# HIGH RESOLUTION AUDIO SYNCHRONIZATION USING CHROMA ONSET FEATURES

*Sebastian Ewert*[*]

Universität Bonn, Informatik III
Römerstr. 164, D-53117 Bonn

*Meinard Müller, Peter Grosche*[†]

Saarland University and MPI Informatik
Campus E1 4, D-66123 Saarbrücken

## ABSTRACT

The general goal of music synchronization is to automatically align the multiple information sources such as audio recordings, MIDI files, or digitized sheet music related to a given musical work. In computing such alignments, one typically has to face a delicate tradeoff between robustness and accuracy. In this paper, we introduce novel audio features that combine the high temporal accuracy of onset features with the robustness of chroma features. We show how previous synchronization methods can be extended to make use of these new features. We report on experiments based on polyphonic Western music demonstrating the improvements of our proposed synchronization framework.

***Index Terms***— Music synchronization, onset features, chroma features, audio alignment

## 1. INTRODUCTION

In digital music libraries and private music collections, there is an increasing number of documents available for a given musical work. These documents may comprise various audio recordings, MIDI files or score representations. Music information retrieval (MIR) aims at developing techniques and tools for organizing, understanding and searching this multimodal information in a robust, efficient and intelligent manner. In this context, various alignment and synchronization procedures have been proposed with the common goal to automatically link several types of music representations, thus coordinating the multiple information sources related to a given musical work [4, 5, 6].

In general terms, *music synchronization* denotes a procedure which, for a given position in one representation of a piece of music, determines the corresponding position within another representation. Depending upon the respective data formats, one distinguishes between various synchronization tasks [5]. For example, *audio-audio* synchronization refers to the task of time aligning two different audio recordings of a piece of music. These alignments can be used to jump freely between different interpretations, thus affording efficient and convenient audio browsing. The goal of *MIDI-audio* synchronization is to coordinate MIDI note events with audio data. The result can be regarded as an automated annotation of the audio recording with available MIDI data.

In the design of synchronization algorithms, one has to deal with a delicate tradeoff between robustness and temporal accuracy. As first contribution, we introduce a novel class of 12-dimensional onset features, which combine the robustness of conventional chroma features [1] with the accuracy of conventional one-dimensional onset features [2]. These features are obtained by identifying pitch-based onset information on the chroma level (Sect. 2). As second contribution, we introduce a synchronization framework that allows for improving the overall synchronization accuracy without losing robustness (Sect. 3). Here, the idea is to making the best of each feature type when combining the various information. Our experiments show that our synchronization procedure, which integrates conventional chroma features as well as our novel onset features, significantly improves the accuracy in particular for piano music while not collapsing for music that does not contain clear note attacks (Sect. 4). Further related work will be discussed in the respective sections.

## 2. AUDIO FEATURES

In order to synchronize different music representations, one needs to find suitable feature representations being robust towards those variations that are to be left unconsidered in the comparison. In this context, chroma-based features have turned out to be a powerful tool for synchronizing harmony-based music, see [4, 5]. In summary, chroma features encode the short-time energy distribution over the 12 traditional pitch classes of the equal-tempered scale encoded by the attributes C, C$^\sharp$, D, . . .,B. Furthermore, chroma features can be made invariant to dynamic variations by normalization. For details we refer to the literature [5]. In the following, the first six measures of the Etude No. 2, Op. 100, by Friedrich Burgmüller will serve us as our running example, see Fig. 1a, denoted by the identifier **Burg2**. Fig. 1b shows a normalized chroma representation of an audio recording of **Burg2**.

In the following, we describe another class of highly expressive audio features that indicate note onsets along with

[*] The author is funded by the German Research Foundation (DFG CL 64/6-1)

[†] The authors are funded by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI)

