Online Appendix to: Quantitative Study of Music Listening Behavior in a Smartphone Context

YI-HSUAN YANG, Research Center for Information Technology Innovation, Academia Sinica
YUAN-CHING TENG, Research Center for Information Technology Innovation, Academia Sinica

A. DETAILS OF THE EXTRACTED AUDIO FEATURES

We provide the details of the 9 audio feature sets listed in Table I of this paper. A flowchart of audio feature extraction is depicted in Fig. 11.

To reduce computation load, music signals are sampled at 22,050 Hz and reduced from stereo to mono-channel by averaging the left and right channels. Moreover, considering the limited source on mobile devices for further application, we only took from each song the segment going from the 30th to the 60th seconds for feature extraction and considered the segment-level feature vector as representative of the information of the whole song, a common practice in MIR [Scaringella et al. 2006].

A.1. Temporal

In music, timbre is often thought of as the quality of sound that makes a particular musical sound different from another, even when they have the same pitch and loudness [Müller et al. 2011]. Timbre has been found to be related to 3 main properties of music signals: temporal evolution of energy, spectral envelope shape (relative strength of the different frequency components) and time variation of the spectrum [G. Peeters and McAdams 2011]. We use the MIRtoolbox to compute zero-crossing rate and low-energy rate of the time-domain music signal to characterize its temporal evolution of energy. The former measures the signal noisiness by counting the number of signal values that cross the zero axis in each time window (i.e.,, sign changes), whereas the latter measures whether a music signal has some very loud frames but lots of silent frames by counting the number of frames showing a root-mean-square (rms) energy that is lower than the average energy over time. We divide the music signal into 50%-overlapping frames of 50 ms and aggregate the frame-level zero-crossing rates and low-energy rates by taking the mean and standard deviation across all the frames.

A.2. Spectral

In addition, we obtain the spectrum of a music signal by short-time Fourier transform (STFT; with half-overlapping 50 ms frames) and then calculate the following 10 frame-level spectral features using the MIRtoolbox.

— spectral centroid is the statistical first central moment (center of gravity) of the spectrum; higher spectral centroid indicates "brighter" audio texture.
— spectral spread is the statistical second central moment of the spectrum, measuring how stretched or squeezed the spectrum is.
— spectral skewness is the statistical third central moment of the spectrum, measuring the symmetry of the spectrum with respect to the spectral centroid.
— spectral flatness is calculated as the ratio between the geometric mean and the arithmetic mean of the spectrum, measuring whether the distribution is smooth or spiky.
— spectral entropy is the relative entropy of the spectrum.

