Transposing Chroma Representations to a Common Key

Joan Serrà, Emilia Gómez and Perfecto Herrera

***Abstract*—Chroma representations of musical excerpts are nowadays very popular and used for a wide variety of applications. Within these, transposition to a common key or tonality can represent an important aspect, usually having a dramatic impact to final system's accuracy. We present and evaluate a new and straightforward way for transposing two chroma representations to a common key that outperforms previous methods based on key estimation and that, without detriment of accuracy, is computationally faster than trying all possible transpositions. In addition, we also provide some insights into the internal organization of this new tool, suggesting it would organize transposition indices in a coherent manner.**

***Index Terms*—Music, Information retrieval, Acoustic signal analysis, Multidimensional sequences, Symbol manipulation**

I. Introduction

TRANSPOSING musical excerpts to a common key or tonality is a necessary feature when comparing melodies, harmonies or any tonal representation of these musical excerpts. This process is specially crucial in many music information retrieval (MIR) tasks related to music similarity such as audio matching and alignment [1], [2], song structure analysis [3] or cover song identification [4], [5], where melodic or harmonic representations of musical excerpts are used. Furthermore, this is a necessary feature for any music retrieval or recommendation engine comparing tonal information.

Chroma features or pitch class profiles (PCP) have become very popular and widely used among these and many other MIR-related tasks (e.g. key and chord estimation [6], [7]) as they provide a description of the audio tonal content that, ideally [8], (a) represents the pitch class distribution of both monophonic and polyphonic signals, (b) considers the presence of harmonic frequencies, (c) is robust to noise and non-tonal sounds, (d) is independent of timbre and played instrument, (e) is independent of loudness and dynamics and (f) is independent of tuning, so that the reference frequency can be slightly different from the standard A 440 Hz. Chroma features (figure 1) are derived from the energy found within a given frequency range (typically from 50 to 5000 Hz) in short-time spectral representations (e.g. 100 msec) of audio signals extracted on a frame-by-frame basis. This energy is usually collapsed into an octave-independent histogram representing the presence (or relative intensity) of each of the 12 semitones of an equal-tempered chromatic scale.

The authors are with the Music Technology Group, Universitat Pompeu Fabra, Ocata 1 (3rd floor), 08003 Barcelona, Spain, phone: +34-93-542-2864, fax: +34-93-542-2202, e-mail: {jserra,egomez,pherrera}@iua.upf.edu

This research has been partially funded by the EU-IP project PHAROS IST-2006-045035 (http://www.pharos-audiovisual-search.eu) and the e-Content Plus project VARIAZIONI ECP-2005-CULT-038264 (http://www.variazioniproject.org).

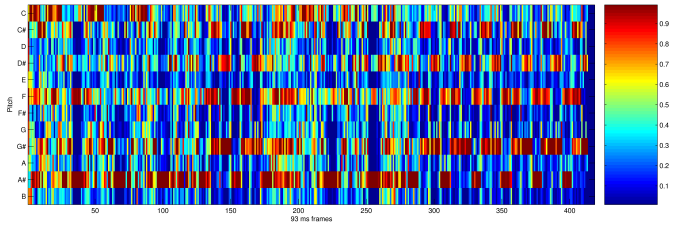

Fig. 1. Example of a chroma feature (vertical) sequence extracted on a frame basis (horizontal) from the song "Roxanne" by *The Police*.

In many applications involving the comparison of musical excerpts, key invariance is achieved by processing all possible relative transpositions between two chroma representations (brute-force method). This is a common strategy followed in many music retrieval or cover song identification systems (e.g. [4], [9]). Alternatively, one may automatically estimate the main key for the musical excerpts being compared and then normalize their chroma representations with respect to this (e.g. [10], [11]). In [12], [5], we have introduced a new and very straightforward way for transposing chroma representations to a common key based on what we call the optimal transposition index (OTI).

In this paper we extend this proposed approach to consider multiple transposition indices and test the effect that these might have in final system's performance. In addition, as an exhaustive comparison between OTI-based transposition methods and alternative approaches has not been properly done, we include in our evaluation the most common methods found in the literature. Finally, the internal organization of these transposition indices is highlighted.