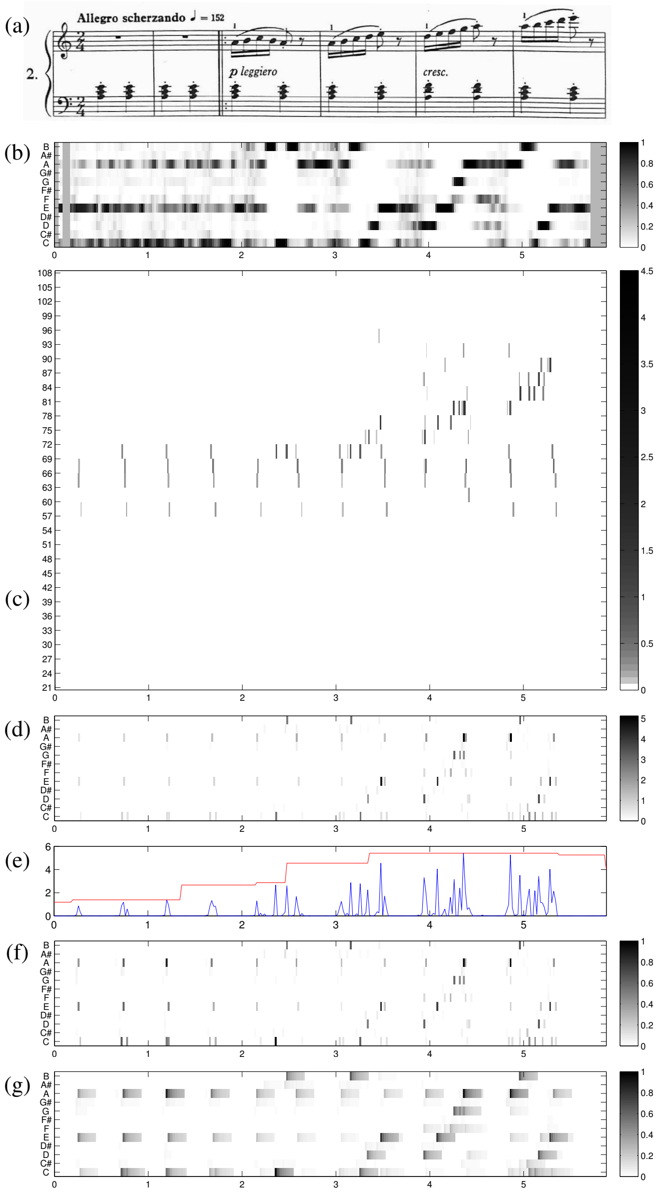

**Fig. 1**. **(a)** First six measures of Burgmüller, Op. 100, Etude No. 2. **(b)** - **(g)** feature representations of a corresponding audio recording (see Sect. 2 for a description).

their respective pitch affiliation [5]. The feature extraction procedure is motivated by the observation that for many instruments such as the piano or the guitar, playing a note results in a sudden energy increase (attack phase).

First, the audio signal is decomposed into 88 subbands corresponding to the musical notes A0 to C8 (MIDI pitches $p = 21$ to $p = 108$) of the equal-tempered scale, as in the chroma feature calculation. Then, 88 local energy curves are computed by convolving each of the squared subbands with a suitable window function. Finally, for each energy curve the discrete temporal derivative is calculated and half-wave rectified (positive part of the function remains). The significant peaks of the resulting curves indicate positions of significant energy increase in the respective pitch subband. An onset feature is specified by the pitch of its subband and by the time position and height of the corresponding peak. Fig. 1c shows the resulting onset representation obtained for our running example **Burg2**. Note that the set of onset features is sparse while providing information of very high temporal accuracy. On the downside, the extraction of onset features is a delicate problem involving fragile peak picking operations.

To enhance the robustness of the pitch-based onset features, we add up the features belonging to pitches of the same pitch class, as motivated by the chroma features. We first evenly split up the time axis into segments of a fixed length and take a suitable logarithm of the onset values, which accounts for the logarithmic sensation of sound intensity. For each segment, we add up the logarithmic values over all pitches that correspond to the same chroma. The resulting 12-dimensional features will be referred to as *CO (chroma onset) features*, see Fig. 1d. To make the CO feature invariant to dynamic variations while keeping low level onsets we employ a locally adaptive normalization strategy. First, we compute the norm of each 12-dimensional CO feature vector, see Fig. 1e (blue curve). Then, for each time frame, we assign the local maxima of the sequence of norms over a time window that ranges one second to the left and one second to the right, see Fig. 1e (red curve). Finally, we divide the sequence of CO features by the sequence of local maxima in a pointwise fashion, see Fig. 1f. The resulting features are referred to as *LNCO (locally adaptive normalized CO) features*. Intuitively, LNCO features account for the fact that onsets of low energy are less relevant in musical passages of high energy than in passages of low energy. In summary, the octave identification makes LNCO features robust to timbre and extraction errors while still encoding 12-dimensional highly accurate onset information. At this point, we emphasize that the opposite variant of first computing chroma features and then computing onsets from the resulting chromagrams is not as successful as our strategy. The major reason for this is that by first changing to the coarser chroma representation one may already loose valuable onset information. For example, suppose a clear onset in the C3 pitch band and some smearing in the C4 band. Then, the smearing may overlay the onset on the chroma level, which may result in missing the onset information. However, by first computing onsets for all pitches separately and then merging this information, the onset of the C3 pitch band will be clearly visible on the chroma level.

