Analysis of Bit Sequence Representation for Sound Classification

Yikang Wang*, Masaki Okawa*, and Hiromitsu Nishizaki*
* University of Yamanashi, Kofu, Japan
E-mail: {wwm1995, yukari_phantasm, nisizaki}@alps-lab.org Tel/Fax: +81-55-220-8361

***Abstract*—In sound classification, commonly used speech perceptual features, such as the Mel-frequency cepstral coefficient and the Mel-spectrogram, ignore information other than the frequency features in raw waveforms. We cannot conclude that these discarded parts are meaningless. To avoid missing information in the time series, we previously proposed the bit sequence representation, which maintained the temporal characteristics of the sound waveform and improved its performance over the original waveform. The present study validated our findings on three datasets, namely two datasets for music/speech classification and one for English speech classification. We also compared the classification performances when the features were not preprocessed with that when the maximum amplitude was restricted. As a result, we found that appropriately limiting the maximum amplitude is effective in improving the classification performance.**

I. Introduction

Sound classification includes tasks, such as speech recognition (ASR), acoustic event detection (AED), acoustic scene classification (ASC), and music/speech discrimination. In these tasks, sound perceptual feature extraction processes are generally required before using classification algorithms, such as the commonly used Mel-frequency cepstral coefficient (MFCC) [1], [2], [3], Mel-spectrogram, perceptual linear predictive coefficient [4], and power-normalized cepstral coefficients [5]. The recognition performance of different features differs depending on specific tasks. For example, in the task of ASC, P. Tilak et al. [6] used an end-to-end approach to perform a direct classification with raw waveforms, which outperformed the MFCCs. In the speech recognition task, experiments demonstrated the MFCCs like features have a significantly better recognition performance [3], [5]. In other words, we need a generic feature that can be used in different tasks.

After investigation, we found that all the commonly used feature extraction approaches only analyzed the frequency component and discarded the information in the acoustic wave other than the frequency and the amplitude. However, we do not assert that the discarded portion contains no useful information.

Under the assumption that the information discarded by the frequency features is useful, we abandoned using the Mel-spectrogram and the MFCCs in our previous work [7]. Instead, we used a temporal-based feature, called bit sequence representation, which deformed from the original acoustic sample, to propose a deep neural network-based sound classification approach. We used two different bit sequence representations to prove that the feature has a higher recognition performance than the MFCCs and a good robustness in both AED and music/speech discrimination tasks.

This paper discusses the validity of bit sequence representation in music/speech discrimination tasks and an English utterance classification task through two sets of experiments. In the first set of experiments, we compare the differences in the classification performances of the three datasets under a similar network structure using the original waveform, the bit sequence representation of the acoustic waveform without any pre-processing for end-to-end recognition, and the pre-processed MFCCs features. In the second set of experiments, we compare the classification performances of the bit sequence representation and the raw waveforms by adjusting the waveform's amplitude maximum.

Based on the preceding experiments, we investigate and analyze the effectiveness of the end-to-end approach using bit sequence representation in different tasks and find the effect of adjusting the maximum amplitude of the waveform on the classification performance. The contributions of this paper are as follows:

- This study presents three different forms of bit sequence representations (int16, float16, and float32) that improve the feature performance.
- This feature does not lose information from the waveform. Compared to the MFCCs, which require complex calculations and convolution with filters, the bit sequence representation only needs to change the method of reading the raw waveforms.
- By plotting the spectrograms, the significance of the high and low-bit features in the bit sequence representation is confirmed to lay the foundation for further research.
- The bit sequence representation has the same performance level as the MFCCs in the English utterance classification task. Combined with the conclusions of our previous studies, the bit sequence representation has better cross-tasking capabilities than MFCCs.
- The experiments show that restricting the maximum amplitude of the raw waveforms can effectively improve the recognition effect, regardless of which feature is used.

The remainder of this paper is organized as follows: the next section describes three forms of bit column performance and the neural network architectures used for two types of audio

classification tasks; Section 3 describes the dataset details, experimental setups for the two tasks, and the results; and Section 4 provides our conclusions.

