Using a Spectral Flatness Based Feature for

Audio Segmentation and Retrieval

Abstract

A method that utilizes a spectral flatness based tonality feature for segmentation and content-based retrieval of audio is outlined. The method uses the tonality measure which is derived from the discrete bark spectrum as a means of detecting transitions between tonal and noise-like parts of the audio input. The meaning of 'tonal' in this context is different from the music-theoretical meaning and implies that there are dominant sinusoidal components in the spectrum, but, does not indicate that they are consonant or harmonic in any sense. Segmentation is performed by determining the times of these transitions, hence providing reference points for search purposes. Search is carried out by pivoting the query information on these reference points. The cumulative distance between the tonality pattern in successive frames of the query and the candidate sound fragments is used as a measure of similarity.

In order to quantify the tonality, the input is processed as follows : the signal is sampled at 22050 Hz and a 2048-point FFT is performed on each frame using a Hanning analysis window. The window is hopped every 71 msecs. which corresponds to approximately 30 % overlap with the previous window. A pre-emphasis filter is applied to compensate for the reduced sensitivity of the human ear at low frequencies. The bark-band filter outputs are calculated from the FFT output by integration of the power spectral density within the critical bands. This is followed by a stage in which tonality determination is carried out for each critical band. The Spectral Flatness Measure (SFM) and the corresponding tonality coefficient (Johnston 1988) are used to quantify the tonal quality, i.e. how much tone-like the sound is as opposed to being noise-like. SFM is defined by the ratio of the geometric mean to the arithmetic mean of the power spectral density components in each critical band.

A tonality vector is defined to be the collection of tonality coeficients for a single frame. More specifically, the tonality vector contains a tonality coeficient for each critical band. In order to perform segmentation, the tonal-to-noise and the noise-to-tonal transitions are obtained. These transitions are calculated, as a sequence $\boldsymbol{b_j}$ (j is the time index), from the rate of change of the smoothed tonality vector with respect to time. The smoothing occurs due to averaging of the tonality vectors across several frames. By detecting rapid changes in the summary tonality in either direction, segmentation decisions are made. The sequence $\boldsymbol{b_j}$ is an indicator for the overall tonal quality change of the input signal. Whenever tonal inputs become dominant in the input signal, for example, with the start of a vocal or solo instrument, the value of the indicator increases during the transition. That is to say, an increase in the value of $\boldsymbol{b}$ (over time) is interpreted as a transition from a noisy part to a more tonal part. Conversely, the decrease in the value of $\boldsymbol{b}$, possibly extending to the negative extreme, indicates a transition from tonal to noisy spectra.

The search for an audio fragment is done by comparing a short sequence of tonality vectors in the database with the same number of tonality vectors in the search sequence. As an exhaustive search is very costly for this purpose, a selective search that uses segments corresponding to relatively high perceptual entropy is performed. The determination of these segments is performed by the use of the $\boldsymbol{b}$ sequence. The starting times of these segments are called anchor times and are classified as either 'tonal-to-noise', or, 'noise-to-tonal'. The noise-to-tonal anchor

time is determined by a local maximum in $\boldsymbol{b}$. In order for it to qualify as an anchor point, the difference between this maximum and the first minimum that precedes it must be larger than a threshold, ρ_{nt}. Similarly, the tonal-to-noise anchor is found for a minimum with a threshold, ρ_{tn}.

As new audio files are added to the database they are processed to obtain the anchor times and tonality vectors for all frames. This information is stored, to be referenced later in the search phase. A query consists of a short audio fragment. Once a search is initiated, the query is processed to obtain the compatible form of information in the search database, i.e. tonality vectors and anchor times. The tonality vectors starting from the anchor times in the query are aligned with the anchor times of the audio in the database. The search is based on an objective of finding the minimum distance between the tonality variation pattern of the query and the variation pattern in fragments of the audio data in the database. The distance measure is simply the sum of differences of the corresponding tonality values between a candidate sound fragment and the search sequence.

The search for a fragment that is already in the database leads to an exact match and therefore is found without error. An experiment carried out using 14 2-3 minute long pieces from pop, classical and jazz recordings showed that an exact match is always found. When a copy of the query does not exist in the database the most similar fragment is pulled out. The similarity determined by this method is based solely on the variation of tonality in the audio fragments. This means that characteristics such as pitch, loudness or consonance are not dealt with. Hence, the type of similarity found by this method, in general, is characterized by the unfolding of bark-band spectral envelopes and more specifically, reveals vocal or instrumental onset pattern resemblance and likeness of onset patterns of percussive perturbations to steady tonal sounds.

The work described here explores the applicability of a tonality feature based on the Spectral Flatness Measure to segmentation and content-based retrieval for audio data. As explained above, audio fragments that are found to be similar in this categorization are similar in terms of tonal and noise characteristics spanning time and frequency. For various applications, multi-feature systems have been reported to perform successfully on arbitrary audio signals (Scheirer and Slaney 1997; Tzanetakis and Cook 1999). Correspondingly, the tonality feature described here could be combined with other features to further narrow down the 'most similar' list and make the resemblance perceptually more relevant.

Author Information

Ozgur Izmirli
Center for Arts and Technology,
Department of Mathematics and Computer Science,
Connecticut College
oizm@conncoll.edu

References

Johnston, J. D. 1988. "Transform Coding of Audio Signals Using Perceptual Noise Criteria," IEEE Journal of Selected Areas in Communication, Vol. 6, pp. 314-323.

Scheirer, E. and Slaney, M. 1997. "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," Proceedings ICASSP'97, pp.1331-1334.

Tzanetakis G. and Cook, P. 1999. "Multifeature Audio Segmentation for Browsing and Annotation," Proceedings IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 103-106.