inter.noise
HAMBURG 2016

DNN-based Environmental Sound Recognition with Real-recorded and Artificially-mixed Training Data

Yasutaka NAKAJIMA[1]; Masahiro SUNOHARA[2]; Taisuke NAITO[3]; Norihito SUNAGO[4];
Toshiya OHSHIMA[5]; Nobutaka Ono[6]

[1-5] RION CO., LTD., Japan

[6] National Institute of Informatics, Japan

ABSTRACT

In this paper, we report on our investigation of environmental sound recognition using a deep neural network (DNN). Preparing a sufficient amount of training data is generally important in machine learning On the other hand, different environmental sounds such as cicada and ambulance sounds occur while overlapping each other. As a result, training data including mixtures of different sounds are necessary for environmental sound recognition. However, it is difficult to obtain all combinations of different sounds in real-recorded data. In this study, we increased the amount of training data using artificially-mixed sounds. First, some distinctive single sounds which were recorded on different days near the sound sources individually were selected, or others were separated from the real-recorded data by extracting appropriate parts in the time domain. Those sounds which mainly consisted of a single sound source were applied filters to reduce other sounds in the frequency domain. Next, they were mixed with different ratios of sound levels, simulating a variation of possible mixings in the real environment. Finally, both the real-recorded data and the artificially-mixed sound data were used for training the DNN and we then conducted on environmental sound recognition. We show that this approach achieves more accurate results than the ones using only real-recorded data.

Keywords: DNN Sound Recognition I-INCE Classification of Subjects Number(s): 74.4, 74.8, 74.9

1. INTRODUCTION

Recognizing environmental sounds is important to know not only noise pollution such as road traffic noise and aircraft noise but also bird twitters, insect sounds, the sounds of the wind and rain, and daily life sounds that depend on the season. It is also important to identify what kind of sound sources affect a particular sound environment at a particular place to take proper countermeasures.

However, the number of data becomes huge if the number of monitoring points is large and the measurement term is long. Operators have a hard time accurately distinguishing sounds successively. Consequently, an automatic recognition is important for a large number of data.

In a previous paper, we compared the results of both linear discriminant analysis and deep neural network (DNN)[1]. Preparing a sufficient amount of training data is generally important in machine learning. However, it takes a lot of time and effort.

In this study, we used both artificially mixed data and actually recorded data as training data and aimed to improve the recognition rate and to reduce the time and effort for annotation or labeling sounds.

[1] yasutaka@rion.co.jp
[2] suno@rion.co.jp
[3] tnaito@rion.co.jp
[4] sunago@rion.co.jp
[5] t-ohshima@rion.co.jp
[6] onono@nii.ac.jp

2. MEASUREMENT

2.1 Sound Environment of Measurement Point

The measurement point was located in a company parking area behind several two-story houses, approximately 30 meters behind a typical suburban bus road[2]. A fire department branch office is 300 m west of the measurement point, and we could sometimes hear ambulance and fire truck sirens. Major sound events were cicadas, outside air conditioners, road traffic noise, and neighborhood noise such as garage doors opening and aircraft flying overhead. One single sound rarely occurred separately, and usually more than two sounds occurred simultaneously.

A sound level meter, RION NL-52, with the NX-42RT program was placed at the point, and sound pressure waveforms and short-time L_{eq} data for each 1/3 octave band were continuously recorded on a SD memory card (resolution of 24 bits, 48 kHz with a frequency weighting 'Z') By selecting 24-bit resolution, wide dynamic range sound events such as loud aircraft noise or quiet insect sounds were able to be recorded over a long time period.

2.2 Frequency Analysis

We used one-third octave band levels of every one second, which can be easily measured by sound level meters. Each input signal was passed through 1/3 octave band-pass filters from 12.5 Hz to 20 kHz (33 bands). The output signals of each filter was converted to L_{eq} averaged every one second interval.

3. TRAINING AND TEST

3.1 Training Data

We herein compare two methods: The previous method used only actually recorded data as training data. Our proposed method used both actually recorded data and artificially mixed data as training data. We heard actually recorded data and labeled them using 55 categories, as shown in Table 1. Some of index number of sound sources were not used because there were no sound events associated with some sound sources.

If more than two sounds occurred simultaneously, we labeled them as multiple sounds. We prepared 8700 seconds or 8700 pieces of data as training data as shown in Figure 3. We used 18:00–20:00 data on another day as test data; these data were different from the training data because the test data were used for a recognition evaluation.

Table 1 - Sound source category list.

