IDMT-Traffic: An Open Benchmark Dataset for Acoustic Traffic Monitoring Research

Jakob Abeßer
Semantic Music Technologies
Fraunhofer IDMT
Ilmenau, Germany
jakob.abesser@idmt.fraunhofer.de

Saichand Gourishetti
Industrial Media Applications
Fraunhofer IDMT
Ilmenau, Germany

András Kátai
Industrial Media Applications
Fraunhofer IDMT
Ilmenau, Germany

Tobias Clauß
Industrial Media Applications
Fraunhofer IDMT
Ilmenau, Germany

Prachi Sharma
Semantic Music Technologies
Fraunhofer IDMT
Ilmenau, Germany

Judith Liebetrau
Industrial Media Applications
Fraunhofer IDMT
Ilmenau, Germany

***Abstract*—In many urban areas, traffic load and noise pollution are constantly increasing. Automated systems for traffic monitoring are promising countermeasures, which allow to systematically quantify and predict local traffic flow in order to to support municipal traffic planning decisions. In this paper, we present a novel open benchmark dataset, containing 15,706 2-second long stereo audio clips, which were extracted from 4718 vehicle passing events captured with both high-quality sE8 and medium-quality MEMS microphones. This dataset is well suited to evaluate the use-case of deploying audio classification algorithms to embedded sensor devices with restricted microphone quality and hardware processing power. In addition, this paper provides a detailed review of recent acoustic traffic monitoring (ATM) algorithms as well as the results of two benchmark experiments on vehicle type classification and direction of movement estimation using four state-of-the-art convolutional neural network architectures.**

I. Introduction

A world-wide rise in population and a steady urbanization trend cause people to move from rural areas to bigger cities. With more and more active vehicles, travelling times increase and so do noise and air pollution levels. Intelligent transportation systems (ITS) are effective countermeasures to reduce and optimize traffic flow by adapting to local traffic situations. In the past decade, several automatic methods for traffic monitoring were developed for application scenarios such as controlling traffic light cycles, traffic accident detection, logistics monitoring, and other smart city application.

Traffic monitoring systems use various sensor modalities to measure traffic flow, which range from camera sensors for visual object detection and tracking, magnetic loop sensors for counting passing vehicles, to measurement systems based on radio waves (Radar) and light waves (Lidar). While such systems can be installed as distributed sensor networks to cover large areas, installation and maintenance costs are often high. Acoustic traffic monitoring (ATM) systems provide a less expensive alternative for non-intrusive traffic measurements and is the main focus here.

This paper has three main contributions. First, we present a compact state-of-the-art review of recent ATM systems. As a second contribution, we introduce the `IDMT-Traffic` dataset, a novel dataset for traffic monitoring that includes 15,706 2-second long stereo audio clips, which were extracted from multi-microphone audio recordings of 4718 annotated passing vehicles. The dataset is intended as public benchmark to further stimulate research on acoustic traffic monitoring. Finally, we present the results of two benchmark experiments for vehicle type classification and direction of movement estimation using four different convolutional neural network (CNN) architectures.

This paper is structured as follows: We first review recent ATM algorithms in Section II before Section III describes the `IDMT-Traffic` dataset in details. Then, Section IV discusses the experimental procedure and the results of the two benchmark experiments. Finally, Section V concludes this work.

II. Related Work

The audible sound on a road is emitted by several sound sources such as engines, exhausts, wheels and air turbulence, which occurs when vehicles pass by [1]. By breaking down this complex audio analysis scenario, researches approached traffic monitoring from different perspectives. In this section, we categorize existing ATM algorithms based on the applied data acquisition and statistical modeling approaches.

Moving sound sources such as vehicles can be detected based on their emitted sound patterns if they are recorded with at least two separate microphones. Therefore, most ATM methods analyze either *stereo audio signals* [1], [2] [3] [4] [5] [6] or *multi-channel audio recordings* [7] [8] [9] [10], which are recorded with microphone arrays [7]–[9]. Microphones are commonly integrated into smaller sensor units that are placed either at the roadside [3] [4] [8] [2] or mounted on light poles at a height between 0.5 and 3 meters [1] [7] [6].

