ENHANCING SOUND TEXTURE IN CNN-BASED ACOUSTIC SCENE CLASSIFICATION

Yuzhong Wu, Tan Lee

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China

ABSTRACT

Acoustic scene classification is the task of identifying the scene from which the audio signal is recorded. Convolutional neural network (CNN) models are widely adopted with proven successes in acoustic scene classification. However, there is little insight on how an audio scene is perceived in CNN, as what have been demonstrated in image recognition research. In the present study, the Class Activation Mapping (CAM) is utilized to analyze how the log-magnitude Mel-scale filter-bank (log-Mel) features of different acoustic scenes are learned in a CNN classifier. It is noted that distinct high-energy time-frequency components of audio signals generally do not correspond to strong activation on CAM, while the background sound texture are well learned in CNN. In order to make the sound texture more salient, we propose to apply the Difference of Gaussian (DoG) and Sobel operator to process the log-Mel features and enhance edge information of the time-frequency image. Experimental results on the DCASE 2017 ASC challenge show that using edge enhanced log-Mel images as input feature of CNN significantly improves the performance of audio scene classification.

Index Terms— Convolutional neural network, acoustic scene classification, sound texture, class activation map, edge enhancement

1. INTRODUCTION

Large amount of multimedia information becomes easily accessible nowadays. The performance of speech and image recognition systems has been significantly improved with the use of deep neural networks and exploding amount of training data. Audio-related tasks, e.g., Acoustic Scene Classification (ASC) [1, 2, 3], Sound Event Detection (SED) [4, 5, 6] and Audio Tagging [7, 8, 9, 10], have also received increasing attention in recent years. They have many real-world applications. For example, context-aware mobile devices could provide better responses to their users in accordance with the acoustic scene. A smart home-monitoring system could detect unusual incidences by using audio. An audio search engine is able to retrieve information efficiently from massive online recordings.

Acoustic scene classification (ASC) is the process of identifying the type of acoustic environment (scene) where a given audio signal was recorded. It has been a major task in the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) since 2013. In the 2017 ASC challenge, most of the best-performing models were based on convolutional neural networks (CNN). Mun et al. [11] addressed the problem of data insufficiency and proposed to use the Generative Adversarial Network (GAN) [12] to augment training data. Han et al. [13] was focused on preprocessing of input features. Fusion of CNN models with preprocessed input features led to improved overall model performance.

Despite the clearly demonstrated effectiveness of CNN-based models in the ASC task, there is little insight on how an audio scene is perceived in a CNN model. Whilst similar issue has been extensively explored in image classification. In [14], Zeiler & Fergus used the De-convolutional Network [15] to visualize and understand CNN. Springenberg et al. applied the guided backpropagation [16] to obtain sharp visualization of descriptive image regions. The Class Activation Mapping (CAM) [17] was proposed as a means of highlighting the discriminative image regions for specific output classes in CNNs with global average pooling. Selvaraju et al. developed a generalized version of CAM, named the Gradient-weighted Class Activation Mapping (Grad-CAM) [18], which could be applied to a broader range of CNN models.

The input of an audio classification model is usually a time-frequency representation extracted from the raw audio waveform. Among the various types of time-frequency representations, the logarithmic-magnitude Mel-scale filter bank (abbreviated as log-Mel) feature is widely adopted. Similar to spectrogram, a log-Mel feature is a visual representation of the frequency content of sounds as they vary with time. Given an audio signal with audible sound events such as "bird singing", "speech", "applause", these sound events can also be identified in the corresponding log-Mel feature based on their distinct visual patterns. From this perspective, we may call a log-Mel feature as an image. Visualization of CAM using the log-Mel "image" allows the comparison between machine perception and human interpretation.

In this paper, we present an attempt to understand how CNN models learn to identify an acoustic scene from log-Mel feature representations. The investigation starts with benchmark systems with log-Mel features and different CNN models. The method of CAM is used to provide visualization of the CNN activation behavior with respect to input features. The observed CAMs for acoustic scene data suggest that CNN classification models tend to emphasize on the overall background sound texture of log-Mel input features, whilst individual sound events in the scene are of less importance. Hence we propose to use the Difference of Gaussian (DoG) and the Sobel operator to pre-process the log-Mel feature to make the background texture information more salient. We also use the method of background drift removing with medium filter as described in [13] as a comparison to our methods. These texture-enhanced features demonstrate an improved performance on ASC.

