Audio Classification of Bit-Representation Waveform

Masaki Okawa¹, Takuya Saito¹, Naoki Sawada¹, Hiromitsu Nishizaki²

¹Integrated Graduate School of Medicine Engineering, and Agricultural Sciences, University of Yamanashi;
²Graduate School of Interdisciplinary Research, University of Yamanashi

Abstract

This paper investigates waveform representation for audio signal classification. Recently, many studies on audio waveform classification such as acoustic event detection and music genre classification have been increasing. Most studies on audio waveform classification are based on deep neural networks (DNNs) because environmental sound analysis including detection and classification will become a key technology in the near future.

In acoustic event detection tasks, some corpora have been released and they are used at a competition such as DCASE. And, on the music genre classification task, GTZAN and FMA have been released. So we can use so many datasets for studying audio/sound/music processing.

Generally, in an audio detection task, an audio raw waveform is pre-processed before inputting to a neural network. Many pre-processing methods have been tried to use. For example, Mel-Frequency Cepstrum Coefficients (MFCCs) and filter bank outputs are popular and widely used. Most pre-processing approaches transform a time domain waveform into its frequency domain by computing the Discrete Fourier Transform (DFT). Recently, end-to-end approaches have been becoming popular. However, there are a few papers that investigate feature extraction methods from a time-domain waveform. This is because why a neural network architecture for audio classification has mainly two hierarchies: one is composed of convolutional neural network (CNN) layers for feature extraction, the other consists of recurrent-based layer(s) and/or fully-connected (FC) layer(s) to classify the input waveform. The CNN layer(s) can extract features from the input waveform instead of human-designed features. Until recently, therefore, a few studies tackled the pre-processing methods of an audio waveform. Lee et al. proposed a CNN architecture which can learn representations using sample-level filters on the music genre classification task. Sainath et al. also investigated that raw waveforms are directly inputted to the convolution layer.

Against the previous works, we propose a new pre-processing method that converts a raw audio waveform into a bit-based representation. We investigate two sorts of bit-representations of an audio waveform in this paper. This is a perfect novel idea that has never proposed before. We evaluate the effectiveness of the bit-representation of the raw audio by directly classifying audio files.

The contributions of this paper are described as follows:

- The paper first shows that bit-representation waveform makes convolution layers of a neural network to extract more effective features from audio. Therefore, the performances of both the acoustic event detection task and the music/speech binary classification task achieved drastically improved compared to the typical data representation such as MFCCs and power spectrum.
- The paper experimentally shows the bit-representation waveform has noise robustness on the music/sound classification task because time sequences of higher-order bits of the bit representation waveforms were not strongly affected by noise.
- The bit-representation also has robustness against the domain mismatch between training and testing of the neural network.
- The bit-representation of a raw audio waveform is useful on various types of audio classification.

This paper is organized as follows: the next session presents how to transform an audio waveform into bit-representation waveforms. Section 3 shows neural network architectures used in two sorts of audio classification tasks. Section 4 describes the experimental setups for the two tasks and their results, and conclusions are drawn in Section 5.

2. Bit-Representation of Raw Audio Waveform

In this paper, we introduce to use bit-representations of a raw audio waveform. Two types of bit-representations of a raw audio waveform will be reviewed.

Generally, a sampled value is represented as a signed integer. For example, a sampled value varies from −32768 to 32767.
Figure 1: Extraction of bit-based pulse waveforms from a raw sound waveform.

Figure 2: Extraction of bit pattern image from a raw sound waveform.

+32767 when the quantization bit rate is 16-bit per sample. The previous study [18], which dealt with raw sound waveform on the music genre classification task, used raw-sampled value sequence as the input representation to the neural network. Unlike the previous study [18], our method transforms a sampled integer value into a bit vector, a set of bit values of the sampled value. For example, when a sampled integer is “12” (quantization bit rate is 8/sample), the bit vector of “12” is shown as “00001010”. The sequence of bit vector sequence proceeds into a neural network. The bit-representation transformation process can make the more rich representation of a raw audio waveform compared to the integer-based audio waveform representation. Therefore, we guess that convolution layers in a classifier can extract more effective feature maps from a bit-representation waveform to realize high-accurate audio detection/classification.

