ACOUSTIC EVENT SEARCH WITH AN ONOMATOPOEIC QUERY: MEASURING DISTANCE BETWEEN ONOMATOPOEIC WORDS AND SOUNDS

Shota Ikawa[1], *Kunio Kashino*[1,2]

[1] Graduate School of Information Science and Technology, University of Tokyo, Japan
[2] NTT Communication Science Laboratories, NTT Corporation, Japan

ABSTRACT

As a means of searching for desired audio signals stored in a database, we consider using a string of an onomatopoeic word, namely a word that imitates a sound, as a query, which allows the user to specify the desired sound by verbally mimicking the sound or typing the sound word, or the word containing sounds similar to the desired sound. However, it is generally difficult to realize such a system based on text similarities between the onomatopoeic query and the onomatopoeic tags associated with each section of the audio signals in the database. In this paper, we propose a novel audio signal search method that uses a latent variable space obtained through a learning process. By employing an encoder-decoder onomatopoeia generation model and an encoder model for the onomatopoeias, both audio signals and onomatopoeias are mapped within the space, allowing us to directly measure the distance between them. Subjective tests show that the search results obtained with the proposed method correspond to the onomatopoeic queries reasonably well, and the method has a generalization capability when searching. We also confirm that users preferred the audio signals obtained with this approach to those obtained with a text-based similarity search.

Index Terms— audio signal search, onomatopoeia, latent variable, encoder-decoder model

1. INTRODUCTION

Recently, a large amount of audio data is being accumulated in local storage or on the Internet, and the demand for an audio signal search technique has been increasing. Audio signal search methods can be divided into two types according to query types: search with an audio query and search with a text query.

For the former, searches based on audio feature matching is widely utilized [1]. However, except for the cases such as audio fingerprinting, there are generally many cases where the audio signal or feature is difficult to obtain to use as a query. For example, sound engineers who want to find specific sound effects in a sound database will not have the desired signal that can be used as a query. For the latter type of search, sound classification or description tags must be attached to acoustic signals in advance. For example, a video hosting service can use metadata, the anchor texts of incoming links, and comments as text information. However, it is widely known that automatic audio classification or description is not a simple task [2], and therefore, this approach sometimes requires a lot of human labor.

Against this background, we propose the use of onomatopoeias as audio search queries. The application we have in mind is a generic sound search system that allows users to find or locate their desired sounds. For example, it would be useful to be able to spot specific audio samples or events, such as birds' songs, machine failure sounds, or accident sounds, from among a vast amount of stored data.

Onomatopoeias are the words that imitate non-speech sounds within the pronunciation of a certain language system, and there are two modes: written and spoken. Onomatopoeias are widely seen in many languages, including English, Chinese, and Japanese, and they effectively support our daily communication. The use of onomatopoeias helps us to express acoustic information in a form that others can easily understand [3]. In previous studies, they have been effectively used for intuitive audio searches [4], and as a kind of classification tags for acoustic events [5, 6]. Up to now, most research using onomatopoeias for audio search has been text-based, which means it was based on the textual similarities between the onomatopoeic query and the onomatopoeic tags attached to the acoustic signals in advance [7]. However, as detailed in the following section, this approach poses several problems.

To solve these problems, here we propose a method that takes advantage of a latent space. The space is obtained through the learning process of an encoder-decoder model [8, 9] for onomatopoeia generation [10]. The space can be sufficient to allow it to be shared by onomatopoeic and audio signal encoders. This allows us to directly measure the distance between a written or spoken onomatopoeia and a section of an audio signal, which means that we can perform a similarity search for audio signals with an onomatopoeia query, without audio classification, description or transcription.

The rest of this paper is organized as follows. Section 2 discusses the problems of the existing text-based audio search methods. Section 3 introduces our method. Section 4 evaluates our proposed system. Section 5 concludes the paper.

2. PROBLEMS WITH TEXT-BASED AUDIO SEARCH

Previous work on audio signal search with an onomatopoeic query has usually been based on the similarity between the query text and the onomatopoeia tags associated with each audio signal in the database. In addition to the preprocessing, or human labor, needed to attach such tags in the database, this approach essentially poses the following problems.