II. NEURAL NETWORK CLASSIFIER BASED ON THE BIT SEQUENCE REPRESENTATION

In the recent years, DNNs have been very successful at ASR [2], [8], [9], ASC [10], [11], [12], AED [13], [14], [15], and music/speech discrimination [16], [17]. More recently, end-to-end approaches have become increasingly popular. Many sound classification systems using end-to-end approaches have achieved unprecedented results [18], [19]. Our proposed bit sequence representation only changes the method of reading data without losing any information of raw waveforms, does not require pre-emphasis processing, and can be used as an input feature for the end-to-end approach.

Three classification tasks are handled in this study: two music/speech discrimination tasks and an English utterance classification task. In our previous work [7], we proposed two types of bit sequence representations of a raw audio waveform and provided two network structures inspired by convolutional (Conv.), long short-term memory (LSTM), fully-connected deep neural networks [20]. We continued to use these network structures herein with selections and adjustments based on the tasks and chose three types of bit sequence: a 16-bit sequence of integers; a 32-bit sequence of single-precision floating-point numbers; and a 16-bit sequence of half-precision floating-point numbers.

A. Bit Sequence Representation

The bit sequence representation of a sound waveform is expressed not as an integer value but as a bit sequence, and is represented as two-dimensional data of "number of samples" and "digits of the bit sequence." We can perform convolutional processing like an image by representing sound waves in two dimensions.

B. Structure of the Classifier

Fig. 1 shows an overview of the music/speech discrimination experiment model with a 32-bit input bit length. This sound classification model has a structure that extracts features from the bit sequence of a raw audio waveform using three-layer CNNs. The classification is performed in the subsequent bi-directional GRU [21] layer and fully-connected layer. The model uses softmax cross-entropy as the loss function, Adam as the optimizer, with an initial learning rate of 0.0002, and the number of output classes depends on the dataset used in the experiment (2 or 6). Other details of the model parameters are indicated in Table I. Fig. 2 shows a schematic diagram of the model used in the English utterance classification experiments. This model has a structure that is very similar to that of the abovementioned sound classification model, in which the features are extracted from the bit sequences at three Conv. layers, and then classified at the LSTM [22] layer and the fully-connected layer, and we use the same loss function and optimizer as in the speech/music discrimination experiment

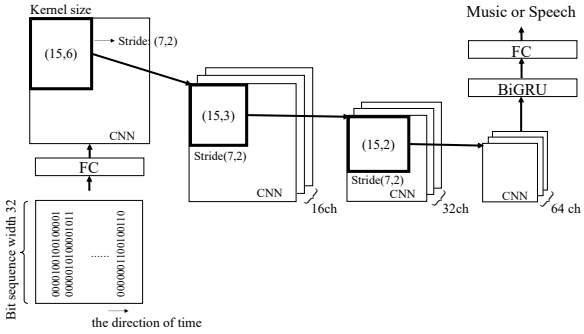

Fig. 1. Music/Speech classifier.

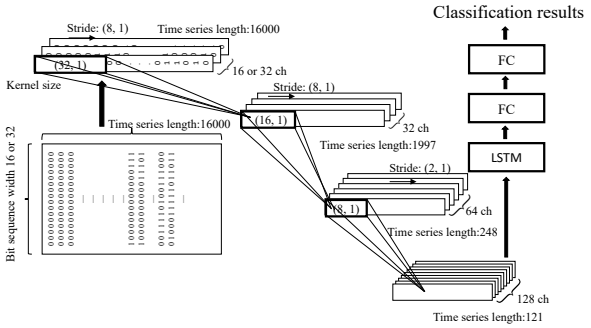

Fig. 2. English utterance classifier.

model. However, the convolution method in the Conv. layers is different. This model uses the bit width as the channel size in the convolution of the audio bit sequences; thus, each digit of the bit sequence is convolved. The details of the model parameters are indicated in Table II.

III. EXPERIMENTS

A. Dataset Details

The first dataset is called the GTZAN dataset [23] in Marsyas. This dataset consists of two music and speech classes. Each class comprises of 64 audio files. Each audio file is in a 30 s WAV format (22,050 Hz ampling frequency, single channel and 16 bits per sample). In our experiments, we divided and downsampled the dataset into a 10 s WAV format (8,000 Hz sampling frequency, single channel, and 16 bits per sample) and used 384 wave-formatted files. In the dataset, 70% of the data was used for the training, and the others were used for the evaluation.

