INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

CAPTDURE: Captioned Sound Dataset of Single Sources

Yuki Okamoto[1], *Kanta Shimonishi*[1], *Keisuke Imoto*[2],
Kota Dohi[3], *Shota Horiguchi*[3], *Yohei Kawaguchi*[3]

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

y-okamoto@ieee.org, is0460rf@ed.ritsumei.ac.jp, keisuke.imoto@ieee.org,
{kota.dohi.gr,shota.horiguchi.wk,yohei.kawaguchi.xk}@hitachi.com

Abstract

In conventional studies on environmental sound separation and synthesis using captions, datasets consisting of multiple-source sounds with their captions were used for model training. However, when we collect the captions for multiple-source sound, it is not easy to collect detailed captions for each sound source, such as the number of sound occurrences and timbre. Therefore, it is difficult to extract only the single-source target sound by the model-training method using a conventional captioned sound dataset. In this work, we constructed a dataset with captions for a single-source sound named CAPTDURE, which can be used in various tasks such as environmental sound separation and synthesis. Our dataset consists of 1,044 sounds and 4,902 captions. We evaluated the performance of environmental sound extraction using our dataset. The experimental results show that the captions for single-source sounds are effective in extracting only the single-source target sound from the mixture sound.

Index Terms: target sound extraction, environmental sound, sound event detection, natural language processing

1. Introduction

Environmental sounds are indispensable to enhance the sense of presence and immersion in media content such as movies and games. One way to prepare a desired single-source sound is to obtain it from a sound database [1]. However, the required environmental sounds may not always exist in the database. Moreover, there are many multiple-source sounds on the Internet, but obtaining only the single-source target sound is not easy.

To obtain only the single-source target sound, environmental sound extraction and separation methods have been proposed [2, 3, 4, 5, 6, 7] for extracting only the single-source target sound from a multiple-source sound. For example, a method has been proposed with which a sound event class is specified and only sounds corresponding to the sound event class are extracted [2, 3, 5, 8]. To extract sounds more accurately than the sound event class, the method using onomatopoeic words that transcribes the characteristics of the sound to be extracted has been proposed [9]. However, since onomatopoeic words are language and culture dependent, the dataset for model training needs to be modified in accordance with the user's native language. To avoid dependence on language and culture, a method of environmental sound extraction and separation using captions for sound has also been proposed [10, 11, 12]. With this method, the characteristics of the multiple-source target sound are expressed by captions and the multiple-source target sound can be extracted. This method enables us to extract the multiple-source target sound, which includes some sound event classes corresponding to the caption. The conventional environmental sound-extraction method using captions is based on a dataset [13, 14, 15] in which captions are assigned to multiple-source sounds. The conventional dataset Table 1 shows captions collected for single- and multiple-source sounds in our dataset. As shown in Table 1, captions collected for multiple-source sounds are less detailed for a single sound event than those collected for a single-source sound. Therefore, it is difficult to extract only a single-source target sound using captioned sound dataset collected for multiple-source sounds. We need the caption for that single-source sound to extract a single-source target sound from multiple-source sounds.

Table 1: *Examples of captions in CAPTDURE. The sound sources are underlined.*

#Sources	Caption
Single	• One or two keyboards continue to be pressed slowly with a light, high-touch sound. • The sound of the buttons on the keyboard being pressed one by one quite slowly. • The electronic tone of the digital alarm clock is high-pitched and continues to sound slowly to gradually faster. • The bright, high-pitched alarm clock bells are ringing.
Multiple	• Keyboard typing and mouse clicking noises are continually heard. • The sound of a toaster rang out as if to counteract the sound of typing on a small keyboard. • The alarm clock is drowned out by the loud noise of the hair dryer. • The alarm clock beeps at equal intervals, with one final sound to close the lock.

Environmental sound synthesis methods, which artificially generate sounds, have also been proposed to obtain the required sound [16, 17, 18, 19, 20]. Recently, there are some studies of environmental sound synthesis using captions [21, 22, 23, 24]. However, no captioned sound datasets represent only a single-source sound, so it is difficult to fine-tune the generated single-source sound. We need the caption corresponding to a single-source sound to control synthesized environmental sound finely.