Sound sources	
0: car	27: roll of thunder
1: truck	30: chattering
2: motorcycle	32: noise of something hitting something else
3: car with modified muffler	33: dog
4: motorcycle with modified muffler	34: street vender loudspeaker
5: door shutting sound	35: air conditioner
6: horn	40: chime by disaster management loudspeaker
7: engine sound	41: announcement by disaster management loudspeaker
8: reverse alarm sound	42: advertising van
10: aircraft sound	45: ambulance
11: helicopter	46: fire truck
20: bird	47: police car
21: crow	50: road construction
22: cicadas	51: construction
25: rain	54: background noise
26: wind	

3.2 Artificially Mixed Data

Regarding environmental sound recognition, different environmental sounds, for example, cicada sounds and ambulance sirens, often occur simultaneously. Therefore, we had to prepare training data which was mixed using multiple sounds at variated different levels. However, recording all the combinations of sounds as actually recorded data is practically difficult because the number of combinations is huge. Moreover, Operators also have practical difficulty in labeling all the data. Applying multiple labels for one sample is possible regarding DNN when there are overlapped sounds, but it is meaningless if we cannot prepare enough samples of multiple sounds occurring simultaneously.

To solve this issue, we added artificially mixed sounds as training data. Here, we individually recorded sounds consisting of one single sound source as close as possible to the sound source on different days. In this study, we used 13 cicada sounds, six ambulance siren sounds, and 35 outside air conditioner sounds as typical sounds at the measurement point.

For sounds that could not be recorded near the sound source, in this case, we inevitably extracted five second data at the measurement point, with the data consisting of only the target sound source. We used frequency filters to suppress non-target sounds if the sound includes non-target sounds. For cicada sounds, we applied a high pass filter with a cut-off frequency of 500 Hz and 24 dB/oct. to suppress outside air conditioner sounds and then to prepare only cicada sounds. For the ambulance siren sounds, we applied a high pass filter with a cut-off frequency of 500 Hz and 24 dB/oct. and a low pass filter with a cut-off frequency of 2.5 kHz and 24 dB/oct. to suppress non-target sounds such as outside air conditioner sounds. No frequency filter was applied for outside air conditioner sounds because they almost always occurred and could be easily recorded.

After that, we artificially mixed those sounds at various levels and prepared a variety of sounds as training data to simulate practically all possible combinations of sound mixtures at possible levels, as shown in Figure 1. The range in which we varied sound levels came from one second L_{eq} level histogram range of actually recorded data for each sound. For cicada sounds, we divided 13 samples into ten and three samples and then varied their sound levels in the form of three patterns to -5 , 0 , and +5 dB, where 0 dB means the typical sound level of cicada sounds at the measurement point. As a result, we made 10 x 3 = 30 and 3 x 3 = 9 samples, respectively. Similarly, for outside air conditioner sounds, we divided 18 samples into three, two, ten, and three samples and then varied their sound levels to -3 , 0 , and +3 dB. As a result, we made nine, six, 30, and nine samples. For ambulance sounds, we divided six samples into two and four samples and then varied their sound levels to -6 , -3 , 0 , +3, and 6 dB. As a result, we made ten and 20 samples.

After we artificially mixed those sounds, finally we got a mixture of cicada and outside air conditioner sounds at different levels, 270 samples; a mixture of all the data, 540 samples; a mixture of ambulance and outside air conditioner sounds, 180 samples; outside only sounds, 30 samples; for a total of 1020 samples.

Moreover, we considered the effect of background noise. We extracted data having no specific dominant sounds as background noise and calculated the L_{95} of each band level. We added the calculated band levels to 1020 samples in the form of power summation as typical background noise in the frequency domain. The duration of one sample was five seconds, or 5100 pieces of data. Finally, we combined both 8700 actually recorded pieces of data and 5100 artificially mixed pieces of data as training data on the DNN as shown in Figure 4.

