TRANSPOSITION-INVARIANT SELF-SIMILARITY MATRICES

Meinard Müller and Michael Clausen
Bonn University
Department of Computer Science III

ABSTRACT

Self-similarity matrices have become an important tool for visualizing the repetitive structure of a music recording. Transforming an audio data stream into a feature sequence, one obtains a self-similarity matrix by pairwise comparing all features of the sequence with respect to a local cost measure. The basic idea is that similar audio segments are revealed as paths of low cost along diagonals in the resulting self-similarity matrix. It is often the case, in particular for classical music, that certain musical parts are repeated in another key. In this paper, we introduce the concept of a transposition-invariant self-similarity matrix, which reveals the repetitive structure even in the presence of key transpositions. Furthermore, we introduce an associated transposition index matrix displaying harmonic relations within the music recording. As an application, we sketch how our concept can be used for the task of audio structure analysis.

1 INTRODUCTION

The general concept of self-similarity matrices, which has been introduced to the music context by Foote [3], reveals the repetitive structure of a time-dependent data streams. One first transforms a given audio recording into a sequence $V := (v_1, v_2, \ldots, v_N)$ of feature vectors $v_n \in \mathcal{F}$, $1 \leq n \leq N$, where $\mathcal{F}$ denotes a suitable feature space (e. g., a space of spectral, MFCC, or chroma vectors). Then, based on a suitable local cost measure $c : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$, one forms an N-square *self-similarity matrix* $\mathcal{S}$ defined by $\mathcal{S}(n, m) := c(v_n, v_m)$, $1 \leq n, m \leq N$, comparing all features in a pairwise fashion. The crucial observation is that a pair of similar segments in the audio recording is revealed as a path of low cost along diagonals in the resulting self-similarity matrix.

As the running example of this paper, we consider the first movement of Beethoven's piano sonata Op. 31, No. 2 ("Tempest") in a recording by Barenboim. The rough musical form of this movement is given by $A_1A_2BA_3C$, where A_1 corresponds to the exposition (measures 0–90), A_2 to the repetition of the exposition, B to the development (measures 93–142), A_3 to the recapitulation (measures 143–217), and C to a short coda (measures 218–228). The musical parts A_1 and A_2, which are mere repetitions in the score, are played by Barenboim in the same fashion and correspond to the time intervals $[0 : 124]$ and $[130 : 251]$ (measured in seconds) of the recording, respectively. However, even though A_3 semantically corresponds to A_1, there are significant variations in structure and key. A musical analysis shows that A_1 has the substructure $A_1 = R_1S_1T_1U_1$, where R_1 represents the first measure, S_1 measures 2–7 (part of the first theme), T_1 measures 8–40 (continuation of the first theme and the transfer to the second theme), and U_1 measures 41–90 (second theme). Similarly, one has substructures $A_2 = R_2S_2T_2U_2$ and $A_3 = R_3X_3S_3T'_3U_3$. Here, the three R- and S-parts more or less coincide. Similarly, the three U-parts closely correspond to each other, however, with one difference: U_3 is a modulated version of U_1 transposed five semitones upwards (and later transposed seven semitones downwards). Furthermore, A_3 contains an additional part X_3 and part T'_3 significantly differs from its counterpart T_1 in structure and key.

A conventional self-similarity matrix as shown in Figure 1 (a) (with respect to chroma-based audio features as discussed in Section 2), reveals only parts of the musical structure. In particular, the path starting at coordinate $(0, 130)$ and ending at $(124, 251)$ indicates the similarity of the time intervals $[0 : 124]$ (part A_1) and $[130 : 251]$ (part A_2). Similarly, there are paths reflecting the similarity of the three R- and S-parts. However, repetitive segments that differ by some transposition are not reflected by the self-similarity matrix.

