APSIPA Transactions on Signal and Information Processing, 2022, 11, e13

Original Paper

Onoma-to-wave: Environmental Sound Synthesis from Onomatopoeic Words

Yuki Okamoto[1], Keisuke Imoto[2], Shinnosuke Takamichi[3], Ryosuke Yamanishi[4], Takahiro Fukumori[1] and Yoichi Yamashita[1*]

[1] *Ritsumeikan University, Shiga, Japan*
[2] *Doshisha University, Kyoto, Japan*
[3] *The University of Tokyo, Tokyo, Japan*
[4] *Kansai University, Osaka, Japan*

ABSTRACT

In this paper, we propose a framework for environmental sound synthesis from onomatopoeic words. As one way of expressing an environmental sound, we can use an onomatopoeic word, which is a character sequence for phonetically imitating a sound. An onomatopoeic word is effective for describing diverse sound features. Therefore, the use of onomatopoeic words as input for environmental sound synthesis will enable us to generate diverse sounds. To generate diverse sounds, we propose a method based on a sequence-to-sequence framework for synthesizing environmental sounds from onomatopoeic words. We also propose a method of environmental sound synthesis using onomatopoeic words and sound event labels. The use of sound event labels in addition to onomatopoeic words enables us to capture each sound event's feature depending on the input sound event label. Our subjective experiments show that our proposed methods achieve higher diversity and naturalness than conventional methods using sound event labels.

*Corresponding author: Yuki Okamoto, y-okamoto@ieee.org. This work was supported by JSPS KAKENHI Grant Number JP19K20304 and ROIS NII Open Collaborative Research 2021 Grant Number 21S0502.

Received 11 October 2021; Revised 05 February 2022
ISSN 2048-7703; DOI 10.1561/116.00000049

Keywords: Environmental sound synthesis, sound event, onomatopoeic word, sequence-to-sequence model

1 Introduction

Environmental sound synthesis is a research field of sound generation and is the task of generating natural environmental sounds. Many environmental sounds are used in the production of movies, games, and other content [13]. However, there is a limit to the amount of environmental sound data that is openly available. In addition, there are cases where the environmental sound that exactly matches the required sound does not exist. Therefore, it is possible to solve these problems by using environmental sound synthesis. Moreover, environmental sound synthesis has great potential for many applications such as supporting movie and game production [9, 13, 21, 23], and data augmentation for sound event detection and scene classification [4, 18].

In recent years, some methods of environmental sound synthesis using deep learning approaches have been developed [9, 11, 15]. One of the methods of environmental sound synthesis uses sound event labels as the input [15]. The method enables the generation of environmental sounds expressing sound events. In this method, since only sound event labels are input to the system, similar sounds are generated for the given sound event class; thus, the generated sounds are not sufficiently varied. Another possibility of environmental sound synthesis is to use onomatopoeic words, which are character sequences that phonetically imitate sounds. According to the studies of Lemaitre and Rocchesso [10] and Sundaram and Narayanan [19], onomatopoeic words are effective for expressing the features of audio samples. For example, when expressing *the sound of a whistle* using onomatopoeic words, we can distinguish the sounds with different durations and pitches using the length of the phoneme sequence, such as "py u" (short whistle) and "p i i i" (long whistle). Based on the idea of mapping onomatopoeic words to environmental sounds, Kawai developed KanaWave [1], software that generates environmental sounds from onomatopoeic words. KanaWave generates environmental sounds by simply connecting multiple sounds corresponding to the input onomatopoeic words, each of which is associated with a specific sound in a one-to-one correspondence. Therefore, the sounds generated by KanaWave do not have sufficient naturalness and diversity. To utilize environmental sounds in media content, such as in animation and movie production, an environmental sound synthesis method that can generate synthesized sounds with high naturalness and diversity is required.

In this paper, we propose environmental sound synthesis from onomatopoeic words using a statistical approach. Statistical methods make it possible to automatically learn the correspondence between environmental sounds and onomatopoeic words from large amounts of data with high diversity.

Even if there is a large dataset, the diversity of generated sounds is limited because the conventional method generates sounds by combining sounds in a dataset. On the other hand, a statistical method enables us to generate more diverse synthesized sounds than conventional methods. In the proposed method, we utilize the sequence-to-sequence conversion framework (seq2seq framework) [20] to generate environmental sounds from onomatopoeic words. The seq2seq framework is often used in some tasks of sequence transformation, such as those in speech synthesis and neural machine translation, and it has shown high performance in many studies [3, 6, 22]. The seq2seq framework uses several layers of recurrent neural network (RNN), which can model time-series information. Therefore, the seq2seq framework enables us to generate environmental sounds by considering the phoneme sequence for an onomatopoeic word. We also propose a method of environmental sound synthesis using sound event labels, which are used in the conventional method, and onomatopoeic words. The purpose of using onomatopoeic words and sound event labels as input is to control diversity and overall acoustic features concerning the type of sound sources, respectively.

The remainder of this paper is structured as follows. In Section 2, we describe the proposed methods of environmental sound synthesis from an onomatopoeic word. In Section 3, we report subjective experiments carried out to evaluate the performance of environmental sound synthesis from an onomatopoeic word. Finally, we summarize and conclude this paper in Section 4.

