Hybrid Constant-Q Transform Based CNN Ensemble for Acoustic Scene Classification

Mou Wang, Rui Wang, Xiao-Lei Zhang, and Susanto Rahardja
Northwestern Polytechnical University, Xi'an, China
E-mail: {wangmou21, wangrui2018}@mail.nwpu.edu.cn, {xiaolei.zhang, susanto}@nwpu.edu.cn

***Abstract*—Acoustic scene classification (ASC) has attracted much attention in recent years. In previous studies, the most common architecture is convolutional neural network (CNN) fed by three main features, i.e. log-mel energies, harmonic-percussive source separation (HPSS) and constant-Q transform (CQT). In this paper, we present a hybrid constant-Q transform (HCQT) based CNN system for ASC. Specifically, we first extract CQT and HCQT from each audio clip as the acoustic features, as well as other several features such as Mel-frequency cepstral coefficients, log-mel energies and its HPSS. Then, we feed those features into 5-layer or 9-layer CNNs with average pooling separately. Considering different features that have complementary information with each other, we further develop several methods to integrate the outputs of the CNNs, including averaging, weighted averaging, random forests and extremely randomized trees. To the best of our knowledge, this is the first time HCQT based method is being used for ASC. Essentially, the method combines two CQTs with different resolutions for remedying the high-frequency bins of the traditional CQT. In addition, we investigate different ensemble strategies of the CNN models thoroughly. We evaluated the proposed system in the DCASE 2019 challenge. Experimental results show that HCQT is more effective than the conventional CQT. Furthermore, the accuracies of our system on the validation and leaderboard datasets are 77.5% and 79.3% respectively, which outperforms the two comparison baselines significantly.**

I. Introduction

Acoustic scene classification (ASC) aims at recognizing the physical or social meanings of the acoustic environments of sounds in places such as beach, restaurant, road, etc [1], [2]. ASC is important in many applications, such as robotic navigation [3], context-aware services [4], and surveillance [5], [6]. Recently, the ASC research has drawn much attention. For example, the technical program committee of the IEEE audio and acoustic signal processing has recently hosted a series of challenges, named detection and classification of acoustic scenes and events (DCASE), for developing sound classification and detection systems [7], [8]. DCASE is the first large-scale challenge of that actively promotes research in ASC.

An ASC method is composed of acoustic features and classifiers. Most ASC research focus on developing powerful classifiers. Early works use conventional machine learning models, such as Gaussian mixture model [9], support vector machine [10], factor analysis (i.e. i-vectors) [11], and non-negative matrix factorization [12]. Recently, state-of-the-art methods are mostly based on convolution neural networks (CNN) [7], [13]–[16]. For example, Kong et al. [7] proposed generic cross-task systems based on CNN for all tasks of DCASE 2019. As a result, classifiers have been largely improved. On the other side, little attention has been paid on developing discriminant acoustic features which are to our knowledge at least as important as the classifiers.

Most previous ASC research uses handcraft features, such as Mel-frequency cepstral coefficients (MFCC) [9], [11], linear prediction cepstral coefficient [17], and perceptual linear predictive [17]. Some features that are widely used for image processing, such as the histogram of gradients and local binary patterns [18], are also effective for ASC. Log-mel energies [7], [13], [14], [19], the harmonic-percussive source separation (HPSS) of log-mel energies [13], [15], and constant Q transformation (CQT) [14], [20] are three widely used features in recent ASC research.

In this paper, we introduce a new acoustic feature, named hybrid constant-Q transform (HCQT), to a CNN-based ASC system. HCQT combines two CQTs with different resolutions for remedying the high-frequency bins of the traditional CQT. The system first extracts HCQT, as well as other several features such as MFCC, log-Mel energies and its HPSS, from an audio clip. Then, it trains a CNN model on each acoustic feature independently. Finally, it fuses the outputs of all CNN models for the final prediction. The fusion methods include averaging, weighted averaging, random forests, and extremely randomized trees. The contributions of this paper lie in the following two aspects:

- First, we introduced the usage of HCQT to the study of ASC for the first time.
- Second, we investigated different fusion strategies of the CNN models thoroughly.
- Third, we proved empirically that HCQT is significantly better than the widely used CQT in all evaluation scenarios.