2.1. Bit-pulse extraction

Figure 1 shows one of the methods for bit-transformation from a raw audio waveform. Figure 1 shows an example when the quantization bit rate is 8/sample. In this transformation method, each sampled integer value at each time (the total number of samples is T) is expanded to a bit-representation vector, and then, bit-pulse waveforms are obtained by extracting each dimension/digit of the bit-representation vector sequence. Finally, the bit-pulse waveforms are inputted to a CNN layer of a classifier, in which one waveform is treated as one input channel.

The key idea of this method is to segment a raw audio waveform (integer value sequence) into several pulse waveforms. For example, in Fig. 1 we can get eight bit-pulse waveforms when the quantization bit rate is 8/sample. The bit-pulse waveform based on the Most Significant Bit (MSB) can capture the dynamic change (plus or minus) of the raw audio waveform. Like the MSB example, each bit-pulse waveform has any role for extracting features of audio. Therefore, we can accurately analyze the raw audio waveform using bit-representation transformation compared to without the transformation process.

2.2. Bit pattern image

The other method for transforming a raw audio waveform into a bit-representation-based sequence is to convert a waveform to a two-dimensional bit pattern image. Figure 2 also shows the other bit-representation method of a raw audio waveform. This figure is the same as Fig. 1 on the quantization bit rate. Unlike the transformation method is shown in Fig. 1, Figure 2 method makes a two dimensional bit pattern image from the bit-representation vector sequence. A bit pattern image is given to a two-dimensional convolution layer.

There are some ideas that an image representation from a raw audio waveform is used to a sound classification task [20, 21] or a speech recognition (key word spotting) task [22]. These studies used a sound spectrogram image from the audio waveform. Our proposed method is entirely different from the previous ones [20, 21, 22]. Our method does not lose the amount of information included in the raw audio waveform.

3. Neural Networks

In this paper, we use two sorts of neural networks depending on the evaluation tasks whose details will be explained in Section 4. This section explains these neural network architectures. There are infinite neural network architectures. Note that, therefore, the neural network architectures adopted in this paper are not necessarily optimum for the tasks. We should explore the optimal neural network architecture as future work. The main purpose of this paper is to investigate data representation for neural network-based classifiers.

3.1. Neural network for bit-pulse waveform

Figure 3 shows a CNN-Long Short-Term Memory (LSTM) network architecture which can accept the bit-representation waveform with multi-channel. This is a two-stage model that is widely used for classification or recognition of time sequential data. The CNN layers extract feature map from the input waveforms, and the LSTM layer with the FC layer classifies the input waveform using the history information.

Bit-pulse waveforms (multi-channel waveforms) are inputted to the first CNN layer, and then the third CNN layer outputs the feature map of 512 channel with size of (1, N). The number of N perfectly depends on the length (duration) of the input waveform. The feature map is divided into N small feature maps each of which has 512 dimensions. Each small feature map proceeds to the LSTM layer in sequential order. Finally, the output of the LSTM layer, when the final (N-th) small feature map is given to the LSTM, is transferred to the FC layer.

3.2. Neural network for bit pattern image

Figure 4 also explains a CNN-Bi-directional Gated Recurrent Unit (BiGRU) network architecture which can accept the bit pattern image of audio waveform. The model architecture is a little similar to the CNN-LSTM network described in Section
Figure 3: Neural network architecture for bit-pulse waveform of audio waveform.

4. Experiments

4.1. Experimental setup

4.1.1. Acoustic event detection task

We used the acoustic event classification database\textsuperscript{2} harvested from Freesound\textsuperscript{1} for the acoustic event detection task in this paper. The database consists of 28 events. The details of the dataset are described in Takahashi’s paper\textsuperscript{12}.

The total number of audio files is 5,223, and its duration is 768.4 minutes. The dataset is separated into two groups: one is for training the CNN-LSTM model, the other is for testing. The number of training and testing audio files are 3,702 and 1,521, respectively. All audio file is converted to 16,000Hz sampling rate with 16-bit quantization rate. The condition for the data separation is perfectly the same as the previous work\textsuperscript{12}. Therefore, we can compare the results between our methods and the Takahashi’s result\textsuperscript{12}.