2 Proposed Method

2.1 Overview of Environmental Sound Synthesis from Onomatopoeic Words

Figure 1 shows the framework of environmental sound synthesis from onomatopoeic words. This approach consists of a model training block and a sound synthesis block. In the model training block, acoustic feature sequence $\boldsymbol{o}$ and phoneme sequence $\boldsymbol{l}$ are extracted from environmental sounds and onomatopoeic words, respectively. Acoustic model parameter $\boldsymbol{\lambda}$ is estimated using extracted features $\boldsymbol{o}$ and $\boldsymbol{l}$ as follows:

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} P(\boldsymbol{o} \mid \boldsymbol{l}, \boldsymbol{\lambda}). \tag{1}$$

In this paper, we propose two model training methods as follows.

1. Model training method using only onomatopoeic words as input to network (Section 2.2.1)
2. Model training method using onomatopoeic words and sound event labels as input to network (Section 2.2.2)

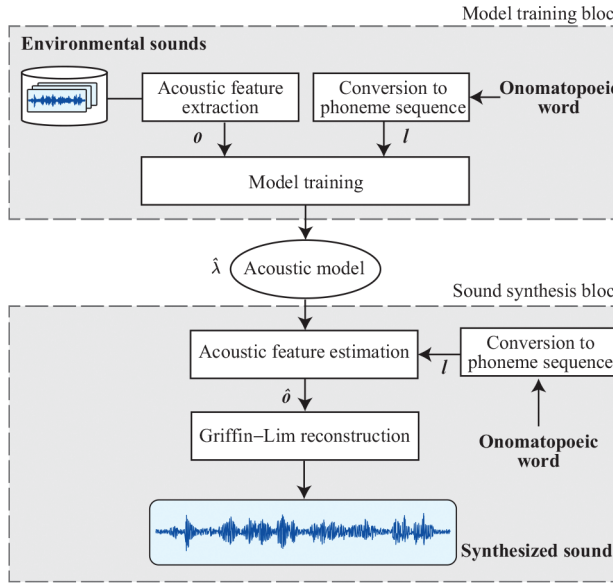

Figure 1: Overview of environmental sound synthesis using onomatopoeia.

We will detail the model training methods in Section 2.2. In the sound synthesis block, phoneme sequence $\boldsymbol{l}$ is converted from an input onomatopoeic word. Acoustic feature sequence $\boldsymbol{o}$ is estimated from a phoneme sequence $\boldsymbol{l}$ of the onomatopoeic word and acoustic model $\hat{\boldsymbol{\lambda}}$ as follows:

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid \boldsymbol{l}, \hat{\boldsymbol{\lambda}}). \tag{2}$$

Finally, we reconstruct an environmental sound wave from estimated acoustic feature sequence $\hat{\boldsymbol{o}}$ using the Griffin–Lim algorithm [5].

2.2 Proposed Model Training Methods

2.2.1 Environmental Sound Synthesis Using Onomatopoeic Words

Figure 2 shows an overview of model training using onomatopoeic words. To synthesize environmental sounds from onomatopoeic words, we employ the seq2seq framework [20]. The seq2seq framework comprises an encoder and a decoder. Our method uses one-layered bidirectional long short-term memory (BiLSTM) as the encoder and two-layered long short-term memory (LSTM) as the decoder. As shown in Figure 2, a phoneme sequence of the onomatopoeic word, $\boldsymbol{l} = \{l_1, \ldots, l_T\}$, is input to the encoder. The encoder extracts feature vectors $\boldsymbol{\nu} = [\boldsymbol{\nu}^f, \boldsymbol{\nu}^b]$ from input sequence $\boldsymbol{l}$. Superscripts f and b indicate

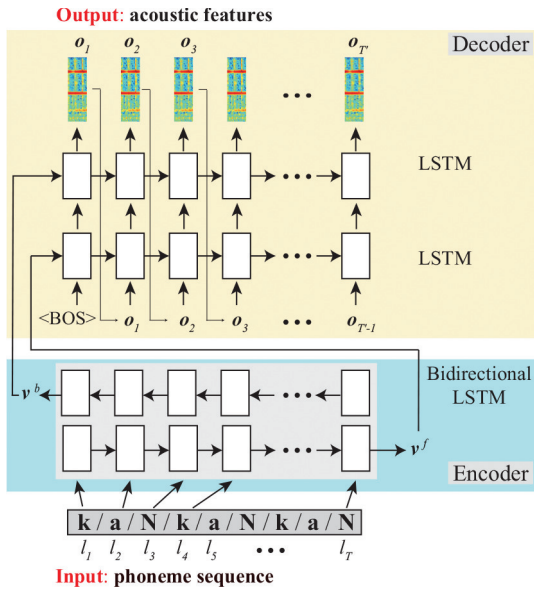

Figure 2: Environmental sound synthesis from onomatopoeic words.

forward and backward networks, respectively. In unidirectional LSTM, the beginning features tend to be lost when the sequence is long. Therefore, using BiLSTM for the encoder, we can expect to extract a feature vector $\boldsymbol{\nu}$ that captures entire onomatopoeic words from past and future directions. The decoder estimates acoustic feature sequence $\boldsymbol{o} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_{T'}\}$ from extracted feature vectors $\boldsymbol{\nu}$ in the encoder as follows:

$$p(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_{T'} \mid l_1, \ldots, l_T) = \prod_{t=1}^{T'} p(\boldsymbol{o}_t \mid \boldsymbol{\nu}, \boldsymbol{o}_1, \ldots, \boldsymbol{o}_{t-1}). \tag{3}$$

Using two-layered LSTM for the decoder, we can expect to estimate acoustic features by considering features in the forward and backward directions of onomatopoeic words extracted by the encoder. The L1 norm between the estimated acoustic feature sequence $\boldsymbol{o}$ and the target feature sequence at each time step is used as the loss function.