2. BACKGROUND

2.1. Class Activation Mapping

The class activation mapping [17] highlights the class-specific discriminative regions in the input image. It can help understand the CNN behavior and visualize the internal representation of CNNs. It can also be used for weakly supervised object localization task. However, the CAM is only applicable to CNNs with global average pooling (GAP). Suppose we have a trained CNN network with global average pooling. There are C output classes. The number of

channels in the last convolutional layer is K. The point (x, y) in the k^{th} feature map before GAP as $f_k(x, y)$. The weight in the output layer is denoted as w_k^c, indicating the importance of the k^{th} feature map for class c. Then the classification score of class c (before the softmax) is given by

$$y^c = \sum_k w_k^c \sum_{x,y} f_k(x, y) = \sum_{x,y} \sum_k w_k^c f_k(x, y). \quad (1)$$

Based on equation 1, the spatial elements of class activation map M_c for class c is given by

$$M_c(x, y) = \sum_k w_k^c f_k(x, y). \quad (2)$$

The Gradient-weighted Class Activation Mapping (Grad-CAM) [18] is a strict generalization of CAM. It replaces the weight of each activation map w_k^c with the average gradient back-propagated to each feature map, which is given by

$$\alpha_k^c = \frac{1}{Z} \sum_{i,j} \frac{\partial y^c}{\partial f_k(i, j)}, \quad (3)$$

where Z is the number of pixels in the feature map. Notice that f_k in Grad-CAM can be from any convolutional layers in CNN, not limited to the last convolutional layer. Thus, the Grad-CAM can be applied to a larger variety of CNN models, such as those with fully connected layers (e.g. AlexNet, VGG).

In this paper, we propose to use Grad-CAM to analyze the trained CNN models for ASC task. Through empirical analysis of class activation maps w.r.t. the ground-truth scene classes, we argue that CNN models are more focusing on the overall background sound texture for classifying acoustic scenes. The distinct sound events (foreground) are usually of less importance in classification.

2.2. Sound Texture

Texture is described as an attribute that characterize spatial arrangement of pixel intensities in specific regions of an image. In the area of computer vision, texture analysis is a well studied topic [19, 20, 21, 22].

For audio signal, the notion of "sound texture" has not been seriously discussed. An visual analogy of sound texture given by Saint-Arnaud et al. [23] is that sound texture is like a wallpaper which has local structure and randomness, while from a large scale the fine structure characteristics must remain constant. There were a number of studies on sound texture modeling [24, 25, 26], and commonly mentioned sound textures refer to wind, traffic, and crowd sounds.

In an acoustic scene, there exist various sound sources, which contribute to a mixture of diverse sound events. In audio recordings from acoustic scenes, persistent environment sounds with certain sound textures, e.g., crowd, traffic, form "background" of the scenes. Whilst certain sparsely occurred sound events, e.g., bird singing, human coughing, are more noticeable and could be regarded as distinct "foreground" sounds.

2.3. Feature Preprocessing Methods

2.3.1. Difference of Gaussian

The Difference of Gaussian (DoG) is a well-known method of edge detection in image processing. Briefly speaking, the DoG filtering includes two steps: blurring an image using two Gaussian kernels of different standard deviations, and subtracting one blurred image from another to obtain the edge image. The purpose of Gaussian kernel is to suppress the high (spatial) frequency information (which serves as a low-pass filter). The value of standard deviation decides the range of frequency being suppressed. DoG essentially acts like a band-pass filter. It removes not only high (spatial) frequency noise, but also homogeneous regions in the image.