First, many search results can give the exactly same similarity to a query. This is due to the fact that onomatopoeias are highly-compressed, coarsely-quantized representation of sounds. This makes it difficult to obtain an appropriately ordered result list. As the database grows in size, the usability can be seriously degraded.

Second, it is generally difficult to determine one correct onomatopoeic tag for an audio signal; that is, one audio signal can be described as different onomatopoeias, depending on the listeners.

This is due to the intrinsic ambiguity in onomatopoeic expressions [11]. For this reason, the quality and quantity of the onomatopoeic tags in the database greatly affect the accuracy, efficiency and usability of the search.

3. SEARCH BASED ON LATENT VARIABLES

3.1. Audio search problem definition

Let $\boldsymbol{z_x}$ be a latent variable derived from an audio signal $\boldsymbol{x}$, and $\boldsymbol{z_l}$ be an onomatopoeic latent variable derived from an onomatopoeia. Here, a latent variable is a fixed-dimensional vector. When $\boldsymbol{z_x}, \boldsymbol{z_l}$ are two points in the shared latent space $V \subset \mathbb{R}^n$, the distance between the audio signal and the onomatopoeia is defined as follows:

$$D(\boldsymbol{x}, \boldsymbol{l}) \equiv \|\boldsymbol{z_x} - \boldsymbol{z_l}\|. \tag{1}$$

$\|\cdot\|$ is norm on V. Here, audio search is defined as finding the nearest audio signals to a query based on the distance given in Eq. 1. Hereafter, we assume the query is given in the form of a written onomatopoeia, although the same framework can be applied to the case of spoken onomatopoeias.

3.2. Extracting latent variables

We employ an onomatopoeia generation model to calculate $\boldsymbol{z_x}$ from the corresponding audio signal. The model is based on the idea that an onomatopoeia phoneme string $\boldsymbol{l}$ is generated according to a conditional probability distribution $p(\boldsymbol{l}|\boldsymbol{z_x})$. That is, it generates the onomatopoeia string $\bar{\boldsymbol{l}}$, which has the highest probability given an audio signal.

$$\bar{\boldsymbol{l}} = \arg\max_{\boldsymbol{l}} p(\boldsymbol{l}|\boldsymbol{z_x}) \tag{2}$$

This estimation is decomposed into: (1) the estimation of a mapping $f : \boldsymbol{x} \rightarrow \boldsymbol{z_x}$, namely the extraction of a latent variable from an audio signal x, and (2) the generation of the most plausible onomatopoeia $\bar{\boldsymbol{l}}$ given the latent variable $\boldsymbol{z_x}$. The former step is used to obtain $\boldsymbol{z_x}$ from x.

The onomatopoeic latent variable $\boldsymbol{z_l}$ is extracted from an onomatopoeia l as follows. With the onomatopoeia generation model, the probability of $\boldsymbol{z_l}$ is given by the conditional probability density distribution $p(\boldsymbol{z}|\boldsymbol{l})$, which is the likelihood function of $\boldsymbol{l}$. Thus, we regard the conditional expectation of $\boldsymbol{z}$ as the onomatopoeic latent variable $\boldsymbol{z_l}$. That is, a mapping $g : \boldsymbol{l} \rightarrow \boldsymbol{z_l}$, namely the extraction of the latent variable from an onomatopoeia, is formulated as:

$$g(\boldsymbol{l}) = \int_V \boldsymbol{z} p(\boldsymbol{z}|\boldsymbol{l}) d\boldsymbol{z}. \tag{3}$$

3.3. Solution with neural networks

Ikawa *et al.* [10] proposed using an encoder-decoder model to obtain onomatopoeic representation from sounds. The encoder corresponds to the mapping f and the decoder corresponds to the estimation of $\bar{\boldsymbol{l}}$ from $\boldsymbol{z_x}$, and they are estimated simultaneously.