In this paper, we constructed a dataset with captions for sound shown in Table 1 for single sound sources that can be used in various tasks that involve environmental sounds. The dataset is called CAPTDURE (CAPTioned sound Dataset of single soURcEs), pronounced as "capture." As an example of the use of the dataset, we conduct an experiment on environmental sound extraction using captions. In our experiment, we show that the use of captions for sound assigned to a single-source sound is effective in extracting only target sounds in a multiple-source sound.

The rest of the paper is organized as follows. In Sec. 2, we describe the creation of CAPTDURE. In Sec. 3, we discuss our

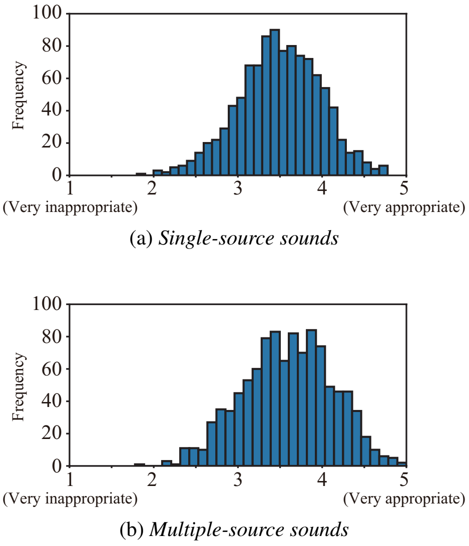
10.21437/Interspeech.2023-577

Table 2: *List of recording equipment*

Recording equipment	Software	Microphone	Audio interface
Apple/Macbook Pro (16-inch 2019)	Audacity[1]	SHURE/MX150B/O-XLR	Roland/Rubix24
TASCAM/DR-44WL	Built-in	Built-in	Built-in
Apple/iPhone SE	PCM recording[2]	Built-in	Built-in
Huawei/dtab d01-G	AudioRec[3]	Built-in	Built-in

Table 3: *Recorded sound event classes*

Sound event class	Description	#Subclasses	#Clips	Duration [s]
Keyboard	Sound of keyboard typing	5	60	465.0
Door	Sound of door opening and closing	4	48	279.0
Mouse	Sound of mouse clicking	7	84	600.0
Water tap running	Sound of water tap running	5	60	431.0
Dryer	Operating sound of dryer	6	252	1,746.0
Ventilation fan	Operating sound of ventilation fan	3	36	284.0
Door lock	Sound of door locking	6	48	245.0
Intercom	Sound of intercom	5	56	350.0
Door knock	Sound of door knock	6	72	430.0
Microwave	Operating sound of microwave	4	48	329.0
Toaster	Operating sound of toaster	5	60	465.0
Cutlery	Sound of hitting cutlery	7	52	362.0
Clock	Operating sound of alarm clock	5	60	420.0
Fan	Operating sound of fan	3	108	779.0
Total		71	1,044	7,185.0



(a) *Single-source sounds*

(b) *Multiple-source sounds*

Figure 1: *Histograms of average appropriateness score*

experiments on environmental sound extraction using language query. Finally, we summarize and conclude this paper in Sec. 4.

2. Creation of CAPTDURE

2.1. Design of CAPTDURE

The CAPTDURE dataset consists of the following contents.

- **Single-source sounds**
  We recorded a total of 1,044 single-source sounds, consisting of 14 types of daily sound events, as shown in Table 3. For each sound event, we recorded approximately 40 to 100 sounds while changing the recording conditions, such as recording equipment. The length of each sound was set to 5 to 9 seconds. Each sound event class was divided into subclasses in accordance with differences in timbre and pitch.
- **Multiple-source sounds**
  We created a total of 1,044 multiple-source sounds using the recorded single-source sounds. We selected two sounds of different sound event classes from the recorded sounds and mixed the sounds so that the signal-to-noise ratio (SNR) was 0 dB. We padded with zero the shorter sound and aligned the sound length of the two sounds.
- **Captions for single-source sounds**
  We collected a total of 4,902 captions (3 or more captions per single-source sound). We describe the details of the captions collection in Sec.2.3.
- **Captions for multiple-source sounds**
  We collected a total of 3,132 captions (3 captions per multiple-source sound).
- **Appropriateness score for each caption**
  We asked crowdworkers to score the appropriateness level for captions transcribed by others. The appropriateness scores enable us to evaluate captions on the basis of the judgment of others. We describe the details of the appropriateness scores in Sec. 2.4.
- **Worker ID**
  CAPTDURE includes anonymized IDs of workers who gave captions and appropriateness scores.