Traffic density is commonly measured on a two-stage scale (congested/non-congested) [5] or on a three-stage scale as either low/free (equivalent to vehicle speeds larger equal to

ISBN: 978-9-0827-9706-0

40 km/h), medium (20-40 km/h), or heavy/jammed (below 20 km/h) [1]–[4]. Traffic density can also be measured by *detecting and counting the number of passing vehicles*. A common approach is to investigate run-time differences between stereo audio signals [7], [9], [10]. Moving sources exhibit a sweep-like peak contour in the temporal development of the cross-correlation function between both signals. While the contour's diagonal alignment indicates the direction of movement, the contour's angle correlates with the speed of a vehicle. Ishida et al. match pre-defined templates with the cross-correlation function to detect for left-right and right-left movements [10].

Heavy traffic can lead to *traffic accidents*, which can be detected by the two sounds tire skidding and car crash [11], [12]. Another approach for traffic monitoring is to distinguish between vehicles in good and bad *mechanical condition* based on emitted sounds [2].

Different audio signal representations are used for solving ATM tasks. While most often raw spectrogram representations are used, some authors compute more advanced audio features such as the Mel-frequency Cepstral Coefficients (MFCC) prior to the modeling and classification steps [2], [3]. In order to train more robust algorithms, Gatto et al. apply data augmentation and mix recorded audio signals with additional noise [5].

ATM algorithms apply various mostly traditional classification algorithms such as Nearest Neighbor classifier [3], Bayer's classifier [4], Random Forest classifier [5], Support Vector Machines (SVM) [2], Artificial Neural Networks (ANN) [2], [3], as well as hybrid approaches such as the Neuro-Fuzzy Classifier [1]. While most above-mentioned tasks require classification algorithms, Djukanović et al. use Support Vector Regression (SVR) [6] to predict the vehicle-to-microphone distances.

Most publications for ATM systems rely on proprietary datasets. However, some publicly available datasets exist and can be applied for traffic monitoring. The MIVIA road audio events data set includes audio recordings of 400 sound events of the two classes tire skidding and car crashes [11]. The MAVD dataset [13] was published for sound event detection (SED) of particular traffic sounds which were recorded from the vehicle classes car, truck, bus, motorcycle in different states such as idling, accelerating, or braking. In the research field of SED, many datasets such as the FSK50k [14] or the AudioSet [15] include general sound classes such as car, truck, or train, whose recognition could be applied in ATM systems. Similarly, acoustic scene classification (ASC) datasets such as the TUT Urban Acoustic Scenes 2018 dataset [16] allow to train algorithms to detect amongst others traffic-related sound scenes such as "Street, traffic", "Bus", "Metro", and "Tram".

III. IDMT-Traffic dataset

In this section, we provide a detailed description of the `IDMT-Traffic` dataset, which is a novel dataset for acoustic traffic monitoring[1] in inner-city and overland road scenarios. The dataset is intended as a public evaluation benchmark for the ATM tasks vehicle detection, vehicle type classification, direction of movement estimation, as well as speed estimation.

The dataset includes a total of 15,706 2-second long stereo audio clips, which were extracted from long-term multi-channel audio recordings. These recordings are time-synchronized stereo audio recordings using both high-quality sE8 microphones[2] as well as lower-budget microelectro-mechanical systems (MEMS) microphones[3]. One subset of the audio clips captures passing vehicle events while another subset captures typical background sounds alongside the road without any audible vehicle sounds. Audio recordings were conducted at four different recording locations including three city traffic locations and one country road location in and around Ilmenau, Germany. The recording scenarios cover both dry and wet street conditions and three different speed limits (30, 50, and 70 km/h), as well as morning and afternoon recordings.

Figure 1a and Figure 1b illustrate the recording setup, which was placed with a distance of 0.5 meters to the adjacent street. Both pairs of sE8 and MEMS microphones are fixed at a distance of 18.5 cm as an approximation of the human ear-distance. The time-synchronized video recordings were used to annotate the vehicle type and direction of movement afterwards. For reasons of data protection, we strictly avoided filming faces and licence plates by aiming the camera at the lower part of the vehicles as shown in Figure 1c and Figure 1d. For each of both microphone types, around 2.5 hours of audio recordings exist with a total of 4718 annotated passing vehicles. The dataset includes four classes: cars (3903 events), trucks (511 events), busses (53 events), and motorcycles (251 events). This distribution reflects the natural imbalance of vehicle types in common traffic scenarios in small-town regions of Germany.

