TRANSFER LEARNING OF WEAKLY LABELLED AUDIO

Aleksandr Diment, Tuomas Virtanen

Tampere University of Technology
Laboratory of Signal Processing
Korkeakoulunkatu 1, 33720, Tampere, Finland
firstname.lastname@tut.fi

ABSTRACT

Many machine learning tasks have been shown solvable with impressive levels of success given large amounts of training data and computational power. For the problems which lack data sufficient to achieve high performance, methods for transfer learning can be applied. These refer to performing the new task while having prior knowledge of the nature of the data, gained by first performing a different task, for which training data is abundant. Shown successful for other machine learning tasks, transfer learning is now investigated in audio analysis. We propose to solve the weakly labelled problem of sound event tagging with small amounts of training data by transferring the abstract knowledge about the nature of audio data from another tagging task. The proposed methods constitute pre-training of a recurrent neural network or its parts to perform one tagging task given abundant and diverse training data, and then using it or its parts for a new task of tagging sound events of different nature, for which the data is limited. Several architectures for such transfer are proposed and evaluated, showing impressive classification accuracy of 83.4% with gains of up to 20 percentage points over the baseline given as little as 36 training samples for the target task.

Index Terms— transfer learning, tagging, weak labels, audio

1. INTRODUCTION

Automatic analysis of everyday sounds, with its numerous potential applications (smart homes, smart cities, context-aware devices), is attracting increasing attention. New datasets are being published [1, 2], and international challenges and workshops are organised [3]. The goal of these activities is to motivate the development of such audio analysis methods that would be able to perform robustly in real-life conditions: a problem currently still unsolved.

The problem of robustness in real-life conditions exists due to the unpredictable nature and endless diversity of noise and distortions of the real-life signals, as well as the diversity of the target signals themselves. In the current age of data-driven machine learning, the intuitive solution is to use large amounts of data to train a deep neural network. However, collecting annotated data with sufficient diversity is very time consuming and costly. On the other hand, various sounds share similar properties, and to learn those, to get a general idea about the structure of acoustic signals, could be much more reasonable and scalable. Obtaining such a general audio-aware tool does still require a lot of data and resources. However, the intention is to be able to reuse it for the new, previously unseen, but still related audio tasks.

Transfer learning [4] has been successfully used for years in various machine learning fields such as text analysis [5], natural language processing [6] and image classification [7], to mention a few. With the advancements of deep learning [8], it has become standard to reuse networks readily pretrained on large datasets, such as ImageNet [9] when solving a new image analysis problem. However, in audio analysis field, such practice is not widespread. We believe that one of the reasons is the following: while it is possible to visualise the features learnt by the first layers of a convolutional neural network trained on image data and to show that those do capture basic shapes and textures, common to images, such approach is less feasible with audio.

Weak labels is another concept, relevant to real-life audio analysis. Collecting large amounts of data required to perform well in diverse conditions is expensive since precise annotations of temporal occurrences and classes of sounds of interest need to be performed by a human listener. With the exploding amounts of raw data available, there is an interest of developing such audio analysis methods, that would not require this level of precision of annotations. Instead, it is desirable to be able to work with so-called weakly labelled data, whose annotations include such information as "somewhere within this temporal region there is a sound of interest occuring". In contrast to sound event detection, this weak label setting refers to sound event tagging. An exciting goal would be to develop such a system, which could be trained with massive weakly labelled data, but capable of outputting strong labels at the analysis stage. Several works have recently addressed the problem of weak labels in audio, such as [10, 11].

Works on problems related to audio transfer learning are yet very limited but promising. Research on emotional audio has successfully employed these ideas, either by transferring between same-domain datasets [12], or even between music and speech domains [13]. For the problems related to environmental audio, one example is self-training of an audio event detector, which is initially trained on a small set of strongly labelled data, followed by a semi-supervised stage with massive unlabelled data [14].

