasets. We also observed that the subclass-conditioned method performed well on the inter-superclass and intra-superclass datasets but did not on the intra-subclass dataset. These results indicate that the performance of sound extraction using an event class as a condition is highly dependent on the fineness of the class definition. The onomatopoeia-conditioned method showed almost the same SDRi on the three datasets. This suggests that an onomatopoeic word can behave like a more fine-grained class than the subclasses, even though it does not require any special domain knowledge for labeling.

Figure 3 shows the spectrograms of the extracted sounds using the subclass-conditioned and onomatopoeia-conditioned methods. For this visualization, we used five samples in the intra-subclass dataset with 0 dB. We observed that the subclass-conditioned method left a significant amount of non-target sounds, while the onomatopoeia-conditioned method extracted only the target sound. Although the onomatopoeia-conditioned method performed better than the superclass- and subclass-conditioned methods, it still did not perform well when the target sound was highly overlapped with interference sounds (cf. “Subclass: Phone4” in Fig. 3). As a result, the mixtures with high overlap ratios resulted in small SDRi and the mixtures with low overlap ratios resulted in large SDRi, and thus the

Table 3. SDRi [dB] for extracted signals

Dataset	Method	SNR				
		-10 dB	-5 dB	0 dB	5 dB	10 dB
Inter-superclass dataset	Superclass-conditioned method	5.11 ± 3.02	4.72 ± 2.75	4.06 ± 2.55	2.70 ± 2.13	1.33 ± 2.12
	Subclass-conditioned method	5.06 ± 2.97	4.75 ± 2.85	4.04 ± 2.52	2.81 ± 2.31	1.25 ± 2.09
	Onomatopoeia-conditioned method	4.63 ± 2.58	4.57 ± 2.69	4.02 ± 2.53	2.77 ± 2.22	1.41 ± 2.12
Intra-superclass dataset	Superclass-conditioned method	2.05 ± 2.37	1.97 ± 2.40	1.86 ± 2.38	1.50 ± 2.19	0.82 ± 1.89
	Subclass-conditioned method	5.03 ± 2.56	4.77 ± 2.59	4.19 ± 2.45	2.74 ± 2.12	1.26 ± 2.06
	Onomatopoeia-conditioned method	5.61 ± 2.78	5.36 ± 2.75	4.73 ± 2.52	3.10 ± 2.27	1.42 ± 2.06
Intra-subclass dataset	Superclass-conditioned method	2.03 ± 2.40	2.06 ± 2.54	1.87 ± 2.37	1.49 ± 2.09	0.79 ± 1.98
	Subclass-conditioned method	3.14 ± 2.78	3.09 ± 2.77	2.84 ± 2.63	2.21 ± 2.29	1.01 ± 2.12
	Onomatopoeia-conditioned method	5.83 ± 2.43	5.68 ± 2.53	5.11 ± 2.58	3.34 ± 2.24	1.64 ± 2.02