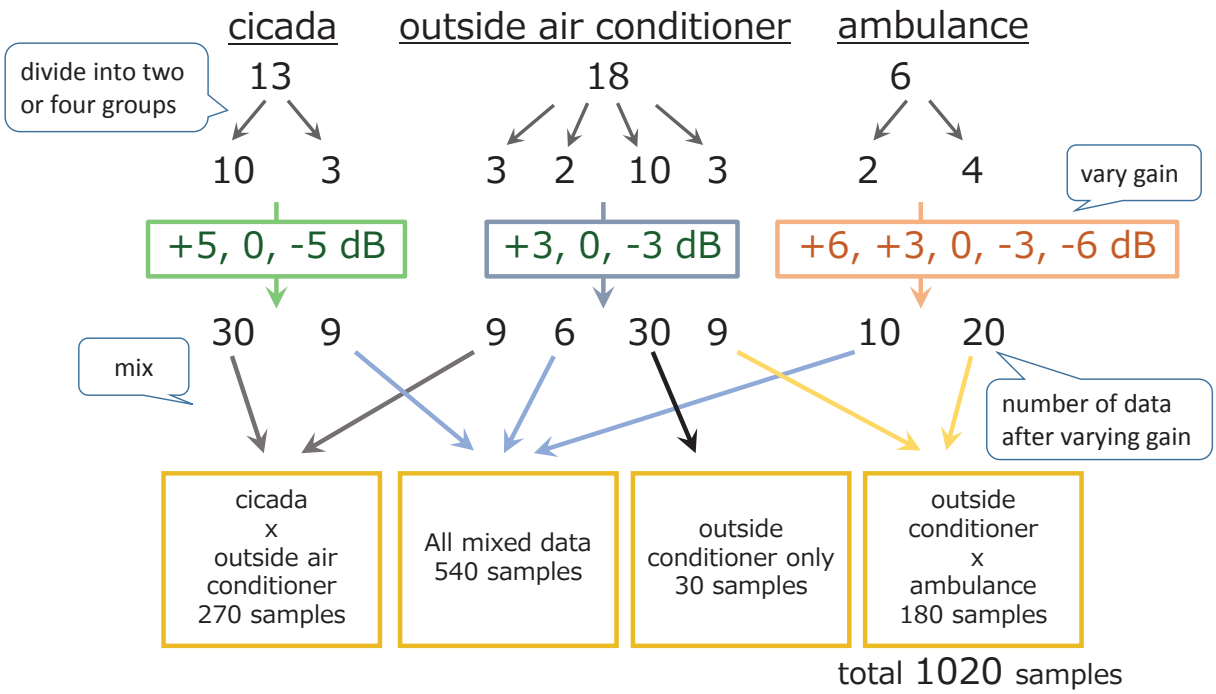


Figure 1 - Artificial mixture procedure of each sound after varying sound levels.

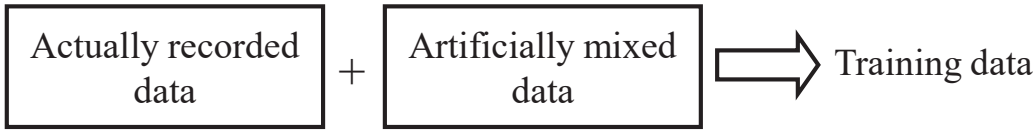


Figure 2 - Combined training data consisting of both actually recorded data and artificially mixed data.

3.3 Deep Neural Network

The number of input features for both the previous method without artificially mixed data and the method with artificially mixed data was 33, and the architecture of the DNN included two hidden layers with 20 and ten neurons. The output layer output a 55 dimensional vectors corresponding to the index numbers of the sound sources. The output value of the DNN means the likelihood that input data contain each sound source associated with each node of the output layer.

As a pre-training of the DNN, each adjacent pair of the neural network (NN) layers was trained as the restricted Boltzmann machine (RBM) in an efficiently greedy, layer-wise technique for initializing the NN weights. After the pre-training, the weights were fine-tuned using the backpropagation algorithm with the supervised labeled data.[3] The conditions of the training are shown in Table 2. A computer calculation by MATLAB was carried out using the DeepLearnToolbox.[4]

Table 2 – Conditions of the DNN training

Number of nodes	Input layer	33
	Hidden layer	20, 10
	Output layer	55
Number of hidden layers		2
Pre-training		100 epochs for each layer
Fine-tuning		300 epochs
Activation function of each layer		Sigmoid

3.4 Test Data

We used sound data from 18:00 to 20:00 on a summer day as test data, as shown in Figure 4. It was a day of sunny weather and maximum temperature of 37 degrees Celsius. It was 29 degrees Celsius at night. Sound ranged from 4 kHz to 16 kHz, the high frequency component red color area indicates cicadas. In suburb areas with some trees, cicadas can be one of the main sound sources during summer in Japan, and sometimes we could not measure and evaluate noise such as road traffic noise and aircraft noise because of cicada sounds. Sunset was at 18:31 on that day. Cicadas stopped chirping around 20 minutes after sunset. A bright vertical line in the lower frequency of the sound spectrogram indicates aircraft and motorcycle sounds. There were outside air conditioners 30 meters away from the measurement point and whirring stationary. We heard the sound of insects at night.

4. RESULTS

4.1 Recognition Results without Artificially Mixed Data – the Previous Method

We show DNN recognition results which used only actually recorded data as training data. Figure 4 shows a spectrogram of the test data. Figure 5 shows the recognition results about sound sources having more than 0.5 likelihood at the output layers. The number of training data is 8700 data, or second which were labeled every one second. Outside air conditioner sounds, sound source No. 35, were almost always constant and cicadas, No. 22, stopped chirping at around 18:50. The results correspond to the real situations. However, ambulance sounds were not able to be recognized adequately except around 18:20.

Figure 6 shows the DNN outputs or the recognition results for some of the typical sound sources. The orange hatching area represents intervals that there were actual sounds which were recognized by humans. For aircraft, the recognition results or the likelihood corresponds to the reality until around 18:40, but the likelihood generally became high after 18:50 and did not correspond to the reality. For cicadas, the likelihood was high until around 18:50, and after that the cicadas stopped chirping. This corresponds to reality very well. It is difficult even for humans to determine when cicadas stop chirping because the number of chirping cicadas decreased gradually at around 18:50. Outside air conditioners made sounds constantly, and the likelihood was always high. For ambulances, the results were correct around 19:22 and 19:50. However, the results were not correct until 18:50, or the interval where cicada chirping overlapped.