The CNN-LSTM model used for this task is for the bit-pulse-based transformation, shown in Fig. 3. Therefore, an input audio waveform is separated into the 16 bit-based pulse waveforms (Fig. 1) before the audio proceeds to the neural network architecture.

![Diagram of Neural Network Architecture for Bit-Pulse Waveform](https://data.vision.ee.ethz.ch/cvl/aedataset)

Figure 4: Neural network architecture for bit pattern image of audio waveform.

Table 1: Training condition of the CNN-LSTM model for pulse waveform.

| Mini-batchsize | 64 |
|----------------|----|
| Num. of epochs | 500 |
| Kernel size at CNN layers | (130) |
| Stride at CNN layers | (1, 10) |
| Channels (1st, 2nd, 3rd CNN) | (128, 256, 512) |
| Activation at CNN layers | ReLU\textsuperscript{26} |
| Dropout | 0.2 at all the CNN layers |
| Batch normalization\textsuperscript{23} | No |
| Loss func. | Softmax cross entropy |
| Optimizer | MomentumSGD |
| Learning rate | 0.002 |

The training condition of the CNN-LSTM is described in Table 1. The hyper-parameters such as kernel size in each CNN layer is heuristically decided.

For this task, we compare our bit-representation method to the typical data representations: MFCC, power spectrum, and raw sound waveform (integer representation). The neural network model with power spectrum representation is the same as\textsuperscript{12}. Note that this model is different from Fig. 3. The 256-dimensional power spectrum and its Δ and ΔΔ are inputted to the CNN layer. In the case of MFCC and raw sound waveform, the network architectures are the same as Fig. 3 however, the parameters are optimized for each input type of feature.

4.1.2. Music/speech classification task

In the second music/speech binary classification task, we used the Marsyas dataset\textsuperscript{23}. It is also used in the previous work\textsuperscript{24}. The dataset consists of music and speech classes each of which has 64 audio files. Each audio file has a WAVE format and the sampling frequency rate is 22,050Hz with 16-bit quantization rate. Each audio file is segmented into 10 seconds audio
files, and they are converted into 8,000Hz sampling frequency with 16-bit quantization rate. Finally, the total number of audio files are 384, and they are separated into two groups; one group (269 files) is for training the CNN-BiGRU and the other (115) is for testing. A 10 seconds waveform is converted to an (80000, 16) size of the bit pattern image.

The neural network model used for the second task is for bit pattern image, shown in Fig. 4. Therefore, an input audio waveform (10 seconds file) is converted to an (80000, 16) size of bit pattern image before the audio file proceeds to the neural network. The training condition of the CNN-BiGRU is described in Table 2.

In this task, we investigate the domain dependency of the CNN-BiGRU model which accepts the bit pattern image of audio waveform. When we use a raw-level information of sound waveform, it is concerned that the neural network may be likely to be over-trained. Therefore, we collected audio files from the radio broadcast in Japan; the total number of radio files is 400 with music/speech label. Duration of each audio file is 10 seconds which is the same as in the case of the Marsyas dataset. And then, another classification model, whose architecture is the same as the CNN-BiGRU, is trained using the radio audio files. Because this model is trained from the out-of-domain dataset on testing, we can investigate the domain dependency of the model with the bit pattern image transformation method.

Moreover, we also investigate the noise robustness of the bit representation. We summarize the experimental conditions as follows:

C01 : The model is trained from the Marsyas dataset, it is evaluated on the 115 files (clean).

C02 : The same model as C01, however, it is evaluated on the 115 files to which 10dB while noise is injected.

C03 : The model is trained from the radio files, it is evaluated on the 115 files (clean).

4.2. Results and discussions
Table 3 shows the classification accuracy rates for each data representation on the acoustic event detection task. We compare the four sorts of data representations. As shown in Table 3, bit-representation of audio waveform got the best accuracy rate of 88.4% among the other (typical) data representation methods. We guess that the CNN layers can extract the more effective feature maps from the bit-pulse waveforms because bit-representation makes the CNN layers to detect a minute change of audio waveform easier than the typical data representation like power spectrum.

