Music/speech classification and detection submission for MIREX 2015

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

ABSTRACT

We briefly describe our submissions to MIREX Music/speech classification and detection tasks. The submissions are based on our method for segmentation of folk music field recordings [1], which was adapted to fit the tasks of classification and detection as defined by the MIREX challenge.

1. CLASSIFICATION

The classification approach is very straightforward. First, the recordings are converted to a single channel and normalized. Then, a number of low-level features is calculated, with a 50ms window size and 50% overlap, including MFCCs, spectral entropy, tonality and four Hertz modulation. Basic statistics (mean, mean absolute, variance, variance/mean) of these features and their deltas are calculated on 3 second long low-level feature chunks with 0.5 seconds step size to produce the final set of features.

Two submissions were posted.

MM1 trains a logistic regression classifier on the MIREX database, where 14 features are selected with forward feature selection, to predict whether a feature vector (statistics of low-level features of a 3 second long audio chunk) contains music or speech. During testing, the classifier is run on the entire piece to be classified and a weighted average of individual predictions is calculated. The weighting is based on the percentage of non-silent frames in each chunk. If the final result does not show a clear winner (difference between probabilities of both classes is under 0.2), an additional logistic regression classifier, trained on folk music field recordings [1] is run and both scores averaged to yield the decision.

MM2 does not train a classifier, but just uses a logistic regression classifier for labeling. The classifier was trained on folk music recordings [1] to predict five classes: solo singing, choir singing, bell chiming, instrumental and speech. It is based on the same set of features, their statistics and window sizes as MM1, as well as the same weighting of predictions to yield the final score.

2. DETECTION

The detection algorithm is based on the MM2 classification approach that yields music/speech classification scores for 3 second long chunks of audio, with 0.1 second step size. It does not handle overlapping speech over music segments and classifies them as either speech or music.

The initial set of segment boundaries $\mathfrak{B}$ is obtained by finding peaks above a threshold in a curve obtained by calculating symmetric KL distances between 15 second long classification score windows to the left and right of each boundary.

Additionally, silent regions are detected and regions longer than 1.5 seconds are added to the set of boundaries $\mathfrak{B}$.

Segments defined by the boundaries in $\mathfrak{B}$ are classified as speech or music by calculating the weighted average of classification scores within each segment as described previously.

The last phase of the algorithm finds all pauses (silent regions) longer than 1 second within each differently labelled segment and decides whether to place additional within-segment boundaries at these pauses. The decision is based on pause length and the KL distance between 15 second long sequences of MFCC (1-10) features (describing sound timbre) to the left and right of each pause. If the distance exceeds a threshold, an additional boundary is created - the rationale being that timbre changes at this point (either between two different music pieces or two different speakers).

REFERENCES

[1] 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.