A NEW ACOUSTIC FEATURE FOR COVER SONG IDENTIFICATION

Ning Chen
East China University of Science and Technology nchen@ecust.edu.cn

ABSTRACT

In this extended abstract, an auditory theory based feature extraction algorithm is proposed for cover song identification task. In our algorithm, first, the time-varying loudness of the input music is counted with gs2002 model to simulate the frequency-dependent transmission characteristics of the outer ear. Next, the output is filtered by the gammachirp filterbank to model the frequency selective characteristic of the basilar membrane. Then, the Discrete Cosine Transform (DCT) is performed on the filterbank's outputs to achieve decorrelation of them. Finally, the Non-negative Matrix Factorization (NMF) is operated on the DCT compacted features to reduce their dimension and maintain the discriminative performance, simultaneously. In addition, the proposed acoustic feature is combined with one of the most efficient similarity method, called Q_{max}, to evaluate its performance in audio cover song identification task in 2014 Music Information Retrieval Evaluation Exchange (MIREX).

1. INTRODUCTION

A cover song is an alternative version, performance, rendition, or recording of a previously recorded musical piece, so it might differ from their original in several musical aspects such as timbre, tempo, song structure, main tonality, arrangement, lyrics, or language of the vocals [1]. Thus, cover song identification task is a big challenge. Most of the available systems use PCP, a feature based on musical information, or its variation as primary source of information, because it is 1) robust to noise and non-tonal sounds, 2) independent of loudness and dynamics, 3) independent of tuning [2]. However, the PCP feature or its variation does not take the listener's auditory perception when presented with music into consideration, so it can not process the music as human's ears do. To solve this problem, a new acoustic feature extraction algorithm is put forward for cover song identification task. To evaluate the performance of the proposed scheme, it is combined with Q_{max} similarity method [1] and sent to the auditory cover song identification task of 2014 Music Information Retrieval Evaluation Exchange (MIREX) [3].

2. PROPOSED ALGORITHM

The schematic diagram for the proposed feature extraction algorithm is shown in Figure 1.

(i) Pre-processing: to make the feature extracting procedure more efficient and faster, the input music is converted from stereo to mono and then resampled to 8 kHz. The pre-processed music is denoted as **s**.

(ii) Framing: the pre-processed music **s** is windowed into short frames $\{\mathbf{s}_i | i = 1, \cdots, N\}$ of 464 ms with no overlap between subsequent windows, because the non-stationary music signal can be assumed to be stationary for such a short interval and it will facilitate the time-frequency analysis. The most important is that it will increase the efficiency of the feature extraction procedure.

(iii) Loudness counting: to simulate the transfer function from the sound field to the oval window of the cochlea, each frame $\mathbf{s}_i$ is filtered by the time-varying loudness counter gm2002 [4] (the code of which can be found on website [1]) to get $\mathbf{s}_{gm\_i}$. The purpose is to compensate for the frequency-dependent transmission characteristics of the outer ear, the tympanic membrane, and the middle ear.

(iv) Gammachirp filterbank [5] filtering and down-sampling: to simulate the frequency selective characteristic of the basilar membrane of human ear, each loudness filtered signal $\mathbf{s}_{gm\_i}$ is passed through a gammachirp filterbank that is composed of N_c channels to get

$$\mathbf{S}_{gc\_gm\_i} = \left[\mathbf{s}^{(1)}_{gc\_gm\_i}, \cdots, \mathbf{s}^{(j)}_{gc\_gm\_i}, \cdots, \mathbf{s}^{(N_c)}_{gc\_gm\_i}\right], \quad i = 1, \cdots, N \tag{1}$$

where $\mathbf{s}^{(j)}_{gc\_gm\_i}$ is obtained by passing $\mathbf{s}_{gm\_i}$ through the j-th channel of the gammachirp filterbank. The center frequency of the filters in the filterbank ranges from 100 Hz to 20 kHz.

The gammachirp filter has a well-defined impulse response (see Eq.(2)), so gammachirp filterbank is an excellent candidate for an asymmetric, level dependent auditory filterbank in time-domain models of auditory processing [5].

$$g_j(t) = at^{n-1}e^{-2\pi bERB(f_j)t}cos(2\pi f_j t + c\ln t + \phi), (t > 0) \tag{2}$$

where t represents time, a, n, b, f_j, c and ϕ are parameters, $\ln$ is the natural logarithmic operator. $ERB(f_j)$ is the Equivalent Rectangular Bandwidth (ERB) of the filter centered at f_j, and at moderate levels $ERB(f_j) = 24.7 + 0.108f_j$ in Hz.

[1] http://code.soundsoftware.ac.uk/projects/aimmat.

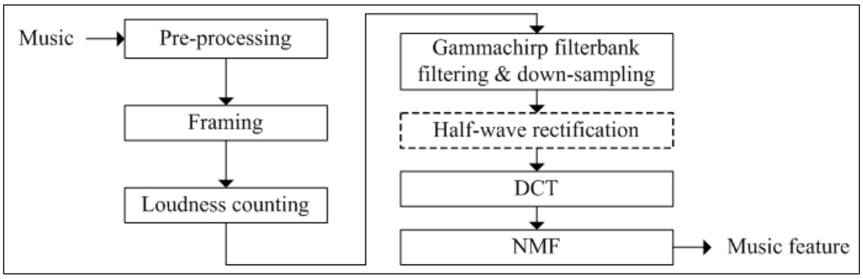

Figure 1. Schematic diagram of the proposed feature extraction algorithm.

