# MULTI-TEMPORAL RESOLUTION CONVOLUTIONAL NEURAL NETWORKS FOR THE DCASE ACOUSTIC SCENE CLASSIFICATION TASK

*Alexander Schindler*

Austrian Institute of Technology
Center for Digital Safety and Security
Vienna, Austria
alexander.schindler@ait.ac.at

*Thomas Lidy, Andreas Rauber*

Vienna University of Technology
Institute of Software Technology
Vienna, Austria
lidy,rauber@ifs.tuwien.ac.at

## ABSTRACT

In this paper we present our DCASE 2017 Challenge on Detection and Classification of Acoustic Scenes and Events contributions. We propose a parallel Convolutional Neural Network architecture for the task of classifying acoustic scenes and urban sound scapes. We propose a Deep Neural Network architecture for the task of acoustic scene classification which harnesses information from increasing temporal resolutions of Mel-Spectrogram segments. This architecture is composed of separated parallel Convolutional Neural Networks which learn spectral and temporal representations for each input resolution. The resolution are chosen to cover fine-grained characteristics of a scene's spectral texture as well as its distribution of acoustic events. The best performing variant of the proposed model scores 90.54% accuracy on the development dataset. This is a 6.81% improvement of the best performing single resolution model and 15.74% of the DCASE 2017 Acoustic Scenes Classification task baseline [1].

## 1. INTRODUCTION

Convolutional Neural Networks (CNN) [2] have become a popular choice in computer vision due to their ability to capture nonlinear spatial relationships which is in favor of tasks such as visual object recognition [3]. Their success has fueled the interest as well in audio-based tasks such as speech recognition and music information retrieval. An interesting sub-task in the audio domain is the detection and classification of acoustic sound events and scenes, such as the recognition of urban city sounds, vehicles, or life forms, such as birds [4]. The IEEE AASP Challenge DCASE is a benchmarking challenge for the "Detection and Classification of Acoustic Scenes and Events". Acoustic Scene Classification (ASC) in urban environments (task 1) is one of four tasks of the 2016 and 2017 competition. The goal of this task is to classify test recordings into one of predefined classes that characterizes the environment in which it was recorded, for example "metro station", "beach", "bus", etc. [1].

The presented approach attempts to circumvent various limitations of Convolutional Neural Networks (CNN) concerning audio classification tasks. By applying CNNs to an audio analysis task it is transposed to the visual computing domain. A common approach is to use Short-Term Fourier Transform (STFT) to retrieve a Spectrogram representation which is in the following interpreted as a gray-scale image. Commonly a Mel-Transform is applied to scale the Spectrogram to a desired input size. In previous work we have introduced a CNN architecture to learn timbral and temporal representations at once. This architecture takes a Mel-Spectrogram as input and reduces this information in two parallel CNN stacks towards the spectral and the temporal dimension. The combined representations are input to a fully connected layer to learn the concept relevant dependencies. The challenge is how to choose the length of the input analysis window. Acoustic events can be single sounds or compositions of multiple sounds. Acoustic scenes could be described by the presence of a single significant acoustic event such as the sound of the waves at the beach or by combinations of different events. The temporal pattern of such combinations varies distinctively across and within the acoustic scenes (see Figure 1 for examples of acoustic scenes). Choosing the wrong size of the analysis window can either prevent from having sufficient timbral resolution or to fail to recognize acoustic events with longer patterns.

Thus, we propose an architecture that trains on multiple temporal resolutions to harness relationships between spectral sound characteristics of an acoustic scene, and its patterns of acoustic events. This would facilitate to learn more precise representation on a high temporal scale to discriminate timbral differences such as diesel engines from trucks and petrol based engines from private cars. On the other hand, low level temporal resolutions with ranges from several seconds can optimize on different patterns of acoustic events such a speech, steps or passing cars. Finally, the representations of the different temporal resolutions, learned by the parallel CNN stacks, are combined to form an input for a fully connected layer which learns the relationships between them to predict the acoustic scenes annotated in the dataset.

