COMBINING CHROMA FEATURES FOR COVER VERSION IDENTIFICATION

Teppo E. Ahonen
Department of Computer Science, University of Helsinki
teahonen@cs.helsinki.fi

ABSTRACT

We present an approach for cover version identification which is based on combining different discretized features derived from the chromagram vectors extracted from the audio data. For measuring similarity between features, we use a parameter-free quasi-universal similarity metric which utilizes data compression. Evaluation proves that combined feature distances increase the accuracy in cover version identification.

1. INTRODUCTION

Measuring similarity in music is an essential challenge in music information retrieval (MIR). However, the definition of similarity is not trivial. Clearly, pieces of music from the same genre are similar in various features such as orchestration, but the essential similarity of the compositions can vary largely within the genre.

Cover version identification provides a valid, objective way to estimate how well similarity in music can be recognized and measured. Cover versions often differ in various musical features, but can still be distinguished to be different performances of one composition by a human listener. Thus, successful cover version identification yields important information on how similarity in music can be measured and how features affecting the similarity can be represented.

We approach the problem of cover version identification by taking into account several features derived from the chromagram. These features are represented using different kinds of discrete alphabets and the similarity between features is calculated using a similarity metric called normalized compression distance (NCD) [9]. Evaluation shows that when using NCD for cover version identification, better identification accuracy can be obtained by taking several features into account instead of just focusing on a single feature.

Cover version identification is an objective way to estimate the performance of a retrieval system based on musical similarity. Cover versions, especially in popular music, are often intentionally different from the original recordings of the composition. Changes are common in such features as the musical keys, structures, tempos and arrangements. Also, the lyrics can be altered, translated to another language, or completely discarded. It is more difficult to estimate the features which do not change, but usually these are the melodic and harmonic features.

When successful, cover version identification provides a reliable content-based way to measure the essential similarity in different pieces of music. This provides various potential targets of applications for such systems and algorithms, ranging from end users to music researchers.

In recent years, cover version identification has gained a significant amount of interest from the MIR community. Although a relatively short time has passed since the problem of cover version identification was addressed, the problem has been studied extensively and with various different kinds of approaches.

The most important feature in cover version identification is the chromagram. Chromagram, also known as the pitch class profile, is a sequence of 12-dimension vectors which describe the relative energy of each semitone pitch class. As such, chromagram captures important tonal information and represents the harmonic and melodic content of the audio file.

Various different methods for measuring similarity between chromagrams or features derived from chromagrams exist. These include dynamic time warping and other edit distance variants, dot product and cross correlation. For an extensive and comparative review on different cover version identification approaches, we refer to [20].

The MIREX (Music Information Retrieval Evaluation eXchange) is a community-driven effort providing evaluation for different MIR applications. Cover version identification has been a MIREX task since 2006, and through the years, several different approaches have participated in the evaluation and significant improvement in the identification performance can be perceived. In 2009, the best performing cover version identification application performed with a mean of average precision value of 0.75 [1], suggesting that there still are several unsolved problems in cover version identification which need to be addressed until the problem can be declared solved.

We propose an approach that uses a similarity metric called normalized compression distance (NCD) [9] for measuring the similarity between features extracted from the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
© 2010 International Society for Music Information Retrieval.

[1] http://www.music-ir.org/mirex/2009/index.php/Audio_Cover_Song_Identification_Results

audio files. For features, we extract several different representations from the chromagram vectors. As data compression works with discrete symbols, we use several different techniques for quantizing the continuous chroma values. Our starting point is that different representations have more distinguishing power when combined than they have when used alone. Also, we assume that when using NCD the chromagram cannot be quantized into a representation which both contains all the required information and is not too noisy. Thus, different features must be represented and measured on their own.

The rest of this paper is organized as follows. In Section 2, we give a brief tutorial on the concepts and theories behind the normalized compression distance. In Section 3 we describe the chroma features we use for identification. The approach is evaluated in Section 4. Finally, we present conclusions and discussion in Section 5.

2. NORMALIZED COMPRESSION DISTANCE

Normalized compression distance (NCD) is a distance metric that has its roots in information theory. The idea is to measure the information in an object using Kolmogorov complexity, the length in bits of the shortest binary program that produces the object as an output. Based on the Kolmogorov complexity, a universal information distance can be calculated. This distance, called normalized information distance [9], is denoted

$$NID(x,y) = \frac{max\{K(x|y), K(y|x)\}}{max\{K(x), K(y)\}} \qquad (1)$$

where $K(x)$ is the Kolmogorov complexity of the string x and $K(x|y)$ is the conditional Kolmogorov complexity, meaning the length of $K(x)$ given the information of y.

