ACOUSTIC SCENE CLASSIFICATION IN DCASE 2019 CHALLENGE: CLOSED AND OPEN SET CLASSIFICATION AND DATA MISMATCH SETUPS

Annamaria Mesaros, Toni Heittola, Tuomas Virtanen

Tampere University
Computing Sciences
Korkeakoulunkatu 10, Tampere, Finland
name.surname@tuni.fi

ABSTRACT

Acoustic Scene Classification is a regular task in the DCASE Challenge, with each edition having it as a task. Throughout the years, modifications to the task have included mostly changing the dataset and increasing its size, but recently also more realistic setups have been introduced. In DCASE 2019 Challenge, the Acoustic Scene Classification task includes three subtasks: Subtask A, a closed-set typical supervised classification where all data is recorded with the same device; Subtask B, a closed-set classification setup with mismatched recording devices between training and evaluation data, and Subtask C, an open-set classification setup in which evaluation data could contain acoustic scenes not encountered in the training. In all subtasks, the provided baseline system was significantly outperformed, with top performance being 85.2% for Subtask A, 75.5% for Subtask B, and 67.4% for Subtask C. This paper presents the outcome of DCASE 2019 Challenge Task 1 in terms of submitted systems performance and analysis.

Index Terms— Acoustic Scene Classification, DCASE 2019 Challenge, open set classification

1. INTRODUCTION

Acoustic scene classification is a task of widespread interest in the general topic of environmental audio analysis, and refers to the specific case of classifying environments based on their general acoustic characteristics [1, 2, 3]. Other closely related and popular directions of research include classification of individual sound events from the environment, sound event detection, localization and tagging. Specific applications for acoustic scene classification include services and devices that can benefit of context awareness [4], services or applications for indexing audio content [5], documentary and archival of everyday experience [6], wearable technology, navigation systems for robotics, etc.

As a research area acoustic scene classification is not novel, but has gained traction in recent years due to the wide availability of user devices and applications. However, is not plausible to be able to record training data with all devices or all types of scenes that may be encountered in use conditions. In such situation, the classifiers require methods to handle device mismatch through e.g. domain adaptation, and the ability to detect acoustic scenes unseen in training.

This work has received funding from the European Research Council, grant agreement 637422 EVERYSOUND

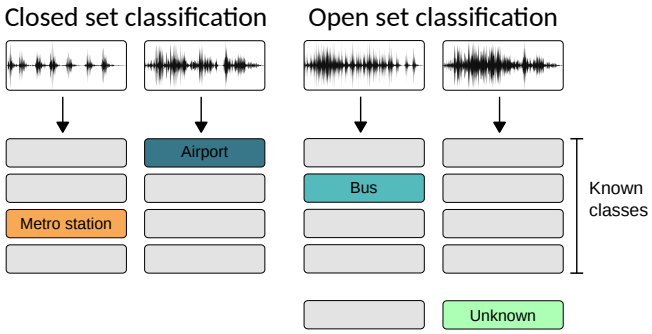

Figure 1: Closed and open-set acoustic scene classification

In DCASE 2019 Challenge, the Acoustic Scene Classification Task includes three subtasks, among which two represent realistic usage cases. Subtask A is a closed-set supervised classification problem where all data is recorded with the same device; Subtask B is a closed-set classification problem with mismatched recording devices between training and evaluation data, and Subtask C is an open-set classification problem in which evaluation data could contain acoustic scenes not encountered in the training.

In this paper we present the task setup and submissions of DCASE 2019 Challenge Task 1. We introduce the three different setups used for the three subtasks, describe the datasets provided for each, and present the challenge submissions. Evaluation and analysis of submitted systems includes general statistics on systems and performance and system characteristics.

The paper is organized as follows: Section 2 presents the task setup including data, rules and baseline system, Section 3 presents the main statistics about the received submissions, while Section 4 presents an analysis of the main trends in the submissions, with details about selected systems. Finally, Section 6 presents the conclusions and ideas for future editions.