In view of synchronization applications, we further process the LNCO feature representation by introducing an additional temporal decay. To this end, each LNCO feature vector is copied $n$ times and the copies are multiplied by decreasing positive weights (in our experiments we chose $n = 10$ with weights $(1, \sqrt{0.9}, \sqrt{0.8}, \ldots, \sqrt{0.1})$). Then, the $n$ copies are arranged to form short sequences of $n$ consecutive feature vectors of decreasing norm starting at the time position of the

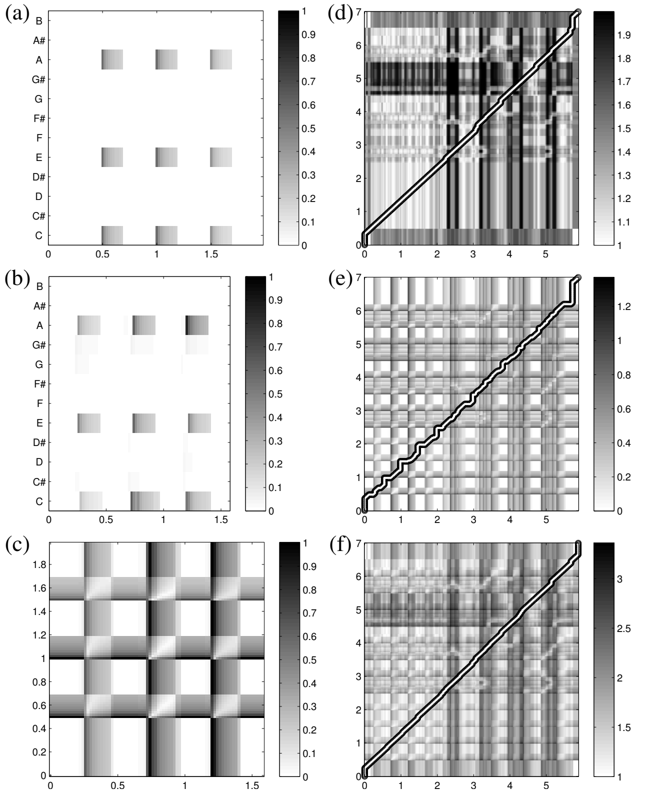

**Fig. 2**. **(a)-(c)** Illustration of the effect of the decay operation on the cost matrix level. **(d)** $C_{\textbf{chroma}}$, **(e)**, $C_{\textbf{DLNCO}}$ **(f)** $C_{\textbf{chroma}} + C_{\textbf{DLNCO}}$ for **Burg2**.

original vector. The overlay of all these decaying sequences results in a feature representation, which we refer to as *DLNCO (decaying LNCO) feature* representation, see Fig. 1g. The benefit of these additional temporal decays will become clear in the synchronization context described in Sect. 3. Note that in the DLNCO feature representation, one does not lose the temporal accuracy of the LNCO features—the onset positions still appear as sharp left edges in the decays.

## 3. SYNCHRONIZATION ALGORITHM

In this section, we show how our novel DLNCO features can be used to significantly improve the temporal accuracy of previous chroma-based synchronization strategies without sacrificing their robustness. While we consider the case of MIDI-audio synchronization in the following, other cases such as audio-audio synchronization may be handled in the same fashion. Most synchronization algorithms [4, 6] rely on some variant of dynamic time warping (DTW) and can be summarized as follows. First, the two music data streams to be aligned are converted into feature sequences, say $V := (v_1, v_2, \ldots, v_N)$ and $W := (w_1, w_2, \ldots, w_M)$, respectively. Then, an $N \times M$ cost matrix $C$ is built up by evaluating a local cost measure $c$ for each pair of features, i. e., $C(n,m) = c(v_n, w_m)$ for $1 \leq n \leq N, 1 \leq m \leq M$. Finally, an optimum-cost alignment path is determined from this matrix via dynamic programming, which encodes the synchronization result. See [5] for a detailed account on DTW in the music context. For an illustration, we refer to Figs. 2d-2f, which show various cost matrices along with optimal alignment paths for our **Burg2** example.