To evaluate the effectiveness of HCQT, a system that utilizes CQT is selected for comparison. The CQT in the system is then replaced with the HCQT and evaluated on the subtask A of the task 1 of the DCASE 2019 challenge. Experimental results show that HCQT produces over 2% absolute classification accuracy improvement over CQT when they are used alone, and produces at least about 1% higher classification accuracy than CQT when they are used together with other acoustic features. Moreover, averaging is the best fusion strategy among the fusion strategies. Finally, our system is

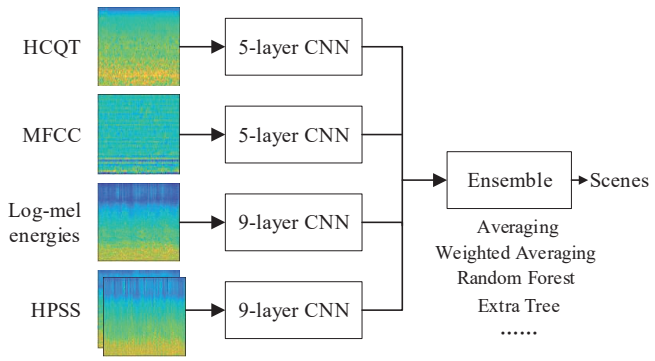

Fig. 1. The architecture of the proposed system.

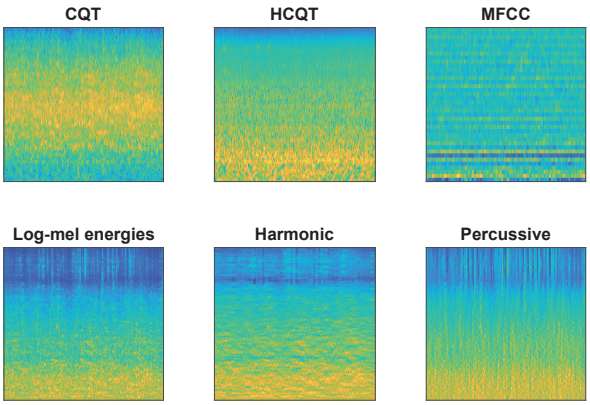

Fig. 2. Visualizations of the acoustic features of the proposed system for the "airport" scene.

significantly better than its individual CNN components and the official baseline.

This paper is organized as follows. Section 2 presents the proposed method. Section 3 presents the experiments. Section 4 concludes this paper.

II. Proposed Method

In this section, an overview of the system is presented in Section II-A. Subsequently, the detailed description of the HCQT features, CNN classifiers and ensemble strategies are explained in Sections II-B, II-C, and II-D respectively.

A. System Overview

As shown in Fig. 1, the proposed system contains three components—feature extractor, CNN classifiers, and classifier ensemble. The feature extractor extracts four kinds of features from each recording, which are HCQT, log-Mel energies, HPSS and MFCC. Then, we input each feature into a CNN. Because the dimensions of HCQT and MFCC are relatively low, their CNNs contain only 5 layers. On the other side, because the dimensions of HPSS and log-Mel energies are high, we train two 9-layer CNN for each of them. Finally, we combine the probability outputs of all CNNs by different ensemble methods including averaging, weighted averaging, random forest or extremely randomized trees.

Here we introduce the log-Mel energies, HPSS and MFCC acoustic features briefly as follows, leaving the description of HCQT in Section II-B. MFCC is a common acoustic feature. In this paper, MFCC consists of 20 Mel-filter bins with a pre-emphasized Hanning window. Log-Mel energies are the most popular feature for ASC [13]. They first extract an STFT spectrogram, and then get Mel-energies by applying a Mel-filter bank of 256 filters to the spectrogram, where the cut-off frequencies of the Mel-filters are from 50 Hz to 14 kHz. At last, the log-Mel energies are obtained by applying the logarithm operation to the Mel-energies. HPSS explores the harmonic and percussive aspects of a sound separately [15]. Here we first apply HPSS to the log-Mel energies, and then concatenate the harmonic and percussive parts of the log-mel energies as the HPSS feature.

