Voice Recognition Using Surrogate Data Method with 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

Voices are often classified by using Recurrent Neural Network and 1-Dimensional Convolutional Neural Network. In general, Recurrent Neural Network and 1-Dimensional Convolutional Neural Network learn original data. 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 Recurrent Neural Network and 1-Dimensional Convolutional Neural Network.

1. Introduction

Neural Network (NN) is a system model on neurons of the human brain nervous system. Among them, Recurrent Neural Network (RNN) and 1-Dimensional Convolutional Neural Network (1D-CNN) are used for voice classification [1]-[2]. They need to learn voice waveform in advance to classify voice. There are various characteristics depending on each sound source. They classify voice by finding them. However, the specific judgment part is unknown. Therefore, it is necessary to find these.

In this study, surrogate data method is used. The surrogate data method is creating surrogate data. Surrogate data is data that 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 learned by RNN and 1D-CNN are replaced from original data with surrogate data. It can be found out which part of the voice waveform is important by comparing the test accuracy at that time.

2. Neural Network

The research on NN was established as an academic field in 1958 [3]. Since then, it has repeated the ice ages and booms many times and now reaches the present. Currently, NN is used in medical field, car field, home electronics field and so on [4]-[6]. NN has two famous classification models. They are RNN and CNN. RNN is used in fields related to time series data. CNN is used in fields related to image recognition. The network structure of NN is divided into an input layer, an intermediate layer and an output layer. The intermediate layer of CNN 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. However, in recent years, CNN has also been use to time series data. In this study, CNN is used for time series data that is one-dimensional data to classify voice.

3. Surrogate Data Method

The surrogate data method was proposed in 1992 for chaos time series analysis [7]. 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 [8]-[10]. 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 original data are important can be found.

4. Dataset

In this study, time series data of voices are classified. Three types of voice are prepared. In this study, they are called voices 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 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.

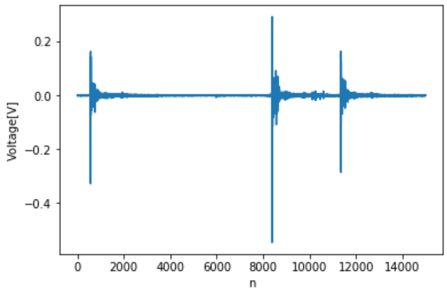

Figure 1: Original data

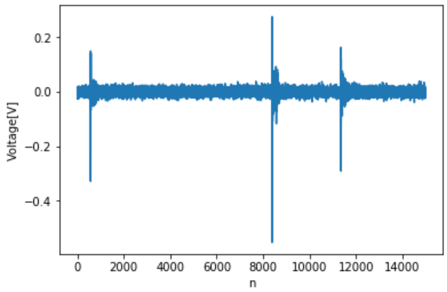

Figure 2: White noise data

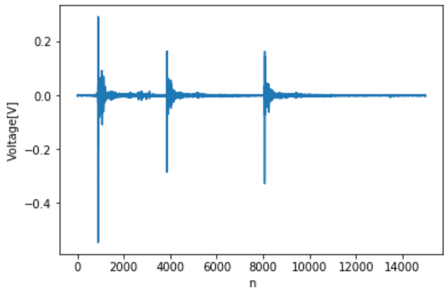

Figure 3: Time shift data

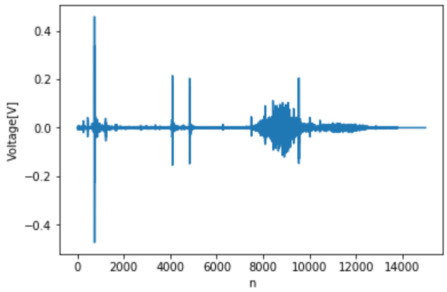

Figure 4: Time stretch data

5. Proposed Method

Four types of surrogate data are created. Surrogate data is a destruction of some information of original data. 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 original data in Fig. 1.

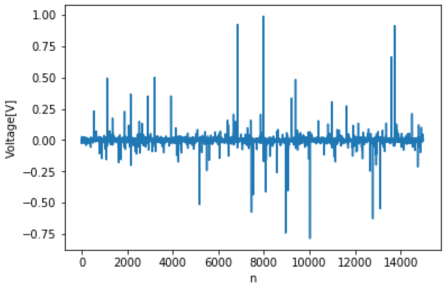

Figure 5: RSSD

(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 original data in Fig. 1.

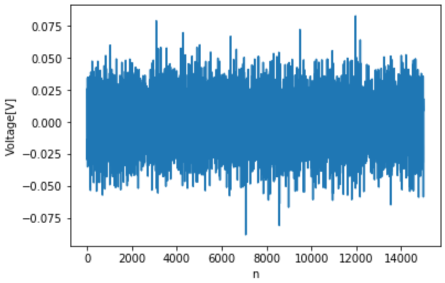

Figure 6: FTSD

(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 original data in Fig. 1.

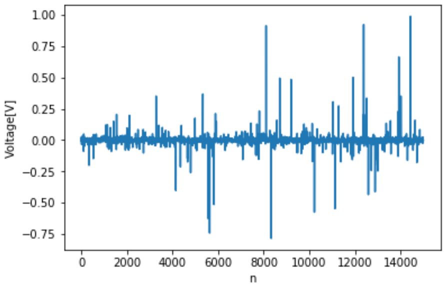

Figure 7: AAFTSD

(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 original data in Fig. 1.