However, Kolmogorov complexity is non-computable, and thus the normalized information distance cannot be calculated. However, Kolmogorov complexity can be approximated using any standard lossless data compression algorithm. The better the compression of a string is, the closer the approximation is to the Kolmogorov complexity.

The normalized compression distance approximates the Kolmogorov complexity with the aid of a data compression algorithm. For strings x and y, the NCD is denoted

$$NCD(x,y) = \frac{C(xy) - min\{C(x), C(y)\}}{max\{C(x), C(y)\}}, \qquad (2)$$

where $C(x)$ is the length of the string x when compressed using a standard lossless data compression algorithm C and xy is the concatenation of the two strings.

NCD is proven to be robust against noise in the data [8], and studies have proven that observing several common pitfalls of the compression algorithms will help to evade problems when measuring the distances [7]. Especially, PPM-based (Prediction by Partial Matching) compression algorithms have been proven to be resistant against noise [8] and perform well in NCD calculation despite the lengths of the files [7].

Normalized compression distance has been used for several tasks in MIR. In the symbolic domain, there has been research at least in melody classification [16], genre classification [9], composer classification [9] and piano music classification [10]. In the audio domain, NCD has been applied for tasks such as structure-based clustering [3], genre classification [6, 17], cover version identification [1] and query by example [12].

3. CHROMA FEATURES

The chromagram seems to be the only valid feature to be used for cover version identification. For example, the MFCC vectors capture the timbral information of the audio file, but this information has very little help in identifying cover versions. The chromagram is robust against the changes in instrumentation and dynamics, and it captures both melodic and harmonic information from the audio file.

The easiest way to measure similarity between chromagrams using NCD would seem to be converting the chromagram into a sequence of characters and calculating the distance between these. However, we noticed that this approach has several drawbacks. If the alphabet used in sequences is small, the information contained in the chromagram will be too reduced and different sequences will turn out too similar, making distinguishing the sequences challenging. A large alphabet that contains most of the information of the chromagram, on the other hand, will make sequences noisy and lead into insignificant compression and thus into impractical identification. Our solution is to extract various feature sets of the chromagram and measure the similarities between each set.

For obtaining chromagram from the audio file, we use MIRToolbox [15], version 1.3. The window length for the Fourier transform needed in obtaining the chromagram is 0.1858 seconds and the hop factor is 0.875. We use a four-octave range of transformation with a minimum frequency of 55 Hz.

We do not have any tempo estimation and beat averaging over the chromagram frames. This is based on the assumption that unsuccessful tempo estimation might lead to even noisier representations and thus to worse identification results. A similar observation was made in [2], where frame-based identification yielded better results than the tactus-based version. Also, in [1] it was suggested that the shorter chroma sequences produced by the beat averaging may have a negative impact on the NCD values, because the error between $K(x)$ and $C(x)$ minimizes as the file length increases [9].

For compression, we use the PPMZ compression algorithm. The PPMZ is a statistical, more efficient compression algorithm than the more commonly used gzip and bzip2. Thus, it provides a better approximation of the Kolmogorov complexity. This may not, however, lead automatically into better NCD values, as the improvements in compression may be different for the different items in the formula and thus cause the NCD value to move away from the NID value [9].

3.1 Chroma Sequence Labeling

In order to measure similarity successfully with a compression algorithm, the continuous chroma vectors need to be quantized. Out of the several existing quantization methods, the hidden Markov model (HMM) has the advantage of taking into account the temporal statistics. The HMM approach has been studied extensively in converting chroma vectors into a discrete representation, and it is a common method when estimating a chord sequence representation from the harmonic content of the audio. The approach can be described as a process of using the chroma vectors as observations for a HMM whose each state represents a triad chord, training the model with the expectation-maximization (EM) algorithm, and finally obtaining the state transition path using the Viterbi algorithm.

Out of the several different methods, we use the one suggested by Bello and Pickens [4]. This means initializing the state transition parameters according to a double-nested circle of fifths and selecting the mean vectors and the covariance matrices on the basis of musical knowledge. When training the model with the EM algorithm, we train only the state distribution and transition parameters and leave the observation parameters untrained.

