Voice Recognition Using Surrogate Method with 1D-Convolutional Neural Network

Tomiyuki Furugaki, Tomoya Takata,
Yoko Uwate and Yoshifumi Nishio
Dept. of Electrical and Electronic Engineering, Tokushima University
2-1 Minami-Josanjima, Tokushima 770–8506, Japan
Email:{furugaki, tomoya, uwate, nishio}@ee.tokushima-u.ac.jp

***Abstract*—Voice classification are often classified by using Recurrent Neural Network and 1-Dimensional Convolutional Neural Network (1D-CNN). 1D-CNN is used for classification model. In general, 1D-CNN learns original data. However, in this study, original data is replaced with surrogate data. Test accuracies at that time are compared. In this way, we search which part of time series data is effective for using 1D-CNN.**

I. Introduction

Neural Network (NN) is a system model on neurons of the human brain nervous system. Among them, 1-Dimensional Convolutional Neural Network (1D-CNN) is used for voice classification. 1D-CNN needs to learn voice waveform in advance to classify voice. There are various characteristics depending on each sound source. 1D-CNN classifies voice by finding them. However, the specific judgment part is unknown because we need to find these.

In this study, the surrogate data method is used. The surrogate data method is creating surrogate data. Surrogate data is preserves some of the statistical properties of time series data and destroys other properties. After that, it is indicated that there is a significant difference between the statistical properties of time series data and the surrogate data. In this way, the method proves the importance of destroyed properties.

In this study, the data that the CNN learns is replaced from the original data with surrogate data. It can be seen that which part of the voice waveform is important by comparing the test accuracy at that time.

II. Convolutional Neural Network

The research on CNN was established as an academic field in 1956. Since then, it has repeated the ice ages and booms many times and now reaches the present. Currently, CNN is diverse in medical field, car field, home electronics field and so on. The beginning of these booms is image recognition. CNN is inspired from the biological process and conceived from the arrangement of the visual cortex of animals. In the field of image recognition, CNN has achieved tremendous performance with many tasks. In addition, CNN is attracting attention. In particular, the intermediate layer of CNN extracts high versatility and splendid feature quantities. The network structure of CNN is divided into an input layer, an intermediate layer and an output layer. The intermediate layer includes convolution layers, pooling layers and fully connected layers. Features of inputs are extracted in the convolution layer, and position invariance is acquired in the pooling layer. Next, it becomes the 1-Dimensional array in fully connected layers and it changes to probability. Finally, CNN outputs classification results by the probability. In recent years, CNN has also been use to audio signal processing. In this study, CNN is used for time series data that is one-dimensional data to classify voice.

III. Surrogate data method

The surrogate data method was proposed in 1992 for chaos time series analysis. There are no necessary and sufficient conditions for chaos. Therefore, the only way to determine chaos is to find out that there is chaoticity. In many cases, chaos is determined by spectral continuity, strange attractors, Lyapunov exponents, bifurcations, and so on. However, it has been pointed out that even with random noise alone. The Lyapunov exponent is positive and noise and chaos cannot be distinguished. Therefore, the surrogate data method is proposed to test whether it is noise or data generated from a deterministic system. With hypothesis testing, it is difficult to say that it is noise if a data passes the test. However, it cannot be asserted that it is chaos because the surrogate data method is based on hypothesis testing in statistics. In this study, surrogate data is created and compared the accuracy of learning surrogate data with the accuracy of learning original data. In this way, what characteristics of the original data are important can be found.

IV. Dataset

In this study, time series data of voices are classified. Three types of voice are prepared. In this study, they are called voice 1, 2 and 3. Forty pieces of data for 6 second each are prepared. Each time series data is sampled at a sampling frequency 3000 [Hz]. Next, the data is augmentationed. There are three types of augmentation. Figure 1 shows the examples of the original time series data. Figures 2, 3 and 4 show the examples of the time series data about the data is added white noise, time shift and time stretch for augmentation.