We calculated the receiver operating characteristic (ROC) curve from TPR and FPR while changing the threshold of the output layer at 0.001 step from 0 to 1 because the recognition results depend on the threshold setting of the output layer.[5] The Y axis represents the sensitivity or TPR= TP/ (TP+FN), which means the probability that sound is recognized for a target sound if one actually occurred. The X axis represents FPR=FP/(FP+TN). TP represents a true positive, FP indicates a false positive, FN means a false negative, and TN is a true negative. We got the area under the curve (AUC) and the equal error rate (EER) about the ROC curve to evaluate recognition performance, as shown in Figure 7.

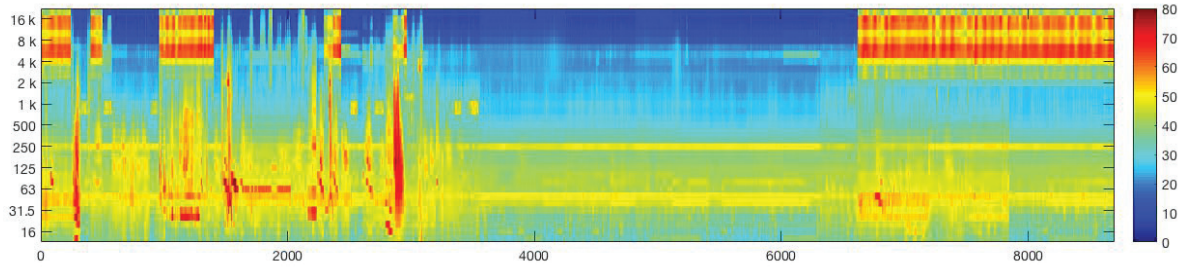


Figure 3 – Sound spectrogram of training data without artificially mixed data.

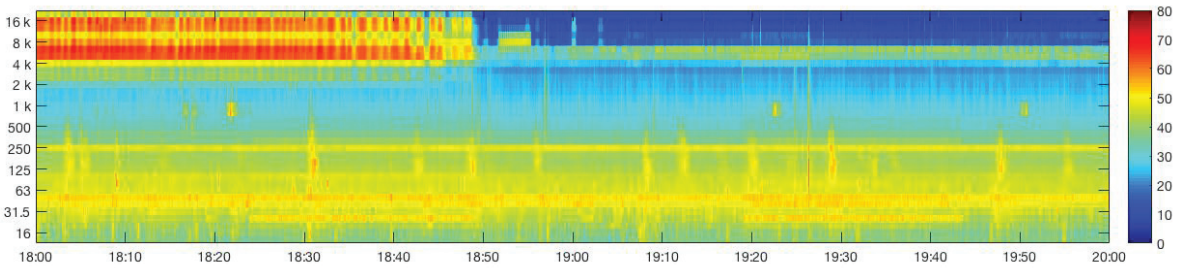


Figure 4 - Sound spectrogram of test data.

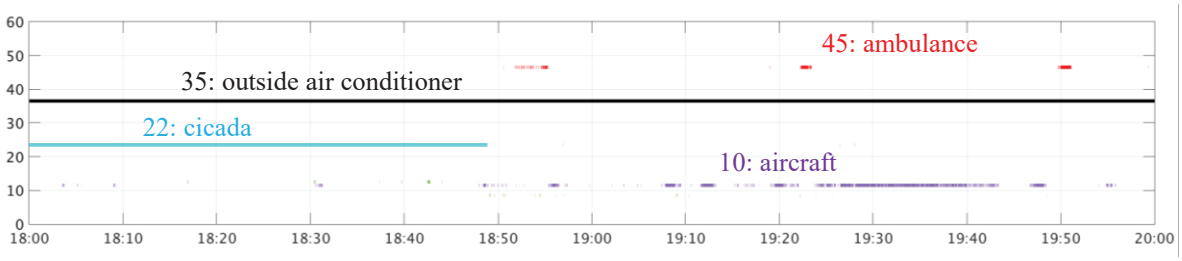
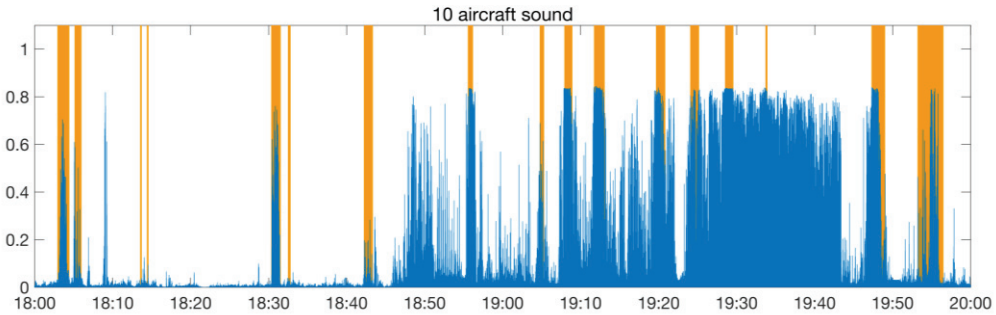
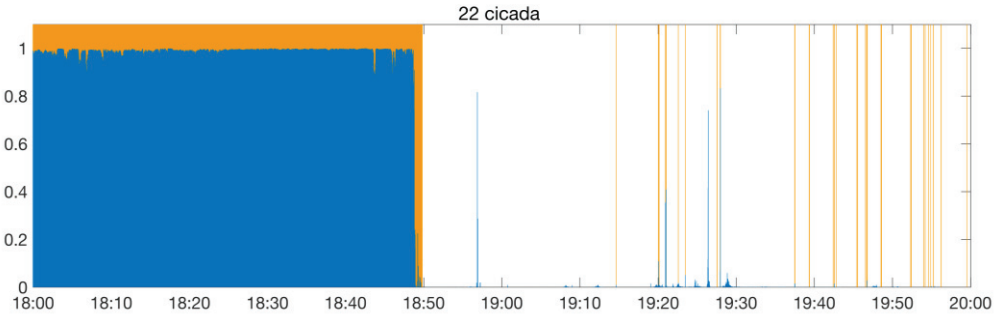