We used the encoder-decoder model shown in Figure 1. The audio latent variable $\boldsymbol{z_x}$ is calculated from acoustic features. Hereafter, we call this part an audio signal encoder. Then, the initial states of the decoder-LSTMs are calculated from $\boldsymbol{z_x}$. Here, the dimension of the latent space V is determined by the number of units of the corresponding layer of the neural network. Using tanh as the activation function, each element of $\boldsymbol{z}$ takes the value $[-1, 1]$.

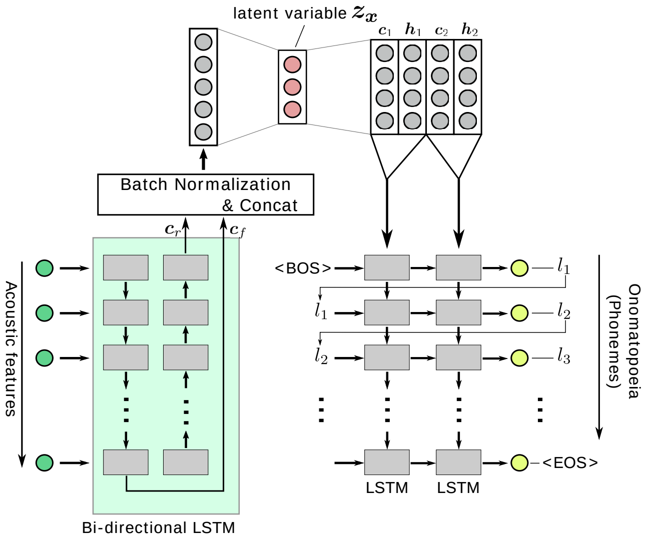

Figure 1: Block diagram of the audio signal encoder-decoder model.

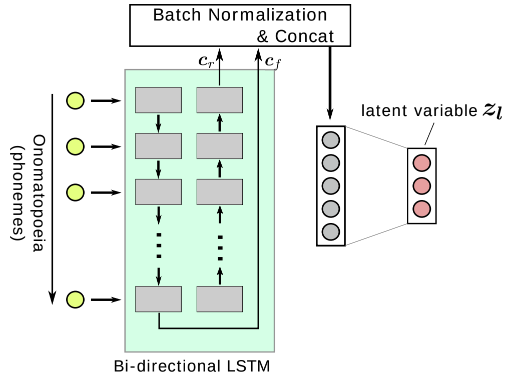

Figure 2: Block diagram of onomatopoeic encoder model.

The mapping g can also be obtained using a neural network (hereafter, onomatopoeic encoder). Figure 2 shows the structure of the onomatopoeic encoder. The estimated mapping $\hat{g} = g_{\hat{\boldsymbol{\theta}}}$ is acquired based on the learned parameters of the onomatopoeic encoder $\hat{\boldsymbol{\theta}}$. With the estimated audio signal encoding mapping $\hat{f}$, the loss function used to train the onomatopoeic encoder is written as:

$$\mathcal{L}(\boldsymbol{\theta}) = \|g_{\boldsymbol{\theta}}(\boldsymbol{l}) - \hat{f}(\boldsymbol{x})\|, \tag{4}$$

where the definition of norm is the same as in Eq. (1).

3.4. Audio signal search

Using the estimated mappings $\hat{f}, \hat{g}$, the audio signal search is realized by measuring the distances between the onomatopoeic query and each audio signal in the database:

$$\hat{D}(\boldsymbol{x}, \boldsymbol{l}) = \|\hat{f}(\boldsymbol{x}) - \hat{g}(\boldsymbol{l})\|. \tag{5}$$

The neural networks are trained with a set of audio signals associated with onomatopoeic tags. Unlike the existing text-based

Table 1: Experimental conditions

LSTM cells	512
Batchsize	256
Output phoneme labels	32
Optimizer	ADAM [14]
MFCC dimension	20
FFT window (MFCC)	2048 samples
FFT shift (MFCC)	512 samples

search methods described in section 2, once the test set is given, our method does not need an onomatopoeic tag for any of the audio signals in the database.