2.2.2 Environmental Sound Synthesis Using Onomatopoeic Words and Sound Event Labels

The method of environmental sound synthesis using only onomatopoeic words is expected to enable the control of the time-frequency structural features of

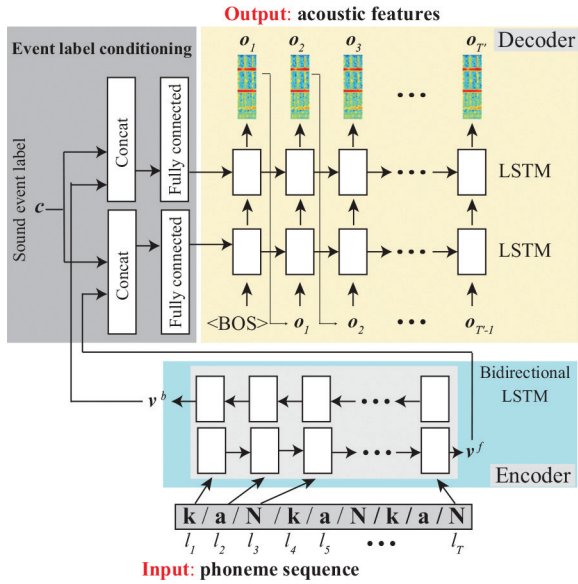

Figure 3: Environmental sound synthesis from onomatopoeic words and sound event labels.

synthesized sounds, such as sound duration. The method of environmental sound synthesis using only onomatopoeic words will generate diverse sounds. However, for example, the onomatopoeic word "p a N" could be considered to fit multiple sound events, such as *the sound of shooting guns* and *balloons breaking*. Therefore, we cannot control the frequency property associated with the sound categories using only onomatopoeic words. To control the frequency characteristics of sound events, we utilize sound event labels in addition to onomatopoeic words.

Figure 3 shows an overview of model training using onomatopoeic words and sound event labels. The method uses the seq2seq framework comprising one-layered BiLSTM as the encoder and two-layered LSTM as the decoder. The seq2seq-based intersequence conversion may involve conditioning on the decoder to control the decoder's output features [2, 7, 8, 17]. In the proposed method, sound event labels $\boldsymbol{c}$ represented as one-hot vectors and extracted feature vectors $\boldsymbol{\nu}$ are concatenated and given as the initial state of the decoder. The decoder estimates acoustic feature sequence $\boldsymbol{o} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_{T'}\}$ from extracted feature vectors $\boldsymbol{\nu}$ in the encoder and sound event labels $\boldsymbol{c}$ as follows:

$$p(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_{T'} \mid l_1, \ldots, l_T) = \prod_{t=1}^{T'} p(\boldsymbol{o}_t \mid \boldsymbol{\nu}, \boldsymbol{o}_1, \ldots, \boldsymbol{o}_{t-1}, \boldsymbol{c}). \tag{4}$$

The L1 norm between the estimated acoustic feature sequence $\boldsymbol{o}$ and the target at each time step is used as the loss function.

3 Experiments

The synthesized sounds designed for background sounds or sound effects in movies and games sounds should have high naturalness and diversity. From this viewpoint, we conducted two types of subjective test. For synthesized sounds, we evaluated their (i) naturalness and (ii) sound diversity as environmental sounds. We aim to achieve the same level of quality as natural sound in terms of both the naturalness and diversity of the generated sound.

3.1 Experimental Conditions

For the evaluation, we used 10 types of sound event (*bell ringing*, *alarm clock*, *manual coffee grinder*, *cup clinking*, *drum*, *maracas*, *electric shaver*, *tearing paper*, *trash box banging*, and *whistle*[1]) contained in the Real World Computing Partnership-Sound Scene Database (RWCP-SSD) [14]. We used a total of 1000 samples (100 samples × 10 sound events), in which 95 samples of each sound event were used for model training and the others were used for the subjective test. For the onomatopoeic words corresponding to each sound sample, we used the dataset in RWCP-SSD-Onomatopoeia [16]. There are many onomatopoeic words that contain some syllables in RWCP-SSD-Onomatopoeia. Each sound sample has more than 15 onomatopoeic words, and we used 15 onomatopoeic words per audio sample for model training for a total of 14,250 onomatopoeic words (15 onomatopoeic words × 950 audio samples). Table 1 shows the experimental conditions and parameters used for the proposed methods. In this study, we use the log-amplitude spectrogram as an acoustic feature. The generated audio samples are available on our web page.[2]

3.2 Subjective Evaluations

Following the evaluation perspective described at the beginning of Section 3, we conducted the following two sets of experiments:

3.2.1 Experiment I: Evaluation of Naturalness for Environmental Sounds

The target sound of this paper is a sound that is comfortable as an environmental sound and that expresses the input onomatopoeic word. There are two perspectives of naturalness that should be satisfied. For this reason, we designed several experiments to evaluate each perspective. In Experiments I-1 and II-2, we presented environmental sounds and the onomatopoeic word used for the input, and evaluated how acceptable or expressive the presented sounds

[1] *Whistle* refers to the sound of whistles such as these carried by policemen to give warning as the need arises.

[2] https://y-okamoto1221.github.io/Onoma_to_wave_Demonstration/

Table 1: Experimental conditions.