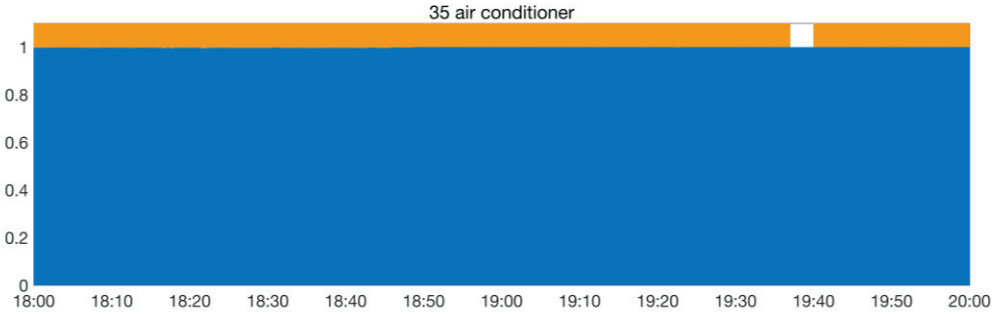
Figure 5 - Recognition results without artificially mixed data: sound sources having more than 0.5 likelihood.



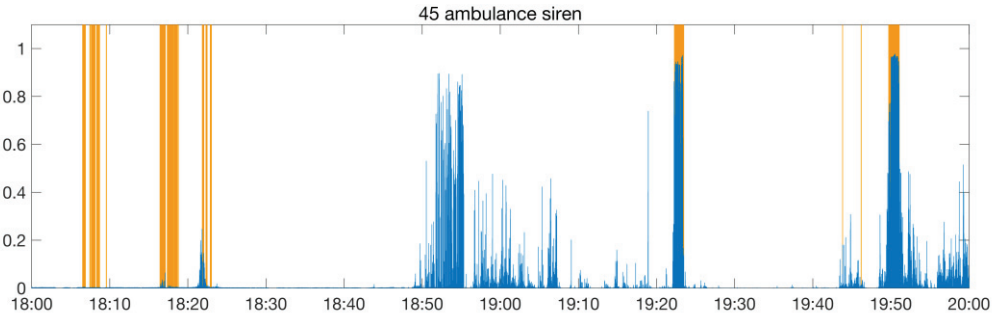
(a) aircrafts



(b) cicadas



(c) outside air conditioners



(d) ambulance sirens

Figure 6 - Recognition results without artificially mixed data.

Vertical orange hatching area: intervals that there were actual sounds recognized by humans

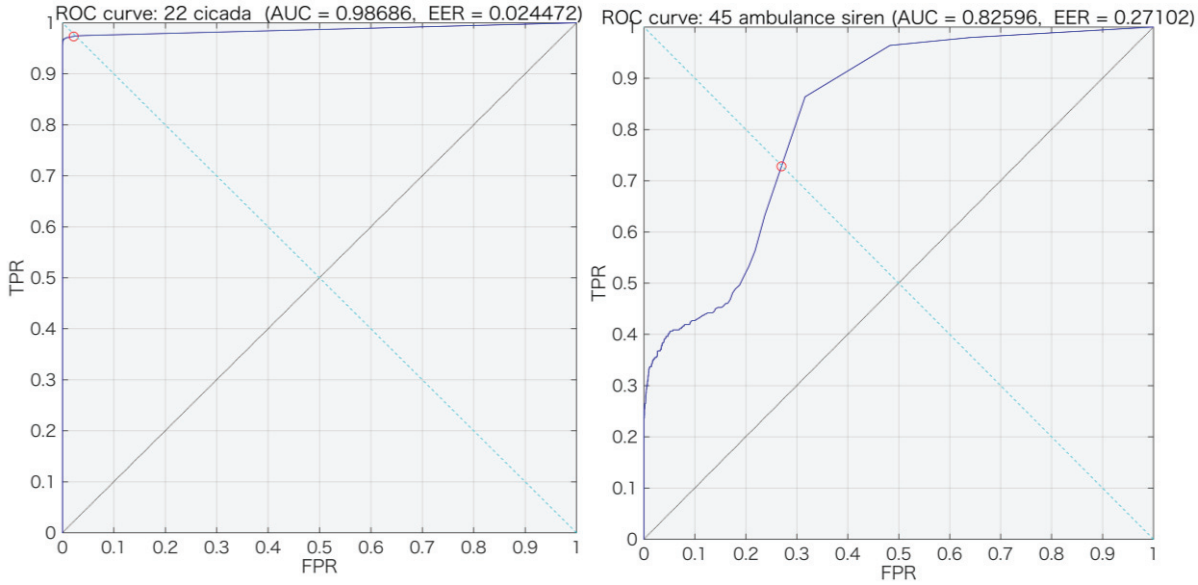


Figure 7 – ROC courve of each sound source without artificially mixed data.