As the main contribution w.r.t. previous ATM datasets listed in Section II, the `IDMT-Traffic` dataset includes vehicle passings recorded at different day times, road conditions, speed limits, and microphones. Such degrees-of-freedoms allow for a systematic evaluation of the robustness of ATM algorithms in order to allow for a better performance in real-world scenarios.

IV. Benchmark Experiments

Using the `IDMT-Traffic` dataset introduced in Section III, we conducted two benchmark experiments for different ATM tasks. Here, we only used the audio recordings recorded with the high-quality sE8 microphones and leave an investigation of the influence of microphone mismatch between the MEMS and sE8 microphones a topic for future research.

A. Audio Representation & Pre-processing

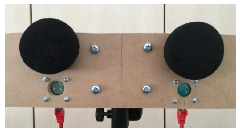
Audio files were processed at a sample rate of 48 kHz. In this work, we address the tasks of detecting vehicles and classifying their type and direction of movement. Therefore,

[1]The dataset can be downloaded at https://www.idmt.fraunhofer.de/en/publications/datasets.html

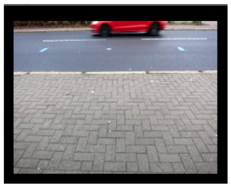
[2]https://www.seelectronics.com/se8-mic

[3]InvenSense ICS-43434

(a) Back view: stereo microphone setup (left) and video camera (right).



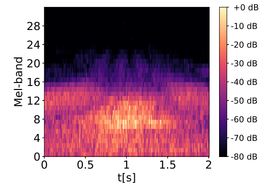
(b) Frontal view: microphone setup (top: sE8, bottom: MEMS microphones.)



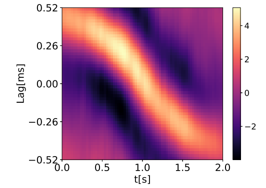
(c) Example video frames for vehicle type annotations.

(d) Example video frames for vehicle type annotations.

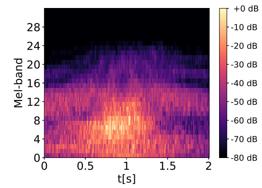
Fig. 1: Recording setup for the dataset creation including two pairs of microphones (a, b) with sE8 microphones (top) and MEMS microphones (bottom) and a digital camera. Two example video frames showing passing vehicles (c, d).



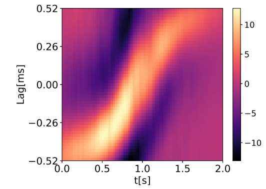
(a) CA, L→R, 50 km/h, MS.



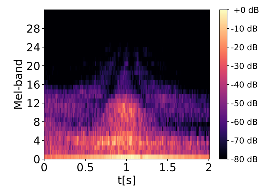
(b) CA, L→R, 50 km/h, CC.



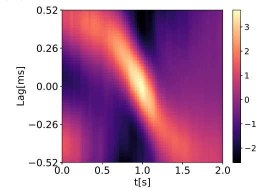
(c) TR, R→L, 50 km/h, MS.



(d) TR, R→L, 50 km/h, CC.



(e) MC, L→R, ≈ 70 km/h, MS.



(f) MC, L→R, ≈ 70 km/h, CC.

Fig. 2: Examples of two-second long patches taken from the `IDMT-Traffic` dataset for the vehicle type classes car (CA), truck (TR), and motorcycle (MC) for both mel-spectrogram (MS) and cross-correlation (CC) features. Furthermore, direction of movement as either left-to-right (L→R) or right-to-left (R→L) as well as approximate vehicle speed (km/h) is provided.

we extract two types of features to be processed by the convolutional neural networks introduced in Section IV-B.

As first feature type, we extract mel-spectrograms using the *librosa* library [17] using an FFT size, a window size, and a hop size of 2048, 1024, and 512 samples, respectively. As pre-processing, we average the left and right audio channels and down-sample the audio signal to a sample rate of 22.05 kHz. In our experiments, we investigate the effect of the number of mel-bands $N_B \in \{16, 32, 64, 128\}$ on the recognition performance. Logarithmic magnitude scaling is applied to compensate for the natural dynamic range of the traffic recordings. Finally, two-second long sub-sequences (patches) are extracted.