In Section 2, we introduce the concept of transposition-invariant self-similarity matrices that are invariant under all transpositions. In particular, we adopt an idea by Goto [4], which is based on the observation that the transpositions can be handled by cyclically shifting the chroma. Here, the *chroma* correspond to the twelve traditional pitch classes of the equal-tempered scale [1]. In Section 3, we sketch how the transposition-invariant self-similarity matrices can be used for automated audio structure analysis. In Section 4, we conclude this paper and give prospects on future work. Further references to related work are given in the respective sections.

2 TRANSPOSITION-INVARIANT SELF-SIMILARITY MATRIX

The properties of a self-similarity matrix $\mathcal{S}$ depend on the kind of audio features extracted from the audio recording as well as on the local cost measure c. In the following, we use chroma-based audio features as described,

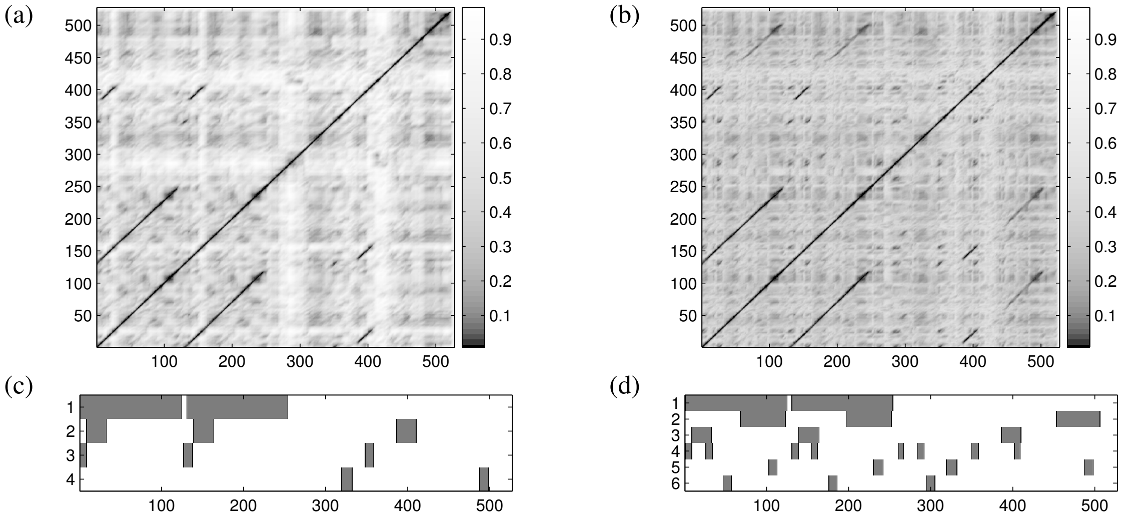

Figure 1. First movement of Beethoven's piano sonata Op. 31, No. 2 ("Tempest") in a recording by Barenboim. **(a)** Self-similarity matrix $\mathcal{S}$. Low costs are indicated by dark colors (cost 0 corresponds to black) and high costs by light colors (cost 1 corresponds to white). **(b)** Transposition-invariant self-similarity matrix $\sigma^{\min}(\mathcal{S})$. **(c)** Groups of mutually similar audio segments obtained from $\mathcal{S}$. **(d)** Groups of mutually similar audio segments obtained from $\sigma^{\min}(\mathcal{S})$.

e. g., in [1, 4, 6]. Assuming the equal-tempered scale, the chroma correspond to the set $\{\mathrm{C}, \mathrm{C}^\sharp, \mathrm{D}, \ldots, \mathrm{B}\}$ that consists of the twelve pitch spelling attributes as used in Western music notation. Note that in the equal-tempered scale different pitch spellings such $\mathrm{C}^\sharp$ and $\mathrm{D}^\flat$ refer to the same chroma. We consider the feature space

$$\mathcal{F} := \left\{ v \in \mathbb{R}^{12} \;\middle|\; \textstyle\sum_{i=1}^{12} v(i)^2 = 1 \right\}$$