We now introduce three different cost matrices, where the third one is a simple combination of the first and second one. The first matrix $C_{\textbf{chroma}}$ is a conventional cost matrix based on normalized chroma features and the cosine distance [4, 5], see Fig. 2d. The second cost matrix $C_{\textbf{DLNCO}}$ is based on DLNCO features as introduced in Sect. 2. To compare two DLNCO feature vectors, $v$ and $w$, we now use the Euclidean distance $c_{\textbf{DLNCO}}(v, w) := \|v - w\|$, see Fig. 2e. At this point, we need to make some explanations. First, recall that each onset has been transformed into a short vector sequence of decaying norm. As an example, Figs. 2a and 2b show DLNCO features for the very beginning of **Burg2** for an audio and a MIDI version, respectively. Using the Euclidean distance to compare two such decaying sequences leads to a diagonal corridor of low cost in $C_{\textbf{DLNCO}}$ in the case that the directions (i. e., the relative chroma distributions) of the onset vectors are similar, see Fig. 2c. This corridor is tapered to the lower left and starts at the precise time positions of the two onsets to be compared. Second, note that $C_{\textbf{DLNCO}}$ reveals a grid like structure of an overall high cost, where each beginning of a corridor forms a small needle's eye of low cost. Third, sections in the feature sequences with no onsets lead to regions in $C_{\textbf{DLNCO}}$ having zero cost. In other words, only significant events in the DLNCO feature sequences take effect on the cost matrix level and the structure of $C_{\textbf{DLNCO}}$ regulates the course of a cost-minimizing alignment path in event-based regions to run through the needle's eyes of low cost.

The cost matrix $C_{\textbf{chroma}}$ accounts for the rough harmonic flow of the two representations, whereas $C_{\textbf{DLNCO}}$ exhibits onsets of the same chroma class. The sum $C = C_{\textbf{chroma}} + C_{\textbf{DLNCO}}$ yields a cost matrix that accounts for both types of information. Note that in regions with no onsets, $C_{\textbf{DLNCO}}$ is zero and the combined matrix $C$ is dominated by $C_{\textbf{chroma}}$. Contrary, in regions with significant onsets, $C$ is dominated by $C_{\textbf{DLNCO}}$. Therefore, the component $C_{\textbf{chroma}}$ regulates the overall course of the cost-minimizing alignment path and accounts for a robust synchronization, whereas the component $C_{\textbf{DLNCO}}$ locally adjusts the alignment path and accounts for high temporal accuracy.

## 4. EXPERIMENTS

In this section, we report on some of our synchronization experiments, which have been conducted on a corpus of harmony-based Western music. To allow for a reproduction of our experiments, we used pieces from the RWC music database [3]. In the following, we consider 16 representa-

| **RWC ID** | (Comp./Interp., Instr.) | **Chroma** | | **DLNCO** | | **Chroma+ DLNCO** | |
|---|---|---|---|---|---|---|---|
| | | mean | std | mean | std | mean | std |
| **Burg2** | (Burgmüller, piano) | 50 | 48 | 21 | 17 | 18 | 14 |
| **C025** | (Bach, piano) | 27 | 33 | 18 | 27 | 14 | 12 |
| **C028** | (Beethoven, piano) | 54 | 58 | 131 | 318 | 29 | 40 |
| **C031** | (Chopin, piano) | 57 | 64 | 22 | 68 | 22 | 33 |
| **C032** | (Chopin, piano) | 30 | 47 | 12 | 9 | 13 | 21 |
| **C029** | (Schumann, piano) | 46 | 72 | 94 | 264 | 15 | 36 |
| **Average over piano** | | 44 | 54 | 50 | 117 | 19 | 26 |
| **C003** | (Beethoven, orchestra) | 116 | 96 | 241 | 338 | 116 | 98 |
| **C015** | (Borodin, strings) | 79 | 68 | 268 | 356 | 82 | 56 |
| **C022** | (Brahms, orchestra) | 50 | 54 | 26 | 52 | 17 | 20 |
| **C044** | (Rimski-Korsakov, flute/piano) | 41 | 17 | 22 | 19 | 27 | 15 |
| **C048** | (Schubert, voice/piano) | 55 | 50 | 70 | 173 | 31 | 34 |
| **Average over non-piano** | | 68 | 57 | 125 | 188 | 55 | 45 |
| **J001** | (Nakamura, piano) | 34 | 59 | 17 | 37 | 14 | 15 |
| **J038** | (HH Band, big band) | 45 | 46 | 85 | 204 | 31 | 64 |
| **J041** | (Umitsuki Quart., sax/bass/perc.) | 39 | 67 | 37 | 117 | 23 | 55 |
| **P031** | (Nagayama, electronic) | 68 | 50 | 124 | 217 | 46 | 43 |
| **P093** | (Burke, voice/guitar) | 91 | 95 | 71 | 103 | 40 | 58 |
| **Average over jazz/pop** | | 55 | 63 | 67 | 136 | 31 | 47 |
| **Average over all** | | 55 | 58 | 79 | 145 | 34 | 38 |