We propose approaches for transferring knowledge about the structure of audio signals, gained by performing one audio analysis task, into a different task, for which the training data is very limited. The source and target tasks of the proposed setup are tagging of sound events of different classes. We use tagging of baby cries as a source task and tagging of glass breaks as a target task. The selection of such classes is motivated by the difference in their acoustic properties (making the transfer more challenging, while better demonstrating its potential), as well as by the applicability of such classifiers in real-life applications (e.g. smart homes [15]). We study whether learning to do tagging of one sound event class given large and diverse data is helpful to later perform tagging of sound events of a different class, for which the training data is limited, even if they do not share much of the acoustic properties (baby cries exhibiting fewer abrupt

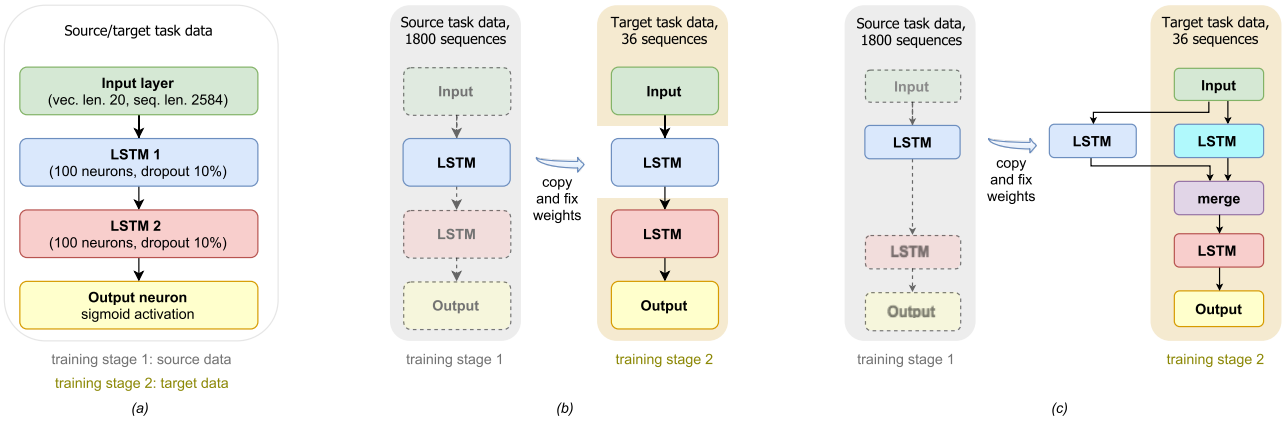

Figure 1: Architectures studied in this work. The baseline with no transfer is implemented with the architecture (a). It is also used in the first method, where the network is initialised on a source task and then fully retrained on a target task. In (b), the network on the left is trained on the source task with large amount of data (stage 1), and then the weights of the first LSTM layer are copied to the corresponding layer of the network on the right, while the remaining weights are fine-tuned for the target task with little training data (stage 2). In (c), the network on the left is first trained on source task data (stage 1), after which its first LSTM layer with pre-learned and fixed weights is incorporated into the network on the right using an additional merge layer. The rest of the weights of the network on the right are trained on the small target task dataset (stage 2).

transients and being more stationary than glass breaks). Transferring the knowledge would therefore stand for transferring the generally useful insight about structure of environmental audio and ways to efficiently perform detection of "interesting" event classes in audio. We simulate the abundant and limited datasets by creating mixtures of recordings of acoustic scenes and sound events.

Next, we present the proposed approaches for incorporating the source task knowledge into the target task problem. Thereupon, details of data acquisition and generation are given, followed by the evaluation results and conclusions.

2. METHODOLOGY

The transfer learning setup consists of a source task and a target task. The source task can be performed well due to the abundance of available training data. To perform the target task given a limited amount of data, potentially useful knowledge learnt from the source task is transferred to the target task. Here, we investigate ways of transferring such knowledge from one audio tagging task to another.

