Analyze Documentation

Prepared by: Tristan Jehan, CTO

July 15, 2010
Version: 2.2

The Echo Nest Corporation 48 Grove St. Suite 206, Somerville, MA 02144 (617) 628-0233 tristan@echonest.com http://the.echonest.com

Introduction

"Analyze" is a music audio analysis tool available as a free public web API (visit developer.echonest.com) and a stand-alone 60 Mb command-line binary program, on Mac, Windows, or Linux platforms for commercial partners (contact biz@echonest.com). The program takes a digital audio file from disk (e.g. mp3, m4a, wav, aif, mov, mpeg, flv), or audio data piped in on the command line. It generates a JSON-formatted text file that describes the track's structure and musical content, including rhythm, pitch, and timbre. All information is accurate to the microsecond (audio sample).

Analyze is the world's only "music listening" API. It uses proprietary machine listening techniques to simulate how people perceive music. It incorporates principles of psychoacoustics, music perception, and adaptive learning to model both the physical and cognitive processes of human listening. The output of analyze contains a complete description of all musical events, structures, and global attributes, such as key, loudness, time signature, tempo, beats, sections, harmony. It allows developers to create applications related to the way people hear and interact with music.

The output data allows developers to 1) *interpret*: understand, describe, and represent music. Applications include music similarity, playlisting, music visualizers, and analytics. 2) *synchronize*: align music with other sounds, video, text, and other media. Some example of applications include automatic soundtrack creation, and music video games. 3) *manipulate*: remix, mashup, or process music by transforming its content. An example application is the automatic beat-matching song-collage website thisismyjam.com

Output Data

- **segments**: a set of sound entities (typically under a second) each relatively uniform in timbre and harmony. Segments are characterized by their perceptual onsets and duration in seconds, loudness (dB), pitch and timbral content.
- **tatums**: a set of tatum times, in seconds. Tatums represent the lowest regular pulse train that a listener intuitively infers from the timing of perceived musical events (segments).
- **beats**: a set of beat times, in seconds. A beat is the basic time unit of a piece of music; for example, each tick of a metronome. Beats are typically multiples of tatums.
- **bars**: a set of bar times, in seconds. A bar, or measure, is a segment of time defined as a given number of beats. Bars also indicate downbeats, the first beat of the measure.
- **sections**: a set of sections times, in seconds. Sections are defined by large variations in rhythm or timbre, e.g. chorus, verse, bridge, guitar solo, etc.
- **time signature**: an estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
- **key**: the estimated overall key of a track. The key identifies the tonic triad, the chord, major or minor, which represents the final point of rest of a piece, or the focal point of a section.
- **mode**: a mode is a type of scale as related to its "tonic", which for example would not include the black keys on a piano (flats and sharps). Mode tells us if a piece is in a major or minor key.
- **tempo**: the overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece, and derives directly from the average beat duration.
- **loudness**: the overall loudness of a track in decibels (dB). Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude).
- **duration**: the duration of a track in seconds as precisely computed by the audio decoder.
- **end of fade in**: the end of the fade-in introduction to a track in seconds.
- **start of fade out**: the start of the fade out at the end of a track in seconds.
- **metadata**: analyze, compute, and track information.
- **timbre**, **pitch**, and **loudness** are described in detail as part of the *segments* interpretation below.

JSON Schema


```
{
  “meta”:
      {
        “analyzer_version”:”3.05a”, "detailed_status":"OK", "filename":"/Users/Jim/
        Desktop/file.mp3", "artist":"Kemp Harris", "album":"Edenton", "title":"Sweet
        Weepin Jesus", "genre":"Blues", "bitrate":320, "sample_rate":44100,
        "seconds":325, "status_code":0, "timestamp":1279120425, "analysis_time":
        4.67276
      },
  “track”:
      {
        "num_samples":7498360, "duration":325.06168,
        "sample_md5":"0a84b8523c00b3c8c42b2a0eaabc9bcd", "decoder":"ffmpeg",
        "offset_seconds":0, "window_seconds":0, "analysis_sample_rate":22050.000,
        "analysis_channels":1, "end_of_fade_in":0.00000, "start_of_fade_out":
        320.32553, "loudness":-11.018, "tempo":104.072, "tempo_confidence":0.557,
        "time_signature":4, "time_signature_confidence":0.311, "key":9,
        "key_confidence":0.771, "mode":0, "mode_confidence":0.553,
        "codestring":"eJwdk8U7m4Rz9Pej...tbtSnk8U7m4Rz980uF", "code_version":3.130
      },
  “bars”:
      [{"start":1.49356, "duration":2.07688, "confidence":0.037}, ...],
  "beats":
      [{"start":0.42759, "duration":0.53730, "confidence":0.936}, ...],
  "tatums":
      [{"start":0.16563, "duration":0.26196, "confidence":0.845}, ...],
  "sections":
      [{"start":0.00000, "duration":8.11340, "confidence":1.000}, ...],
  "segments":
      [{
        "start":0.00000, "duration":0.31887, "confidence":1.000,
        "loudness_start":-60.000, "loudness_max_time":0.10242,
        "loudness_max":-16.511, "pitches":[0.370, 0.067, 0.055, 0.073, 0.108, 0.082,
        0.123, 0.180, 0.327, 1.000, 0.178, 0.234], "timbre":[24.736, 110.034, 57.822,
        -171.580, 92.572, 230.158, 48.856, 10.804, 1.371, 41.446, -66.896, 11.207]
       }, ...]
}
```