2. TASK DESCRIPTION

The goal of acoustic scene classification is to classify a test recording into one of the provided predefined classes that characterizes the environment in which it was recorded. In DCASE 2019 challenge, the Acoustic Scene Classification task presented participants with three different subtasks that required system development for three different situations:

- Subtask A: Acoustic Scene Classification. Classification of data from the same device as the available training data.

- Subtask B: Acoustic Scene Classification with mismatched recording devices. Classification of data recorded with devices different than the training data.
- Subtask C: Open set Acoustic Scene Classification. Classification on data that includes classes not encountered in the training data.

2.1. Dataset

The dataset for this task is the TAU Urban Acoustic Scenes 2019 dataset, consisting of recordings from the following acoustic scenes: airport, indoor shopping mall, metro station, pedestrian street, public square, street with medium level of traffic, travelling by tram, travelling by bus, travelling by underground metro, and urban park. The dataset used for the task is an extension of the TUT 2018 Urban Acoustic Scenes dataset, recorded in multiple cities in Europe. TUT 2018 Urban Acoustic Scenes dataset contains recordings from Barcelona, Helsinki, London, Paris, Stockholm and Vienna, to which TAU 2019 Urban Acoustic Scenes dataset adds Lisbon, Amsterdam, Lyon, Madrid, Milan, and Prague. The recordings were done with four devices simultaneously, denoted in the data as device A (Soundman OKM II Klassik/studio A3 electret binaural microphone), device B (Samsung Galaxy S7), device C (IPhone SE), and device D (GoPro Hero5 Session). The data recording procedure is explained in detail in [13].

Different versions of the dataset are provided for each subtask, together with a training/test partitioning for system development. Generalization properties of systems were tested by presenting in the evaluation set data recorded in cities unseen in training (10 cities in development data, 12 in evaluation). As a special situation, in Subtask C additional data is provided for the open set classification; this consists of the "beach" and "office" classes of TUT Acoustic Scenes 2017 dataset, and other material recorded in 2019. Similarly, data from acoustic scenes other than the 10 mentioned above were present in the evaluation data.

Table 1 summarizes the information about datasets. For each subtask, the development set is split into training/test subsets, created based on the recording location such that the training subset contains approximately 70% of recording locations from each city. The evaluation set was released as audio only, two weeks before the challenge submission deadline; reference annotation is available only to task coordinators for evaluating the systems' performance.

Use of external data was allowed in all subtasks under the conditions that the data is freely accessible and available before the release of the Evaluation dataset. A list of external data sources was provided, and participants had the option to suggest others.

2.2. Evaluation

The submissions were evaluated using classification accuracy calculated as average of the class-wise accuracy, with each segment considered as an independent test sample. Ranking of submissions was done as follows:

- Subtask A used the average accuracy on all evaluation data.
- Subtask B used the average accuracy on devices B and C.
- Subtask C used the weighted average of the accuracy of known classes ACC_{kn} and accuracy of the unknown class ACC_{unk}, with a weight of 0.5 for each:

$$ACC_w = wACC_{kn} + (1 - w)ACC_{unk} \quad (1)$$

During the challenge, public leaderboards were provided for each task through Kaggle InClass competitions. Leaderboards were meant to serve as a development tool for participants, and did not have an official role in the challenge.

2.3. Baseline system

The baseline system implements a convolutional neural network (CNN) based approach using two CNN layers and one fully connected layer, trained using log mel-band energies extracted for the 10-second audio examples. The system is identical to the baseline provided in Task 1 of DCASE 2018 Challenge, and detailed system parameters can be found in [13]. Model selection is done using a validation set of approximately 30% of the original training data. Model performance is evaluated on this validation set after each epoch, and the best performing model is selected.

Specific modifications for subtasks include the use of different training data for the different subtasks and the decision making process for the output. Training of the system for Subtask B was done such that all available audio material (devices A, B and C) was used, with no specific way of treating parallel data. For Subtask C, the system was trained using only the known classes audio material.

The activation function in the output layer for Subtasks A and B is softmax, allowing selection of the most likely class in the closed-set classification problem. For Subtask C, the activation function in the output layer is sigmoid, to allow making the open-set decision based on a threshold; if at least one of the class values is over the threshold of 0.5, the most probable target scene class is chosen, if all values are under the threshold, the unknown scene class is selected.

