Improving score to audio alignment:

Percussion alignment and Precise Onset Estimation

Xavier Rodet, Joseph Escribe and Sébastien Durigon

Ircam – Centre Pompidou – CNRS UMR 9912
Analyse Synthèse

Abstract

Score to audio alignment connects events in a score to points on the time axis of the performance audio signal. By using the score, it facilitates audio indexing. This paper presents some major improvements, particularly alignment of percussive sounds in polyphonic music in addition to harmonic notes and precise time-onset estimation.

A detection algorithm for percussive sounds uses correlation with percussive sounds from a database, to detect template sounds, and finds new references in the performance corresponding to the template sound looked for. Then a local distance for percussive sound and harmonic notes is used by a dynamic time warping (DTW) algorithm to perform the alignment of all notes together.

For precise time-onset estimation, different methods based on the energy of the note are applied on the frames given by the DTW alignment. To remedy for possible frame errors, an algorithm of tempo control is used to detect the notes which are found out of the local forward and backward tempi. The algorithm is evaluated on a reference database including classical, jazz improvisation and popular music. Results show a good reliability and precise time-onset estimation.

1 Introduction

Score to audio alignment, or in brief *score alignment*, connects events in a score to corresponding points on the performance signal time axis. A very similar task is known as *score following,* this term being reserved for the real-time case such as the one where a computer program is used to accompany a musician, and *score alignment* for the non real-time case (Score Following Panel, ICMC 2003). Score alignment can thus use the whole score and the whole audio file if needed to perform the task, while score following specifies a maximum latency between an event in the audio stream and the decision to connect it to an event in the score. By using score information, score alignment permits to perform extensive audio indexing. It allows computing note time-onset, duration, loudness, pitch contour, descriptors and interpretation in general. Automatic score alignment has many applications (Soulez, Schwarz and Rodet, 2003) among which we can mention:

1. Audio listening enriched with links to a symbolic score display.
2. Indexing of continuous media, for content-based access, retrieval or other applications.
3. Musicological comparison of different performances and interpretations.
4. Construction of a score describing as exactly as possible a given performance (enriched MIDI).
5. Audio segmentation into note samples for data base construction, learning, etc.

In the present paper, we focus on applications 3 and 4, which require precise note onset-time estimation. Previous works on score to audio alignment have been presented in (Soulez, Schwarz and Rodet, 2003) and (Turetsky and Ellis 2003). The present paper follows and improves (Soulez, Schwarz and Rodet, 2003) in which estimation of note time-onsets was not very precise and percussive sounds were not treated. Previous work on percussive detection can be found, for example, in (Gouyon, Pachet and Zils, 2002). Following mostly this article, a method for dealing with percussion is explained in Section 2. Improvements of onset-time estimation are described in Section 3. Among other improvements, a local tempo control algorithm to check the coherence of onset-times with respect to the local tempo and its evolution is explained in Section 4. The complete score alignment process is performed as follows:

1. Construction of the score representation by parsing the MIDI file.
2. Extraction of audio features from signal.
3. Detection of kick and snare-like sounds.
4. Calculation of local distance between score and performance for harmonic notes and percussion.
5. Computation of the optimal alignment path, which minimizes the total distance for harmonic and percussive sounds (DTW algorithm).
6. Precise onset-time estimation of aligned notes
7. Tempo control and corrections if needed.

2 Percussion Alignment

2.1 Method and Reference Templates

Tests done by (Gouyon, Pachet, and Delerue 2000) show that correlation is an appropriate method to detect percussive sounds in polyphonic audio signals. The percussion detection algorithm described in this paper has been implemented in our system. It is based on the correlation $\Gamma(\tau)$ of sound signal templates $Z(t)$ with the performance audio signal $S(t)$ being studied. Different sound signal templates are used for the different classes of percussive sounds looked for. Then, in order to improve efficiency and quality of detection, a second normalized correlation $\Lambda(t)$ is computed between $\Gamma(\tau)$ and the autocorrelation of the template $Z(t)$ being tested. Only kick-like and snare-like sounds detection are implemented in the present system. Hi-hats, toms and cymbals sounds raise various problems and are not treated for the

moment.

