A CHROMA-BASED TEMPO-INSENSITIVE DISTANCE MEASURE FOR COVER SONG IDENTIFICATION

Jesper Højvang Jensen
Aalborg University
Dept. Electron. Syst.

Daniel P.W. Ellis
Columbia University
LabROSA

Mads G. Christensen
Aalborg University
Dept. Electron. Syst.

Søren Holdt Jensen
Aalborg University
Dept. Electron. Syst.

ABSTRACT

In the context of music, a cover version is a remake of a song, often with significant stylistic variation. In this paper we describe a distance measure between sampled audio files that is designed to be insensitive to instrumentation, time shift, temporal scaling and transpositions. The algorithm was submitted to the Music Information Retrieval eXchange (MIREX) 2007 audio cover song identification task, where it came fourth of the eight submitted algorithms.

1 INTRODUCTION

As the size of digital music collections increase, navigating such collections become increasingly difficult. One of the goals of the music information retrieval community is to develop signal processing algorithms to facilitate such navigation, for instance by finding cover versions of a song. Comparing different algorithms has been impractical, as copyright issues have prevented the development of standard music collections. The annual MIREX evaluations overcome this problem by having participants submit their algorithms which are then centrally evaluated. This way, distributing song data is avoided. If the test collection will not come to the algorithms, the algorithms will go to the test collection.

The MIREX cover song retrieval contest was first held in 2006, where the algorithm in [3] had the best retrieval performance. This algorithm was especially developed for cover song identification and was computationally relatively expensive. It combined the chromagram with a beat tracker in order to obtain a beat-synchronous chromagram that was insensitive to temporal differences between different versions of a song. Instead of beat tracking, the submission described in this paper uses a feature that is insensitive to time shifting and temporal scaling.

2 OVERVIEW

In Figure 1, a block diagram of the proposed algorithm is shown. The assumptions behind are that a song and its cover versions share the same melody, but might differ with respect to instrumentation, time shifts, tempo and transpositions. The feature extracted as shown in Figure 1 is insensitive to the former three properties, while the distance computation ensures invariance to transpositions. In Figure 2 and 3, the output of different stages in the feature extraction is shown. Note that except a horizontal shift of one band, Figure 2(c) and 3(c) are very similar.

2.1 Chromagram

The chromagram is conceptually a frequency spectrum which has been folded into a single octave [1]. This single octave is divided into 12 logarithmically spaced frequency bins that each correspond to one semitone on the western musical scale. Although only an approximation, we will consider the chromagram a soft decision activation pattern that tells whether a given note is playing. To compute the chromagram, we use the implementation described in [3]. Empirically, we found that using the logarithm of the chromagram increased performance. Let $\boldsymbol{Y}$ be a matrix containing the chromagram where element $(\boldsymbol{Y}_{\log})_{ij}$ measures the strength of semitone i, $i \in 1, 2, \cdots, 12$, in frame j. To avoid numerical problems, we then compute the logarithm as

$$(\boldsymbol{Y}_{\log})_{ij} = \log \frac{(\boldsymbol{Y})_{ij} + \delta}{\delta}, \tag{1}$$

where δ is a small constant.

2.2 Power spectrum

To avoid time alignment problems, we remove all phase information by computing the power spectrum for each row of $\boldsymbol{Y}_{\log}$, i.e., the activation pattern for each semitone. Empirically, the power spectrum performed better than the amplitude spectrum.

2.3 Time-scale invariance

If $x(t)$ is a continuous signal and $X(f) = \mathcal{F}\{x(t)\}$ is its Fourier Transform, then a temporal scaling of $x(t)$ will also cause a scaling in the frequency domain: $\mathcal{F}\{x(kt)\} = X(f/k)$. For cover song detection, it is reasonable to assume that k is bounded, i.e., that two songs do not differ in tempo more than e.g. a factor 1.4, in which case $\frac{1}{1.4} \leq k \leq 1.4$. If either the time or frequency axis is viewed on a logarithmic scale, a time scaling (i.e., $k \neq 1$)

This research was supported by the Intelligent Sound project, Danish Technical Research Council grant no. 26–04–0092, and the Parametric Audio Processing project, Danish Research Council for Technology and Production Sciences grant no. 274–06–0521.

