ICA 2019 AACHEN

PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

[The effect of room acoustics on audio event classification]

Dimitra Emmanouilidou, Hannes Gamper

Microsoft Research, Redmond, US, {dimitra.emmanouilidou, hannes.gamper}@microsoft.com

Abstract

The increasing availability of large-scale annotated databases, together with advances in data-driven learning and deep neural networks, have pushed the state of the art for computer-aided detection problems like audio scene analysis and event classification. However, the large variety of acoustic environments and their acoustic properties encountered in practice can pose a great challenge for such tasks and compromise the robustness of general-purpose classifiers when tested in unseen conditions or real-life applications. In this work we perform a quantitative analysis of the effect of room acoustics on general audio event detection scenarios. We study the impact of mismatches between training and testing conditions in terms of acoustical parameters, including the reverberation time (T60) and the direct-to-reverberant ratio (DRR), on audio classification accuracy and class separability. The results of this study may serve as guidance for practitioners to build more robust frameworks for audio event classification tasks.

Keywords: Sound Event Classification, Reverberation, T60, C50, ESC-50

1 INTRODUCTION

Sound events serve humans as cues for understanding content and contextual information regarding their surroundings. The aim of computerized audio event detection is to effectively process and convert audio signals into descriptive representations that can be used by automatic processes for inference. There has been considerable research interest in audio event detection and classification over the past few decades, resulting in scientific challenges like DCASE [1], in publicly available data sets [2, 3, 4] and open source contributions.

Recent advances in audio event detection and classification have seen adoption in a variety of applications spanning different fields. In health care monitoring scenarios, Ghiasi [5] proposed a system for classifying heart sounds related to coronary artery disease and heart valve defects; in [6] lung sound signals related to pneumonia and other respiratory diseases are detected, and in [7] an assisted living framework for monitoring patients' behaviour is presented. In home security [8] and surveillance applications, Foggia [9] studied automated detection of road accidents in audio streaming scenarios. Other applications include: multimedia database retrieval, where Esling [10] proposed audio retrieval and classification via multi-objective audio matching; and audio tagging and segmentation, with recent work focusing on scenarios with increased number of classes and label uncertainty [11], that better reflect practical system requirements.

Research on audio event detection and classification faces two major challenges: i) limited availability of reliably annotated data and ii) large variability in terms of the recording hardware, noise conditions, and acoustic environment. The first challenge (i) arises from the fact that collecting and carefully annotating large amounts of audio data is a time consuming and costly task. Recently, efforts have been made to exploit data sets with sparse or noisy labels, driven in part by the increasing availability of large databases containing user-contributed audio clips and meta data. Task 4 of the 2018 DCASE challenge addressed the large-scale detection of sound events using weakly labeled data without explicit event time stamps [12]. The goal was to exploit large amounts of unbalanced and unlabeled training data combined with a smaller set of weakly labeled data. In a similar setting, the 2019 DCASE challenge seeks to exploit a small amount of reliably, manually labeled data, together with a large quantity of noisy web data in a multi-label audio tagging task with a large vocabulary of labels.

The second challenge (ii) is exacerbated by an increasing reliance on large, user-contributed data sets, as these sets presumably exhibit high inter- and intra- class variability in terms of the recording equipment, acoustic environment, and background noise conditions compared to data sets collected in a concerted effort by professionals. This variability poses a challenge for audio event classification models. Lopatka [13] studied the deterioration of acoustic event classification in the presence of background noise, and how this effect varies per class type. They show that the sound of glass breaking displayed moderate classification deterioration in low signal-to-noise (SNR) cases in terms of precision and recall, while gunshot sound classification demonstrated a more significant deterioration at low SNRs, and scream sounds showed a big spread between precision and recall for low SNR cases. The organizers of the first DCASE challenge further discussed the inherent difficulty in detecting overlapping sound events [14]. However, to the best of our knowledge, there is little previous work studying the effect of reverberant environments for the task of audio event classification.

