PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Performance analysis of audio event classification using deep features under adverse acoustic conditions

Irene MARTÍN-MORATÓ[(1)], Maximo COBOS[(1)], Francesc J.FERRI[(1)], Javier NARANJO-ALCAZAR[(2)]

[(1)]Department of Computer Science, Universitat de València, Spain
[(2)]Visualfy, Benisanó, Spain

Abstract

Audio event classification has been traditionally performed by extracting standard features based on human perception, such as Mel-frequency cepstral coefficients. However, the trend followed in the last years is primarily based on the information provided by deep features, which are extracted from the responses to complex input patterns learned within deep neural networks. These have been shown to obtain, in general, better performance than the hand-crafted ones. In fact, deep features are known to provide good generalization properties to classify events not seen during training, and can even be extracted from raw audio data. Since the captured audio data is highly dependent on the acoustic properties of the auditory scene, it is important to assess the impact that adverse acoustic conditions have in the final classification performance. In this paper, we analyze the robustness of deep features under controlled acoustic conditions by simulating different degrees of background noise and reverberation. Results show an acute degradation in the performance resulting from adverse acoustic conditions, which suggests there is room for improvement in terms of robustness to different scenarios.

Keywords: audio event classification, deep features, transfer learning, convolutional neural networks

1 INTRODUCTION

The concept of smart cities has popularized the use of smart acoustic applications such as acoustic monitoring [2, 3], automatic indexing and tagging [7], environmental context identification or ambient assisted living systems [6]. These systems can be embedded into distributed devices in order to capture and analyze the different signals present in the environment. However, there is a large variety of acoustic conditions that may affect the performance of such systems under different scenarios. Most of the above applications are based on audio classification and machine learning techniques. For that reason, classifiers must be robust against undesired background noises and other adverse conditions. In fact, even with a considerable amount of training data, achieving good performance is not guaranteed when deployment scenarios do not match the training acoustic conditions. Thus, a good set of features providing sufficient robustness and generalization capabilities is needed.

Traditionally, Audio Event Classification (AEC) has been performed using hand-crafted features such as Mel-Frequency Cepstral Coefficients (MFCC), Mel-Frequency Energies (MFE) or MPEG descriptors [15, 20, 24, 5]. Although such features have been shown to work nicely for some classification tasks, selecting appropriately the best features for a specific problem is not straightforward, which motivates the use of feature selection methods [13]. At last, it is the level of expertise of the researcher what determines the final features used, and this in turn may lead to an arduous tuning process with a final result that might be far from optimal.

In recent years, the most common approach for audio analysis, either for classification or detection tasks, is the use of Deep Neural Networks (DNNs), since these have shown superior performance over more traditional machine learning systems [9, 16]. Particularly, Deep Convolutional Neural Networks (CNNs) are capable of capturing energy modulation patterns across time and frequency [8, 10], with internal layers that act as feature extractors from the input and final layers providing class likelihoods given the internal activations for the observed input. The outputs of these internal layers are known as deep features or internal representations and, in contrast to hand-crafted features, they are directly learned by the system with the aim of discriminating among a massive set of input examples. The input to a CNN must be some form of sound representation. Usually,

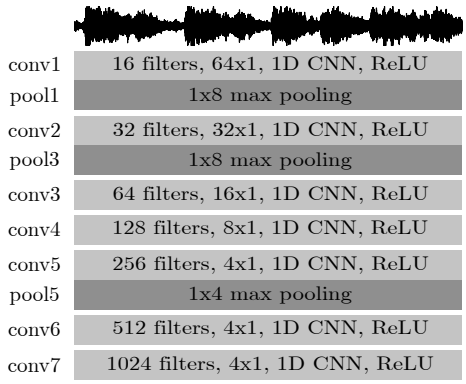


Figure 1. SoundNet architecture

most systems have been designed to accept as input a given time-frequency representation of the audio signal. However, the selected audio representation is independent from the classification stage and it might discard information that can be important for classification. New approaches tend to use the raw audio input inspired by the image recognition task, where the raw RGB pixel values are used as the inputs feeding the CNN [11]. It has been shown in [22] and [21] that CNNs are able to learn time-domain filters from raw acoustic data, performing similarly to the ones used by traditional features.