Our program uses different sound signal templates $Z(t)$ collected from music of various styles to take into account many percussion sounds. A low pass filter impulse response is also used as a template for kick-like sounds, and a band pass filter impulse response as a template for snare-like sounds. Template duration is fixed to 80 ms, a compromise between too short templates which would not describe the sound very well, and too long templates which increase computation time and could be too specific.

2.3 New Templates from Audio Signal

The first detection step uses a fixed set of sound signal templates. Then, new templates are extracted from the performance signal, in order to enhance detection (Gouyon, and Herrera 2001). The second step is effective only for performance signals with a low level of noise. Thus, the choice of executing this step or not is left to the user according to the signal-to-noise ratio in the performance signal.

Extraction of new templates is done as follows. Among the local maxima of the correlation $\Lambda(t)$, the four largest values are retained as possible new template positions. Then, around each of the four maxima, several temporal envelopes are computed. For each envelope, the maximum slope segment in the envelope defines a line L. The intersection of L with the time axis is considered as an estimate of the attack time. The attack time for a new template is the mean of the various estimates. For the possible new templates, the zero crossing rate (ZCR) is computed between the attack time previously defined and the maximum (Gouyon, Pachet, and Delerue 2000). For kick-like sounds, the retained template is the one with the lowest ZCR value, whereas for snare-like sounds, it is the one with the highest ZCR value. This process provides a new template to be used again for correlation and so on. The loop is stopped when the number of percussive notes does not change from one iteration to the next or when a maximum number of iterations is reached.

2.4 Providing Percussion Detection Data to the DTW Algorithm

Percussive notes are used for alignment by the DTW algorithm together with harmonic notes. The DTW algorithm is detailed in (Soulez, Rodet and Schwarz, 2003). It is based on a local distance between a note of the score and a frame of the performance signal. The smaller the distance between note N and frame k in the signal, the higher the probability that note N is being played at frame k. Thus, the correlation value (local maximum of $\Lambda(t)$) has to be transformed into a distance value d_p for percussive notes. This correlation is weighted by a coefficient w in order to provide values comparable with the distance of harmonic notes (Soulez, Schwarz and Rodet, 2003). Some maxima of the correlation function with low value are artifacts in the sense that they do not designate percussion sounds. To enhance large values and to decrease the impact of artifacts, the square of the normalized correlation Λ is used. Then, the percussion distance is the weighted minimum in the frame of the squared inverse of Λ:

$d_p = w*minimum\_in\_the\_frame(1/(\Lambda(t)*\Lambda(t)))$

Finally, the sum of the harmonic-note distance (Soulez, Schwarz and Rodet, 2003) and the percussion-note distance is used by the DTW algorithm to perform the alignment.

3 Note Onset Time Estimation

3.1 Purpose

The DTW score alignment algorithm (section I, step 5.) provides, for each note (or chord) of the score, a frame which is supposed to contain the attack of the note. For various reasons, efficiency, frequency resolution and robustness, this algorithm cannot localize note onsets precisely enough for some applications such as resynthesis. Therefore, a precise onset estimation procedure is necessary after DTW alignment. Four methods have been compared to estimate this onset-time.

3.2 Various Energy Models

In (Soulez, Rodet and Schwarz, 2003), for a given note, two values $E(t)$ and $\Delta(t)$ related to energy are defined and used. The energy $E(t)$ is defined as the sum of the energies in band pass filters centered on the harmonic partials of the note. The *energy step in a band* at time t is defined as the absolute energy variation since the previous extremum and the *energy step for a note* $\Delta(t)$ is then defined as the sum of the energy steps in each band. In the present work, two other values have also been tested, the energy derivative $E'(t)$ and a spline model $S(t)$ of the energy E(t).

A spline model of order 2, composed of two connected segments, is fitted on the energy $E(t)$ between the beginning of the frame and the maximum of $E(t)$ in the frame. The time of the connection point defines the estimated onset-time.

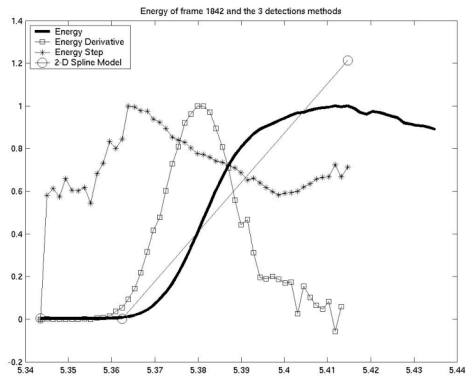

Figure 1. The 4 energy related values in a frame.

