AUDIO FINGERPRINTING SYSTEM: MIREX 2015 SUBMISSIONS

Zhichao Wang
Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Science
wangzhichao@hccl.ioa.ac.cn

ABSTRACT

This submission is a music identification system using fingerprinting technology. The feature extraction is based on the Philips audio fingerprint [1] method, additionally we introduce MASK features as described in [2] to increase the robustness of the system. In the retrieval process, frame skip is used to increase the searching speed.

1. SYSTEM DESCRIPTION

The audio fingerprinting system consists of two parts: first a hash table is built as database using the training set which usually contains different style of music. After the database is built, the searching part search the fingerprints extracted from the query audio to see if there is a matched one in the database. Figure 1 shows the framework of the system.

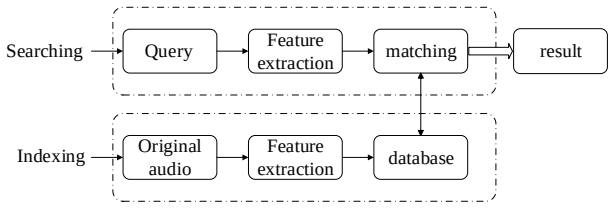

Figure 1. Audio Fingerprint System.

1.1 Feature Extraction

The Philips audio fingerprint has been used for years and the power difference feature has been proven very robust to many kinds of noise. So the submission uses the Philips fingerprint to represents audio signal. The input audio signal is firstly down sampled to 5kHz. Then a Hanning window is used to segment the signal into frames. After that Fourier Transform is performed to each frame to transform the signal from time domain to the frequency domain. This spectrum is then grouped into 33 bands that are logarithmically spaced from 300 Hz to 2000 Hz. Finally a fingerprint for each frame is a 32-bit number encoded by the power difference along the frequency axis and time axis as defined below:

$$F(n,m)=\begin{cases}1, & if\ ED(n,m)>0\\ 0, & else\end{cases} \quad (1)$$

$$ED(n,m)=E(n,m)-E(n,m+1)-(E(n-1,m)-E(n-1,m+1)) \quad (2)$$

Where F(n, m) means the m-th bit of frame n and E(n, m) represents the energy of the band m of frame n.

Instead of using fixed frame length of 0.37 seconds and frame step of 11.6 milliseconds. Our system has two configuration options "DNAFrameSampleNum" and "DNAFrameStepSampleNum" to control the frame length and the frame step more flexible. The "DNAFrameSampleNum" option controls the sample number within one frame while the "DNAFrameStepSampleNum" controls the sample points number of the step between two adjacent frames.

1.2 Mask Feature

The mask feature is computed based on the Philips fingerprint. The bits encoded from the frequency spectrum where the power differences are close to zero are most vulnerable to external noise. So after the power differences are calculated, they are ranked from 1 to 32 by the absolute value. 1 denotes the least reliable bit as it is most easily corrupted by noise while 32 means the bit is most robustness to noise. There is a configuration option "MinLevel" in the submission system which is used to set the threshold of the bits' reliability. That is, the smallest MinLevel bits are unreliable. So the mask feature is also a 32-bit number and each bit represents the reliability of the corresponding bit of the fingerprint as defined below:

$$MF(n,m)=\begin{cases}1 & if\ F(n,m)\ is\ realiable\\ 0 & else\end{cases} \quad (3)$$

where MF(n, m) is the m-th bit of the n-th frame. Then in the matching process different weight is implied to the fingerprint bits according to the mask feature.

1.3 Matching

Assuming Q and R are two fingerprint blocks that derived from a query and a reference signal respectively, and M is the mask feature extracted from the reference signal. The bit error rate (BER) between Q and R is calculated as below:

$$BER = \frac{\frac{1}{s}\sum_{n=1}^{N}\sum_{m=1}^{32}(Q(n,m) \oplus R(n,m)) \,\&\, M(n,m)}{a \times RB + b \times (32 \times N / s - RB)} \quad (4)$$

where N represents the total frames, RB is the number of reliable bits, a and b are the penalty weights given to the reliable bits and unreliable bits. a/b is greater than 1 as the reliable bits is more noise resistant and if a mismatch happens at a reliable bit, it is more likely that the reference is not match to the query, s is the frame skip step, in order to speed up the matching process, the comparison between Q and R occurs every s frames.

2. CONFIGURATION

We provide several configuration options for users to control the audio fingerprint system. The "MinLevel" controls the threshold of the mask feature as described in 1.2. The "AcceptThres" option controls the threshold of the accept BER, it is limited 0-0.4 as larger "AcceptThres" may increase the false positive rate. "TopN" controls the number of music selected from the database. "DNAFrameSampleNum" and "DNAFrameStepSampleNum" control the frame length and the frame step in the framing process. If the query segment is very short such as less than one second, then we can choose smaller "DNAFrameSampleNum" and "DNAFrameStepSampleNum" to generate enough frames.

3. REFERENCES

[1] Haitsma J, Kalker T: "A Highly Robust Audio Fingerprinting System," *ISMIR*, pp.107-115,2002.

[2] Coover, Bob, and Jinyu Han: "A Power Mask based audio fingerprint," International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.