To reduce the dimension of the feature, the output of each channel $\mathbf{s}^{(j)}_{gc\_gm\_i}$ is down sampled to 100 Hz to get $\hat{\mathbf{s}}^{(j)}_{gc\_gm\_i}$. Thus, for each frame $\mathbf{s}_i$, we get an auditory feature matrix, denoted as $\hat{\mathbf{S}}_{gc\_gm\_i}$

$$\hat{\mathbf{S}}_{gc\_gm\_i} = \left[\hat{\mathbf{s}}^{(1)}_{gc\_gm\_i}, \cdots, \hat{\mathbf{s}}^{(N_c)}_{gc\_gm\_i}\right] \tag{3}$$

(v) Half-wave rectification: To simulate the response of hair cell to the output of the basilar membrane and keep it phase-locked to the peaks in the wave, the half-wave rectification is applied on the downsampled output of the gammachirp filterbank to get the auditory feature matrix $\hat{\mathbf{S}}_{gc\_gm\_i}$. This step is only included in version NC3, but not in version NC2. This is the only difference between these two versions.

(vi) Discrete Cosine Transform (DCT): to achieve energy compaction, DCT is applied on the auditory feature matrix $\hat{\mathbf{S}}_{gc\_gm\_i}$ (see Eq.(4)) to get $\hat{\mathbf{S}}_{dct\_gc\_gm\_i}$.

$$\hat{\mathbf{S}}_{dct\_gc\_gm\_i} = \hat{\mathbf{S}}_{gc\_gm\_i} \times \mathbf{X}_{DCT} \tag{4}$$

where the elements $X_{DCT}(m_1, m_2)$ of the $\mathbf{X}_{DCT}$ is defined as

$$X_{DCT}(m_1, m_2) = \sqrt{\frac{2}{N_c}} cos\left(\frac{\pi}{2N_c} \cdot m_1 \cdot (2m_2 - 1)\right), \quad m_1, m_2 = 1, \cdots, N_c \tag{5}$$

Only the front half of each colunm in matrix $\hat{\mathbf{S}}_{dct\_gc\_gm\_i}$ is reserved, thus, we get $\hat{\mathbf{S}}_{half\_dct\_gc\_gm\_i}$.

(vii) NMF: To reduce the dimension of compacted auditory feature matrix further and maintain its discriminative performance at the same time, rank 1 NMF is performed on $\hat{\mathbf{S}}_{half\_dct\_gc\_gm\_i}$ with Eq.(6).

$$\hat{\mathbf{S}}_{half\_dct\_gc\_gm\_i} \approx \mathbf{w}_i \times \mathbf{f}_i \tag{6}$$

Then, $\mathbf{f}_i = [f_i(q)|q = 1, \cdots, N_c/2]$ is the acoustic feature of the frame $\mathbf{s}_i$. And the feature matrix $\mathbf{F}$ of the whole song is gotten as follow.

$$\mathbf{F} = \left[\mathbf{f}_1{}^T, \cdots, \mathbf{f}_i{}^T, \cdots, \mathbf{f}_N{}^T\right] \tag{7}$$

where T indicates matrix transposing.

To evaluate the performance of the proposed feature in an audio cover song identification task, it was combined with one of the most efficient similarity measures, called Q_{max} [1], and submitted to the audio cover song identification task in MIREX 2014.

3. CONCLUSIONS

A new feature extraction scheme is proposed for an audio cover song identification task. This scheme is different from the available ones because: i) Most of the available schemes extract feature based on music theory, while the proposed one composes feature based on auditory model. ii) Different dimension reduction methods, such as resampling, DCT and NMF, are combined in the proposed scheme to reduce the feature's dimension to the largest extent to make the similarity measuring much faster. Our future work is to study the feature extraction algorithm combining the merits of auditory model and music theory.

4. REFERENCES

[1] J. Serrà and X. Serra and R.G. Andrzejak. " Cross recurrence quantification for cover song identification. *New Journal of Physics*, Vol. 11, No. 9, pp. 1–20, 2009.

[2] J. Serrà, and E. Gómez and P. Herrera. *Advances in Music Information Retrieval*, Springer-Verlag Berlin Heidelberg, 2010.

[3] J. S. Downie. "The music information retrieval evaluation exchange (20052007): a window into music information retrieval research. ," *Acoustical Science and Technology*, Vol. 29, No. 4, pp. 247–255, 2008.

[4] B.R. Glasberg and B.C.J. Moore "A model of loudness applicable to time-varying sounds," *Journal of the Audio Engineering Society*, Vol. 50, pp. 331–342, 2002.

[5] T. Irino and R.D. Patterson. "A time-domain, level-dependent auditory filter: the GammaChirp," *the journal of the acoustical society of America*, Vol. 101, No. 1, pp. 412–419, 1997.