of normalized 12-dimensional chroma vectors $v = (v(1), v(2), \ldots, v(12))$, where $v(1)$ corresponds to chroma C, $v(2)$ to chroma $\mathrm{C}^\sharp$, and so on. Then, the given audio signal is decomposed into a sequence $V = (v_1, v_2, \ldots, v_N)$ of normalized chroma vectors $v_n \in \mathcal{F}$, $1 \leq n \leq N$, which expresses the signal's local energy distribution among the 12 pitch classes. Such a chroma representation can be obtained, e. g., from a spectrogram by suitably pooling Fourier coefficients [1] or by using multirate filter bank techniques [6]. Chroma-based audio features absorb variations in parameters such as dynamics, timbre, and articulation and closely correlate to the short-time harmonic content of the underlying audio signal. In the following, we use a feature sampling rate of 1 Hz, i. e., each vector corresponds to one second of the original audio signal.

Furthermore, we use the local cost measure $c : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ defined by $c(v, w) := 1 - \langle v, w \rangle$ for $v, w \in \mathcal{F}$. Since v and w are normalized, the inner product $\langle v, w \rangle$ coincides with the cosine of the angle between v and w. Actually, in the following, we use an enhanced version of the local cost measure by incorporating contextual information, see [5] for details. The resulting self-similarity matrix will be denoted by $\mathcal{S}$ and is shown in Figure 1 (a) for our Beethoven example.

To account for transpositions, we revert to the observation by Goto [4] that the twelve cyclic shifts of a 12-dimensional chroma vector naturally correspond to the twelve possible transpositions. In contrast to previous approaches, we incorporate all transpositions into a single self-similarity matrix. To this end, let $\sigma : \mathcal{F} \to \mathcal{F}$ denote the *cyclic shift* defined by

$$\sigma((v(1), v(2), \ldots, v(12))) := (v(2), \ldots, v(12), v(1))$$

for $v := (v(1), \ldots, v(12)) \in \mathcal{F}$. Then, for a given audio data stream with chroma-based feature sequence $V := (v_1, v_2, \ldots, v_N)$, the *$i$-transposed self-similarity matrix* $\sigma^i(\mathcal{S})$ is defined by

$$\sigma^i(\mathcal{S})(n, m) := c(v_n, \sigma^i(v_m)),$$

for $1 \leq n, m \leq N$ and $i \in \mathbb{Z}$. Obviously, one has $\sigma^{12}(\mathcal{S}) = \mathcal{S}$. Intuitively, $\sigma^i(\mathcal{S})$ describes the similarity relations between the original audio data stream and the audio data streams transposed by i semitones upwards (modulo 12). Taking the minimum over the twelve different cylic shifts, we obtain the *transposition-invariant self-similarity matrix* $\sigma^{\min}(\mathcal{S})$ defined by

$$\sigma^{\min}(\mathcal{S})(n, m) := \min_{i \in [0:11]} \Big(\sigma^i(\mathcal{S})(n, m) \Big).$$

Furthermore, we store the minimizing shift indices in an additional N-square matrix $\mathcal{I}$, which is referred to as *transposition index matrix*:

$$\mathcal{I}(n, m) := \operatorname{argmin}_{i \in [0:11]} \Big(\sigma^i(\mathcal{S})(n, m) \Big).$$

We illustrate this concept by means of two examples. Figure 1 (b) shows the transposition-invariant self-similarity matrix of our Beethoven example. The most striking difference to the conventional self-similarity matrix shown in Figure 1 (a) are the two additional paths in the upper left part. (Due to the symmetry of $\mathcal{S}$ and $\sigma^{\min}(\mathcal{S})$, we only consider the part above the main diagonal in the following discussion.) The first of these paths

