SHUFFLING AND MIXING DATA AUGMENTATION FOR ENVIRONMENTAL SOUND CLASSIFICATION

Tadanobu Inoue[1], *Phongtharin Vinayavekhin*[1], *Shiqiang Wang*[2], *David Wood*[2], *Asim Munawar*[1], *Bong Jun Ko*[2], *Nancy Greco*[2], *Ryuki Tachibana*[1]

[1] IBM Research, Tokyo, Japan, {inouet, pvmilk, asim, ryuki}@jp.ibm.com
[2] IBM Research, Yorktown Heights, NY, USA, {wangshiq, dawood, bongjun_ko, grecon}@us.ibm.com

ABSTRACT

Smart speakers have been recently adopted and widely used in consumer homes, largely as a communication interface between human and machines. In addition, these speakers can be used to monitor sounds other than human voice, for example, to watch over elderly people living alone, and to notify if there are changes in their usual activities that may affect their health. In this paper, we focus on the sound classification using machine learning, which usually requires a lot of training data to achieve good accuracy. Our main contribution is a data augmentation technique that generates new sound by shuffling and mixing two existing sounds of the same class in the dataset. This technique creates new variations on both the temporal sequence and the density of the sound events. We show in DCASE 2018 Task 5 that the proposed data augmentation method with our proposed convolutional neural network (CNN) achieves an average of macro-averaged F1 score of 89.95% over 4 folds of the development dataset. This is a significant improvement from the baseline result of 84.50%. In addition, we also verify that our proposed data augmentation technique can improve the classification performance on the Urban Sound 8K dataset.

Index Terms— Domestic Activities, Data Augmentation, Deep Learning, Convolutional Neural Network

1. INTRODUCTION

In recent years, there is an increasing popularity in installing smart speakers in a home environment due to its capability to interact and activate home appliances through its voice interface. The low cost of these smart speakers encourages the use of more than one device to cover a larger area of a home. The technology in smart speakers, Micro Electro Mechanical Systems (MEMS) array microphones, can be additionally used for monitoring sounds other than human voice. The smart speaker capability can be adapted through machine learning to monitor and detect human activities in daily life routine [1, 2].

We consider human activity monitoring and detection as a multi-class classification problem [3]. The task is to identify acoustic scenes and events using environmental sounds [4]. A supervised machine learning technique, deep learning based on convolutional neural networks (CNNs) to be specific, is used as a classifier. CNNs have been widely used in acoustic scene classification tasks due to their promising performance [5, 6, 7, 8, 9].

It is well-known that deep learning requires a large amount of data to train an accurate model. To increase the amount of training data and reduce overfitting, numerous data augmentation methods have been studied in the acoustic literature. Some musically inspired deformations such as pitch shifting and time stretching are adopted to augment training sound data [6, 10]. Jaitly and Hinton [11] showed that the data augmentation based on vocal tract length perturbation (VTLP) is effective to improve the performance of automatic speech recognition (ASR). Takahashi *et al.* [12] mixed two sound sources within the same class to generate a new sound. Tokozume *et al.* [13] proposed a method to mix two sound sources from different classes. Both labels and sounds are mixed and referred to as between-class data. They train the model solely using the generated data without using the original data. Zhang *et al.* [14] proposed a similar approach to use between-class data, but they also use mixing of sounds from the same class in the training.

In these previous works, the temporal order of the sound events is kept and does not generate new variations on the sound sequence. In addition, mixing by linearly combining two sounds [12, 13, 14] usually increases the number of sound events (event density) which could introduce bias in the model.

In contrast to the existing approaches mentioned above, we propose a method that increases the variation in the training samples on *both the temporal sequence and event density* (the number of the sound events in a time period) of the sound events. Our proposed method can both increase and decrease the density of sound events, while keeping the overall average density of events the same as in the original sound and thus introducing *no bias* to the model. Our proposed method augments input acoustic data by combining sounds from two sound sources of the same class. Each sound source is di-

https://doi.org/10.33682/wgyb-bt40

vided into multiple segments, and the new sound is generated by shuffling and mixing these segments of two sounds from different sound sources. This is based on our observation that environmental sound is generally composed of background sound and events; each event often occurs discretely in a sound sequence without any temporal relation with others. The fact that mixing is randomized keeps the overall average event density the same as in the original sounds.