In speech recognition, noise and overlapping sounds are known to negatively affect model performance. Prior work on the effect of reverberation suggests that parameters such as the clarity index (C50) as well as the direct-to-reverberant ratio (DRR) strongly affect speech recognition performance [15]. A recently published DCASE task involves localization and recognition of individual sound events within various reverberant and noisy conditions [16]. This task may ignite interest in addressing the issue of reverberation in audio event classification. Here we study the effect of the acoustic environment, in terms of reverberation parameters, on the performance of a machine-learning based audio event classification model.

2 DATA CORPUS AND METRICS

To determine the effect of reverberation on audio event classification, we rely on an audio event classification corpus, and a large set of measured acoustic impulse responses (AIRs) to simulate various acoustic conditions.

2.1 Sound event data set

The Environmental Sound Classification (ESC-50) data set consists of 2 000 audio recordings [3]. Each file in the data set was recorded at 44 100 Hz and has a duration of 5 seconds. The clips were annotated using a crowd sourcing platform, were judges were presented with 50 classes, under the categories animals, natural soundscapes, human non-speech sounds, domestic sounds and exterior noises. All classes in the data set are balanced, containing 40 examples each, and are split into 5 folds for cross validation.

2.2 Impulse Responses

A large corpus of acoustic impulse responses was compiled from real measurements from proprietary and public data set sources: ACE Challenge Corpus [17], PORI Concert Hall Impulse Responses [18], REVERB Challenge corpus [19], Echothief Impulse Response Library [20], SOFA [21], SMARD [22], Real Acoustic Environments Working Group database [23], and Multichannel Acoustic Reverberation Database at York [24].

2.3 Impulse response parameter estimation

The *reverberation time* (T60) describes the time it takes for the energy of an AIR to decay by 60 dB. It is estimated here using a method by Karjalainen et al. [25]. A related parameter known to be perceptually relevant is the *early decay time* (EDT). It can be estimated by fitting a line to the energy decay curve (EDC), from the point where the EDC drops below -5 dB to where it drops below -15 dB.

Given an AIR, $h[n]$, the *direct-to-reverberant ratio* (DRR) is the ratio of the energy of the direct path, estimated in a 2.5 ms window around the maximum amplitude point of the impulse response, to the energy of the reflected paths outside this window [17]. With $n_d = \arg\max_n |h[n]|$, DRR is given as:

$$\mathrm{DRR} = 10\log_{10}\left(\frac{\sum_{n=n_d-n_w}^{n_d+n_w} h[n]^2}{\sum_{n=n_d+n_w}^{\infty} h[n]^2}\right), \qquad (1)$$

where and n_w is the number of samples in a 2.5 ms window at the given sampling rate. Note that n_d and (1) are slightly modified compared to the definitions given by Eaton [17]. The *clarity* index (C50) measures the

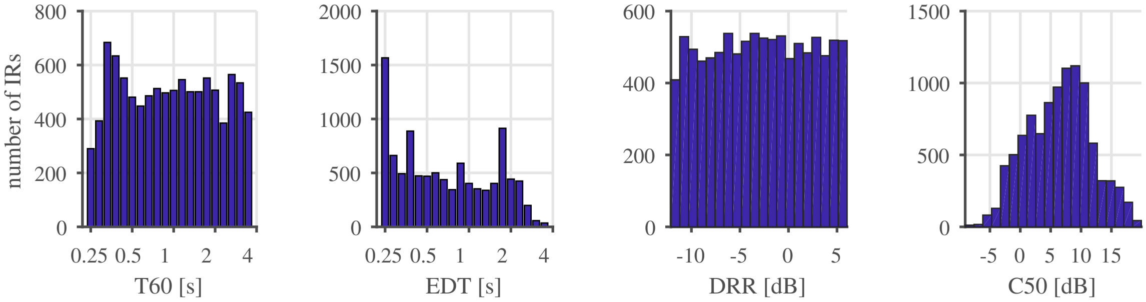

Figure 1. Distribution of impulse response parameters.

