Proposal and experimental results of a method for removing mixed voice using U-Net

Jian LIN
Dept. of Information and Computer Sciences
Kanagawa Institute of Technology
Atsugi, Japan
s1921063@cco.kanagawa-it.ac.jp

Shota SANO
Graduate School of Engineering
Kanagawa Institute of Technology
Atsugi, Japan
s2185008@cco.kanagawa-it.ac.jp

Taiga WATANABE
Dept. of Information and Computer Sciences
Kanagawa Institute of Technology
Atsugi, Japan
s1921077@cco.kanagawa-it.ac.jp

Yuusuke KAWAKITA
Dept. of Information and Computer Sciences
Kanagawa Institute of Technology
Atsugi, Japan
kwkt@ic.kanagawa-it.ac.jp

Tsuyoshi MIYAZAKI
Dept. of Information and Computer Sciences
Kanagawa Institute of Technology
Atsugi, Japan
miyazaki@ic.kanagawa-it.ac.jp

Hiroshi TANAKA
Dept. of Information and Computer Sciences
Kanagawa Institute of Technology
Atsugi, Japan
h_tanaka@ic.kanagawa-it.ac.jp

***Abstract*— The mixing of other people's voices in daily conversations and remote meetings causes a large obstacle to the voice that you originally want to hear, regardless of the volume. In this paper, the voice of another specific speaker added to the voice of the speaker is regarded as noise, and a method of removing the noise and the experimental result are shown. We created a learning model for removing the mixed noise using U-Net and the data obtained by converting the voice data before and after the noise was mixed into an image by STFT. Then, the effect was confirmed by comparing the image data, and voice data obtained by the reverse STFT with the ones before the removal.**

Keywords*— *Voice, Noise removal, Spectrogram, U-Net, Normalized squared difference

I. INTRODUCTION

Due to the COVID-19 pandemic, remote work and remote meetings have been becoming more common. There are many life sounds we do not want to be transmitted, such as the voices of people at home and the sounds of pets. Moreover, it seems that there are many situations in which the voices of other people nearby are worrisome in daily life.

For noise removal, many methods such as spectral subtraction and Wiener filter have been proposed [1] [2]. Regarding the removal of the voice of another person added to the human voice, it is considered that these methods do not always provide sufficient performance, because the characteristics of the human voices are similar. A method using sound source separation [3] such as the MUSIC method by arranging multiple microphone sensors at appropriate positions is also adaptable, but the plural sensors are required and their positions are restricted to keep performance.

The authors created a noise removal model using STFT (Short Time Fourier Transformation) and U-Net, limiting it to train running noise. This noise was superimposed on human voice data and created model was applied to remove the train noise. It was confirmed that good noise removal characteristics were obtained[4]. In this paper, we experimentally try to apply this method using the voice of another person as noise and verify its performance. By limiting the noise to be removed, it is possible to reduce the load of collecting training data and creating a learned model. It is also considered that the requirements for operating requirements for processing devices are also relaxed.

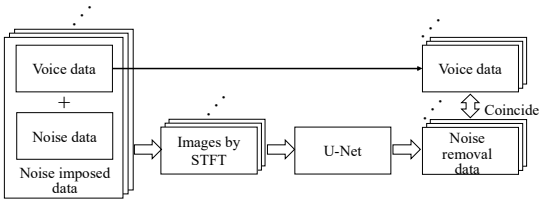

Fig.1 Learning model creation method

In this paper, we first show how to create a learning model for noise removal. Then, as an evaluation of noise removal using the learning model, the degree of difference between the image data before and after noise mixing and the image after noise removal is shown. The effect is confirmed by actually listening to the audio data restored by the inverse STFT conversion. By comparing them, it is shown that the proposed method is effective.

II. NOISE REMOVAL METHOD

As a model for noise removal, U-Net, which is a deep learning network model whose input data is the image, was used. There are already many research practices, and the authors have applied them to the noise reduction of trains while running, and have obtained good results. Fig.1 shows how to create a learning model by using U-Net. When U-Net inputs data with superimposed noise, it creates a learning model so that its output reproduces the image data before the superimposed noise.

This time, we will create a learning model that realizes the function of removing the voice of a specific speaker. By limiting the voice to be removed, it is expected to ensure the performance under the condition of low SN ratio.

At first, time-series voice data is converted into images by STFT. These image data were used as learning data. Table I summarizes various parameters when creating a learning model. As the voice data for learning, 20 men from the proven database [5], 10 to 15 minutes each, were used. As the voice to be removed, the voice of one of the authors when reading a book was recorded. The recording conditions are sampling 16kHz and monaural recording. It is assumed to remove the voices of other people nearby in the office and the voices of family members when attending remote meetings while working from home.