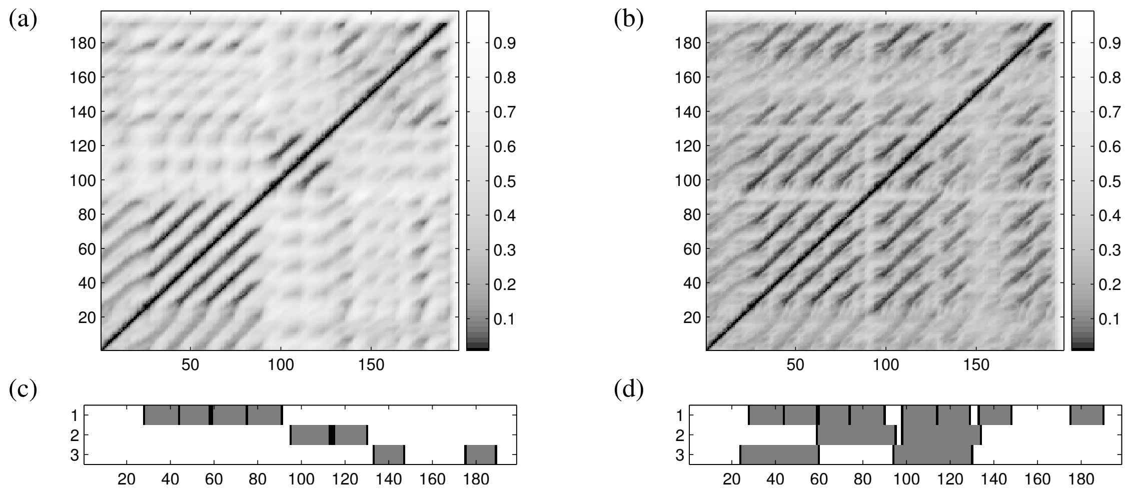

Figure 2. Zager & Evans, "In the year 2525". **(a)** Self-similarity matrix $\mathcal{S}$. **(b)** Groups of mutually similar audio segments obtained from $\mathcal{S}$. **(c)** Transposition-invariant self-similarity matrix $\sigma^{\min}(\mathcal{S})$. **(d)** Groups obtained from $\sigma^{\min}(\mathcal{S})$.

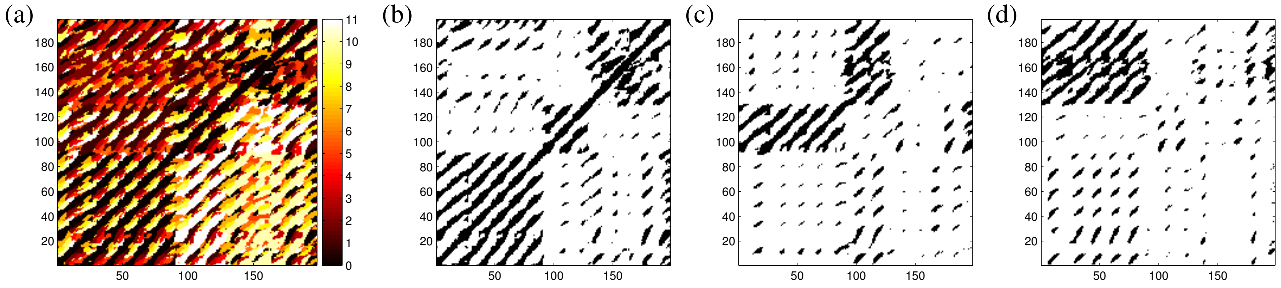

Figure 3. Zager & Evans, "In the year 2525". **(a)** Color-coded representation of the transposition index matrix $\mathcal{I}$. The three black-white images indicate the positions (black color), where the minimizing index is **(b)** $i = 0$ corresponding to no shift, **(c)** $i = 1$ corresponding to one semitone upwards, **(d)** $i = 2$ corresponding to two semitones upwards.

starts at coordinate $(67, 452)$ and ends at $(120, 504)$ indicating the similarity of the time intervals $[67 : 120]$ (part U_1) and $[452 : 504]$ (part U_3). Similarly, the second of these paths starts at coordinate $(196, 452)$ and ends at $(247, 504)$ indicating the similarity of the time intervals $[196 : 247]$ (part U_2) and $[452 : 504]$ (part U_3). Thus, these paths reveal the modulated repetition of the second theme in the recapitulation. We will continue our discussion of further additional path relations in Section 3.