We conduct experiments with two acoustic datasets. First, we applied the proposed method to DCASE 2018 Task 5 dataset [3, 15], which includes sounds in a home environment. Our proposed method alleviates the effect of unbalanced classes in the dataset, and significantly increases the classification performance (F1 score) and is a main ingredient in building the system [16] that won the challenge[1]. Second, the proposed method is applied to Urban Sound 8K dataset [17], where the results show that our proposed method produces comparable results to other data augmentation techniques that are designed for this dataset.

This paper starts by describing the proposed data augmentation technique in Section 2. Experimental results on two datasets are described in Section 3. Finally, the conclusion is given in Section 4.

2. SHUFFLING AND MIXING DATA AUGMENTATION

In this section, we introduce the shuffling and mixing data augmentation to increase variation of training samples for training a deep learning model. We augment sound data based on two assumptions.

First, based on our observation, we assume that environmental sound is generally composed of background sounds and foreground event sound. The foreground events often occur discretely and have no temporal relation with each other. For example, let us consider *eating* sounds as shown in Fig. 1. Foreground event sounds can be caused by the sound of dishes or kitchen utensils; however, these events are temporally independent of each other. In other words, even when the order of these sound events is swapped, the sound can still be categorized as *eating*. Therefore, it is possible to generate a new sound clip by shuffling the order of sound segments. Second, we assume that mixing two sound sources within the same class results in a new sound in the same category. This assumption has been also used in previous works [12, 14].

Based on these two assumptions, we propose a simple but effective data augmentation technique, which is comprised of two steps: (a) shuffling, and (b) mixing two sounds of the same class, as shown in Fig. 2. To simplify the explanation, let us consider two sound clips of the same class and the

[1]http://dcase.community/challenge2018/task-monitoring-domestic-activities-results

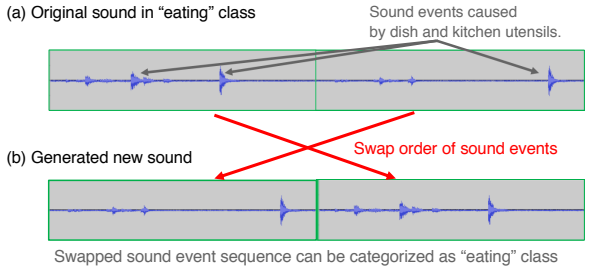

Figure 1: Swapping the order of sound events creates a sound in the same class.

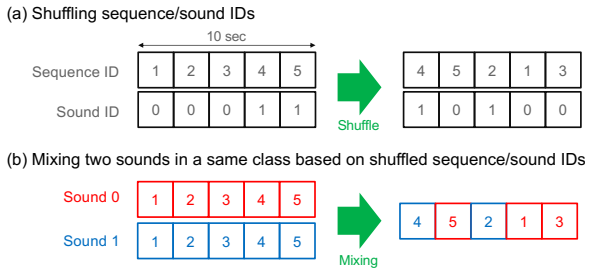

Figure 2: Generating new data based on shuffling and mixing.

same length of 10 seconds. We divide them into segments. The length of each segment can be arbitrary and is considered as a hyper-parameter that represents an estimated length of sound that contains at least one atomic foreground event. In the above example, the length of each segment is 2 seconds. We define two arrays to keep sequence IDs and sound IDs respectively and shuffle them as shown in Fig. 2(a). The sequence ID represents the order of sound segments, that is, when a sequence ID array is shuffled from $(1, 2, 3, 4, 5)$ to $(4, 5, 2, 1, 3)$, it means that the forth segment of the original sound is used as the first segment of the new sound, the fifth segment is used as the second segment and so on. The sound ID represents sound source from two sounds, *Sound*-0 or *Sound*-1, and how to mix them. For example, a sound ID $(0, 0, 0, 1, 1)$ represents a 60% *Sound*-0 mixing ratio, which is also a hyper-parameter of the method. When the sound ID is shuffled to $(1, 0, 1, 0, 0)$, the first segment of the new sound is picked from *Sound*-1 and the second segment is picked from *Sound*-0, and so on. We mix two sounds of the same class based on the shuffled sequence/sound IDs as shown in Fig. 2(b).

