SPEECH MUSIC DETECTION AND CLASSIFICATION

Reinhard Sonnleitner
Department of Computational Perception, Johannes Kepler University, Linz, Austria

ABSTRACT

This document describes the speech/music classification (discrimination) and detection method of my submissions. For both tasks, classification and detection, the same set of features is computed from the audio files and used to train random forest classifiers. The detection task uses a two class model, while the classification task computes a binary model. To make results comparable, the submissions do no not use pre-trained models; The classifiers are exclusively trained on the supplied data of either task.

1. INTRODUCTION

The music speech classification task poses the problem of discriminating between music and speech. For a number of short audio snippets, the method has to decide on whether the audio represents speech, or music. In this task, a given audio snippet never consists of both classes, but can include silence and noise.

The music speech detection task consists of detecting segments in the audio, which contain speech, music, both or none of these. This submission treats that as a two class problem, where one class represents non-music and music, and the other represents non-speech and speech.

To begin with, all audio files are converted to monaural signals sampled at 22.05kHz, and represented as STFT spectrograms using a window size of 4096 samples and a hop size of 512 samples. The spectrogram is then scaled using an approximate cent-filterbank starting at 150Hz and a hop of 50 cents. Frequencies lower than 150Hz are discarded. The resulting spectrogram represents frequencies from $\approx$ 150Hz to $\approx$ 11025Hz, compressed into 149 frequency bins, and is the basis for subsequent feature extraction computations.

Feature extraction is performed at a rate of 5Hz, i.e. the feature vectors are extracted at five evenly spaced time points per second over the duration of the audio. Each of these time points is the center of a so-called observation window of width ≈ 1.23 seconds (53 frames). These windows represent the context for feature extraction.

2. FEATURE EXTRACTION

The feature set consists of features for speech and music detection.

To detect speech, I use a version of the speech detection feature described in [1], that is trivially extended to expose the direction of the detected frequency trajectories and is parameterized as described above. Throughout this document I refer to the extended feature as "CFT".

Once the CFT is computed, the feature is also computed for a number of 27 overlapping frequency sub-bands of a bandwidth that corresponds to 10 frequency bins and an overlap of 5 frequency bins. This results in a total of 28 CFT vectors. To reduce the overall feature dimension, all sub-band CFT vectors are summarized by their variances and zero crossing rates. The CFT-part of the feature set thus consists of 53 values for the CFT vector and 2 values (var and czr) for each summarized CFT vector. This results in a feature dimensionality of $53 + 28 * 2 = 109$.

To extend the feature set for music detection, a spectrogram binarization similar to the preprocessing step described in [2] is performed. For this, the lower 100 frequency bins of the cent-scaled log-magnitude spectrogram excerpt are smoothed using a sliding uniform filter with a window size of 3 frequency bins and 11 time frames. The smoothed spectrogram is subtracted from the log-magnitude spectrogram and the result is binarized using a threshold of 0.1. The frequency bins are summed along the time axis, and normalized. This vector of frequency activation intensities (it consists of 100 values) is appended to the CFT vector, resulting in a total feature dimensionality of 209.

3. MUSIC/SPEECH CLASSIFICATION TASK

The above feature extraction process is performed for each train audio snippet, and a ground truth vector is created by replicating the supplied class label according to the feature extraction frequency (i.e. 5Hz). Once all train audio files are processed, a random forest classifier is trained that treats the two class labels in binary way: Either it is speech or non-speech, which in this task translates to music.

For predicting the class of unseen audio snippets, the feature extraction is performed and for each feature vector a class probability is predicted. The final prediction for the whole audio file is computed by comparing the median of the predicted class probabilities to a threshold of 0.5. If larger, speech is predicted, otherwise music.

4. MUSIC/SPEECH DETECTION TASK

The experiment setup is basically as described in [1]. Each audio file is subjected to the same feature extraction process as explained above, and the supplied ground truth segments for the classes speech and music are aligned to the time points of the extracted features. For this task, the random forest is trained for two classes, music and speech. The ground truth for both classes represents four possible combinations: none, music only, speech only, music and speech. Predictions for either class combination are post processed by median smoothing.

5. REFERENCES

[1] R. Sonnleitner, B. Niedermayer, G. Widmer and J. Schlüter: "A Simple And Effective Spectral Feature For Speech Detection In Mixed Audio Signals," *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx12)*, 2012.

[2] K. Seyerlehner, T. Pohle, M. Schedl, G. Widmer "Automatic music detection in television productions," *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx07)*, 2007.