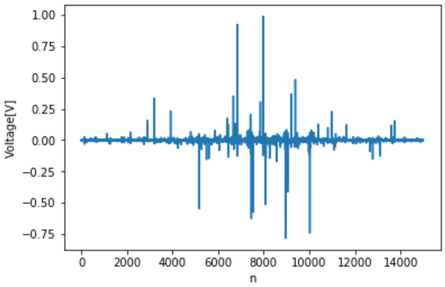

Figure 8: IAAFTSD

6. Architecture

RNN and 1D-CNN are used for convention architecture. Figures 8, 9 show the structure of RNN and 1D-CNN we used.

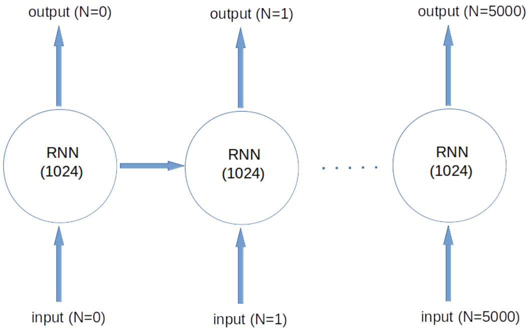

Figure 9: structure of RNN

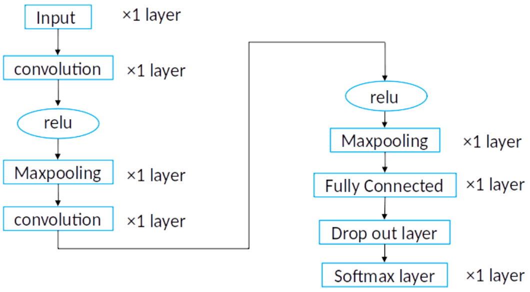

Figure 10: structure of 1D-CNN

RNN has 1024 nodes in the intermediate layer. RNN can store past information.

Two convolutional layers and two maxpooling layers were used for 1D-CNN. Drop out layer was prepared to prevent over learning. We derived the probability to calculate classification results by the softmax activation function. Equation (3) shows softmax function.

$$\rho(x) = \frac{exp(x_j)}{\sum_{i=1}^{n} exp(x_i)} \tag{3}$$

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

7. 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 1 shows the number of the train data and test data. Table 2 and 3 show the results of our research.

Table 1: The number of the train data and test data

	train data	test data
voice 1	360	90
voice 2	360	90
voice 3	360	90

Table 2: Test accuracy of RNN

	test accuracy [%]
original data	74.1
RSSD	43.6
FTSD	39.5
AAFTSD	42.5
IAAFTSD	45.0

Table 3: Test accuracy of 1D-CNN

	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. Tables 2 and 3 show that test accuracies of surrogate data are lower than test accuracies of original time series data. The decay rate of test accuracy of RNN of RSSD is lower than that of FTSD. The decay rate of test accuracy of 1D-CNN of FTSD is lower than that of RSSD. Test accuracies of IAAFTSD of RNN and 1D-CNN are higher than those of the other surrogate data. FTSD did not store frequency distribution of original data. RSSD did not store correlation of original data. Therefore, it was understandable that correlation is more important than frequency distribution for RNN. On the other hand, it was understandable that frequency distribution is more important than correlation for 1D-CNN.

8. Conclusion

In this study, three types classification were carried out with surrogate data. Then, we understood that correlation is more important than frequency distribution for RNN. On the other hand, it was understandable that frequency distribution is more important than correlation for 1D-CNN.

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

References

[1] Gelly, Gregory / Gauvain, Jean-Luc, “Minimum word error training of RNN-based voice activity detection”, In Interspeech-2015, pp. 2650-2654.

[2] Abhishek Sehgal, Nasser Kehtarnavaz, ”A Convolutional Neural Network Smartphone App for Real-Time Voice Activity Detection”, Access IEEE, vol. 6, pp. 9017-9026, 2018.

[3] F. Rosenblatt,“The Perceptron：a Probabilistic Model for Information Storage and Organization in the Brain”, Psychol. Rev, Vol. 65, no. 6, pp. 386-408, 1958.

[4] Shih-Chung B. Lo, 1 Heang-Ping Chan, 2 Jyh-Shyan Lin, 1 Huai Li, 1 Matthew T. Freedman and Seong K. Mun 1, “Artificial Convolution Neural Network for Medical Image Pattern Recognition”, Neural Networks, Vol. 8, No. 7/8, pp. 1201-1214, 1995.

[5] S. Draghici. “A neural network-based artificial vision system for license plate recognition”. Int. J. Neural Syst. January, pp. 113-126, 1997.

[6] Ameena Saad al-sumaiti, Mohammed Hassan Ahmed Magdy M. A. Salama. “Smart Home Activities: A Literature Review”. Electric Power Components and Systems, pp. 294-305, Februaly 2013.

[7] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D.Farmer, Physica (Amsterdam) 58D, pp. 77 1992.

[8] 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.

[9] 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.

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