4. EXPERIMENTS

We evaluate the proposed method from two standpoints: the appropriateness of the search results and by comparing it with a text-based search. In both cases, the task is to find the nearest neighbor signal. For an audio signal database $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ and an onomatopoeic query $\boldsymbol{l}$, the nearest neighbor signal $\bar{\boldsymbol{x}}(\boldsymbol{l})$ is represented as:

$$\bar{\boldsymbol{x}}(\boldsymbol{l}) = \operatorname*{arg\,min}_{\boldsymbol{x}_i \in X} D(\boldsymbol{x}_i, \boldsymbol{l}). \tag{6}$$

4.1. Dataset

We used a subset of the audio signals contained in the Real World Computing Partnership (RWCP) sound scene database [12] to train the neural networks. The database includes various sound samples recorded without background noise in an unechoic environment and digitized at 48 kHz, with linear PCM of 16 bit accuracy.

For the training, we chose 709 signals, including those made by bells, coins, and hitting wood with a stick. The number of the class labels (bell, coin, etc.) was 81, and 7 to 10 signals were sampled for each class.

To build the training set, onomatopoeic tags were collected from human listeners. Considering the ambiguity of onomatopoeia, multiple onomatopoeic tags were attached to each audio signal. To accomplish this, 73 Japanese speakers were asked to produce three onomatopoeias for each sound using katakana, which is a Japanese syllabary. Each katakana answer was converted to a string based on the International Phonetic Alphabet (IPA) [13] and used as an onomatopoeic tag. In Japanese, onomatopoeias are usually written in katakana, and it is straightforward to convert katakana to IPA, and vice versa. We associated 12 onomatopoeias for each audio signal in the dataset.

Note that we used the IPA symbols as a simple universal representation of pronunciation in the experiments, but any phonogram sequences, or texts, can be used in our framework.

4.2. Learning of the encoder-decoder onomatopoeia generation model

Table 1 lists the experimental conditions. For simplicity, we used a series of mel-frequency cepstral coefficients (MFCC) as the input. As output phonemes, we used 29 kinds of symbols that consist of the standard IPA phonetic symbols and Japanese-specific ones: "ɴ" for moraic nasal, "H" for the second mora of a long vowel and "Q" for a moraic silence when emphatic. In addition, we used three special symbols: "BOS (beginning of the sequence)," "EOS (end of the sequence)," and "UNK (unknown)."

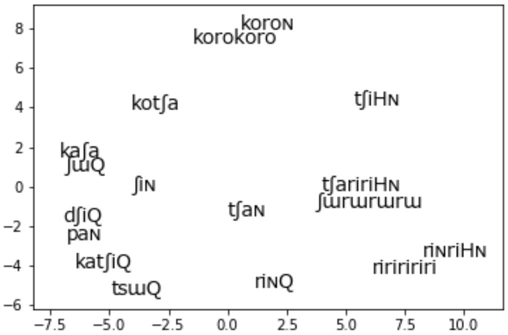

Figure 3: Onomatopoeic latent variables whose dimensions are reduced from 128 to 2 by using PCA. It is shown that onomatopoeias with similar characteristics (e.g. "koroɴ" and "korokoro") are closely located.

From the preliminary experiments, we chose 128 as the number of latent variable dimensions. After 34 epochs of learning, the audio signal encoder-decoder model achieved a 9.9% word error rate and a 4.0% mean phoneme error rate for an onomatopoeia generation task [10], with a test set consisting of 101 audio signals, which were again sampled from the RWCP dataset excluding the ones used for the network training.

4.3. Learning of the onomatopoeic encoder

The same training data were used for training the onomatopoeic encoder as in the previous section. L1-norm was employed as the loss function (Eq. (4)). Figure 3 shows an example of the resulting distribution of onomatopoeic latent variables. It is observed that the onomatopoeias with similar characteristics are localized to each other in the latent space.

4.4. Experimental setups and results

Experiment 1: Suitability of signals found for queries

This experiment was designed to confirm whether the found signals correctly corresponded to the onomatopoeic queries.