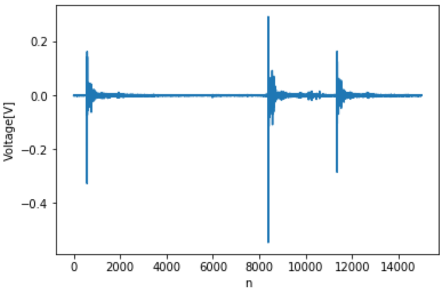

Fig. 1: Original data

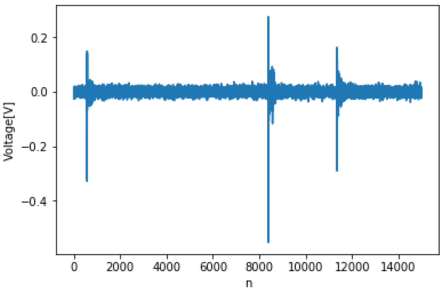

Fig. 2: White noise data

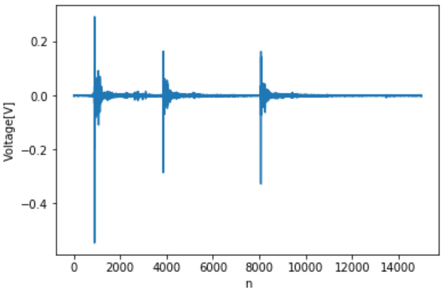

Fig. 3: Time shift data

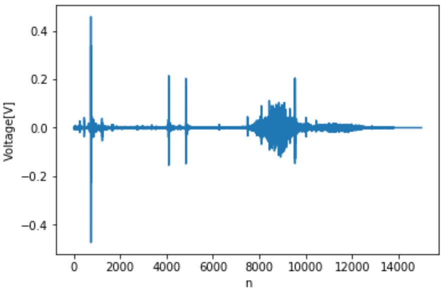

Fig. 4: Time stretch data

V. Proposed Method

Four types of surrogate data are created. Surrogate data is destroyed some information. The following explanations (a), (b) , (c) and (d) describe how to create four types of surrogate data.

(a) Random Shuffle Surrogate Data (RSSD)
$x(n)$ means time function. n means time. It is RSSD data that changes the order of n at random. The correlation of original data is broken by converting the data into RSSD. Figure 5 shows RSSD of the original data in Fig. 1.

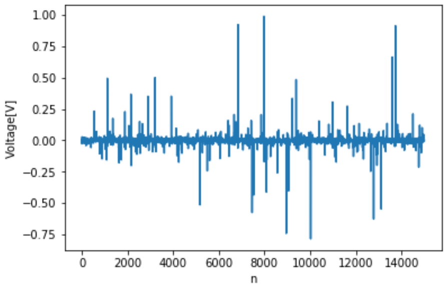

Fig. 5: RSSD

(b) Fourier Transform Surrogates Data (FTSD)

$$X(\omega) = \sum_{n=1}^{n} x(n) e^{-i\frac{2\pi kn}{N}} \tag{1}$$

$$x(n) = \frac{1}{N} \sum_{n=1}^{n} X(\omega) e^{i\frac{2\pi kn}{N}} \tag{2}$$

Equations (1) and (2) show discrete Fourier Transform (DFT) and Inverse Discrete Fourier Transform (IDFT). k means frequency. N (= 15000) means the number of the samples.

Step 1. Calculate DFT $X(\omega)$ of $x(n)$.

Step 2. Randomize the phase of $X(\omega)$.

Step 3. Calculate IDFT randomized $X(\omega)$.

FTSD is made in this way. The frequency distribution of original data is broken by converting the data into FTSD. Figure 6 shows FTSD of the original data in Fig. 1.

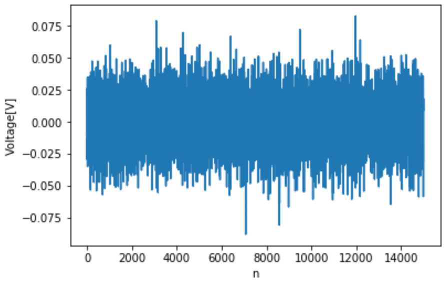

Fig. 6: FTSD

(c) Amplitude Adjusted Fourier Transform Surrogates Data (AAFTSD)

Step 1. Prepare random numbers $R(n)$ according to the standard normal distribution.

