ACOUSTIC SCENE CLASSIFICATION: FROM A HYBRID CLASSIFIER TO DEEP LEARNING

Anastasios Vafeiadis[1], Dimitrios Kalatzis[1], Konstantinos Votis[1], Dimitrios Giakoumis[1], Dimitrios Tzovaras[1], Liming Chen[2] and Raouf Hamzaoui[2]

[1]Information Technologies Institute, Center for Research & Technology Hellas, Thessaloniki, Greece
[2] Faculty of Technology, De Montfort University, Leicester, UK

Abstract

We investigated two approaches for the acoustic scene classification task. Firstly, we used a combination of features in the time and frequency domain and a hybrid Hidden Markov Model-Support Vector Machines (HMM-SVM) classifier to achieve an average accuracy over 4-folds of 80.9% on the development dataset and 61.0% on the evaluation dataset. Secondly, by exploiting data augmentation techniques and using the whole segment (as opposed to splitting into sub-sequences) mel-spectrogram as an input, the accuracy of our Convolutional Neural Network (CNN) system was boosted to 95.9%. However, due to the small number of kernels used for the CNN and a failure of capturing the global information of the audio signals, it achieved an accuracy of 49.5% on the evaluation dataset. Our two approaches outperformed the DCASE baseline method, which uses log-mel band energies for feature extraction and a Multi-Layer Perceptron (MLP) to achieve an average accuracy over 4-folds of 74.8%.

Acoustic Scene Classification Framework

An Acoustic Scene Classification (ASC) framework includes the process of audio signal acquisition, feature extraction and classification (Fig.1). The detection module first segments the sound events from the continuous audio signal. Then features are extracted to characterize the acoustic information. Finally, classification matches the unknown features with an acoustic model, learnt during a training phase, to output a label for the segmented sound event.

