ADAPTIVE DISTANCE NORMALIZATION FOR REAL-TIME MUSIC TRACKING

Andreas Arzt, Gerhard Widmer

Johannes Kepler University
Department of Computational Perception
Linz, Austria

Simon Dixon

Centre for Digital Music
Queen Mary University of London
London, England

ABSTRACT

The goal of real-time music tracking is to follow a musical performance on-line and at any time report the current position in the score. To achieve this, both the score and the performance have to be represented in a suitable way. In this paper, we first evaluate the performance of some well-known features and then propose a simple but effective distance normalization strategy for onset-emphasized features, which greatly improves the alignment results. Finally, we combine both harmonic and onset-emphasized features in a fashion known from off-line audio alignment, resulting in a combination which outperforms each individual feature regarding robustness and accuracy.

Index Terms— Music Processing, Real-time Music Tracking, Audio Alignment, Audio Features, Feature Normalization

1. INTRODUCTION

The task of a real-time music tracking system (a.k.a. a score follower) is to listen to a musical performance and automatically recognize at any time the current position in the musical score. A system that achieves this task accurately and robustly promises to be useful in a wide range of applications (e.g., by automatically accompanying or interacting with musicians or by creating visualizations of their performance). There exist various approaches to this problem, most notably the systems by Raphael [1] and Cont [2], which both are based on different kinds of graphical models. In contrast to these approaches our tracker is based on an on-line version of the well-known dynamic time warping algorithm.

In a real-time music tracking system the way the two time series (the score and the live performance to be aligned) are represented plays a vital role. Generally we can distinguish two kinds of features: features which describe the general energy distribution (the harmonic content) and features which describe increases in energy (modelling the onsets). While harmonic features like the normalized chroma representation are well established in the field of audio alignment, features based on onsetness have so far been somewhat neglected. A difficulty with 'onsetness' features for alignment purposes are different energy levels of onsets, which may lead to big distances between otherwise similar frames and thus to alignment errors. In this paper, we will tackle this problem by introducing a simple distance normalization strategy for onset features that tries to reduce the influence of the energy levels on the distance computation process, while still preserving the onset information. We will show that using onset features computed with this normalization procedure leads to better tracking results than using the prevailing normalized chroma features. In a final step, we combine distances based on harmonic and onset features into one distance measure, forming a combination which outperforms the individual features by far.

While the features and methods described in this paper are presented and evaluated in the context of our real-time music tracker, they may of course also be useful for the closely related task of off-line audio alignment.

2. REAL-TIME MUSIC TRACKING VIA ON-LINE DYNAMIC TIME WARPING

Our real-time audio tracking system is based on an on-line version of the dynamic time warping algorithm. It takes two time series consisting of feature vectors as input – one known completely beforehand (the score) and one coming in real-time (the live performance) –, computes an on-line alignment, and at any time returns the current position in the score. As the focus of this paper is on the feature representations used for the alignment process, we will only give a short intuitive description of this algorithm and refer the reader to [3] for further details.

Dynamic Time Warping (DTW) is an off-line alignment method for two time series based on a local cost measure and an alignment cost matrix computed using dynamic programming. Each cell of the alignment cost matrix contains the cost of the optimal alignment up to this cell. When the complete matrix is computed, the optimal alignment path is obtained by tracing the dynamic programming recursion back-

This research was supported by the Austrian Science Fund (FWF): TRP109-N23 and Z159.

wards (backward path).

In [3] Dixon proposed an on-line version based on the standard DTW algorithm which has two important properties that make it useable in real-time systems: the alignment is computed incrementally by always expanding the matrix into the direction (row or column) containing the minimal costs (forward computation), and it has linear time and space complexity due to the fact that instead of the whole matrix in each step only a fixed number of cells is computed.

Subsequently, improvements to this basic algorithm were proposed. This includes an extension called the 'backward-forward strategy' [4], which reconsiders past decisions and tries to improve the precision of the current score position hypothesis, and relatively simple tempo models [5] which are used to stretch or compress the score representation accordingly and therefore reduce differences in absolute tempo between the score representation and the live performance. For the evaluation runs in this paper, we used the simpler of the two tempo models presented in the above-mentioned paper.