Generating new training samples in this way results in more variations of the temporal event location in the sound source. It also creates more variation in the number of sound events in a time period (event density). If the new sound is composed of multiple segments each containing a small number of sound events, it results in a decrease of event density. Similarly, if it is composed of multiple segments each containing a large number of sound events, the new sound will have higher event density. This is in contrast to the pre-

vious methods [12, 14] that mix the two sound sources by overlaying on top of each other, where the resulting sound keeps the same event order and tends to have higher event density than the original sound.

3. EXPERIMENTS

In this section, we evaluate our proposed data augmentation technique on two datasets with different characteristics. DCASE 2018 Task 5 dataset [15] is based on continuous recording sounds of a single person living in a vacation home over a period of one week [3]. It is composed of nine sound classes. Most of the sounds are created by one particular person and are relatively low volume except for *vacuum cleaning*. On the other hand, Urban Sound 8K dataset is created by downloading sounds from an online sound repository, Freesound.org. The recorded sounds come from various sound sources, containing ten classes of urban environmental sounds. Most of the sounds are quite noisy compared to the sounds in DCASE 2018 Task 5 dataset.

3.1. DCASE 2018 Task 5 dataset

The DCASE 2018 Task 5 dataset contains sound data captured in the living room. Each individual sound data is recorded using a single microphone array (with four microphones). There are microphone arrays at seven undisclosed locations. The dataset is divided into a development dataset and evaluation dataset. We focus on the development dataset in this paper. Each sound is 10 seconds long consisting of 4-channel 16-bit data sampled at 16 kHz. There are unequal numbers of samples in different classes, which possibly reflects the frequency of activities in real life. The amount of data in the following six classes: *cooking*, *dishwashing*, *eating*, *other*, *social activity*, and *vacuum cleaning*, is extremely small compared to the other three classes: *absence*, *watching TV*, and *working*.

The proposed data augmentation approach is used to increase the training data of the six classes to create a more balanced training set. Each sound data is divided into five segments with two seconds in length and mixed with 3-to-2 (60%) mixing ratio. Fig. 3 illustrates the amount of data in each class before and after applying our shuffling and mixing augmentation on Fold 1 of the development dataset.

As shown in Fig. 3, 30% of the training data is selected as validation data. All sounds recorded in the same session are only in either the training or validation data. This corresponds to how it is done in the baseline system. We converted the 10 second sound waveform into log-scaled mel-spectrogram (logmel) of size 40×501 matrix and used as the input to the deep learning model. More details of preprocessing are in our technical report of the DCASE 2018 challenge [16].

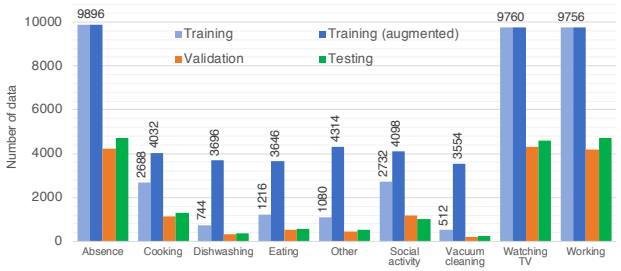

Figure 3: Number of data before and after data augmentation in Fold 1 of the development dataset. The augmentation is conducted on training data only.

Table 1: Proposed network architecture.

Layer	Output size
Input	$40 \times 501 \times 1$
Conv(7×1, 64) + BN + ReLU	$40 \times 501 \times 64$
Max pooling(4×1) + Dropout(0.2)	$10 \times 501 \times 64$
Conv(10×1, 128) + BN + ReLU	$1 \times 501 \times 128$
Conv(1×7, 256) + BN + ReLU	$1 \times 501 \times 256$
Global max pooling + Dropout(0.5)	256
Dense	128
Softmax output	9

In addition to data augmentation, we designed a new deep neural network architecture, where the main characteristics is that it starts with multiple convolutional layers across frequency axis where the kernel size on the time axis is fixed to one and then it followed by a convolutional layer across time where the kernel size on the frequency axis is fixed to one. This allows the network to look for local patterns across frequency bands and also the short-connected temporal components which represent sound events in the input data. In addition, the network also maintains the size of the time axis of the logmel until the final pooling layer. The complete network architecture and parameters are shown in Table 1.

In the dataset, one test sample has 4-channels. Sound in each channel is pre-processed and passed through the classifier independently. We average these four softmax predictions of each channel to calculate the final probability prediction for each test sample.

