CLASSIFICATION OF BUILDING NOISE TYPE/POSITION VIA SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper presents noise type/position classification of various impact noises generated in a building which is a serious conflict issue in apartment complexes. For this study, a collection of floor impact noise dataset is recorded with a single microphone. Noise types/positions are selected based on a report by the Floor Management Center under Korea Environmental Corporation. Using a convolutional neural networks based classifier, the impact noise signals converted to log-scaled Mel-spectrograms are classified into noise types or positions. Also, our model is evaluated on a standard environmental sound dataset ESC-50 to show extensibility on environmental sound classification.

1 INTRODUCTION

Some conflicts between residents originated from incorrect source localization by human hearing. Also, correctly identifying noise types/locations is the first step for the noise reduction. Therefore, noise type/position classification is a technique required to identify impact noise.

Various impact noises such as footstep and hammer hitting in a living space incur annoyance to residents (Park et al., 2016). Chronic noise in a living space is a significant threat to resident's health (Park et al., 2017; Miedema, 2004). In some case, impact noise arises conflict between residents. Since more than 60 % of the residential buildings in Korea are apartment housings (Shin et al., 2015), the conflict has become serious social issue (Lee & Haan, 2011; Park et al., 2016). In 2012, the Korea government established the Floor Noise Management Center under Korea Environment Corporation affiliated with the Ministry of Environment (Floor-Noise-Management-Center, 2012) for impact noise identification and conflict mediation. The center has handled 119,500 civil complaints of impact noise over 6 years (Floor-Noise-Management-Center, 2018).

There are several related works on noise reduction (Choi et al., 2004; Lee et al., 2014), annoyance measurement (Park et al., 2016), and noise measurement (Jeon et al., 2006; Park et al., 2017). However, impact noise classification is studied only in our previous work (Anonymous-authors, 2018). Our previous work studies classification of the impact noises using a convolutional neural networks (CNN) based model. Our model classifies impact noise recordings into labeled categories. It shows extensibility of CNN to impact noise classification. But, our model is evaluated on the limited data generated on limited positions. And, the labels of dataset are categorized into noise type-position combined form.

In order to improve our previous work, this study expands the previous work as follows. First, $1,000$ impact noise data is newly gathered on 10 more positions in the building. The new data is set as test set to validate robustness of our model. Second, the classification problem is divided into following two problems: noise type classification problem and position classification problem. This form is considered as more adequate problem definition. Also, the number of samples per category is increased which is expected to improve performance of our model. Third, our model is validated on a standard environment sound dataset. This validation can show the extensibility of our model to other problems.

We expect that this work can contribute to other fields. Expected fields are noise type/position classification in a very complex structure, and environmental sound classification.

2 IMPACT NOISE DATASET

Since a dataset for noise type/position classification of impact noise does not exist, we built an impact noise dataset in our past work (Anonymous-authors, 2018). It is composed of audio clips of impact noise recorded by a smartphone microphone (Samsung Galaxy S6). In this work, we gathered impact noise data on 10 more postions (19 locations in total) in the building to expand the dataset.

We planned dataset collection based on a report by the Floor-Noise-Management-Center (2018). In the report, from 2012 to 2018, the center received 119,550 complaints from victims suffering from impact noise. The center visited 28.1% of the victims to identify impact noise. 79.4% of the complaints were caused by the upper floor residents and 16.3% of the complaints were by the lower floor residents. Identified noise types are listed in the following order: footstep (71.0%), hammering (3.9%), furniture (3.3%), home appliances (vacuum cleaner, laundry machine, and television) (3.3%), door (2.0%) and so on. Unidentified or unrecorded sources account for 10.1% of the total. Based on these results in the report, we focused on generation of impact noise on the upper floor(3F) and the lower floor(1F). In addition to them, impact noises on the 2F(the middle floor) are also recorded to check whether our model can distinguish the noise generated on this floor from the noises on the other floors. Also, top four noise types which occupy 81.5% of the identified noise types are selected.