Note that several estimated parameter come with a confidence value, ranging between 0 and 1.

Interpretation

Rhythm

Beats are subdivisions of bars. Tatums are subdivisions of beats. That is, bars always start at the same time as a beat and ditto tatums. *Confidence* is a value between 0 and 1 (inclusive) that indicates the reliability of its corresponding attribute. Note that a low confidence does not necessarily mean the value is inaccurate. Exceptionally, a *confidence* of -1 indicates "no" value: the corresponding element must be discarded. A *track* may result with no *bar*, no *beat*, and/or no *tatum* if no periodicity was detected. The *time signature* ranges from 3 to 7 indicating time signatures of 3/4, to 7/4. A value of -1 may indicate no time signature, while a value of 1 indicates a rather complex or changing time signature.

Pitch

The *key* is a number ranging from 0 to 11 and corresponds to one of the 12 keys: C, C#, D to B. If no key was detected, the value is -1. The *mode* is equal to 0 or 1 for "minor" or "major" and may be -1 in case of no result. *Confidence* values always range from 0 to 1. Note that the major key (e.g. C major) could more likely be confused with the minor key at 3 semitones lower (e.g. A minor) as both keys carry the same pitches. Harmonic details are given in *segments* below.

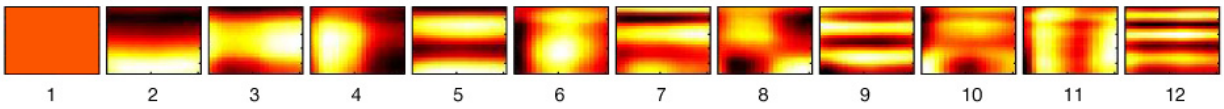
Segments

Beyond timing information (start, duration), segments include *loudness*, *pitch*, and *timbre* features.

- *loudness* information (i.e. attack, decay) is given by two data points, including dB value at onset and maximum dB value. The dB value at offset is equivalent to the dB value at onset for the following *segment*. The last segment specifies a dB value at offset as well.

- *pitch* content is given by a "chroma" vector, corresponding to the 12 pitch classes C, C#, D to B, with values ranging from 0 to 1 that describe the relative dominance of every pitch in the chromatic scale. For example a C Major chord would likely be represented by large values of C, E and G (i.e. classes 0, 4, and 7). Vectors are normalized to 1 by their strongest dimension, therefore noisy sounds are likely represented by values that are all close to 1, while pure tones are described by one value at 1 (the pitch) and others near 0.

- *timbre* is the quality of a musical note or sound that distinguishes different types of musical instruments, or voices. It is a complex notion also referred to as sound color, texture, or tone quality, and is derived from the shape of a segment's spectro-temporal surface, independently of pitch and loudness. Our *timbre* feature is a vector that includes 12 unbounded values roughly centered around 0. Those values are high level abstractions of the spectral surface, ordered by degree of importance. For completeness however, the first dimension represents the average loudness of the segment; second emphasizes brightness; third is more closely correlated to the flatness of a sound; fourth to sounds with a stronger attack; etc. See an image below representing the 12 basis functions (i.e. template segments). The actual timbre of the segment is best described as a linear combination of these 12 basis functions weighted by the coefficient values: timbre = c1 x b1 + c2 x b2 + ... + c12 x b12, where c1 to c12 represent the 12 coefficients and b1 to b12 the 12 basis functions as displayed below. Timbre vectors are best used in comparison with each other.



