Human Interaction and Emerging Technologies (IHIET 2023), Vol. 111, 2023, 270–277
https://doi.org/10.54941/ahfe1004033

Automatic Text-to-Sound Generation by Doc2Vec

Kakeru Iwamoto, Hironori Uchida, Yujie Li, and Yoshihisa Nakatoh

Kyushu Institute of Technology, 1–1 Sensui-cho, Tobata-ku, Kitakyushu-shi, Fukuoka, Japan

ABSTRACT

Nowadays, the market size for games and video viewing services has been expanding. The demand for sound effects to produce content using these services is also rising. However, sound-effect production often requires expensive software or hardware, exceptional equipment use, and experience. Therefore, this study aims to reduce the cost of sound effects production and generate imaginable sound effects outputs. It examines a method for automatically generating sound effects based on text input using Doc2Vec. The Natural Language Processing Model (NLP) calculates the similarity between the input text and the labels in the dataset. The Natural Language Processing Model (NLP) is created by Doc2Vec pre-loading sound-related language expressions (labels). The model is pre-trained using labels from VGG-Sound data. The calculated highly similar sound are downloaded from VGG-Sound, a specified number of sound datasets. The data downloaded from highly similar data is synthesized in similarity order, and the audio is output. Furthermore, this paper verifies the correlation between the sentences used in the proposed method and the generated sound effects. The paper conducted an experiment presenting the generated sound and the sentences used to create them and had the participants rate them on a 5-point scale. The sentences used for a generation are those that live in the dataset, those that lived in a small number in the data set, those that did not exist in the data collection, those to which information such as location or scene has been added, and those that contain multiple events. The results show that the more sound in the dataset, the higher the rating, and that sentences with numerous information or numerous events produce lower ratings.

Keywords: Doc2vec, NLP, Text-to-sound, Automatic-generation

INTRODUCTION

Japan's Ministry of Economy, Trade and Industry (METI, 2020) announced in FY2020 the Size of the Global Content Markets, as shown in Figure 1. According to this data, the global market size is increasing and is expected to exceed US$1.2995 trillion by 2023. Additionally, Figures 2 show the "Percentage of the Global Content Markets by Content Sector" published by the Ministry of Economy, Trade and Industry (2020). The percentage of the global content market for video, games, and music, which require sound effects, increased from 2014 to 2018, with a particularly high rate of games showing an increasing trend. It also prove that while there will be no significant percentage change in music from 2018 to 2023, there is expected to be a large increase in the video and game sectors. Additionally, in the Japanese content market, there is no significant change in the percentages of video and music but a marked increase in the percentage of the game field. Based on the above, demand for sound effects in video, games and music is expected to increase.

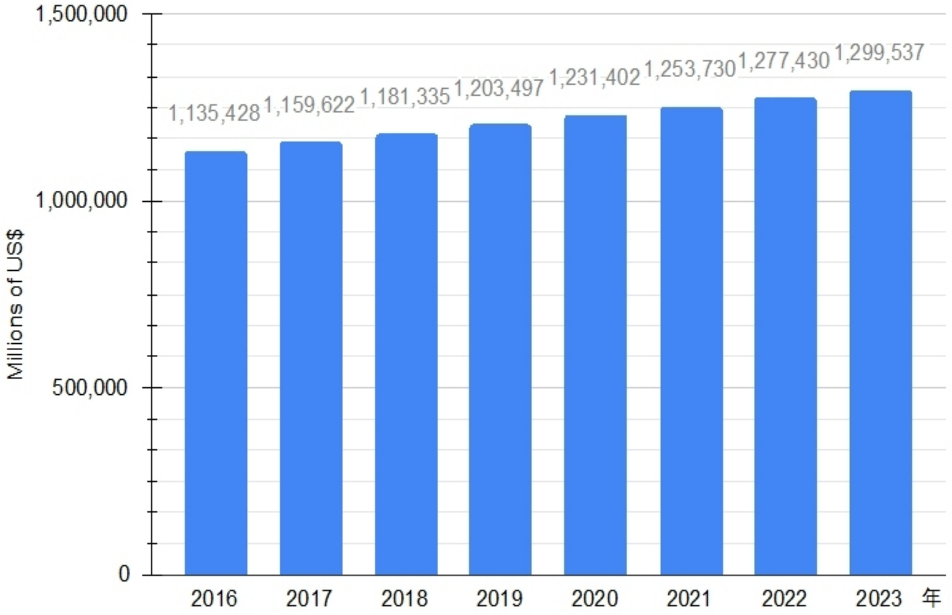

Figure 1: Size of the global content markets (METI, 2020).