Step 2. Sorting $R(n)$ in the same size relation as $x(n)$.

Step 3. Create $R'(n)$ which is FTSD of sorted $R(n)$.

Step 4. Sorting $x(n)$ in the same size relation as $R'(n)$.

AAFTSD is made in this way. The correlation of original data is broken by converting the data into AAFTSD. However AAFTSD has similar correlation than that of RSSD. Figure 7 shows AAFTSD of the original data in Fig. 1.

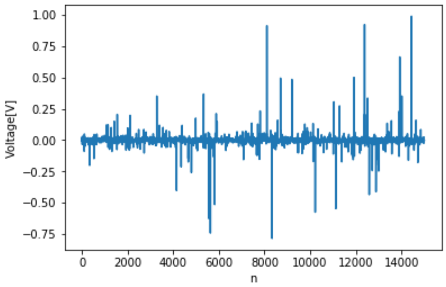

Fig. 7: AAFTSD

(d) Iterated Amplitude Adjusted Fourier Transform Surrogates Data (IAAFTSD)

Step 1. Prepare $s^{(0)}$ which is RSSD of original data as the initial value.

Step 2. Calculate DFT $S^{(i)}$ of $s^{(i)}$.

Step 3. Replace amplitude of $S^{(i)}$ with amplitude of original. Put it as $\overline{S}^{(i)}$

Step 4. Calculate IDFT $\overline{s}^{(i)}$ of $\overline{S}^{(i)}$.

Step 5. Sorting $\overline{s}^{(i)}$ in the same size relation as original data.

Step 6. Add 1 to i.

Step 7. Repeat until $i = 7$.

IAAFTSD is made in this way. IAAFTSD saves the frequency distribution and has similar correlation than that of AAFTSD. Figure 8 shows IAAFTSD of the original data in Fig. 1.

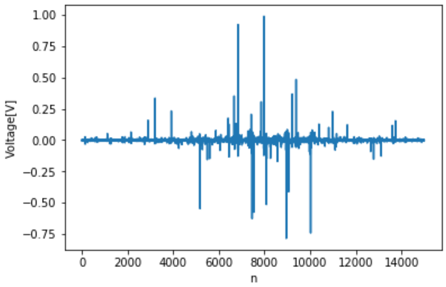

Fig. 8: IAAFTSD

VI. Architecture

1D-CNN is used for convention architecture. Figure 8 shows the structure of 1D-CNN we used.

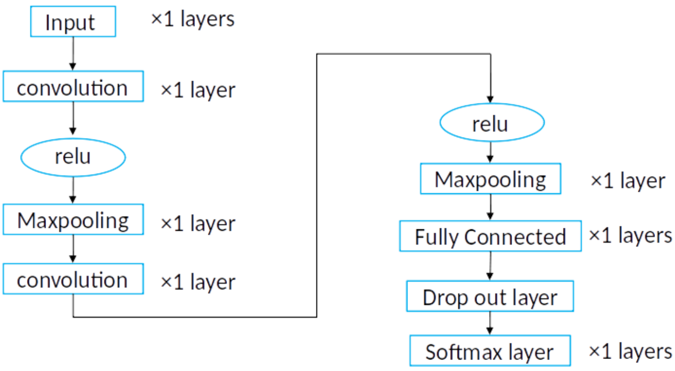

Fig. 9: structure of 1D-CNN

A convolutional layer is composed with two learnable parameters weight and bias. This is called a filter. The filter performs a convolution calculation along the entire input. Equation (3) shows convolution formula.

$$(I * K)(i,t) = \sum_{m} \sum_{n} I(m,n)K(i-m,j-n) \qquad (3)$$

I is the input data and K is the filter. Maxpooling compresses information to transform input data into a more manageable form Two convolutional layers and two maxpooling layers are used. Drop out layer is prepared to prevent over learning. By using the relu function, gradient disappearance hardly occurs and easy to calculate. The probability is derived to calculate classification results by the softmax activation function. Equations (4) and (5) show relu function and softmax function.

$$f(x) = x^{+} = max(0,x) \qquad (4)$$

$$\rho(x) = \frac{exp(x_l)}{\sum_{i=1}^{n} exp(x_l)} \qquad (5)$$

$\rho(x)$ is the probability of being classified as l. n is the total number of classes.