Sound length	1–2 s
Sampling rate	16,000 Hz
Waveform encoding	16-bit linear PCM
Acoustic feature	log-amplitude spectrogram
Window length for FFT	0.128 s (2048 samples)
Window shift for FFT	0.032 s (512 samples)
Encoder LSTM layers	1
# units in encoder LSTM	512
Decoder LSTM layers	2
# units in decoder LSTM	512, 512
Event label dimensions	10
Teacher forcing rate	0.6
Batch size	5
Optimizer	RAdam [12]

were in relation to the onomatopoeic word. In Experiment I-3, only the sound was presented to evaluate its naturalness as an environmental sound, and the sound itself was simply evaluated in terms of "quality as an environmental sound."

- **Experiment I-1: acceptance level of synthesized sounds for onomatopoeic words**
  We presented pairs of a sound (natural or synthesized) and an onomatopoeic word. The listener graded the acceptance level of the synthesized and natural sounds for onomatopoeic words on a scale of 1 (highly unacceptable) to 5 (highly acceptable).

- **Experiment I-2: expressiveness of synthesized sounds for onomatopoeic words**
  We presented pairs of a sound (natural or synthesized) and an onomatopoeic word. The listener graded the expressive level of the synthesized and natural sounds for onomatopoeic words on a scale of 1 (very unexpressive) to 5 (very expressive).

- **Experiment I-3: naturalness of environmental sounds**
  We presented a natural or synthesized sound. The listener graded the naturalness of the synthesized and natural sounds on a scale of 1 (very unnatural as an environmental sound) to 5 (very natural as an environmental sound).

Table 2: Number of synthesized sounds used for subjective test.

Experiment	# labels	# samples in each label	# listeners	# total samples
Exp. I-1	10	10	30	3000
Exp. I-2	10	10	30	3000
Exp. I-3	10	5	30	1500
Exp. II-1	5	5	30	750
Exp. II-2	10	2–3	50	1300

3.2.2 Experiment II: Evaluation of Sound Diversity for Environmental Sounds

To evaluate diversity for synthesized sounds, we conducted two types of subjective evaluation as follows:

- **Experiment II-1: diversity of synthesized sounds for each sound event**
  We presented two sounds synthesized by the same method to listeners. In the proposed method, sounds are generated using randomly selected onomatopoeic words from the overall dataset as the input in each sound event. The listener graded the dissimilarity level between two presented sounds on a scale of 1 (very similar) to 5 (very dissimilar).

- **Experiment II-2: diversity of synthesized sounds for the same onomatopoeic words**
  We presented listeners with synthesized sound, and the listeners selected the sound event label that best represents the sound from ten choices.

Each experiment was conducted using a crowdsourcing platform. Table 2 shows the numbers of audio samples and listeners in each experiment. The dataset used in these experiments, RWCP-SSD-Onomatopoeia, contains only onomatopoeic words collected from Japanese speakers. Onomatopoeic words given to a sound differ depending on the native language. Thus, the sounds generated were evaluated only by Japanese speakers.

To compare the synthesis methods, we evaluated the sounds synthesized by the conventional method using WaveNet [15] and KanaWave [1]. In the conventional environmental sound synthesis method using WaveNet, the sound event label is used as the input to the system, and the waveform sample in the next time is predicted by finding $\boldsymbol{x}$, where the following equation is maximized.

$$p(\boldsymbol{x} \mid \boldsymbol{c}) \approx \prod_{t=1}^{T} p(x_t \mid x_{t-R}, \ldots, x_{t-1}, \boldsymbol{c}), \tag{5}$$

where $\boldsymbol{x}, \boldsymbol{c}$, and R indicate the generated waveform, sound event label represented as one-hot vectors, and receptive field, respectively. The receptive

Table 3: List of synthesis methods evaluated for each evaluation metric.

Method	Exp. I-1	Exp. I-2	Exp. I-3	Exp. II-1	Exp. II-2
Natural sound	✓	✓	✓		
WaveNet			✓	✓	
KanaWave	✓	✓	✓		
Seq2seq **(proposed)**	✓	✓	✓		✓
Seq2seq + event label **(proposed)**	✓	✓	✓	✓	✓

field is a parameter related to the number of samples of waveforms required to go back to in the past. The method using seq2seq estimates the acoustic features and restores the waveform on the basis of the estimates, while the method using WaveNet estimates the waveform directly. The conventional method using WaveNet [15] does not input onomatopoeic words. Therefore, we evaluated synthesized sounds by WaveNet in only experiments I-3 and II-1. KanaWave is the conventional non-statistical method of generating environmental sounds from only onomatopoeic words. KanaWave generates sounds by simply connecting multiple sounds corresponding to the input onomatopoeic words, each of which is associated with a specific sound in a one-to-one correspondence. The system generates a waveform when onomatopoeic words in *katakana*, which is a Japanese syllabary are used as input. There are several parameters that can be set to adjust the pitch of the sound. However, it is low naturalness because it generates sounds by simply connecting multiple sounds. The list of synthesis methods evaluated in each experiment is shown in Table 3.

3.3 Experimental Results and Discussion

3.3.1 Experiment I

Experiments I-1 and I-2: the average acceptance and expressiveness scores of synthesized and natural sounds for onomatopoeic words and their standard deviations are respectively shown in Figures 4 and 5. The results show that our proposed methods can generate environmental sounds that are a better representation of onomatopoeic words than those generated by KanaWave.