The other dataset is a radio dataset collected by ourselves for music/sound discrimination tasks. The collected data were from actual Japan broadcasted programs. This dataset consists of 17,973 training audio files and 115 evaluation audio files. Each file is in a 10 s WAV format (8,000 Hz sampling frequency, single channel, and 16 bits per sample). For the experiments, the radio data were divided into six categories: music, music and speech, speech, laughter, silence, and environmentally sound. We assigned one of the six labels to an audio file.

We also used the Google Speech Commands Dataset ver. 2 [24] to investigate the applicability of the bit sequence

TABLE I
STRUCTURE DETAILS OF MUSIC/SPEECH CLASSIFIER.

Layers	Features			
	Bit (int16/float16)	Bit (float32)	Raw waveforms	MFCC
Dropout	0.1	0.1	-	0.1
Dense	16	32	-	39
Dropout	0.2	0.2	0.2	0.2
Reshape	(80,000, 16, 1)	(80,000, 32, 1)	(80,000, 1, 1)	(1,001, 39, 1)
Conv. 1	(15, 6), (7, 2), 16	(15, 6), (7, 2), 16	(15, 1), (7, 1), 16	(15, 12), (7, 6), 16
Dropout	0.2	0.2	0.2	0.2
Conv. 2	(15, 3), (7, 1), 32	(15, 3), (7, 2), 32	(15, 1), (7, 1), 32	(15, 4), (7, 2), 32
Dropout	0.2	0.2	0.2	0.2
Conv. 3	(15, 1), (7, 1), 64	(15, 2), (7, 2), 64	(15, 1), (7, 1), 64	(15, 1), (7, 1), 64
Reshape	(231, 256)	(231, 192)	(231, 64)	(1, 64)
BiGRU	256	256	256	256
Dense (softmax)	2 or 6	2 or 6	2 or 6	2 or 6

TABLE II
STRUCTURE DETAILS OF ENGLISH UTTERANCE CLASSIFIER.

Layers	Features		
	Bit	Raw waveforms	MFCC
Mask	-1	0	-
Reshape	(16,000, 1, 16)	(16,000, 1, 1)	(197, 39, 1)
Conv. 1	(32, 1), (8, 1), 32	(32, 1), (8, 1), 32	(25, 1), (6, 1), 128
Dropout	0.4	0.4	0.4
Conv. 2	(16, 1), (8, 1), 64	(16, 1), (8, 1), 64	(6, 1), (2, 1), 256
Dropout	0.4	0.4	0.4
Conv. 3	(8, 1), (2, 1), 128	(8, 1), (2, 1), 128	-
Dropout	0.4	0.4	-
Reshape	(121, 128)	(121, 128)	(12, 256)
LSTM	128	128	128
Dropout	-	-	0.4
Dense	128	128	128
Dropout	-	-	0.4
Dense (softmax)	35	35	35

representation of a sound wave to the utterance classification asides from music/speech discrimination. The dataset consists of 35 different types of utterances, which totaled to 105,829 audio files. Each audio data file is in the WAV format (16,000 Hz sampling frequency, single channel, and 16 bits per sample) with less than 1 s duration. Approximately 10% of the audio files was less than 1 s, and we adjusted the data to be 1 s. In the experiment, 84,843 sound files were used for learning; 11,005 sound files were used for the evaluation, and the remaining files were used for the verification. The audio files were distributed such that the same speaker will not be included in the training, evaluation, or verification set at the same time.

The GTZAN dataset is referred to herein as the **D1**. The self-collected radio dataset is referred to as **D2**. The Google Speech Commands Dataset used for the word utterance classification is referred to as **D3**.

We conducted two sets of experiments to investigate the effectiveness of the end-to-end classification method, which directly classifies bit sequences without pre-processing, and the influence of the sound wave amplitude.