3. FEATURE REPRESENTATION

In order to align a live performance to a score suitable feature representations are needed. We first convert a MIDI version of the score into a sound file using a software synthesizer. Thus we actually treat this task as an audio-to-audio alignment problem, with additional knowledge about the score audio file (e.g., the exact timing of each note onset). Both streams to be aligned are represented as sequences of analysis frames, computed via a short-time Fourier transform (STFT) of the signal with a hamming window of size 92ms and a hop size of 23ms. Then each frame resulting from the STFT is mapped onto a musically more meaningful representation better suited for the task of audio alignment.

A natural musically motivated choice for representing the tonal/harmonic content is to map the data into frequency bins with semitone spacing. As we would like this representation to be invariant to dynamic variations, we normalize each vector to sum up to 1. We will refer to this representation as normalized semitone features (*NS*).

For features representing the note onsets we again map the STFT data to the semitone scale but now only store the increase in energy in each bin relative to the previous frame. As described in [6] for chroma onset features we now first take a suitable logarithm of the values in the vector, motivated by the logarithmic sensation of sound in humans, and then normalize each vector by the maximum norm in a fixed window around this vector. While in the original paper this window is centered on the vector to be normalized, we had to shift the window to only use data up to the current vector to make it computable on-line. We will refer to this representation as locally adaptive normalized semitone onset features (*LNSO*).

Most recent audio alignment systems are based on different variants of chroma-based harmonic features (e.g., [7, 8, 9]). In general chroma vectors consist of 12 elements per time frame, corresponding to pitch classes; their values are computed by mapping the frequency bins of the STFT to the 12 pitch classes and summing up the energies. There also exist more sophisticated ways of computing chroma features and for this paper we are in fact using the method described in [10]. Again each vector finally is normalized to sum up to 1 to make it invariant to dynamic variations (normalized chroma features (*NC*)).

It is also possible to compute chroma onset features by a mapping of the LNSO features described earlier to the chroma representation, as recently done in [6]. The authors refer to this representation as locally adaptive normalized chroma onset features (*LNCO*). Additionally, as this introduces a desirable property for off-line (= backward) audio alignmen,t they also introduce an extra temporal decay to these features (decaying LNCO (*DLNCO*)), which should not be favourable for the on-line task in question. Nonetheless we will also evaluate the effect of this delay in our on-line (= forward) audio alignment algorithm.

Finally, having defined a number of possible feature representations, a function determining the alignment cost of 2 frames (distance between 2 frames) is needed. In this paper we will use the L_1 distance:

$$d(I, J) = \sum_{k=1}^{n} |I_k - J_k|, \tag{1}$$

where I and J are either semitone or chroma frames.

4. ADAPTIVE DISTANCE NORMALIZATION

One problem with using the aforementioned features based on onset information (LNSO, LNCO and DLNCO) directly is that they are only normalized relative to their local context within their audio streams. There can be huge differences in onset strength between the two audio streams, especially when the score audio stream is generated from a deadpan MIDI file without loudness information (= with the same velocity for every note). In contrast to this, there are all kinds of variations in dynamics in the live performance. Take for example a piano performance in which the performer emphasizes the melody while playing the accompaniment very softly, or 'blurs' the onsets by using the sustain pedal, as is often the case. Then the correct alignment of the louder melody notes would lead to minimal distances, but there would occur substantial alignment costs for each of the accompanying notes, possibly leading to alignment errors.

A simple solution to this problem is to compute a normalized distance d_n of two frames by dividing d by the sum of their L_1-norms:

$$d_n(I, J) = \frac{d(I, J)}{|I|_1 + |J|_1}. \tag{2}$$

ID	Composer	Piece Name	# Perf	Eval. Type
CE	Chopin	Etude Op. 10 No. 3 (excerpt)	22	Match
CB	Chopin	Ballade Op. 38 No. 1 (excerpt)	22	Match
MS	Mozart	1^{st} Mov. of Sonatas KV279, KV280, KV281, KV282, KV283, KV284, KV330, KV331, KV332, KV333, KV457, KV475, KV533	1	Match
RP	Rachmaninoff	Prelude Op. 23 No. 5	3	Man. Annotations

Table 1. The data set used for the evaluation.

Evidently this simple approach has its drawbacks. It introduces a lot of noise to the distance matrix by heavily up-scaling small distances between frames with low energy in them. Still this normalization step greatly improves the alignment results and actually makes the semitone onset features useable – alignments based on the unnormalized distances got lost most of the time (see Table 3).