Figure 6 shows spectrograms of sounds synthesized by our methods using only onomatopoeic words. The phoneme representation /q/ in Figure 6 is a pronunciation called a double consonant. As shown in Figure 6, the proposed

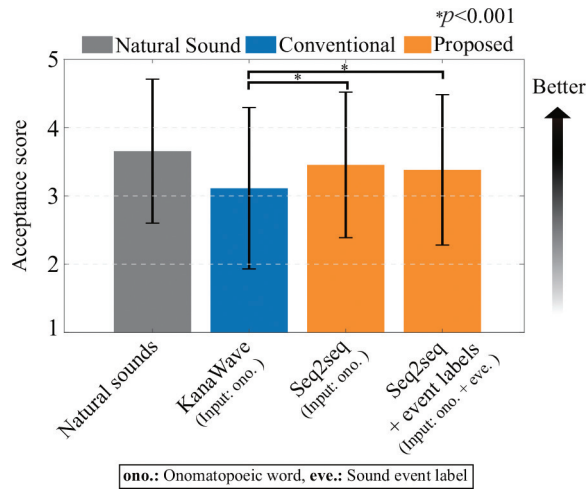

Figure 4: Acceptance scores of natural and synthesized sounds.

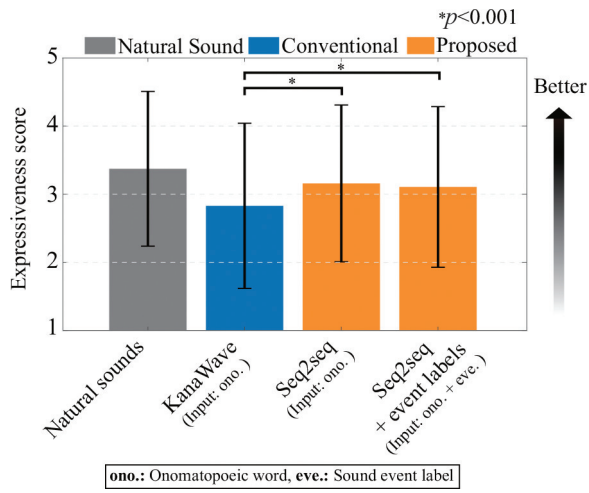

Figure 5: Expressiveness scores of natural and synthesized sounds.

method can generate diverse environmental sounds. Also, the longest sound (right) is not the sound given by simply stretching the other sounds (left and center). Thus, onomatopoeic words are useful for generating diverse sounds with different characteristics, such as sound duration.

Figure 7 shows the spectrograms of sounds synthesized by KanaWave and the proposed method using both onomatopoeic words and sound event labels.

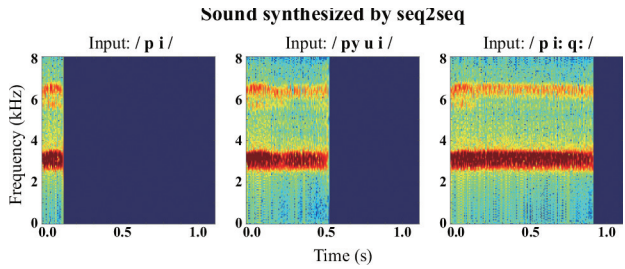

Figure 6: Spectrograms of environmental sounds synthesized using only onomatopoeic words.

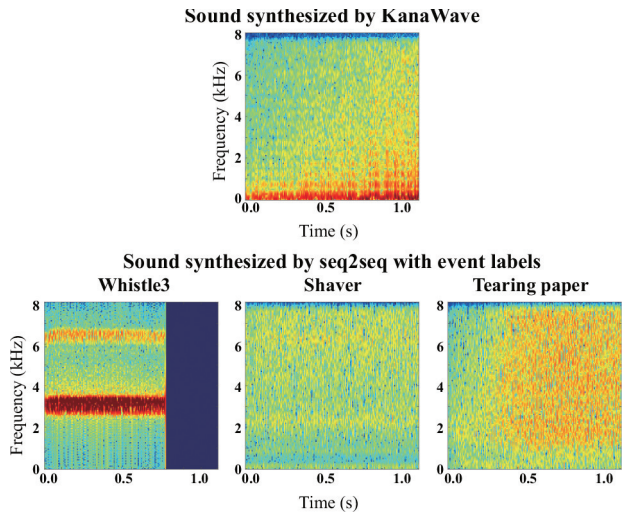

Figure 7: Spectrograms of environmental sounds synthesized by KanaWave and the proposed method using onomatopoeic words and sound event labels.

In Figure 7, each synthesized sound is generated from a phoneme sequence of the onomatopoeic word "b i i i i i i" input to the system. In the proposed method using both onomatopoeic words and sound event labels, we used sound event labels of *whistle*, *electric shaver*, and *tearing paper*. KanaWave can only generate one type of sound from the same onomatopoeic word. Therefore, the sound synthesized by KanaWave does not have diversity. On the other hand, the proposed method using onomatopoeic words and sound event labels can generate various sounds from the same onomatopoeic word by changing the input sound event labels.