3. CHALLENGE SUBMISSIONS

The task has received a total number of 146 submissions from 46 teams (maximum 4 submissions per team allowed). Subtask A was the most popular, as expected, with 98 submissions; Subtask B has received 29 submisions, and Subtask C 19.

Subtask A had the best performance of 85.2%, with a 95% confidence interval (CI) between 84.4 and 86.0. Zhang et al. [14] are authors of the four best systems. Koutini et al. [15] ranked 5th - 8th, with their best system having an accuracy of 83.8% (CI 82.9 - 84.6). The McNemar test between the top system of Zhang et al. and Koutini et al. shows that they are significantly different, establishing Zhang et al. as the top system. The baseline system with a performance of 63.3% ranks very low, with only 5 of the 98 submitted systems performing lower.

In Subtask B, Kosmider et al. [16] obtained the highest performance of 75.3% (74.3 - 76.3) on data from devices B and C, and submitted the four best systems. Tied on 4th rank, the system by McDonnell et al. [17] obtained an accuracy of 74.9% (73.9 - 75.9), while on rank 5, the system by Eghbal-zadeh et al. [15] has an accuracy of 74.5% (73.5 - 75.5). McNemar's test shows that the top system by Kosmider et.al and the system by McDonnell et al. make significantly different errors; also McDonnell et al. and Eghbal-zadeh et al. are significantly different according to the same test, therefore even though their confidence intervals overlap, their order in ranking is justified. The baseline system ranks last with a significant gap to the second last, as no effort was made in it to deal with the device mismatch.

The top system in Subtask C, by Zhu et al. [18], has an accuracy of 67.4% (66.8 - 68.1) calculated according to (1), and again the four best systems were submitted by the same team. On rank 5 is

	Subset	Hours	Devices	Observations
Subtask A dataset:	Dev [7]	40	A	Binaural audio, data balanced between classes
TAU Urban Acoustic Scenes 2019	Eval [8]	20	A	Introduced two unseen cities
Subtask B dataset:	Dev [9]	46	A, B, C	Single channel audio, 3h of parallel data provided
TAU Urban Acoustic Scenes 2019 Mobile	Eval [10]	30	A, B, C, D	Introduced two unseen cities, unseen device D
Subtask C dataset:	Dev [11]	44	A	Single channel audio, 4h "unknown" class data
TAU Urban Acoustic Scenes 2019 Open set	Eval [12]	20	A	Introduced two unseen cities, unknown class data

Table 1: Summary of datasets. Complete details for each are included with the data package.

a system by Rakowski et al. [19] with a performance of 64.4% (CI 63.8 - 65.1); this is outside of the confidence interval of the top system, therefore the top system performs significantly better. In this subtask too, the baseline system ranks last.

Figure 2 presents the performance of the top ten teams for Subtasks A and B, and top 5 for Subtask C. Best system per team is selected for the illustration. In the bottom panel, additional details on the system performance are presented: seen vs unseen cities in Subtask A, other devices including unseen device D in Subtask B, and the known vs unknown scenes in Subtask C.

4. ANALYSIS OF SUBMISSIONS

A large majority of submissions for all subtasks used as features log mel energies and used classifiers based on convolutional neural networks. The following statistics are based on the information reported by participants.

4.1. Acoustic Scene Classification

Subtask A includes 85 systems of the 99 (including baseline) that use log mel energies, as standalone features or in combination with other features; the other most common preprocessing technique was harmonic/percussive source separation [20] used by 8 systems of 3 teams. From the 99 systems, 58 reported using mixup [21] as a data augmentation method.

CNNs were part of 82 systems, in many cases as ensemble. Ensembles were very common, with 75 systems reporting use 2 to 40 subsystems. Many ensembles are just multiple CNNs, while in some cases combinations of specific architectures like VGG, Inception and ResNet were used. The most number of systems in an ensemble is 40, with one billion parameters [22]. The system uses a model pre-trained on AudioSet [23] and obtains an accuracy of 80.5%, ranking 15th among the 99 systems.