As a second feature type, we compute the local cross-correlation between the left and right audio channel at the original sample rate of 48 kHz. Therefore, we extract blocks of 200 ms duration from the audio signals using a hopsize of 25 ms. From the cross-correlation function between the left and right channel of the b-th sample block, we keep the center part $c_b \in \mathbb{R}^{51}$ with a margin of 25 lags around zero-lag index. We derive a two-dimensional feature representation by stacking the block-wise cross-correlation functions for two-second long patches as for the mel-spectrograms.

Features are standardized (zero mean and unit variance) per bin, i. e., per frequency bin for the mel-spectrogram patches and per time lag for the cross-correlation patches, over all patches of a given dataset. This normalization procedure is performed independently for the training set, validation set, and test set. Figure 2 shows both the mel-spectrogram as well as the cross-correlation features for three examples, showing vehicle passings of a car, a truck, and a motorcycle.

B. Neural Network Architectures

In our benchmark experiments, we test three different convolutional neural network architectures, which will be detailed in the following sections. Table I summarizes the number of parameters per model.

1) VGGNet: The `VGGNet` model proposed by Takahashi et al. in [18] uses four pairs of 3x3 convolutional layers with intermediate pooling only between the layer pairs. This way, the spatial resolution is decreased while the number of filters is simultaneously increased from 32 to 256. Several regularization strategies such as batch normalization, dropout, as well as L2 regularization in the penultimate dense layers are applied to improve the model's generalization towards new data.

2) ResNet: The `ResNet` is the "RN1" as proposed by Koutini et al. in [19]. It includes five residual blocks with two convolutional layers each. The network was designed to have a reduced receptive field and was evaluated in the original paper for the task of acoustic scene classification.

3) SqueezeNet: The `SqueezeNet` architecture was introduced in [20] and implements several model compression strategies. As a first strategy, 3x3 filters are replaced by 1x1 filters in the convolutional layers. As a second strategy, the network includes a number fire modules, which use a squeeze-

Model	# Parameters	Classification Task
`VGGNet`	$1,442,788$	Vehicle type
`ResNet`	$3,259,012$	Vehicle type
`SqueezeNet`	$1,171,652$	Vehicle type
`MobileNetMini`	$15,363$	Direction of movement

TABLE I: Summary of the compared neural network architectures, their number of parameters, as well as the classification tasks, they have been evaluated for.

Dataset	Car	Truck	Motorcycle	No vehicle
Training Set	2471	290	132	2393
Validation Set	275	32	15	266
Test Set	1157	189	99	1412

TABLE II: Number of patches per class in the training set, validation set, and test set for vehicle type classification.

	Car	Truck	Motorcycle	No vehicle
Car	**97.29**	2.62	0.02	0.09
Truck	60.21	**38.84**	0.63	0.32
Motorcycle	3.23	1.21	**95.35**	0.2
No vehicle	0.23	0.01	0.11	**99.65**

TABLE III: Confusion matrix for vehicle type classification using `VGGNet` with $N_B = 16$ (all values in percent).

Dataset	L→R	R→L	No vehicle
Training Set	1445	1448	2393
Validation Set	161	161	266
Test Set	678	767	1412

TABLE IV: Number of patches per class in the training set, validation set, and test set for direction of movement estimation.

and-expand approach to reduce the depth of feature maps while maintaining their size.

4) MobileNetMini: The `MobileNetMini` model is a miniaturized version of the MobileNet architecture proposed in [21]. It includes one convolutional layer and one depth-wise convolutional layer with batch normalization and ReLU activation functions each followed by a global max pooling operations and a final softmax dense layer.

C. Experimental Procedure

From the `IDMT-Traffic` dataset, we first select audio files recorded using the sE8 microphones at two recording locations with speed limits of 30 and 50 km/h. Both sets are combined, then shuffled and split into training set (90 %) and validation set (10 %). Recordings from the third location having a speed limit of 70 km/h was used as test set.[4] Using this data partition, we aim to test the robustness of the ATM algorithms against different vehicle speeds and the corresponding changes in the vehicle sound characteristics. We trained all neural networks using the Adam optimizer [22] for 250 epochs with a learning rate of 10^{-5}. Early stopping with a patience of 50 epochs is used on the validation loss to monitor the training process.