The purpose of this paper is to analyze the capability of a pre-trained CNN to extract features from the raw audio input robust to adverse acoustic conditions. Specifically, the output from different internal layers of the SoundNet system [4] will be used as features tested under a transfer learning framework. Different datasets will be considered, which will be augmented to simulate different acoustic conditions in terms of event-to-background ratio and reverberation time. The results show that the features extracted from the internal layers of the CNN suffer severely the effects of noise and reverberation, which translates into a significant decrease in the classification performance of the system. These empirical results complement those obtained in previous work [14] by analyzing the effect over different layers of the network and considering different classification contexts with datasets of different size.

The paper is organized as follows. Section 2 describes the structure of the SoundNet baseline system used as feature extractor in this work. Section 3 presents our experimental layout, where we modify the original datasets in order to simulate adverse acoustic conditions. In Section 4 we discuss the robustness of the extracted deep features against different acoustic scenarios. Finally, conclusions are summarized in Section 5.

2 SOUNDNET SYSTEM

SoundNet is a CNN-based system for audio event and sound scene classification [4] that was trained using two million videos from Flickr, using a parallel visual recognition system as a teacher working on unlabeled video as a bridge. The network architecture considers seven convolutional and three maxpooling layers, as seen in Figure 1. The model can be used as a feature extractor even for classes unseen during the training process by using the output of a selected hidden layer as the input to a classifier of choice. It is a well-known fact that layers from different depths tend to learn features sensitive to different description levels. The first layers of the network specialize in extracting low-level descriptors of the input signals, which are less representative of the different sound classes. This can be compared to visual recognition networks, where the first layers capture rough features such as vertical lines, edges or contours. As we go deep through the network, the filters will capture middle-level and high-level features, with an increasing degree of complexity.

In order to get a better understanding of the SoundNet generalization capabilities to different acoustic conditions, we have selected four levels of depth. Three of them correspond to the outputs from a set of convolutional

blocks, composed by a batch-normalization followed by a Rectified Linear Activation Function (ReLu). We will refer to them as *conv5*, *conv6* and *conv7*, where the number indicates the depth of the block within the full architecture. The fourth output corresponds to the one obtained from the last pooling layer (*pool5*) which is the one suggested by the authors of SoundNet as a good choice for performing transfer learning.

2.1 Classification system

Once the deep features from the raw audio files have been extracted from a training dataset, a classifier must be selected and adjusted. Since our aim is to study the impact of the learnt network representations on the final classification performance under different acoustic conditions, it is more appropriate to select a classification model with a small number of parameters to tune, such a linear Support Vector Machine (SVM). This model has been shown to provide good classification results [4] and, according to our experimentation, using more complex classifiers does not lead to significant performance improvements. The linear SVM used in our experiments is trained with $C = 0.01$ following a one-vs-all strategy for the multi-class problem, similar to [25].

3 EXPERIMENTAL SETUP

In order to study the robustness of deep features, several acoustic scenarios were synthetically simulated. The baseline scenario made use of the original datasets, which contain clean data (without noise or reverberation). The adverse acoustic conditions are generated by corrupting the baseline datasets with background noise and/or reverberation. Since we are studying the performance of the system from a classification point of view, we use Accuracy (*Acc*) as performance metric. It is calculated class-wise (macro-averaging) as $Acc = TP/NP$, where NP is the total amount of positive samples and TP the number of correctly classified examples, measured per class. For comparison purposes with previous works we also compute the F-score (*F1*) metric as:

$$F1 = \frac{2 \times Pre \times Rec}{Pre + Rec}, \tag{1}$$

where $Pre = TP/(TP+FP)$ and $Rec = TP/(TP+FP)$, are the Precision and Recall, respectively, calculated averaging over all the instances (micro-averaging). For this metric, TP is the number of correct, FN missed and FP false alarm instances of the complete dataset.

3.1 Synthetically degraded data

For carrying out our evaluation under adverse conditions, two datasets are selected, the first one is ESC-50 (dataset for Environmental Sound Classification) [17], which is a collection of 2000 audio events with a duration of 5 seconds. The events are equally balanced into 50 classes (with 40 examples per class). The second dataset is ESC-10, which is a sub-set of 10 classes (*dog barking, rooster, rain, sea waves, fire crackling, crying baby, sneezing, clock ticking, helicopter and chainsaw*) extracted from the previous one. Both datasets are prearranged into five folds for cross-validation evaluation, hence the results are averages over these folds.