This dataset is freely available online[4].

2.2. Recording environment and setup

The environmental sounds included in CAPTDURE were recorded in two types of soundproof rooms. The first soundproof room is a length of 3.1 m, a width of 5.4 m, and a height of 2.7 m with a reverberation time (T_{60}) of 0.2 s. The second soundproof room is a length of 4.9 m, a width of 4.0 m, and a height of 2.5 m with a reverberation time (T_{60}) of 0.2 s. Some environmental sounds, such as "water tap running" and "door" that are difficult to record in a soundproof room, were recorded at locations other than these rooms. Table 2 lists the equipment used for the sound recording. The sampling frequency was 48 kHz, and the quantization bit rate was 16 bits. The microphone was set at a distance of approximately 0.3 to 0.5 m from the sound source being recorded.

2.3. Captions collection

We used Lancers[5], which is a crowdsourcing service in Japan, to collect captions for sounds from Japanese crowdworkers. In the pre-experiment, we tried using Amazon Mechanical Turk[6] (MTurk). However, comparing Lancers and MTurk, we confirmed that the quality of crowdworkers was higher in Lancers.

The collection of captions was conducted by presenting five sounds to one worker. In the pre-experiment, the five sounds presented to the crowdworkers were from different sound event classes. However, the caption of each sound mainly consisted of the caption of the sound class it belongs to. Therefore, the five sounds presented to the crowdworkers consisted of the same-sound event class, and we instructed the crowdworkers to write a different sentence as the caption given to each sound. Crowdworkers were also instructed to avoid providing only with the type of sound, such as "the sound of the microwave," or only with onomatopoeic words, such as /k a N k a N/.

We collected a total of 4,902 captions (3–5 captions per single-source sound) for the single-source sounds. Moreover, we collected a total of 3,132 captions (3 captions per multiple-source sound) for the multiple-source sounds. In addition to the captions collected in Japanese, CAPTDURE also contains

Table 4: *Statistics of train, validation, and test set of CAPTDURE*

Dataset	#Clips	#Captions	#Words (en) / #Characters (ja)
Single-source sound			
train	795	3,774	10.53 (en) / 22.80 (ja)
validation	82	374	10.62 (en) / 22.91 (ja)
test	167	501	10.43 (en) / 23.05 (ja)
Multiple-source sound			
train	795	2,385	16.83 (en) / 36.47 (ja)
validation	82	246	16.60 (en) / 35.53 (ja)
test	167	754	17.79 (en) / 38.06 (ja)

captions translated into English using the DeepL API[7].

2.4. Captions evaluation

We asked crowdworkers to score the level of appropriateness for captions transcribed by others. We present pairs of audio samples and the corresponding caption to a crowdworker. The crowdworker then gives an appropriateness score for each caption. The appropriateness score is on a five-point scale, 1 (very inappropriate) to 5 (very appropriate), for each caption by those who did not create the caption. The appropriateness scores for each caption were collected from more than three crowdworkers.

Figure 1 shows the histogram of each collected appropriateness score calculated as the average per sound for single and multiple-source sounds. Most of the captions were given an appropriateness score of 3 or higher. This fact indicates that we were able to collect many appropriate captions to express each sound.

2.5. Data splitting

We split the single- and multiple-source sounds of each subclass as approximately 7:1:2 for the train, validation, and test sets. The captions corresponding to each sound are also included in each set. The statistics of each set are shown in Table 4. We artificially split the sounds of each set to balance the number of words/characters of captions.

3. Experiments

To demonstrate the efficiency of CAPTDURE, we evaluated the performance of environmental sound extraction using caption.