Figure 1 illustrates noise(source)/receiver positions in the building for impact noise generation. 9 solid circles are selected (noise) positions for training and validation of a model. 10 circles in grid pattern are (noise) positions not for training a model but for checking the limitation of the model. Five noise types are selected to cover the top four noise types: a medicine ball dropped from 1.2 m of height hitting the floor (MB), a hammer dropped from 1.2 m of height hitting the floor (HD), hitting the floor with a hammer (HH), dragging a chair on the floor (CD), and a vacuum cleaner (VC). Generating real footstep noise many times on every source position is challenging work. Furthermore, it could hurt a person who generates the noise. Thus, usually, an impact ball (2.5 kg, 185 mm) or a bang machine (7.3 kg) is used to produce low frequency noise of footstep noise (Jeon et al., 2006). Instead of using them, a medicine ball (2.0 kg, 200 mm) is used to produce the low frequency noise. Since a laundry machine and a television are hard to transport and install at the noise location, only vacuum cleaner is used to generate noise. VC is generated only on 2F because vacuum cleaner noise on the upper floor and the lower floor are barely audible at the receiver position in the building. Sampling frequency and sample duration are set as 44,100 Hz and approximately 5 s, respectively.

Table 1 is summary of the finalized impact noise dataset. It contains 2,950 floor impact noises in total and they can be classified into 59 categories. Each category contains 50 recordings of floor impact noise.

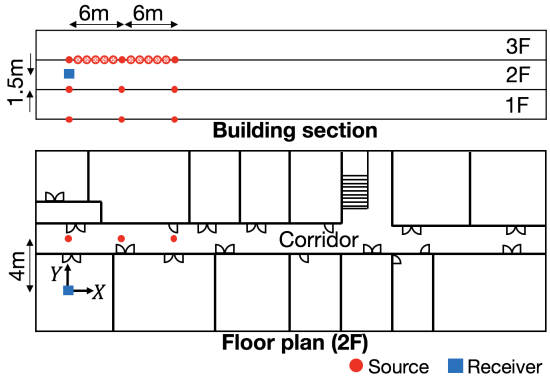

Figure 1: Noise(source)/receiver positions for generation of impact noise dataset.

Table 1: Summary of the impact noise dataset (*The first row indicates X direction)

		0m	1m	2m	3m	4m	5m	6m	7m	8m	9m	10m	11m	12m
3F	**MB**	50	*50*	*50*	*50*	*50*	*50*	50	*50*	*50*	*50*	*50*	*50*	50
	HD	50	-	-	-	-	-	50	-	-	-	-	-	50
	HH	50	*50*	*50*	*50*	*50*	*50*	50	*50*	*50*	*50*	*50*	*50*	50
	CD	50	-	-	-	-	-	50	-	-	-	-	-	50
2F	**MB**	50	-	-	-	-	-	50	-	-	-	-	-	50
	HD	50	-	-	-	-	-	50	-	-	-	-	-	50
	HH	50	-	-	-	-	-	50	-	-	-	-	-	50
	CD	50	-	-	-	-	-	50	-	-	-	-	-	50
	VC	50	-	-	-	-	-	50	-	-	-	-	-	50
1F	**MB**	50	-	-	-	-	-	50	-	-	-	-	-	50
	HD	50	-	-	-	-	-	50	-	-	-	-	-	50
	HH	50	-	-	-	-	-	50	-	-	-	-	-	50
	CD	50	-	-	-	-	-	50	-	-	-	-	-	50

3 LEARNING IMPACT NOISE

In this section, we explain our noise type/position classifier for impact noise generated in a building. In Section 3.1, applications of CNN(convolutional neural networks) in audio area are briefly reviewed. Noise type and position classifications are presented in details in Section 3.2.1 and Section 3.2.2, respectively. In Section 3.3, our method is applied to classification of other standard environment sound dataset (ESC-50) to examine that our method can be extended to environmental sound classification problems.