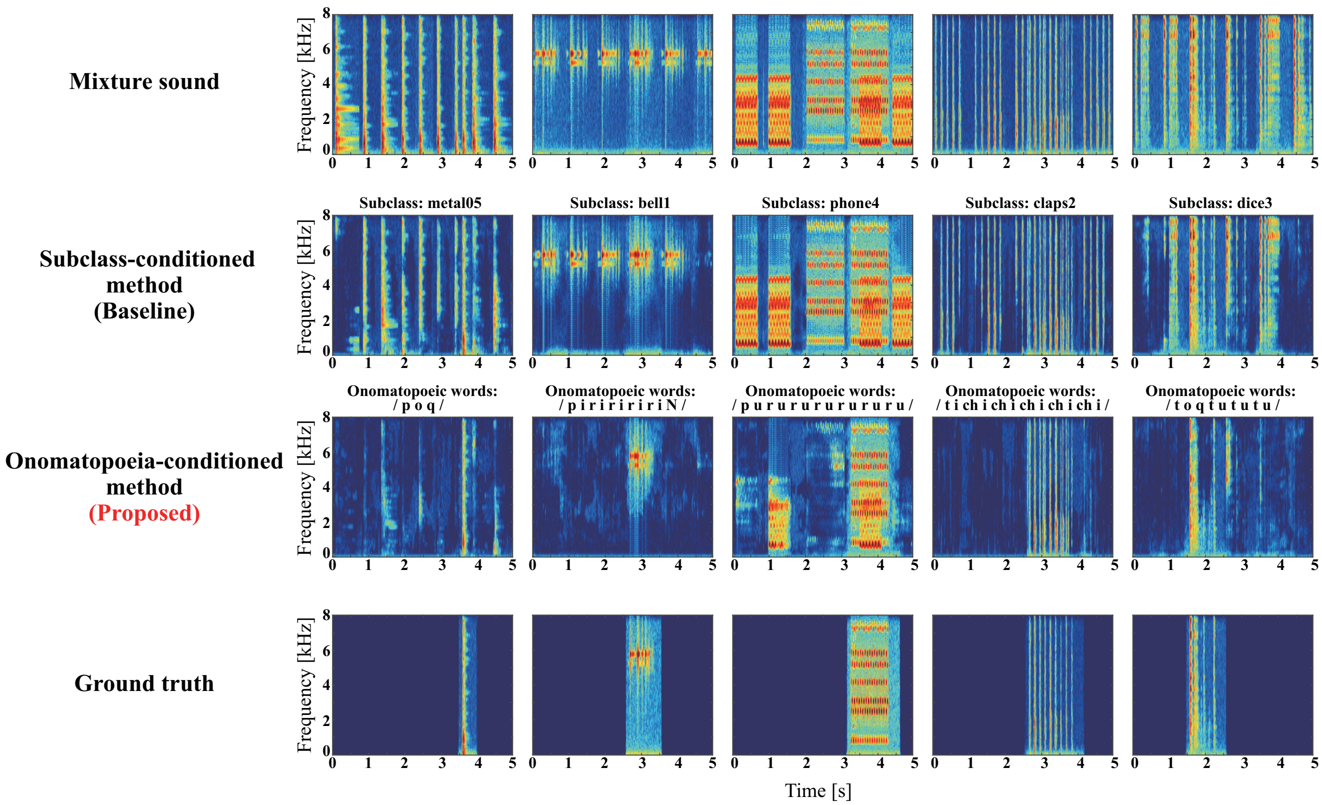

Fig. 3. Examples of environmental sound extraction using intra-subclass dataset. Mixture spectrogram (first row), results of subclass-conditioned sound extraction (second row), results of onomatopoeia-conditioned sound extraction (proposed) (third row), and ground truth spectrogram (fourth row).

standard deviations in Table 3 are large overall. The extraction of overlapping sounds requires further study. The extracted sounds are available on our web page[1].

5. CONCLUSION

We proposed an environmental-sound-extraction method using onomatopoeic words. The proposed method estimates a time-frequency mask of the target sound specified by an onomatopoeic word with the U-Net encoder-decoder architecture. The experimental results indicate that our proposed method extracts specific sounds from mixture sounds by using an onomatopoeic word as a condition. Our proposed method outperformed conventional methods that use a sound-event class as a condition. The results indicate that onomatopoeic words can behave like more fine-grained classes than sound-event classes, even though it does not require any special domain knowledge for labeling. In the future, it will be necessary to verify the effectiveness of the proposed method for onomatopoeic words assigned by speakers of different languages.

[1] https://y-okamoto1221.github.io/Sound_Extraction_Onomatopoeia/

6. REFERENCES

[1] K. Imoto, "Introduction to acoustic event and scene analysis," *Acoust. Sci. Tech.*, vol. 39, no. 3, pp. 182–188, 2018.

[2] Y. Zhang, J. Hu, Y. Zhang, B. Pardo, and Z. Duan, "Vroom!: A search engine for sounds by vocal imitation queries," in *Proc. Conference on Human Information Interaction and Retrieval* (*CHIIR*), 2020, pp. 23–32.

[3] Y. Zhang and Z. Duan, "IMISOUND: An unsupervised system for sound query by vocal imitation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2016, pp. 2269–2273.

[4] B. Kim and B. Pardo, "Improving content-based audio retrieval by vocal imitation feedback," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2019, pp. 4100–4104.

[5] Y. Zhang, B. Pardo, and Z. Duan, "Siamese style convolutional neural networks for sound search by vocal imitation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 429–441, 2019.

[6] S. Ikawa and K. Kashino, "Acoustic event search with an onomatopoeic query: measuring distance between onomatopoeic words and sounds," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), 2018, pp. 59–63.

[7] G. Lemaitre and D. Rocchesso, "On the effectiveness of vocal imitations and verbal descriptions of sounds," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 862–873, Feb. 2014.

[8] S. Sundaram and S. Narayanan, "Vector-based representation and clustering of audio using onomatopoeia words," in *Proc. American Association for Artificial Intelligence* (*AAAI*) *Symposium Series*, 2006, pp. 55–58.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention* (*MICCAI*), 2015, pp. 234–241.

[10] G. M.-Brocal and G. Peeters, "Conditioned-U-Net: Introducing a control control mechanism in the U-Net for multiple source separations," in *Proc. International Society for Music Information Retrieval* (*ISMIR*), 2019, pp. 159–165.

[11] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Environmental sound segmentation utilizing Mask U-Net," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), 2019, pp. 5340–5345.

[12] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. International Society for Music Information Retrieval* (*ISMIR*), 2017, pp. 745–751.

[13] Q. Kong, Y. Wang, X. Song, Y. Cao, W. Wang, and M. D. Plumbley, "Source separation with weakly labelled data: an approach to computational auditory scene analysis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, pp. 101–105.

[14] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to what you want: Neural network-based universal sound selector," in *Proc. INTERSPEECH*, 2020, pp. 1441–1445.

[15] J. H. Lee, H.-S. Choi, and K. Lee, "Audio query-based music source separation," in *Proc. International Society for Music Information Retrieval* (*ISMIR*), 2019, pp. 878–885.

[16] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. L. Roux, and J. R. Hershey, "Universal sound separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), 2019, pp. 175–179.

[17] O. Slizovskaia, L. Kim, G. Haro, and E. Gomez, "End-to-end sound source separation conditioned on instrument labels," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2019, pp. 306–310.

[18] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[19] P. Smaragdis and G. J. Mysore, "Separation by "humming": User-guided sound extraction from monophonic mixtures," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), 2019, pp. 69–72.

[20] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[21] S. Nakamura, K. Hiyane, F. Asano, and T. Endo, "Sound scene data collection in real acoustical environments," *The Journal of the Acoustic Society of Japan (E)*, vol. 20, no. 3, pp. 225–231, 1999.

[22] Y. Okamoto, K. Imoto, S. Takamichi, R. Yamanishi, T. Fukumori, and Y. Yamashita, "RWCP-SSD-Onomatopoeia: Onomatopoeic words dataset for environmental sound synthesis," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), 2020, pp. 125–129.

[23] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proc. International Conference on Learning Representation* (*ICLR*), 2020, pp. 1–13.

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