In Section 2 and 3 we provide a detailed description of our method and the applied data augmentation methods. Section 4 describes the evaluation on the development dataset and results while results are presented and discussed in Section 5. Finally, Section 6 summarizes the paper and provides conclusions.

## 2. METHOD

The presented approach analyses multiple temporal resolutions simultaneously. The design of this architecture is based on the hypothesis that acoustic scenes are composed of the spectral texture or timbre of a location such as the low-frequent humming of refrigeration units in supermarkets as well as a sequence of acoustic events. These events can be unique for certain locations such as the sound of breaking waves at the beach, but usually the characteristics of a location is described by mixtures of multiple events or sounds. Spectral texture or timbre analysis requires high temporal resolutions. To distinguish the trembling fluctuations of a truck's diesel engine from a private car an analysis window of several milliseconds is required. Acoustic events, as exemplified in Figure 1,

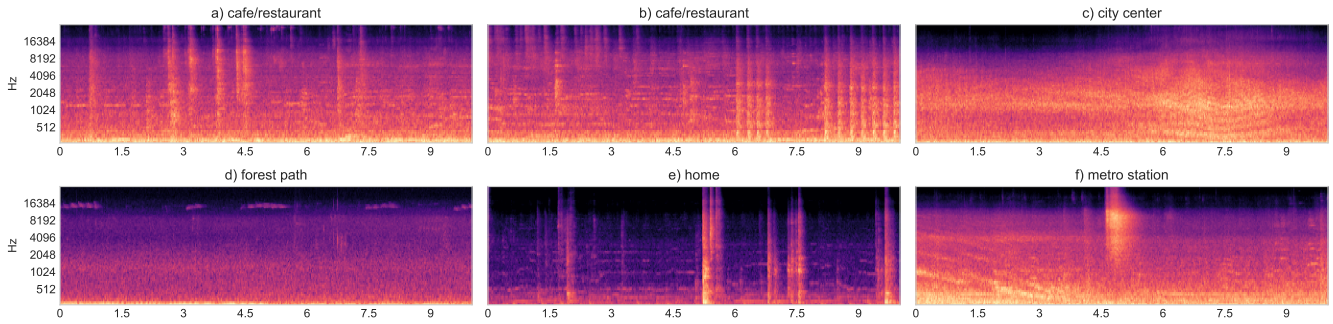



Figure 1: Example Mel-Spectrograms to visualize variances in length and shape of different acoustic events. a) dropping coins into the cash-box, b) beating coffee grounds out of the strainer, c) Doppler effect with Lloyd's mirror effect [5] of a passing car, d) chirping bird, e) opening and closing of cupboards and drawers in the kitchen, f) arriving subway with pneumatic exhaust.

happen on a much broader temporal scale. The pattern of beating the coffee grounds out of the strainer of an espresso machine in a caf (see Figure 1 b) requires an analysis window of 0.5 to 1 seconds. Up to 5 seconds are required for the very significant dropping sound of a decelerating Metro engine with the pneumatic exhaust of the breaks at full halt (see Figure 1 f).

Figure 2 visualizes different spectral resolutions at a fixed start-offset from audio content recorded in a residential area. Figure 2 a) visualizes the low-frequent urban background hum at a very high temporal resolution. At this level a CNN can learn a good timbre representation for acoustic scenes, but it is not able to recognize acoustic events that are longer than 476 milliseconds. Patterns such as speech (see Figure 2 c) or combinations of patterns such as people talking while a car is passing (see Figure 2 e) require much longer analysis windows, up to several seconds. The problem with single-resolution CNNs is, that a decision has to be made concerning the length and precision of the analysis window. A high temporal resolution prevents from recognizing long events while a low resolution is not able to effectively describe timbre. Increasing the size of the input segment to widen the analysis window would also increase the size of the model, its number of trainable parameters and the number of required training instances to avoid overfitting. If pooling-layers are extensively used to reduce the size of the model, the advantage of the high temporal resolution gets lost in these data-reduction steps.