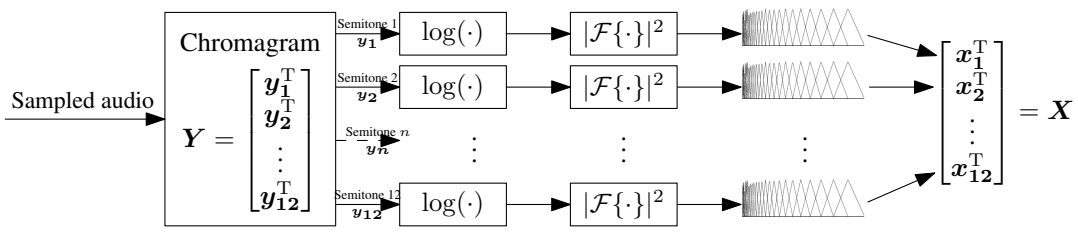
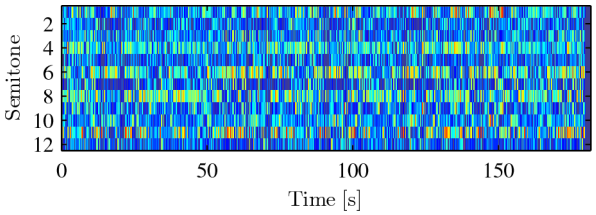
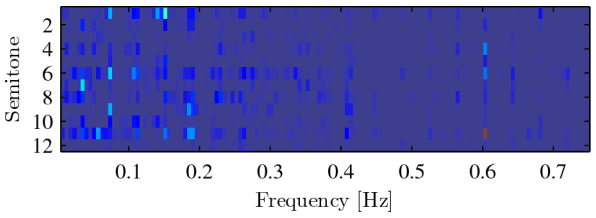


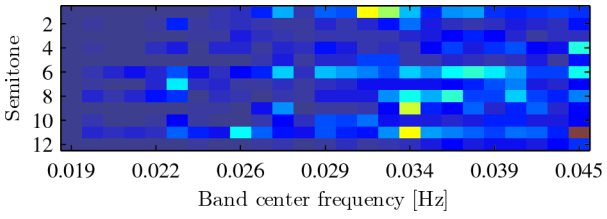
Figure 1. Block diagram of the feature extraction.



(a) Chromagram after taking the logarithm.

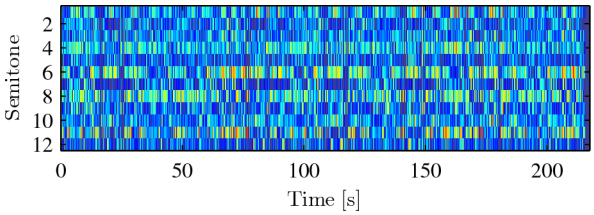


(b) Power spectrum of the chromagram rows.

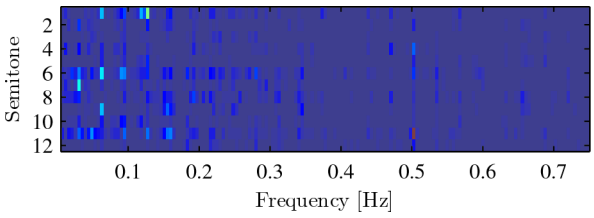


(c) Energy in the 25 exponentially spaced bands

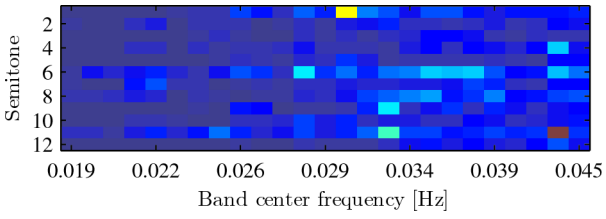
Figure 2. Different stages of feature extraction from MIDI song with duration 3:02.



