Binary Attribute Embeddings for Zero-Shot Sound Event Classification

1st Yihan Lin
Kobe University
Graduate School of System Informatics
Kobe, Japan
yihan@stu.kobe-u.ac.jp

2nd Xunquan Chen
Kobe University
Graduate School of System Informatics
Kobe, Japan
xunquan@stu.kobe-u.ac.jp

3rd Ryoichi Takashima
Kobe University
Graduate School of System Informatics
Kobe, Japan
rtakashima@port.kobe-u.ac.jp

4th Tetsuya Takiguchi
Kobe University
Graduate School of System Informatics
Kobe, Japan
takigu@kobe-u.ac.jp

***Abstract*—In this paper, we introduce a zero-shot learning method for sound event classification. The proposed method uses a semantic embedding of each sound event class and measures the compatibility between the semantic embedding and the input audio feature embedding. For semantic embedding, we newly define attribute vector that explains several attribute information of a sound event class, such as pitch, length, material of the sound source, etc. In the experiments, the proposed method showed higher accuracy than a conventional method using word embedding as the semantic embedding.**

***Index Terms*—Sound Event Classification, Zero-Shot Learning, Semantic Embedding, Attribute Information**

I. Introduction

Sound Event Classification (SEC) is a task in which we classify active sound events in a recording, such as the sound of running water, footsteps, or a moving car. One of the problems with this task is that training data is hard to come by for some events. For example, anomaly data in the anomaly event detection is difficult to be corrected because the anomaly event rarely happens. In this work, we investigate the zero-shot learning method for SEC, that is, recognizing a sound event which has no training data.

Whereas the zero-shot learning has been studied widely in the field of computer vision [1], [2], there are few researches of zero-shot learning for SEC [3]. In the previous work [3], they have used semantic embeddings of the event class instead of the class label itself, and they have recognized an unseen event evaluating the compatibility between the semantic embeddings and the input audio feature embedding. For the semantic embedding, this method used a word embedding generated by Word2Vec [4] from a class label. However, representing class information by word embedding is considered insufficient for sound classification, because it reflects the semantic similarity of each word, but not the audio similarity.

In this study, in order to explore more appropriate semantic embeddings that reflect the similarity of sound, we propose attribute information (such as the material of the sound source and high/low pitch) of each sound event.

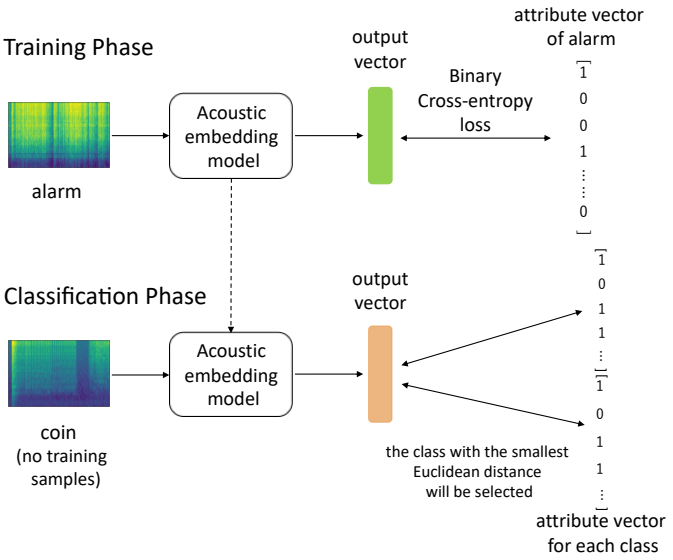
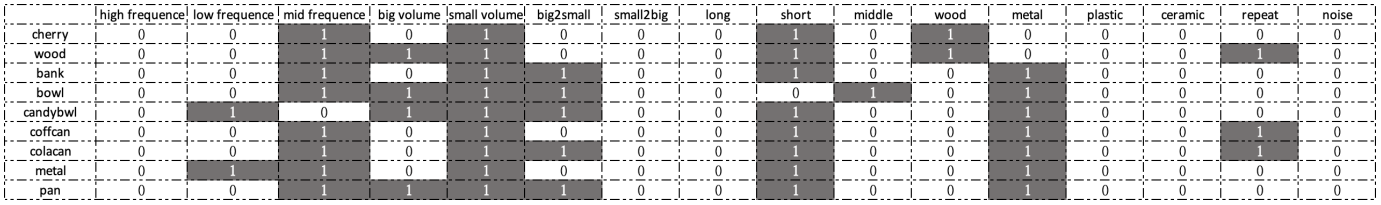


Fig. 1. Overview of the proposed zero-shot SEC framework

II. Proposed Method

Fig. 1 shows the overview of the proposed method. During training, the acoustic embedding model is trained to output attribute vector that explains the attribute information of the event class. For example, the sound of a hair dryer is like white noise, and the sound of an alarm is a repeating sound.