In order to study the robustness of deep features under real adverse acoustical conditions, we modified the test audio files adding reverberation and different levels of noise. The reverberant conditions were simulated using the image-source method [1], considering a wall reflection factor of $\rho = 0.8$ in a rectangular room of dimensions $10\,\text{m} \times 4\,\text{m} \times 3\,\text{m}$, resulting in a reverberation time of $T_{60} = 0.31\,\text{s}$. We fixed the microphone at the position $(2.5, 1.0, 2.0)$, and the sound source location was randomly varied inside the room, following the same procedure as in [14]. Noisy conditions were generated by adding different levels of white Gaussian noise, leading to various Event to Background Ratios (EBRs), $\text{EBR} \in \{-6, 0, 6\}$ dB, as defined in [12]:

$$\text{EBR} = 20\log_{10}\left(\frac{E_{rms}}{B_{rms}}\right), \tag{2}$$

where E_{rms} and B_{rms} are the root mean square values of the event and the background, respectively. Taking the above into account, we generated four acoustic scenarios:

- **Original**: the examples of the test folds are perfectly isolated and there is no artificial additive noise. Thus, the examples are not modified with respect to the original dataset.
- **Reverberant**: the examples of the test folds are convolved with synthetic reverberant impulse responses, but no noise is artificially added.
- **Original+noise**: the examples of the test folds include different levels of background noise, but reverberation effects are omitted.
- **Reverberant+noise**: besides including reverberation, the examples of the test folds contain different levels of background noise.

As explained in Section 2, the results are presented for different network depths in order to evaluate the performance according to different representation levels learnt by the original SoundNet system. Finally, a third database is used in our analysis, in order to study how well SoundNet works as feature extractor on a bigger and more realistic dataset. The data used is from UrbanSound8k [19], which contains 8732 labeled audio examples from 10 different urban sound categories: *air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music*. The data is 4 seconds long and the dataset is pre-sorted in ten folds for easy reproduction and comparison.

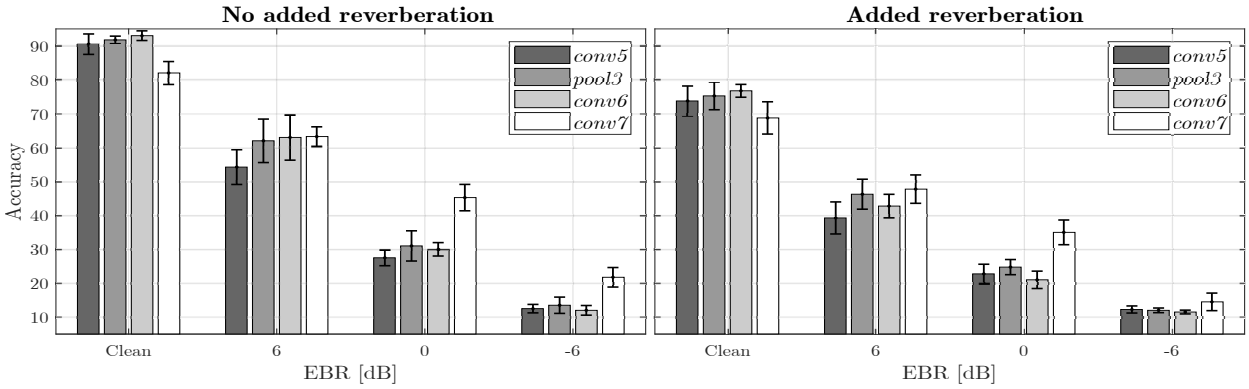


Figure 2. Accuracy values for ESC-10 under adverse scenarios, along with the respective standard deviation.

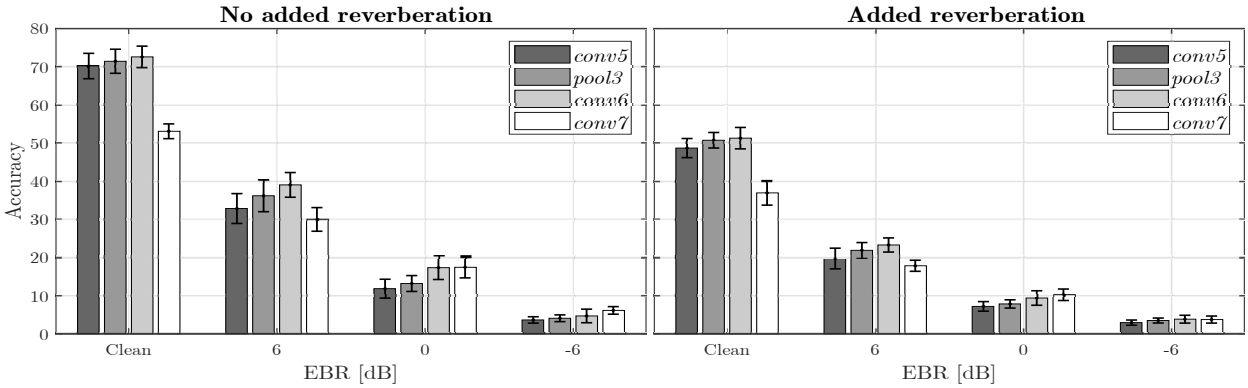


Figure 3. Accuracy values for ESC-50 under adverse scenarios, along with the respective standard deviation.