Experiment I-3: the average MOS scores for the naturalness of synthesized and natural sounds, and their standard deviations are shown in Figure 8. The results indicate that sounds synthesized by the proposed methods achieve

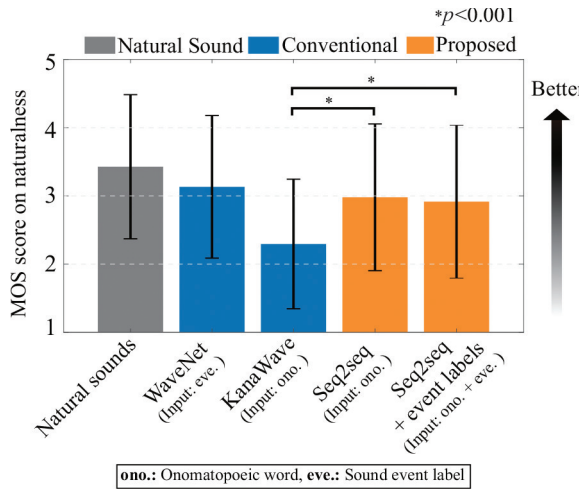

Figure 8: MOS scores for naturalness of natural and synthesized sounds.

higher naturalness than those synthesized by KanaWave. The experimental results also show that sounds synthesized by our methods had a similar sound quality to those synthesized by WaveNet. Thus, the proposed methods achieve environmental sound synthesis from onomatopoeic words without degrading the sound quality compared with conventional methods. In addition, natural sounds still have higher naturalness than sounds synthesized by the proposed methods. From these results, it is still necessary to develop a method of environmental sound synthesis that can provide quality equivalent to that of natural sounds.

3.3.2 Experiment II

Experiment II-1: the average dissimilarity score of the synthesized sound for each sound event is shown in Figure 9. In this experiment, a high dissimilarity means that there is a rich diversity of synthesized sounds within the same type of event. The experimental results indicate that the method using seq2seq with onomatopoeic words and sound event labels as input can generate more diverse sounds than the method using seq2seq with only sound event labels as input. This result shows that onomatopoeic words enable us to generate diverse sounds. By comparing the conventional method using WaveNet with only sound event labels as input and the proposed method with onomatopoeic words and sound event labels as input, we found that the proposed method can generate more diverse sounds for *drum* and *shaver*. On the other hand, the sounds synthesized for *cup* and *maracas* by our proposed method had a similar

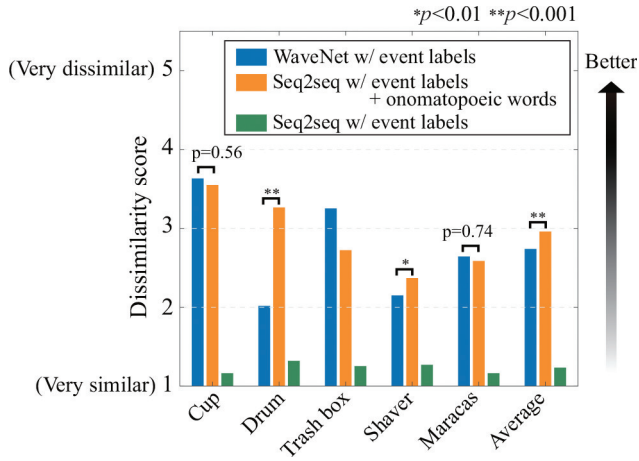

Figure 9: Dissimilarity scores of synthesized sounds.

diversity to those synthesized by WaveNet. The conventional method using WaveNet tends to include noise in the generated sound. The sound generated by WaveNet tends to get a high dissimilarity score owing to these noises. On the other hand, the proposed method can generate diverse and comparatively clear sounds with low noise. Thus, the proposed method enables us to generate diverse environmental sounds by using onomatopoeic words.

Experiment II-2: part of the distributions of sound event labels given to the synthesized sound from each onomatopoeic word are shown in Figure 10. The sounds synthesized by our method using only onomatopoeic words tend to be given only one sound event label. On the other hand, the sounds synthesized by our method using both onomatopoeic words and sound event labels tend to be given various sound event labels. The entropies of the distribution of a given acoustic event label are 1.70 bit for the method using only onomatopoeic words and 1.82 bit for the method using both onomatopoeic words and sound event labels. In this experiment, the maximum entropy is 3.02 bit when 10 types of sound event labels equally appear for each synthesized sound. This result shows that using both onomatopoeic words and sound event labels can represent multiple sound events for the same onomatopoeic word.

Figure 11 shows spectrograms of natural and synthesized sounds. In Figure 11, each synthesized sound is generated from a phoneme sequence of the onomatopoeic word "b i: i q" input to the system. In the proposed method using both onomatopoeic words and sound event labels, we used the sound event labels of *whistle*, *electric shaver*, and *tearing paper*. As shown in Figure 11, using only onomatopoeic words as an input generates sounds with