The top system by Zhang et al. used log mel energies and CQT as feature representations, generative neural network-based augmentation, and an ensemble of 7 CNNs having in total 48M parameters [14]. Among the 7 subsystems, one uses an adversary city adaptation branch that classifies the test samples into the target city, and a gradient reverse layer that makes the output of convolutional layers similar for the same scene class over various city domains. The system has a performance of 77.9% on data from unknown cities, compared to 86.7% on the cities enocuntered in training.

The runner-up team proposed a variety of Receptive-Field-Regularized CNN classifiers, among which one submission was single-model. Separate predictions of the model (snapshots) taken every 5 epochs after 300 epochs in training were used as a way to incorporate more opinions on the test data,resulting in a system with 35M parameters. The best system of the team uses among others a new convolutional layer to create Frequency-Aware Convolutional Neural Networks [15], with filters more specialized in certain frequencies. This ensemble of 7 subsystems has 71M parameters, and its confidence intervals overlap with the single model system.

4.2. Acoustic Scene Classification with mismatched devices

Subtask B has a total of 30 systems including the baseline, of which 29 are CNN-based, the other one using support vector machines. Among all, 25 systems use log mel energies, four use perceptually weighted power spectrogram (one team), and one uses mel-frequency discrete wavelet coefficients; 20 systems use mixup, and 14 have a parameter count over 10M. The most number of systems in an ensemble is 124, belonging to the top system, while the highest parameter count is 727M for an ensemble of 11 systems using 20 snapshots each [15]. Methods for dealing with the device mismatch include domain adaptation and transfer learning, feature transform, spectrum correction, and regularization.

The top systems from Kosmider et al. [16] are based on large ensembles and use a spectrum correction method to account for different frequency responses of the devices in the dataset. The method uses the special feature of the provided development data, namely the temporally aligned recordings from different devices and calculated correction coefficients for devices, using as a reference the average spectrum of devices B and C. The method obtains an accuracy of 75.3 on the data from devices B and C (ranking metric), 80.8% on device A, and 38.6% on the unseen device D.

Also in the top is a simple two-system CNN ensemble by McDonnell et al. that uses multiple forms of regularization that involves aggressively large value for weight decay and not learning batch normalization scale and offset, along with mixup and temporal crop augmentation [17]. The two CNNs use deep residual networks with two pathways, one for high frequencies and one for low frequencies, that were fused prior to the network output. Authors point out that the temporal and frequency axes in spectrograms represent fundamentally different information than for images, and choose no to downsample the frequency axis within the networks. No specific processing of the parallel data from different devices is reported, but the system obtains nevertheless a very balanced performance of the four different devices: 74.9% (73.9 - 75.9) on devices B and C, performs with 79.8% on device A, and 65.2% on device D, which is the highest accuracy obtained on device D among all systems.

4.3. Open set Acoustic Scene Classification

Subtask C has a total of 20 entries, all using log mel energies, with all but the winner team using CNNs. Only six systems have better accuracy for the known classes than the unknown class, indicating a tendency towards optimization for detection of the unknown class.