**Table 1**. Alignment accuracy for the three different synchronization procedures (Chroma, DLNCO, Chroma+DLNCO) on the test database obtained from the RWC database [3]. All values are given in milliseconds.

tive pieces, which are listed in Table 1. These pieces are divided into three groups: six classical piano pieces, five classical pieces of various instrumentations, and five jazz pieces and pop songs. Note that while pure piano music typically comprises the concise note attacks the DLNCO features are designed for, such information is often missing especially in string and general orchestral music. We now show that our extended synchronization framework leads to significant improvements for piano music, while not losing on accuracy for music lacking in clear note attacks.

In the following, we use three different synchronization procedures based on chroma features only, on DLNCO features only, and a combination of these features (Chroma+DLNCO), see Sect. 3. In each experiment we use 50 features per second, i.e., the features have a temporal resolution of 20ms. To automatically determine the accuracy of our synchronization procedures, we used pairs of MIDI and audio versions of each of the 16 pieces listed in Table 1. Here, the audio versions were generated from the MIDI files using a high-quality synthesizer. Thus, for each synchronization pair, the note onset times in the MIDI file are perfectly aligned with the onset times in the respective audio recording. We randomly distorted the MIDI files by splitting up the MIDI files into $N$ segments of equal length (in our experiment we used $N = 20$) and stretching or compressing each segment by a random factor within an allowed distortion range (in our experiments we used a range of $\pm 30\%$). We refer to the resulting MIDI file as the *distorted MIDI file* in contrast to the original *annotation MIDI file*. We synchronized the distorted MIDI file and the associated audio recording and used the resulting alignment path to adjust the note onset times in the distorted MIDI file and to obtain a third MIDI file referred to as *realigned MIDI file*. The accuracy of the synchronization result is then determined by comparing onset times of corresponding notes in the realigned MIDI file and the annotation MIDI file. For each of the 16 pieces and for each synchronization procedure Table 1 shows the mean value and the standard deviation over all absolute onset differences. Note that using a combination of chroma and DLNCO features significantly improves the synchronization accuracy: the average onset error for piano music drops from 44ms (Chroma) to 19ms (Chroma+DLNCO). For orchestral or pure string music without clear note attacks, the DLNCO features do not yield any valuable information. For example, in the case of Borodin's String Quartet (C015), the onset error increases from 79ms (Chroma) to 269ms (DLNCO) when using only the onset features. However, in the combined case, the chroma features overrule the corrupt DLNCO features leading to an onset error of 82ms (Chroma+DLNCO) that is comparable to the chroma only case.

In conclusion, our experiments show that the combination of using chroma and DLNCO onset features significantly improve the synchronization accuracy for music with clear note attacks and does not degrade for music which lacks this information. At this point, one may object that one typically obtains better absolute synchronization results for synthetic audio material (which was used to completely automate our evaluation) than for real audio recordings. We also tested our synchronization on real audio recordings of all 16 pieces, which actually led to similar results as the synthesized examples. Sonifications of the MIDI-audio synchronization results for the real audio files of the 16 pieces have been made available on the website `http://www-mmdb.iai.uni-bonn.de/projects/syncDLNCO/`.

For the future, we will incorporate other types of features that capture local rhythmic information and smooth note transitions for orchestral, string, or brass music [7]. Here, our synchronization framework allows for making the best of each feature type when combining the various information.

## 5. REFERENCES

[1] M. Bartsch and G. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Trans. on Multimedia*, vol. 7, no. 1, pp. 96–104, Feb. 2005.

[2] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," in *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, Sept. 2005.

[3] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical and jazz music databases," in *Proc. ISMIR, Paris, France*, 2002.

[4] N. Hu, R. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proc. IEEE WASPAA, New Paltz, NY*, October 2003.

[5] M. Müller, *Information Retrieval for Music and Motion*, Springer, 2007.

[6] R. J. Turetsky and D. P. W. Ellis, "Force-aligning MIDI syntheses for polyphonic music transcription generation," in *Proc. ISMIR, Baltimore, USA*, 2003.

[7] W. You and R. Dannenberg, "Polyphonic music note onset detection using semi-supervised learning.," in *Proc. ISMIR, Vienna, Austria*, 2007.