2.3.2. Sobel Operator

The Sobel operator [27] is commonly used for edge detection in computer vision. It comprises two 3×3 convolution kernels, which are used to obtain the gradient approximations in the horizontal direction (G_x) and vertical direction (G_y). For an image A, we have

$$G_x = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} * A, \quad G_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * A. \quad (4)$$

The gradient approximations in different directions can be combined as G, as the result of Sobel filtering:

$$G = \sqrt{G_x^2 + G_y^2}. \quad (5)$$

2.3.3. Removing Background Drift Using Medium Filter

Median filtering is useful in distinguish objects in an image with transitional background. By subtracting the medium-filtered image from the original one, the background drift is removed and those sharp changes (edges) are preserved [28]. For the ASC task, median filtering was found to be very effective in feature pre-processing [13], though the determination of kernel size for optimal performance is not straightforward.

3. ACOUSTIC SCENE CLASSIFICATION SYSTEM

3.1. System Design

Experiments on scene visualization and classification are all based on the TUT Acoustic Scenes 2017 database [2]. This database was adopted for the DCASE 2017 ASC challenge. It has two subsets: the development dataset (for model training and cross validation) and the evaluation dataset (for performance evaluation).

All audio samples in the dataset are 10-second long. They are cut into 1-second segments with 0.5 second overlapping. Short-Time Fourier Transform (STFT) is applied to each of the 1-second segments, with window length of 25ms, window shift of 10ms and FFT length of 2048. 128-dimension log-Mel filterbank features are derived from the FFT spectrum for each frame. Feature components of all frequency bins are normalized to have zero mean and unit variance based on training data statistics.

The CNN model receives the log-Mel feature image of a 1-second segment as the input, and generates a classification score for the segment. The classification score for a 10-second audio sample is obtained by averaging the segment-level scores.

3.2. Model Structure

We examine the performance of two different CNN models. The CNN-FC model as detailed in Table 1 is inspired by the AlexNet [29] and VGG [30] model. After the last convolutional layer, the feature maps are flattened to obtain the input for the fully connected layer.

Table 1: The CNN-FC model structure.

	Input 1x100x128
1	3x3 Convolution (pad-1, stride-1)-64-BN-ReLU
2	3x3 Max Pooling (stride-2)
3	3x3 Convolution (pad-1, stride-1)-192-BN-ReLU
4	3x3 Max Pooling (stride-2)
5	3x3 Convolution (pad-1, stride-1)-384-BN-ReLU
6	3x3 Max Pooling (stride-2)
7	3x3 Convolution (pad-1, stride-1)-256-BN-ReLU
8	3x3 Convolution (pad-1, stride-1)-256-BN-ReLU
9	3x3 Max Pooling (stride-2)
	Flattening
10	Dropout (p=0.5)
11	Fully Connected (dim-2048)-BN-ReLU
12	Dropout (p=0.5)
13	Fully Connected (dim-2048)-BN-ReLU
14	15-way SoftMax

Table 2: The CNN-GAP model structure.

	Input 1x100x128
1	3x3 Convolution (pad-1, stride-1)-64-BN-ReLU
2	3x3 Max Pooling (stride-2)
3	3x3 Convolution (pad-1, stride-1)-192-BN-ReLU
4	3x3 Max Pooling (stride-2)
5	3x3 Convolution (pad-1, stride-1)-384-BN-ReLU
6	3x3 Convolution (pad-1, stride-1)-256-BN-ReLU
7	3x3 Convolution (pad-1, stride-1)-256-BN-ReLU
8	3x3 Max Pooling (stride-2)
9	Global Average Pooling
10	15-way SoftMax

The CNN-GAP model described in Table 2 is constructed by replacing the fully connected part in CNN-FC model with a global average pooling layer, and removing one of the max pooling layer to obtain higher resolution for CAMs. Global average pooling (GAP) has been proven to be a good regularizer for CNNs in image classification [31]. GAP is also used in CNNs with audio input feature [32, 33, 34, 35]. The same setup for training and testing is adopted for both models unless stated otherwise.

4. VISUALIZATION WITH CLASS ACTIVATION MAPS

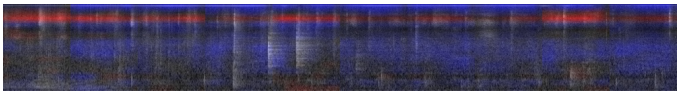
For a given audio segment (1-second long in this study), the short-time log-MEL features could be viewed as a gray-scale image, with the x axis and the y axis representing time and frequency respectively. The image is combined with class activation maps for localizing the discriminative time-frequency regions. The activations are derived for the ground-truth scene class, and thus can be used to represent the input patterns learned by the CNN.