12 basis functions for the timbre vector: x = time, y = frequency, z = amplitude

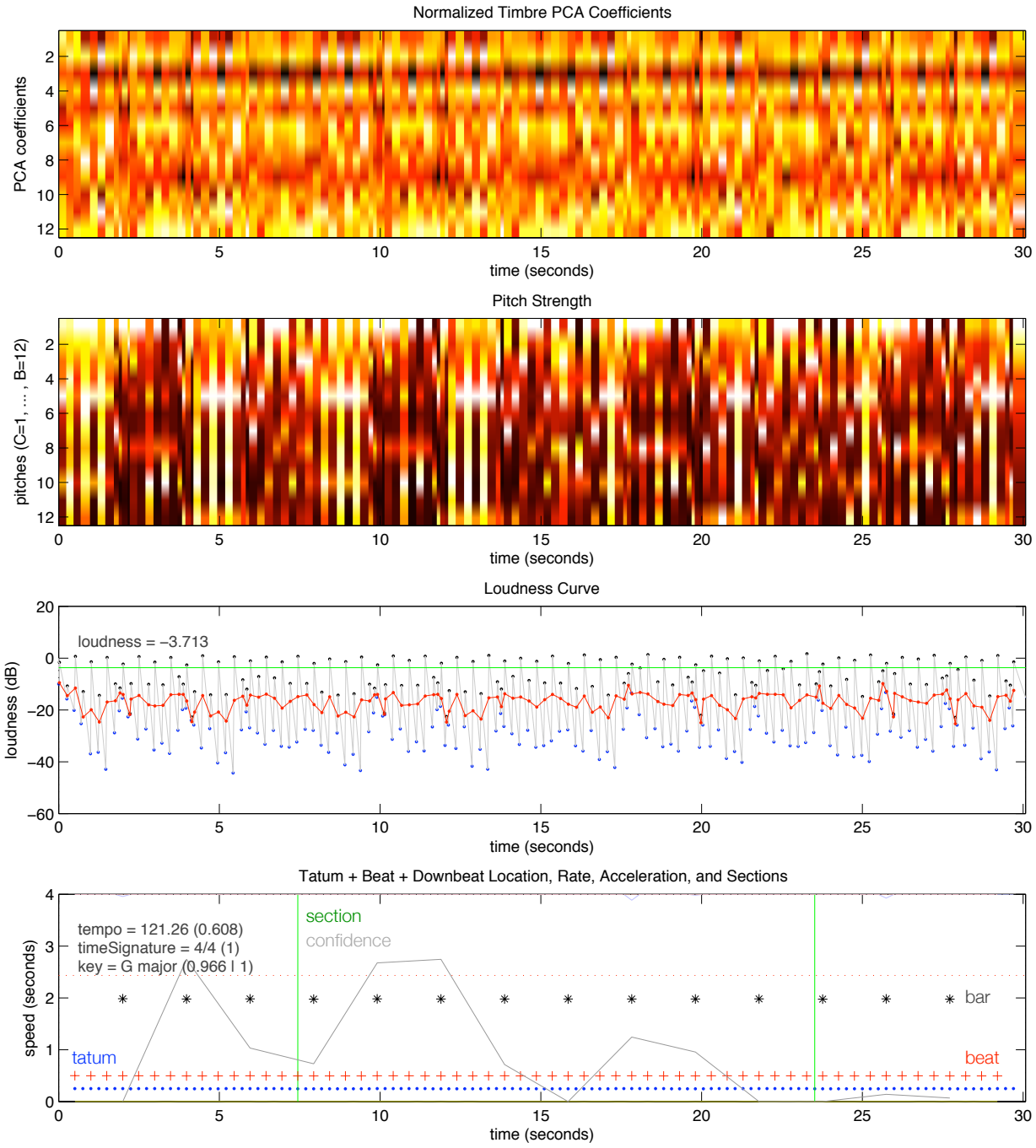
Command-Line Analyzer Usage

$ analyze [options] /Path/To/Filename.ext /Path/To/Output.ext

(ext: a valid audio extension, e.g. mp3, m4a, wav, aif, etc. and output, e.g. xml, json)

where options include:
-d decoder_name
-o offset_in_seconds
-s seconds_to_analyze

Display Example



Plot of the XML data of a 30-second excerpt of “around the world” by Daft Punk.