Thus, we propose to use multiple inputs at different temporal resolutions to have separate CNN models learn acoustic scene representations at different scales which are finally combined to learn the categorical concepts of the acoustic scene classification dataset.

### 2.1. Deep Neural Network Architecture

The presented architecture consists of identical but not shared Convolutional Neural Network (CNN) stacks - one for each temporal resolution. These stacks are based on the parallel architectures initially described in [6] and further developed in [7, 8]. The fully connected output layers of each parallel CNN stack, which is considered to contain the learned representation for the corresponding temporal resolution, are combined to the multi-resolution model.

**The Parallel Architecture:** This architecture uses a parallel arrangement of CNN layers with rectangular shaped filters and Max-Pooling windows to capture spectral and temporal relationships at once [6]. The parallel CNN stacks use the same input - a log-amplitude transformed Mel-Spectrogram with 80 Mel-bands spectral and 80 STFT frames temporal resolution. The variant used in this paper (see Figure 3b) is based on the deep architecture presented in [8]. The first level of the parallel layers are similar to the original approach [7]. They use filter kernel sizes of $10\times23$ and $21\times10$ to capture frequency and temporal relationships. To retain these characteristics the sizes of the convolutional filter kernels as well as the feature maps are sub-sequentially divided in halves by the second and third layers. The filter and Max Pooling sizes of the fourth layer are slightly adapted to have the same rectangular shapes with one part being rotated by $90^\circ$. Thus, each parallel layer subsequently reduces the input shape to $2\times10$ dimensions - one layer reduces the spectral while preserving the temporal information, the other performs the same reduction on the temporal axis. The final equal dimensions of the final feature maps of the parallel model paths balances their influences on the following fully connected layer with 200 units.

**Multi-Temporal Resolutions CNN:** The proposed architecture instantiates one parallel architecture for each temporal resolution (see Figure 3b. Their fully connected output layers are concatenated. To learn the dependencies between the sequences of spectral and temporal representations of the different temporal resolutions an intermediate fully connected layer with 512 units is added before the Softmax output layer.

**Combined Max-Average Pooling Layers:** Max-Pooling is a data reduction layer which keeps the highest local value within a given surrounding. This property has shown to be advantageous for image processing tasks because it preserves salient objects and edges. In the acoustic domain sound is not predominately defined by its loudest value. A Max-Pooling layer though only keeps those peaking values which in turn mask all others. In the worst case, a single short spectral burst with its spectral energy spread over all frequencies would mask any other spectro-temporal pattern of an audio signal independent from its length. Average-Pooling on the other hand filters out peaking events which removes the attacks of a sound which are in combination with decay an important property of timbre. The proposed Max-Average-Pooling layer applies Max- and Average-Pooling on the inputs. The resulting output is stacked which corresponds to a doubling of the input feature maps (see Figure 3c).

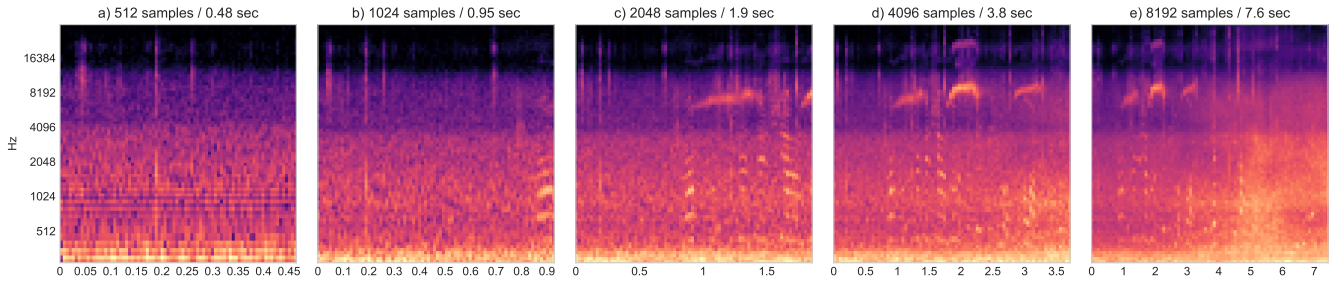



Figure 2: Input Segments for the Convolutional Neural Networks with 80 Mels spectral and five different temporal resolutions with fixed start-offset. a) spectral texture of residential area background noise, b) person saying a word (vertical wave-line), c) person talking, tweet of a bird (horizontal arc), d) person talking, bird tweeting, e) person talking, bird tweeting, car passing (light cloud to the right)