The experiments are carried out using the 4-fold cross validation setting of the development dataset. This corresponds to the test protocol of the DCASE 2018 Challenge. The model is trained with Adam optimizer [18] and an initial learning rate of 0.0001. We use a batch size of 256 samples and train the classifier for 500 epochs. The network weights which result in the best accuracy on the validation data is used to evaluate the test data. We examine the following configurations and compare the result with the baseline system: i) proposed CNN without data augmentation, ii) baseline CNN with proposed data augmentation, and iii)

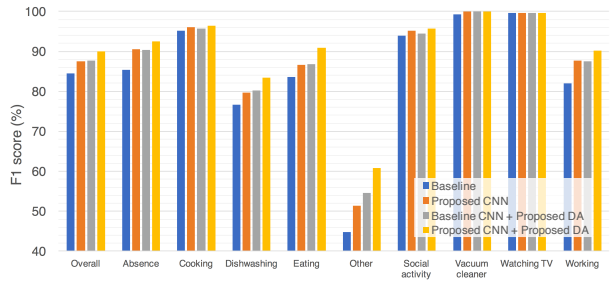

Figure 4: Comparison of macro-average F1 scores and F1 scores of each class.

proposed CNN with proposed data augmentation.

Fig. 4 shows the overall F1 scores and also for each class separately. We can see that the proposed network architecture and data augmentation approach each improves the classification performance and the combination of them gives the best performance. The overall F1 score by the proposed system is 89.95%, while the overall F1 score by the baseline is 84.50%. The proposed system improves F1 scores in all classes, especially the F1 score of *other* class.

3.2. Urban Sound 8K dataset

Urban Sound 8K dataset contains 10 sound classes of urban environmental sounds and has been widely used in acoustic classification literatures. Salamon and Bello [6] has investigated the effect of various data augmentation techniques and could be considered as a command baseline in this dataset. This experiment aims to compare an effect of those techniques the proposed shuffling and mixing data augmentation.

In the previous work [6], the details of how to augment each sound data is provided; however, implementation of the model and training procedure is not given. We attempted to replicate the result and our implementation achieved a mean accuracy of 71.6% across 10 folds for a baseline without data augmentation. Additional details are listed below: (a) We padded all sound data to 4 seconds by repeating the sound (self-concatenating) if required. During training, a 3 second segment is randomly chosen for each data sample in each epoch. However, during inference on a test sample, we slice a 3 second window with 1-frame hop in temporal axis of log-mel, pass them through the network, and ensemble the probability by averaging. (b) We replicated SB-CNN, use glorot uniform initialization for all layers and add batch normalization after each CNN layer [19]. (c) Model is trained for 50 epochs, with a minibatch of 100 samples. In each epoch, all sounds are considered in the training while undersampling method is applied to balance the number of data between all class. The model weights that performed best on the validation set are chosen for the final weights. (d) We strictly followed 10-fold cross-validation protocol[3]. During testing

[3]https://urbansounddataset.weebly.com/urbansound8k.html

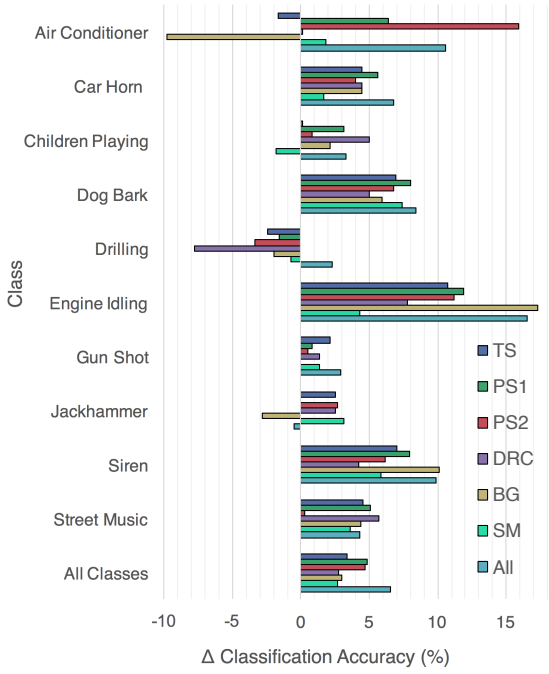

Figure 5: Overall and class-wised Δaccuracy from baseline of each data augmentation: Time Stretch (TS), Pitch Shift (PS1, PS2), Dynamic Range Compression (DRC), Background noise (BG), the proposed Shuffling and Mixing (SM), and All combined (All).

the Nth fold, the $(N-1)$th fold is used as validation.