B. Experimental Setup

1) End-to-end experiments with bit sequence representation: We conducted the first experiment to investigate the effectiveness of the end-to-end classification approach, which directly classifies bit sequences without pre-processing by comparing three different bit sequence representations, MFCCs with pre-processing, and the raw waveform input without conversion to bit sequences. The evaluation scale was based on the classification accuracy of the evaluated data. Three kinds of bit sequence representations referred to an integer value converted to a 16-bit sequence, a single-precision floating-point value converted to a 32-bit sequence, and a half-precision floating-point value converted to a 16-bit sequence.

2) Amplitude variation experiment: The second experiment was conducted to investigate the influence of the waveforms amplitude change on the bit sequence input by varying the waveform amplitudes and evaluating each classification using the accuracy rate. The amplitudes were varied by changing the maximum amplitude of each waveform such that the amplitudes of the entire waveform were aligned. The maximum value of the amplitude to be adjusted was 2^n, where n ranges from 0 to 15. The maximum sound waveform value was set to a power of 2 to make it easier to understand how much bit length affects the sound classification considering that the bit length representing 2^n changes as n changes. All values after the decimal point were rounded down to an integer value, when the waveform amplitude varied. Thus, the waveform amplitudes were aligned and converted into three different bit

TABLE III
CLASSIFICATION ACCURACIES [%] FOR VARIOUS SORTS OF INPUT FORMATS.

Input Feature	Dataset		
	D1	D2	D3
int16	**96.5**	89.6	85.0
float16	94.7	88.7	84.3
foat32	**96.5**	87.8	84.9
Raw waveforms	85.1	88.7	22.7
MFCC	93.9	**90.4**	**91.1**

sequences used as input to the sound classification model. We also investigated a method of classifying raw waveforms to compare them with the bit sequence.

C. Experimental Results and Visualization Analysis

Table III shows the classification accuracy for each dataset by each input representation.

The bit sequence accuracy for D1 was higher for MFCCs and the raw waveforms. On the contrary, the classification accuracy for D2 and D3 was higher for MFCCs than the bit sequence representation. However, no significant difference was found in the classification accuracy between the MFCCs and the bit sequences in all datasets; therefore , the bit sequences were an effective input for the sound classification. No significant change in the classification accuracy was observed as the bit sequence was changed; thus, no problem was encountered in using any type of bit sequence representation, and we can achieve a fair classification accuracy. Furthermore, the accuracy of the bit sequence representation in D1 was better than that of the MFCCs. In other words, the bit sequences can achieve a higher classification accuracy than the preprocessing method depending on the tasks to be handled. However, in D3, the classification accuracy of the direct input of raw waveforms was significantly lower than that of the bit sequence representation.Even in cases where classifying raw waveforms was difficult, depending on the task, a highly accurate sound classification can be achieved by converting the waveforms to bit sequences.

Table IV shows the classification accuracy for each input type when the maximum amplitude was aligned. The classification accuracy in all datasets was improved by aligning the maximum amplitude of the sound waves compared to the results shown in Table III, which shows the classification accuracy without aligning the maximum amplitude.

This improvement restricted the maximum value of all sound files during sound classification to prevent some parts of the sound files from being outliers with extremely large amplitudes, thereby reducing misclassification results. Therefore, aligning the maximum amplitude of the sound waveform is considered to be helpful when using the raw waveforms as the input and can be used as a new normalization method of raw waveforms. Moreover, the classification accuracy did not significantly change when the amplitude of the bit sequence representation was aligned, indicating that no significant difference existed in the sound classification, regardless of which bit sequence representation was used. However, the classification

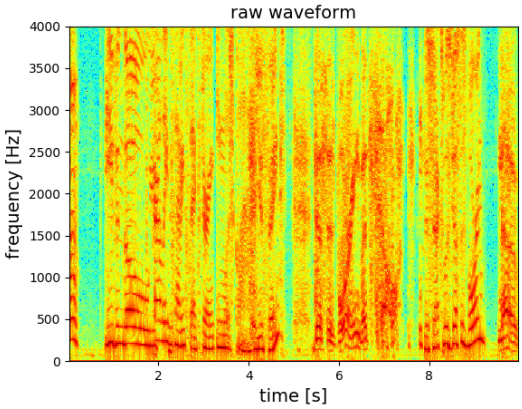

Fig. 3. Spectrogram of a raw waveform.

accuracy was markedly lower when the maximum amplitude was 1 or 2 than when the maximum amplitude was four or more. When converting an integer to a bit sequence, an amplitude of 2 can only be expressed in two bits. A high classification performance cannot be achieved in any dataset with only two bits of information. Therefore, more than 3 bits of information must be used to realize accurate sound classification.