## 3. DATA AUGMENTATION

The most challenging characteristics of the provided dataset is its low variance. Table 1 depicts that for each class audio content of 3120 seconds length is provided. Nevertheless, this content originates from only 13 to 18 different locations per class. To create more data instances these recordings have been split into 10 seconds long audio files, but this does not introduce more variance due to very high self-similarity within a location. This low variance leads complex neural networks with a large number of trainable parameters to over-fit on the training data. Further, the limitation of 10 seconds per file prevents from using larger analysis windows. To circumvent these shortcomings data augmentation using the following methods is applied:

**Split-Shuffle-Remix of audio files:** To create additional audio content by increasing the length of an audio file its content is segmented by non-silent intervals. To create approximately 10 segments the Decibel-threshold is iteratively increased until the desired quantity is reached. These segments are duplicated to retrieve four identical copies which corresponds to 40 seconds of audio. All segments are then randomly reordered and remixed into a final combined audio file.

**Remixing Places:** To introduce more variance in the provided data, additional training examples are created by mixing files of the same class. Based on the assumption that classes are composed of a certain spectral texture and a set of acoustic events, mixing files of the same class would generate new recordings of this class. For each possible pairwise combination of locations within a class, a random file for each location is selected. The recordings are mixed by averaging both signals.

**Pitch-Shifting:** The pitch of the audio signal is increased or decreased within a range of 10% of its original frequency while keeping its tempo the same. The 10% range has been subjectively assessed. Larger perturbations sounded unnaturally.

**Time-Stretching:** The audio signal is speed up or slowed down randomly within a range of 10% at maximum of the original tempo while keeping its pitch unchanged.

**Noise Layers:** A data-independent augmentation method to increase the model's robustness. The input data is corrupted by adding Gaussian noise with a probability of $\sigma = 0.1$ is to the Mel-Spectrograms. The probability $\sigma$ has been empirically evaluated in preceding experiments using different single-resolutions models. From the tested values [0.05, 0.1, 0.2, 0.3] a $\sigma$ of 0.1 improved the model's accuracies most.

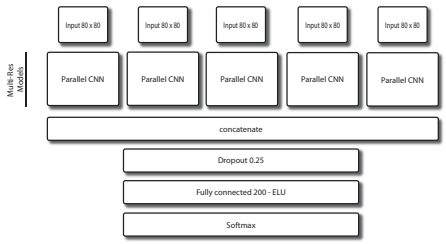
## 4. EVALUATION

The presented approach was evaluated on the *development dataset* of the *TUT Acoustic Scenes 2017 dataset* [1]. The dataset consists of 15 classes representing typical urban and rural acoustic scenes (see Table 1). 4-fold cross-validation was applied using grouped stratification which preserved the class distribution of the original ground-truth assignment in the train/test splits as well as ensured that files of the same location are not split across them. The performance was measured in classification accuracy on a per-instance-level (*raw*) for every extracted Mel-Spectrogram as well as on a per-file-level (*grouped*) by calculating the average Softmax response for all Mel-Spectrograms of a file. For each audio file 10 log-amplitude scaled Mel-Spectrograms with 80 Mels times 80 frames are extracted from the normalized input signal using random offsets and increasing FFT window sizes of 512, 1024, 2048, 4096, 8192 samples with 50% overlap. To augment the data, additional 10 random input segments were extracted for *time-stretched*, *pitch-shifted place-wise remixed* audio content. *Split-Shuffle-Remix* augmentation preceded all feature extraction processes. The neural networks were trained using Nadam optimization [9] with *categorical crossentropy loss* at $10^{-5}$ learning rate and a batch-size of 32. The learning rate was reduced by 10% if the validation loss did not improve over 3 epochs maintaining a minimum rate of $5 * 10^{-6}$.