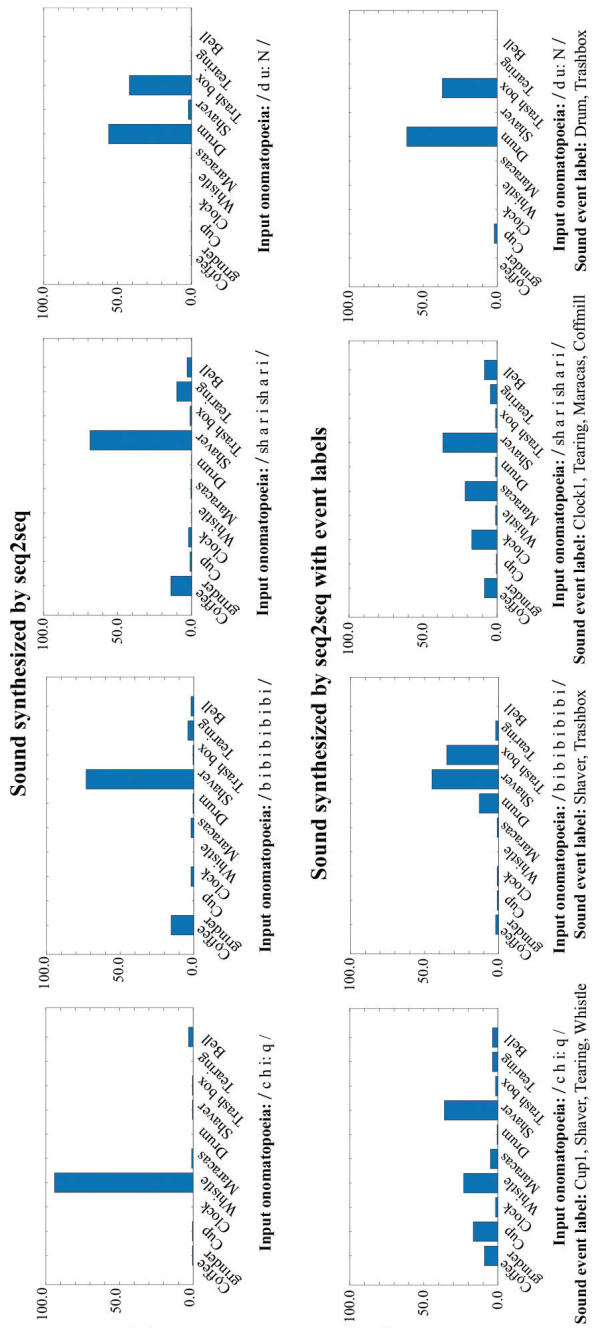

Figure 10: Number of responses of sound event labels to each sound synthesized by our method using onomatopoeic words.

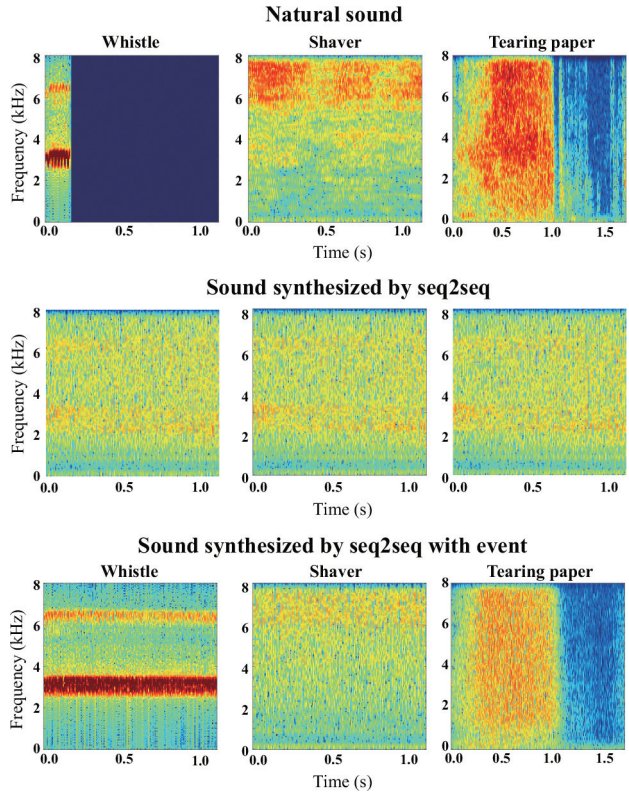

Figure 11: Spectrograms of synthesized environmental sounds, which are generated from a phoneme sequence of the onomatopoeic word "b i: iq", and natural sounds.

similar features when the initial values of model parameters in model training are changed. On the other hand, using both onomatopoeic words and sound event labels, it is possible to generate sounds that capture each sound event's feature depending on the input sound event label. These results also show that using sound event labels can control sound events of sound synthesized from onomatopoeic words.

4 Conclusion

In this paper, we proposed environmental sound synthesis from onomatopoeic words. We found that the proposed methods can generate sounds with high naturalness and diversity. The experimental result indicates that the use of sound event labels in addition to onomatopoeic words as input enables us to

control the sound events of generated sounds and to generate diverse sounds. In the future, we will generate environmental sounds from onomatopoeic words using more types of sound event.

Biographies

Yuki Okamoto received his B.E. and M.E. degrees from Ritsumeikan University in 2019 and 2021, respectively. He is currently pursuing a Ph.D. degree at the Graduate School of Information Science and Engineering, Ritsumeikan University. His research interests are environmental sound synthesis and sound event detection. He is a member of the IEEE Signal Processing Society and the Acoustic Society of Japan (ASJ).

Keisuke Imoto received his B.E. and M.E. degrees from Kyoto University in 2008 and 2010, respectively. He received his Ph.D. degree from SOKENDAI (The Graduate University for Advanced Studies) in 2017. He joined the Nippon Telegraph and Telephone Corporation (NTT) in 2010 and the Ritsumeikan University as an Assistant Professor in 2017. He moved to Doshisha University as an Associate Professor in 2020. He has been engaged in research on sound event detection, acoustic scene analysis, anomalous sound detection, and microphone array signal processing. He is a member of IEEE, EURASIP, APSIPA, ASJ, and IEICE.