Table 1. F-score for ESC-10 averaged over the five folds, training with clean data

	No added reverberation				*Added reverberation*			
Depth	clean	6dB	0dB	-6dB	clean	6dB	0dB	-6dB
conv5	92.22 (2.76)	63.05 (5.24)	36.84 (2.89)	18.81 (1.66)	77.42 (3.88)	48.88 (5.46)	31.63 (3.73)	18.52 (1.43)
pool5	93.51 (0.89)	70.84 (4.94)	40.99 (4.99)	20.18 (3.26)	79.30 (3.41)	57.12 (4.36)	34.24 (2.88)	18.18 (0.94)
conv6	94.31 (1.85)	70.46 (4.97)	39.91 (2.09)	18.16 (1.97)	81.24 (2.18)	53.65 (3.46)	29.80 (3.12)	17.49 (0.77)
conv7	85.08 (2.83)	71.57 (2.66)	55.09 (4.38)	30.30 (3.32)	74.73 (5.22)	58.34 (3.79)	46.22 (3.83)	21.52 (3.55)

Table 2. F-score for ESC-50 averaged over the five folds, training with clean data

	No added reverberation				*Added reverberation*			
Depth	clean	6dB	0dB	-6dB	clean	6dB	0dB	-6dB
conv5	75.79 (2.72)	42.09 (3.99)	17.19 (3.39)	5.50 (1.20)	56.68 (2.06)	27.33 (3.52)	10.63 (1.76)	4.35 (0.94)
pool5	76.79 (2.38)	46.47 (4.65)	19.20 (2.97)	6.17 (1.33)	58.59 (1.78)	30.20 (2.52)	11.66 (1.55)	5.16 (0.91)
conv6	77.85 (2.31)	48.96 (3.28)	24.45 (4.02)	7.15 (2.64)	58.62 (2.45)	31.57 (1.77)	13.78 (2.52)	5.81 (1.53)
conv7	59.54 (1.37)	37.68 (3.51)	22.59 (3.60)	7.96 (1.47)	43.53 (3.36)	23.16 (1.71)	13.33 (2.16)	4.66 (1.26)

4 RESULTS AND DISCUSSION

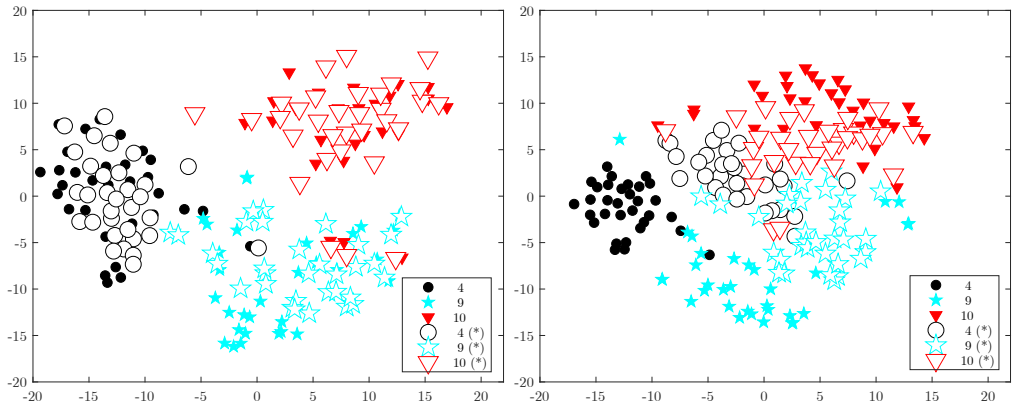
This section discusses the results for the experiments described in the previous section. The results will be presented for two different training situations. First, it is considered that the system has been trained using the original examples, but tested under different degrees of noise and reverberation. Then, the performance will be compared to the case when the training conditions match those of the test, but considering as well different acoustic scenarios. The results here presented have been obtained using the code provided by the authors of SoundNet [25]. For inference, the scores of the different temporal frames resulting from a given example are averaged, selecting as predicted label the class with the highest mean score.