B. Hybrid Constant-Q Transform

HCQT is a variant of CQT. CQT was originally proposed in [21]. Its importance has been demonstrated in music analysis. Recently, CQT has been frequently used in the ASC research [22], [23]. It mimics the behavior human auditory system. Specifically, suppose it has K logarithmically spaced filters whose center frequencies and spectral widths are denoted as $\{f_k\}_{k=1}^K$ and $\{\delta_k\}_{k=1}^K$ respectively. The spectral widths of the filters have the following connection:

$$\delta_k = 2^{1/b} \cdot \delta_{k-1} = (2^{1/b})^k \cdot \delta_{\min}, \tag{1}$$

where $\delta_{\min}$ is the band width of the filter with the lowest center frequency, and b is the number of filters per octave. Quality factor is defined as the ratio of the center frequency to the band-width, i.e. $Q = f_k/\delta_k$. The Q-factors of all frequency bins of CQT are a constant. Suppose the sampling frequency is s, then the window length of each bin is a function of the bin's index k:

$$N[k] = \frac{s}{\delta_k} = Q\frac{s}{f_k}. \tag{2}$$

Because s/f_k is the number of samples processed per cycle at frequency f_k, Q equals to the number of integer cycles processed at frequency f_k. Therefore, the digital frequency in CQT becomes $\frac{2\pi Q}{N[k]}$. For example, the Hanning window for CQT is:

$$W[k,n] = 0.5 - 0.5\cos\frac{2\pi n}{N[k]-1}, \quad 0 \le n \le N[k]-1 \tag{3}$$

Finally, CQT is calculated as follows:

$$X_{\text{CQT}}[k,m] = \frac{1}{N[k]}\sum_{n=0}^{N[k]-1} W[k,n-m]x[n]e^{\frac{-j2\pi Qn}{N[k]}}. \tag{4}$$

where $x[n]$ is the n-th sample of a speech frame in the time domain. CQT is essentially a wavelet transform. It has high time-resolution at the high frequency bins, while it has high frequency-resolution at the low frequency bins.

HCQT on the other hand consists of two variants of CQT. We assume that the frame shift contains L samples in the time

TABLE I
ARCHITECTURE OF THE 5-LAYER CNN. THE NUMBERS IN THE BRACKETS ARE THE KERNEL SIZE AT THE CONVOLUTION/POOLING LAYERS. THE NUMBERS 128, 256, AND 512 ARE THE NUMBER OF FILTERS IN THE CONVOLUTION BLOCKS. THE NUMBER "10" IS THE NUMBER OF THE HIDDEN UNITS IN THE OUTPUT LAYER.

Input: features
(5x5) Conv2D-128-BN-ReLU
(2x2) Average Pooling
(5x5) Conv2D-128-BN-ReLU
(2x2) Average Pooling
(5x5) Conv2D-256-BN-ReLU
(2x2) Average Pooling
(5x5) Conv2D-512-BN-ReLU
(2x2) Average Pooling
Dense-10-SoftMax

TABLE II
ARCHITECTURE OF THE 9-LAYER CNN.

Input: features
(3x3) Conv2D-64-BN-ReLU
(3x3) Conv2D-64-BN-ReLU
(2x2) Average Pooling
(3x3) Conv2D-64-BN-ReLU
(3x3) Conv2D-64-BN-ReLU
(2x2) Average Pooling
(3x3) Conv2D-64-BN-ReLU
(3x3) Conv2D-64-BN-ReLU
(2x2) Average Pooling
(3x3) Conv2D-64-BN-ReLU
(3x3) Conv2D-64-BN-ReLU
(2x2) Average Pooling
Dense-10-SoftMax

domain. We further select the k_c-th filter that satisfies:

$$N[k_c] = 2L. \tag{5}$$

We regard the frequencies higher than f_{k_c} as high frequencies, and the frequencies lower than f_{k_c} as low frequencies. For the high frequency part of HCQT, it filters the STFT spectrogram by the filter bank of the high frequency part of CQT. For the low frequency part of HCQT, it uses the standard CQT directly. Compared to CQT, HCQT is more computationally efficient. In this paper, we set both HCQT and CQT to 84 bins, among which 51 bins are low frequency ones and the rest 33 bins are high frequency ones. A visualized comparison of the extracted features of the proposed system is shown in Fig. 2.