As second example, we consider the song "In the year 2525" by Zager & Evans, which has the musical form $AB_1^0B_2^0B_3^0B_4^0C_1^0B_5^1B_6^1C_2^1B_7^2EB_8^2F$. The song starts with a slow intro, which is represented by the A-part. The chorus of the song, which is represented by the B-parts, is repeated 8 times. In particular, B_5^1 and B_6^1 are transpositions by one semitone upwards and B_7^2 and B_8^2 are transpositions by two semitones upwards of the first four B-parts B_1^0 to B_4^0. The respective transposition indices have been indicated by the additional superscripts. Similarly the two transitional C-parts are shifted versions from each other. Figure 2 (a) shows the conventional self-similarity matrix. The path relations reveal the similarities of the four audio segments corresponding to the first four B-parts as well as the similarity between the audio segments corresponding to B_5^1 and B_6^1 and to B_7^2 and B_8^2, respectively. However, the pairwise similarity relations between all eight B-parts only become visible in the transposition-invariant self-similarity matrix shown in Figure 2 (b).

The transposition index can be read off from the transposition index matrix $\mathcal{I}$, which is shown in Figure 3 (a) for the song "In the year 2525" in a color-coded form. Note that, opposed to the self-similarity matrices, $\mathcal{I}$ is not symmetric along the main diagonal. Actually, a minimizing index i at coordinate (n, m) induces a minimizing index $12 - i$ at coordinate (m, n). For the sake of a better visualization, the three separate black-white images shown in Figure 3 (b)–(d) indicate by the black color all coordinates (n, m), where the minimizing index in the definition of $\sigma^{\min}(\mathcal{S})(n, m)$ is $i = 0$, $i = 1$, and $i = 2$, respectively. We first discuss the case $i = 0$ as shown in Figure 3 (b). Here, the black color at coordinate (n, m) indicates that $c(v_n, \sigma^i(v_m))$ assumes a minimal values for $i = 0$. In other words, the chroma vector v_n is closer to v_m than to any other shifted version of v_m. This constitutes a necessary condition for the short-term harmonic content of the audio signal at time position m to be close to the one at time position n. (However, this condition is not sufficient in the sense that the cost $c(v_n, v_m)$ may still be high in absolute terms.) Therefore, as expected, the minimizing index is $i = 0$ at all positions, where the conventional self-similarity matrix reveals paths of low cost. Analogously, Figure 3 (c) reveals all coordinates (n, m), where the short-term harmonic at time position m relates by one semitone upwards to the the short-term harmonic at time position n. Thus, the black regions of Figure 3 (c) reveal the upper semitone relation of $B_5^1 B_6^1$ to the first four B-parts. In addition, they also reveal upper semitone relation between $B_7^2 B_6^8$ and $B_5^1 B_6^1$. Figure 3 (d) has a similar interpretation. Finally, the regularly placed patches (short paths) in Figure 3 (c) and (d) reveal interesting substructures of the B-parts. Indeed, each B-part itself consists of four subparts which are harmonically correlated: the second subparts is a shifted version of the first one going one semitone downwards. The third subpart is shifted a further semitone downwards, before the melody is going upwards again in the fourth subpart.