We propose three methods for transfer learning of weakly labelled sounds. They all are based on the following many-to-one recurrent neural network architecture: two LSTM [16] layers (100 neurons each, tanh activation, 10% dropout for input gates and no reccurent dropout) connected to one output neuron (sigmoid), which predicts whether the sound event was detected somewhere in the input sequence of audio frames (2584 frames in a sequence, corresponding to 30 seconds of audio). See Figure 1 (a) (one-stage version) for the illustration of the basic architecture.

The first approach (also in Figure 1 (a), the two-stage version, referred also to as *pre-training of all layers*) is simply initialising the network with the source task data and then using it as the starting state for the target task. During the target task learning stage, the network would be tuned across all its layers. That is, we investigate the usefulness of pre-training the network to do a different but conceptually similar task. A potential drawback of such setup is that there is no guarantee that the network will not "forget" everything it has learnt during the source task stage after it has been tuned for the target task. Still, we believe that initialising on a diverse audio data is more helpful than a data-unaware initialisation.

The second approach, referred to as *low-level pre-train and fix*, as seen in Figure 1 (b), consists of partially enforcing the importance of retaining the knowledge learnt from performing the source task during the transfer. The network of the base architecture is first trained for the source task (training stage 1). Thereupon, the low-level features learnt from this stage are fixed, and only the higher-level features are fine-tuned for the target task (training stage 2). In practice, this means fixing the weights of the first LSTM layer when performing the retraining of the rest of the network on the target task data. The intention is that the network on its first LSTM layer would learn the structure of audio data in general, and that detection of different sound event classes would happen on a higher level of abstraction. The assumption here is that the first LSTM layer would be capable of capturing generally useful features for audio analysis tasks when trained to perform only one but with diverse enough data. Fine tuning for a particular new unseen class of sound events could therefore be done with a smaller training dataset, if this assumption holds.

Finally, the third approach, *parallel lower level layers*, as in Figure 1 (c), is the combination of the first two: it enables both adaptation to the target task on all the abstraction levels and retaining of the knowledge about the general structure of audio data learnt while performing the source task. That is, we allow training on all the levels of abstraction for the new task as well as fixing the lower level features learnt from the source task. In practice, we propose to implement such an arrangement with two parallel lower-level LSTM layers, one of which will have weights learnt from the source task and then fixed, while the other will be trainable on the new data. The outputs of these layers are merged by concatenation prior to the next LSTM layer.

All the networks are trained using binary cross-entropy loss and rmsprop optimiser (learning rate 0.001). For data-unaware initialisation of the weights, Glorot normal initializer [17] is used. We do not use validation data based early stopping due to the scarcity of target task data, but rather terminate each training stage after 300 epochs. The implementation uses keras package [18]. Run on a Tesla K80 GPU, one epoch takes approximately 100 seconds at the pre-training stage and 20 seconds at the fine-tuning stage.

3. DATA

A dataset of weakly labelled data was generated for this work. It consists of real-life recordings of various acoustic scenes [2] with sound events artificially added to half of the recordings at random temporal locations with various event-to-background ratios. Given sufficient diversity of underlying real-life recordings, a dataset of mixtures was generated, large enough for the task of investigating the applicability of the proposed methodology for transfer learning. Here, we describe the details of the dataset and its generation. Both the source data and software for mixture generation are released[1] for academic research as part of the DCASE 2017 challenge [19].

3.1. Source recordings

The dataset was generated using background recordings from 15 different acoustic scenes and sound event recordings from two classes (baby cry and glass break), obtained from Freesound [20]. The source files were accompanied by information to perform train-test splits and source-target task splits. For the background recordings, it is the location id (so that recordings from the same location would not come in both training and test sets) and acoustic scene class (so that source and target tasks would not share the same classes of acoustic scenes). For the sound events, the train-test split is to be done over the authors of the original recordings at Freesound (so that no same user name would appear in both training and test sets).

The background recordings are an almost exact copy of TUT Acoustic scenes 2016, development dataset [2], with the exception of recordings naturally containing target class events, annotated by a human listener and then excluded. The classes of the acoustic scenes are the following: bus, cafe/restaurant, car, city centre, forest path, grocery store, home, beach, library, metro station, office, residential area, train, tram, and park. The dataset totals to 39 minutes of audio.