There are two main types of sound effect production methods. The first is the method of recording real-life objects. This is one of the simplest methods of creating sound effects in which the sound source is registered with a recording device, and the sound effect is converted into waveform data to be used as a sound effect. The sound effect is not used as it is, but the recorded waveform data is often processed by editing software to change the pitch or synthesize sound. Examples include human footsteps, applause, doors opening and closing, and animal noises. In addition, sounds may be artificially recorded by humans as if they were different sounds. This method is often used in movies, where plastic bags are deformed to sound like firewood or eggs are cracked to create the sound of eating snacks. The second method is to create an imaginary sound that does not exist in reality. An example of this method is to prepare an electronic sound that can be used as a basis and then process it using a synthesizer. Examples include particular sounds such as a beam being fired, magic being cast, or a button being pressed. However, all the above methods require high-performance recording equipment, editing software, synthesizers, and the knowledge and experience to handle them. Therefore, creating original sound effects for independent games and video content produced by individuals or small groups is extremely expensive.

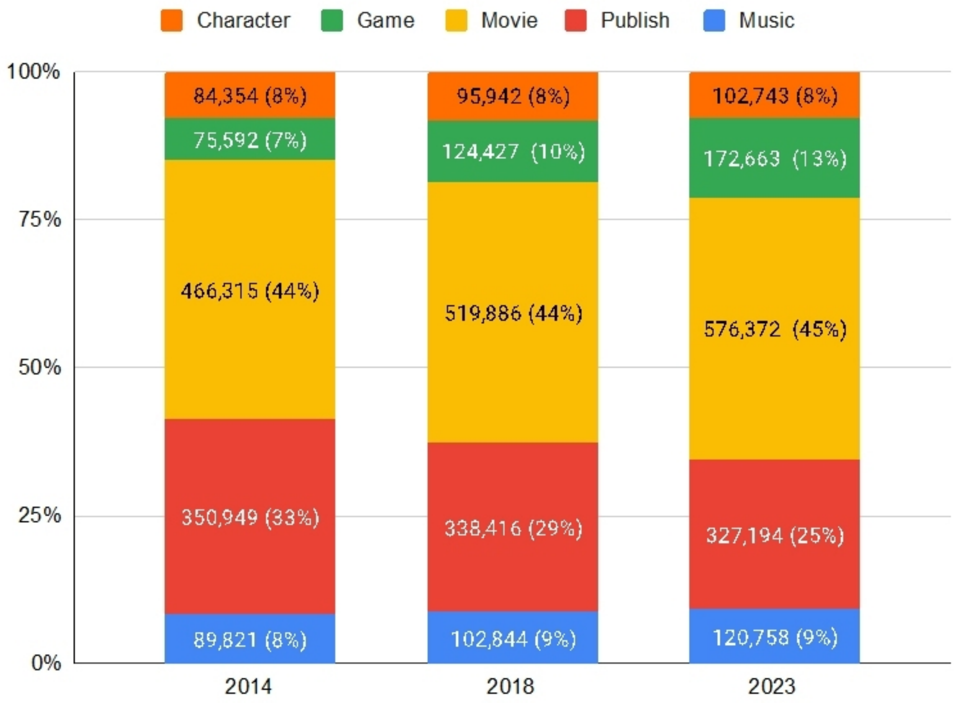

Figure 2: Percentage of the global content markets by content sector (METI, 2020).

Therefore, this study aims to reduce the cost of sound effect creation. A system that automatically generates sound effects from input text was studied and developed as a first step.