We visualized the spectrogram of each digit of the bit sequence to check what information is contained in a bit sequence representation. For comparison, we also present herein the spectrogram of the untransformed raw waveforms. Fig. 3 shows the spectrogram of the raw waveforms. Fig. 4 illustrates the spectrogram of the sound waveform converted to a 16-bit integer string. Fig. 5 displays the spectrogram of the sound waveform converted to a 16-bit half-precision floating-point string, where ch is the digit of each bit sequence. The higher the value, the higher the bit sequence. For example, ch0 indicates the lowest bit, whereas ch15 indicates the highest bit. If we convert the sound waveform into a 16-bit string of integers (Fig. 4), we can confirm that the sign bit, which is the highest-order bit, and the three high-order bits of ch10 to ch12 have features that are similar with those in Fig. 3. Therefore, the upper three bits of the bit sequence representation were found to contain features similar to the raw waveforms. As a result, sound classification with high accuracy is challenging when a bit sequence is used for the sound classification, unless it contains the upper 3 bits.

This tendency is shown in Table IV to be true not only for 16-bit integers, but also for floating-point numbers. In other words, a bit sequence with a maximum amplitude of approximately 3 bits is essential for classification. Fig. 5 shows that the upper bits of ch10 to ch15 had characteristics similar to those of Fig. 3, which depicts a spectrogram of the raw waveforms. The highest-order bits of the floating-point number were the sign bit and the exponent part containing the information necessary for the sound classification. Table IV shows that the information necessary for the sound classification were included in the sign bit and the exponent

TABLE IV
CLASSIFICATION ACCURACIES [%] FOR THE AMPLITUDE-RESTRICTED AUDIO.

Input Feature		Maximum amplitude value															
		1	2	4	8	16	32	64	128	256	512	1024	2048	4096	8192	16384	32678
D1	int16	59.6	93.9	97.4	97.4	97.4	97.4	98.2	98.2	97.4	98.2	**99.1**	**99.1**	98.2	96.5	**99.1**	96.5
	float16	57.9	95.6	97.4	97.4	97.4	96.5	97.4	**99.1**	**99.1**	97.4	97.4	**99.1**	98.2	97.4	97.4	98.2
	float32	63.2	93.0	97.4	97.4	**99.1**	96.5	98.2	**99.1**	96.5	98.2	97.4	97.4	97.4	98.2	97.4	97.4
	Raw waveforms	57.9	87.7	96.5	97.4	97.4	98.2	**99.1**	98.2	98.2	95.6	96.5	95.6	96.5	84.2	78.9	83.3
D2	int16	60.9	80.9	80.0	88.7	**90.4**	89.6	87.8	88.7	87.0	87.8	88.7	87.8	89.6	89.6	87.8	88.7
	float16	87.0	87.0	87.0	86.1	85.2	87.8	87.8	87.8	86.1	87.0	**90.4**	87.0	86.1	86.1	89.6	86.1
	float32	86.1	86.1	86.1	86.1	87.0	87.8	**91.3**	87.8	87.0	87.8	85.2	86.1	87.0	87.0	88.7	87.0
	Raw waveforms	60.0	87.0	86.1	**92.1**	91.3	87.8	87.8	87.8	87.0	87.0	87.0	88.7	87.0	88.7	87.8	87.0
D3	int16	3.8	49.7	73.1	83.4	86.2	**87.8**	87.2	87.6	86.9	85.9	86.6	84.3	86.3	85.1	86.8	82.0
	float16	3.7	49.8	73.8	84.3	86.2	89.2	**89.3**	89.1	87.5	86.4	87.9	84.2	87.1	84.8	86.5	85.2
	float32	3.8	47.0	73.8	82.6	86.6	88.3	88.0	**89.0**	87.5	87.0	86.7	85.9	84.5	85.2	85.7	83.9
	Raw waveforms	4.5	49.9	76.9	85.8	88.5	89.7	**90.1**	89.8	89.1	87.2	85.4	78.0	64.5	33.5	16.8	7.7