DOI: http://dx.doi.org/10.1145/0000000.0000000

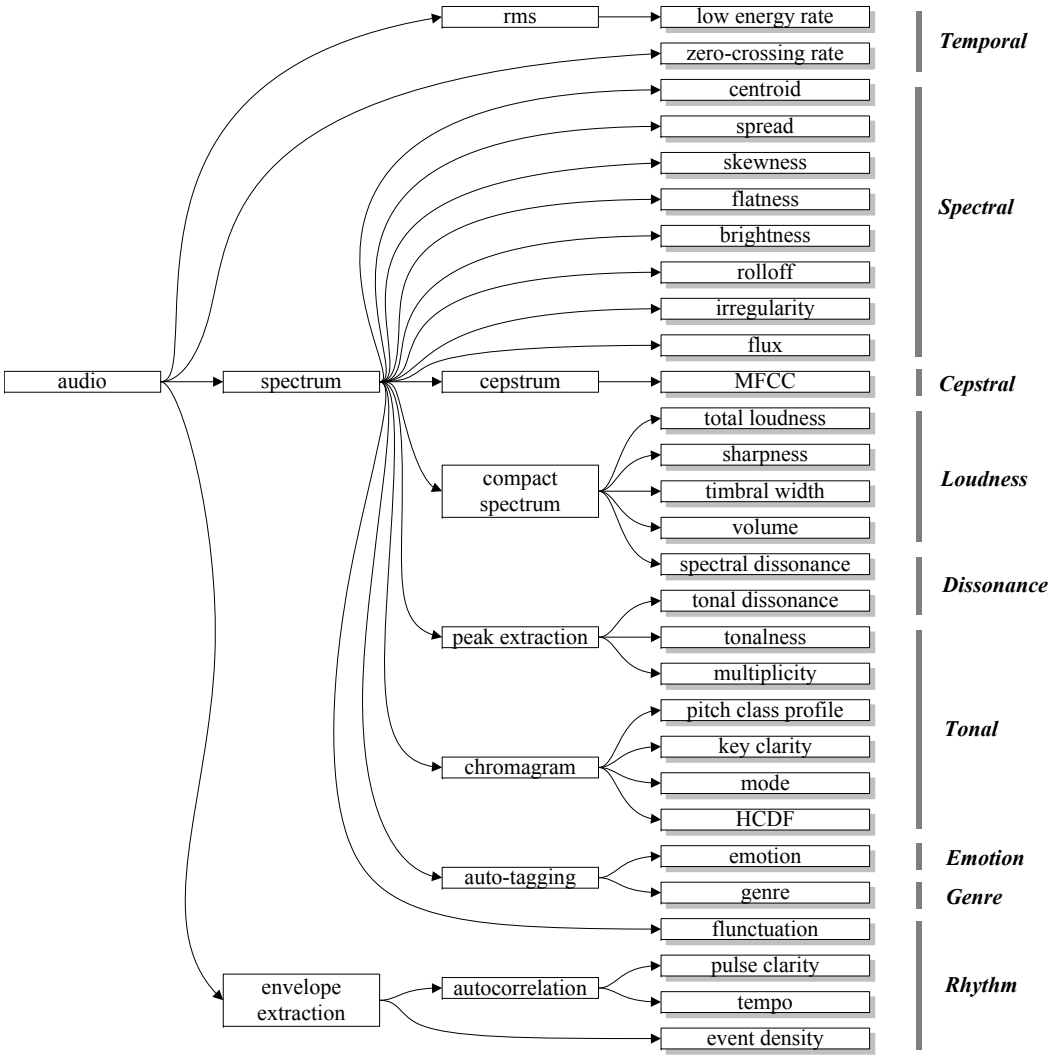

Fig. 11. Flowchart of audio feature extraction.

— spectral brightness is defined as the amount of spectral energy corresponding to frequencies higher than 1,500 Hz.
— spectral rolloff at 0.85 or 0.95 is the frequency below which a certain fraction (i.e., 85% or 95%) of the total energy is contained for a spectrum.
— spectral irregularity measures the degree of variation of the successive peaks of the spectrum; specifically, we adopt the Jensen's model [Jensen 1999] and compute the sum of the square of the difference in amplitude between adjoining partials.
— spectral flux, as an estimate of the amount of local spectral change over time, is calculated as the Euclidean distance between the spectrum of each successive frames.

A.3. Cepstral

We also employ the MIRtoolbox [Lartillot and Toiviainen 2007] and compute the Mel-frequency cepstral coefficients (MFCC), which is arguably the most widely used audio

feature representation. MFCC offers a description of the spectral shape of the sound. It is computed by taking the coefficients of the discrete cosine transform of each short-time log-power spectrum expressed on a non-linear perceptual-related Mel-frequency scale [Davis and Mermelstein 1980]. Typically only the 10–20 lowest DCT coefficients are retained and the rest are discarded in order to make the timbre descriptor invariant to pitch information present in the higher coefficients [Müller et al. 2011]. Following many previous work, we take the first 13 DCT coefficients for each short-time frame and the first-order and second-order derivatives of the 13 coefficients over time.

A.4. Loudness

Loudness is the perceptual intensity of sound. It depends primarily on the physical intensity of sound (i.e.,, sound pressure level), but it is also related to other physical properties of sound, such as frequency and duration. We adopt the Moore and Glasberg's psychoacoustic loudness model [Moore et al. 1997] implemented by PsySound [Cabrera 1999] to model the sensation of loudness by human. The model takes a detailed approach to model the outer and middle ear transfer function and uses a set of auditory filters to model the response of the basilar membrane within the cochlea of the inner ear and the masking effect. The resultant distribution of loudness sensation across the 108-bin compact spectrum is referred to as the specific loudness function (SLF). We then compute the total loudness as the integral of the SLF, the sensational sharpness of sound (from dull to sharp) using the Zwicker (Z)'s model and the Aures (A)'s model, the timbral width as the flatness of the SLF, and the sensational volume of sound (from small to large) using the Stevens' model [Cabrera 1999].

A.5. Dissonance

Dissonance measures the harshness or roughness of the acoustic spectrum [Cabrera 1999]. The dissonance generally implies a combination of notes that sound harsh or unpleasant to people when played at the same time. Empirically, many musical pieces involve a balanced combination of consonance and dissonance sounds, e.g., the release of harmonic tension might create pleasure [Parncutt and Hair 2011]. PsySound calculates dissonance in 4 ways: either the Hutchinson and Knopoff (H&K)'s model or the Sethares (S)'s model can be employed, and the model can be applied to either all the components of the spectrum or only to the spectral peaks [D. Cabrera and Schubert 2007]. The resulting spectral and tonal dissonance measures the degree deviating from the noisiness of the sound and the dissonance among tonal components, respectively.