II. Experiment

A. Cover song identification

Cover songs (or versions) consist in different performances of the same underlying musical piece. A change of the main tonality of the song is a common feature between its different covers. This is usually done to adapt the original composition to a different singer or solo instrument, or just for 'aesthetic' reasons. Transposition to a common key has been elucidated to be a very important feature for any cover song identification system, providing a deep impact on final system's accuracy (up to 17% difference in standard evaluation measures, depending on the method chosen [5]). Therefore, it seems appropriate to study the effect of different transposition methods on this task. To do so, we use the same algorithm described in [12], [5] (figure 2) but with new modifications in the OTI generation and song transposition modules that will be explained in the

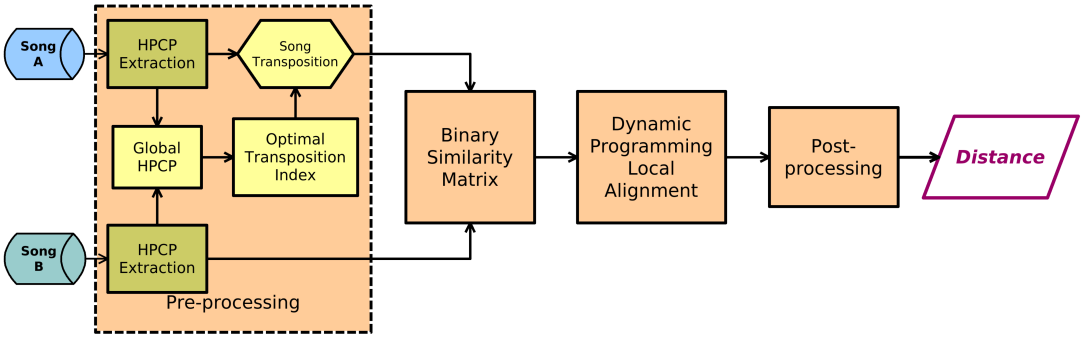

Fig. 2. Block diagram of the cover song identification system.

following subsection. We now briefly summarize the overall cover song identification system while emphasizing the part under test: chroma transposition. It is not the objective of this paper to discuss the details about our cover song identification algorithm since they have been extensively exposed in the literature, so, for further details, we refer the interested reader to [12], [5].

The system starts computing harmonic pitch class profiles (HPCP) [8]. The HPCP is an enhanced pitch class distribution (or chroma) feature, computed in a frame-by-frame basis using only the local maxima of the spectrum within a certain frequency band. In addition, HPCPs are tuning independent (so that the reference frequency can be different from the standard A 440 Hz), and they consider the presence of harmonic frequencies. An HPCP sequence (e.g. figure 1) is extracted for each of the two songs being compared. With this sequence, we compute what we call a global HPCP by averaging all feature vectors in a sequence as:

$$\overrightarrow{g_A} = \frac{\sum_{i=1}^{N} \overrightarrow{h_{A,i}}}{max\{\sum_{i=1}^{N} \overrightarrow{h_{A,i}}\}} \quad (1)$$

where $\overrightarrow{g_A}$ represents the global HPCP for song A, $\overrightarrow{h_{A,i}}$ corresponds to the extracted HPCP vector for frame i in song A, and N is the total number of frames. The result is normalized by the maximum in order to have values between 0 and 1. An analogous process is followed for song B ($\overrightarrow{g_B}$).

With the global HPCPs for the two songs we calculate the optimal transposition index (OTI). This represents the number of positions that an HPCP vector needs to be circularly shifted to have maximal resemblance to the other, where resemblance is measured by the dot product ('$\cdot$'):

$$OTI(\overrightarrow{g_A}, \overrightarrow{g_B}) = \underset{1 \leq j \leq M}{\operatorname{argmax}} \{\overrightarrow{g_A} \cdot Circshift_R(\overrightarrow{g_B}, j-1)\} \quad (2)$$

M is the HPCP size considered (usually the 12 semitones of the equal tempered scale), and $Circshift_R(\overrightarrow{v}, j)$ is a function that rotates a vector ($\overrightarrow{v}$) j positions to the right. A circular shift (to the right) of one position is a permutation of the entries in a vector where the last component becomes the first one and all the other components are shifted to the right.

Then, to transpose one song to the key of the other one, for each HPCP vector i in the whole sequence we compute:

$$\overrightarrow{h_{A,i}^{Tr}} = Circshift_R(\overrightarrow{h_{A,i}}, OTI) \quad (3)$$

where superscript ^{Tr} denotes musical transposition.