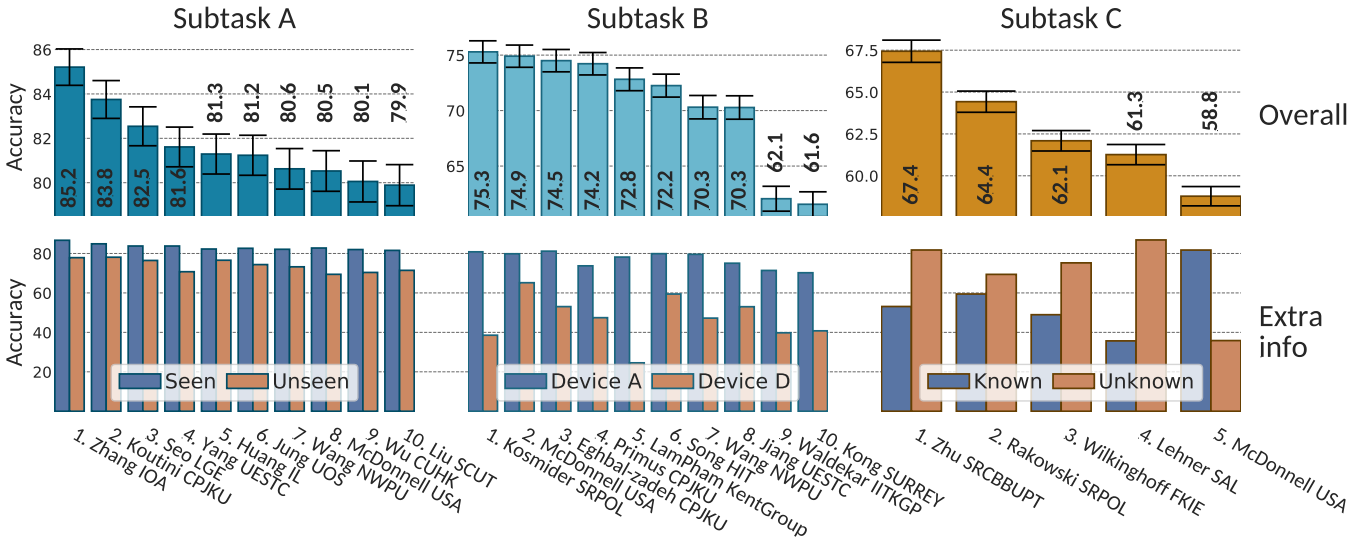

Figure 2: Performance of top teams in each subtask, including confidence intervals

The highest reported number of subsystems in an ensemble for this subtask is 17, but most have only 2 to 4 subsystems.

The top ranked system by Zhu et al. uses CRNNs with self-attention mechanism which are trained on different time divisions of the mel spectrogram. The decision for the unknown class is guided by a threshold of 0.4 on the output layer probabilities for the classes [18]. Their choice of threshold results in a 81.8% accuracy on the unknown class, with 53.1% on the known 10 classes.

Rakowski et al. [19] employed a frequency-aware CNN that preserves the location of features on the frequency axis by applying global pooling only across the temporal dimension, similar to the observations of [17] in Subtask B. The approach resulted in a relatively balanced performance on the known and unknown classes, 59.5% and 69.4% respectively. Lehner et al. [24] used a rejection option for the identification of unknown class, based on the most likely of the ten known classes. They note that the weighted average accuracy used for ranking Subtask C favors aggressive rejection, and for this reason chose the threshold for rejection as the maximum score on the validation data. As a result, they obtained an accuracy of up to 91% on the unknown class, but considerably lower performance on the ten known classes, only 30%.

A notably different approach was proposed by Wilkinghoff et al. [25], which treats the open set classification problem as a combination of convolutional neural networks for closed-set classification and deep convolutional autoencoders for unknown class detection. The method results in a high accuracy on the unknown class at the expense of low accuracy in the closed set (75.2% vs 48.9%).

5. DISCUSSION

One immediate observation about the submissions is that there was little use of external data, with only the four mentioned systems of one team using pretrained models [22]. This is contrary to the feedback of previous challenges that indicated participants wanted to use external data. It is possible that the datasets provided for the task are considered large enough to warrant robust modeling, and therefore use of external data is not necessary.

Compared to 2018 Challenge, novel approaches tailored to use of parallel data have emerged for solving the device mismatch. Among all, the spectrum correction has provided the best performance on the target devices [16], but the best generalization over the four devices was obtained by extensive regularization procedures [17]. The open set classification was tackled by participants in few different ways, with most systems using a threshold. The more distinct approaches treated the unknown class as a separate class [17] or as a subproblem [25]. In most cases, the optimization resulted in emphasis on getting good performance on the unknown class, at the expense of the performance on the ten known classes.

We also want to highlight two approaches to cross-task solutions, one very basic and another one including many techniques to achieve robustness. [26] consists of a generic CNN architecture, similar to the baseline but with more layers, and obtains average performance in Subtask A (53th with 70.5%) but is only better than the baseline system in Subtask B and Subtask C. In contrast, [17] uses more specialized networks and extensive regularization and data augmentation techniques, resulting in a system highly robust to device mismatch (4th in subtask B), having a performance 10 to 15% higher than [26] in all subtasks. These results show that a generic approach, while generally appropriate for straightforward tasks such as the closed set classification, is not suitable for dealing with the same problem in realistic settings, but requires additional techniques to obtain satisfactory performance.

