MUSIC AND/OR SPEECH DETECTION MIREX 2018 SUBMISSION

Blai Meléndez-Catalán
Music Technology Group,
Universitat Pompeu Fabra
BMAT Licensing S.L.
bmelendez@bmat.com

Emilio Molina
BMAT Licensing S.L.
emolina@bmat.com

Emilia Gómez
Music Technology Group,
Universitat Pompeu Fabra
emilia.gomez@upf.edu

ABSTRACT

In this extended abstract, we briefly describe our submission to the MIREX 2018 *Music Detection* and the *Music Relative Loudness Estimation* tasks. We present the same algorithm to both tasks. It is based on a convolutional neural network (CNN) that estimates the proportion of the loudness that corresponds to the music content at each frame. By thresholding the loudness estimation, we perform the final classification into the classes of each task. The algorithm is trained using a dataset of about 30 hours of broadcast audio annotated using BAT [3], which allows for the annotation of relative loudness.

1. INTRODUCTION

Music detection refers to the task of finding music events in an audio recording [1]. According to the literature, the two main applications of music detection algorithms are: automatically indexing and retrieving auditory information based on its audio content, and the monitoring of music for copyright management [1, 2, 4, 5].

In the current copyright management business paradigm, whether the music is used in the foreground or background, i.e., its loudness in relation to other simultaneous content, is a relevant factor [2]. In this case, the music detection task falls short and we need to create algorithms that are able to estimate the loudness of the music with respect to other simultaneous sounds.

In section 2, we present the feature extraction process that transform the audio into the model's input. Then, in section 3, we describe the dataset that we have used to train the model. Finally, in section 4, we provide a general explanation of the architecture and the most important parameters of the model.

2. FEATURE EXTRACTION

We extract the features that we use as input to the model from a logarithmic mel-spectrogram, computed from audio at 8000 Hz and 16 bits per sample, with 128 frequency values for each frame. Frames and hop sizes are 512 and 128 samples, respectively. To generate the input we cut the logarithmic mel-spectrogram in blocks of 128 frames. The resulting input has a size of 128x128 and covers approximately 2 seconds of audio. Before entering the network, we normalize the input using min-max normalization.

3. TRAINING DATASET

The audio data used for the training contains 2-minutes excerpts of broadcast audio from TV and radio channels of several countries that amount to a total of approximately 30 hours. The dataset was annotated using BAT [3], which makes it possible to annotate the relative loudness of simultaneous events. With this information, we generate a ground truth that consists in an array of length 2 that contains the proportion of loudness corresponding to the musical content and the non-musical content, respectively, contained in each network's input. For training, 80% of the data goes to the training split, while the remaining part is split equally between the development and the test sets.

4. PROPOSED MODEL

The model consists in a standard CNN: several convolutional layers, each of them followed by a max-pooling layers, that lead to a set of dense layers. Each convolutional layer is followed by a ReLU activation function. The output layer has 2 neurons and a softmax activation function. We train this model for 100 epochs with mean squared error as the loss function and using the ADAM optimizer. No regularization method is applied. We save the model that produces the lowest loss for the development set.

For the task of Music Detection, we use one threshold to transform the regression output of the network into a 2 classes classification output: *Music* and *No Music*. For the Music Relative Loudness Estimation task, we add a second threshold that separates between *Foreground Music* and *Background Music*. Finally, we apply a set of rules related to the temporal context of each input to smooth the network's output.

[1] http://www.music-ir.org/mirex/wiki/2015:Music/Speech_Classification_and_Detection#Music.2FSpeech_Detection

[2] https://createurs-editeurs.sacem.fr/brochures-documents/regles-de-repartition-2017

5. REFERENCES

[1] Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. Music tracking in audio streams from movies. In *Proceedings of the IEEE 10th Workshop on Multimedia Signal Processing (MMSP)*, pages 950–955, 2008.

[2] Tomonori Izumitani, Ryo Mukai, and Kunio Kashino. A background music detection method based on robust feature extraction. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13–16, 2008.

[3] Blai Meléndez-Catalán, Emilio Molina, and Emilia Gómez. BAT: An open-source, web-based audio events annotation tool. In *3rd Web Audio Conference*, 2017.

[4] Klaus Seyerlehner, Tim Pohle, Markus Schedl, and Gerhard Widmer. Automatic music detection in television productions. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-07)*, 2007.

[5] Yongwei Zhu, Qibin Sun, and Susanto Rahardja. Detecting musical sounds in broadcast audio based on pitch tuning analysis. In *IEEE International Conference on Multimedia and Expo*, pages 13–16, 2006.