USING CONTINIOUS WAVELET TRANSFORM AND HIDDEN MARKOV MODELS FOR AUDIO MUSIC SIMILIRATY ESTIMATION

Roman Aliyev
Belarusian State University
romanaliyev@gmail.com

ABSTRACT

In this submission an audio music similarity estimation technique is presented. The proposed method is based on two algorithms – chroma feature extraction using continuous wavelet transform and modeling music structures using hidden Markov models. During analysis stage each music track is characterized as a vector – set of probabilities of accepting each music structural model. Thus the similarity between two tracks is seen as a distance between their characterizing vectors.

1. METHOD OVERVIEW

The proposed method is illustrated in Figure 1. To predict audio similarity between music tracks each one flows through the sequence of 4 stages. During the first one time series of chroma features are extracted from the input audio waveform using continuous wavelet transform (CWT). Each feature is then quantized to the closest element from the codebook. The next stage is the estimation of the probabilities that codeword indices are matched by each music structural model. At this point, each model is represented by a hidden Markov model (HMM). The final result is a similarity rate per input pair of audio tracks.

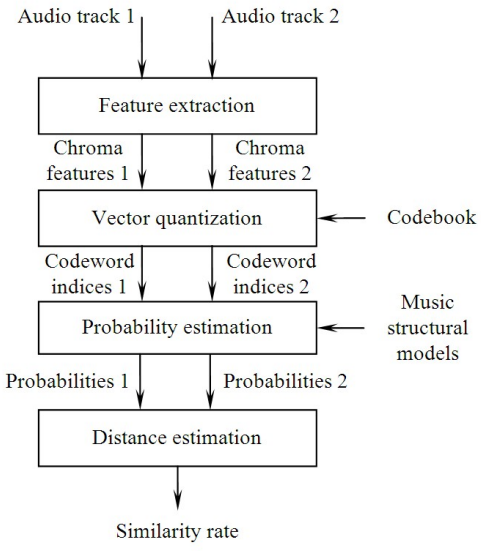

Figure 1. Diagram of the proposed method.

There are two more stages – codebook and model training – computation of the key chroma features and the set of HMMs. The details about all stages are given above.

1.1 Feature extraction

The algorithm of CWT is obtained to convert a music signal into the sequence of spectral feature vectors. It uses Morlet mother wavelet, optimized for music signals [1]. The result is a log-frequency spectrogram with semitone resolution and a 100 ms time step. Then all the bins corresponding to each separate chroma are added together in order to build 12-bin chroma features.

1.2 Codebook training

The codebook has been trained with 100 hours of music data, recorded from 10 different internet radio stations (10 hours for each station). The data included hip-hop, rock, pop, dance, chanson, blues, hard rock, jazz, classic and folk music. All chroma features of the training data have been clustered using K-means algorithm with 4096 clusters.

1.3 Vector quantization

After applying the feature extraction every chroma feature is replaced with the index of the closest element from the codebook using Manhattan distance metric.

1.4 Model training

Parameters of music structural models have been computed with 347 small-scale annotations from SALAMI project [2]. Thus, each fully connected HMM corresponds to one annotation.

1.5 Probability estimation

During this stage the forward algorithm is obtained to estimate the set of 347 probability values. Each value is the probability that the sequence of indices has been generated by particular HMM.

1.6 Similarity estimation

Finally, the set of probabilities is seen as a vector. Thus, the similarity between a pair of audio tracks can be considered as Manhattan distance between their vectors.

2. ACKNOWLEDGEMENTS

Special thanks to the authors of SALAMI project and Jordan Smith for the support in particular.

3. REFERENCES

[1] R. Aliyev: “Optimization of short-time Fourier and continuous wavelet transforms for spectral analysis of musical signals,” *Journal of Digital Signal Processing*, No. 2, pp. 16–19, 2013.

[2] Jordan B. L. Smith, J. Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie: “Design and creation of large-scale database of structural annotations,” *Proceeding of the International Society for Music Information Retrieval Conference*, pp. 555–560, 2011.