3.1. Network architecture

To extract the sound that corresponds to caption $\boldsymbol{l}$ from multiple-source sound $\boldsymbol{x}$, we trained a neural network based on Conv-TasNet [25]. The T-length multiple-source sound $\boldsymbol{x} \in \mathbb{R}^{1\times T}$ is fed to the encoder, which consists of a one-dimensional (1-D) convolution layer as follows:

$$\boldsymbol{W} = \mathsf{Encoder}\left(\boldsymbol{x}\right) \in \mathbb{R}^{C\times T'}. \tag{1}$$

At the same time, the sound caption $\boldsymbol{l}$ is fed to the Bidirectional Encoder Representations from Transformers (BERT) [26]. In this experiment, we used the pre-trained English[8] and Japanese[9] BERT models. The vector obtained by BERT is compressed by linear transformation $\mathsf{Linear}(\cdot)$ to the C-dimension as follows:

Table 5: *Experimental conditions*

Waveform	
Multiple-source-sound length	10 s
Sampling rate	16 kHz
Waveform encoding	16-bit linear PCM
Parameters of Conv-TasNet	
# of filters in autoencoder	256
Length of filters in samples	20
# of channels in bottleneck	256
# of channels in convolutional block	512
Kernel size in convolutional block	3
# of convolutional blocks in each repeat	8
# of repeats	4
Parameters of model training	
Batch size	1
Training epoch	120
Learning rate	0.0001
Optimizer	RAdam [27]

$$\boldsymbol{o} = \mathsf{Linear}\left(\mathsf{BERT}\left(\boldsymbol{l}\right)\right) \in \mathbb{R}^{C}. \tag{2}$$

The soft-mask $\boldsymbol{M}$ is calculated from the obtained matrix $\boldsymbol{W}$ and vector $\boldsymbol{o}$ as

$$\boldsymbol{M} = \mathsf{Separator}\left(\ \boldsymbol{W} \odot \underbrace{[\boldsymbol{o}, \boldsymbol{o}, \ldots, \boldsymbol{o}]}_{T'}\ \right) \in (0,1)^{C\times T'}, \tag{3}$$

where $\odot$ denotes the Hadamard product and $\mathsf{Separator}(\cdot)$ is K-stacked 1-D convolutional layers. Finally, the Hadamard product of soft-mask $\boldsymbol{M}$ and $\boldsymbol{W}$ is fed to the decoder, which consists of a 1-D convolution layer, to obtain the target signal corresponding to the caption as follows:

$$\hat{\boldsymbol{y}} = \mathsf{Decoder}\left(\boldsymbol{M} \odot \boldsymbol{W}\right) \in \mathbb{R}^{1\times T}. \tag{4}$$

The network was trained to minimize the L1 norm between target sound $\boldsymbol{y}$ and estimated sound $\hat{\boldsymbol{y}}$.

3.2. Dataset construction

We randomly selected three single sounds with different sound event classes from CAPTDURE to create a mixture sound. The training and validation sets consisted of 2,385 and 246 mixture sounds, respectively. We used three randomly selected captions for each sound for our experiments.

We constructed two evaluation datasets using the test set of single-source sounds from CAPTDURE: the inter-event-class dataset and the intra-event-class dataset. Each sound in those datasets is a mixture sound of a target sound and an interference sound, but those are from different sound event classes in the inter-event-class dataset and from the same sound event class in the intra-event-class dataset. The signal-to-interference ratio was varied by $\{-10, -5, 0, 5, 10\}$ dB. Each evaluation set consisted of 501 mixture sounds for each SNR.

