Combination of Onset-Features with Applications to High-Resolution Music Synchronization

Peter Grosche[1], Meinard Müller[1], Sebastian Ewert[2]

[1] *Saarland University and MPI Informatik, Saarbrücken, Germany, {pgrosche,meinard}@mpi-inf.mpg.de*

[2] *Bonn University, Bonn, Germany, ewerts@cs.uni-bonn.de*

Introduction

Many different methods for the detection of note onsets in music recordings have been proposed and applied to tasks such as music transcription, beat tracking, tempo estimation, and music synchronization. Most of the proposed onset detectors rely on the fact that note onsets often go along with a sudden increase of the signal's energy, which particularly holds for instruments such as piano, guitar, or percussive instruments. Much more difficult is the detection of onsets in the case of more fluent note transitions, which is often the case for classical music dominated by string instruments. In this paper, we introduce improved novelty curves that yield good indications for note onsets even in the case of only smooth temporal and spectral intensity changes in the signal. We then show how these novelty curves can be used to significantly improve the temporal accuracy in music synchronization tasks.

Onset Detection

The most characteristic property going along with a note onset is a sudden increase in the signal's energy. However, simultaneously occurring events in polyphonic music may lead to masking effects that even out the energy ascents and prevent an observation of a significant energy increase. To circumvent this masking effects, detection functions were proposed that analyze the signal in a band-wise fashion [4] and try to extract transients occurring in a specific frequency region of the signal. A widely used approach to onset detection in the frequency domain is the *spectral flux* or *novelty* method described in [1] that analyzes the lapse of the spectral content of the signal and thus adds the possibility for detecting changes of pitch or timbre of the signal. As a side-effect of a sudden energy increase, there often is an accompanying broadband noise burst appearing in the signal's spectrum. This effect is mostly masked by the signal's energy in lower frequency regions but well detectable in the *high-frequency content* [2] of the spectrum.

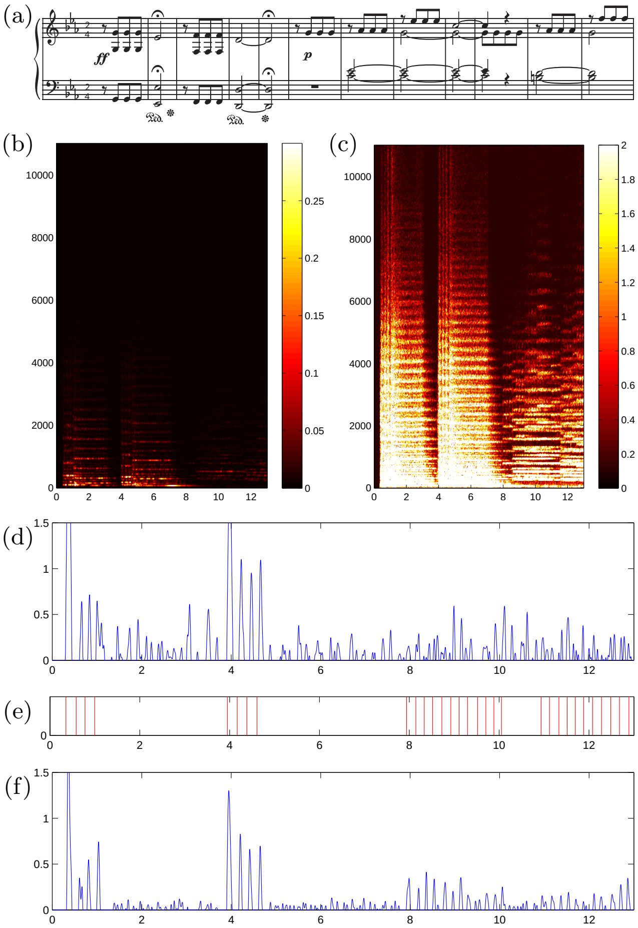

Figure 1: First 12 measures of Beethoven's Symphony No. 5 (Op. 67) interpreted by Bernstein. **(a)** Score representation. **(b)** Linear magnitude spectrogram $|X|$. **(c)** Logarithmic magnitude spectrogram Y. **(d)** Novelty curve $\bar{\Delta}_{\mathrm{mag}}$ based on the magnitude spectrogram $|X|$. **(e)** Ground truth annotation of onsets. **(f)** Novelty curve $\bar{\Delta}_{\mathrm{comp}}$ based on the compressed spectrogram Y.