3.1 CONVOLUTIONAL NEURAL NETWORKS IN AUDIO DOMAIN

CNN is well known for its remarkable performance than those of conventional machine learning techniques in visual recognition tasks. CNN is also widely used in audio domain, such as environmental sound classification (Piczak, 2015a; Salamon & Bello, 2017; Tokozume & Harada, 2017) and music classification (Dieleman & Schrauwen, 2014; Lee & Nam, 2017; Lee et al., 2018). Input features of their models are time-frequency patch or raw waveform instead of using RGB color space image. But, their design pattern is fundamentally same with that used in visual recognition task which is composed of convolutional layers, pooling layers, and fully connected layers.

There are several works which employ a model for visual recognition task to audio domain. Hershey et al. (2017) showed state-of-the-art models for visual recognition perform well on audio event classification. Amiriparian et al. (2017) employed VGG19 and AlexNet for snore sound classification. Ren et al. (2018) employed VGG16 for phonocardiogram classification.

Usually, a CNN based model contains a large number of learnable parameters and its performance is limited if dataset is small (Oquab et al., 2014). In such a situation, transfer learning, known as a technique to improve the model performance, can be introduced (Van Den Oord et al., 2014; Oquab et al., 2014; Marmanis et al., 2016; Soekhoe et al., 2016). The technique trains parameters of networks on a training data in source task. In target task, the parameters are transferred and fine-tuned on a target data. Pre-training of parameters in source task offers efficient learning because the parameters are pre-initialized in the source task (Soekhoe et al., 2016). Amiriparian et al. (2017) and Ren et al. (2018) pre-trained their models on ImageNet dataset and transferred the parameters to models in target tasks. Although these studies are visual knowledge transfer, the models perform well in audio domain.

3.2 CONVOLUTIONAL IMPACT NOISE NETWORKS

VGG16 by Simonyan & Zisserman (2014) is selected for this study instead of designing a new network architecture. There are several reasons why we select the model as a baseline model in this study. First, the model performs well in audio domain. In particular, the performance difference between the state-of-the-art model is not large(at most 0.024 area under curve) in (Hershey et al.,

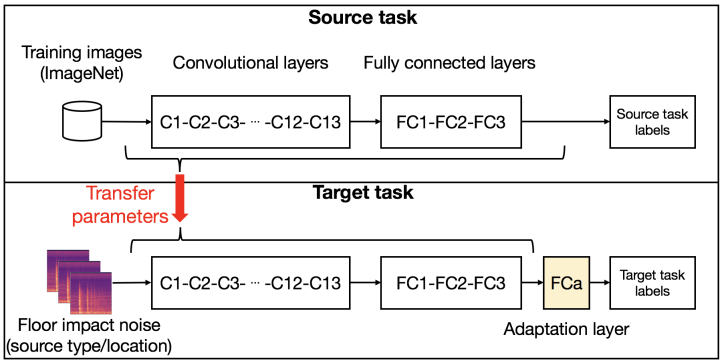

Figure 2: Transferring pre-trained parameters. C and FC represent convolutional layer and fully connected layer, respectively (Oquab et al., 2014).

2017). Second, its pre-trained parameters are accessible on Visual Geometry Group (VGG) website and managed by the group.

Figure 2 illustrates the model used for this study. The impact noise dataset contains smaller samples per category than a very large scale dataset. Therefore, this shortage of the dataset can limit performance of our model for classification of noise type/position. In order to overcome the limitation, pre-trained parameters by Simonyan & Zisserman (2014) on ImageNet are transferred to VGG16. An adaption layer which reduces output dimension to the number of categories is added and all the parameters of the model are fine-tuned on noise types or positions of the impact noises. We named the model as VGG16-PRE.

