Music/Speech Discrimination based on Chroma Vector Analysis

Aiko Uemura and Jiro Katto
Department of Computer Science and Engineering, Waseda University, Japan
{uemura, katto}@katto.comm.waseda.ac.jp

ABSTRACT

We present a music and speech part detection method incorporating chroma vector analysis. We apply image-based mask filters to the time-series chroma components and detect music parts. The envelopes of chroma components of music signals tend to have a horizontal (i.e. temporal) correlation in time-frequency representation because music signals have periodic chord sequences. Based on this fact, we analyze time series of chroma components and attempt to segment music and speech parts in audio signals.

1. INTRODUCTION

We focus on chroma vectors, defined as a 12-dimensional vector that represents the intensity of the 12 semitones pitch classes of the chromatic scale irrespective of octaves [1]. Chroma vectors accommodate harmonic musical structure. They are often used for chord recognition. One might note the horizontal (i.e. temporal) correlation of chroma components in the music scene, but not in a speech scenes because music has harmonic contents and consists fundamentally of chord sequences. Speech scene includes pauses and has no salient periodicity.
In this task, we present a proposed method as shown in Fig. 1. We extract envelopes of chroma vectors in time series to define a mask filter and compute an energy ratio of masked chroma peaks to total peaks for each frame. Chroma peaks are peak components filtered by the mask and total peaks are peaks of chroma that is not filtered. The mask filter is derived using image processing techniques that are applied to time-frequency representation of acoustic inputs. Then the acoustic signals are classified into two parts: music and speech.

2. ALGORITHM

2.1 Chroma vector

For this study, we calculate chroma vectors of three types: Chroma Pitch-base (CP), Chroma Log Pitch (CLP), and Chroma DCT-Reduced log Pitch (CRP). Each element of chroma vectors is represented by 8-bit. First, we calculate the CP by adding up the corresponding values that belong to the same pitch class. Second, we extract the CLP features by applying a logarithmic compression of a pitch representation. A multirate filter bank is designed to calculate the pitch representation of each pitch. It passes all frequencies around the respective center frequency, disregarding

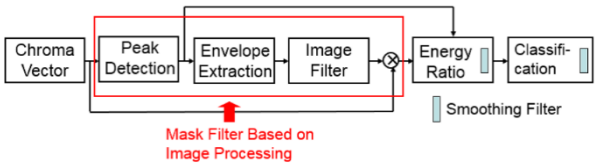

Figure 1. Overview of our proposal

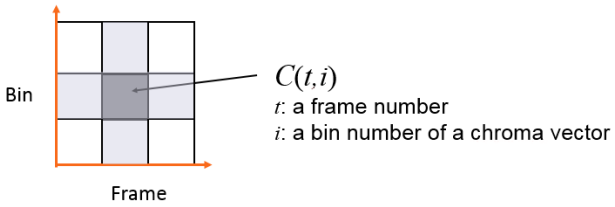

Figure 2. Envelope extraction using four neighborhood bins at peak bin

all other frequencies. The sampling frequencies change to decrease the time resolution naturally for lower frequencies. Then, each energy value e of the pitch representation is calculated using the value $\log(\eta \cdot e + 1)$, where η is a positive constant. We use the number of upper coefficients $\eta = 100$, as described in an earlier report [2]. Third, to calculate CRP, DCT is applied to the logarithmic pitch representation. The upper coefficients of the pitch-frequency cepstrum are employed to the inverse DCT. For our experiments, we use the number of upper coefficients p=55 similar to a report of the literature [2]. The tool used to calculate CP, CLP and CRP features is also provided by chroma toolbox [3].

2.2 Mask filter based on image processing

We detect peaks of chroma components in each frame as shown in Fig. 1, because peaks are regarded as a harmony transition. Then, we calculate curvatures to extract envelopes using four neighborhood bins at peak $C(t, i)$, where t is a frame number and i is a bin number of a chroma vector, as shown in Fig. 2. An initial mask is defined using a binary mask, where 1 represents the existence of temporal consecutiveness of the chroma, and 0 means no existence. This initial mask is shown by

$$m(t,i) = \begin{cases} 1 & \text{if } \kappa_t < \kappa_i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where κ_t checks horizontal (temporal) correlation and κ_i checks vertical correlation, defined respectively as

$$\kappa_t = \frac{C(t-1,i)\cdot C(t+1,i)}{C(t,i)^2}, \quad \kappa_i = \frac{C(t,i-1)\cdot C(t,i+1)}{C(t,i)^2} \tag{2}$$

for the curvature. We furthermore apply bilateral filters [4] to the initial mask to eliminate noise in the speech part.

2.3 Energy ratio calculation and classification

We calculate chroma energy ratio of the masked peak components to total components in each frame. The energy ratio $R(t)$ is calculated by

$$R(t) = \frac{\sum_i \hat{m}(t,i)C(t,i)}{\sum_i C_{peak}(t,i)} \tag{3}$$

where $C_{peak}(t, i)$ is the detected chroma peak. This $R(t)$ is then smoothed by taking one second median filter, which is formulated as

$$\hat{R}(t) = \text{median}\{R(t)\} \tag{4}$$

The window size of the median filter is determined by auxiliary experiment. The ratios of the music scene differ from those of the speech. This energy ratio is then used to classify the acoustic signals as music parts or speech. Herein, this classification is done by simple thresholding:

$$label(t) = \begin{cases} 1 & \text{if } \hat{R}(t) \geq threshold \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where 1 and 0 represent the music scene and other parts in each frame, respectively, and the threshold is pre-specified. All labels are smoothed similarly by taking one second average.

3. REFERENCES

[1] T. Fujishima: "Realtime Chord Recognition of Musical Sound: a System using Common Lisp Music," *Proceedings of the International Computer Music Association*, pp. 464–467, 1999.

[2] M. Müller, S. Ewert: "Towards Timbre-invariant Audio Features for Harmony-based Music," *IEEE Transaction on Audio, Speech, and Language Processing*, Vol. 18, No. 3, pp. 649–662, 2010.

[3] M. Müller and S. Ewert: "Chroma Toolbox: MATLAB Implementations for Extracting Variants of Chroma-Based Audio Features," *Proceedings of the International Symposium on Music Information Retrieval*, pp. 215–220, 2011.

[4] C. Tomashi, R. Manduchi: "Bilateral Filtering for Gray and Color Images," *Proceedings of the International Conference on Computer Vision*, pp. 839–846, 1998.