The 24-chord estimation provides a robust but slightly noisy representation of the harmonic content of the audio file. When observing the representations we noticed that the estimated chords were occasionally oscillating between major and minor chords of the same root note. This suggests that the third of the chord can harm the sequence labeling. Similar observation can be derived from the MIREX chord detection task where average overlap scores usually become better when the major and minor chords are merged (see for example the results of the MIREX Chord Detection Task 2009 [2]). This led us to an experiment with a 12-state HMM, where the triad of the chord is discarded from the chord templates. In the 12-state HMM, the initial parameters are set in a similar manner as with the 24-state HMM, but with respect to the simpler model and reduced chords. As such, the state sequence provided by the Viterbi algorithm can be seen as a "power chord" representation. Such representation is clearly too reduced and inaccurate to distinguish the versions on their own, but it seems to improve the identification performance when used in parallel with the 24-state HMM representation. In Figures 1 and 2 we display state sequences derived from a single audio file using 24- and 12-state HMMs, respectively.

3.2 Chromagram Flux

In addition to the chromagram vectors themselves, we experimented on whether the distance between subsequent chromagram vectors might have any effect. A somewhat similar approach was presented in [14], where a 12-dimension dynamic chroma vector feature called delta chroma was utilized. The delta chroma describes the degree of chroma changes on all possible intervals.

Here, we do not consider the delta chroma, but instead we calculate the distances between successive chroma vectors. A similar approach was utilized in [21], where correlation between adjacent chroma vectors was used as a feature in identification. We discovered empirically that the Manhattan distance (the city-block distance) had more distinguishing power for our work than Eucledian or cosine distances.

The Manhattan distances between chroma vectors of a musical piece can be seen as a time series. To discretize the time series, we use SAX (Symbolic Aggregate approXimation) [18]. In short, SAX discretizes the continuous values by first reducing their dimensionality using piecewise aggregate approximation and then discretizing the values according to a Gaussian curve. We chose SAX after experimenting with several quantization methods. Also, SAX has been used successfully for quantization when calculating similarity between time series using NCD [13].

Selection of the SAX parameters is not a trivial task. As we want to represent the whole chromagram flux as a string of characters, the sliding window is set to the length of the chromagram. The alphabet size and SAX accuracy parameters are more difficult to choose. We set the alphabet size to four and the number of frames per character to ten. These were chosen empirically, and thus are open to discussion.

3.3 Strongest Tone Sequence

The chromagram represents not only harmonic, but also melodic information contained in the audio file. We tested several methods to have more melodic information from the chromagram to be presented in a format suitable for NCD, but as with the chromagram quantization, different representations proved either to be too noisy or too reducing.

However, a straightforward way to represent some of the mid-level melodic information proved to increase the identification accuracy. We took the index of the strongest pitch class of a chroma vector (for a normalized chroma vector, the pitch class with the value of one), and represented the piece of music as a sequence of the strongest pitch class components. For a less densely orchestrated piece of music, this representation provides some information of the predominant melody of the piece. Even with more dense arrangements, it provides a representation that displays information different from the sequence labeling.

3.4 Transposition

Because cover versions are occasionally performed in a different key, the distance between chroma features can turn out large if key invariance is not addressed, even if the chroma features would otherwise be fairly similar. To obtain key invariance, a possible solution is to calculate distances between all 12 transpositions of the candidate version, but this is time-consuming. Another solution is to transpose the chromagrams into a common key using key estimation, but as with the tempo estimation, key estimation can fall short and lead to even worse identification results. We do not estimate the keys from the chromagrams,

[2] http://www.music-ir.org/mirex/2009/index.php/Audio_Chord_Detection_Results

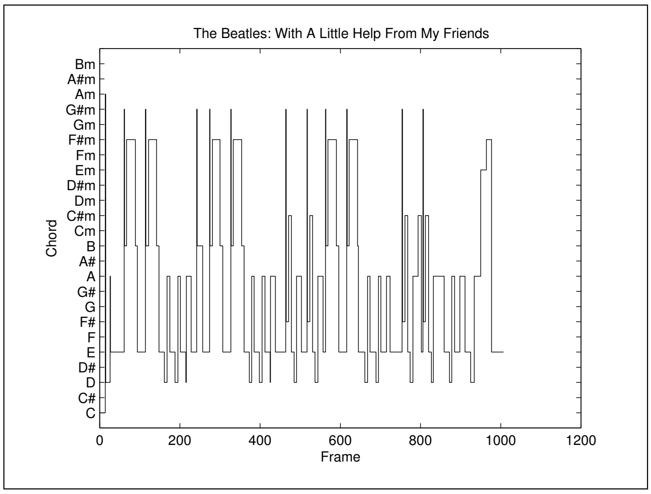

Figure 1. 24-state HMM Viterbi path for *With A Little Help From My Friends* performed by The Beatles.