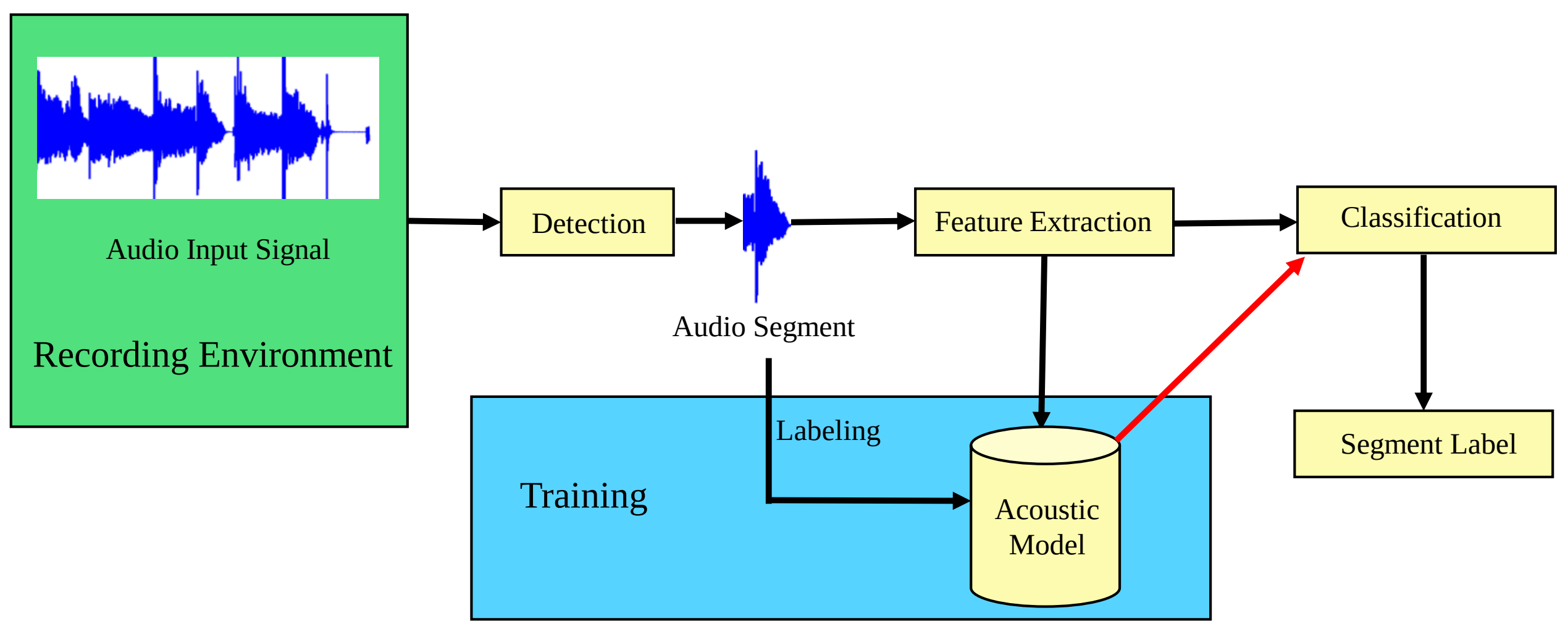

Figure 1: A typical Acoustic Scene Classification framework

HMM-SVM Approach

All audio files are transformed into the frequency domain through a 2048-sample Short-Time Fourier Transform (STFT) with 50% overlap. Each frame has a window size of 40 ms with a 20 ms hop size from the next one. In our HMM-SVM approach, we firstly converted the 24-bit depth stereo audio recordings to mono, then we divided the spectrum into 40 mel-spaced bands, and the following features are extracted for each band: *Spectral Rolloff (SR), Spectral Centroid (SC), Mel-Frequency Cepstral Coefficients (MFCC) (static, first and second order derivatives) and Zero-Crossing Rate (ZCR).*

- 39 MFCCs (static + first order derivative + second order derivative)
- Average ZCR
- SC & SR

We have aggregated all the features by taking the mean, variance and skewness. Finally, we applied Sequential Backward Selection (SBS) to select the most important features for each recording class. The HMM-SVM model is shown in Fig.2. We used 3 hidden states for the HMM (beginning, middle, end of a recording) and for the SVM part we used the Radial-Basis Function (RBF) kernel and after performing grid search, we found that the best parameters were $\sigma = 0.1$ and $C = 100$.

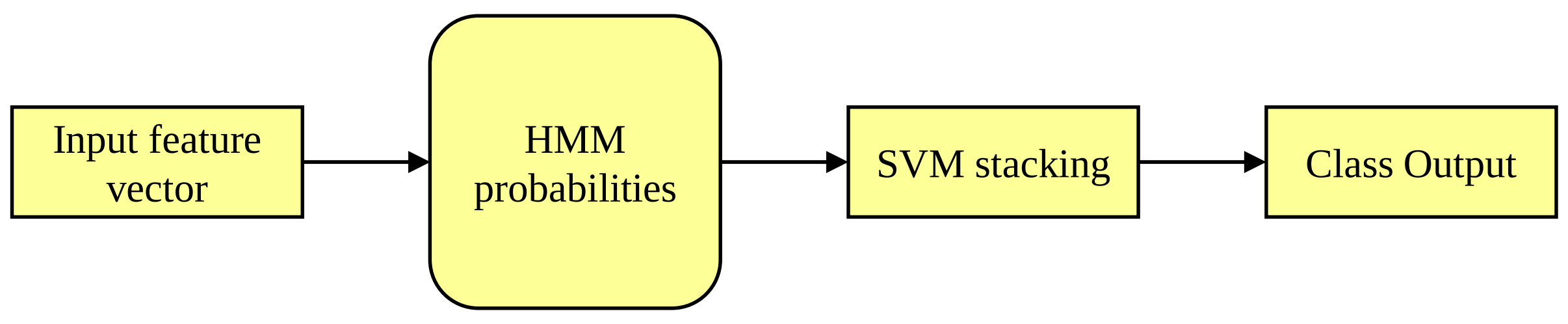

Figure 2: HMM-SVM architecture

CNN Approach

Environmental audio recordings have different temporal properties. In our CNN approach we produced two additional recordings from the original ones, since there was a strong need for more training data to be applied to the deep learning model, as following:

- Gaussian noise over the 10 seconds; average time domain value of zero
- Resampled the original signal from 44.1kHz to 16kHz

Hence the total training audio files of each fold were increased from 3510 to 10530 and the testing from 1170 to 3510.

We used the mel-spectrogram with 128 bins which is a sufficient size to keep spectral characteristics while greatly reduces the feature dimension. Each frame has a window size of 40 ms with a 20 ms hop size from the next one. Our network architecture consists of 4 convolutional layers (Fig.3). In detail, the first layer performs convolutions over the spectrogram of the input segment, using 3x3 kernels.