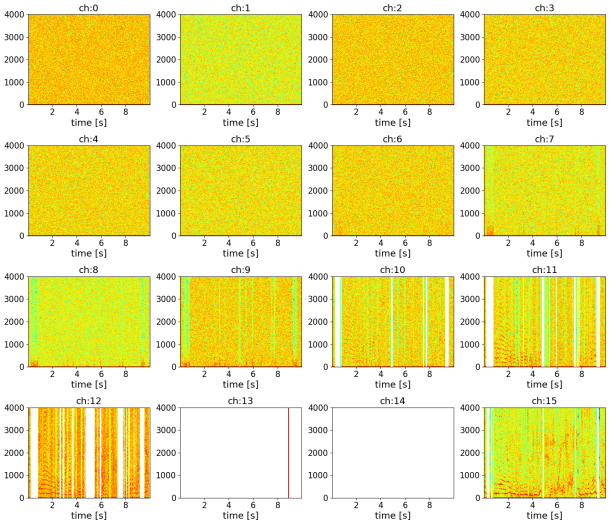

Fig. 4. Spectrograms of each bit sequence (int16 representation) of a raw waveform.

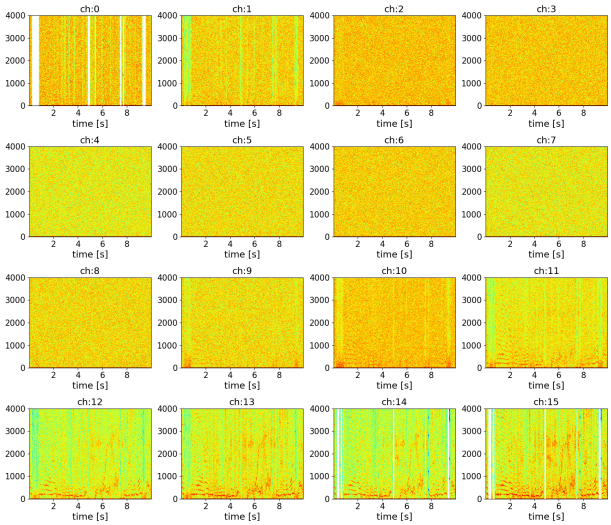

Fig. 5. Spectrograms of each bit sequence (float16 representation) of a raw waveform.

part when the maximum amplitude was 4 or more, even in a floating-point number. These results show that the sound wave was normalized by aligning the maximum amplitude in sound classification, which was effective in classifying sound with raw waveforms. However, almost no difference in the sound classification was found no matter which bit sequence is used. When the maximum amplitude value was restricted in the 16-bit integer representation, at least three bits must be considered to achieve a high sound classification accuracy.

IV. CONCLUSIONS

This study aimed to analyze sound classification characteristics using bit sequence representations as the input. We conducted two experiments using an end-to-end approach to investigate the accuracy of the music/speech classification and English utterance tasks and the effects of aligning the amplitude maximums of the sound waves on classification.

In the end-to-end experiments using bit sequence representations, the same classification accuracy was confirmed by using any type of bit sequence compared with the pre-processed MFCCs. The experimental results for D1 showed a bit sequence representation with a higher classification accuracy than the MFCCs. Bit sequence representation can be applied to various sound classification tasks. Moreover, the sound classification accuracy is improved by aligning the maximum amplitude of the raw waveforms. However, achieving a high-accuracy sound classification is difficult if the maximum amplitude of the raw waveforms is converted to a 16-bit integer string. We also confirmed herein that achieving a high-accuracy classification is difficult when the maximum amplitude is tiny.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (B) Grant Number 17H01977. Besides, a part of this work was also supporoted by Hoso Bunka Foundation.

REFERENCES

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,

IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366, 1980.

[2] G. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, 2012.

[3] A. F. Abka and H. F. Pardede, "Speech recognition features: Comparison studies on robustness against environmental distortions, In *Proceedings of the 2015 International Conference on Computer, Control, Informatics and Its Applications*, pp. 114-119, 2016.