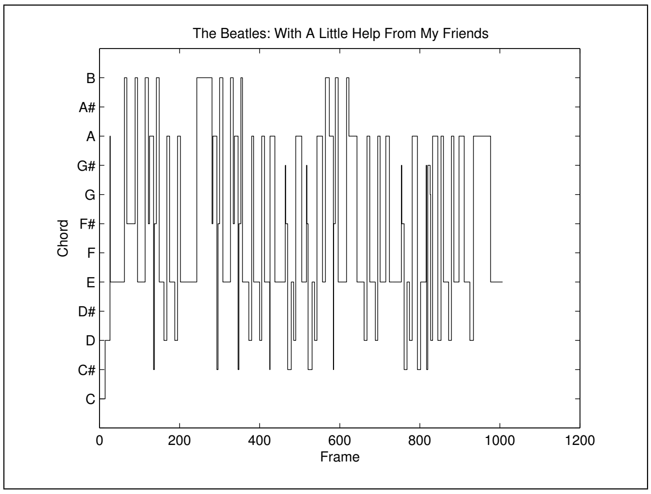

Figure 2. 12-state HMM Viterbi path for *With A Little Help From My Friends* performed by The Beatles.

but instead use Optimal Transposition Index (OTI) [19] to transpose the chroma sequences into a common key. In OTI, the transposition index is selected by first taking the global chroma vectors (by summing and normalizing the chroma vectors) of the two pieces of music. Then, the transposition index is selected by rotating the candidate global chroma vector 12 times and calculating the dot product between each pair of the target and candidate global chroma vectors. The rotation with the highest dot product is selected as the transposition index and the whole chromagram of the candidate is rotated according to the index. Fast and straightforward, OTI has also been proven to provide better identification accuracy than using the key estimation [19]. We apply OTI before any feature extraction.

3.5 Total Distance

After the distances between all the features of the pieces of music are calculated, the total distance for a pair of performances is obtained by simply taking the mean of all the feature distances. The distances could be weighted according to the importance of the features. To reduce the possible bias in the mean values caused by outliers, we also measured the total distance as the median of all measured feature distances.

4. EVALUATION

4.1 Test Data

To evaluate the performance of our approach, we collected a data set of original performances and their cover versions. For each original piece of music we included five cover versions. The data set included 25 such six-song sets and to complete the collection, a total number of 600 unique pieces of music were included, thus making the collection a total of 750 pieces of music with 150 possible queries.

The material was obtained from personal music collections and contains mostly western popular music, but with cover versions ranging from classical music renditions to world music and electronic versions. Apart from studio cover versions by different artists, the data set also includes live versions and a few remixes of the original versions. The complete detailed content of the data set can be requested from the authors.

4.2 Results

We used each of the 150 versions in the dataset as a query. From the output distance matrix, we calculated the total number of identified covers in the top five (TOP-5), the mean of average precisions (MAP) and the mean rank of the first identified cover (RANK). The results, using the mean as the total distance, are depicted in Table 1.

To present the effect of each different feature in the identification, we ran the algorithm for the whole test data set using only selected features of the feature set. The results for different feature sets, using the mean as the total distance, are depicted in Table 2.

The difference between using the mean and median values as the total distance is depicted in Figure 3. Generally, using the median as the total distance provided smaller distances. This suggests that outliers do exist in the feature distances and overall identification could be improved by taking them into account. However, using the mean as the total distance provided slightly better identification accuracy with a TOP-5 rating of 263 against the TOP-5 rating of 243 of the median distance.

4.3 Comparison to the LabROSA Cover Song Detection System

To see how well our approach performs in comparison with another cover version identification approach, we ran our test data with the LabROSA Cover Song Identification software [11]. To our knowledge, this is the only cover version identification application that is freely distributed and available online [3] .

[3] http://labrosa.ee.columbia.edu/projects/coversongs/

Measure	Value	Range
TOP-5	263	[0–750]
MAP	0.410	[0–1]
RANK	4.795	[1–745]

Table 1. Results of the 150 query evaluation.