The target sound event recordings originate from Freesound and are accompanied here by precise annotations of their temporal occurrences. We collected the isolated sound events from Freesound in the following manner. All the recordings matching the target class name query and with a sampling rate $f_s \geq 44100$ Hz were downloaded using the API with python wrapper.[2] The target sound events were then isolated from the recordings using a two-step procedure. First, a semi-supervised segmentation [21] was performed with an SVM model trained to distinguish between high-energy and low-energy short-term frames and then applied on the whole recording. A dynamic thresholding was used to detect the active segments.

The events were then screened by a human listener and only those clearly corresponding to the target class were retained (e.g. baby cry recordings included sounds of baby sighs and coughs, which were discarded). The temporal annotations were manually refined for all the isolated events with a step of 100 ms in such a way that there would not be abrupt clicks on the boundaries, but no silence regions before or after the events either. The statistics of the isolated events are:

- baby cry: 106 training and 42 test instances, mean duration 2.25 seconds (standard deviation 0.98),
- glass break: 96 training and 43 test instances, mean duration 1.16 seconds (standard deviation 0.71).

[1] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-rare-sound-event-detection

[2] https://github.com/xavierfav/freesound-python-tools

Due to the nature of these sounds, there still are some regions of silence inside the annotated events (e.g. a baby cry consisting of two phrases, annotated as one cry). Performing a frame-level annotation on target event presence was deemed unfeasible. However, events with pauses longer than one second were eliminated.

The detailed temporal annotations, albeit not of the highest importance for the weak labelling task, were nevertheless useful. They allowed generation of mixtures in a highly controlled manner. Each mixture contains at most one target sound event (as defined above, series of two sound event with a pause shorter than one second in between is referred to as one), and multiple sound events in a longer source file can be used separately in different mixtures.

3.2. Mixture generation

The background recordings were segmented into 30-second long sequences. Thus, a classification instance in this work is a 30-second long recording, referred to as *mixture*, which is a background recording with or without an added target sound event. During mixture generation, the event presence rate was set to 0.5, yielding balanced sets in all cases. The event-to-background ratios (EBR) were set to a random value from a list of -6, 0 and 6 dB for each mixture with a target event. This value refers to the logarithmic ratio of the RMS energies of the event and the corresponding background segment, onto which the event was mixed.

For each mixture, the background instance, the event instance, the event presence flag, its timing in the mixture and the EBR value were all selected randomly and uniformly, allowing for a generation of an infinite number of mixtures, which do, however, share the underlying source data. This is not an issue for the cross-validation setup, since the split is performed in terms of the origin of the source data.

For the sound event files with sampling rate different from the target 44100 Hz (higher was allowed), resampling to 44100 Hz was performed prior to the summation. To avoid clipping, the mixtures were scaled with a factor of 0.2 (found experimentally suitable for the given dataset parameters). All the mixtures were scaled, not only the clipping ones, so that the dynamics were preserved. To avoid introducing quantisation noise, the files were saved in 24 bit format.

3.3. Training and test sets

The training sets contain 1800 instances of 30-second audios (backgrounds with or without target class sound events) for the *full data* case and 36 instances for the *limited data* case. The test set contains 500 instances and is generated with different but fixed underlying source data (as described above, different location ids for backgrounds and different authors of recordings for the target events). Both in all the training sets and in the test set, the target sound events occur in exactly half of the cases.

There is an additional dimension of the split. Namely, the source and target tasks never share underlying data used in training and test mixtures. For the source task of baby cry tagging, we use background material from one set of contexts (beach, bus, cafe/restaurant, car, city centre, forest path and grocery store), while the mixtures generated for the target task use background recordings of different contexts (home, library, metro station, office, park, residential area, train). This way, we ensure that the potential benefit of transfer learning does not originate from network memorising the background data, but rather learning to perform tagging in general, in various diverse settings.