The proposed CAM visualization of audio segment is created by mixing 3 image components. The first component is the gray-scale log-MEL image. The time-frequency regions that positively influence the classification score of the ground-truth scene class are indicated by a semi-transparent image in red color. The negative activations are viewed as another semi-transparent image of blue color. Being different from [18], both positive and negative activations are included for the observation of acoustic scene features.

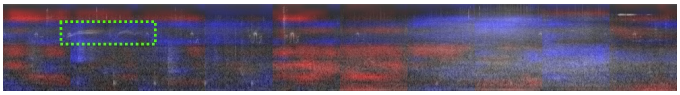
Figure 1 gives a few examples of gradient-weighted CAM visualization derived from the 8^{th} layer feature maps in CNN-FC model. These 10-second audio samples are from the training set of DCASE 2017 dataset. In Figure 1b, the white horizontal lines during the first 3 seconds (inside the green dashed line rectangle) are "bird singing" sounds. It is noted that these distinct sound events are not associated with strong positive (red) or negative (blue) activation. In other words, in CNN classification, these sounds are not regarded as representative patterns for the residential area scene.

Figure 2 shows the examples of CAM visualization derived from the 7^{th} layer feature maps in CNN-GAP model. As we can see from these examples, the magnitudes of positive activations are much higher than those of negative activations. The activations are concentrated on a few frequency bins, unlike the CNN-FC model. In addition, the eye-catching bright lines (foreground sounds) in the log-MEL images are associated with low activation intensity. This suggests that the CNN-GAP model performs classification based on the background "texture" of input image.

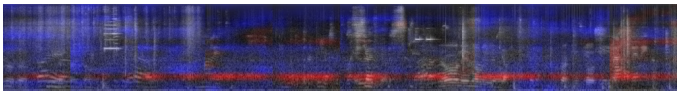
It frequently happens that the regions of distinct sound events in the log-MEL images have small activation intensity, which might be counter-intuitive. However, further investigation is needed to find out if these sound events are really trivial for classification, or it is because the CNN models fail to learn these patterns.



(a) metro station



(b) residential area



(c) train

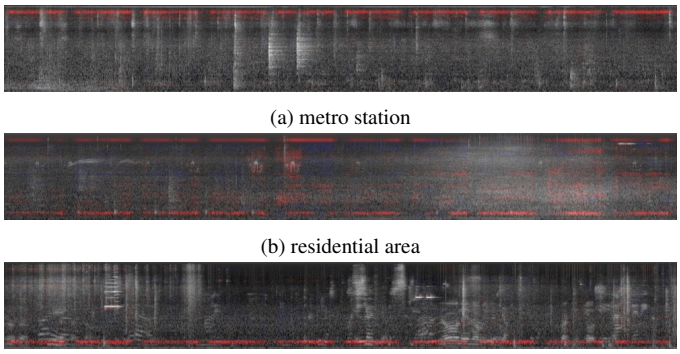
Fig. 1: CAM visualizations w.r.t the ground-truth scene classes. They are derived from the CNN-FC model with log-Mel input. The 3 audio samples are recorded in (a) metro station, (b) residential area and (c) train respectively. High energy regions (distinct sound events) are not strongly activated. It seems that the model is trying to learn the texture of background sounds.

5. ENHANCING THE EDGE INFORMATION IN LOG-MEL IMAGES

5.1. Edge-Enhanced Features

We propose to use DoG and Sobel operator to enhance the edge information in the input images, making the background texture more salient. Figure 3 gives a few examples of edge-enhanced images and the corresponding unenhanced ones.

To obtain the DoG enhanced image, we apply Gaussian filter with standard deviation 1 to the original log-Mel image. Then we apply another Gaussian filter with standard deviation $\sqrt{2}$ on the original image to obtain another blurred image. Subtraction between these two blurred image gives the result of DoG. DoG is able to remove high spatial-frequency components and homogeneous regions of images.