The subjects were presented with an onomatopoeia in katakana on a display, which was a query, and then with an audio signal, which was a result of the nearest neighbor search using the proposed method. They were then asked to choose one of five options: "very suitable," "relatively suitable," "neutral," "relatively unsuitable," and "very unsuitable." We performed the experiment using two different audio databases:

- A database consisting of the above-mentioned "training" set sampled from the RWCP database (hereafter, the RWCP test set)
- A database consisting of the sounds sampled from another dataset (hereafter, the external test set)

The former was used to verify the basic behavior, and the latter was used to evaluate the generalization performance of the proposed method.

For the external test set, we used part of Free Sound Dataset Kaggle 2018 [15], which is a subset of FSD [16] and is used for

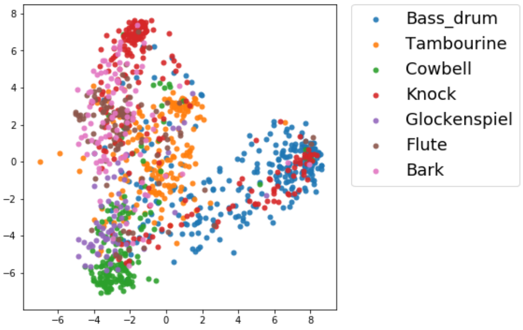

Figure 4: Distributions of audio latent variables for the external test set. Dimensions of each latent variable are reduced from 128 to 2 by using PCA. It is observed that the samples that belong to the same class tend to be localized in this space.

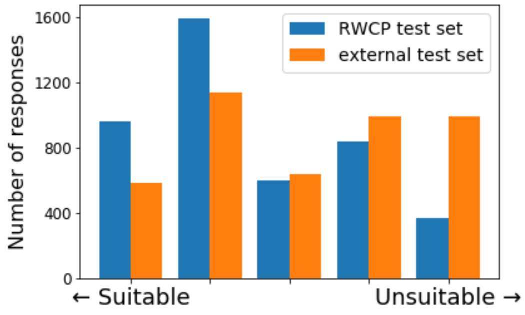

Figure 5: Suitability of the audio signals presented by the search

the general-purpose audio tagging task in the DCASE 2018 CHALLENGE. There were 11,719 audio signals. The MFCC sequence for each audio signal was calculated according to Table 1 after converting the sampling frequency from 44.1 kHz to 48 kHz by using FFmpeg [17]. Figure 4 shows the distributions of the audio latent variables of some of the external test set obtained by the trained model visualized by PCA.

We chose 217 Japanese onomatopoeias as the queries for each test set. There were 20 subjects, and the total number of responses for each test set was 4,340.

Figure 5 shows the result. The most frequent response for both test sets was "relatively suitable". For the RWCP test set, 58.7% of the responses were "suitable," showing that the proposed method worked effectively. For the external test set, the "suitable" responses amounted to 39.7%, which is fewer than the RWCP case. This is because the number and variations of audio signals included in the external test set is much greater than that of the training set. However, this still means that the proposed model has a generalization ability even for the external test set, because if the audio samples were randomly presented in this test, most responses must have been "very (or relatively) unsuitable".

Experiment 2: Comparison of the proposed method and the text-based method

We used the audio database that consisted of the same audio samples as those used in the network training in order to evaluate whether the search results obtained with the proposed method were preferable to those obtained with the text-based method. The subjects were presented with one onomatopoeia on a display and two audio signals, "A" and "B." They were then asked to choose one of five options: "A is much better," "A is relatively better," "no difference," "B is relatively better," and "B is much better." Either "A" or "B" (randomly selected) was the search result obtained with the proposed method and the other was the one obtained with the text-based method.

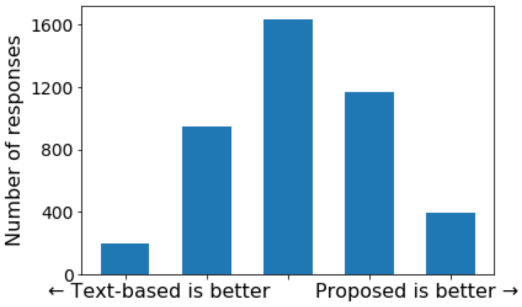

Figure 6: Comparison of the proposed method and the text-based method

The text-based method in this experiment was based on the similarity measured by the edit (Levenshtein) distance between the IPA strings. In the audio database, multiple audio signals can correspond to the same onomatopoeic tag, yielding multiple search results for one query with the same similarity. In such cases, one signal was randomly chosen as the result.

