COGNITIVE FEATURES FOR COVER SONG RETRIEVAL AND ANALYSIS

Jan Van Balen, Frans Wiering, Remco Veltkamp
Dept. of Information and Computing Sciences, Utrecht University, the Netherlands
J.M.H.VanBalen, F.Wiering, R.C.Veltkamp@uu.nl

1. INTRODUCTION

This article presents the first results of a study on large-scale automatic identification of derivative works in early popular music. We present a music cognition-inspired approach to (audio) version detection that is (1) tailored for early 20th century recordings, (2) based on simple, indexable feature representations and (3) allows for interpretation of the resulting musical descriptions. For this purpose, two new descriptor types are introduced: pitch bihistograms and chroma correlation coefficients.

2. MOTIVATION

Musical heritage collections such as folk music archives may contain large numbers of closely related documents, such as exact duplicates of a recording, different renditions of a song, or loose variations on a theme. The particularities of such variations are of great interest in the study of music genealogies, oral transmission of music, and other aspects of music studies (Volk et al. [2012]).

The field of Music Information Retrieval has produced a lot of research on automatic systems for cover version identification. The large majority of these systems focuses on popular music from the last 50 years. An overview of version detection techniques is found in Serrà [2011].

Also, a large majority of systems is based on alignment. This means that, for the system to assess the similarity between two document, it needs to compute the optimal way of aligning a pair of descriptive time series for those two documents. In other words, when analyzing a complete collection of recordings, a computationally expensive comparison of time series is required for every pair of songs in that collection. For large datasets, this is far from practical, and on the scale of the type of corpus required to study patterns and trends in musical, infeasible: a full-blown comparison of 10,000 documents at 1 pairwise comparison per second would take over a year (578 days).

Outside of MIR's audio community, interesting work has been done on the analysis of folk songs and tune families based on their scores (Kranenburg [2010], Bohak & Marolt [2009]). Unfortunately, symbolic transcriptions of music are not always available.

In our study, we aim to construct and test a cognition-inspired type of music features that is global and indexable, can be computed from (polyphonic) audio, and allows to reliably identify derivative works in a collection of recordings. Additionally, we want the features to be interpretable and easy to extend for use in analysis (e.g. clustering songs into genres, creating a phylogenetic tree of music pieces or quantifying the schematic musical expectations in a corpus). Interpretability is an important necessary requirement for applications in analysis, and a well-known issue with many currently used features Aucouturier & Bigand [2013]. In this study, we aim to address this by drawing inspiration from descriptors used in music cognition.

3. FEATURES

We propose a pair of simple, audio-based analogues to the well-known bigram representation often used in symbolic music processing. The first representation relates to melody and approximates a histogram of pitch bigrams (pairs of consecutive pitches) weighted with duration. The second, harmony-related representation is an abstraction of the co-occurrence matrix of chord notes in a song.

Many authors have proposed analyses based on pitch bigrams, most of them from the domain of cognitive science or otherwise interested in the information dynamics at play in music (Li & Huron [2006], Müllensiefen & Frieler [2006], Rodriguez Zivic et al. [2013]). This is not surprising: distributions of bigrams effectively encode a form of first-degree expectations: if the relative frequency of bigrams in a piece is conditioned on the first pitch in the bigram, we obtain the conditional frequency of a pitch given the one preceding it. Expectations of this kind have been linked to melodic complexity, familiarity and even preference ratings (Huron [1999]).

Marginalized and non-marginalized variants have been proposed for pitch events corresponding to pitch heights, durations, pitches with height and duration, pitch intervals, and scale degrees. The first new feature we introduce will follow the latter paradigm: it is a distribution over pairs of (chromatic) scale degrees $\{1, 2, \dots 12\}$. It will be referred to as the **pitch bihistogram**, a bigram representation that can be computed from continuous pitch data.

Assume that a pitch time series $P(t)$, quantized to semitones and folded to one octave, can be obtained. If a pitch histogram is defined as:

$$H(p) = \sum_{P(t)=p} \frac{1}{n},$$

with n the length of the time series and $p \in \{1, 2, \dots 12\}$,

the pitch bihistogram is then defined:

$$B(p_1, p_2) = \sum_{\substack{P(t_1)=p_1 \\ P(t_2)=p_2}} w(t_2 - t_1)$$

where

$$w(x) = \begin{cases} \frac{1}{d}, & \text{if } 0 < x < d. \\ 0, & \text{otherwise.} \end{cases}$$

and d is the size of the look-ahead window. In this study we will consistently use pitch contours that have been aligned to an estimate of the piece's overal tonic.

The second feature representation we propose focuses on vertical rather than horizontal pitch relations.

$$C(p_1, p_2) = \text{corr}(c(t, p_1), c(t, p_2)),$$

where $c(t, p)$ is a 12-dimensional chroma time series (also known as pitch class profile) computed from the song audio (Gomez [2006]). From this chroma representation of the song $c(t, p)$ we compute the correlation coefficients between each pair of chroma dimensions to obtain a 12×12 matrix of **chroma correlation coefficients** $C(p_1, p_2)$. Again, the chroma features are all transposed to the same tonic (e.g. A) based on an estimate of the song's overall key.

For key detection, a global chroma feature is computed from a full chroma representation of the song. This global profile is then correlated with all 12 modulations of the standard diatonic profile to obtain the tonic. The binary form (ones in the 'white key' positions and zeros in the others) is used. For this study, pitch contours and chroma features where computed using Melodia (by Salamon & Gomez [2010]), and HPCP (Gomez [2006]) respectively, with default settings. [1] For efficiency in computing the pitch bihistogram, the pitch contour was median-filtered and downsampled to 10 frames / s.