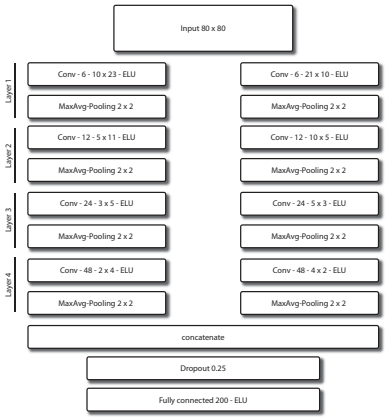
The evaluation is divided into single- and multi-resolution experiments. First, for each of the combined model's resolutions a separate *parallel CNN model*, and second, the full multi-resolution model is evaluated. Both experiments are performed using unaugmented (*raw*) and augmented input data.

### 4.1. Evaluation Dataset

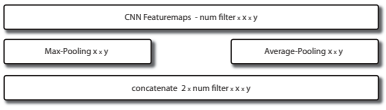
The final predictions on the evaluation dataset have been estimated using the trained models of each of the cross-validation folds on the development dataset. From each evaluation dataset audio file 10 log-scaled Mel-Spectrograms have been extracted after expanding its length to 40 seconds using the described *Split-Shuffle-Remix* augmentation method. Each data instance was predicted by each of the four optimized models. The prediction results were weighted by the fold accuracy and finally the average of all predictions for a file was calculated. The following submissions have been evaluated on the evaluation dataset:




(a) Multi-Resolution Model




(b) Parallel CNN Architecture




(c) Combined Max- and Average-Pooling Layer

Figure 3: The *Multi-Resolution Model* (a) which consists of one *Parallel CNN Architecture* (b) per temporal resolution.

**Schindler_AIT_task1_1:** The multi-resolution architecture as described in Section 2 using Max-Pooling for data reduction.

**Schindler_AIT_task1_2:** The same multi-resolution architecture as described in Section 2 using the combined Max- and Average-Pooling layers.

## 5. RESULTS AND DISCUSSION ON THE DEVELOPMENT DATASET

As shown in Table 2 the proposed multi-resolution model clearly outperforms the best performing single-resolution models by 3.56%. Although an improvement can already be observed on unaugmented (*raw*) data, the high complexity of the model especially gains from the added variance of augmented data. An improvement of 6.81% was reached using the combined Max- and Average pooling layers. This represents an improvement of 12.49% and 15.74% over the DCASE 2017 Acoustic Scenes Classification task baseline [1]. Grouping and averaging the predictions for a file of all single-resolution models (see *'grouped single'* in Table 2) does not increase the performance of these models, nor is it comparable to the multi-resolution model. It was further observed that lower

Table 1: Per class dataset Overview. Number of different locations, complete length as well as min/max/mean length of audio content.

| Label | num diff locations | Audio length (in seconds) sum | min | max | mean |
|---|---|---|---|---|---|
| beach | 17 | 3120 | 120 | 210 | 183.5 |
| bus | 18 | 3120 | 60 | 300 | 173.3 |
| cafe/restaurant | 16 | 3120 | 120 | 300 | 195.0 |
| car | 17 | 3120 | 90 | 270 | 183.5 |
| city_center | 15 | 3120 | 150 | 270 | 208.0 |
| forest_path | 18 | 3120 | 60 | 300 | 173.3 |
| grocery_store | 17 | 3120 | 120 | 270 | 183.5 |
| home | 16 | 3120 | 90 | 300 | 195.0 |
| library | 16 | 3120 | 150 | 240 | 195.0 |
| metro_station | 17 | 3120 | 90 | 300 | 183.5 |
| office | 13 | 3120 | 150 | 300 | 240.0 |
| park | 17 | 3120 | 120 | 210 | 183.5 |
| residential_area | 17 | 3120 | 120 | 240 | 183.5 |
| train | 17 | 3120 | 90 | 270 | 183.5 |
| tram | 17 | 3120 | 60 | 300 | 183.5 |