Combining these ideas, we now describe an approach for computing a novelty curve that indicates note onset candidates. As it turns out, our novelty curve is suited for detecting percussive as well as pitched-percussive onsets even if there is only a weak attack phase. Given a music recording, a short-time Fourier transform is used to obtain a spectrogram $X = (X(k,t))_{k,t}$ with $k \in [1 : K] := \{1, 2, \ldots, K\}$ and $t \in [1 : T]$. Here, K denotes the number of Fourier coefficients, T denotes the number of frames, and $X(k,t)$ denotes the k^{th} Fourier coefficient for time frame t. Note that the Fourier coefficients of X are linearly spaced on the frequency axis. Using suitable binning strategies, various approaches switch over to a logarithmic spaced frequency axis, e.g., by using mel-frequency bands or pitch bands, see [4]. Here, we keep the linearly spaced frequency axis, since it puts greater emphasis on the high-frequency regions of the signal, thus accentuating the afore mentioned noise bursts visible in the high-frequency content. Next, we

[1] The authors are funded by the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University.

[2] The author is funded by the German Research Foundation (DFG CL64/6-1).

apply a logarithm to the magnitude spectrogram $|X|$ of the signal yielding $Y := \log(1 + C \cdot |X|)$ for a suitable positive constant $C > 1$, see [5]. The advantages of such a compression step may be summarized as follows. First, it accounts for the logarithmic sensation of human sound intensity. Furthermore, the compression factor C allows for adjusting the dynamic range of the signal. By increasing C, the low-intensity values within the spectrogram are more and more highlighted, thus preventing a masking by high-intensity values. This effect is clearly visible in Fig. 1, which shows the magnitude spectrogram $|X|$ and the compressed spectrogram Y for a Bernstein recording of Beethoven's Fifth Symphony. Here, the logarithmic compression enhances the clarity of the weak transients, especially in the high-frequency content. On the downside, a large compression factor C may also amplify non-relevant low-energy noise components. In our experiments, we use the value $C = 1000$, but our results as well as the findings reported by Klapuri et al. [5] show that the specific choice of C does not effect the final result in a substantial way.

To compute novelty curves, one basically applies a first order differentiator to the magnitude spectrum $|X|$ or compressed magnitude spectrum Y. More precisely, we sum up only positive intensity changes to emphasize onsets while discarding offsets. We define the novelty function $\Delta_{\mathrm{mag}} : [1 : T - 1] \to \mathbb{R}$ for the magnitude spectrum $|X|$ as follows:

$$\Delta_{\mathrm{mag}}(t) := \sum_{k=1}^{K} \Big| |X(k, t+1)| - |X(k,t)| \Big|_{>0} \tag{1}$$

for $t \in [1 : T - 1]$, where $|x|_{>0} := x$ for a positive real number x and $|x|_{>0} := 0$ for a negative real number x. To obtain our final novelty function $\bar{\Delta}_{\mathrm{mag}}$, we subtract the local average and only keep the positive part (half wave rectification). The final curve is shown in Figure 1d for our Beethoven example. Similarly, we define the novelty function $\Delta_{\mathrm{comp}} : [1 : T - 1] \to \mathbb{R}$ for the compressed spectrum Y as follows:

$$\begin{aligned} \Delta_{\mathrm{comp}}(t) \quad &:= \quad \sum_{k=1}^{K} \Big| Y(k, t+1) - Y(k,t) \Big|_{>0} \\ &= \quad \sum_{k=1}^{K} \Big| \log \Big(\frac{1 + C|X(k, t+1)|}{1 + C|X(k,t)|} \Big) \Big|_{>0} \qquad (2) \\ &= \quad \log \Big(\prod_{k=1}^{K} \Big| \frac{1 + C|X(k, t+1)|}{1 + C|X(k,t)|} \Big|_{>1} \Big) \end{aligned}$$

for $t \in [1 : T - 1]$, where $|x|_{>1} := x$ for a real number $x > 1$ and $|x|_{>1} := 1$ for a real number $x \leq 1$. From Δ_{comp} we obtain the final novelty function $\bar{\Delta}_{\mathrm{comp}}$ as above, see Figure 1f.

Comparing $\bar{\Delta}_{\mathrm{mag}}$ and $\bar{\Delta}_{\mathrm{comp}}$ shown in Figure 1 clearly illustrates the benefits of the compression step. Note that the logarithmization gives higher weight to an absolute intensity difference within a quiet region of the signal than within a louder region, which follows the psychoacoustic principle that a just-noticeable change in intensity is roughly proportional to the absolute intensity [6]. Furthermore, the compression leads to a better temporal localization of the onset, because the highest relative slope of the attack phase approaches the actual onset position and noticeably reduces the influence of amplitude changes (e.g. tremolo) in high intensity regions.