4. DATASET AND METHOD

The above features were tested on a set of 150 recordings digitized especially for this study: 100 45-rpm records from the 50's and 60's, and 50 78-rpm records, most of them from before 1950. The corpus they belong to is a real-world 'popular music heritage' collection that was, until recently, only accessible through manually transcribed metadata (such as titles, artists, original title, composer).

Amongst these records are 50 pairs of songs that correspond to the same composition. All of these tunes are translated covers or interpretations of melodies with a different text. Such songs are especially interesting since they garantee some deviation from the source, which is desirable when models of music similarity are tested.

A retrieval experiment on these songs was carried out: for each song that is part of a pair, the song was taken out of the corpus and used as a query to which all the remaining 149 songs were ranked (e.g. using a cosine distance). The rank r of the other song of the pair then determined the precision $p = \frac{1}{r}$. This was done 100 times to obtain the Mean Average Precision (MAP). A random baseline for this task was established at MAP $= 0.036$, with a standard deviation of 0.010 over 100 randomly generated distance matrices.

[1] mtg.upf.edu/technologies

Weights:	H	B	C	MAP
	1	0	0	0.27
	0	1	0	0.43
	0	0	1	0.42
	0	1	1	0.53

Table 1: Overview of results. H = pitch histogram, C = chroma correlation coefficients, B = pitch bihistogram.

The main experiment parameters for the features described above were d, the look-ahead window of the bihistogram, and the weighting of each of the three features when all are combined. d was set to $0.500s$ after a quick optimisation. The weights were restricted to 0 and 1 (feature used or not used) as can be seen in the results summary in Table 1.

5. RESULTS

The precision obtained using only pitch histograms is already substantial, about 0.27. However using only the pitch bihistogram feature, a MAP of around 0.43 can be obtained, compared a very competitive 0.42 for using just the chroma correlations. Finally, when the last two features are combined, the MAP goes up to 0.53. In the latter configuration, 44 of the 100 queries retrieve their respective cover version in first place: the 'precision at 1' is 0.440.

A thorough comparison with existing version detection systems will be performed later in this study, but a survey of reported performances of cover detection systems shows that precisions of this order of magnitude make a promising start. For example, the intervalgram method by Walters et al. [2013] achieves a precision at 1 performance of 0.538 on the covers80 dataset (160 songs). This method uses an indexable representation for pruning the candidate set but relies on alignment for the steps thereafter. Methods relying only on pairwise comparison have reported both higher and lower precisions on the same dataset.

In the current experiments, no indexing or look-up was performed. For actual indexing, the proposed features would have to be quantized and reduced in dimensionality. Locality Sensitive Hashing does precisely this, as also described in Walters et al. [2013]. We refer to Slaney et al. [2012] for an in-depth account of the relationship between LSH and high dimensional distances. The features we propose are indeed very suitable for such applications.

6. CONCLUSIONS AND FUTURE WORK

In this abstract, we have proposed two new features for the description and retrieval of early popular music: pitch bihistogram features and chroma correlation coefficients. The features are evaluated on a newly compiled dataset

of variations and translations in early popular music. We demonstrate a promising performance of MAP $= 0.53$. A quantitive comparison to state-of-the-art indexable representations and alignment-based retrieval systems will be carried out in the future. Other future work includes the separate evaluation and optimisation of the pitch extraction and key extraction components. Finally, we hope to include a representation of rhythm in the model.

7. ACKNOWLEDGEMENTS

The authors would like to thank Dimitrios Bountouridis and Marcelo Rodriguez.

8. REFERENCES

Aucouturier, J.-J. & Bigand, E. (2013). Seven problems that keep MIR from attracting the interest of cognition and neuroscience. *Journal of Intelligent Information Systems*, *41*(3), 483–497.

Bohak, C. & Marolt, M. (2009). Calculating similarity of folk song variants with melody-based features. In *Proc Int Society Music Information Retrieval Conf (ISMIR)*, number Ismir, (pp. 597–601). Citeseer.

Gomez, E. (2006). *Tonal Description of Music Audio Signals*. PhD thesis, Universitat Pompeu Fabra.

Huron, D. (1999). Musical Expectation. In *The 1999 Ernest Bloch Lectures*.

Kranenburg, P. v. (2010). *A Computational Approach to Content-Based Retrieval of Folk Song Melodies*. PhD thesis, Utrecht University.

Li, Y. & Huron, D. (2006). Melodic modeling: A comparison of scale degree and interval. In *Proc. of the Int. Computer Music Congerence*.

Müllensiefen, D. & Frieler, K. (2006). Evaluating different approaches to measuring the similarity of melodies. *Data Science and Classification*.

Rodriguez Zivic, P. H., Shifres, F., & Cecchi, G. a. (2013). Perceptual basis of evolving Western musical styles. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(24), 10034–8.

Salamon, J. & Gomez, E. (2010). Melody extraction from polyphonic music signals using pitch contour characteristics. In *IEEE Trans. on Audio, Speech and Language Processing*.

Serrà, J. (2011). *Identification of versions of the same musical composition by processing audio descriptions*. PhD thesis, Universitat Pompeu Fabra.

Slaney, M., Lifshits, Y., & He, J. (2012). Optimal parameters for locality-sensitive hashing. *Proceedings of the IEEE*, *100*(9), 2604–2623.

Volk, A., de Haas, B., & van Kranenburg, P. (2012). Towards modelling variation in music as foundation for similarity. In *Proc. 12th Int. Conf. Music Perception and Cognition*.

Walters, T. C., Ross, D. A., & Lyon, R. F. (2013). The intervalgram: An audio feature for large-scale cover-song recognition. In *From Sounds to Music and Emotions: 9th International Symposium, CMMR 2012, London, UK, June 19-22, 2012, Revised Selected Papers* (pp. 197–213).