All the impact noise signals are converted to log-scaled Mel-spectrograms using LibROSA(version 0.5.1) (McFee et al., 2015). Size of the log-scaled Mel-spectrogram is fixed to 224×224 by VGG16 whose input dimension is $224 \times 224 \times 3$. Log-scaled Mel-spectrogram is obtained by the following steps. $\boldsymbol{s}$ with time duration of 3 s is extracted from each recording in the dataset. The time duration covers almost of floor impact noise duration. *Event start* in the metadata is referred for finding an initial location of each recording. $\boldsymbol{S}$ is squared magnitude of short time Fourier transform of $\boldsymbol{s}$ using $2,048$ point fast Fourier transform (FFT), window size of 591, and hop size of 591. The window size offers high time resolution of the time-frequency patch avoiding overlapping for the given input size and the time duration. $\boldsymbol{FS}$ gives a Mel-spectrogram $\boldsymbol{M}$, where $\boldsymbol{F}$ is a Mel-filter bank. Frequency range of the Mel-filterbank is set as $0 - 22,050\ Hz$. The Mel-spectrogram is converted to a log-scaled Mel-spectrogram $\boldsymbol{P} = 10 \log \boldsymbol{M}/M_m$, where M_m is the maximum element of $\boldsymbol{M}$. Since VGG16-PRE has 3 input channels, $\boldsymbol{P}$ is supplied to all the channels.

3.2.1 NOISE TYPE CLASSIFICATION

The impact noises are labeled into 5 noise types: MB, HD, HH, CD, and VC. Dimension of the adaptation layer (FCa) is set as 5 and the pre-trained parameters are transferred to VGG16-PRE. L_2-regularization is applied to the last layer with penalty value of 0.01. VGG16-PRE is fine-tuned on the impact noises whose number of recordings are not written italics in Table 1. We named this dataset as **TV-set**(training and validation set). The others are not used for the fine-tuning but purely used for testing the fine-tuned model. Since they are generated out of the positions used for fine-tuning, it can be used for testing the robustness of noise type classification. We named this dataset as **TS-set**(test set).

VGG16-PRE is evaluated using 5-fold cross validation. Usually, it is used for evaluating a model when a dataset is small. Also, every model fine-tuned on k-th fold of TV-set is tested against TS-set. The fine-tuning minimizes cross-entropy loss with logits using mini-batch gradient descent with learning rate of 0.001 and mini-batch size of 30. The global mean value of the input channel is changed to the mean of the training data. The parameters of VGG16-PRE are not frozen for all the layers. A softmax classifier is employed. A model with the highest validation accuracy is saved during 30 epochs of training on each fold. Validation accuracy on each fold of TV-set and test accuracy on TS-set are measured, respectively.

3.2.2 POSITION CLASSIFICATION

The impact noises in TV-set are labeled into 9 positions depending on their impact positions: 1F00m, 1F06m, 1F12m, 2F00m, 2F06m, 2F12m, 3F00m, 3F06m, and 3F12m, where the first two characters represents floor and the followings are distance from the receiver position in X direction. One unique point of position classification is that TS-set is composed of impact noises generated out of the 9 positions used for fine-tuning. So, it is an interesting point that to observe classification of TS-set into the 9 positions by a model fine-tuned on TV-set.

For fine-tuning a model, dimension of the adaptation layer (FCa) is set as 9 and the pre-trained parameters are transferred to VGG16-PRE. The later steps including optimization and evaluation methods are same with those in Section 3.2.1 except performance measurement on TS-set.

We suggest a performance test for position classification on TS-set as follows. Figure 3 illustrates noise(source) positions on 3F where the impact noises are generated. Intuitively, two dashed lines can divide the positions into three groups. These two dashed lines are assumed as virtual boundaries. The impact noises generated on 3F3m and 3F9m are excluded in performance measurement because they are on the boundaries. True label of an impact noise in TS-set is assumed as the closest position in TV-set. For example, true label of an impact noise whose source position is 3F8m is assumed as 3F6m. Finally, test accuracy is measured using the assumed labels.

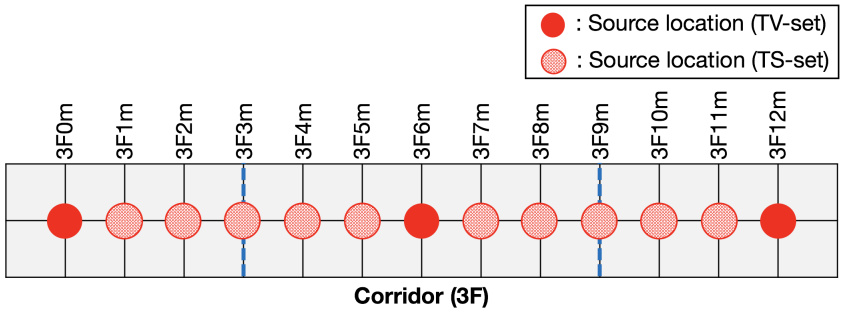

Figure 3: Source positions on 3F.