C. Convolutional Neural Networks

We follow the CNN framework in [7]. Specifically, because the input acoustic features have different dimensions, it is needed to train two kinds of CNNs with different depths—5-layer CNNs and 9-layer CNNs. The 5-layer CNN is designed for the low-dimensional HCQT and MFCC. Its parameter setting, which is similar to AlexNet [24], is summarized in Table I. Specifically, it consists of 4 convolutional layers, each of which consists of a convolutional kernel with a kernel size of 5×5, batch normalization (BN), rectified linear units (ReLU), and average pooling with a kernel size of 2×2. The output layer is a fully connected softmax layer. The cross entropy loss is used as the training criterion.

If the acoustic features are high-dimensional, such as the log-Mel energies and HPSS, then we seek ways to train deeper CNNs than that in Table I. Motivated by the VGG network [25], we decompose each convolutional layer with the 5×5 kernel in Table I to a convolutional block that consists of two cascaded convolutional layers with 3×3 kernels. The architecture of the 9-layer CNN model is described in Table II.

D. Ensemble Methods

An ensemble of the CNN models leads to better performance than its components. However, the problem of identifying a simple and good ensemble strategy seems not well explored. In this paper, we investigate the following four ensemble strategies:

- **CNN-AVER:** Suppose the probabilistic outputs of the CNNs as $P = [p_1, ..., p_m, ..., p_M]$ where p_m refer to the output of the m-th CNN and M is the number of CNNs. CNN-AVER averages the probabilistic outputs P of the CNNs directly [26]. The predicted class of an audio scene segment is:

$$c = \arg\max \frac{1}{M} \sum_{m=1}^{M} p_m. \tag{6}$$

- **CNN-WAVER:** We denote the classification accuracies of the CNN models on a validation set as $A = [a_1, ..., a_m, ...a_M]$, where a_m is the accuracy of the m-th CNN. Then, the predicted class of an audio scene segment produced by CNN-WAVER [26] is calculated by:

$$c = \arg\max \frac{1}{M} \sum_{m=1}^{M} a_m \circ p_m. \tag{7}$$

  where the operator "$\circ$" denotes the element-wise product.
- **CNN-RF:** We take the probabilistic outputs of the CNNs on a training set, denoted as $P_{\text{train}} = [p_1, ..., p_m, ..., p_M]$, as a new feature for training a random forest [27].
- **CNN-ET:** We take P_{train} as a new feature for training an extremely randomized tree [28].

III. EXPERIMENTS

A. Datasets

We conducted experiments on the subtask A of task 1 of the DCASE 2019 challenge. The task requires the participants to use the TAU Urban Acoustic Scenes 2019 dataset [29] as the development set, and evaluates the performance of the models on a leaderboard set. The development set consists of 10 acoustic scenes, including airport, shopping mall, metro station, pedestrian street, public square, street with traffic, tram, bus, metro and urban park. It consists of 14400 audio segments, each of which is 10 seconds long. Therefore, the total time of the dataset is 40 hours. It was partitioned into a training subset and a test subset by the DCASE organizers, which contain 9185 and 4185 segments, respectively.

TABLE III
ACC COMPARISON OF THE CNN MODELS WITH CQT AND HCQT RESPECTIVELY ON THE TEST SUBSET.

	5-layer CNN	9-layer CNN
CQT	0.643	0.652
HCQT	0.679	0.667

TABLE IV
ACC COMPARISON OF THE ENSEMBLE MODELS THAT FUSE THE CQT/HCQT-BASED CNN WITH THE MEL/HPSS/MFCC-BASED CNN ON THE TEST SUBSET. THE SYMBOL "+" MEANS THE FUSION OF THE CNN MODELS WITH DIFFERENT ACOUSTIC FEATURES. THE LINE WHERE "—" EXISTS PRESENTS THE PERFORMANCE OF THE CNN MODELS WITH THE MEL, HPSS, OR MFCC FEATURE ALONE. THE TERM "MEL" DENOTES THE LOG-MEL ENERGIES.

	+ Mel	+ HPSS	+ MFCC
—	0.721	0.724	0.663
CQT	0.737	0.726	0.713
HCQT	**0.746**	**0.746**	**0.721**
CQT + HCQT	0.729	0.746	0.721

We reduced the sampling rate of the audio recordings from 48 kHz to 32 kHz, given the fact that the audio recordings with a sampling rate of 32 kHz contain most of the energy [7]. We set the frame length to 64 ms and frame shift to 15 ms, respectively.