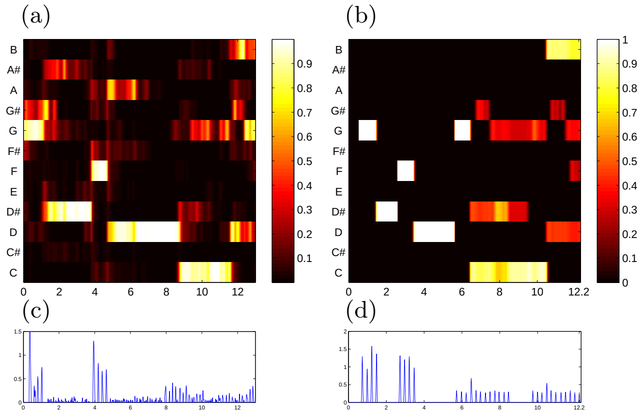

Figure 2: Chroma representations of the Beethoven example. **(a)** Audio recording. **(b)** MIDI file. **(c-d)** Corresponding novelty curves $\bar{\Delta}_{\mathrm{comp}}$.

Music Synchronization

As an application, we now show how novelty curves can be used for improving the temporal accuracy in music synchronization tasks. 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 [3]. For example, the goal of MIDI-audio synchronization is to coordinate MIDI events with audio data. The result can be regarded as an automated annotation of the audio recording with available MIDI data.

Most synchronization algorithms [7, 3, 9] 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. Note that N and M do not have to be equal, since the two versions typically have different lengths. 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. Our synchronization approach follows these lines using the standard DTW algorithm, see [3] for a detailed account on DTW in the music context. For an illustration, we refer to Fig. 3a, which shows a cost matrix along with an optimal alignment path.

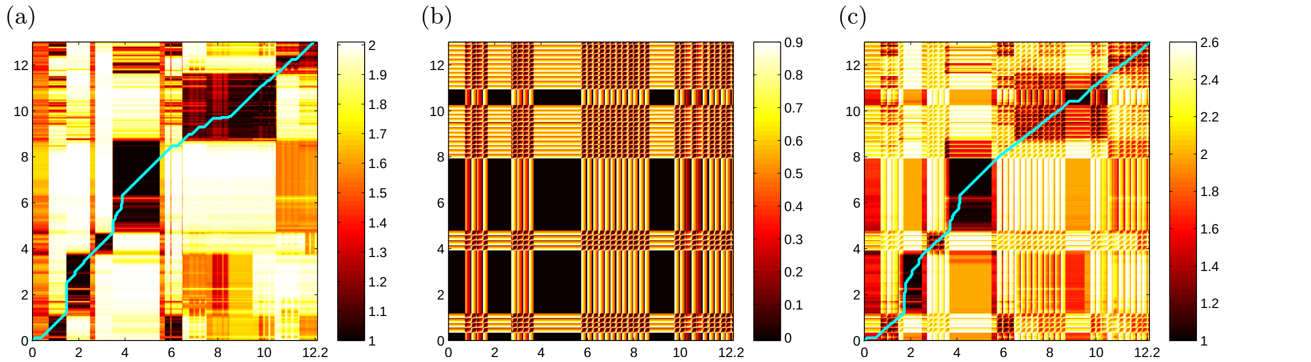

Figure 3: (a) Conventional chroma-based cost matrix C_{chroma}. **(b)** Cost matrix of novelty curves C_{novelty}. **(c)** Combined cost matrix $C_{\text{chroma+novelty}}$.

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 [3, 8]. Here, the chroma refer to the 12 traditional pitch classes of the equal-tempered scale encoded by the pitch spelling attributes $\mathrm{C}, \mathrm{C}^\sharp, \mathrm{D}, \ldots, \mathrm{B}$. Representing the short-time energy content of the signal in each of the 12 pitch classes, chroma features show a high degree of robustness to variations in timbre and articulation [8]. Furthermore, normalizing the features makes them invariant to dynamic variations. For details on how to derive chroma features from audio and MIDI files, we refer to the cited literature. For our Beethoven example, Fig. 2 shows chroma representations of an audio and MIDI version.

