5D-05

Audio feature extraction based on correlations between undecimated wavelet coeficients

Takuya Kobayashi, Akira Kubota
Department of Electrical, Electronic and Communication Engineering
Chuo University
Tokyo, Japan

***Abstract*—We present a novel low-level audio features that are effective for music genre classification.**

Under the assumption that SVM is used for classifier learning, the experimental results on GTZAN data set showed that the proposed method demonstrated the best accuracy of 81.5%, outperforming the conventional methods.

I. Introduction

Due to development of Information technology, music streaming service has become widespread. And it became possible to trial listening, obtain and hold a large number of songs. On the other hand, it is difficult for the user to search for music desired.

Music information retrieval (MIR), feature extraction and classification audio features have been constantly studied. These features are categorized into low-level, mid-level and top level. We focus on the low-level features that are effective for music genre classification. While the low-level feature has been proven to be effective for music genre classification, classification accuracy using a support vector machine（SVM) is set to 70% or more and less 80% [1].

In the present paper, we present a novel low-level feature based on correlations of wavelet coefficients of audio signals effective for music genre classification. Experimental results using GTZAN data set show SVM [2].

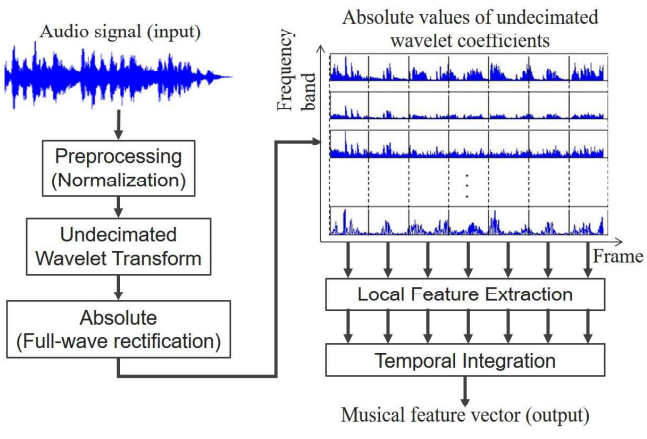

Fig. 1. Flow of the proposed feature extraction method

II. Proposed musical feature extraction

A. Overview

The flow of the proposed feature extraction is illustrated in Fig. 1. The proposed method extracts the musical feature vector using undecimated wavelet transform (UWT) of the input audio signal.

Firstly, the input audio signal is normalized in the preprocessing step and then decomposed into multiple signals in different resolution levels (i.e., sub-band signals) using UWT. Each obtained signal is fully rectified such that it has the absolute values and split into multiple local frames. In each frame, some basic statistics of all the sub-band signals are computed and extracted as local features, referred to as timbre features. These local features over all the fames are finally combined into a single vector as the feature vector using some statistical moments.

B. Preprocessing

Let $x[n]$ $(n = 0, 1, \ldots, N-1)$ denote an input audio signal with N samples. In the preprocessing step, the input audio signal is normalized by its mean absolute value

$$A = \frac{1}{N}\sum_{n=1}^{N} |x[n]|, \tag{1}$$

and therefore the normalized signal is given by

$$y[n] = \frac{1}{A}x[n]. \tag{2}$$

This mean absolute value A is also used as an element of the output feature vector.

C. Undecimated wavelet transform and framing

Using UWT, the normalized signal $y[n]$ is decomposed into the sub-band signals $\{y^{(j)}[n]\}_{j=0,1,\ldots,J-1}$, where J is the number of the sub-bands and the lower j means the higher frequency sub-bands. Each sub-band signal is fully rectified and split into K frames; therefore the signal of k'th frame $(k = 0, 1, \ldots, K-1)$ in j'th sub-band signal, $v_k^{(j)}[m], (m = 0, 1, \ldots, N/K-1)$, is obtained to be

$$v_k^{(j)}[m] = |y^{(j)}[kN/K + m]|. \tag{3}$$

assuming N/K becomes an integer, the number of samples.

D. Local feature extraction

In each frame $v_k^{(j)}[m]$, the local features are extracted as the mean and the coefficient of variation with respect to m, which are respectively computed as

$$\alpha_k^{(j)} = \frac{1}{M}\sum_{m=1}^{M} v_k^{(j)}[m], \tag{4}$$

and

$$\beta_k^{(j)} = \frac{1}{\alpha_k^{(j)}}\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(v_k^{(j)}[m]-\alpha_k^{(j)}\right)^2}, \tag{5}$$

where $M = N/K$.