6. CONCLUSIONS

The 2019 Challenge has introduced realistic problems that included evaluation data that contained unseen cities in Subtask A, unseen devices in Subtask B and unseen acoustic scene classes in Subtask C. Acoustic Scene Classification remains the favorite task in the DCASE Challenge, as it offers a textbook problem for audio classification, suitable for beginners in the field. With multiple subtasks of different complexity, the task has also attracted the attention of experienced researchers, and state of the art methods for audio classification are continuously developed within the framework of acoustic scene classification.

7. REFERENCES

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct 2015.

[2] 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.

[3] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, Feb 2018.

[4] B. N. Schilit, N. Adams, and R. Want, "Context-aware computing applications," in *IN PROCEEDINGS OF THE WORKSHOP ON MOBILE COMPUTING SYSTEMS AND APPLICATIONS*. IEEE Computer Society, 1994, pp. 85–90.

[5] Y. F. Phillips, M. Towsey, and P. Roe, "Revealing the ecological content of long-duration audio-recordings of the environment through clustering and visualisation," *PLOS ONE*, vol. 13, no. 3, pp. 1–27, 03 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0193345

[6] M. Droumeva, "Curating everyday life: Approaches to documenting everyday soundscapes," *M/C Journal*, vol. 18, no. 4, 2015.

[7] T. Heittola, A. Mesaros, and T. Virtanen, "TAU Urban Acoustic Scenes 2019, Development dataset," Mar. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2589280

[8] ——, "Tau urban acoustic scenes 2019, evaluation dataset," June 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3063822

[9] ——, "TAU Urban Acoustic Scenes 2019 Mobile, Development dataset," March 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2589332

[10] ——, "TAU Urban Acoustic Scenes 2019 Mobile, Evaluation dataset," June 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3063980

[11] ——, "TAU Urban Acoustic Scenes 2019 Openset, Development dataset," March 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2591503

[12] ——, "TAU Urban Acoustic Scenes 2019 Openset, Evaluation dataset," May 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3064132

[13] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13.

[14] H. Chen, Z. Liu, Z. Liu, P. Zhang, and Y. Yan, "Integrating the data augmentation scheme with various classifiers for acoustic scene modeling," DCASE2019 Challenge, Tech. Rep., June 2019.

[15] K. Koutini, H. Eghbal-zadeh, and G. Widmer, "Acoustic scene classification and audio tagging with receptive-field-regularized CNNs," DCASE2019 Challenge, Tech. Rep., June 2019.

[16] M. Kośmider, "Calibrating neural networks for secondary recording devices," DCASE2019 Challenge, Tech. Rep., June 2019.

[17] W. Gao and M. McDonnell, "Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths," DCASE2019 Challenge, Tech. Rep., June 2019.

[18] H. Zhu, C. Ren, J. Wang, S. Li, L. Wang, and L. Yang, "DCASE 2019 challenge task1 technical report," DCASE2019 Challenge, Tech. Rep., June 2019.

[19] A. Rakowski and M. Kośmider, "Frequency-aware CNN for open set acoustic scene classification," DCASE2019 Challenge, Tech. Rep., June 2019.

[20] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *2008 16th European Signal Processing Conference*, Aug 2008, pp. 1–4.

[21] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[22] J. Huang, P. Lopez Meyer, H. Lu, H. Cordourier Maruri, and J. Del Hoyo, "Acoustic scene classification using deep learning-based ensemble averaging," DCASE2019 Challenge, Tech. Rep., June 2019.

[23] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 776–780.

[24] B. Lehner and K. Koutini, "Acoustic scene classification with reject option based on resnets," DCASE2019 Challenge, Tech. Rep., June 2019.

[25] K. Wilkinghoff and F. Kurth, "Open-set acoustic scene classification with deep convolutional autoencoders," DCASE2019 Challenge, Tech. Rep., June 2019.

[26] Q. Kong, Y. Cao, T. Iqbal, W. Wang, and M. D. Plumbley, "Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems," DCASE2019 Challenge, Tech. Rep., June 2019.