Features used	TOP-5	MAP
24-state HMM	216	0.356
24- and 12-state HMM	242	0.378
HMMs and Chroma flux	249	0.399
All features	263	0.410

Table 2. The effect of combining different features.

The comparison between the results of our approach and the LabROSA application is depicted in Table 3. The results show that the performance of our application is comparable with the performance of the LabROSA system. However, we are aware that the LabROSA application was introduced several years ago and is possibly not comparable with some of the state-of-the-art approaches. For comparing the performance of our approach with more state-of-the-art approaches, we refer to the future MIREX cover song identification task where our application will be submitted.

5. CONCLUSIONS

We have presented an approach for cover version identification that combines different features derived from the chromagrams extracted from the audio files. To discretize continuous values, several techniques such as HMM and SAX have been used. The similarity between discretized features is calculated using a distance metric called normalized compression distance, which uses data compression to approximate the Kolmogorov complexities of the objects and as such is a quasi-universal, parameter-free similarity metric.

Based on the results, it is evident that the chroma feature combination together with the NCD metric can be used for cover version identification. As our results proved, combining different features and composing the final distance based on the distances between these features provides more accurate identification with the NCD. The algorithm was tested with competent results against a large data set consisting of various different kinds of versions from original performances.

The biggest obstacle for using normalized compression distance for cover version identification is the process of converting continuous features to discrete representations. Extracting features from audio is likely to yield noisy representations, and although NCD has been proved to be resistant against noise [8], it still affects the identification.

Our approach has more emphasis on the harmonic features, and observing the results supports this: pieces of music with distinctive, recognizable harmonic content are easily identified even when arrangements and structures vary. Also, as stated, we comprise the total distance simply as a mean of all distances, but this could be improved by weighting the different distances according to their relevance. Using the median as the total distance also gave a finding of the bias caused by the outliers.

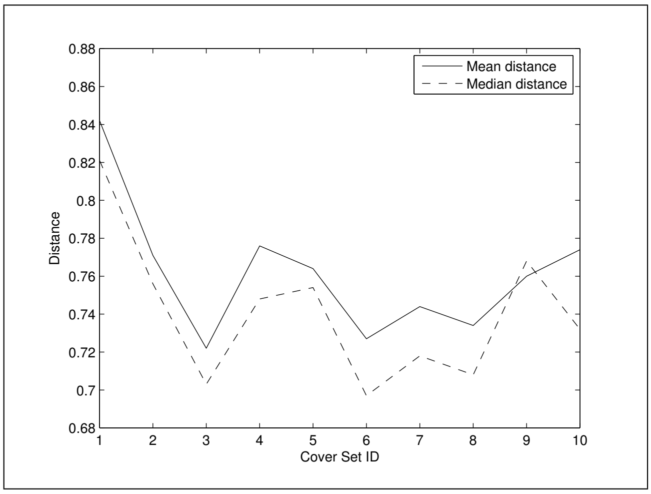

Figure 3. Mean distances for the first identified covers in ten 6-version cover sets using mean and median total distances.

Another issue demanding attention is that the phases of the measuring process each have a wide selection of parameters. Parameter selection is present in every phase of the identification process: from selecting the parameters of the Fourier transform when obtaining the chromagram to the choice of the compression algorithm used for calculating the NCD values. It is unclear if the parameters we have selected are optimal for the identification task and also the possibility of overfitting is evident. Future work addressing the parameter selection is under consideration.

5.1 Remark on Different Versions

As the cover version dataset also included live renditions and remixed versions of the original recordings, we took a closer look at the cases of these versions.

Live versions, either by the performers of the original versions or by a different performer, were in most cases identified very well. We see two reasons for this. First, live versions are often quite similar to original versions, having only slight modifications such as key, tempo or small structural differences (lengthier introductions or solo sections). Second, the live versions are less densely produced and arranged, whereas the studio versions are usually far more orchestrated. This makes the chroma features derived from live versions less noisy, which in turn benefits the similarity measuring. All in all, live version detection can be seen as a somewhat easier case of cover version identification. Thus, developing and testing cover version identification algorithms using predominantly live renditions may lead to slightly biased results.