3.3 VALIDATION ON A STANDARD ENVIRONMENTAL SOUND DATASET

The impact noise can be considered as environmental sound. In this section, VGG16-PRE is evaluated on a standard environmental sound dataset ESC-50 (Piczak, 2015b). Actually, this evaluation is out of scope for impact noise identification. However, through the evaluation, VGG16-PRE can be verified on a standard sound dataset. Also, robustness and extensibility of VGG16-PRE to environmental sound classification can be shown.

ESC-50 is composed of 50 categories and each category contains 40 environmental sounds. ESC-50 is pre-arranged into five folds for fair performance comparison. Time duration and sampling frequency of each audio clip are 5 s and $44,100$ Hz, respectively. They are converted to log-scaled Mel-spectrograms by the method in Section 3.2. Window size and hop size are set as 985 in order to use all time range of audio clip avoiding overlapping.

VGG16-PRE is fine-tuned on each fold for 10 epochs. Mini-batch size and learning rate are set as 30 and 0.001, respectively. Also, validation accuracy is measured.

4 RESULTS AND DISCUSSIONS

4.1 NOISE TYPE CLASSIFICATION RESULTS

Table 2 shows accuracies of the noise type classifier on TV-set and TS-set. The first column of the table represents dataset. The first row of the table represents noise types of the impact noises. Validation accuracy on TV-set is measured as 99.7 %. Test accuracy on the TS-set is measured

as 99.2 %. Since the classifier is trained only on the TV-set, test accuracy can be lower than the validation accuracy.

One notable result is, for noise type classification, VGG16-PRE shows robustness on position change. As shown in Table 1, impact positions used for generating the TS-set are out of those used for TV-set generation, but the accuracy difference between the validation accuracy and test accuracy is 0.5 %.

Table 2: Accuracy of noise type classifier on TV-set and TS-set

	MB	**HD**	**HH**	**CD**	**VC**
TV-set	0.998	0.989	0.998	1.000	1.000
TS-set	0.992	-	0.992	-	-

4.2 POSITION CLASSIFICATION RESULTS

Table 3 shows validation accuracy of the position classifier on TV-set. The first row of the table represents the 9 positions used for generating the TV-set. The second row shows the corresponding validation accuracies to the 9 positions. Average of the accuracies is 94.1 %. When the accuracies are divided into 3 groups by floor, then validation accuracy on 1F is lower than that on 3F.

Table 3: Validation accuracy of source position classifier on TV-set

1F00m	**1F06m**	**1F12m**	**2F00m**	**2F06m**	**2F12m**	**3F00m**	**3F06m**	**3F12m**
0.900	0.875	0.940	1.000	0.956	0.976	0.940	0.915	0.970

Table 4 shows test accuracy of the position classifier on TS-set, where the first row represents the true labels assumed in Section 3.2.2. The second row of the table shows the corresponding test accuracies to the assumed true labels. Average of the accuracies is 69.6 %. Since positions of the impact noises in TS-set are different from those in TV-set, the test accuracy can be lower than the validation accuracy. If the position classification is changed to floor classification, then the test accuracy is raised to 98.8 %.

Table 4: Test accuracy of position classification on TS-set

3F00m	**3F06m**	**3F12m**
0.816	0.556	0.854

Figure 4 shows confusion matrices drawn with the validation and the test results. The confusion matrix at the left is drawn with the validation results. In the confusion matrix, the followings are observed. Most of the errors are the misclassifications to neighboring positions on the same floor. Especially, impact noises at $X = 6\ m$ are classified to the nearby locations. It is also observed in Table 3.