The values computed by the four methods appear on fig. 1 for a given note and detected frame. The energy *E(t)* shows an increase at a note beginning but the precise onset is difficult to determine. A large value of *Δ(t)* usually appears around the onset time and its use in the *attack distance* AD (Soulez, Rodet and Schwarz, 2003) has improved the DTW score alignment. However, this is not precise enough for note onset-time estimation. The energy derivative *E'(t)* usually shows a local maximum around the onset time, which can be used as an onset-time estimator. Still, some larger maxima can appear in the frame and cause estimation errors. These four values have been evaluated and compared (Section 5).

4 Tempo Control

The methods seen above are useless for onset estimation if the frame given by the DTW algorithm does not contain the onset of the searched note. To remedy for these errors, an algorithm named *Tempo Control* has been developed. Many tempo estimation algorithms have been published (e.g. Scheirer, 1998).

For each note, the Tempo Control algorithm first estimates a local tempo according to the score and the onset of neighboring notes. Then, the local tempo is used to find the frame where the onset of the note under study should occur. More precisely, for each note, two local tempi are estimated. Linear regression applied on the next (respectively previous) *N* onset-times provides a forward tempo *FT* (respectively backward, *BT*). If local tempo does not vary much around the considered note, i.e. *FT* close to *BT*, then the onset-time should occur in accordance with the local tempo. Otherwise, the note has probably been misplaced by DTW and its frame is eventually corrected to be coherent with local tempo. Then the *Precise Onset-Time* algorithm re-estimates onset of notes in the new frame. Tempo Control followed by precise onset-time estimation is applied a number of times fixed by the user.

5 Evaluation And Results

Rigorous evaluation of score alignment raises several difficulties. Firstly, a correct transcription of the score into MIDI is needed for the evaluation to be meaningful. Secondly, given an audio file, precise onset-time reference for each note is extremely difficult to obtain. We have used three evaluation methods. The first is to informally listen a resynthesis computed from the MIDI aligned score and judge closeness to the original. Onset-times errors are easily detected in fast rhythmic passages. The second method is to synthesize a new audio file from a MIDI file, including tempo and interpretation changes, reverberation, etc. Evaluation compares the MIDI file onset-time references with the results of the synthetic audio alignment. However, the quality of these references depends on the properties of the MIDI-to-audio synthesizer. The third method is to handmark spectrogram and audio signal. However, this difficult task can only be done on few audio files. Moreover, the precision of this hand marking is limited (for real recordings, uncertainty often exceed 10 ms) and needs to be estimated on synthetic audio files. A small evaluation database has been built with 4 audio files synthesized from MIDI with and without reverb, and 3 hand-marked audio files. Then the differences between the onset-time references and the onset-time estimations are computed. Since our main goal is audio resynthesis from enriched MIDI, the *mean* of the differences is less relevant (it reduces to a different time origin). But the *standard deviation* is a measure of how synthetic notes are played late and in advance. For fast phrases, the difference has to be kept very low for a good resynthesis. Evaluation of the three best algorithms on our database gives the following standard deviations:

Step *Δ*	: 23 ms
Energy derivative *E'*	: 19 ms
Spline model *S*	: 18 ms

Adding the tempo control algorithm and other recent improvements result in the following values:

Energy derivative *E'*	: 18 ms
Spline model *S*	: 13 ms,

while the mean deviation is reduced to only 5 ms.

A larger database of various titles, MIDI and audio performances has been built. It contains 35 pieces of classical, jazz, popular and electronic music, with percussion, harmonic and voice sounds. The system has been informally evaluated

on this data base by listening at sounds resynthesized from the MIDI aligned score, and by judging closeness to the original. Most often resynthesis is heard very close to original, even in difficult cases where interpretation varies note onset-times far from MIDI and tempo values.

6 Conclusion And Future Work

A complete system has been built for score to audio alignment. Given a MIDI file (the score) and a performance audio recording of the piece, it connects events in the score and corresponding points on the audio performance time axis. The system is able to treat classical, jazz, popular and electronic music including percussion sounds and harmonic instruments or the voice. To deal with percussion sounds in addition to harmonic instruments, a special treatment has been developed and tested. The main applications in sight are musicological comparison of different interpretations and construction of a new score describing as exactly as possible a given performance (enriched MIDI). For these applications, a precise note onset-time estimation algorithm has been developed. It is shown to largely improve the results to the point where resynthesis rarely presents audible artifacts in terms of note onset-times.