In the classification phase of the zero-shot task, the event class to be recognized has no available training samples but has only attribute information. The input sound is converted to the acoustic embedding by the trained model. Then, the Euclidean distance between the output acoustic embedding and



	high frequence	low frequence	mid frequence	big volume	small volume	big2small	small2big	long	short	middle	wood	metal	plastic	ceramic	repeat	noise
cherry	0	0	1	0	1	0	0	0	1	0	1	0	0	0	0	0
wood	0	0	1	1	1	0	0	0	1	0	1	0	0	0	1	0
bank	0	0	1	0	1	1	0	0	1	0	0	1	0	0	0	0
bowl	0	0	1	1	1	1	0	0	0	1	0	1	0	0	0	0
candybwl	0	1	0	1	1	1	0	0	1	0	0	1	0	0	0	0
coffcan	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1	0
colacan	0	0	1	0	1	1	0	0	1	0	0	1	0	0	1	0
metal	0	1	1	0	1	0	0	0	1	0	0	1	0	0	0	0
pan	0	0	1	1	1	1	0	0	1	0	0	1	0	0	0	0

Fig. 2. Example of binary attribute embeddings of some classes in our experiments

the attribute vector of each class, and the class having the smallest distance is selected as the recognition result.

As there has been no previous work addressing the attribute information of the sound event, we newly define the attribute information named as "binary attribute embeddings". The binary attribute embedding consists of a total of 16 attributes, such as "metallic sound or not", "repeated sound or not", "high-pitched sound or not", etc., designating the sound as "1" if it is true and "0" if it is not. Fig. 2 shows examples of binary attribute embeddings of some classes used in our experiments.

III. Experiment And Results

We conducted our experiments on the RWCP-SSD [5] dataset, which contains various real-world environmental sounds. In this dataset, we selected 20 classes for training set and 6 classes as testing set as shown in Table I. For each test data, the SEC system select the recognition result from the classification candidates, that is the 6 unseen classes defined for the test set. In other words, we evaluate the proposed method in a 6-class classification task. VGGish [6] was used as the acoustic embedding model. All the sounds were trimmed to 3 seconds in duration, and then they were transformed into a 40-dimensional log Mel-filter bank features.

TABLE I
THE CLASSES IN TRAINING SET AND TEST SET

Training Data : 20 classes (4,049 data in total)
cherry, bank, bowl, candybwl, coffcan, colacan, metal, pan, trash-box, case, dice, bottle, china, cup, pump, spray, clap, alarm, dryer, tear

Test Data : 6 classes (1,560 data in total)
bell, tambourine, coin, clock, wood, particle

We compared the classification performances when using our proposed binary attribute embedding and when using conventionally used word embedding. The results using the word embedding and using the binary attribute embedding are shown in Table II and Table III, respectively. When the word embedding is used, most of testing samples were classified as coin or wood. As a result, the classification accuracy for coin and wood was high, whereas the accuracy for other classes was 0. On the other hand, by using the binary attribute embedding, the classification accuracy was improved in most of the classes.

TABLE II
CONFUSION MATRIX OF TEST CLASSES USING WORD EMBEDDING[%]. THE AVERAGE ACCURACY WAS 51.1%.

		Predicted label					
		bell	coin	tambourine	clock	wood	particle
True label	bell	0.0	5.4	0.0	0.0	94.6	0.0
	coin	0.0	92.0	0.0	0.0	8.0	0.0
	tambourine	0.0	1.4	0.0	0.0	84.6	0.0
	clock	0.0	0.0	0.0	0.0	100.0	0.0
	wood	0.0	1.4	0.0	0.0	98.6	0.0
	particle	0.0	61.0	0.0	0.0	39.0	0.0

TABLE III
CONFUSION MATRIX OF TEST CLASSES USING BINARY ATTRIBUTE EMBEDDING[%]. THE AVERAGE ACCURACY WAS 64.3%.

		Predicted label					
		bell	coin	tambourine	clock	wood	particle
True label	bell	60.6	1.2	0.7	37.1	0.0	0.2
	coin	4.2	71.3	5.0	13.5	0.3	5.7
	tambourine	86.5	13.5	0.0	0.0	0.0	0.0
	clock	0.0	0.0	0.0	100.0	0.0	0.0
	wood	0.0	0.0	0.0	0.3	99.7	0.0
	particle	22.	20.0	8.0	30.0	10.0	10.0

IV. Conclusion

In this paper, by defining the attribute information, we achieved higher zero-shot classification performance than the conventional method using word embedding. For future work, we will continue to explore more appropriate attribute information and the acoustic embedding model.

References

[1] W. Wang, et al., "A survey of zero-shot learning: Settings, methods, and applications," ACM Transactions on Intelligent Systems and Technology (TIST), 10(2):13, 2019.

[2] C. H. Lampert, et al., "Learning to detect unseen object classes by between-class attribute transfer," in Proc. ICCV, 2009.

[3] H. Xie, et al., "Zero-shot audio classification via semantic embeddings," IEEE/ACM Trans. on ASLP, vol. 29, pp. 1233-1242, 2021.

[4] T. Mikolov, et al., "Distributed representations of words and phrases and their compositionality," in Proc. NIPS, pp.3111-31119, 2013.

[5] S. Nakamura, et al., "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in Proc. LREC2000, pp. 965-968, May. 2000.

[6] R. S. Hershey, et al., "CNN architectures for large-scale audio classification," in Proc. ICASSP, 2017.