RELATED RESEARCH

A method on automatic sound effect generation similar to this research is Diffsound (Dongchao *et al.*) This is an investigation of sound generation from text input. At the inquiry stage, the system consisted of a text encoder, a Virtual Quantified Variational Autoencoder (VQ-VAE), a decoder, and a vocoder. The flow of this system is as follows. First, the text encoder extracts text features. Next, the extracted features are converted into a mel-spectrogram by VQ-VAE. Even more, the extracted features are converted into a mel-spectrogram by VQ-VAE. Lastly, a vocoder generates waveforms from the mel-spectrogram to produce sound. Experimentation with the above system revealed that the decoder's performance significantly impacts the sound produced, so this study focused on the design of a sound decoder. Initially, we began by using an auto-regressive decoder (AR decoder) as our decoder. However, since the AR decoder predicts input values one element at a time, as shown in Figure 3, it was found that the AR decoder was prone to bias and error accumulation and that the time to generate the decoder increased in proportion to the length of the speech. This study proposed Diffsound, a non-autoregressive decoder based on a discrete-diffusion model, to solve these problems. The configuration of the system containing Diffsound is shown in Figure 4. Diffsound predicts all input elements at once, as shown in Figure 5, and then generates the predictions of each component in a single step. Since the accuracy of the factors predicted in the next stage is improved, the overall accuracy could be raised by repeating this process several times. In addition, an evaluation of the system shown in Figure 4 revealed that it produced a Mean Opinion Score (MOS) rating of 3.65 generating speech making it five times faster at than only an AR decoder was used. In addition, it was found that sentences with a single event were easier to develop vocabulary than sentences with multiple possibilities.

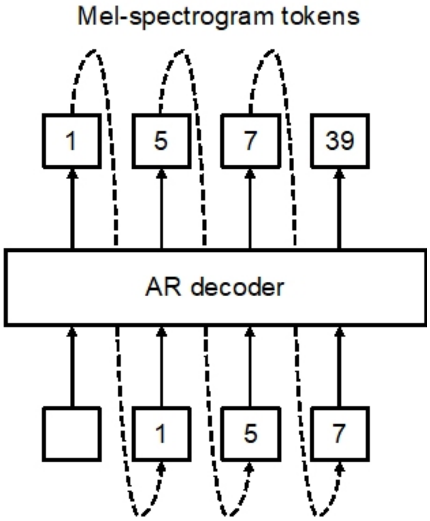

Figure 3: Example of spectrogram generation by an AR decoder (Dongchao *et al.*).

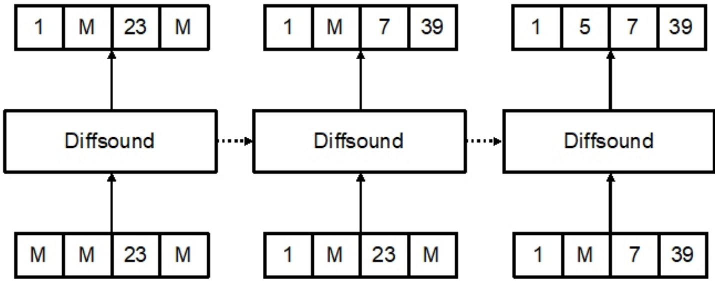

Figure 4: Example of spectrogram generation by Diffsound (Dongchao *et al.*).

PROPOSED METHOD

The flowchart of the proposed method is shown in Figure 6. First, a natural language processing model is used to calculate the similarity of the input sentences. The natural language processing model is created by loading the language expressions (labels) associated with the sound effects on the data set

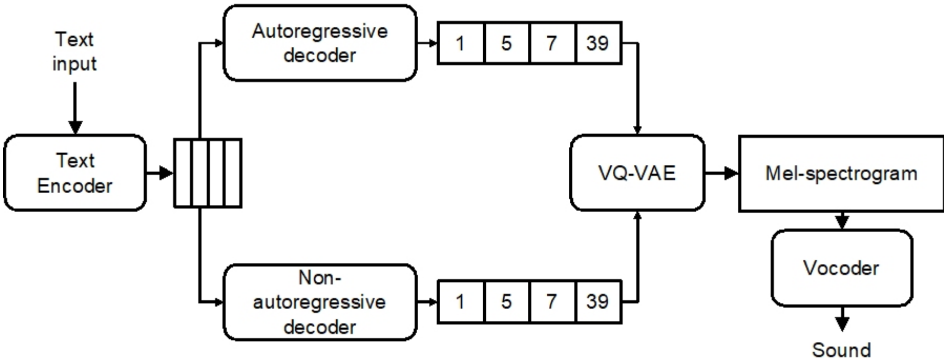

Figure 5: Configuration of the system including Diffsound (Dongchao *et al.*).

in advance by Doc2Vec. Next, a specified number of sounds with high similarity calculated are downloaded from the URL listed on the VGG-Sound data set of sound (Honglie *et al.*). Finally, the downloaded data are synthesized from those with the highest similarity, and the audio is output.