After one chroma sequence is transposed to the key of the other one, a binary similarity matrix is computed. This is used as a local cost function for a dynamic programming local alignment (DPLA) algorithm (figure 2), which finds the best subsequence matches between all possible ones while considering tempo deviations and sequence gaps. The best local alignment of the two songs is finally used to obtain a dissimilarity measure between them, which will assess us in knowing if they are covers or not.

B. Transposition methods

To further explore the capabilities of OTI-based transposition, we modify expressions 2 and 3 in order to account for more than one transposition option. We compute the dot products between $\overrightarrow{g_A}$ against all possible circular shifts of $\overrightarrow{g_B}$ and we store the results in a resemblance array $\overrightarrow{R}$ such that:

$$R_j(\overrightarrow{g_A}, \overrightarrow{g_B}) = \overrightarrow{g_A} \cdot Circshift_R(\overrightarrow{g_B}, j-1) \quad (4)$$

for $j = 1, \ldots, M$, where M is again the HPCP size considered. R_j represents the similarity between song's A global HPCP ($\overrightarrow{g_A}$) and the $j-1$-th circularly shifted song's B global HPCP ($\overrightarrow{g_B}$). We sort this array in descending order and store the permutation indices to obtain a list of transposition indices ranked from best to worse:

$$\overrightarrow{OTI} = PermutationIndices(Sort_{Desc}(\overrightarrow{R})) \quad (5)$$

It is now easy to see that the optimal transposition index defined in equation 2 corresponds to the first position of the array (OTI_1), and that all possible transposition indices are contained in $\overrightarrow{OTI}$ (length M). Next, as OTI has changed from a single value to an array of values (from equation 2 to 5), we have to modify expression 3 in order to be able to consider different transposition indices:

$$\overrightarrow{h_{A,i}^{Tr}} = Circshift_R(\overrightarrow{h_{A,i}}, OTI_k) \quad (6)$$

and the rest of the method summarized in section II-A is done as many times as the total number transposition indices k considered (NTIC). The best similarity measure between transposed chroma representations is kept as the final decision. In this way, if NTIC=1, we are computing a cover song similarity measure for just one transposition (the OTI-based transposition, OTI_1) and, if NTIC=M, we compute a cover song similarity measure for all possible transpositions ($OTI_{1,\cdots,M}$). Consequently, we are able to test two methods for transposing chroma representations to a common key: OTI-based transposition and brute-force method. Furthermore, we are able to test intermediate steps such as NTIC=$2,\ldots,M-1$ ($OTI_{2,\ldots,M-1}$) and the effect of no transposition (NTIC=0). In addition, as baseline for subsequent evaluations, we also include in our tests a random $\overrightarrow{OTI}$ generator, which aleatorically generates transposition indices.

The remaining alternative for chroma transposition (as summarized in section I) consists on using a key estimation algorithm and then transposing the song to a predefined key (e.g., C major or A minor). Therefore, a method for musical excerpts comparison can be applied without any transposition post-processing. In the case of our cover song identification system, equations 1 to 3 wouldn't be necessary, and that's what we do in our study for this particular transposition method. To automatically estimate the key we use a state-of-the-art algorithm [6], [8] which had an accuracy of 75% for real audio pieces, and scored among the first classified algorithms in the MIREX 2005 contest[1] with an accuracy of 86% with synthesized MIDI files.

C. Evaluation

We evaluate the transposition methods explained in section II-B with a music collection of 90 songs comprising 15 groups of 6 covers each (the original one + 5 covers). This database has been used in previous experiments [12], [5] and contains several cover songs that change the key with respect to the original ones. Although we don't have the key manually annotated and validated for all the songs in the database, we will see in section III that a qualitative idea of the number of transpositions found in the database can be obtained.

We query all the songs in the music collection and obtain a 90×90 distance matrix that is further processed to obtain several evaluation measures. Here, for assessing identification accuracy, we report results for recall (R) and mean average precision (MAP) [13]. Since we have a maximum number of 5 possibly retrieved covers per query, we report the recall achieved within the first 5 retrieved items (R@5). These two measures are widely used for evaluating many information retrieval systems. In particular, they are used by the Music Information Retrieval Evaluation eXchange (MIREX[2]), an international initiative to develop formal, common evaluation standards for MIR, where a cover song identification task has been run in 2006 and 2007[3].

