THE ECHO NEST MUSICAL FINGERPRINT

Daniel P.W. Ellis
Columbia University LabROSA
dpwe@ee.columbia.edu

Brian Whitman, Tristan Jehan, Paul Lamere
The Echo Nest Corporation
{brian, tristan, paul}@echonest.com

ABSTRACT

We will discuss the methodology and accuracy of the Echo Nest Musical Fingerprint (ENMFP), an open-source fingerprint code generator and query server that works on music files. We define fingerprinting as matching an arbitrary audio signal to its underlying song in a 10 million song database.

1. MUSIC FINGERPRINTING

"Fingerprinting" of audio files [1] [3] is becoming a necessary feature for any large scale music understanding service or system. Online music stores want to resolve existing user catalog against their cloud storage to save bandwidth. Music indexing tools want to adjust poor metadata. Music listeners want to remove duplicates and see necessary context next to their audio. The Echo Nest Corporation recently implemented a fingerprint scheme that is based on portions of their Analyze product and will be released as an open source service for any use– with the intended goal to catalog every music file on any node of the internet.

We define fingerprinting as converting an unknown audio query Q, represented as a series of time domain samples to a match of a known song S which is represented as an ID, metadata (artist, title, etc) and list of matching tracks T. Each T is itself an audio file of any bitrate, compression type, volume or quality. Our intended goal is to match on files, not in-the-air recordings such as Shazam [3].

2. FEATURES AND CODE GENERATION

Our "code generator" is a modified version of the hosted service Analyze [2], which computes segmentation, timbre and chroma features, beat and section and other audio-derived attributes from any audio file. The code generator itself only computes the segmentation and chroma features in Analyze, which results in a $d = 12$ vector $\mathbf{C}$ for each discovered segment in Q. On average, a pop song will emit roughly 4 segments per second of audio, but the rate widely varies with the song's complexity and speed. This segment-aligned chroma vector $\mathbf{C}$ is then vector quantized

to a 10-bit code $\mathbf{c}$. (The VQ code table was trained on 10,000 random songs from a large collection.)

Successive chains of three $\mathbf{c}$ are hashed together to a 30-bit key by simply shifting $\mathbf{c_0}$ to bitspace 30-21, $\mathbf{c_1}$ to 20-11, $\mathbf{c_2}$ to 10-1. Each code is stored alongside a time offset value $\mathbf{t}$, which is the position of the start of the segment and is quantized to 32ms increments.

3. QUERIES IN PRACTICE

We treat each code as a term in a document in an inverted index. The track ID from T is stored as the document ID with each hashed 30-bit key stored as a term (with term offset stored natively as $\mathbf{t}$.) This allows fast lookup of a series of query hashes in the inverted index. The underlying data store has a lookup table keyed on each of the possible 30-bit keys (2^{30}) with list of track IDs associated to each key. A multi-code query (in practice, we expect roughly 30 seconds of audio as Q, which would be on average 120 hash keys) simply returns the track ID with the most matches of each code term. If the query server returns more than one strong match, we compute a histogram of all time offset $\mathbf{t}$ differences per matching key. We then use the total of the top two histogram buckets to inform the "true score." This allows us to ensure that the codes occur in order even if Q is from a different section of the song and thus has a different absolute time offset.

4. EVALUATION

We will describe our evaluation framework and show error rates for the fingerprint using different audio fidelity reductions.

5. REFERENCES

[1] J. Haitsma and T. Kalker. A highly robust audio fingerprinting system with an efficient search strategy. *Journal of New Music Research*, 32(2):211–221, 2003.

[2] Tristan Jehan. *Creating music by listening*. PhD thesis, Massachusetts Institute of Technology, 2005.

[3] A. Wang. An industrial strength audio search algorithm. In *International Conference on Music Information Retrieval (ISMIR)*.