B. Comparison Methods

First of all, we use the feature name to represent a CNN method with the acoustic feature. For example, we denote "HCQT" as a CNN model with the HCQT feature, and denote "HCQT+CQT" as an ensemble of two CNNs with the HCQT and CQT as the acoustic features respectively. We denote the proposed method with all four acoustic features as "*HCQT-based ensemble*". We further denote the comparison method that is the same as the proposed one except that it uses CQT to replace HCQT as "*CQT-based ensemble*", and the comparison method that uses all four acoustic features together with the CQT feature as "*HCQT+CQT-based ensemble*".

To evaluate the general performance of the proposed method. We compared the proposed method with two public baselines. The first baseline, denoted as *baseline-DCASE*, is an official baseline provided by the DCASE2019 organizer [29]. It uses the log-Mel energies with 40 bins as the acoustic feature, and trains a 3-layer CNN to do ASC. Readers are encouraged to read [29] for the details of the comparison method.

The second baseline, denoted as *baseline-cvssp*, is a system proposed by Kong et al. [7]. It first extracts 64-dimensional log-Mel energies with a frame-length of 32 ms and a frame-shift of 16 ms, and then uses a CNN with the same architecture as that in Table II for ASC.

We take classification accuracy (ACC) as the evaluation metric. If not specified, all ensemble methods use averaging as the ensemble strategy.

TABLE V
ACC COMPARISON OF FIVE CNN ENSEMBLE METHODS ON THE TEST SUBSET AND LEADERBOARD DATASET.

	Test	Leaderboard
Baseline-DCASE [29], [30]	0.625	0.643
Baseline-CVSSP [7], [30]	0.703	0.693
CQT-based ensemble	0.764	—
HCQT-based ensemble	**0.775**	**0.793**
CQT+HCQT-based ensemble	0.770	—

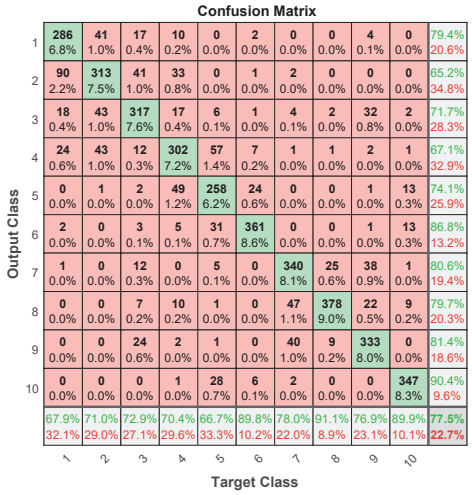

Fig. 3. Confusion matrix of the HCQT-based ensemble on the test subset. X-axis indicates the predicted label. Y-axis indicates the ground-truth label.

C. Results

To show the advantage of HCQT over CQT, we take CQT and HCQT as the acoustic feature of two CNN models respectively. The comparison result is shown in Table III. From the table, we see that (i) HCQT leads to better performance than CQT, and (ii) the conclusion is consistent across different architectures of CNN. As a byproduct, we find that, when HCQT is used, the 5-layer CNN performs better than the 9-layer CNN. Table IV shows the comparison result the CNN ensembles that take different groups of acoustic features as the input. From the table, we observe that HCQT is a better supplement to the other acoustic features than CQT, and combining CQT and HCQT together does not improve the overall performance. As a conclusion, HCQT is a better feature than CQT for ASC.

We further compared CQT and HCQT in the system level. The comparison result is shown in Table V. The results of the baseline-DCASE system are from the DCASE website [29] and leaderboard in Kaggle [30] respectively, and The results of the baseline-CVSSP system are from [7] and the same leaderboard. From the table, we further manifest the advantage of HCQT over CQT, and moreover, the proposed HCQT-based ensemble method performs the best among all comparison methods. To further understand the HCQT-based ensemble, we show its confusion matrix on the test subset in Fig. 3.