To get rid of this up-scaling effect, we introduce a weight describing the 'onsetness' of the two frames involved. When chosen correctly, this weight can be seen as a dampening factor: avoiding the scale-up effect for small numbers while still normalizing the distance when enough energy is involved. Recall that due to the locally adaptive normalization step the L_1-norm of each frame is a measure of its 'onsetness', ranging from 0 to 1. Thus the mean of the L_1-norms of two frames is a natural measure of their combined 'onsetness'. Based on experimental results we chose to apply a suitable function to this value, leading to the desired dampening effect and resulting in the normalized and weighted distance d_{nw}:

$$d_{nw}(I,J) = d_n(I,J) * \sqrt[4]{\frac{|I|_1 + |J|_1}{2}}. \qquad (3)$$

It is important to note that the main point of the formula above is not the application of exactly the 4^{th} root – this function merely gave the best results in our evaluation, but only by a very small margin. We achieved very similar results with other functions (the square root, cubic root or also a function based on the logarithm), as long as the function fulfilled the intended dampening task described above.

4.1. Results

The performance of each feature configuration was thoroughly tested on various pieces of piano music (see Table 1). This table also indicates how the ground truth data was prepared, where 'match' means that the we have access to very accurate data about every note onset, as these were recorded on a computer-monitored grand piano. The Tables 2 and 3 show the percentage of correctly aligned notes for the different configurations mentioned in the text. A note is accepted as correctly aligned if the computed time differs from the actual onset time not more than 250 ms (see also [11] for more details on the evaluation of real-time audio-to-score alignment systems).

ID	$\mathbf{NS}^{dn}$	$\mathbf{NC}^{dn}$
CE	82.01%	87.78%
CB	75.04%	79.97%
MS	90.19%	91.20%
RP	75.61%	81.45%

Table 2. Real-time alignment results for the harmonic features (see text).

Regarding the harmonic features, the suitability of the NC features for audio alignment purposes is well established (see e.g., [12]) and again confirmed by our experiments (see Table 2). In contrast to that the performance of the LNSO features, which work far better than the related LNCO features, may come as a bit of a surprise (see Table 3). It seems that when it comes to modelling onsets the mapping to the chroma scale destroys crucial information (the absolute height of the onsets). As also shown in this table, the normalization step for onset features is indispensable. While with the unnormalized distances the tracker in many cases gets completely lost or at least produces a lot of errors both normalized versions lead to robust and accurate alignments, the weighted one even clearly outperforming the NC features.

Interestingly the basic DLNCO features outperformed the LNSO and LNCO features. But while the normalization process also has a positive influence on the DLNCO features in general (interestingly, the Rachmaninov Prelude is an exception) they do not benefit to the same extent. When comparing the evaluation runs using the distance normalization process the LNSO features are clearly preferable at least for on-line trackers such as ours.

5. MIXING CHROMA AND ONSET INFORMATION

Having evaluated their accuracy when used individually it seems reasonable to try to combine the two presented feature types (harmonic and onset) into one feature set, something which has already been suggested in [6] in a different (off-line) alignment setting. There the authors mix NC features with the DLNCO features described above by simply computing 2 distinct local cost matrices, and finally summing up both matrices to get a distance measure which accounts for

ID	LNSO^{d}	LNSO^{dn}	LNSO^{dnw}	LNCO^{d}	LNCO^{dn}	LNCO^{dnw}	DLNCO^{d}	DLNCO^{dn}	DLNCO^{dnw}
CE	9.68%	91.93%	**96.09%**	43.66%	92.78%	**95.85%**	77.01%	89.46%	**93.26%**
CB	5.8%	91.79%	**95.61%**	28.92%	86.74%	**93.72%**	65.97%	79.05%	**85.27%**
MS	1.2%	**97.41%**	93.76%	18.23%	90.28%	**90.91%**	48.18%	79.14%	**85.18%**
RP	2.28%	78.71%	**86.25%**	20.77%	42.67%	**71.77%**	**33.73%**	2.91%	23.56%

Table 3. Real-time alignment results for the onset features (see text).

ID	NC^{dn}	LNSO^{dnw}	NC^{dn}+LNSO^{dnw}
CE	87.78%	96.09%	**96.13%**
CB	79.97%	95.61%	**96.38%**
MS	91.20%	93.76%	**98.20%**
RP	81.45%	86.25%	**93.73%**

Table 4. Real-time alignment results for the single best features and their combination (see text).

both types of information.