(a) metro station

(b) residential area

(c) train

Fig. 2: CAM visualizations w.r.t the ground-truth scene classes. They are derived from the CNN-GAP model with log-Mel input. The audio samples are the same as those in Figure 1. Different from CNN-FC model, there is little negative activation for CAMs derived from CNN-GAP model.

The result of Sobel operator is an image with pixel values being equal to the gradient magnitudes of the respective pixels in the original image. Comparing to DoG, the images enhanced by Sobel operator have more fine-grained texture.

The above mentioned edge-enhanced images are compared to the one obtained by median filtering. The kernel size of medium filter is empirically set to $(51, 7)$ where 51 refers to time frames (about 0.5 second) and 7 refrs to frequency bins. It is noted that median filtering process has a high computation cost, i.e., requiring to calculate the median of each $(51, 7)$ input window for each output pixel.

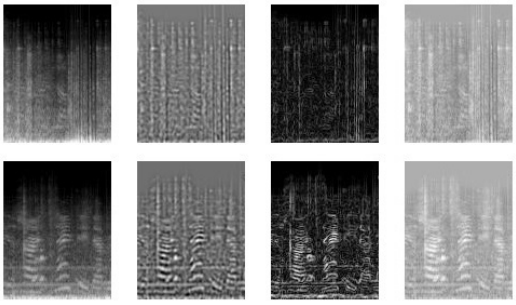


Fig. 3: Illustration of edge-enhanced input features for CNNs. (From left to right) First column: Original log-Mel image; second column: edge-enhanced images with DoG; third column: edge-enhanced images with Sobel operator; fourth column: background drift removed images with medium filter.

5.2. Evaluating the Edge-Enhanced Features

Table 3 shows the accuracy (averaged over 3 trials) for different types of input features. The "LogMel-128" means the 128 dimensional log-Mel feature, which is considered as benchmark feature. "DoG" and "Sobel" refer to the DoG enhanced and Sobel operator enhanced LogMel-128 features, respectively. "Medium" refers to the background-drift-removed LogMel-128 feature (using medium filter). The baseline system accuracy is provided by the DCASE 2017 ASC challenge [36]. It can be seen that applying edge-enhancement techniques leads to significant improvement of classification performance. We also check the CAM visualizations of CNN models with the edge-enhanced input images, and the observations in Section 4 are still valid. While the CNN-FC model and CNN-GAP model are different in visualized patterns of CAM, they show similar performance given the same input feature.

Table 3: Experiment results for CNN-FC and CNN-GAP models trained with different input features. The evaluation data in TUT Acoustic Scenes 2017 database is used for evaluation. Table element is the overall classification accuracy (averaged over 3 trials). The accuracy for the baseline setup is from the DCASE 2017 ASC challenge [36].

Feature\Model	CNN-FC	CNN-GAP	Baseline
Baseline	-	-	0.610
LogMel-128	0.658	0.681	-
DoG	0.720	0.722	-
Sobel	0.701	0.716	-
Medium	0.757	0.754	-

While the performance of using "DoG" feature is not as good as the "Medium" feature, DoG is computationally much more efficient than median filtering. For edge-enhanced features from 100 log-Mel images of size $(1000, 128)$, computing "Medium" features takes 272.02 seconds with kernel size $(51, 7)$ in our computer. If the kernel size is changed to $(3, 3)$, the computation time is 5.7 seconds. On the other hand, applying DoG and Sobel operator takes 0.46 and 0.30 second respectively.

6. CONCLUSION

In this paper, we illustrate the use of class activation mapping for analysis of CNN behavior towards audio features. We find that the distinct sound events in log-Mel features are usually not regarded as representative patterns of acoustic scenes. Regarding ASC task as a sound texture classification problem, we use the DoG, Sobel operator and background drift removing to enhance the edge information in the log-Mel image. Using these methods, the model performance is improved significantly compared to the benchmark.

7. ACKNOWLEDGEMENTS

This research is partially supported by a GRF project grant (Ref: CUHK14227216) from the Hong Kong Research Grants Council and a direct grant from Research Committee of the Chinese University of Hong Kong.

8. REFERENCES

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, May 2015.

[2] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference 2016*, Budapest, Hungary, 2016.