Left: cicadas, Right: ambulance sirens.

4.2 Recognition Result with Artificially Mixed Data – the Proposed Method

We show recognition results in which both actually recorded data and artificially mixed data were used as training data, as shown in Figure 8. The artificially mixed data include ambulance, cicada and outside air conditioner sounds. We also show recognition results of aircraft as an example not included as artificially mixed data.

The proposed method shows a much lower likelihood than the previous method in intervals where aircraft sounds were evident before 18:50, as shown in Figure 10. The likelihood increased when no aircraft sounds were evident after 19:00. The proposed method shows good results for cicadas—almost the same results until 18:50. After that, the proposed method had a higher likelihood than the previous method because cicadas sometimes chirped suddenly even at night because of high temperatures. The proposed method has better recognition results for ambulances between 18:00 and 18:50, even if ambulance siren sounds were overlapping with cicada sounds. For ambulance sounds, the EER of the proposed method improved from about 0.271 to 0.168, as shown in Figure 12. For cicadas, the EER of both the previous method and the proposed method were about 0.024 and 0.029, respectively and already low because actually recorded data had included a lot of cicada sounds, and the number of training data was enough for cicada sounds in the previous method. For aircraft, the EER slightly decreased from 0.757 to 0.694 and the both EER were high. Aircraft sounds were not included as artificially mixed data.

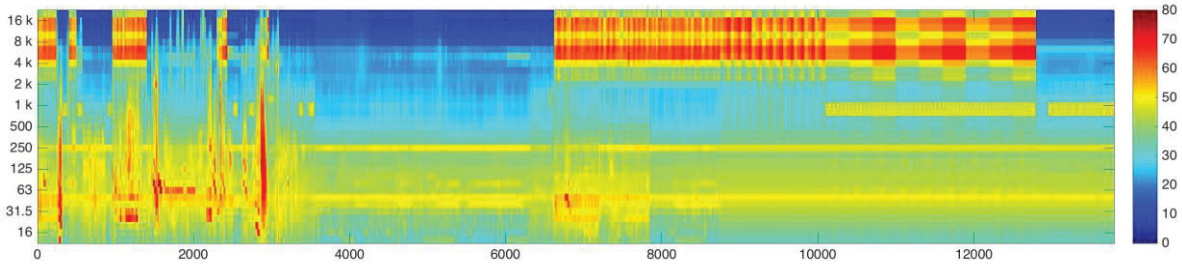


Figure 8 – Sound spectrogram of training data with artificially mixed data.

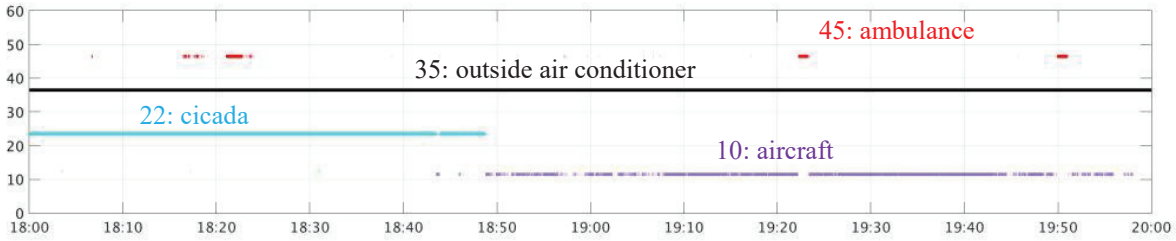
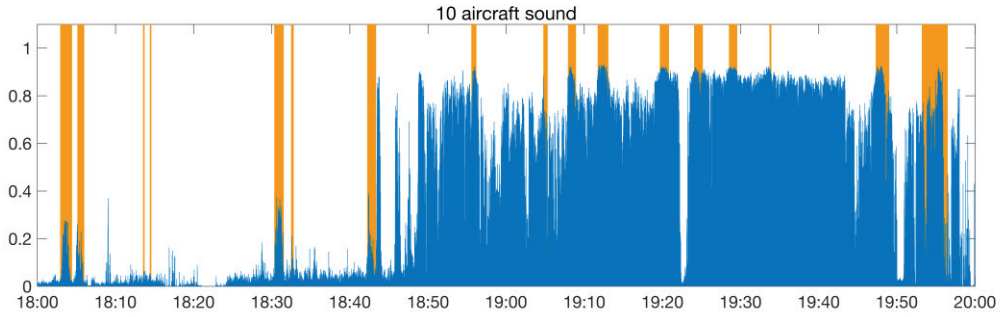
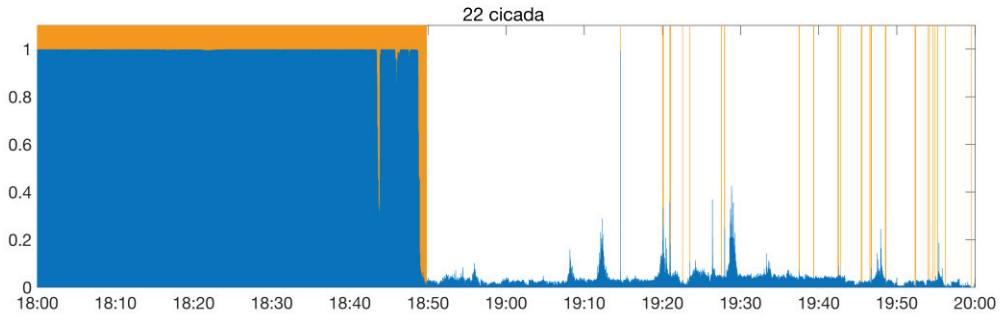