Table 2: Experimental results (classification accuracy with standard deviation over cross-validation folds). Single-resolution model results provided on top, multi-resolution with Max-Pooling (MP) and Max-Average-Pooling (MAP) models at the bottom.

| fft win size | instance raw | grouped raw | instance augmented | grouped augmented |
|---|---|---|---|---|
| 512 | 64.14 (2.84) | 70.32 (2.96) | 69.06 (4.33) | 76.63 (4.44) |
| 1024 | 66.32 (2.58) | 71.27 (3.06) | 71.70 (5.46) | 77.06 (5.46) |
| 2048 | 66.83 (1.52) | 70.23 (1.99) | 76.24 (2.53) | 80.46 (3.30) |
| 4096 | 69.50 (2.83) | 71.92 (3.23) | 79.20 (3.03) | 81.66 (3.29) |
| 8192 | 69.66 (2.58) | 71.47 (2.95) | 82.26 (2.40) | 83.73 (2.63) |
| grouped single | | 73.12 | | 83.19 |
| multi-res MP | 72.23 (4.15) | 74.30 (4.81) | 85.22 (2.11) | 87.29 (2.02) |
| multi-res MAP | | | 88.82 (2.92) | 90.54 (3.33) |

temporal resolutions perform better than higher. This could indicate that the higher contrast of peaking spikes in the spectrograms makes it easier for algorithms to learn better and more discriminative representations than from the noise-like pattern of higher temporal resolutions. As already reported in preceding studies [7, 10, 8] the grouped accuracy outperforms instance based (*raw*) prediction. Averaging over multiple predicted segments of a test file balances outliers in the classification results. The custom dropout which dropped the output of two random resolution CNN stacks showed little effect on the general performance of a model. Conventional Dropout with a probability of $\sigma = 0.25$ seemed sufficient.

In the DCASE 2017 challenge on task 1, the multi-res model Schindler_AIT_task1_1 and Schindler_AIT_task1_1 (using max-average pooling) both scored 61.7 % accuracy on the evaluation set.

## 6. CONCLUSIONS AND FUTURE WORK

The presented study introduced a Convolutional Neural Network (CNN) architecture which harnesses multiple temporal resolutions to learn dependencies between timbral properties of an acoustic scene as well as its temporal pattern of acoustic events. The experimental results showed that the proposed multi-resolution model outperforms the all single-resolution and combined models by at least 6.81%. Future work woul concentrate on improved data augmentation models, including evaluations on which augmentation methods have an improving/degrading effect on the classes (e.g. grocery store) and which methods can be applied to make the lower per-

forming classes more discriminative.

## 7. REFERENCES


[1] 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.

[2] Y. LeCun, Y. Bengio, *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[4] B. Fazekas, A. Schindler, T. Lidy, and A. Rauber, "A multimodal deep neural network approach to bird-song identification," *Working Notes of CLEF*, vol. 2017, 2017.

[5] K. W. Lo, S. W. Perry, and B. G. Ferguson, "Aircraft flight parameter estimation using acoustical lloyd's mirror effect," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 38, no. 1, pp. 137–151, 2002.

[6] J. Pons, T. Lidy, and X. Serra, "Experimenting with musically motivated convolutional neural networks," in *Proceedings of the 14th International Workshop on Content-based Multimedia Indexing (CBMI 2016)*, Bucharest, Romania, June 2016.

[7] T. Lidy and A. Schindler, "CQT-based convolutional neural networks for audio scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 60–64.

[8] A. Schindler, T. Lidy, and A. Rauber, "Comparing shallow versus deep neural network architectures for automatic music genre classification," in *9th Forum Media Technology (FMT2016)*, vol. 1734. CEUR, 2016, pp. 17–21.

[9] T. Dozat, "Incorporating nesterov momentum into adam," 2016.

[10] T. Lidy and A. Schindler, "Parallel convolutional neural networks for music genre and mood classification," Music Information Retrieval Evaluation eXchange (MIREX 2016), Tech. Rep., August 2016.