(a) Chromagram after taking the logarithm.



(b) Power spectrum of the chromagram rows.



(c) Energy in the 25 exponentially spaced bands

Figure 3. Feature extraction from the same MIDI song as in Figure 2, except it is stretched to have duration 3:38.

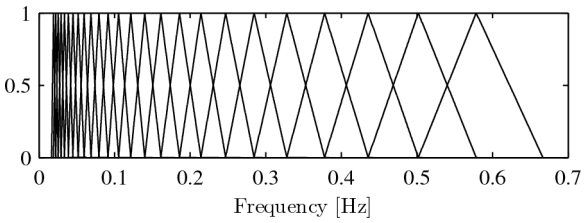


Figure 4. Bandwidths of the 25 logarithmically spaced filters.

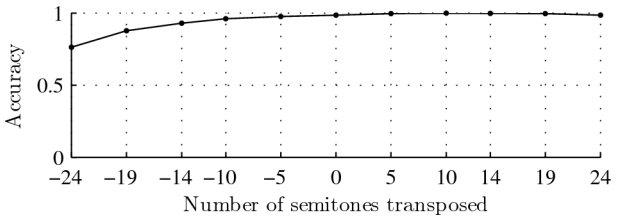


Figure 5. Effect of transpositions on melody recognition accuracy.

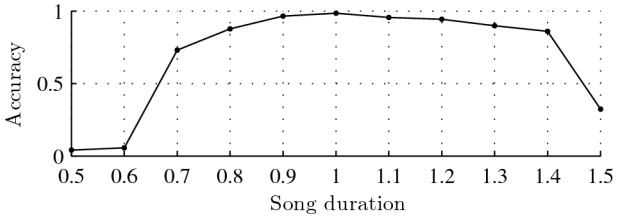


Figure 6. Effect of lengthening or shortening a song on melody recognition accuracy. Duration is relative to original song.

will show up as an offset. This is used in e.g. [5] to make the distance between the fundamental frequency and its harmonics independent of the fundamental frequency itself. As we assume k is bounded, then so will the offset be. Thus, on the logarithmic scale, we could represent the signal by the output of a number of equally spaced bands with effective bandwidth corresponding to the potential offset. Alternatively, on the linear scale, a set of exponentially spaced bands with relatively large effective bandwidth would do the trick. In Figure 4, the 25 logarithmically spaced bands with 50% overlap that we used are shown. The lowest band start at 0.017 Hz, and the highest band end at 0.667 Hz, thus capturing variations on a time scale between 1.5 s and 60 s. The amount of temporal scaling allowed is further increased when computing the distance. The resulting feature is a 12×25 matrix where component i, j reflects the activity for semitone i in band j.

2.4 Distance

We compute the distance between the two feature matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ by normalizing them to unit norm and minimizing the Frobenius distance over the allowed transpositions and frequency shifts. First, we normalize to unit Frobenius norm:

$$\boldsymbol{X}'_1 = \boldsymbol{X}_1 / \|\boldsymbol{X}_1\|_{\mathrm{F}}, \tag{2}$$

$$\boldsymbol{X}'_2 = \boldsymbol{X}_2 / \|\boldsymbol{X}_2\|_{\mathrm{F}}. \tag{3}$$