The confusion matrix at the right is drawn with the test results. The true labels are separately represented into two noise types: HH and MB, in order to observe the position classification to noise types. The predicted labels are the true labels assumed in Section 3.2.2. The dotted lines indicate a subset of the 9 positions used for training the model. When test accuracy is separately calculated depending on noise type, test accuracies are 74.1 % for HH and 65.0 % for MB.

4.3 VALIDATION RESULTS ON A STANDARD ENVIRONMENT SOUND DATASET

On ESC-50 repository, evaluation results of other models designed for environmental sound classification are reported (Piczak, 2015a). Table 5 shows validation accuracies of our model on ESC-50 and the top-ranked models on ESC-50 repository. Our model shows 12.3 % higher validation accuracy than the best model Sailor et al. (2017). This experimental result supports that visual knowledge transfer can be effective to environmental sound classification.

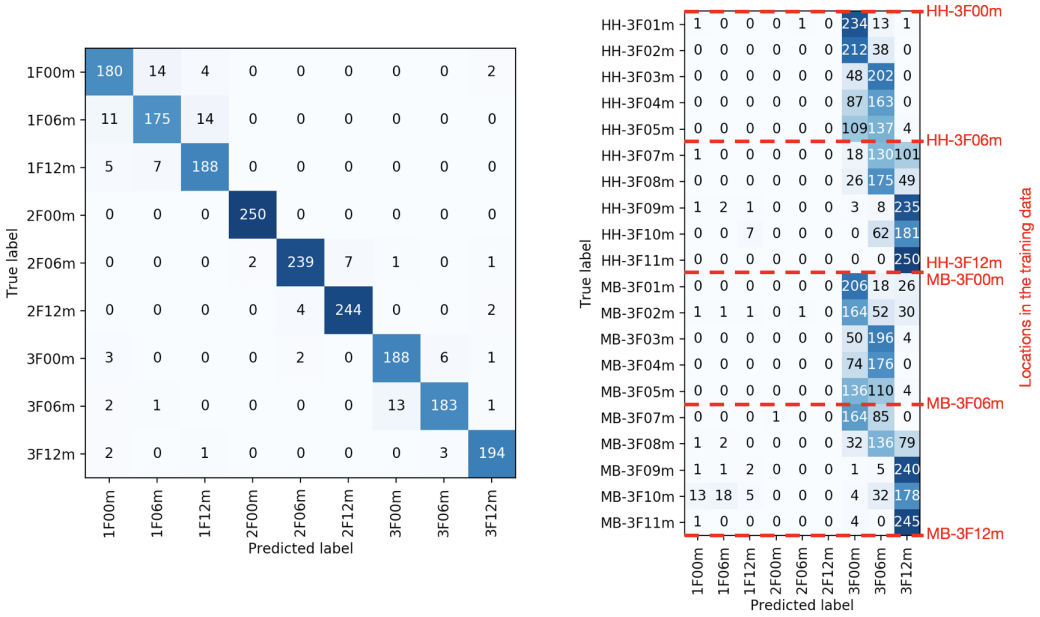

Figure 4: (Left) Confusion matrix drawn with the validation results. (Right) Confusion matrix drawn with the test results.

Table 5: Validation accuracy of models on ESC-50

Model	**Validation accuracy**
VGG16-PRE	0.988
ConvRBM (Sailor et al., 2017)	0.865
EnvNet-v2 (Tokozume et al., 2017)	0.849
CNN pre-trained on AudioSet (Kumar et al., 2017)	0.835

Figure 5 shows confusions of VGG16-PRE on ESC-50. In the confusion matrix, confusions between ESC-50 categories can be observed. Also, validation accuracy to each category can be observed. The most confusing category is engine. The categories of ESC-50 can be loosely rearranged into 5 major categories: Animals, Natural soundscapes & water sounds, Human/non-speech sounds, Interior/domestic sounds, and Exterior/urban noises. The most confusing major category is Exterior/urban noises (validation accuracy is 97.6 %).