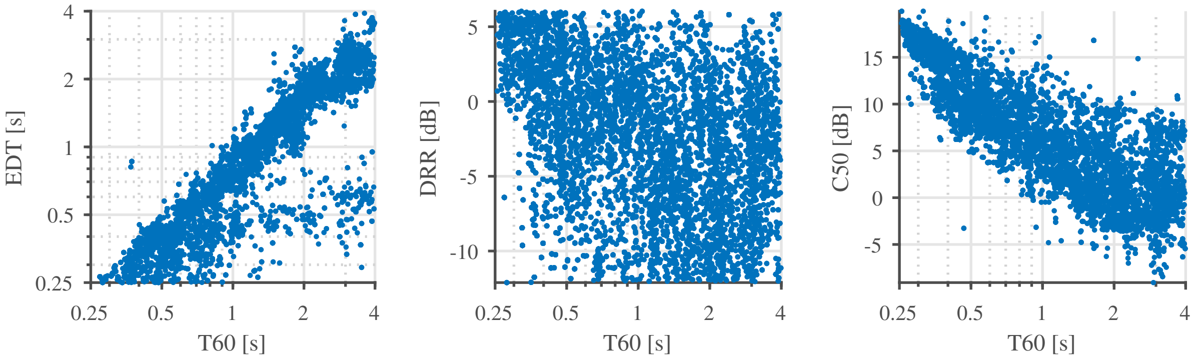

Figure 2. Distribution of impulse response parameters.

energy ratio between early and late parts of the impulse response [26]:

$$\mathrm{C50} = 10 \log_{10} \left(\frac{\sum_{n=n_0}^{n_0+n_{50}} h[n]^2}{\sum_{n=n_{50}}^{\infty} h[n]^2} \right), \tag{2}$$

where n_0 is defined as the sample with the largest drop in the EDC, which was found to be a relatively robust measure for determining the direct path, and n_{50} is the number of samples corresponding to a 50 ms window at the given sampling rate.

2.4 Corpus generation

All samples and AIRs were resampled to 16 000 Hz for further processing. After pruning AIRs with measurement artifacts, low sampling rates, or extreme reverberation parameters (e.g., reverberation times longer than 4 seconds), we compiled a corpus of 11 684 AIRs. For training and evaluation of the audio event classification, we created two separate sets: a *raw* set (**raw**) and a *reverberant* set (**rev**). The **raw** set consists of the raw ESC-50 samples. The **rev** set was created by convolving the raw ESC-50 samples with AIRs from our AIR corpus, to generate audio event examples with varying acoustic conditions. To ensure a uniform and dense sampling of the acoustic parameters, we generated 10 000 **rev** examples from the 2 000 **raw** samples, by convolving each sample with 5 AIRs drawn randomly from a uniform distribution between 0.25 and 4 seconds for T60, and between -12 and 6 dB for DRR. Note that for T60, the distribution was chosen to be uniform on a logarithmic scale, as we hypothesize that to be more in line with the expected effect of T60 on classification. It should also be stated that the acoustic conditions of the **raw** set are unknown. We assume that these unknown conditions are randomly distributed in terms of their acoustic parameters, and that their effect can be mitigated through averaging of the classification results. Figure 1 illustrates the distribution of the four acoustic parameters studied

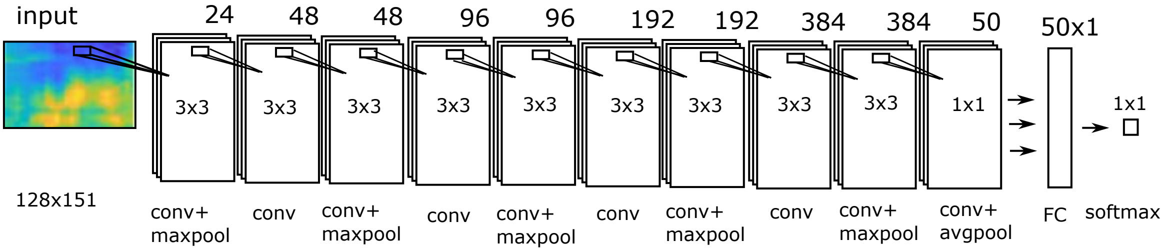

Figure 3. Class prediction model: CNN with 10 convolutional layers and a fully connected layer.

here, for the randomly drawn AIRs. As can be seen, EDT has a distribution similar to T60, while C50 exhibits a Gaussian distribution. Plotting T60 against the other parameters confirms the desired uniform distribution with DRR (Figure 2, center), while revealing a strong correlation between T60 and EDT (Figure 2, left) as well as C50 (Figure 2, right). While this correlation is not surprising given the physical processes underlying AIRs, it should be considered when studying the effects of these AIR parameters in isolation.