3.4. Features

As features, we use 20 MFCC [22] coefficients, extracted in frames of 2048 samples (46 ms with sample rate 44100 Hz) with hop size of 512 samples (12 ms). This results in sequences of 2584 frames for each 30-second mixture. Delta features are not used, as motivated by the recurrent properties of the employed neural network. Librosa [23] framework is used for feature extraction.

We do not perform any optimisation of the feature extraction parameters in this work, since the primary objective is to investigate the improvement introduced by the proposed transfer learning architecture over the baseline, while keeping the rest of the variables fixed. We use MFCCs as a feature, shown applicable for various audio analysis problems, including real-life sound event detection [24].

4. EVALUATION

In this section, we evaluate the proposed architectures for transfer learning and compare their performance to the baselines. As a lower boundary, we have a network as shown in Figure 1 (a), trained on the limited dataset of target task (18 positive and 18 mixtures only). Evaluated on the 500 mixtures of the test set, it shows 62.4 classification accuracy. Incorporating the proposed methods for transferring the insight about the audio data in general from a source task is expected to outperform this baseline method.

As the top boundary, we show the results of the basic network (Figure 1 (a), with only one stage) when full dataset (1800 mixtures) is available for the target task. Such network achieves classification accuracy score of 94.6. Getting even remotely closer to this value in the limited data transfer learning case would be considered a success.

The evaluation results are presented in Table 1. We see that pre-initialising the target task network by first performing source task training with the full dataset (Figure 1 (a)) is, indeed, helpful. We achieve 10 pp (percentage points) performance boost in this set up, and the network does not "forget" what it has learnt from the source task after 300 epochs, even though all the weights are allowed to be retrained. This is an encouraging result despite its simplicity.

The second proposed method, referred to as *low-level pre-train and fix* (Figure 1 (b)) consisted of fixing the weights of the first LSTM layer trained on the full source task dataset, and retraining only the rest of the network with limited data of target task. The evaluation shows a 4 pp degradation of performance with this set up compared to the baseline. From this negative result, we learn that the first LSTM layer is not capable of capturing all the possible properties of acoustic data when trained to perform one task, even though the data is abundant. Apparently, the sounds of baby cries and glass breaks are acoustically so different, that both tasks need at least some low-level descriptors fine-tuned. Performing the first training stage on a multiclass problem with events of different acoustic properties is a natural potential solution to this problem. Also, convolutional neural networks seem a valid alternative for the lower level layer.

Finally, the third proposed method with parallel low-level LSTM layers, each trained for their own task, as shown in Figure 1 (c), shows impressive results. A boost of more than 20 pp in accuracy is achieved. This result is noticeably closer to the upper boundary full-data case than it is to the limited data baseline. We see that, indeed, the network managed to some extend to learn to perform audio tagging in general during the first stage, and then transferred the obtained knowledge to the new task. Keeping half of the lower-level weights trainable in the second stage showed to be useful.

We also evaluate the performance of the proposed *parallel lower layers* architecture with a variable amount of training data for the target task. The results are presented in Figure 2. One-shot learning (with only one positive training example) appears to beat the *from scratch* case slightly, but such small difference might be explained by random effects. However, the improvement provided by transfer learning remains apparent with increasing amount of data, showing most prominence in the case of 32 samples. We see the validity of the proposed approach for the cases of limited training data.

method	**test acc., %**
Limited training data (36 examples)	
from scratch	62.4
pre-training of all layers (a)	72.4
low-level pre-train and fix (b)	58.6
parallel lower level layers (c)	**83.4**
Full training data (1800 examples)	
from scratch	94.6

Table 1: Evalutaion results. *From scratch* refers to the architecture (a) in Figure 1 in one stage, without transfer learning. *pre-training of all layers* refers to architecture (a), first trained on full source task data and then using the obtained weights as initialisation for training all the layers on limited target task data. *Low-level pre-train and fix* is architecture (b), and *parallel lower level layers* is architecture (c).