Remixed versions, on the other hand, were often far more difficult to identify. In many cases, remixed versions

System	TOP-5	MAP
Our approach	263	0.410
LabROSA	256	0.405

Table 3. The results between our approach and LabROSA system.

share only limited elements similar to the original performance, usually combining audio elements of the original performances with completely different, and often electronic, instrumentation. Whereas live versions usually have very little changes in structures and a stripped-down instrumentation, the situation is often completely vice versa with remix versions: the original structure is often completely discarded and the instrumentation is usually even more dense than the original performance. We feel free to say that remix version identification is a far more difficult case of cover version identification. Thus, it would be interesting to see how well cover version identifiers perform when the task is specifically remix version identification. To our knowledge, version identification specialized in remix identification has been done only on a small scale [5].

6. ACKNOWLEDGEMENTS

This work has been supported by the Academy of Finland, grant #129909. The author would like to thank Kjell Lemström and Rainer Typke for their insightful comments on earlier versions of this paper. The author is also grateful to Päivi Suomalainen for her valuable help in collecting and organizing the test data.

7. REFERENCES

[1] T. E. Ahonen. Measuring harmonic similarity using PPM-based compression distance. In *WEMIS'09*, Corfu, Greece, 2009.

[2] J. P. Bello. Audio based cover song retrieval using approximate chord sequences: Testing shifts, gaps, beats and swaps. In *ISMIR'07*, Vienna, Austria, 2007.

[3] J. P. Bello. Grouping recorded music by structural similarity. In *ISMIR'09*, Kobe, Japan, 2009.

[4] J. P. Bello and J. Pickens. A robust mid-level representation for harmonic content in music signals. In *ISMIR'05*, London, UK, 2005.

[5] M. Casey and M. Slaney. Fast recognition of remixed music audio. In *ICASSP-07*, Hawaii, USA, 2007.

[6] Z. Cataltepe, Y. Yaslan, and A. Sonmez. Music genre classification using MIDI and audio features. *EURASIP Journal on Applied Signal Processing*, 2007(1), 2007.

[7] M. Cebrián, M. Alfonseca, and A. Ortega. Common pitfalls using the normalized compression distance: what to watch out for in a compressor. *Communications in Information and Systems*, 5(4):367–384, 2005.

[8] M. Cebrián, M. Alfonseca, and A. Ortega. The normalized compression distance is resistant against noise. *IEEE Transactions on Information Theory*, 53(5):1895–1900, 2007.

[9] R. Cilibrasi and P. M. B. Vitányi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, 2005.

[10] R. Cilibrasi, P. M. B. Vitányi, and R. de Wolf. Algorithmic clustring of music based on string compression. *Computer Music Journal*, 28(4):49–67, 2004.

[11] D. P. W. Ellis and G. E. Poliner. Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In *ICASSP-07*, Hawaii, USA, 2007.

[12] M. Helén and T. Virtanen. A similarity measure for audio query by example based on perceptual coding and compression. In *DAFx-07*, Bordeaux, France, 2007.

[13] E. Keogh, S. Lonardi, and C. A. Ratanamahatana. Towards parameter-free data mining. In *KDD'04*, Seattle, Washington, USA, 2004.

[14] S. Kim and S. Narayanan. Dynamic chroma feature vectors with applications to cover song identification. In *MMSP'08*, Cairns, Australia, 2008.

[15] O. Lartillot and P. Toiviainen. A MATLAB toolbox for musical feature extraction from audio. In *DAFx-07*, Bordeaux, France, 2007.

[16] M. Li and R. Sleep. Melody classification using a similarity metric based on Kolmogorov complexity. In *SMC'04*, Paris, France, 2004.

[17] M. Li and R. Sleep. Genre classification via an LZ78-based string kernel. In *ISMIR'05*, London, UK, 2005.

[18] J. Lin, E. Keogh, S. Lonardi, and B. Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *DMKD'03*, San Diego, California, USA, 2003.

[19] J. Serrà, E. Gómez, and P. Herrara. Transposing chroma representations to a common key. In *IEEE CS Conference on The Use of Symbols to Represent Music and Multimedia Objects*, 2008.

[20] J. Serrà, E. Gómez, and P. Herrara. *Audio Cover Song Identification and Similarity: Background, Approaches, Evaluation, and Beyond*. Springer-Verlab Berlin / Heidelberg, 2010.

[21] Y. Yu, M. Crucianu, V. Oria, and L. Chen. Local summarization and multi-level LSH for retrieving multivariant audio tracks. In *MM'09*, Beijing, China, 2009.