2.5 Evaluation metrics

As the ESC-50 data set contains a balanced number of samples per class, the weighted accuracy (WA) measure was used for evaluating classification performance. WA is given by

$$\mathrm{WA} = \frac{\sum_{j=1}^{\mathrm{J}} N_{\mathrm{corr},j}}{\mathrm{N}}, \tag{3}$$

where N is the total number of samples, and $N_{\mathrm{corr},j}$ is the number of correct predictions for class $j = [1, ..., \mathrm{J}]$.

3 AUDIO EVENT CLASSIFICATION MODEL

ESC-50 was created from a set of user-uploaded data and includes challenging examples with large ambiguity between classes. Previous work has shown promising results on this set by using transfer learning and extracting audio features using models pre-trained on large data sets [27, 28]. However, to analyze the effect of reverberation on both training and testing performance of an audio event classification model, we rely only on features and embeddings extracted directly from our data corpus, described in Section 2.

3.1 Model architecture

The classification model and data processing used here closely follow the AclNet convolutional neural network (CNN) architecture in [29], which was shown to provide near state-of-the-art performance on ESC-50. Unlike AclNet, which operates directly on time domain input signals, we use a Mel-frequency spectrogram as input to the network, as described in Section 3.2. A block diagram of the model used here is shown in Figure 3. The network consists of 10 convolutional layers with rectified linear unit (ReLU) activation, batch normalization after all but the first and last layer, a kernel size of 3×3 and a stride of 1. Dropout is added for regularization before layers 4, 8, and 10, with a rate of 0.2. Max pooling over 2×2 patches with a stride of 2 is performed after layers 1, 3, 5, 7, and 9. After each max pooling layer, the number of CNN filters doubles, from 24 to 48, 96, 192, and 384. The final CNN layer has 50 filters, equal to the number of classes. It is followed by a single 2×4 average pooling to reduce the number of outputs to 50. The classification result is obtained at the output of a single fully connected linear layer.

3.2 Data augmentation and feature extraction

To increase the amount of available training data, a common technique is to apply transformations to the input signals or features, a process referred to as data augmentation. Here we perform augmentation online, that is, transformations are applied to all samples as they are retrieved for training or testing. As proposed by Huang [29], we first extract a random 2-second segment from the 5-second audio clips. As the ESC-50 clips contain silent segments, we discard segments whose amplitude never exceeds 10% of the overall maximum amplitude. The 2-second segment is then stretched in time through resampling with a random factor drawn uniformly from [0.8, 1.25]. The resulting segment is cropped to 1.5 seconds, and a random gain drawn uniformly from [-6, 6] dB is applied. A Mel-spectrogram is extracted from the resulting clip using an FFT size of 512 samples and an overlap of 160 samples, yielding 128 spectral and 151 temporal bins. This feature matrix is fed as input to the CNN model, after taking the logarithm and applying a constant bias and gain for normalization. During testing, the same feature matrix is calculated for the 1.5-second segment with the highest energy.

4 EXPERIMENTAL EVALUATION

4.1 Train and test conditions

To explore the effect of reverberation on classification performance, we used the following conditions:

1. Train on **raw**, test on **raw** (**TrRaw-TeRaw**); this serves as the baseline, and corresponds to the typical experimental setup in prior work using the ESC-50 data set.

2. Train on **raw**, test on **rev** (**TrRaw-TeRev**); this reveals the performance impact of testing on acoustically more challenging conditions than the network was trained on.