[1]http://www.music-ir.org/mirex/2005/index.php/Audio_and_Symbolic_Key_Finding

[2]http://music-ir.org/mirexwiki

[3]http://www.music-ir.org/mirex/2007/index.php/Audio_Cover_Song_Identification_Results

III. Results

In table I we show the general accuracies for the different transposition variants tested. We can appreciate that all transposition methods improve the accuracy of the cover song identification task (up to relative values higher than 40% compared with simply not considering any transposition). The key estimation method performs worse among the three transposition methods tested. This might be due to the fact that automatic key estimation algorithms are not completely reliable, what, for sure, introduces errors to our cover song identification engine. Furthermore, as we query all songs against all, these errors might be propagated among queries (if we fail in determining the key of one song, we will not retrieve the covers of it, and neither retrieve it as a cover of others). As expected, the brute-force method (trying all possible transpositions) presents the best accuracy, followed by the OTI-based transposition method.

TABLE I
IDENTIFICATION ACCURACY FOR DIFFERENT TRANSPOSITION VARIANTS TESTED.

Transposition method	R@5	MAP
Random transposition	0.131	0.164
No transposition	0.478	0.516
Key estimation	0.496	0.532
OTI-based transposition	0.653	0.698
Brute-force method	0.682	0.729

To evaluate the capabilities of the new transposition method proposed in equations 4, 5 and 6, we tested all possible number of considered transposition indices (NTIC). The results for R@5 and MAP are plotted in figure 3. Note that just considering two transpositions (NTIC=2), we are able to achieve the same accuracy than with the brute-force method (NTIC=12). Thus, instead of computing all possible cost matrices and alignments, we just have to compute the ones corresponding to the two most probable or optimal transposition indices. This is quite remarkable as we decrease by a factor of 6 the number of operations done by the cover song identifier.

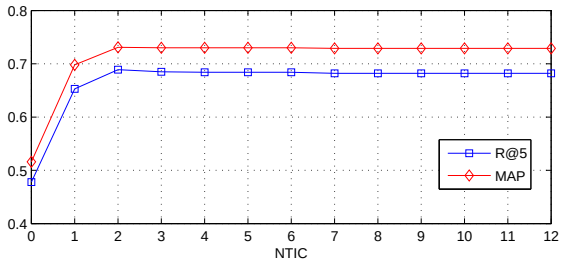

Fig. 3. R@5 and MAP curves of the cover song identification algorithm for all possible number of transposition indices considered (NTIC).

A histogram of the optimal transposition index (OTI_1), reveals that the most frequent transposition index is 0 (no transposition), followed by two semitone transpositions (b2 and b7). These are very common transpositions between covers, specially for adapting a song to a different singer. Other common intervals for transposing songs are the fifth and the fourth.

In order to understand the information provided by the OTI array (equations 4 and 5) and the rise in accuracy from NTIC=1 to NTIC=2 (from NTIC=$3, \ldots, 12$ the accuracies are the same, see figure 3), we study the relationship between the first transposition index and the other ones. To do that, we plot a histogram of the intervals between the different transposition options OTI_k and OTI_1 (figures 4 and 5). That is, for instance, if OTI_2 is 10 and OTI_1 was 7, this corresponds to a 3 semitone interval (a minor third, b3).

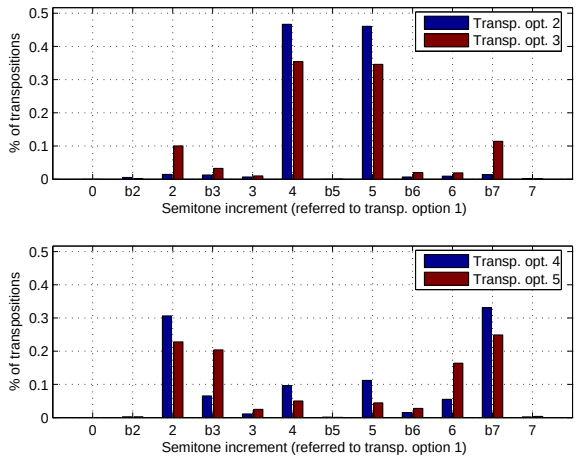

Fig. 4. Histograms of interval differences between OTI_2 and OTI_3 referred to OTI_1 (upper panel) and between OTI_4 and OTI_5 referred to OTI_1 (lower panel).