[4] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech, *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738-1752, 1990.

[5] C. Kim, and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24, pp. 1315-1329, 2016.

[6] P. Tilak, A. Agrawal and V. Ramasubramanian, "Acoustic scene classification using deep CNN on raw-waveform technical Report, In *Proceedings of DCASE2018*, 4 pages, 2018.

[7] M. Okawa, T. Saito, N. Sawada and H. Nishizaki, "Audio classification of bit-representation waveform, In *Proceedings of INTERSPEECH 2019*, pp. 25532557, 2019.

[8] C. Lscher, E. Beck, K. Irie, M. Kitza, W. Michel, A. Zeyer, et al., "RWTH ASR Systems for LibriSpeech: Hybrid vs Attention–w/o Data Augmentation, *arXiv preprint* arXiv:1905.03072, 2019.

[9] K. H. Lee, W. H. Kang, H. Lee, and N. S. Kim, "Stochastic DNN-HMM Training for Robust ASR, In *Proceedings of APSIPA ASC 2018*, pp. 177-182, 2018.

[10] H. Chen, Z. Liu, Z. Liu, P. Zhang, and Y. Yan, "Integrating the Data Augmentation Scheme with Various Classifiers for Acoustic Scene Modeling, *arXiv preprint* arXiv:1907.06639, 2019.

[11] J. Cho, S. Yun, H. Park, J. Eum, and K. Hwang, "Acoustic Scene Classification Based on a Large-margin Factorized CNN, *arXiv preprint* arXiv:1910.06784, 2019.

[12] L. Pham, H. Phan, T. Nguyen, R. Palaniappan, A. Mertins, and I. McLoughlin, "Robust Acoustic Scene Classification using a Multi-Spectrogram Encoder-Decoder Framework, *arXiv preprint* arXiv:2002.04502, 2020.

[13] Y. Guo, M. Xu, Z. Wu, J. Wu and B. Su, Multi-Scale Convolutional Recurrent Neural Network with Ensemble Method for Weakly Labeled Sound Event Detection, In *Proceedings of 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 1-5, 2019.

[14] M. Mulimani, A. B. Kademani, and S. G. Koolagudi, "A Deep Neural Network-Driven Feature Learning Method for Polyphonic Acoustic Event Detection from Real-Life Recordings, In *Proceedings of ICASSP 2020*, pp. 291-295, 2020.

[15] J. Wang, and S. Li, "Comparing the influence of depth and width of deep neural network based on fixed number of parameters for audio event detection, In *Proceedings of ICASSP 2018*, pp. 2681-2685, 2018.

[16] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification, In *Proceedings of ICASSP2017*, pp. 23922396, 2017.

[17] G. K. Birajdar, and M. D. Patil, "Speech/music classification using visual and spectral chromagram features, *Journal of Ambient Intelligence and Humanized Computing*, 11.1, pp. 329-347, 2020.

[18] T. Tanaka, R. Masumura, T. Moriya, T. Oba, and Y. Aono, "A Joint End-to-End and DNN-HMM Hybrid Automatic Speech Recognition System with Transferring Sharable Knowledge, In *Proceedings of INTERSPEECH 2019*, pp. 2210-2214, 2019.

[19] C. Kim, M. Shin, A. Garg, and D. Gowda, "Improved Vocal Tract Length Perturbation for a State-of-the-Art End-to-End Speech Recognition System, In *Proceedings of INTERSPEECH 2019*, pp. 739-743, 2019.

[20] T. N. Sainath, R. J. Weiss, A. W. Senior, and O. Vinyals, "Learningthe speech front-end with raw waveform CLDNNs, In *Proceedings of INTERSPEECH 2016*, pp.15, 2015.

[21] K. Cho, B.V. Merrienboer, . Glehre, D. Bahdanau, F. Bougares, H. Schwenk, et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation, In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.17241734, 2014.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory, *Neural Computation*, vol.9, 17351780, 1997.

[23] G. Tzanetakis and P. Cook, "MARSYAS: A framework for audio analysis, *Organised Sound*, vol.4, pp. 169175, 2000.

[24] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition, *arXiv preprint* arXiv:1804.03209 , 2018.