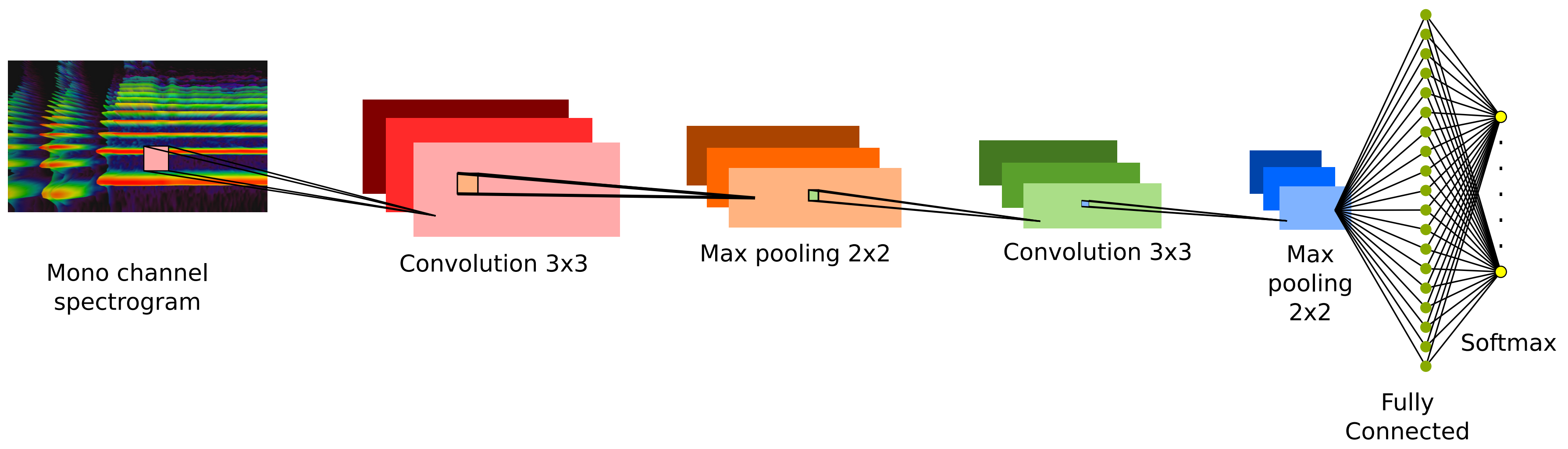

Figure 3: CNN architecture

Results

Tables 1 and 2 summarize the performance achieved by our methods and compare their accuracy with the baselines.

Table 1: Performance comparison (averaged over 4-folds) between the DCASE2017 baseline based on GMM and our hybrid SVMHMM approach

Class	Baseline GMM w/ MFCC feautres (%)	Our approach SVM-HMM w/ MFCC, ZCR, SR. SC features (%) (development dataset)	Our approach SVM-HMM w/ MFCC, ZCR, SR. SC features (%) (evaluation dataset)
Beach	75.0	78.8	23.1
Bus	84.3	90.1	42.6
Cafe/Restaurant	81.7	68.3	58.3
Car	91.0	94.2	66.7
City center	91.0	91.3	77.8
Forest path	73.4	85.6	86.1
Grocery store	67.9	80.8	64.8
Home	71.4	74.5	94.4
Library	63.5	65.7	39.8
Metro station	81.4	89.1	92.6
Office	97.1	99.0	54.6
Park	39.1	59.0	20.4
Residential area	74.7	79.8	72.2
Train	41.0	63.8	81.5
Tram	79.2	85.6	39.8
Average	**74.1**	**80.9**	**61.0**

Table 2: Comparison of recognition accuracy between the proposed system and a baseline system based on Log-mel band energies and MLP for the DCASE 2017 dataset averaged over 4-folds

Class	Baseline Log-mel band energies MLP (%)	Our System (with data augmentation) Log-mel spectrogram CNN (%) (development dataset)	Our System (with data augmentation) Log-mel spectrogram CNN (%) (evaluation dataset)
Beach	75.3	97.8	35.2
Bus	71.8	92.3	23.1
Cafe/Restaurant	57.7	96.2	58.3
Car	97.1	97.4	63.0
City center	90.7	99.6	90.7
Forest path	79.5	100.0	90.7
Grocery store	58.7	99.6	57.4
Home	68.6	98.3	61.1
Library	57.1	95.3	20.4
Metro station	91.7	92.3	38.0
Office	99.7	100.0	53.7
Park	70.2	90.6	25.9
Residential area	64.1	90.2	45.4
Train	58.0	93.2	59.3
Tram	81.7	97.0	48.1
Average	**74.8**	**95.9**	**49.5**

Both of our systems severely underperformed on the evaluation dataset. We attribute this to a combination of inadequate feature extraction and model capacity. While our extracted features were adequate enough to encode information present in the development set (and thus lead to good development held out performance) they seem to have captured mostly local information, or at least failed to encapsulate the global structure hidden in the data. The performance of the system could significantly be improved, using the stereo and binaural recordings. Finally, the relatively small capacity of our model (only 5 convolutional kernels) played a significant role in the worsening of the model's performance in the evaluation set.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 676157.

varlab.iti.gr

anasvaf@iti.gr