D. Experiment 1 - Vehicle Type Classification

We consider a four-class classification scenario where we include the three vehicle types cars, trucks, and motorcycles as well a no-vehicle class, which includes spectrogram patches without any passing vehicles. The patches for the first three classes are centered around the annotated passing times. Non-vehicle patches were randomly sampled in between annotated vehicle passings in the audio recordings of the `IDMT-Traffic` dataset. The number of patches per class as well as their partition to training set, validation set, and test set is given in Table II. It can be observed that the classes no-vehicle and car have most patches followed by truck and motorcycle.

[4]The corresponding file split is provided alongside with the `IDMT-Traffic` dataset.

We observe that all models perfectly recognize the no-vehicle patches therefore allow for a robust vehicle detection (binary classification task) based on the high-quality sE8 audio recordings. Concerning the model performance, `VGGNet` and `ResNet` perform comparably well and slightly outperform the `SqueezeNet` model. Interestingly, the results show that a frequency resolution of only 16 mel-bands ($N_B = 16$) is sufficient to classify between vehicle types.

Table III illustrates as an example the confusion matrix for the `VGGNet` with $N_B = 16$. It becomes apparent that the truck-to-car confusion is the most prominent misclassification. We assume that since both vehicles have only small differences in their geometric size, they emit similar sound patterns, which complicate their distinction.

E. Experiment 2 - Direction of Movement Estimation

In this experiment, we evaluate the performance of the `MobileNetMini` architecture on the cross-correlation features for detecting the direction of movement. Here, we include patches across different vehicle types for the classes left-to-right and right-to-left and add no-vehicle patches as third class to simulate the detection task. The number of patches per class as well as their distribution among training, validation, and test sets is given in the Table IV. As can be seen in the confusion matrix in Table V, the direction of movement can be easily determined using a very small MobileNet architecture and the cross-correlation features. This result confirms the findings from the scientific literature [7], [9], [10]. As the only distinction, our method relies on automatic feature learning as part of the CNN model.

V. Conclusions

In this paper we showed that acoustic traffic monitoring provides a low-cost and non-invasive alternative to traffic monitoring approaches based on other sensor modalities such as vision or radar. After providing a thorough review of scientific

	L→R	R→L	No vehicle
L→R	**96.61**	0.83	2.57
R→L	0.26	**98.64**	1.1
No vehicle	0	0.21	**99.79**

TABLE V: Confusion matrix for direction of movement estimation using the `MobileNetMini` (all values in percent).

publications on acoustic traffic monitoring, we presented the novel `IDMT-Traffic` dataset, which is a freely-accessible benchmark dataset intended to stimulate further research in acoustic traffic monitoring.

In our baseline experiments, which used solely the high-quality audio recordings in the dataset, we showed that state-of-the-art convolutional neural networks already achieve high performance scores for vehicle type classification and direction of movement estimation. Furthermore, the results show that vehicle detection can be implemented easily with either the mel-spectrogram or the cross-correlation features.

Having the goal of a real-world deployment of an ATM system in mind, several challenges need to be addressed in future research. The first challenge arises from the microphone mismatched between high-quality and low-quality microphones used in mobile sensor devices. By including audio recordings from both high-quality sE8 microphones as well as lower-quality MEMS microphones, the `IDMT-Traffic` dataset provides a suitable test-bed to develop new algorithmic strategies for domain adaptation. A second challenge comes from computational performance constraints of mobile sensor devices, which might require to compress the neural network models. In addition to these challenges, possible future research directions include a precise speed estimation of vehicles as well as an improved classification of passing trucks.

Acknowledgements

This research was supported by the Fraunhofer Innovation Program "Trusted Resource Aware ICT" (TRAICT). The authors would like to thank Ke Shen, Yishi Sun, Jiaming Tan, Julia Aileen Katharina Duong, and Thanh Phong Duong for their assistance in the dataset recording process.