We will now combine the best features of both classes according to our evaluation runs in the same fashion. For this we picked the NC features (see Table 2) and the LNSO features with the distance normalization procedure described above (see Table 3). Thus as the total distance d_{tot} of 2 frames we get:

$$d_{tot}(I,J) = d_{nw}^{LNSO}(I,J) + d_{n}^{NC}(I,J). \quad (4)$$

We also experimented with an additional weighting of the distances such that in case of onsets $d_{nw}^{LNSO}(I,J)$ is dominant and $d_{n}^{NC}(I,J)$ otherwise, but – despite some promising results with some of the 'weaker' features – we did not achieve further improvements by this strategy.

5.1. Results

As expected, these features, which complement one another in a natural way, lead to a substantial increase in alignment precision and robustness compared to their individual results (see Table 4). The combination outperformed each configuration with single features we tested for this paper.

Table 5 gives a more in-depth comparison of the combined features to the individual ones based on the cumulative frequency of errors. It again confirms that in this natural combination the NC features mainly add robustness (i.e., these features show fewer extreme errors larger than 1 second than the LNSO features). On the other hand the LNSO features greatly improve the precision (e.g., 86.95% of the notes are aligned with an error smaller or equal 0.1 seconds, compared to 67.64% when using the NC features).

Err. (sec)	NC^{dn}	LNSO^{dnw}	NC^{dn}+LNSO^{dnw}
≤ 0.05	35.53%	44.69%	**46.24%**
≤ 0.10	67.64%	86.95%	**89.00%**
≤ 0.15	78.23%	90.49%	**94.13%**
≤ 0.20	83.75%	92.42%	**96.03%**
≤ 0.25	87.12%	93.32%	**96.93%**
≤ 0.30	89.75%	94.05%	**97.57%**
≤ 0.35	91.65%	94.58%	**97.96%**
≤ 0.40	92.91%	94.99%	**98.28%**
≤ 0.45	93.98%	95.36%	**98.47%**
≤ 0.50	94.77%	95.66%	**98.71%**
≤ 1.0	98.16%	97.44%	**99.59%**

Table 5. Real-time alignment results for the single best features and their combination on the whole test set (79178 notes in total) shown as cumulative frequencies of errors of matching pairs of notes.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented a detailed evaluation of different feature configurations for real-time music matching. The main contribution of this paper is an adaptive distance normalization strategy for onset features. The final configuration, a combination of normalized chroma features and locally normalized semitone onset features, together with this distance normalization strategy led to a huge increase in alignment precision. So far we only presented evaluation results on piano music. First experiments on other kinds of music (e.g., orchestral music) show promising results, and we are now collecting the necessary ground truth data for a larger scale evaluation. In our opinion, at least regarding our real-time music tracking system, the possibilities of signal processing are exhausted and further serious improvements both in robustness and in precision are only possible by introducing musical knowledge. We will be working on this in the form of anticipative tempo models and explicit event anticipation.

7. REFERENCES

[1] C. Raphael, "Current directions with music plus one," in *Proc. of the Sound and Music Computing Conference*, 2009.

[2] A. Cont, "A coupled duration-focused architecture for realtime music to score alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 837–846, 2010.

[3] S. Dixon, "An on-line time warping algorithm for tracking musical performances," in *Proc. of the International Joint Conference on Artificial Intelligence*, 2005.

[4] A. Arzt, G. Widmer, and S. Dixon, "Automatic page turning for musicians via real-time machine listening," in *Proc. of the European Conference on Artificial Intelligence*, 2008.

[5] A. Arzt and G. Widmer, "Simple tempo models for real-time music tracking," in *Proc. of the Sound and Music Computing Conference*, 2010.

[6] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.

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

[8] M. Müller, H. Mattes, and F. Kurth, "An efficient multiscale approach to audio synchronization," in *Proc. of the International Conference on Music Information Retrieval*, 2006.

[9] B. Niedermayer and G. Widmer, "A multi-pass algorithm for accurate audio-to-score alignment," in *Proc. of the International Conference on Music Information Retrieval*, 2010.

[10] D. P. W. Ellis and G. E. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[11] A. Cont, D. Schwarz, N. Schnell, and C. Raphael, "Evaluation of real-time audio-to-score alignment," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, Vienna, Austria, 2007.

[12] C. Joder, S. Essid, and G. Richard, "A comparative study of tonal acoustic features for a symbolic level music-to-score alignment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2010*. IEEE, 2010, pp. 409–412.