3. Train on **rev** and **raw**, test on **raw** (**TrRev-TeRaw**); applying reverberation to (some) training data could potentially be seen as a form of data augmentation.

4. Train on **rev** and **raw**, test on **rev** (**TrRev-TeRev**); this scenario illustrates the benefit of training the model on acoustic conditions similar to the ones encountered during testing.

The same classification model (see Section 3.1) is trained and tested on all 4 conditions outlined above. The effect of the reverberation is determined by analyzing the classification performance as a function of the reverberation parameters of the AIRs used to generate the **rev** samples (see Section 2).

4.2 Model training

The CNN model is implemented in PyTorch [30] and trained using stochastic gradient descent, with a learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.0002, over 500 epochs for **TrRaw-TeRaw** and **TrRaw-TeRev**, and 130 epochs for **TrRev-TeRaw** and **TrRev-TeRev**. Five-fold cross-validation is performed per the ESC-50 recommendations, with 4 folds used for training and 1 fold for testing.

4.3 Results

For the baseline condition **TrRaw-TeRaw**, the model achieved an average classification accuracy of 68.1%. As seen in Table 1, performance dropped significantly for **TrRaw-TeRev**, i.e., when testing on reverberant data, to 45.6%. Including **rev** data for training improved performance for both **raw** and **rev** test sets, with **TrRev-TeRaw** and **TrRev-TeRev** achieving 71.2% and 62.4% accuracy, respectively. This indicates that adding reverberant examples to the training data can be useful both for data augmentation and for reducing potential mismatches between acoustic conditions in training and testing.

Table 1. Average classification performance of the CNN (see Section 3.1) for all experimental conditions.

	TrRaw-TeRaw	**TrRaw-TeRev**	**TrRev-TeRaw**	**TrRev-TeRev**
WA (%)	68.1	45.6	71.2	62.4

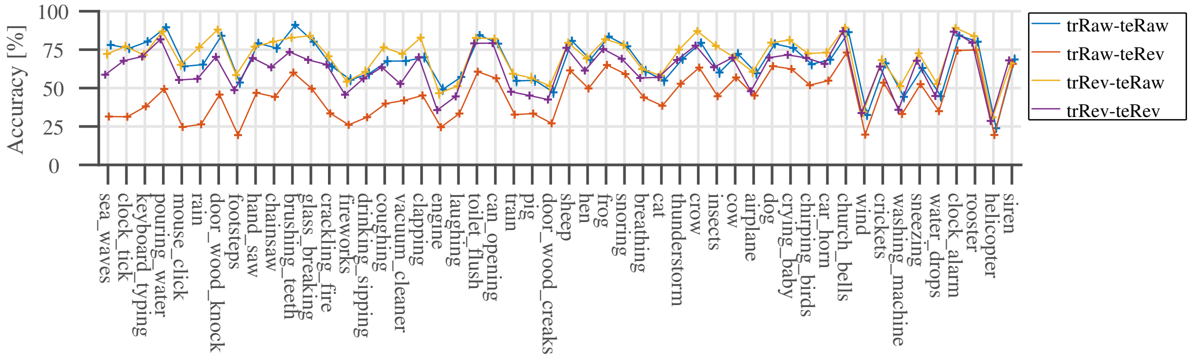

Figure 4. Classification accuracy per class for all experimental conditions.

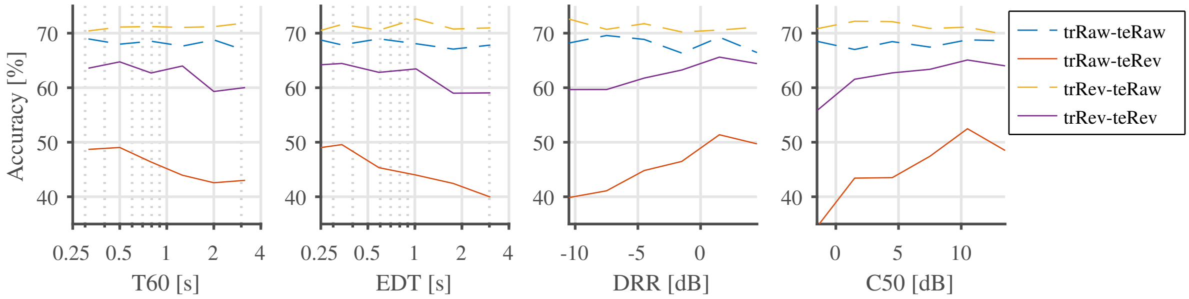

Figure 5. Classification accuracy as a function of IR parameters.