One of the most critical points is note onset-times. New methods are being tested to still improve this estimation so that the system would allow a perfect resynthesis from this point of view. For the main applications in sight, other descriptor estimation algorithms from audio are developed: note duration, loudness, pitch contour and timbre descriptors. The goal is to obtain an enriched MIDI file so that resynthesis from this score would be very close to the original recording while keeping most interpretation details in a reduced and, as much as possible, symbolic representation.

Acknowledgments

The authors thank Axel Roebel, Emmanuel Vincent and Diemo Schwarz for help and valuable discussions.

References

Soulez F., Rodet X. and Schwarz D. (2003) "Improving Polypophonic and Poly-Instrumental Music to Score Alignment." In *Proceedings ISMIR 2003,* **Baltimore.**

Turetsky, Robert J. and Daniel P.W. Ellis (2003) "Ground-Truth Transcriptions of Real Music from Force-Aligned MIDI Syntheses" In Proceedings of the International Symposium on Music *Information Retrieval (ISMIR),* **Baltimore.**

Zils A., Pachet F., Delerue O. and Gouyon F. (2002) "Automatic Extraction of Drum Tracks from Polyphonic Music Signals" In *Proceedings of International Conference on Web Delivering of Music,* **Darmstadt, Germany.**

Duxburry C., Sandler M., and Davies M. (2002) "A hybrid approach to musical note onset detection." In *Proceedings of the 5th Int. Conference on Digital Audio Effects (DAFx-02),* **Hamburg, Germany.**

Gouyon F., Pachet P., and Delerue O. (2000) "On the use of zero crossing rate for an application of classification of percussive sounds." In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00),* **Verona, Italy.**

Gouyon F., and Herrera P. (2001), "Exploration of techniques for automatic labeling of audio drum tracks instruments " In *Proceedings of MOSART: Workshop on. Current Directions in Computer Music.*

Scheirer, E. D. (1998). " Tempo and beat analysis of acoustic musical signals." *J. Acoust. Soc. Am. 103 (1)* **588-601.**

Mayor, O. " An adaptative real-time beat tracking system for polyphonic pieces of audio using multiple hypotheses." In *Proceedings of MOSART Workshop on Current Research Directions in Computer Music*, **Barcelona.**

Foote J., and Uchihashi S. (2001), "The beat spectrum: a new approach to rythm analysis" *IEEE International Conference on Multimedia & Expo 2001.*

Dannenberg R., and Hu N. (2003) "Polyphonic audio matching for score following and intelligent audio editors." In *Proc. Int. Comp. Music Conf. (ICMC),* **pp. 27-33, San Francisco: International Computer Music Association.**

Loscos A., Cano P., and Bonada J. (1999) "Score-Performance Matching using HMMs." " **In** *Proc. Int. Comp. Music Conf. (ICMC),* **pp 441-444.**

Orio N. and Schwarz D. (2001) "Alignment of Monophonic and Polypophonic Music to a Score." In *Proc. Int. Comp. Music Conf. (ICMC),* **Havana, Cuba.**

Klapuri A. (1999) "Sound Onset Detection by Applying Psychoacoustic Knowledge," In *Proceedings. IEEE Conf. Acoustics, Speech and Signal Proceesing (ICASSP,'99).*

Jaillet F., and Rodet X. (2001) "Detection and modeling of fast attack transients," In *Proc. Int. Comp. Music Conf. (ICMC),* **Havana, Cuba**.

Daudet L., Molla S., and Torresani B. (2001) "Transient detection and encoding using wavelet coefficient trees," In *Proceedings of the GRETSI'01 conference.*

Thornburg H. and Gouyon F. (2000) "A Flexible Analysis Synthesis Method for Transients," In *Proc. Int. Comp. Music Conf. (ICMC),* **Berlin**.

Roebel A. (2003) " Transient detection and preservation in the phase vocoder " In ***Pro.*** *Int. Comp. Music Conf. (ICMC),* **Singapore, p.247-250.**

Durigon S. (2003) "Rapport de stage de fin d'études: Détection automatique du tempo et des évenements percussifs dans les signaux audios polyphoniques", Ircam/EFREI.