Music/speech classification and detection submission for MIREX 2018

Matija Marolt
University of Ljubljana
matija.marolt@fri.uni-lj.si

ABSTRACT

We briefly describe our submissions to the MIREX music/speech classification and detection tasks. We submitted three algorithms, which differ significantly in the way they classify the audio into speech or music.

1. MM1

MM1 is our approach submitted to the MIREX 2015 music/speech detection task [1] and is based on our ISMIR 2009 algorithm for segmentation of field recordings [2]. We refer the reader to these publications for details.

2. MM2

MM2 uses our classification approach for segmentation of field recordings based on deep residual networks, presented at the 2018 workshop for Folk Music Analysis [3].

The audio is transformed into time-frequency space with a FFT (46 ms window size), followed by transformation into 64 channel log-scaled mel-scale spectrograms (50-8000 Hz). 2 second long feature blocks are used as neural network inputs.

The deep residual network (see the referenced paper for details on its architecture) classifies such feature blocks into four classes: speech, solo singing, choir singing (more than 1 voice) and instrumental. It outputs a probability distribution over the four classes for each input block and thus cannot handle overlapping categories.

The network was trained on a database of short excerpts gathered from a variety of recordings from ethnomusicological (and related) archives that put their collections online in recent years. The sources include: the British Library world & traditional music collection, Alan Lomax recordings, sound archives of the CRNS and a number of recordings from the Slovenian sound archive Ethnomuse and the Norwegian national library, which are not available online, but were made available to us by ethnomusicologists with the respective institutions. These field recordings were augmented by the well-known GTZAN music/speech collection and the Mirex 2015 music/speech detection public dataset. Altogether 7,000 5 second long excerpts were extracted from these sources, manually labelled and used for training/testing the network.

Segmentation is based on the output class labels and is very simple: class labels are first attributed to speech (separate category) or music (solo+choir+instrumental), then median filtered (2 second window). Regions with detected silence are set to the “no class” label, then speech or music regions, which are separated by less than 3 seconds, are merged, and in the end regions, which are shorter than 3 seconds, removed.

3. MM3

MM3 uses a music/speech classification based on deep residual networks.

The audio is transformed into time-frequency space with a FFT (46 ms window size), followed by transformation into 64 channel log-scaled mel-scale spectrograms (50-8000 Hz). 2 second long feature blocks are used as neural network inputs.

The deep residual network (its architecture is similar as in MM2) classifies such feature blocks two classes: speech and music. A sigmoid output layer is used and thus overlapping classes can be detected.

The network was trained on a large set of speech and music examples from the AudioSet [4] dataset, augmented with samples from field recordings (see MM2), as well as a collections of background noises and sound effects.

Segmentation is based on the output class labels and is very simple: class labels are median filtered (2 second window), regions with detected silence are set to the “no class” label, then speech or music regions, which are separated by less than 3 seconds, are merged, and in the end regions, which are shorter than 3 seconds, removed.

REFERENCES

[1] M. Marolt, "Music/speech classification and detection submission for MIREX 2015," *MIREX 2015,* Available: http://www.music-ir.org/mirex/abstracts/2015/MM1.pdf

[2] M. Marolt, "Probabilistic Segmentation and Labeling of Ethnomusicological Field Recordings," in *ISMIR,*

10th International Society for Music Information Retrieval Conference, Kobe, Japan, 2009, pp. 75-80, 2009.
[3] M. Marolt, "Going Deep with Segmentation of Field Recordings," presented at the *8th International Workshop on Folk Music Analysis*, Thessaloniki, Greece, 2018.
[4] J. F. Gemmeke *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," New Orleans, LA, 2017.