5 Conclusions

In this study, a convolutional neural networks based model is proposed for noise type/position classification of impact noise. An impact noise dataset is built for evaluation of our model. The dataset is built based on a report by the Floor Management Center. The dataset is divided into a training-validation set and a test set. The models for noise type and position classifications are separately designed, but their architectures are fundamentally same except the dimension of the adaptation layers. VGG16 with an adaptation layer is employed for the tasks instead of designing a new model. Since the impact noise dataset is small, parameters of VGG16 pre-trained on ImageNet are transferred to a model. All the parameters of the model are fine-tuned on noise types or positions. The model shows 99.7 % of validation accuracy and 99.2 % of test accuracy for noise type classification. For position classification, the model shows 94.1 % of validation accuracy and test accuracy of 69.6 %. If the position classification is changed to floor classification, then test accuracy is improved to 98.8 %.

The model used for impact noise identification is evaluated on the ESC-50 to compare the evaluation results with other state-of-the-art results. Validation accuracy of the model on ESC-50 is 98.8 %. It

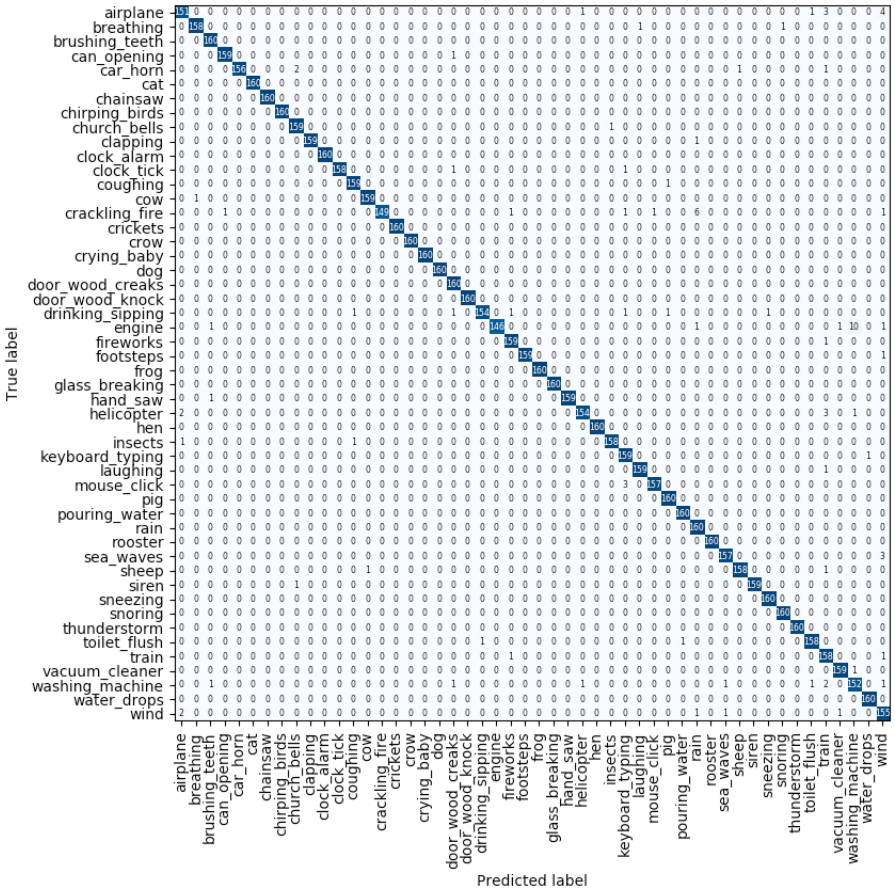

Figure 5: Confusions of VGG16-PRE on ESC-50.

is the best accuracy ever reported on ESC-50 repository. The result shows potential of the method to environmental sound classification as well as impact noise classification.

Future works include impact noise generation at other buildings and apartment houses, and evaluation of the model on another standard environmental sound dataset.

REFERENCES

Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. Snore sound classification using image-based deep spectrum features. In *Proceedings INTERSPEECH*, pp. 3512–3516, 2017.

Anonymous-authors. Hidden title. In *Hidden*, pp. xxx–xxx. Hidden, 2018.

Gyoung-Seok Choi, Hyun-jung Choi, Kwan-Seop Yang, and Kyoung-Woo Kim. Evaluation of floor impact sound performance according to the reduction methods. *Transactions of the Korean Society for Noise and Vibration Engineering*, 14(9):811–818, 2004.

Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 6964–6968. IEEE, 2014.

Floor-Noise-Management-Center. Floor noise management center homepage. `http://www.noiseinfo.or.kr/`, 2012. [Online; accessed 9-September-2018].

Floor-Noise-Management-Center. 2018 report. http://www.noiseinfo.or.kr/about/data_view.jsp?boardNo=213&keyfield=whole&keyword=&pg=1, 2018. [Online; accessed 9-September-2018].

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 131–135. IEEE, 2017.

Jin Yong Jeon, Jong Kwan Ryu, Jeong Ho Jeong, and Hideki Tachibana. Review of the impact ball in evaluating floor impact sound. *Acta Acustica united with ACUSTICA*, 92(5):777–786, 2006.

Anurag Kumar, Maksim Khadkevich, and Christian Fugen. Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. *arXiv preprint arXiv:1711.01369*, 2017.

Jongpil Lee and Juhan Nam. Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging. *IEEE signal processing letters*, 24(8):1208–1212, 2017.

Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, and Juhan Nam. Samplecnn: End-to-end deep convolutional neural networks using very small filters for music classification. *Applied Sciences*, 8(1):150, 2018.

Won-Hak Lee and Chan-Hoon Haan. Floor impact noise characteristics depending on the experimental conditions using impact ball. *The Journal of the Acoustical Society of Korea*, 30(2):92–99, 2011.

Won-hak Lee, Kyoung-woo Kim, and Seock-ho Lim. Improvement of floor impact sound on modular housing for sustainable building. *Renewable and Sustainable Energy Reviews*, 29:263–275, 2014.

Dimitrios Marmanis, Mihai Datcu, Thomas Esch, and Uwe Stilla. Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1):105–109, 2016.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pp. 18–25, 2015.

Henk ME Miedema. Relationship between exposure to multiple noise sources and noise annoyance. *The Journal of the Acoustical Society of America*, 116(2):949–957, 2004.

Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1717–1724, 2014.

Sang Hee Park, Pyoung Jik Lee, Kwan Seop Yang, and Kyoung Woo Kim. Relationships between non-acoustic factors and subjective reactions to floor impact noise in apartment buildings. *The Journal of the Acoustical Society of America*, 139(3):1158–1167, 2016.

Sang Hee Park, Pyoung Jik Lee, and Byung Kwon Lee. Levels and sources of neighbour noise in heavyweight residential buildings in korea. *Applied Acoustics*, 120:148–157, 2017.

Karol J Piczak. Environmental sound classification with convolutional neural networks. In *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*, pp. 1–6. IEEE, 2015a.

Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018. ACM, 2015b.

Zhao Ren, Nicholas Cummins, Vedhas Pandit, Jing Han, Kun Qian, and Björn Schuller. Learning image-based representations for heart sound classification. In *Proceedings of the 2018 International Conference on Digital Health*, pp. 143–147. ACM, 2018.

Hardik B Sailor, Dharmesh M Agrawal, and Hemant A Patil. Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification. *Proc. Interspeech 2017*, pp. 3107–3111, 2017.

Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.

Jaemin Shin, Hyomin Song, and Yoonseok Shin. Analysis on the characteristic of living noise in residential buildings. *Journal of the Korea Institute of Building Construction*, 15:123–131, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Deepak Soekhoe, Peter van der Putten, and Aske Plaat. On the impact of data set size in transfer learning using deep neural networks. In *International Symposium on Intelligent Data Analysis*, pp. 50–60. Springer, 2016.

Yuji Tokozume and Tatsuya Harada. Learning environmental sounds with end-to-end convolutional neural network. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 2721–2725. IEEE, 2017.

Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Learning from between-class examples for deep sound recognition. *arXiv preprint arXiv:1711.10282*, 2017.

Aäron Van Den Oord, Sander Dieleman, and Benjamin Schrauwen. Transfer learning by supervised pre-training for audio-based music classification. In *Conference of the International Society for Music Information Retrieval (ISMIR 2014)*, 2014.