As in Experiment 1, 217 onomatopoeias were used as the queries, and 20 subjects undertook the evaluation. The total number of responses was 4,340.

Figure 6 shows the result. It is shown that the proposed method is preferable to the text-based method, since the distribution is clearly biased to the right from the center. For a quantitative analysis, we assign scores of 2, 1, 0, -1, -2, according to the five kinds of responses, so that the score become larger when the proposed method receives a higher evaluation. The mean of the score appears to be 0.145, and from the t test, it is not 0 at the 1% significance level. This means that the proposed method produced significantly better results than the text-based method.

5. CONCLUSION

This paper proposed a novel method for finding audio signals with an onomatopoeic query. With our method, the distance between an audio signal and an onomatopoeic symbol sequence is directly measured in the latent space. We showed the effectiveness of the proposed method by performing subjective experiments. This paper focused on the use of written onomatopoeias, but we expect that it is straightforward to train the network so that it accepts spoken onomatopoeias as queries. Our future work will also include tests with languages other than Japanese, and a usability study with practical senarios.

6. ACKNOWLEDGMENTS

We thank Prof. Hiroshi Saruwatari and Dr. Shinnosuke Takamichi for their valuable comments and support.

7. REFERENCES

[1] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE MultiMedia*, vol. 3, no. 3, pp. 27–36, Fall 1996.

[2] 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 ICASSP 2017*, New Orleans, LA, 2017.

[3] F. Wang, H. Nagano, K. Kashino, and T. Igarashi, "Visualizing video sounds with sound word animation to enrich user experience," *IEEE Transactions on Multimedia*, vol. 19, no. 2, pp. 418 – 429, 2017.

[4] T. A. Sanae Wake, "Sound retrieval with intuitive verbal descriptions," *IEICE Trans. Inf. and Syst.*, vol. E84-D, no. 11, pp. 1568 – 1576, 2001.

[5] S. Sundaram and S. S. Narayanan, "Classification of sound clips by two schemes: using onomatopoeia and semantic labels," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Hannover, Germany, jun 2008, pp. 1341–1344.

[6] K. Hayashida, Y. Mizoguchi, J. Ogawa, M. Morise, T. Nishiura, and Y. Yamashita, "The acoustic sound field dictation with hidden markov model based on an onomatopeia," vol. 5, 01 2010.

[7] K. Okamoto, R. Yamanishi, and M. Matsushita, "Sound-effects exploratory retrieval system based on various aspects," *IEEJ Transactions on Electronics, Information and Systems*, vol. 136, no. 12, pp. 1712–1720, 2016.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3104–3112. [Online]. Available: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[9] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *CoRR*, vol. abs/1406.1078, 2014. [Online]. Available: http://arxiv.org/abs/1406.1078

[10] S. Ikawa and K. Kashino, "Generating sound words from audio signals of acoustic events with sequence-to-sequence model," in *Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 346 – 350.

[11] K. Ishiguro, Y. Tsubota, and H. Okuno, "Automatic transformation of environmental sounds into sound-imitation words based on japanese syllable structure," in *Proc. EUROSPEECH*, 2003, pp. 3185–3188.

[12] S. Nakamura, K. Hiyane, F. Asano, T. Yamada, and T. Endo, "Data collection in real acoustical environments for sound scene understanding and hands-free speech recognition," in *Proc. EUROSPEECH*, Sep. 1999, pp. 2255–2258.

[13] I. P. Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representation (ICLR)*, 2015.

[15] https://www.kaggle.com/c/freesound-audio-tagging/data.

[16] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, Suzhou, China, 2017, pp. 486–493.

[17] https://www.ffmpeg.org/.