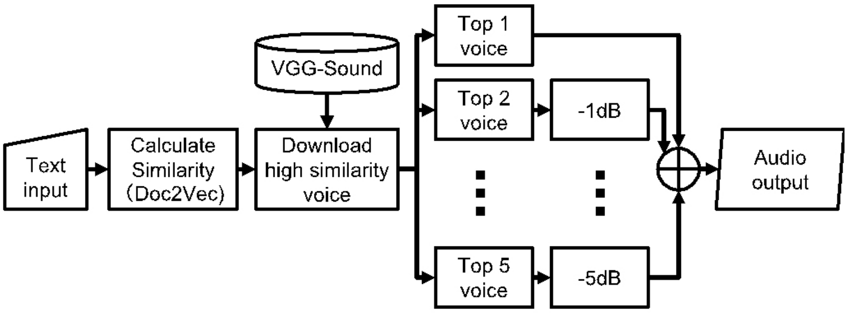

Figure 6: Flowchart of the proposed method.

Doc2Vec is a technology that can convert documents into vectors of a certain length and is an extension for documents of Word2Vec, which can acquire distributed representations of words. The advantages of Doc2Vec are that no teacher data is required for training and that semantic expressions are included in the vector. In addition, Doc2Vec has two methods for creating models: dmpv (distributed memory) and DBOW (distributed bag-of-words). As shown in Figure 7, Dmpv is trained in three layers using one-shot vector representations and document IDs as input. This one-shot vector representation is a vector representation of words where the number of dimensions N is the number of all words in the vocabulary. Only one dimension of a comment is set to 1, while all other sizes are set to 0. Since Doc2Vec requires a document ID in addition to the one-hot vector representation, a one-hot vector representation of the document itself is generated and used as an input value. Input values are input as shown in Figure 7, and by passing them through the document vector matrix D in the middle layer, the predicted words are output for learning. This method is characterized by the fact that the document ID preserves the context of the entire document. Learning here and optimizing the numerical values of the document vector matrix D make it possible to vectorize documents. In this study, dmpv was used to determine the accuracy of the similarity, and the parameters were set, as shown in Table 1 (Jey *et al.*).

Table 1. Parameters.

Vector Size	Window Size	Min Count	Sub-Sampling	Negative Sample	Epoch
300	5	1	10^{-6}	6	100

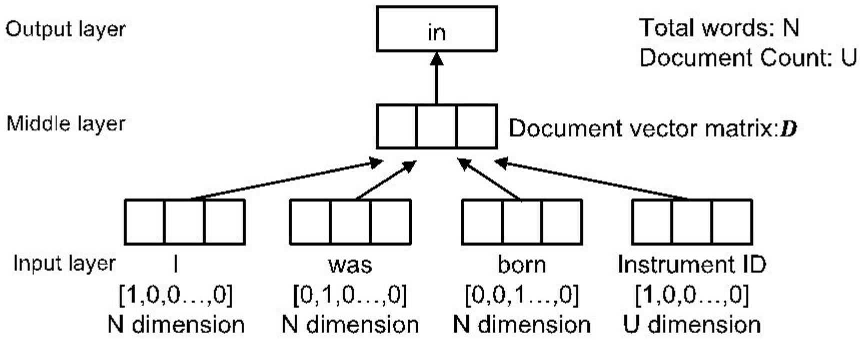

Figure 7: Overview diagram of dmpv.

This study uses VGG-Sound as the dataset (Honglie *et al.*). VGG-Sound is an audio/video dataset consisting of short clips extracted from videos uploaded to YouTube. This dataset is extensive, with over 310 different scenes labeled and over 200,000 audio clips, each 10 seconds long, totaling over 550 hours of audio clips. The dataset is also provided as a text file in CSV format and is defined as YouTube ID, start seconds, label, train/test split, from left to right

EXPERIMENTAL METHOD

To verify the degree of between the input sentences and the sound effects generated by the proposed method, 19 subjects (sixteen males and three females) were presented with the input sentences and asked to listen to the sound-caused impacts afterward and to rate the degree of coincidence between the sentences and the sound effects on a 5-point scale. The subjects were asked to rate the sound effects on a 5-point scale, with 5 as "very much in line with the text," 4 as "in line with the text," 3 as "neither in line with the text," 2 as "not in line with the text," and 1 as "not in line with the text at all." The sound were generated from five types of sentences (sentences that were present in the label of the data set, penalties that were present in the dataset, sentences that were not present in the dataset, sentences with additional information such as location or scene, and sentences that contained multiple events).