In addition, correlations between any two different sub-band signals are extracted. These correlations are computed as the coefficients of correlations:

$$\gamma_k^{(i,j)} = \frac{\sum_{m=1}^{M}(v_k^{(i)}[m]-\alpha_k^{(i)})(v_k^{(j)}[m]-\alpha_k^{(j)})}{\sqrt{\sum_{m=1}^{M}\left(v_k^{(i)}[m]-\alpha_k^{(i)}\right)^2}\sqrt{\sum_{m=1}^{M}\left(v_k^{(j)}[m]-\alpha_k^{(j)}\right)^2}}. \tag{6}$$

The combination numbers of $\gamma_k^{(i,j)}$ is given by ${}_J\mathrm{C}_2$.

E. Feature integration

The local features are integrated over all the frames into a single feature vector. For $\alpha_k^{(j)}$ and $\beta_k^{(j)}$, the mean and the coefficient of variation with respect to k are computed as $\mu_\alpha^{(j)}$, $\nu_\alpha^{(j)}$, $\mu_\beta^{(j)}$ and $\nu_\beta^{(j)}$, respectively.

For $\gamma_k^{(i,j)}$, the mean and the variance (in stead of the coefficient of variation) with respect to k are computed as $\mu_\gamma^{(i,j)}$ and $\sigma_\gamma^{(i,j)}$, respectively.

Adding the mean absolute value A, totally seven features are used to create the output feature vector. The dimension of the vector becomes $2\,{}_J\mathrm{C}_2+4J+1$, independent of the number of frames, K.

III. Experiments

In order to evaluate the proposed feature extraction method, we conducted a test of music genre classification using GTZAN dataset and compared the classification accuracy among the conventional methods. Through the experiment, we used SVM for the classifier and training.

A. Dataset

We used GTZAN dataset provided by Tzanetakis and Cook [3]. This dataset contains 1,000 audio signals in ten genres (Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae, and Rock), 100 signals each. The audio signal is 16-bit mono data of 30 seconds with 22,050 Hz sampling.

TABLE I
Classification accuracy of the conventional feature extraction methods with SVM on GTZAN dataset (for CA data, we refer [1])

Method	CA [%]
{MFCC}×FP	77.7
{MFCC}×GMM	70.4
{STFT+MFCC}×MuVar+beat+pitch	72
DWCH+STFT+MFCC×MuVar	78.5
MFCC×MuCov	78.6
STFT+MFCC×MuVar^2	79.8
CR×NTF	78.2
Proposed method	**81.5**

TABLE II
Genre confusion matrix

Bl	Cl	Co	Di	Hi	Ja	Me	Po	Re	Ro	Genres
87	0	3	4	0	1	2	1	2	0	Blues
0	98	0	0	0	1	0	0	0	1	Classic
1	1	79	4	0	3	4	1	1	6	Country
1	1	3	81	4	0	1	2	7	0	Disco
1	0	0	3	81	1	1	6	5	2	Hiphop
5	7	1	0	0	82	2	0	2	1	Jazz
0	0	1	1	0	0	95	1	0	2	Metal
3	1	6	1	6	0	2	77	1	3	Pop
7	1	2	6	7	1	0	2	72	2	Reggae
3	0	9	5	1	4	8	3	4	63	Rock

B. Results

We set the number of frames and sub-bands to be $K = 256$ and $J = 14$, respectively. The dimension of the extracted feature vector is 239.

The classification accuracy (CA) was calculated using a ten-fold cross-validation. The result is shown in Table I, compared with the conventional feature extraction methods when SVM was used for classification. This result shows that the proposed method demonstrated the best performance. And the CA of each feature ($\alpha_k^{(j)}$, $\beta_k^{(j)}$, $\gamma_k^{(i,j)}$) of the proposed method is 60.8%, 61.1%, 71.8%, respectively.

IV. Conclusions

Although not shown in Table I can be said to be all in the music classification feature extraction methods with SVM, the proposed method resulted in higher accuracy than these results. As a point different from the conventional method, coefficient of correlations between any two different sub-band signals is used as a feature and it is found to be valid in music classification.

References

[1] Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang, "A Survey of Audio-Based Music Classification and Annotation," *IEEE Trans. on Multimedia*, Vol.13, No.2, pp.303–319, 2011.

[2] B. E. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. ACM Conf. Computational Learning Theory*, pp. 144–152, 1992.

[3] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293.302, 2002.

[4] F. Pachet and D. Cazaly, "A classification of musical genre," in *Proc. RIAO Content-Based Multimedia Information Access Conf.*, France, Mar. 2000.