[1] https://www.audacityteam.org/

[2] https://ko-yasui.com/

[3] https://audiorec.jp.aptoide.com/app

[4] https://sites.google.com/view/yuki-okamoto/dataset

[5] https://www.lancers.jp/

[6] https://www.mturk.com/

[7] https://www.deepl.com/en/docs-api

[8] https://huggingface.co/bert-base-uncased

[9] https://huggingface.co/cl-tohoku/bert-base-japanese

Table 6: *SDRi [dB] for extracted signals*

		SNR				
Dataset	Model-training method	−10 dB	−5 dB	0 dB	5 dB	10 dB
Inter-event-class dataset	Training using multiple-source caption (ja)	5.12 ± 4.22	4.39 ± 4.04	3.00 ± 4.21	0.93 ± 4.92	−1.85 ± 5.95
	Training using multiple-source caption (en)	4.67 ± 4.03	3.88 ± 3.88	2.43 ± 4.04	0.17 ± 4.71	−2.78 ± 5.67
	Training using single-source caption (ja)	7.28 ± 5.19	6.26 ± 5.02	4.63 ± 5.18	2.09 ± 5.71	−1.08 ± 6.54
	Training using single-source caption (en)	6.42 ± 4.84	5.14 ± 4.66	3.13 ± 4.89	0.40 ± 5.54	−2.91 ± 6.40
Intra-event-class dataset	Training using multiple-source caption (ja)	3.22 ± 2.78	2.49 ± 2.37	1.11 ± 2.31	−1.01 ± 3.08	−3.93 ± 4.34
	Training using multiple-source caption (en)	3.09 ± 2.50	2.38 ± 2.11	1.15 ± 2.09	−0.86 ± 2.90	−3.61 ± 4.22
	Training using single-source caption (ja)	4.36 ± 3.60	3.24 ± 3.18	1.36 ± 3.01	−1.33 ± 3.66	−4.69 ± 4.73
	Training using single-source caption (en)	4.47 ± 3.09	3.23 ± 2.41	1.29 ± 2.49	−1.48 ± 3.27	−4.87 ± 4.42

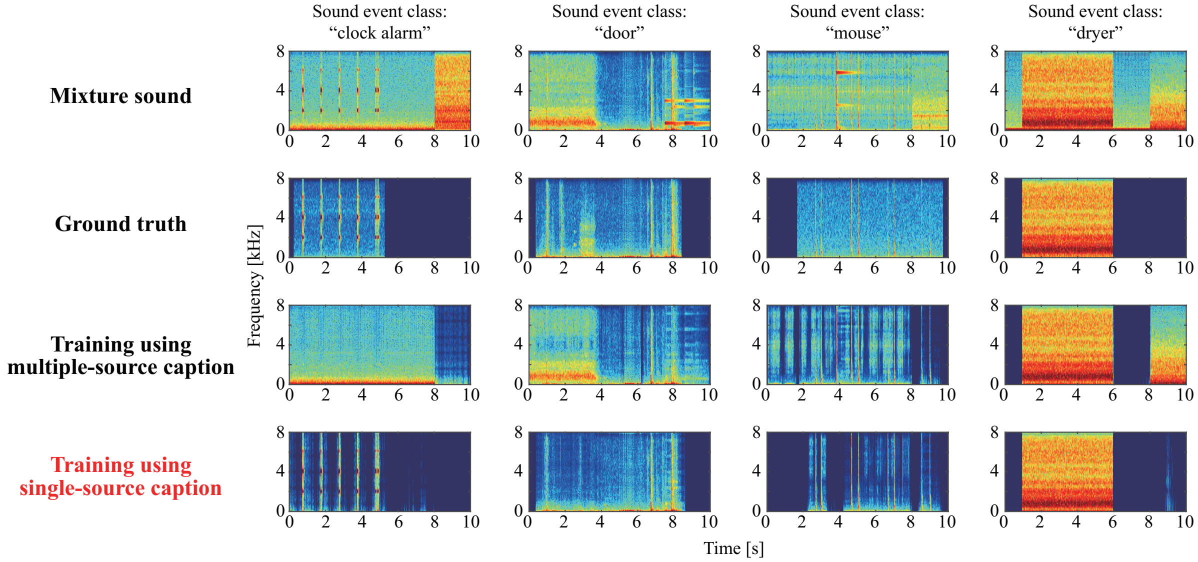


Figure 2: *Examples of environmental sound extraction using captions for the inter-event-class dataset. Mixture spectrogram (first row), ground truth spectrogram (second row), results of model training using multiple-source caption (third row), and results of model training using single-source caption (proposed) (fourth row).*

3.3. Training and evaluation setup

Table 5 lists the experimental conditions and parameters used for this experiment. In this experiment, we compared two types of model training methods as follows:

- **Training using multiple-source captions**
  This model training method uses the captions corresponding to the multiple-source sounds. This method is equivalent to the model training method using the conventional dataset.
- **Training using single-source captions**
  This model training method uses captions corresponding to single-source sounds.

In the evaluation, we evaluate the sound extraction performance using the evaluation dataset created in Sec. 3.2 for each model training method.

To evaluate each model-training method, we used signal-to-distortion ratio improvement (SDRi) [28] as an evaluation metric, which is defined as the difference between the SDR of the target sound to the mixture sound and that of the target sound to the extracted sound. We evaluated in terms of the SDRi on each evaluation dataset.

