ALL-CONV NET FOR FRAME LEVEL MUSIC AND SPEECH SEGMENTATION

Rhythm Bhatia
University of Eastern Finland
rbhatia@student.uef.fi

Anshul Thakur
Indian Institute of Technology, Mandi
anshul_thakur@students.iitmandi.ac.in

EXTENDED ABSTRACT

An all-convolutional neural network (all-conv net) [3] architecture designed for frame level music/speech classification with input as raw audio and processed using Kapre layers [2]. Proposed method is focused on music/speech segmentation using mirex 2018 dataset [1]. The proposed network comprises of Kapre layers followed by eight pairs of convolution and learned pooling layers, two (1,1) convolution layers. The input to network is raw audio files which is then processed to 40×1 normalized melspectrogram by Kapre layers. In convolutional layers, we have used kernel size of 5×5 with a stride of 1×1 and only 16 filters. For pooling, convolution with kernel size of 5×1 and 2×1 with a stride of 5×1 and 2×1. In the end of the network, two 1×1 convolutional layers with 196 and 2 filters are used. The output of these layers is a two-dimensional feature map, converted into a probabilistic score vector using soft-max function. This vector signifies the probability of the presence of music or speech in a specified collection of frames (here we have considered 10 frames at a time) of any raw input of audio. Fig. 1. is code that we used to compute melspectrogram and normalize it per frequency.

Convolutional fully connected layers

At the end of network, fully connected layers are implemented by using 1×1 convolutions with number of filters in each of them equivalent to the number of neurons in a fully connected layer. First layer consists of 196 filters. We reduce our feature map to 2 dimensions to get the probability for the presence of speech/music.

Activations

ReLU (Rectified Linear Unit) activation has been applied over all the convolutional and learned pooling layers. For the fully connected layer of 196 filters we have applied sigmoid activation. Softmax is applied over the final 2-dimensional output of the network.

Performance Analysis

Performance measure as listed in table 1 is F-score.

Discriminating nature of the learned features

First 10 layers of the proposed network learn discriminatory features, while last 2 dense layers perform classification. We analyzed discriminating ability of 16 dimension feature vector generated by the model at 10th layer. Fig. 2. demonstrates feature representation of 2 different audio files. First file contains music whereas second file consists of speech only. Fig. 2 clearly demonstrates that filters i.e. 2 to 7, 10 to 12 and 15 to 16 are high for speech only. On the other hand, filters 0 to 3, 6 to 8, 8 to 10 and 12 to 16 are high for music

```
model = Sequential()
model.add(Melspectrogram(sr=sr, n_mels=40,
n_dft=1024, n_hop=882*2, input_shape=input_shape,
return_decibel_melgram=True,trainable_fb=False,
trainable_kernel=False, name='melgram'))
model.add(Normalization2D(str_axis='freq'))
```

Figure 1. Kapre Code Snippet

Name	f-score
GMM(+HMM)	89.9
RF	90.7
SVM I	90.7
Caffenet S,I,A	96.1
CNN	95.5
All Conv	97.8

Table 1. Performance Comparison

only. Similar, behaviour was observed for other recordings as well. We propose to further use reinforcement

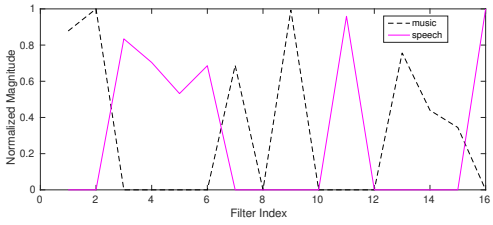

Figure 2. 16-dimensional feature vectors obtained before the dense layers

learning along with the network. With the help of reinforcement, we can optimise our network to perform music and speech segmentation in real-time.

ACKNOWLEDGMENTS

I would like to thank Gabriel Meseguer Brocal (IRCAM, France) who is my mentor at WiMIR for his constant support and to WiMIR society for the scholarship and ISMIR society for organizing this conference.I am immensely thankful to the University of Eastern Finland, IIT Mandi for the knowledge and guidance.

REFERENCES

[1] Mirex 2018 music and/or speech detection challenge. `https://www.music-ir.org/mirex/wiki/2018:Main_Page`. Accessed: 2019-09-15.

[2] Keunwoo Choi, Deokjin Joo, and Juho Kim. Kapre: On-gpu audio preprocessing layers for a quick implementation of deep neural network models with keras. *arXiv preprint arXiv:1706.05781*, 2017.

[3] Arjun Pankajakshan, A Thakttr, Daksh Thapar, Padmanabhan Rajan, Aditya Nigam, et al. All-cony net for bird activity detection: Significance of learned pooling. In *19TH ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION*, 2018.