INA'S MIREX 2018 MUSIC AND SPEECH DETECTION SYSTEM

David Doukhan **Eliott Lechapt** **Marc Evrard** **Jean Carrive**
French National Institute of Audiovisual (Ina), Paris, France
{ddoukhan,elechapt,mevrard,jcarrive}@ina.fr

ABSTRACT

A convolutional neural network (CNN) based architecture is proposed for MIREX 2018 music and speech detection challenge. The system uses log-mel filterbank features. It has 4 convolutional and 4 dense layers. It is part of the inaSpeechSegmenter open-source framework, which was designed for conducting gender equality studies.

1. INTRODUCTION

This paper presents the system submitted to Mirex 2018 Speech and/or Music detection task. This system is based on the `inaSpeechSegmenter` open-source framework (MIT license) [4]. The full framework is available on GitHub [1] and is packaged as a python3 pip module [2] .

It was designed for conducting digital humanities studies describing men and women speaking-time ratios across TV and radio channels. These large-scale descriptions were used as an estimate of gender equality in medias [3, 5].

The framework is composed of two segmentation modules. The first module, which was used for the speech and/or music detection task, is in charge of segmenting audio stream into speech and music. The system was designed for segmentation rather than detection. With respect to the aim of this framework (estimating men and women speech-time ratio), speech-over-music is labeled as speech. Therefore, this module is better suited for speech than for music detection. The second module of the framework (not evaluated in this challenge) is in charge of splitting and labeling the resulting speech segments according to their corresponding gender class.

2. ALGORITHM

The processing pipeline is composed of 4 main steps described in the following subsections.

2.1 Signal Activity detection

A baseline activity detection system – based on adaptive energetic threshold – is used. The threshold is defined as the $\log(3\%)$ of the mean log energy found in a given sound file. Filtering procedures are then used to obtain the segments showing activity from these frame-level activity estimates.

2.2 Feature extraction

The Mel-scaled filter-banks representation of the signal is computed on 25ms sliding windows with 10 ms shift using SIDEKIT [7]. 21 Mel filterbanks sampled between 100 and 4000 Hz are extracted. This rather low maximum frequency limit was chosen to handle low-quality signals that may occur in TV and radio streams: i.e., telephone-quality signal with no energy above 4000 Hz.

Resulting features were grouped into patches concatenating 68 adjacent windows. Patches were normalized to have 0 mean and unit variance, in order to increase the robustness to volume variations between recordings.

2.3 CNN frame-level detection

Figure 1 shows the CNN architecture used for speech-music discrimination, which was implemented using Keras [2]. The CNN input is composed of patches of dimension 68×21 (time × feature dimension), accounting for an input analysis window of 695 ms. The model has 4 convolutional and 4 dense layers. All these layers are followed by a batch-normalization stage and ReLU activation layers. The dense layers are also followed by dropout layers, with dropout rates increasing according to the network depth. The first convolutional layer is associated with the biggest width (5), in order to capture the horizontal patterns typically found in music signals. The last pooling layer operates on the temporal dimension, in order to focus on the pattern showing the biggest activation in this relatively large time-interval. The output is implemented using a softmax activation layer, allowing to obtain a probability estimated for each supported class (i.e., speech, music).

2.4 Viterbi decoding

The CNN output is composed of instantaneous frame-level probabilities for speech and music. These probabilities are used to feed a 2-states Hidden Markov Model, aimed at inferring the most likely sequence of hidden states from these frame-level estimates. State transition probabilities were defined empirically through a grid-search procedure performed on development datasets (section 4).

[1] `https://github.com/ina-foss/inaSpeechSegmenter`

[2] `https://pypi.org/project/inaSpeechSegmenter/`

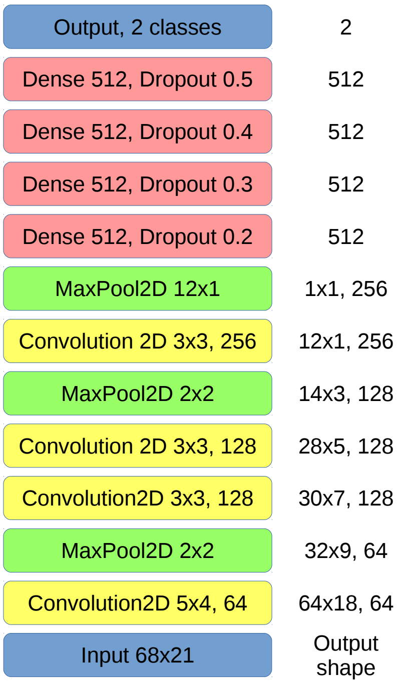

Figure 1. CNN speech-music classification architecture.

3. TRAINING DATASETS

Several datasets aimed for classification were used to train the proposed convolutional neural network model. Three datasets dedicated to speech versus music classification were used: GTZAN [11], Scheirer-Slaney [9] and MUSAN [10]. Additional music data was obtained from the GTZAN musical genre corpus [11]. 72 excerpts corresponding to *a cappella* singing, which was identified as a difficult musical genre, were obtained using the free music archive API [3] . Lastly, we used Ina's speaker dictionary to increase speech data with 2300 distinct speakers [8, 12].

Data labelled as speech in GTZAN, Scheirer-Slaney and Ina's speaker dictionary was possibly mixed with music, while data labelled as music does not contain speech. Therefore, the corresponding classification system is able to discriminate speech, including speech over music, from music alone.

4. DEVELOPMENT DATASETS

The models most suited to frame-level classification were not necessarily suited to the segmentation and the detection tasks. These were evaluated and tuned to optimize frame-level f-measure on 2 distinct detection datasets.

We used the REPERE challenge dataset, which was designed for speech transcription and speaker identification tasks [6]. This dataset does not distinguish between speech and speech-over-music.

Lastly, we used the Muspeak sample data provided for the MIREX-2015 speech/music detection challenge [1].

[3] `http://freemusicarchive.org`

5. ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780069 (MeMad project).

6. REFERENCES

[1] Mirex 2015 music/speech classification and detection and challenge. Visited on 2017-03-07.

[2] François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

[3] David Doukhan and Jean Carrive. Description automatique du taux d'expression des femmes dans les flux télévisuels français. In *XXXIIe Journées d'Études sur la Parole*, pages 496–504, 2018.

[4] David Doukhan, Jean Carrive, Félicien Vallet, Anthony Larcher, and Sylvain Meignier. An open-source speaker gender detection framework for monitoring gender equality. *Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[5] David Doukhan, Géraldine Poels, and Jean Carrive. Describing gender equality in french audiovisual streams with a deep learning approach (accepted). *Journal of European Television History and Culture (VIEW)*, 2018.

[6] Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard. The repere corpus: a multimodal corpus for person recognition. In *LREC*, pages 1102–1107, 2012.

[7] Anthony Larcher, Kong Aik Lee, and Sylvain Meignier. An extensible speaker identification sidekit in python. In *Acoustics, Speech and Signal Processing (ICASSP)*, pages 5095–5099. IEEE, 2016.

[8] François Salmon and Félicien Vallet. An effortless way to create large-scale datasets for famous speakers. In *LREC*, pages 348–352, 2014.

[9] Eric Scheirer and Malcolm Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97.*, volume 2, pages 1331–1334, 1997.

[10] David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

[11] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.

[12] Félicien Vallet, Jim Uro, Jérémy Andriamakaoly, Hakim Nabi, Mathieu Derval, and Jean Carrive. Speech trax: A bottom to the top approach for speaker tracking and indexing in an archiving context. In *LREC*, 2016.