3.4. Experimental results

Table 6 lists the SDRi on each evaluation dataset. The training using single-source captions enables us to extract only the target sound from a mixture sound compared with the training using multiple-source captions. Moreover, from the results of the intra-event-class dataset, the caption for a single-source sound is effective to extract only the target sound, even when the interference sound from the same sound event class appears in the mixture sound.

Figure 2 shows the spectrogram of the extracted sounds using single-source captions. We used four audio samples in the inter-event-class dataset with 0 dB. We observed that the training using multiple-source captions left a significant amount of non-target sounds. The training using single-source captions extracted only the single-source target sound. On the other hand, the model trained using single-source captions could extract only the single-source target sound. Thus, the captions for a single-source source are effective in extracting only the single-source target sound.

4. Conclusion

We constructed a dataset with captions for a single-source sound that can be used in various tasks that use environmental sounds. We recorded a total of 1,044 single-source sounds and collected a total of 4,902 captions for recorded single-sound sources. The experimental results indicate that the use of sound captions assigned to a single-sound source is effective in extracting only the target sounds in mixture sounds. Verifying our dataset's effectiveness for other tasks involving environmental sounds is necessary.

5. References

[1] J. F. Gemmeke., D. P. W. Ellis., D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2017, pp. 776–780.

[2] 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.

[3] 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.

[4] 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.

[5] M. Delcroix, J. B. Vázquez, T. Ochiai, K. Kinoshita, Y. Ohishi, and S. Araki, "Soundbeam: Target sound extraction conditioned on sound-class labels and enrollment clues for increased performance and continuous learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 121–136, 2023.

[6] 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.

[7] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, and D. P. W. Ellis, "Improving universal sound separation using sound classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, pp. 96–100.

[8] 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.

[9] Y. Okamoto, S. Horiguchi, M. Yamamoto, K. Imoto, and Y. Kawaguchi, "Environmental sound extraction using onomatopoeic words," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2022, pp. 221–225.

[10] X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Separate what you describe: Language-queried audio source separation," in *Proc. INTERSPEECH*, 2022, pp. 1801–1805.

[11] K. Kilgour, B. Gfeller, Q. Huang, A. Jansen, S. Wisdom, and M. Tagliasacchi, "Text-driven separation of arbitrary sounds," in *Proc. INTERSPEECH*, 2022, pp. 5403–5407.

[12] H.-W. Dong and N. Takahashi and Y. Mitsufuji and J. McAuley and T. Berg-Kirkpatrick, "Clipsep: Learning text-queried sound separation with noisy unlabeled videos," in *Proc. International Conference on Learning Representation* (*ICLR*), 2023.

[13] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proc. the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 119–132.

[14] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an audio captioning dataset," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, pp. 736–740.

[15] M. Wu, H. Dinkel, and K. Yu, "Audio caption: Listen and tell," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2019, pp. 830–834.

[16] 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, pp. 925–929.

[17] Y. Okamoto, K. Imoto, S. Takamichi, R. Yamanishi, T. Fukumori, and Y. Yamashita, "Onoma-to-wave: Environmental sound synthesis from onomatopoeic words," *APSIPA Transactions on Signal and Information Processing*, vol. 11, e13, 2022.

[18] X. Liu, T. Iqbal, Z. Turab, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Conditional sound generation using neural discrete time-frequency representation learning," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing* (*MLSP*), 2021, pp. 1–6.

[19] P. Chen, Y. Zhang, M. Tan, H. Xiao, D. Huang, and C. Gan, "Generating visually aligned sound from videos," *IEEE Transactions on Image Processing*, vol. 29, pp. 8292–8302, 2020.

[20] 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, pp. 3550–3558.

[21] F. Kreuk and G. Synnaeve and A. Polyak and U. Singer and A. Défossez and J. Copet and D. Parikh and Y. Taigman and Y. Adi, "AudioGen: Textually guided audio generation," in *Proc. International Conference on Learning Representation* (*ICLR*), 2023.

[22] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[23] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *arXiv preprint arXiv:2207.09983*, 2022.

[24] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," *arXiv preprint arXiv:2301.12661*, 2023.

[25] 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.

[26] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1*, 2019, pp. 4171–4186.

[27] 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.

[28] 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.