In the second task, the music/speech binary classification task, the classification accuracy rates are summarized in Table 4. Under the C01 condition (matched condition), the raw audio waveform was slightly better than the bit-representation. However, there are no significant differences between them. On the other hand, raw-level representations got much better performances against the power spectrum and MFCC.

As shown in Table 3 and 4, the bit-representation of the audio waveform is very useful for the audio classification tasks in the viewpoint of the feature map extraction and the robustness for noise and training conditions of the neural network. We also applied the bit-representation to the MNIST (hand-writing number classification) task using the VGG-16 network as an additional experiment for investigating the effectiveness of bit-representation of data. The bit-representations of the hand-writing images achieved 99.73% of classification accuracy against 99.58% from the gray-scale images. The bit-representation may be adequate for any image classification task.

5. Conclusions
This paper proposed the novel data representation method for audio classification using a deep neural network. The proposed method transformed a raw audio waveform (integer value sequence) into either bit-based multi-channel pulse waveforms or a bit pattern image depending on the classification task before inputting to a neural network. The experimental results on both the two tasks showed that the bit-representation of an audio waveform got the best results among the other data preprocessing methods.

In future work, we are going to explore the optimal neural network architecture for bit-representation waveform on various audio sets. Moreover, we also apply our method to other tasks such as speech recognition and speaker recognition.

Table 2: Training condition of the CNN-BiGRU model for bit pattern image.

| Mini-batchsize | 32 |
| Num. of epochs | 300 |
| Kernel size at CNN layers | 3x3 |
| Stride at CNN layers | 1 |
| Channels (1st, 2nd, 3rd CNN) | (269, 512, 1024) |
| Activation at FC and CNN layers | ReLU |
| Dropout | 0.2 at all the CNN layers |
| Batch normalization | No |
| Loss func. | Softmax cross entropy |
| Optimizer | MomentumSGD |
| Learning rate | 0.01 (half every 30 epochs) |

Table 3: Classification accuracy rates for each data representation on the acoustic event detection task.

| Data representation | C01 | C02 | C03 |
|---------------------|-----|-----|-----|
| Bit-representation (bit-pulse) | 92.8 | 88.4 | 81.2 |
| MFCC | 57.4 | 57.4 | 57.4 |
| Power spectrum | 80.3 | 80.3 | 80.3 |
| Raw waveform (integer) | 34.6 | 34.6 | 34.6 |

Table 4: Segment-based classification accuracy rates [%] for each data representation on the music/speech classification task.

| Data | Experimental conditions | C01 | C02 | C03 |
|------|-------------------------|-----|-----|-----|
| Bit-representation | 94.7 | 81.2 | 92.8 |
| Raw waveform (integer) | 95.4 | 71.8 | 83.0 |
| Power Spectrum | 90.5 | 90.5 | 90.5 |
| MFCC with RBM | 91% | — | — |

* This value is from Table 3 in [12].
6. References

[1] M. D. Plumbley, C. Kroos, J. P. Bello, G. Richard, D.P.W. Ellis, A. Mesaros, “Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018),” Tampere University of Technology, Laboratory of Signal Processing, 2018.

[2] J. Nam, K. Choi, J. Lee, S.-Y. Chou, Y.-H. Yang, “Deep Learning for Audio-Based Music Classification and Tagging,” IEEE Signal Processing Magazine, Vol.36, No.1, pp.41–51, 2019.

[3] G. Tzanetakis, P. Cook, “Musical genre classification of audio signals,” IEEE Trans. Speech and Audio Processing, Vol. 10, No. 5, pp. 293-302, 2002.

[4] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, T. Virtanen, “DCASE 2016 acoustic scene classification using convolutional neural networks,” Proc. of DCASE 2016 Workshop, pp. 95-99, 2016.

[5] S. H. Bae, I. Choi, N. S. Kim, “Acoustic scene classification using parallel combination of LSTM and CNN,” Proc. DCASE 2016 Workshop, pp. 1115, 2016.

[6] Y. Guo, M. Xu, J. Wu, Y. Wang, K. Hoashi, “Multi-scale convolutional recurrent neural network with ensemble method for weakly labeled sound event detection,” Proc. of DCASE 2018 Workshop, pp. 98–102, 2018.