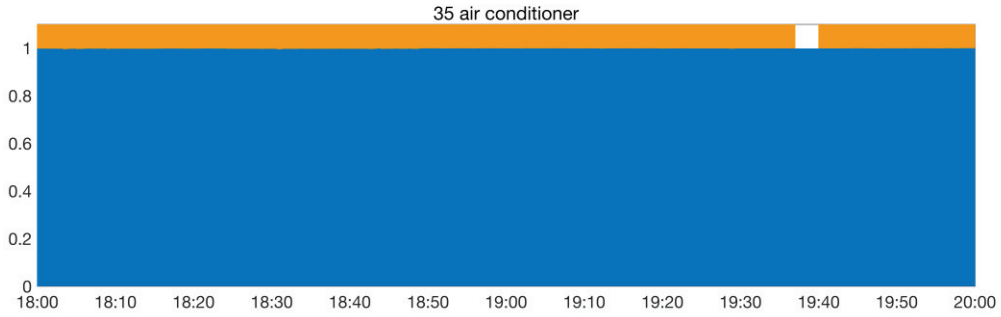
Figure 9 - Recognition results with artificially mixed data: sound sources having more than 0.5 likelihood.



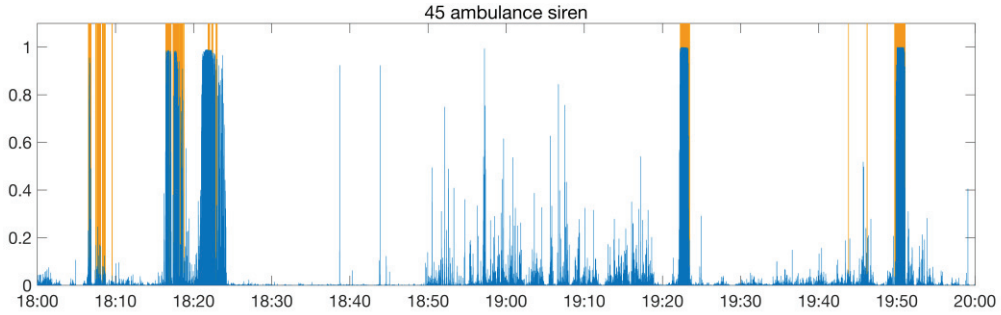
(a) aircrafts



(b) cicadas



(c) outside air conditioners



(d) ambulance sirens.

Figure 10 - Recognition results with artificially mixed data:

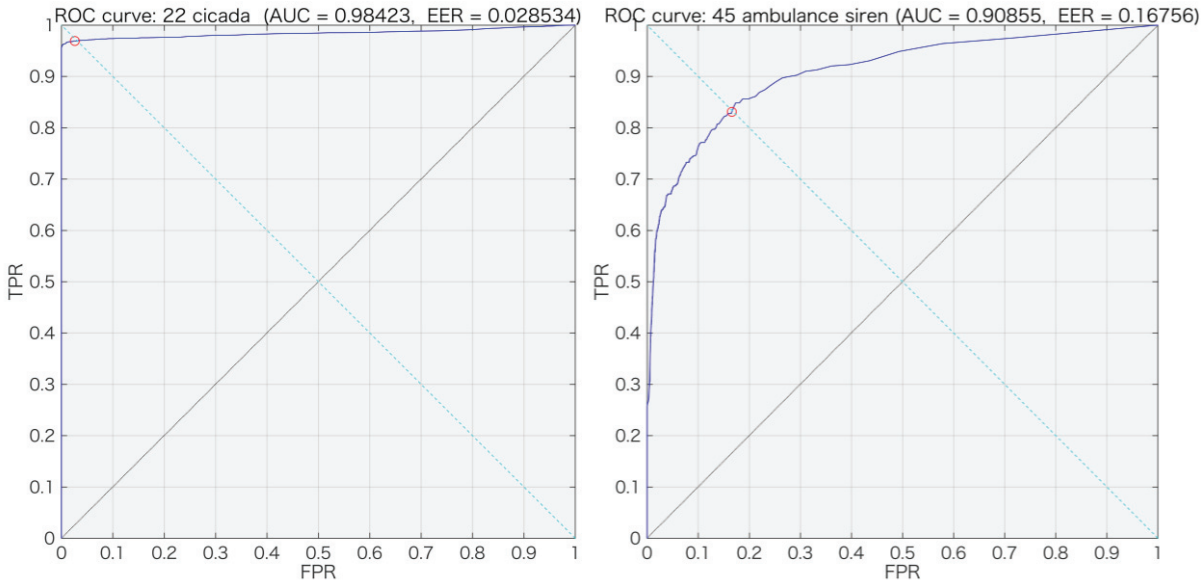


Figure 11 – ROC courve of each sound source with artificially mixed data.

Left: cicadas, Right: ambulance sirens.

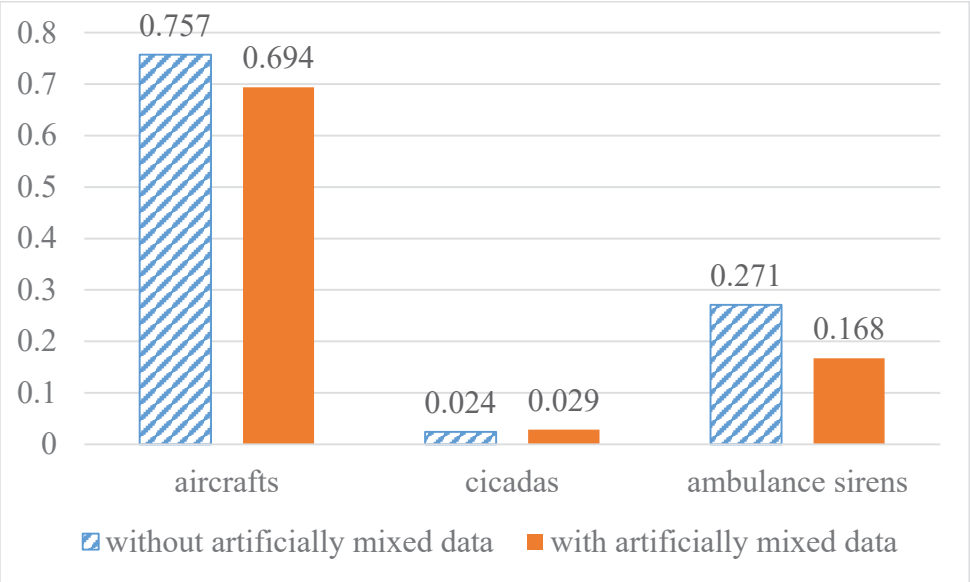


Figure 12 - Equal error rate

5. CONCLUSIONS

The proposed method showed improved recognition results for ambulances than the previous method in terms of EER. The occurrence rate of ambulances was low and the number of the mixture sound pattern as training data regarding ambulances was also low. Artificially mixed data was useful to decrease EER. However, the proposed method showed higher EER for aircraft that is not included as artificially mixed data. Sound targets to be automatically recognized have to be included as artificially mixed data for better recognition results.

The proposed method has the following advantages. First, regarding sound sources such as ambulances, which have a lower occurrence rate, we used artificially mixed sounds to increased the number of training data of ambulances that we would like to recognize and then we got improve recognition results

Second, preparing all the combinations of sounds was difficult in the conditions where various sounds overlap because the number of combinations increased; that is, it was difficult to actually record all the data. We could easily prepare overlapping sound source data that were difficult to record actually, if we recorded target-sounds near sound sources and then artificially mixed them with each other.

Finally, labeling all the sounds data was not necessary because we already know the sound source labels of the sounds that were artificially mixed. We can reduce time and effort for labeling a lot of data consisting of combinations of various sounds.

For future work, we have to consider the effects of the transfer characteristic between a sound source and a measurement point to expand this recognition method to different points because measured sounds depend not only on sound source characteristics, which could already be put on a sound database as near sound sources but also on the transfer characteristic between a sound source and a measurement point.

REFERENCES

1. Nakajima Y., Sunohara M., Naito T., Sunago N., Ohshima T. Environmental noise recognition using DNN. Proc INTER-NOISE 2015; 9–12 August 2015; San Francisco, Japan.
2. Ohshima T. et al, Evaluation of environmental sound quality considering meteorological conditions and masking effects of background noises. Proc INTER-NOISE 2013; 15–18 August 2013; Innsbruck, Austria.
3. Hinton G. E., Osindero S., and Teh Y. A fast learning algorithm for deep belief nets. Neural Computation, 18:1527–1554, (2006).
4. Palm R. B. Prediction as a candidate for learning deep hierarchical models of data. Master's thesis, 2012.
5. Murphy K. P. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012, pp. 180–182.