References

[1] P. Borkar and L. G. Malik, "Acoustic Signal based Traffic Density State Estimation using Adaptive Neuro-Fuzzy classifier," in *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Hyderabad, India, 2013, pp. 1–8.

[2] M. Bhandarkar and T. Waykole, "Vehicular Mechanical Condition Determination and On Road Traffic Density Estimation Using Audio Signals," in *Proceedings of the 6th International Conference on Computational Intelligence and Communication Networks*, Bhopal, India, 2014, pp. 395–401.

[3] V. P. Warghade and M. S. Deshpande, "Road Traffic Condition Estimation Based on Road Acoustics," in *Proceedings of the International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, Pune, India, 2017, pp. 1–5.

[4] V. Tyagi, S. Kalyanaraman, and R. Krishnapuram, "Vehicular Traffic Density State Estimation Based on Cumulative Road Acoustics," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1156–1166, 2012.

[5] R. C. Gatto and C. H. Q. Forster, "Audio-Based Machine Learning Model for Traffic Congestion Detection," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–8, 2020.

[6] S. Djukanovic, J. Matas, and T. Virtanen, "Robust Audio-Based Vehicle Counting in Low-to-Moderate Traffic Flow," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, Las Vegas, NV, USA, 2020, pp. 1608–1614.

[7] Shiping Chen, Ziping Sun, and B. Bridge, "Traffic Monitoring Using Digital Sound Field Mapping," *IEEE Transactions on Vehicular Technology*, vol. 50, no. 6, pp. 1582–1589, 2001.

[8] Y. Na, Y. Guo, Q. Fu, and Y. Yan, "An Acoustic Traffic Monitoring System: Design and Implementation," in *Proceedings of the IEEE UIC-ATC-ScalCom-CBDCom-IoP*, Guangzhou, China, 2015, pp. 119–126.

[9] B. Barbagli, G. Manes, and R. Facchini, "Acoustic Sensor Network for Vehicle Traffic Monitoring," in *Proceedings of the International Conference on Advances in Vehicular Systems, Technologies and Applications (VEHICULAR)*, Venice, Italy, 2012, pp. 1–6.

[10] S. Ishida, S. Liu, K. Mimura, S. Tagashira, and A. Fukuda, "Design of acoustic vehicle count system using DTW," in *Proceedings of the ITS World Congress*, Melbourne, Australia, 2016, pp. 1–10.

[11] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio Surveillance of Roads: A System for Detecting Anomalous Sounds," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 279–288, 2016.

[12] Y. Li, X. Li, Y. Zhang, M. Liu, and W. Wang, "Anomalous Sound Detection Using Deep Audio Representation and a BLSTM Network for Audio Surveillance of Roads," *IEEE Access*, vol. 6, pp. 58 043–58 055, 2018.

[13] P. Zinemanas, P. Cancela, and M. Rocamora, "MAVD: A Dataset for Sound Event Detection in Urban Environments," in *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, New York, NY, USA, 2019, pp. 263–267.

[14] E. Fonseca, S. Member, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: an Open Dataset of Human-labeled Sound Events," *arXiv preprint arXiv:2010.00475*, vol. 14, no. 8, pp. 1–21, 2020.

[15] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 776–780.

[16] A. Mesaros, T. Heittola, and Tuomas Virtanen, "A Multi-Device Dataset for Urban Acoustic Scene Classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*, Surrey, UK, 2018, pp. 9–13.

[17] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," in *Proceedings of the Scientific Computing with Python conference (Scipy)*, Austin, Texas, 2015, pp. 18–24.

[18] G. Takahashi, T. Yamada, N. Ono, and S. Makino, "Performance Evaluation of Acoustic Scene Classification using DNN-GMM and Frame-Concatenated Acoustic Features," in *Proceedings of the 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Honolulu, Hawaii, USA, 2018, pp. 1739–1743.

[19] K. Koutini, H. Eghbal-Zadeh, M. Dorfer, and G. Widmer, "The receptive field as a regularizer in deep convolutional neural networks for acoustic scene classification," *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pp. 1–5, 2019.

[20] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and $<$0.5MB model size," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France, 2016, pp. 1–13.

[21] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[22] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015, pp. 1–15.