Onset-Enhanced Synchronization

Using chroma features for music synchronization leads to reasonable alignment results. However, since the chroma features account only for the rough harmonic flow of a piece the temporal accuracy of alignments from purely chroma-based approaches is often not sufficient depending on the respective application and on the type of music.

To address this issue, the authors of [10] introduce a synchronization strategy that employs a combination of two cost matrices to account for complementary musical information. One cost matrix is based on conventional chroma features and the second one on chroma-based onset features. The onset features proposed in [10] are particularly designed for piano music, where certain characteristics of the piano sound are exploited. The combination of the two cost matrices leads to a significantly enhanced precision of the alignments for this kind of music. However, because of the simple energy based onset detection method, the alignment accuracy is not improved for music comprising instruments with soft onsets like strings.

Using our novelty curves, one can significantly improve the temporal accuracy of synchronization results even in the case of instruments that have a weak attack phase. Again, we combine two cost matrices. The first cost matrix C_{chroma} is based on chroma features. Accounting for the rough harmonic progression of the two representations to be synchronized, this matrix is used to regulate the overall course of the cost-minimizing alignment path and to assure a robust synchronization, see Fig. 3a. For the second matrix, the novelty curves $\bar{\Delta}_{\text{comp}}$ of both, the audio and MIDI file, are used, see Fig. 2c-d. The novelty curve for the MIDI file can directly be derived from the encoded onset times and velocities. The cost matrix is computed in a similar way as described in [10], where we use the Euclidean distance as local cost measure c. The resulting cost matrix C_{novelty}, which is shown in Fig. 3b for our Beethoven example, provides a rich structure. As a first observation, note that horizontal and vertical lines in C_{novelty} of an overall high cost indicate onset positions in the two versions to be synchronized. Second, at the crossing of a vertical and a horizontal line, a small diagonal "corridor" of low cost can be found in C_{novelty} indicating correspondences of onset positions. Third, sections in the feature sequences with no onsets lead to regions in C_{novelty} having zero cost. The purpose of this second matrix is to locally refine the alignment path without affecting the rough overall course of the alignment path.

We now combine the two introduced cost matrices to create a third cost matrix $C_{\text{chroma+novelty}} := C_{\text{chroma}} + C_{\text{novelty}}$, see Fig. 3c. The cost matrix obtained from the novelty curves reveals a grid-like structure of high costs, which is superimposed on top of the coarse-grained cost matrix obtained from the chroma features. The rough course of the alignment path is more or less determined by the chroma features, while the small diagonal corridors of low costs only locally regulate the alignment path to run through corresponding onset positions. Comparing the resulting cost-minimizing alignment paths of C_{chroma} in Fig. 3a and of $C_{\text{chroma+novelty}}$ in Fig. 3c, one can observe a significant improvement in temporal accuracy. Especially in regions with only marginal changes of the harmonic content, C_{chroma} fails to yield a precise alignment. Here, the cost matrix $C_{\text{chroma+novelty}}$ guides the alignment path through corresponding onset positions, thus leading to significantly improved synchronization results.

Conclusion

In this contribution, we described an approach for extracting novelty curves from music signals yielding good indicators for note onsets even in the case of only smooth temporal and spectral intensity changes within the signal. Furthermore, we showed how these features can be used for enhancing conventional music synchronization strategies, resulting in significant improvements with regard to the temporal accuracy of the alignments especially for music comprising instruments with a soft attack phase. For the future, we plan to employ the improved synchronization results for tasks such as performance analysis, where one objective is to extract expressive tempo information from music recordings.

References

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

[2] P. Masri and A. Bateman. Improved Modelling of Attack Transients in Music Analysis-Resynthesis. *Proc. of the International Computer Music Conference (ICMC)*, 1996.

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

[4] A. Klapuri. Sound onset detection by applying psychoacoustic knowledge. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999.

[5] A. P. Klapuri, A. J. Eronen and J. Astola. Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, 2006.

[6] E. Zwicker and H. Fastl, *Psychoacoustics — Facts and Models*, Springer, 1990.

[7] N. Hu, R. Dannenberg and G. Tzanetakis. Polyphonic audio matching and alignment for music retrieval. *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003.

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

[9] F. Soulez, X. Rodet and D. Schwarz, Improving polyphonic and poly-instrumental music to score alignment, *Proc. ISMIR, Baltimore, USA*, 2003.

[10] S. Ewert, M. Müller and P. Grosche, *High Resolution Audio Synchronization using Chroma Onset Features*, to appear in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.