Figure 4 illustrates the performance of the model for all classes and experimental conditions. To reduce variability, the results are averaged over all 5 folds, 3 identical models trained with slightly different (learning rate, dropout rate) settings, i.e., (0.01, 0.2), (0.005, 0.2), and (0.005, 0.5), as well as 6 training epochs from epochs 450 to 500 for **TrRaw-TeRaw** and **TrRaw-TeRev**, and from epochs 90 to 140 for **TrRev-TeRaw** and **TrRev-TeRev**. Vertical lines indicate 95% confidence intervals. As shown, the per-class classification performance ranges from about 20% for the "sea waves" and "clock tick" to about 90% for "siren".

To highlight the effect of mismatches in terms of the reverberation parameters between training and testing, the results are sorted for the condition **TrRaw-TeRev** in terms of the average classification performance relative to the baseline condition, **TrRaw-TeRaw**, that is, from the class most affected by a reverberation parameter mismatch to the class least affected. There are several possible explanations for the per-class performance impact of reverberation. Some classes may exhibit acoustic features less affected by reverberation. These features may include slow spectro-temporal changes that are not masked by reverberation (e.g., "siren"), or spectro-temporal patterns that are sufficiently distinct to be recognizable regardless (e.g., "clock alarm"). Conversely, samples with distinct transient features, including "clock tick" and "keyboard typing", may be negatively impacted by reverberation. Furthermore, the raw ESC-50 samples do not exhibit random acoustic conditions, but that there is some correlation between the class and the typical acoustic conditions in which its samples are recorded. For example, "sea waves" is not typically subject to reverberation, but "washing machine" might be.

However, the response of complex machine learning models to even subtle changes in the input features can be rather non-intuitive, especially in the case of mismatches between training and testing. Thus, it may be more insightful to look for trends averaged over all classes. Figure 5 shows how the total accuracy varies with respect to the different AIR parameters. The results are binned and averaged over all samples of the experimental conditions with reverberant test sets, **TrRaw-TeRev** and **TrRev-TeRev**. For reference, we also show the accuracy of models **TrRaw-TeRaw** and **TrRev-TeRaw**, binned and averaged over the same samples, even though no reverberation was applied to those samples. As can be seen, for these conditions the performance is relatively constant across all bins, i.e., any effects visible for the reverberant test conditions, **TrRaw-TeRev** and **TrRev-TeRev**, are most likely a result of the added reverberation. For **TrRaw-TeRev**, performance is significantly worse compared to other conditions across all AIR parameters. Furthermore, classification performance seems to decrease with T60 and EDT and increase with DRR and C50, by a margin of about 10%. Adding reverberation during training boosts performance by about 16% (see Table 1), an indicator that matching acoustic conditions during training and testing are important for achieving high classification performance.

5 CONCLUSIONS

We present an exploratory study on the effect of reverberation on sound event classification for the Environmental Sound Classification (ESC-50) data set. A convolutional neural network (CNN) based on AclNet [29] was trained and tested on a combination of raw and artificially reverberated ESC-50 samples. For the given model and samples, we observed an average classification performance drop of 22.5% for a model trained on raw ESC-50 samples and tested on reverberant samples. The performance drop ranged from close to 0% to about 50% depending on the class. Our results indicate a correlation between this drop and reverberation time (T60) and early decay time (EDT), as well as direct-to-reverberant ratio (DRR) and clarity (C50). Adding artificially reverberated samples to the training data reduced the performance gap and even improved performance on the raw ESC-50 samples, suggesting adding reverberation may further be useful for data augmentation. A more detailed analysis of the impact of reverberation on class-dependent features is left for future work.