Shinnosuke Takamichi received the B.E. degree from Nagaoka University of Technology, Nagaoka, Japan, in 2011, and the M.E. and Ph.D. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan, in 2013 and 2016, respectively. He is currently an Assistant Professor at The University of Tokyo. He has received more than 20 paper/achievement awards including the 2020 IEEE Signal Processing Society Young Author Best Paper Award.

Ryosuke Yamanishi received his B.E., M.E., and Ph.D. from Nagoya Institute of Technology, Japan, 2007, 2009, and 2012, respectively. He joined in Ritsumeikan University as a research associate in 2012, a research assistant professor in 2013, an assistant professor in 2014, and a lecturer in 2018. During this period, he visited the laboratory of computational intelligence of UBC, Canada as a visiting assistant professor. In 2020, he has joined Kansai University as an associate professor. He is interested in computational culture and arts. He is a member of IEICE, IPSJ, JSAI, JSKE, ASJ, SAS, ACM, and ACL.

Takahiro Fukumori received his B.E., M.E., and Ph.D. degrees from Ritsumeikan University in 2010, 2012, and 2015, respectively. From 2012 to 2015,

he was a JSPS research fellowship for young scientists (DC1). From 2015 to 2020, he was an assistant professor at Ritsumeikan University. He is currently a lecturer at Ritsumeikan University. His current research interests include speech recognition and speech enhancement. He is a member of IEEE, ASJ, IEICE, IPSJ, and RSJ.

Yoichi Yamashita received his B.E., M.E. and Ph.D. degrees from Osaka University in 1982, 1984, and 1993, respectively. He has worked for the Institute of Scientific and Industrial Research of Osaka University as a Technical Official, a Research Associate, and an Assistant Professor from 1984 to 1997. In 1997, he joined Ritsumeikan University as an Associate Professor in the College of Science and Engineering. He is currently a Professor in the College of Information Science and Engineering. His research interests include speech recognition, speech synthesis, acoustic signal processing, and spoken document processing. He is a member of IEICE, ASJ, IPSJ, JSAI, ISCA, and IEEE.

References

[1] "KanaWave," https://www.vector.co.jp/soft/win95/art/se232653.html.

[2] E. Cooper, C. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-Shot Multi-Speaker Text-To-Speech with State-Of-The-Art Neural Speaker Embeddings," in, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, 6184–8.

[3] K. Drossos, S. Advanne, and T. Virtaren, "Automated Audio Captioning with Recurrent Neural Networks," in, *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), 2017, 374–8.

[4] F. Gontier, M. Lagrange, C. Lavandier, and J. F. Petiot, "Privacy Aware Acoustic Scene Synthesis Using Deep Spectral Feature Inversion," in, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, 886–90.

[5] D. Griffin and J. Lim, "Signal Estimation from Modified Short-time Fourier Transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2), 1984, 236–43.

[6] S. Ikawa and K. Kashino, "Generating Sound Words from Audio Signals of Acoustic Events with Sequence-to-Sequence Model," in, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2018, 346–50.

[7] S. Ikawa and K. Kashino, "Neural Audio Captioning Based on Conditional Sequence-to-Sequence Model," in, *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), 2019, 99–103.

[8] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis," in, *Proc. Advances in Neural Information Process. Systems* (*NIPS*), 2018, 4480–90.

[9] Q. Kong, Y. Xu, T. Iqbal, Y. Cao, W. Wang, and M. D. Plumbley, "Acoustic Scene Generation with Conditional SampleRNN," in, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2019, 925–9.

[10] G. Lemaitre and D. Rocchesso, "On the Effectiveness of Vocal Imitations and Verbal Descriptions of Sounds," *The Journal of the Acoustical Society of America*, 135(2), 2014, 862–73.

[11] J.-Y. Liu, Y.-H. Chen, Y.-C. Yeh, and Y.-H. Yang, "Unconditional Audio Generation with Generative Adversarial Networks and Cycle Regularization," *arXiv preprint arXiv:2005.08526*, 2020.

[12] 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, 1–13.

[13] D. B. Lloyd, N. Raghuvanshi, and N. K. Govindaraju, "Sound Synthesis for Impact Sounds in Video Games," in, *Proc. Symposium on Interactive 3D Graphics and Games. ACM*, 2011, 55–61.

[14] S. Nakamura, K. Hiyane, F. Asano, and T. Endo, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," *Proc. Language Resources and Evaluation Conference* (*LREC*), 2000, 965–8.

[15] Y. Okamoto, K. Imoto, T. Komatsu, S. Takamichi, T. Yagyu, R. Yamanishi, and Y. Yamashita, "Overview of Tasks and Investigation of Subjective Evaluation Methods in Environmental Sound Synthesis and Conversion," *arXiv preprint arXiv:1908.10055*, 2019.

[16] 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, 125–9.

[17] J. Park, K. Zhao, K. Peng, and W. Ping, "Multi-speaker End-to-end Speech Synthesis," *arXiv preprint arXiv:1907.04462*, 2019.

[18] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A Library for Soundscape Synthesis and Augmentation," in, *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), 2017, 344–8.

[19] 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, 55–8.

[20] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in, *Proc. Advances in Neural Information Process. Systems* (*NIPS*), 2014, 3104–12.

[21] K. Wang, H. Cheng, and S. Liu, "Efficient Sound Synthesis for Natural Scenes," in, *Proc. IEEE Virtual Reality* (*VR*), 2017, 303–4.

[22] Y. Wang, R. S. Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards End-To-End Speech Synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[23] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, "Visual to Sound: Generating Natural Sound for Videos in the Wild," in, *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2018, 3550–8.