Let $\boldsymbol{T}_{12}$ be the 12×12 permutation matrix that transposes $\boldsymbol{X}'_1$ or $\boldsymbol{X}'_2$ by one semitone:

$$(\boldsymbol{T}_{12})_{i,j} = \begin{cases} (\boldsymbol{I})_{i+1,j} & \text{for } i < 12, \\ (\boldsymbol{I})_{1,j} & \text{for } i = 12, \end{cases} \tag{4}$$

where $\boldsymbol{I}$ is the identity matrix. To compensate for possible transpositions, we minimize the Frobenius distance over all possible transpositions:

$$d'(\boldsymbol{X}'_1, \boldsymbol{X}'_2) = \min_{t \in \{1,2,\cdots,12\}} \|\boldsymbol{T}_{12}^t \boldsymbol{X}'_1 - \boldsymbol{X}'_2\|_{\mathrm{F}}. \tag{5}$$

To allow even further time scaling than permitted by the effective bandwidths, we also allow shifting the matrix:

$$d(\boldsymbol{X}'_1, \boldsymbol{X}'_2) = \min_{s \in \{-2,-1,0,1,2\}} d'(\boldsymbol{X}'^{(s)}_1, \boldsymbol{X}'^{(-s)}_2), \tag{6}$$

where

$$\boldsymbol{X}'^{(s)}_l = \begin{cases} \begin{bmatrix} \mathbf{0}_s & \boldsymbol{X}'_l \end{bmatrix} & \text{if } s \geq 0, \\ \begin{bmatrix} \boldsymbol{X}'_l & \mathbf{0}_{-s} \end{bmatrix} & \text{if } s < 0, \end{cases} \tag{7}$$

and where $\mathbf{0}_s$ is a $12 \times s$ matrix of zeros.

3 RESULTS

We have evaluated the distance measure by using a nearest neighbor classifier on two different datasets, namely a set of MIDI files [4] and the covers80 set [2]. The basic set of MIDI files consists of 900 MIDI songs that are the combinations of 30 different melodies of length 180 seconds played with 30 different instruments. To measure the sensitivity to transpositions and variations in tempo, the set in which the nearest neighbor is found is replaced by transposed versions of the songs and lengthened/shortened versions. In Figure 5 the effect of transpositions is shown, and in Figure 6 the effect of changing the tempo is shown. It is seen that transposing songs hardly affect performance, and that changing the tempo between a factor 0.7 and 1.4 also does not affect performance too seriously.

The covers80 dataset consists of 80 titles each in two different versions, i.e., a total of 160 songs. With this set, a song's nearest neighbor was the cover version in 36% of the cases. However, as parameters have been tweaked using this dataset, some degree of overtraining is inevitable. The algorithm was also submitted for the MIREX 2007 audio cover song identification task. The results of this evaluation are shown in Table 1.

Rank	Participant	Avg. prec.	Covers in top 10
1	Serrà & Gómez	0.521	1653
2	Ellis & Cotton	0.330	1207
3	Bello, J.	0.267	869
4	*Jensen, Ellis, Christensen & Jensen*	*0.238*	*762*
5	Lee, K. (1)	0.130	425
6	Lee, K. (2)	0.086	291
7	Kim & Perelstein	0.061	190
8	IMIRSEL	0.017	34

Table 1. MIREX 2007 Audio Cover Song Identification results. In comparison, the 2006 winner [3] identified 761 cover songs in top 10.

4 REFERENCES

[1] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: using chroma-based representations for audio thumbnailing," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 2001, pp. 15 – 18.

[2] D. Ellis and G. Poliner, "Identifying cover songs with chroma features and dynamic programming beat tracking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2007, pp. 1429–1432.

[3] D. P. Ellis, "Identifying 'cover songs' with beat-synchronous chroma features," in *Music Information Retrieval Evaluation eXchange*, 2006.

[4] J. H. Jensen, M. G. Christensen, and S. H. Jensen, "A framework for analysis of music similarity measures," in *Proc. European Signal Processing Conf.*, 2007, pp. 926–930.

[5] S. Saito, H. Kameoka, T. Nishimoto, and S. Sagayama, "Specmurt analysis of multi-pitch music signals with adaptive estimation of common harmonic structure," in *Proc. Int. Symp. on Music Information Retrieval*, 2005, pp. 84–91.