3 AUDIO STRUCTURE ANALYSIS

We will now sketch, how transposition-invariant self-similarity matrices can be used for efficient *audio structure analysis*. Here, the goal is to automatically extract the repetitive structure or, more generally, the musical form of the underlying piece of music, see, e. g., [1, 2, 4, 6]. In our experiments, we used an implementation of the approach described in [6], which computes groups of audio segments within an audio file that are similar in harmonic progression. This is achieved by running through the following general steps:

1. Extract chroma-based features from the audio signal and compute the transposition-invariant self-similarity matrix $\sigma^{\min}(\mathcal{S})$ as well as the transposition index matrix $\mathcal{I}$. By incorporating contextual information at various tempo levels into the cost measure, the structural properties of the matrix are enhanced, see [5].

2. Extract off-diagonal paths from $\sigma^{\min}(\mathcal{S})$ using a greedy strategy. Each path encodes a pair of similar segments. This step takes care of relative differences in the tempo progression between musically similar segments.

3. Derive the global repetitive structure from the similarity pairs by using suitable clustering techniques. In particular, we employ a one-step transitivity clustering procedure, which balances out the inconsistencies introduced by inaccurate and incorrect path extractions, see [6].

As output, the algorithm delivers a list of groups with each group representing a set of mutually similar audio segments. The final result for the Beethoven example is shown in Figure 1 (d). Each of the six rows corresponds to a group of mutually similar audio segments, where each segment is represented by a gray bar. For example, the first row reveals the similarity between the exposition A_1 and its repetition A_2. The second row reveals the similarity between the three U-parts corresponding to the second theme, where U_3 is a transposed version of U_1 and U_2. Note that this similarity group is not detected when using the conventional self-similarity matrix $\mathcal{S}$, cf. Figure 1 (c). Furthermore, the fourth row, which consists of 8 segments, reveals some interesting substructure: the slow introduction R_1 (first measure) of the "Tempest" is repeated several times throughout the piece in different keys.

Similarly, Figure 2 (d) shows the final result of the extracted global repetitive structure for the song "In the year 2525". The first row encodes a group of eight mutually similar audio segments, which are exactly the eight B-parts. In contrast, when using the conventional self-similarity matrix $\mathcal{S}$, this group is split up into three different groups as illustrated by Figure 2 (c). Furthermore, the second and third row of Figure 2 (d) reveal some superstructure, which are not present in Figure 2 (c). For example, the second row reveals the similarity between $B_3^0 B_4^0 C_1^0$ and $B_5^1 B_6^1 C_2^1$.

4 CONCLUSIONS

In this paper, we have introduced transposition-invariant self-similarity matrices, which reveal the repetitive audio structure even in the presence of key changes. Note that previous approaches to structure analysis such as [4] achieve transposition invariance by computing similarity groups for all twelve transpositions separately, which are then suitably merged in a postprocessing step. In contrast to this, we incorporate all transpositions into a single self-similarity matrix, which then allows for performing a singly joint path extraction and clustering step only. Our experiments showed that such a joint procedure not only significantly increases the efficiency of the overall algorithm, but also stabilizes the clustering step for deriving the similarity groups. An interesting yet open problem is to consider not only transpositions but also other types of modulations such as changes from major to minor keys and vice versa.

5 REFERENCES

[1] M. A. BARTSCH AND G. H. WAKEFIELD, *Audio thumbnailing of popular music using chroma-based representations*, IEEE Trans. on Multimedia, 7 (2005), pp. 96–104.

[2] R. DANNENBERG AND N. HU, *Pattern discovery techniques for music audio*, in Proc. ISMIR, Paris, France, 2002.

[3] J. FOOTE, *Visualizing music and audio using self-similarity*, in ACM Multimedia, 1999, pp. 77–80.

[4] M. GOTO, *A chorus-section detecting method for musical audio signals*, in Proc. IEEE ICASSP, 2003, pp. 437–440.

[5] M. MÜLLER AND F. KURTH, *Enhancing similarity matrices for music audio analysis*, in Proc. IEEE ICASSP, 2006.

[6] ———, *Towards structural analysis of audio recordings in the presence of musical variations*, EURASIP Journal on Advances in Signal Processing, (2007). Article ID 89686, 18 pages.