We applied the proposed shuffling and mixing augmentation to the data by dividing each sound clip into 2 segments with 2 seconds in length and mixed them with 1-to-1 (50%) mixing ratio. Fig. 5 shows the difference for each class in the classification accuracy when adding each data augmentation compared to using only the original training set. Our proposed technique improves the accuracy compared to the baseline, although pitch shifting gives the best result for this dataset. The suitability of different data augmentation techniques to different datasets is worth studying in the future.

4. CONCLUSIONS

We have proposed a data augmentation technique that shuffles and mixes two sounds of the same class in training datasets. This data augmentation can generate new variations on both the sequence and the density of sound events. The proposed method is applied to DCASE 2018 Task 5 dataset and the Urban Sound 8K dataset. In general, the method improves classification results in both datasets. Specifically, it is a part of the system that won the DCASE 2018 Task 5 challenge and it also shows comparable results to other data augmentation techniques in the Urban Sound 8K dataset.

5. REFERENCES

[1] Michel Vacher, Francois Portet, Anthony Fleury, and Norbert Noury, "Development of audio sensing technology for ambient assisted living: Applications and challenges," in *International Journal of E-Health and Medical Communications (IJEHMC)*, 2011.

[2] Lode Vuegen, Peter Van Den Broeck, Bertand Karsmakers, Hugo Van hamme, and Bart Vanrumste, "Automatic monitoring of activities of daily living based on real-life acoustic sensor data: a preliminary study," in *Proc. Fourth workshop on speech and language processing for assistive technologies (SLPAT)*, 2013.

[3] Gert Dekkers, Steven Lauwereins, Bart Thoen, Mulu Weldegebreal Adhana, Henk Brouckxon, Toon van Waterschoot, Bart Vanrumste, Marian Verhelst, and Peter and Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *DCASE 2017*, November 2017.

[4] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley, "Detection and classification of acoustic scenes and events," in *IEEE Transactions on Multimedia*, 2015, vol. 17, pp. 1733–1746.

[5] Karol J. Piczak, "Environmental sound classification with convolutional neural networks," in *MLSP 2015*, 2015.

[6] Justin Salamon and Juan P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," in *IEEE Signal Processing Letters*, March 2017, vol. 24, pp. 279–283.

[7] Yoonchang Han, Jeongsoo Park, and Kyogu Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in *DCASE 2017*, 2017.

[8] Sharath Adavanne, Pasi Pertilä, and Tuomas Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *ICASSP 2017*, 2017.

[9] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das, "Very deep convolutional neural networks for raw waveforms," in *ICASSP 2017*, 2017.

[10] Brian McFee, Eric J. Humphrey, and Juan P. Bello, "A software framework for musical data augmentation," in *ISMIR 2015*, 2015, pp. 248–254.

[11] Navdeep Jaitly and Geoffrey E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *ICML 2013*, 2013.

[12] Naoya Takahashi, Michael Gygli, Beat Pfister, and Luc Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event recognition," in *INTERSPEECH 2016*, September 2016.

[13] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada, "Learning from between-class examples for deep sound recognition," in *ICLR 2018*, 2018.

[14] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR 2018*, 2018.

[15] Gert Dekkers, Lode Vuegen, Toon van Waterschoot, Bart Vanrumste, and Peter Karsmakers, "DCASE 2018 Challenge - Task 5: Monitoring of domestic activities based on multi-channel acoustics," Tech. Rep., KU Leuven, July 2018.

[16] Tadanobu Inoue, Phongtharin Vinayavekhin, Shiqiang Wang, David Wood, Nancy Greco, and Ryuki Tachibana, "Domestic activities classification based on CNN using shuffling and mixing data augmentation," Tech. Rep., DCASE 2018 Challenge, September 2018, non-peer-reviewed technical report.

[17] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in *Proc of 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[18] Diederik P. Kingma and Jimmy Lei Ba, "Adam: a method for stochastic optimization," in *ICLR 2015*, 2015.

[19] Rui Lu, Zhiyao Duan, and Changshui Zhang, "Metric learning based data augmentation for environmental sound classification," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017.