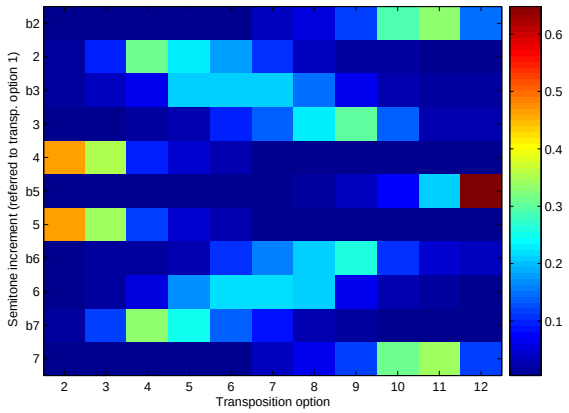

Fig. 5. Histogram (vertical axis) array of interval differences between OTI_k and OTI_1. Percentage of transpositions is indicated in the right-hand side colorscale.

In figure 4 we can see that the most prominent relation between OTI_2 and OTI_3 with respect to OTI_1 is a fifth or a fourth interval (which can also be considered a fifth lower) more than 70% of the time. We can also see that for OTI_4 and OTI_5 these percentages are more spreaded among different intervals (major second and minor seventh being the most prominent ones). A set of these histograms for NTCI=$2, \ldots, 12$ can be seen in figure 5. In there, we can appreciate the internal organization of the OTI array and the relation between OTI_1 and subsequent transposition options. We can again see that the second and third transposition options correspond to fifth and fourth intervals in respect to the firstly proposed one. This is quite remarkable if we think that most common mistakes in automatic key extraction algorithms (and even in human judgements) are related to these intervals [8]. In addition, we note that the worst ranked transposition options basically correspond to very dissonant intervals (minor second, major seventh and augmented fourth). With this, we can hipotesize that the OTI array ($\overrightarrow{OTI}$) would have the ability to arrange transposition indices in an 'intelligent' manner from most to less probable ones [14].

IV. Conclusion

We have presented a new, fast and straightforward way of transposing two chroma representations to a common key that outperforms a state-of-the-art key transposition method. Furthermore, we have demonstrated that as much accuracy as the brute-force method can be reached with 6 times less computational effort. Finally, we have looked at the organization of these transposition indices and shown that they are 'coherently' sorted, which could be the object of further studies.

References

[1] N. Hu, R. B. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," *IEEE Workshop on Apps. of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 185–188, 2003.

[2] M. Müller, F. Kurth, and M. Clausen, "Audio matching via chroma-based statistical features," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 288–295, 2005.

[3] B. S. Ong, E. Gómez, and S. Streich, "Automatic extraction of musical structure using pitch class distribution features," *Workshop on Learning the Semantics of Audio Signals (LSAS)*, pp. 53–65, 2006.

[4] D. P. W. Ellis and G. E. Poliner, "Identifying cover songs with chroma features and dynamic programming beat tracking," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 1429–1432, April 2007.

[5] J. Serrà, E. Gómez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Trans. on Audio, Speech and Language Processing*, July 2008, in press.

[6] E. Gómez and P. Herrera, "Estimating the tonality of polyphonic audio files: cognitive versus machine learning modelling strategies," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 92–95, 2004.

[7] A. Sheh and D. P. W. Ellis, "Chord segmentation and recognition using em-trained hidden markov models," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 183–189, 2003.

[8] E. Gómez, "Tonal description of music audio signals," Ph.D. dissertation, MTG, Universitat Pompeu Fabra, Barcelona, Spain, 2006.

[9] J. P. Bello, "Audio-based cover song retrieval using approximate chord sequences: testing shifts, gaps, swaps and beats," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 239–244, September 2007.

[10] Ö. Izmirli, "Tonal similarity from audio using a template based attractor model," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 540–545, 2005.

[11] M. Marolt, "A mid-level melody-based representation for calculating audio similarity," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 280–285, 2006.

[12] J. Serrà and E. Gómez, "Audio cover song identification based on sequences of tonal descriptors," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2008.

[13] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. ACM Press Books, 1999.

[14] D. Temperley, *The cognition of basic musical structures*. MIT Press, 2001.