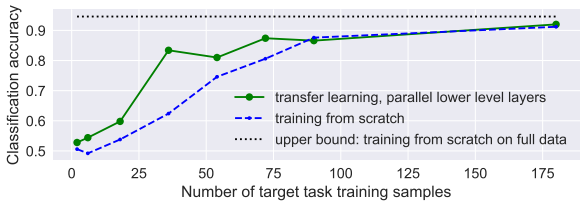

Figure 2: Evaluation results of the *parallel lower level layers* and *from scratch* methods depending on the amount of training data.

5. CONCLUSIONS

We have proposed intuitive yet effective techniques for transfer learning of weakly labelled audio: a simple pre-training of an LSTM network on a different yet conceptually similar tagging task, as well as a network with parallel lower level layers, each trained on a separate task. Pre-training of the whole network has shown 10 pp improvement of the accuracy. The parallel architecture, trained only on 36 target class samples, has performed successfully (20 pp boost).

Many exciting extensions of the approach are to come. The ultimate goal is an architecture for a general-purpose neural network, which could be pre-trained on abundant audio data to have an "idea" about sound signals in general and then fine-tuned for a new task rapidly and with very small amounts of data.

6. ACKNOWLEDGMENT

The authors thank Giambattista Parascandolo for the fruitful discussions about neural networks and related matters. Generous GPU resources, provided by CSC – IT Center for Science, Finland, are gratefully acknowledged. The collection of the acoustic scene recordings used in this work was funded by the European Research Council under Grant Agreement 637422 EVERYSOUND.

7. REFERENCES

[1] 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 *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[2] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference 2016 (EUSIPCO 2016)*, Budapest, Hungary, 2016.

[3] T. Virtanen, A. Mesaros, T. Heittola, M. Plumbley, P. Foster, E. Benetos, and M. Lagrange, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*. Tampere University of Technology. Department of Signal Processing, 2016.

[4] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[5] R. Raina, A. Y. Ng, and D. Koller, "Constructing informative priors using transfer learning," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 713–720.

[6] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proceedings of the 2006 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2006, pp. 120–128.

[7] P. Wu and T. G. Dietterich, "Improving SVM accuracy by training on auxiliary data sources," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 110.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 05 2015. [Online]. Available: http://dx.doi.org/10.1038/nature14539

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR09*, 2009.

[10] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "A joint detection-classification model for audio tagging of weakly labelled data," *Proceedings of ICASSP 2017*, 2017.

[11] A. Kumar and B. Raj, "Audio event and scene recognition: A unified approach using strongly and weakly labeled data," *CoRR*, vol. abs/1611.04871, 2016. [Online]. Available: http://arxiv.org/abs/1611.04871

[12] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 511–516.

[13] E. Coutinho, J. Deng, and B. Schuller, "Transfer learning emotion manifestation across music and speech," in *Neural Networks (IJCNN), 2014 International Joint Conference on*. IEEE, 2014, pp. 3592–3598.

[14] A. Shah, R. Badlani, A. Kumar, B. Elizalde, and B. Raj, "An approach for self-training audio event detectors using web data," *arXiv preprint arXiv:1609.06026*, 2016.

[15] S. Sigtia, A. M. Stark, S. Krstulovi, and M. D. Plumbley, "Automatic environmental sound recognition: Performance versus computational cost," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2096–2107, Nov 2016.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[17] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." in *Aistats*, vol. 9, 2010, pp. 249–256.

[18] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[19] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, R. Badlani, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017.

[20] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *ACM International Conference on Multimedia (MM'13)*, ACM. Barcelona, Spain: ACM, 21/10/2013 2013, pp. 411–412.

[21] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PLOS ONE*, vol. 10, no. 12, pp. 1–17, 12 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0144610

[22] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.

[23] B. McFee, M. McVicar, O. Nieto, S. Balke, C. Thome, D. Liang, E. Battenberg, J. Moore, R. Bittner, R. Yamamoto, and et al., "librosa 0.5.0," Feb 2017.

[24] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in *18th European Signal Processing Conference (EUSIPCO 2010)*, Aalborg, Denmark, 2010, pp. 1267–1271.