EXPERIMENTAL RESULTS

Figure 8 shows the average value of the 5-point scale for each type of sound presented. From the previous results and this Figure 4.6, it can be considered that the sound generated have a sufficient congruency if they are sentences that exist in the data set. The low rating for the "playing castanets" in Figure 8 may be due to the dataset's low relationship between the sound and labels.

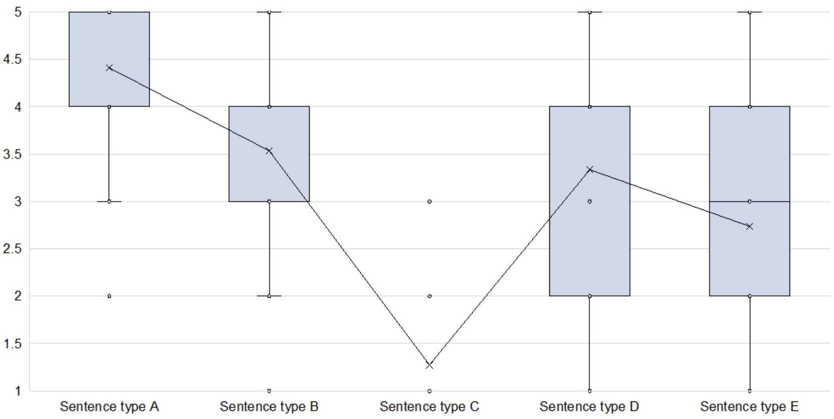

Figure 8: Experimental result.

CONCLUSION

This study aimed to reduce the cost of sound effect production, which is expected to be in increasing demand worldwide. As a first step, this studied and constructed a method to automatically generate sound effects from input sentences. To evaluate the way, 19 subjects were asked to listen to sounds generated from different types of sentences (sentences that exist in the label of the data set, sentences that exist in the data set in small numbers, sentences that do not exist in the dataset, sentences with modifying information such as location and scene, and sentences that contain multiple events) The participants were asked to rate the degree of unity between the sentences and the presented sound on a 5-point scale.

As shown in Figure 4.6, the average value of the 5-point evaluation of the generated presentation sound was 4.412 for "many sentences in the data set," 3.537 for "few sentences in the data set," and 1.274 for "no sentences in the data set. The score was 3.337 for sentences with modifiers added to the sentences in the dataset and 2.737 for sentences containing multiple events. From these results, it can be said that generating a sound that matches the sentence is possible if a sentence exists in the dataset. Furthermore, when information such as location and scene was added to the sentences, the average value of the 5-point evaluation exceeded 3, indicating that the sound was produced to some extent by the sentences but not be suitable for sentences containing multiple events.

Based on the experiment results, two future tasks and improvement measures are listed, respectively. The first is the generation of sound with sentences that do not exist in the dataset. As shown in Figure 4.6, the sentences generated by sentences that do not exist in the dataset are rated low. To improve this situation, it may be possible to increase the number of datasets to accommodate a wide variety of sentences or to generate sound effects even from sentences that do not exist in the dataset by performing machine learning on the input sentences and the generated sound effects. The second is the generation of sound for sentences containing multiple events. As shown in Figure 4.6 under "Multiple Events," the sentences generated by sentences containing various events are not highly rated. To improve this situation, it would be possible to create appropriate sounds by inputting sentences by natural language processing into events and generating sound for each decomposed sentence.

ACKNOWLEDGMENT

I would like to thank my deepest gratitude to Professor Yoshihisa Nakatoh for her guidance and encouragement in carrying out this research.I would also like to thank my sincere gratitude to Assistant Professor Yujie Li and Mr. Hiroki Uchida for their enthusiastic cooperation and guidance.

REFERENCES

Dongchao, Y., Jianwei, Y., Helin, W., Wen, W., Chao, W., Yuexian, Z and Dong, Y. (2022), "Diffsound: Discrete diffusion model for text-to-sound generation", CoRR, abs/2207.09983.

Honglie, C., Weidi, X., Andrea, V. and Andrew, Z. "VGGSOUND: A LARGE-SCALE AUDIO-VISUAL DATASET", ICASSP, (2020).

Jey, H. L. and Timothy, B. (2016), "An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation", *proceedings of the 1st Workshop on Representation Learning for NLP*, pp. 78–86,

METI (2020), "Kontentu no Sekai Sizyou . Nihon Sizyou no Gaikan (Overview of the global and Japanese markets for content)", available at: https://www.meti.go.jp/policy/mono_info_service/contents/downloadfiles/report/202002_contentsmarket.pdf (accessed 27 April 2023).