[7] H. G. Kim, J. Y. Kim, “Acoustic Event Detection in Multichannel Audio Using Gated Recurrent Neural Networks with HighResolution Spectral Features”, ETRI Journal, Vol.39, No.6, pp. 832–840, 2017.

[8] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. L. Roux, K. Takeda, “Duration-Controlled LSTM for Polyphonic Sound Event Detection,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 25, No. 11, pp. 2059–2070, 2017.

[9] K. Choi, G. Fazekas, M. Sandler, K. Cho, “Convolutional recurrent neural networks for music classification,” Proc. of ICASSP 2017, pp. 2392-2396, 2017.

[10] A. Mesaros, T. Heittola, T. Virtanen, “TUT database for acoustic scene classification and sound event detection,” Proc. of the 24th European Signal Processing Conference (EUSIPCO 2016), pp. 11281132, 2016.

[11] K. J. Piczak, “ESC: Dataset for environmental sound classification,” Proc. of the 23rd ACM international conference on Multimedia, pp. 10151018, 2015.

[12] N. Takahashi, M. Gygli, B. Pfister, L. V. Gool, “Deep Convolutional Neural Networks and Data Augmentation for Acoustic Event Recognition,” Proc. of INTERSPEECH 2016, pp. 2982–2986, 2016.

[13] J. F. Gemmeke, D. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, M. Ritter, “Audio Set: An ontology and human-labeled dataset for audio events,” Proc. of ICASSP 2017, pp. 776–780, 2017.

[14] M. Defferrard, K. Benzi, P. Vandergheynst, X. Bresson, “FMA: A dataset for music analysis,” Proc. of Int. Society for Music Information Retrieval Conf., pp. 316323, 2017.

[15] E. Cakir, E. C. Ozan, T. Virtanen, “Filterbank learning for deep neural network based polyphonic sound event detection,” Proc. of the 2016 International Joint Conference on Neural Networks (IJCNN), pp.3399–3406, 2016.

[16] E. Cakir, T. Virtanen, “End-to-End Polyphonic Sound Event Detection Using Convolutional Recurrent Neural Networks with Learned Time-Frequency Representation Input,” Proc. of the 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–7, 2018.

[17] J. Lee, J. Park, L. Kim, J. Nam, “Sample-level Deep Convolutional Neural Networks for Music Auto-tagging Using Raw Waveforms,” Proc. of the 14th Sound & Music Computing Conference, pp.220-226, 2017.

[18] J. Lee, J. Park, K. L. Kim, J. Nam, “SampleCNN: End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification,” Applied Sciences, Vol.8, No.1, 150, 2018.

[19] T. N. Sainath, R. J. Weiss, A. W. Senior, O. Vinyals, “Learning the speech front-end with raw waveform CLDNNs,” Proc. of INTERSPEECH2016, pp.1–5, 2015.

[20] Y.M.G. Costa, L. S. Oliveirab, C. N. Silla Jr., “An evaluation of Convolutional Neural Networks for music classification using spectrograms,” Applied Soft Computing, Vol.52, pp. 2838, 2017.

[21] J. Dennis, H. D. Tran, E. S. Chng, “Analysis of Spectrogram Image Methods for Sound Event Classification,” Proc. of INTERSPEECH 2014, pp.2533–2537, 2014.

[22] S. K. Gouda, S. Kanetkar, D. Harrison, M. K Warmuth, “Speech Recognition: Key Word Spotting through Image Recognition,” arXiv:1803.03759, 2018.

[23] G. Tzanetakis, P. Cook, “MARSYAS: a framework for audio analysis,” Organised Sound, Vol.4, No.3, pp.169–175, 2000.

[24] A. Pikrakis, S. Theodoridis, “Speech-music discrimination: A deep learning perspective,” Proc. of the 22nd European Signal Processing Conference (EUSIPCO 2014), 5 pages, 2014.

[25] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv:1502.03167v3, 2016.

[26] X. Glorot, A. Bordes, Y. Bengio, “Deep Sparse Rectifier Neural Networks,” Proc. of the 14th International Conference on Acritical Intelligence and Statistics, pp. 315–323, 2011.

[27] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. of the IEEE, Vol. 86, No. 11, pp. 2278–2324, 1998.

[28] K. Simonyan, A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Proc. of ICLR 2015, 2015.