4.1 Testing with synthetically degraded data

As mentioned before, different hidden layers may be used as feature extractors providing alternative levels of abstraction. Figures 2 and 3 show the accuracy of the classifier when it is fed with features extracted from different network depths considering the ESC-10 and ESC-50 datasets, respectively. As a reference, Table 1 specifies the F-score for the same testing conditions in ESC-10. As a general observation, SVM-based transfer learning from the pre-trained SoundNet model achieves state-of-the-art performance when the datasets are not artificially degraded (leftmost *clean* column, i.e. no artificial noise nor reverberation). In this ideal case, all the layer outputs present similar performance metrics except *conv7*, which seems to perform considerably worse than the rest. A possible explanation may come from the high level of specialization of this layer. Since it has the highest depth, the features may be more overfitted to the SoundNet original classes. When noise and reverberation are artificially added to the test folds, the performance decreases substantially for all the layers. The addition of reverberation has also a significant effect on the final accuracy, contributing strongly to the observed performance loss. The trend followed by the results is very similar both for ESC-10 and ESC-50, with an expected offset between them due to the higher number of classes of the latter. Note, however, that even when the number of classes is relatively small (ESC-10), going from the clean case to an EBR of 6 dB produces a performance decrease of approximately 30% in accuracy. In addition, the standard deviation (given between brackets on the tables) increases with respect to the ideal scenario whenever noise or reverberation are added. Surprisingly, the *conv7* layer is the one that seems to be more robust to noise, especially in the ESC-10 dataset. This suggests that the best features considering an ideal test scenario may not be necessarily the most robust when examples are observed under changing acoustic conditions.

Table 3. Performance values when training data has the same conditions as test

EBR	ESC-10				ESC-50			
	No added reverberation		*Added reverberation*		*No added reverberation*		*Added reverberation*	
[dB]	Acc	F1	Acc	F1	Acc	F1	Acc	F1
clean	93.00 (1.43)	94.31 (1.85)	84.25 (2.88)	87.19 (2.39)	72.55 (2.81)	77.85 (2.31)	63.10 (4.17)	69.53 (3.61)
6	89.75 (3.79)	91.70 (3.69)	77.75 (4.37)	82.04 (4.16)	67.80 (2.51)	73.51 (2.21)	56.45 (2.31)	63.32 (1.91)
0	86.50 (4.28)	88.71 (3.88)	74.25 (5.70)	79.37 (5.21)	61.50 (1.83)	68.37 (1.57)	49.95 (2.00)	56.84 (2.04)
-6	82.50 (2.50)	85.71 (2.18)	68.75 (2.50)	75.14 (2.95)	53.00 (2.48)	60.20 (1.79)	39.75 (1.87)	46.36 (2.46)



(a) (b)

Figure 4. t-SNE embeddings using deep features from conv6. (a) original data versus synthetically degraded with 6dB EBR (b) original data versus synthetically degraded with 6dB EBR and reverberation. Degraded data marked with an asterisk in both legends.

4.2 Traning and testing with identical conditions

For comparison purposes, Table 3 shows the performance of the *conv6* layer when the training conditions match those of the test conditions. This layer has been selected for further analysis for being the one providing the best average performance across all the tested conditions with both datasets. The results for all cases are, in general, considerably better than those of Tables 1 and 2, as now the training and test examples should ideally follow similar distributions. However, despite matching conditions, the performance in adverse scenarios is below the one obtained for the ideal clean case. In general, F-score gives better results than accuracy because it is computed over all the instances of the dataset, reducing the impact of the worst performing classes. This situation can be graphically illustrated by embedding some of the represented events onto a two dimensional space using t-SNE [23]. In Figure 4 (a) we show 32 examples from three classes (*sea waves* (4), *helicopter* (9) and *chainshaw* (10)) jointly embedded with their noisy (6dB) version, we can see how the features only differ slightly from the non-artificially degraded ones. The same is shown in Figure 4 (b) but adding also reverberation. In this case, the features from the modified files are considerably more separated form their original ones.

Finally, Table 4 collects the results for the Urbansound8k dataset, again with matching conditions in training and test without any alteration on the original dataset. While the number of classes is 10, the amount of examples per class is considerably higher than in ESC-10, showing as well great variability within the examples of each class. The results obtained achieve state-of-the-art performance [18, 21], confirming that deep features from SoundNet are able to generalize well to different audio event datasets. Note that, again in this case, it is *conv6* the layer providing the best performance.