VII. Simulation results

Forty pieces of data are prepared. Each data was augmentationed to three types. After sampling them, surrogate data is created. Furthermore, each data is divided into three pieces. In this way, the number of data became four hundred. Table I shows the number of the train data and test data. Table II shows the results of our research.

TABLE I: The number of the train data and test data

	train data	test data
voice 1	360	120
voice 2	360	120
voice 3	360	120

TABLE II: Test accuracies

	test accuracy [%]
original data	87.9
RSSD	56.9
FTSD	75.9
AAFTSD	57.2
IAAFTSD	78.7

We investigate the average of ten times of test accuracy. Table II shows each test accuracy of surrogate data are lower than that of the original time series data. The rates of test accuracies decline of FTSD and IAAFTSD are lower than those of RSSD and AAFTSD. Test accuracy of IAAFTSD is higher than that of FTSD. These results show that CNN classifies the voices using frequency distribution and correlation.

VIII. Conclusion

In this study, three types classification were carried out with surrogate data. Then, we understood that each test accuracy of surrogate data were lower than that of the original data. Moreover, the declined rates of test accuracies of FTSD and IAAFTSD were lower than those of RSSD and AAFTSD. FTSD did not store frequency distribution of original data. Therefore, it was understandable that frequency distribution was more important than correlation for 1D-CNN. Furthermore, IAAFTSD has correlation and the most similar frequency distribution of original data. However, test accuracy of IAAFTSD was lower than that of original data. From the above, 1D-CNN may recognize the chaoticity of the original data.

In the future, we will find that relationship between similarity of correlation and test accuracy.

References

[1] James Theiler, Paul Linsay, and David M.Rubin, “Detecting Nonlinearity in Data with Long Coherence Times”. In Andreas S. Weigend and Neil A. ershenfeld, editors, “Time Sereis Prediction Forecasting the Future and Understandung the Past”, pp.429-445. Addison Wesley, 1993. a Proceeding Volume in the Santa Fe Institute Studies in the Science of Complexity.

[2] James Theiler, Bryan Galdrikian, Andr Longtin, Stephen Eubank, and J. Doyne Farmer, “Using Surrogate Data to Detect Nonlinearity in Time Series”. In Martin Casdagli and Stephen Eubank, editors, “Onlinear Modeling and Forecasting”, pp. 163-188. Addiison Wesley, 1992. a Proceeding Volume in the Santa Fe Institute Studies in the Science of Complexity.

[3] Daniel T. Kaplan and Richard J. Cohen, “Is Fibrillation Chaos?”. Circulation Research, Vol. 67, No. 4, pp. 886-892, October 1990.

[4] James Theiler. “Some Comments on the correlation dimension of $1/f^{\alpha}$ noise Physics LetterS A, Vol. 155, No. 8, 9, pp. 480-493, May 1991.

[5] James Theiler, Stephen Eubank, Andre Longtin, Bryan Galdrikian, and J. Doyne Farmer, “Testing for nonlinearity in time series: the method of surrogate data” Physica D, Vol. 58, pp. 77-94, 1992.

[6] Dean Prichard and James Theiler. “Generating Surrogate Data for with Several Simultaneously Measured Variables”. Physical Review Letters, Vol. 73, No. 7, pp. 951-954, August 1994.

[7] Thomas Schreiber and Andreas Schmitz. “Improved Surrogate Data for Nonlin- earity Tests”. Physical Review Letters, Vol. 77, No. 4, pp. 635-638, 1996.