REFERENCES

[1] A. Mesaros *et al.*, “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, 2018.

[2] J. F. Gemmeke *et al.*, “Audio set: An ontology and human-labeled dataset for audio events,” in *IEEE ICASSP*, 2017.

[3] K. J. Piczak, “ESC: Dataset for environmental sound classification,” in *International Conference on Multimedia*, 2015.

[4] F. Font, G. Roma, and X. Serra, “Freesound technical demo,” in *International Conference on Multimedia*, 2013.

[5] V. N. Varghees and K. I. Ramachandran, “Effective heart sound segmentation and murmur classification using empirical wavelet transform and instantaneous phase for electronic stethoscope,” *IEEE Sensors Journal*, 2017.

[6] D. Emmanouilidou *et al.*, “Computerized lung sound screening for pediatric auscultation in noisy field environments,” *IEEE TBME*, vol. 65, pp. 1564–1574, July 2018.

[7] R. M. Alsina-Pagès *et al.*, “Homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring,” *Sensors (Basel, Switzerland)*, vol. 17, April 2017.

[8] J.-C. Wang *et al.*, “Gabor-based nonuniform scale-frequency map for environmental sound classification in home automation,” *IEEE T-ASE*, 2014.

[9] P. Foggia *et al.*, “Audio surveillance of roads: A system for detecting anomalous sounds,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, Jan 2016.

[10] P. Esling and C. Agon, “Multiobjective time series matching for audio classification and retrieval,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, Oct 2013.

[11] E. Fonseca *et al.*, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," in *Proceedings of DCASE2018*, November 2018.

[12] R. Serizel *et al.*, "Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments," *DCASE2018 Workshop*, July 2018.

[13] K. Łopatka *et al.*, "Evaluation of sound event detection, classification and localization in the presence of background noise for acoustic surveillance of hazardous situations," in *Multimedia Communications, Services and Security*, Springer, 2014.

[14] D. Giannoulis *et al.*, "Detection and classification of acoustic scenes and events: Ieee aasp challenge," in *IEEE WASPAA*, 2013.

[15] K. Kinoshita *et al.*, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP*, vol. 2016, no. 1, 2016.

[16] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Submitted to DCASE 2019*.

[17] J. Eaton *et al.*, "The ACE challenge—corpus description and performance evaluation," in *IEEE WASPAA*, 2015.

[18] "Concert hall impulse responses Pori, Finland: Reference." http://legacy.spa.aalto.fi/projects/poririrs/docs/poriref.pdf, 2005. Accessed: 2019-02-26.

[19] K. Kinoshita *et al.*, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *IEEE WASPAA*, 2013.

[20] "Echothief impulse response library." www.echothief.com/downloads/. Accessed: 2019-02-26.

[21] "Sofa general purpose database." www.sofaconventions.org/mediawiki/index.php/Files.

[22] J. K. Nielsen *et al.*, "The single-and multichannel audio recordings database (SMARD).," in *Proc. IWAENC*, 2014.

[23] S. Nakamura *et al.*, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition.," in *LREC*, 2000.

[24] J. Y. Wen *et al.*, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. IWAENC*, 2006.

[25] M. Karjalainen *et al.*, "Estimation of modal decay parameters from noisy response measurements," *J. Audio Eng. Soc*, 2002.

[26] G. A. Soulodre and J. S. Bradley, "Subjective evaluation of new room acoustic measures," *J. Acoust. Soc. Am.*, vol. 98, no. 1, 1995.

[27] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016.

[28] A. Kumar, M. Khadkevich, and C. Fugen, "Knowledge Transfer from Weakly Labeled Audio using Convolutional Neural Network for Sound Events and Scenes," in *IEEE ICASSP*, pp. 326–330, 2018.

[29] J. J. Huang and J. J. A. Leanos, "AclNet: efficient end-to-end audio classification CNN," *CoRR arxiv*, 2018.

[30] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.