Table 4. Performance values for Urbansound8k averaged over the ten folds

Acc			*F1*		
pool3	conv6	conv7	pool3	conv6	conv7
68.22 (5.93)	71.81 (5.05)	64.23 (4.41)	73.76 (5.18)	76.85 (3.97)	72.44 (4.19)

5 CONCLUSION

In this paper, we studied the robustness of deep features under several adverse conditions. As expected, performance drops dramatically with the degradation level specially if the final predictor used has been trained with data that does not match the specific acoustic conditions. We have analyzed different internal representations with different levels of adaptation and have observed that the performance drop has a different impact depending of the depth of these representations. The results obtained in this and previous studies suggest that the performance in practical and realistic scenarios could be improved by specifically considering different models of degradation, not only to obtain the final predictor but also to extract even more robust features.

ACKNOWLEDGEMENTS

This work has been partially supported by FEDER and Spanish Gov. through projects TIN2014-59641-C2-1-P, TIN2014-54728-REDC, RTI2018-097045-B-C21, FPU14/06329. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 779158.

REFERENCES

[1] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Amer.*, 65(3):943–950, 1979.

[2] R. M. Alsina-Pagès, J. Navarro, F. Alías, and M. Hervás. homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring. *Sensors*, 17(4), 2017.

[3] S. P. P. Aravinda, S. Gunawardene, and N. Kottege. An acoustic wireless sensor network for remote monitoring of bird calls. In *2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 1–4, Dec 2016.

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

[5] S. Chu, S. Narayanan, and C. C. J. Kuo. Environmental sound recognition with time-frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1142–1158, Aug 2009.

[6] M. Cobos, J. J. Perez-Solano, and L. T. Berger. *Chapter 9 Acoustic-Based Technologies for Ambient Assisted Living*, pages 159–180. CRC Press, 2016.

[7] K. Ellis, E. Coviello, A. B. Chan, and G. Lanckriet. A bag of systems representation for music auto-tagging. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(12):2554–2569, Dec 2013.

[8] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani. Exploiting spectro-temporal locality in deep learning based acoustic event detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):26, Sep 2015.

[9] O. Gencoglu, T. Virtanen, and H. Huttunen. Recognition of acoustic events using deep neural networks. In *2014 22nd European Signal Processing Conference (EUSIPCO)*, pages 506–510, Sept 2014.

[10] L. Hertel, H. Phan, and A. Mertins. Comparing Time and Frequency Domain for Audio Event Recognition Using Deep Learning. *ArXiv e-prints*, Mar. 2016.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[12] G. Lafay, M. Lagrange, M. Rossignol, E. Benetos, and A. Roebel. A morphological model for simulating acoustic scenes and its application to sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1854–1864, Oct 2016.

[13] I. Martin-Morato, M. Cobos, and F. J. Ferri. A case study on feature sensitivity for audio event classification using support vector machines. In *26th IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2016, Vietri sul Mare, Salerno, Italy, September 13-16, 2016*, pages 1–6, 2016.

[14] I. Martin-Morato, M. Cobos, and F. J. Ferri. On the robustness of deep features for audio event classification in adverse environments. In *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pages 562–566, Aug 2018.

[15] I. Martín-Morató, M. Cobos, and F. J. Ferri. Adaptive mid-term representations for robust audio event classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2381–2392, Dec 2018.

[16] Y. Petetin, C. Laroche, and A. Mayoue. Deep neural networks for audio scene recognition. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 125–129, Aug 2015.

[17] K. J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

[18] K. J. Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Sept 2015.

[19] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA, Nov. 2014.

[20] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, Oct 2015.

[21] Y. Tokozume and T. Harada. Learning environmental sounds with end-to-end convolutional neural network. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2721–2725, March 2017.

[22] Z. Tüske, P. Golik, R. Schlüter, and H. Ney. Acoustic modeling with deep neural networks using raw time signal for lvcsr. In *INTERSPEECH*, 2014.

[23] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *JMLR*, 9:2579–2605, 2008.

[24] E. Vozáriková, J. Juhár, and A. Čižmár. Acoustic events detection using mfcc and mpeg-7 descriptors. In A. Dziech and A. Czyżewski, editors, *Multimedia Communications, Services and Security*, pages 191–197, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[25] A. T. Yusuf Aytar, Carl Vondrick. Soundnet: Learning sound representations from unlabeled video.