To discuss the effects of different ensemble strategies on

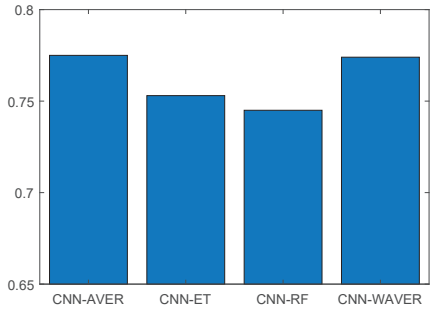

Fig. 4. ACC comparison of the HCQT-based ensemble with four ensemble strategies.

performance, we replace the CNN-AVR strategy of the HCQT-based ensemble by other ensemble strategies in Section II-D. The result is shown in Fig. 4. From the figure, we find that CNN-AVR strategy is the best ensemble strategy.

IV. Conclusions

In this paper, we have proposed a new acoustic feature—HCQT—for ASC. HCQT combines two CQTs with different resolutions together for remedying the high-frequency bins of the traditional CQT. We have further proposed the HCQT-based CNN ensemble. The ensemble method first extracts HCQT, MFCC, log-Mel energies and its HPSS as the input of four CNN models respectively, and then fuses the outputs of all CNN models for the final prediction. Finally, we have discussed four ensemble strategies for ASC, which are CNN-AVR, CNN-WAVER, CNN-RF, and CNN-ET, respectively. We have conducted a systematical comparison on the subtask A of task 1 of the DCASE 2019 ASC challenge. The comparison results show that (i) HCQT is a better feature than the commonly-used CQT, (ii) CNN-AVR is not only simple but also the best ensemble strategy, and (iii) the HCQT-based CNN ensemble is the best system among all comparison method. Particularly, the proposed system is over 7% absolute ACC higher than the public baselines.

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] T. Zhang and J. Wu, "Constrained learned feature extraction for acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1216–1228, Aug 2019.

[3] S. Chu, S. Narayanan, C. . J. Kuo, and M. J. Mataric, "Where am i? scene recognition for mobile robots using audio features," in *2006 IEEE International Conference on Multimedia and Expo*, July 2006, pp. 885–888.

[4] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, Jan 2006.

[5] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, Oct 2005, pp. 158–161.

[6] A. Rakotomamonjy, "Supervised representation learning for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1253–1265, June 2017.

[7] Q. Kong, Y. Cao, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems," *arXiv preprint arXiv:1904.03476*, 2019.

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

[9] G. Takahashi, T. Yamada, S. Makino, and N. Ono, "Acoustic scene classification using deep neural network and frame-concatenated acoustic feature," DCASE2016 Challenge, Tech. Rep., September 2016.

[10] B. Elizalde, A. Kumar, A. Shah, R. Badlani, E. Vincent, B. Raj, and I. Lane, "Experiments on the DCASE challenge 2016: Acoustic scene classification and sound event detection in real life recording," DCASE2016 Challenge, Tech. Rep., September 2016.

[11] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU submissions for DCASE-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks," DCASE2016 Challenge, Tech. Rep., September 2016.

[12] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Supervised nonnegative matrix factorization for acoustic scene classification," DCASE2016 Challenge, Tech. Rep., September 2016.

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

[14] H. Zeinali, L. Burget, and H. Cernocky, "Convolutional neural networks and x-vector embedding for dcase2018 acoustic scene classification challenge," DCASE2018 Challenge, Tech. Rep., September 2018.

[15] Y. Han and J. Park, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," DCASE2017 Challenge, Tech. Rep., September 2017.

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

[17] S. Park, S. Mun, Y. Lee, and H. Ko, "Score fusion of classification systems for acoustic scene classification," DCASE2016 Challenge, Tech. Rep., September 2016.

[18] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of timecfrequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 142–153, Jan 2015.

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

[20] Z. Weiping, Y. Jiantao, X. Xiaotao, L. Xiangtao, and P. Shaohu, "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion," DCASE2017 Challenge, Tech. Rep., September 2017.

[21] J. C. Brown, "Calculation of a constant q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1998.

[22] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of timecfrequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 142–153, Jan 2015.

[23] J. Ye, T. Kobayashi, M. Murakawa, and T. Higuchi, "Acoustic scene classification based on sound textures and events," in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: ACM, 2015, pp. 1291–1294. [Online]. Available: http://doi.acm.org/10.1145/2733373.2806389

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[26] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, March 1998.

[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.
[29] http://dcase.community/challenge2019/.
[30] https://www.kaggle.com/c/dcase2019-task1a-leaderboard/.