A.6. Tonal

The tonal and harmonic aspect of music can be described in terms of the relationship between two or more simultaneous pitches [Müller et al. 2011]. In music, pitches can usually be divided into twelve pitch classes, in which a pitch class encompasses all pitches that are a whole number of octaves apart (e.g., `C4` and `C5`). As pitches belonging to the same pitch class are perceived as having similar quality, we employ the MIRtoolbox [Lartillot and Toiviainen 2007] to project the spectrum onto twelve bins representing the twelve distinct semitones (or chroma) of the musical octave. The resultant representation of pitch is usually referred to as the pitch class profile (PCP). We compute 3 additional features from the PCP using the MIRtoolbox: the key clarity indicates how clearly the set of pitches in the frame is organized in a harmonic structure; the musical mode measures the likelihood for the music signal to be played in major key instead of minor key; and the harmonic change detection function (HCDF) indicates large difference in harmonic content between consecutive frames, such as chord changes, strong melody or bass line movement [C. Harte and Gasser 2006]. Moreover,

we use PsySound to compute tonalness (how tone-like the sound is) and multiplicity (the number of pitches heard for each frame) [Cabrera 1999].

A.7. Rhythmic

We extract the following 5 rhythm features using the MIRtoolbox. The first two features are derived from the fluctuation pattern proposed by Pampalk *et al.* [Pampalk et al. 2002], which is based on the Fastl's model of the perceived fluctuation for amplitude modulated tones [Fastl 1982]. we take the peak and centroid of the fluctuation pattern computed over time to characterize the rhythmic periodicities of the music signal. The last 3 features are calculated by extracting the envelop of the time-domain signal and then detecting the onsets, i.e., the starting time of each musical events (notes). The third feature, event density, is calculated as the number of note onsets per second. From the onset detection curve we further apply autocorrelation function to estimate the average tempo (in beats per minute) and the pulse clarity (i.e., the strength of the beats) [O. Lartillot and Fornari 2008].

A.8. Emotion & Genre

We use the supervised approach proposed in [Yang 2013] to train audio-based music auto-taggers for 190 music emotions and 140 music genres. We then use the probability estimate for each of the music emotion or music genre class as features.

Specifically, we use the MER31k dataset to train the music emotion auto-taggers [Yang and Liu 2013]. The dataset contains 31,427 tracks labeled with music emotion tags defined by AllMusic (http://www.allmusic.com/). We queried for related songs for these 190 emotions using the function `Tag.getTopTracks()` provided by the API of Last.fm (www.last.fm/api), which provides a social platform for on-line users to tag music. For each of the 190 emotion tags, a song list was obtained by searching for the top songs labeled with that tag in Last.fm, and the audio previews were downloaded using the 7digital API (http://developer.7digital.net/). On average, there are 165.4 ± 40.4 songs for each emotion tag (max: 248, min: 28). This dataset was developed in [Yang and Liu 2013] for studying the relationship between user mood and music emotion.

On the other hand, we use the CAL10k data set [Tingle et al. 2010] to train the music genre auto-taggers. The dataset comprises 10,870 partially annotated songs by 4,597 artists.Each song is annotated using a vocabulary of up to 1,053 tags by expert musicologists of the music service company Pandora). We are able to collect the 30-second audio previews of 7,799 songs in this collection using the 7digital API. We use the whole dataset to train linear SVM classifiers for the 140 tags related to genres/styles, such as "bebop," "deathcore metal," "flamenco," "live," "oldies," "opera," and "salsa."

For both music emotion and genre, we use randomized clustering forest (RCF) [Moosmann et al. 2008] for constructing bag-of-audio words and linear SVM for classifier training [Fan et al. 2008]. This approach has been shown to achieve good balance between the accuracy of auto-tagging and the efficiency of feature extraction and classification [Yang 2013]. In our implementation, the AUC for auto-tagging the 190 music emotions is on average 0.7157 ± 0.0747, according to the train/validation/test split specified in [Yang and Liu 2013] for inside-dataset cross validation. On the other hand, the AUC for auto-tagging the 140 music genres is on average 0.8084 ± 0.0909, according to an inside-dataset 5-fold cross-validation protocol set by [Tingle et al. 2010]. We see that the accuracy is higher for music genre than for music emotion.

Given the trained SVM models, we predict the music emotion and genre of the 12,248 unique songs in our experiment data. We then use the likelihood for being associated each of the music emotion or genre classes predicted by SVM as the feature representation, leading to a 190-dimensional 'Emotion' feature vector and a 140-dimensional 'Genre' feature vector for each song.