[3] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and svm for acoustic scene classification,"

in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2013, pp. 1–4.

[4] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1, Jan 2013.

[5] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *2015 International Joint Conference on Neural Networks (IJCNN)*, July 2015, pp. 1–7.

[6] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent Neural Networks for Polyphonic Sound Event Detection in Real Life Recordings," *ArXiv e-prints*, Apr. 2016.

[7] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, et al., "CNN architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing*. 2017.

[8] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "A joint detection-classification model for audio tagging of weakly labelled data," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 641–645, 2017.

[9] Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, et al., "Unsupervised feature learning based on deep models for environmental audio tagging," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1230–1241, June 2017.

[10] E. Cakir, T. Heittola, and T. Virtanen, "Domestic audio tagging with convolutional neural networks," Tech. Rep., DCASE2016 Challenge, September 2016.

[11] S. Mun, S. Park, D. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane," Tech. Rep., DCASE2017 Challenge, September 2017.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014.

[13] Y. Han and J. Park, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, November 2017, pp. 46–50.

[14] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," *ArXiv e-prints*, Nov. 2013.

[15] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 2018–2025.

[16] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for Simplicity: The All Convolutional Net," *ArXiv e-prints*, Dec. 2014.

[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," *ArXiv e-prints*, Dec. 2015.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *ArXiv e-prints*, Oct. 2016.

[19] L. Liu and P. Fieguth, "Texture classification from random features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 574–586, March 2012.

[20] D. Puig, M. A. Garcia, and J. Melendez, "Application-independent feature selection for texture classification," *Pattern Recogn.*, vol. 43, no. 10, pp. 3282–3297, Oct. 2010.

[21] X. Chen, X. Zeng, and D. van Alphen, "Multi-class feature selection for texture classification," *Pattern Recogn. Lett.*, vol. 27, no. 14, pp. 1685–1691, Oct. 2006.

[22] A. H. Bhalerao and N. M. Rajpoot, "Discriminant feature selection for texture classification," in *Proc. British Machine Vision Conference*, 2003.

[23] N. Saint-arnaud and K. Popat, "Analysis and synthesis of sound textures," in *Readings in Computational Auditory Scene Analysis*, 1995, pp. 125–131.

[24] D. Schwarz, "State of the art in sound texture synthesis," in *Proceedings of the 14th International Conference on Digital Audio Effects*, September 2011.

[25] M. Athineos and D. P. W. Ellis, "Sound texture modelling with linear prediction in both time and frequency domains," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[26] W. Liu, G. Liu, X. Ji, J. Zhai, and Y. Dai, "Sound texture generative model guided by a lossless Mel-frequency convolutional neural network," *IEEE Access*, vol. 6, pp. 48030–48041, 2018.

[27] I. Sobel and G. Feldman, "An isotropic 3x3 gradient operator," in *Stanford Artificial Intelligence Project (SAIL)*, 1968.

[28] A. W. Moore and J. W. Jorgenson, "Median filtering for removal of low-frequency background drift," *Analytical Chemistry*, vol. 65, no. 2, pp. 188–191, 1993.

[29] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *ArXiv e-prints*, Apr. 2014.

[30] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *ArXiv e-prints*, Sept. 2014.

[31] M. Lin, Q. Chen, and S. Yan, "Network In Network," *ArXiv e-prints*, Dec. 2013.

[32] Y. Sakashita and M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," Tech. Rep., DCASE2018 Challenge, September 2018.

[33] M. Dorfer, B. Lehner, H. Eghbal-zadeh, H. Christop, P. Fabian, et al., "Acoustic scene classification with fully convolutional neural networks and I-vectors," Tech. Rep., DCASE2018 Challenge, September 2018.

[34] H. Zeinali, L. Burget, and H. Cernocky, "Convolutional neural networks and X-vector embedding for DCASE2018 acoustic scene classification challenge," Tech. Rep., DCASE2018 Challenge, September 2018.

[35] Y. Wu and T. Lee, "Reducing model complexity for DNN based large-scale audio classification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, 2018.

[36] T. Heittola and A. Mesaros, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," Tech. Rep., DCASE2017 Challenge, September 2017.