Table I Parameters in creating learned model

	Parameters	Values
Voice data	Number of data	Male 20 persons
	Voice length	10 – 15 minutes
	Sampling frequency	16kHz
	Noise (imposed sound)	Voice of author
	Voice length	3 minutes
	SN ratio	0/5dB
Spectrogram	Window function	hann
	Window width	1024
	Window shift width	128
	Image size	256 * 256
	Sliding ratio of image cut	0.8
Data set	Data for learning	7073 images
	Data for validation	1769 images

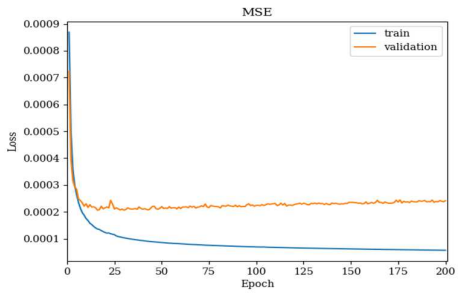

Fig.2 Convergence in learning process

As for the parameters, as a preliminary study, we basically compared the results of multiple parameters and selected the one that seems to be the best. Fig.2 shows the state of convergence of learning. The execution environment this time was OS: Windows10, CPU: Intel (R) Core (TM) i7-8700K CPU @ 3.70GHz 24GB, GPU: NVIDIA GeForce GTX 1070Ti 8GB. The learning in Fig. 2 took about 20 hours as learning time.

III. Noise Removal Results and Their Evaluation

The noise removal experiment was performed using the voice data of 1 man who is different from the voice data of 20 men used in the creation of the learning model, and the voice data of the same person but different voice data as noise. Fig.3 shows a comparison of removal performance by the created learning model. We compared the images and sounds before and after adding the noise voice and after removing the noise.

For learning models created with data of different SN ratios, noise reduction performance was evaluated using the difference between the spectrogram image after noise removal and the data before noise mixing. The training model A was created with SN ratio of 0 dB, and the training model B was with SN ratio of 5 dB. Fig.4 shows the performance of both Model A and Model B. Here, performance was obtained from the normalized square difference between the two images.

As the test data, the voice data of the speaker not used for learning was used. As the noise voice, the voices of different recording sections of the same speaker used as noise voice were used. The SN ratio when mixing noise is 0, 5, 10, 15 dB respectively. As shown in the graph, the noise removal effect was confirmed at any SN ratio except 15dB by numerically.

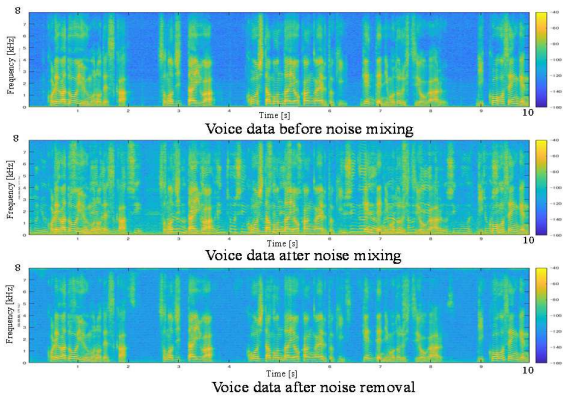

Fig.3 Example of noise removal effect

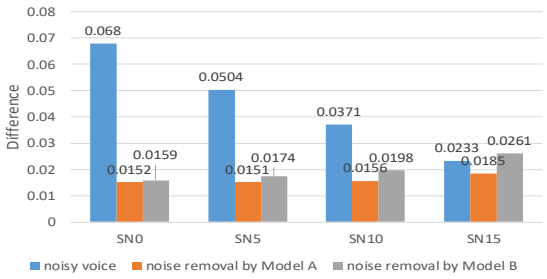

Fig.4 Normalized square difference

The performance of Model A is better than that of Model B. The degree of difference between the spectrogram after removal and before noise mixing with SN ratios of 0, 5, and 10 dB is about 0.015. The removal performance with SN ratio of 15 dB cannot be confirmed from this value, but actually restoring the sound and listening, the noise removal effect was confirmed.

IV. Summary

From the viewpoint of reducing the collection of learning data and the load of learning model creation, we examined removing the voice of a specific speaker using the U-Net model. By converting voice data to images and applying U-Net, the noise removal performance of the learning model was examined. It was confirmed that noise could be removed both numerically and when listening. In the future, not only the author but also many people will evaluate by listening to noise removal voices. Furthermore, the created model will be incorporated into a single board computer to realize a configuration that takes actual use into consideration, and performance will be evaluated in the practical environment.

References

[1] Y. Hashizume, et al., "Study and evaluation of noise reduction method for alarm sound classification," RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP 2020), 1A-M1-1-4, pp. 271-274, 2020.

[2] S.V. Vaseghi, "Advanced Digital Signal Processing and Noise Reduction," Second Edition, Wiley, 2000.

[3] K. Nakadai, et al., "Design and Implementation of Robot Audition System 'HARK'", Advanced Robotics, Vol. 24, No. 5–6, pp. 739–761, 2010.

[4] J. LIN et al., "Investigation of Noise Removal by Using U-Net and Voice Recognition Performance Improvement－For Train Running Noise－," IEICE Technical Report (SeMI), July, 2022, to be appeared (in Japanese).

[5] JNAS: Japanese Newspaper Article Sentences, http://www.mibel.cs.tsukuba.ac.jp/_090624/jnas/.