A MULTI-PASS ALGORITHM FOR ACCURATE AUDIO-TO-SCORE ALIGNMENT

Bernhard Niedermayer[1]
[1]Department for Computational Perception
Johannes Kepler University Linz, Austria
music@jku.at

Gerhard Widmer[1,2]
[2]Austrian Research Institute for Artificial Intelligence
Vienna, Austria
music@ofai.at

ABSTRACT

Most current audio-to-score alignment algorithms work on the level of score time frames; i.e., they cannot differentiate between several notes occurring at the same discrete time within the score. This level of accuracy is sufficient for a variety of applications. However, for those that deal with, for example, musical expression analysis such micro-timings might also be of interest. Therefore, we propose a method that estimates the onset times of individual notes in a post-processing step. Based on the initial alignment and a feature obtained by matrix factorization, those notes for which the confidence in the alignment is high are chosen as anchor notes. The remaining notes in between are revised, taking into account the additional information about these anchors and the temporal relations given by the score. We show that this method clearly outperforms a reference method that uses the same features but does not differentiate between anchor and non-anchor notes.

1. INTRODUCTION

There are several scenarios in which one wants to know the exact parameters (such as onset time, loudness, and duration) of each individual note within a musical performance. Most of these scenarios can occur in musicology, where data from different performances is used to extract general performance rules or to analyze individual artists' expressive styles. Other applications of such data are pedagogical systems or augmented audio editors and players. Unless the pieces under consideration are played on special computer-monitored instruments, audio recordings are the only sources of data describing expression within actual musical performances.

Our aim here was to extract timing (note onset) parameters from a great variety of classical piano music performances automatically. The most general method for this would be blind audio transcription, but current state of the art methods in this field are not reliable enough to base performance analysis on their results. However, since in classical music the piece and score corresponding to an audio recording can be assumed to be known, we can address the much simpler task of audio-to-score alignment.

Here, most state of the art algorithms start by extracting features (mainly chroma vectors) from each time frame of the audio signal as well as from the score representation. To obtain an optimal alignment between the two feature sequences, a distance measure between the feature vectors is then used as input either for Dynamic Time Warping (DTW) or for graphical models, such as Hidden Markov Models.

However, an inherent shortcoming of these methods is that – since only time frames are matched – they cannot distinguish individual onsets of notes that occur simultaneously in the score. This impedes the analysis of expressive elements, such as arpeggios or the asynchronies between a pianist's hands or within a chord. To resolve this shortcoming, the method proposed here extracts an onset time estimation for each individual note. In order to do so, notes for which the timing can be extracted with a relatively high confidence level are marked as "anchor notes". In a second pass, the system then tries to refine the timings of the remaining notes by combining the expected position between the anchors with spectral information.

Section 2 gives an overview of related work. We explain the extraction of anchor notes in Section 3, and describe the refinement method applied to the notes between such anchors in Section 4. Section 5 presents our experimental results and Section 6 provides the conclusion and an outlook on future work.

2. RELATED WORK

Online versus offline differentiation aside, state-of-the-art audio-to-score alignment algorithms can be clustered into two main approaches. One is based on statistical, graphical models built from the score, such as those in [1,2,11]. The other one uses Dynamic Time Warping (DTW) in order to align sequences of features calculated from both the audio and the score representation [5,8].

The latter method normally uses chroma vectors as feature, resulting in relatively robust global alignments. However, their temporal accuracy cannot compete with other features which are used in onset detection. In [3], so-called DLNCO-features were introduced, which in essence combine chroma vectors and (pitch-wise) spectral flux. More-

over, a very high temporal feature resolution is used. This is not trivial, since DTW is of complexity $O(n^2)$, and computational costs constrain the number of frames that are aligned. A multi-scale approach introduced in [8], however,allows the temporal resolution to be increased iteratively.

Another combination of chroma-based alignment with onset detection was presented in [6]. Here, an initial alignment is used in order to train an onset detector. Results from the onset detection are then used iteratively to refine the alignment and to train a better onset detector on this more accurate data. This allows the use of supervised machine learning techniques without the need for external training data.

In [7] and [9], results of a DTW-based alignment are refined in a second pass. Both approaches place a search window around the tentative onset time of a note. This window is then scanned for a compatible note onset. While the first method relies on an STFT spectrogram, the latter uses a dictionary-based decomposition of the spectrum in order to differentiate between spectral energies induced by individual notes.

Like the method proposed here, these two approaches can distinguish between the onsets of notes that occur at the same discrete time within the score. This is different from most other systems, since it is inherent to both the DTW algorithms and to the graphical models used in [2, 11] that they work on features representing discrete time steps and that they cannot differentiate between two events occurring within the same time step.

3. ANCHOR NOTE ESTIMATION

At first the anchor notes are extracted using a two-pass procedure as proposed in [9]. In the first step, a state-of-the-art audio-to-score alignment based on chroma vectors and Dynamic Time Warping (DTW) is performed. Then, a dictionary of tone models is used in order to extract each note's activation function. Notes for which a significant rise in activation energy can be found near the corresponding estimated onset are selected to serve as anchors.

3.1 Chroma Features and DTW

In [5], chroma vectors were found to perform best amongst several features in the context of audio matching and alignment. They consist of 12 elements per time frame corresponding to the pitch classes (C, C#, D,...). The values are calculated from an audio recording by mapping the frequency bins of a short-time Fourier transform to pitch class indices i using

$$i = \left[\text{round}\left(12\log_2\left(\frac{f_k}{440}\right)\right) + 9\right] \mod 12 \qquad (1)$$

where f_k represents the center frequency of the k^{th} bin. The term inside the brackets gives the number of the pitch ($A4 \mathrel{\widehat{=}} 0$) that is nearest to f_k, and applying the module gives the pitch class. The summand 9 shifts the indices such that $i = 0$ corresponds to the pitch class C. The actual values of the chroma vectors are then obtained by summing up the energies of all bins mapped to a certain pitch class.

There are two approaches to calculating chroma features from score representations [5]. One of them is to render the score using a software synthesizer to reduce the problem to the one described above. The other method calculates the chroma vector directly from the score. Here, the mapping becomes trivial, since the pitches are already given. However, one must make assumptions about pitch energies and either use constant energies or a decay model. In our experiments, we compared both methods – the first one using the free software synthesizer *timidity* [1] , and the second one using constant midi note energies. Preliminary experiments showed that the resulting alignments did not differ significantly between the two approaches. Therefore, we decided to use the second one, since it is much cheaper in terms of computational costs.

Given two sequences of feature vectors, a cost function must be defined which accounts for the error made when aligning one specific frame within the first sequence to another specific frame within the other sequence. Our experiments showed that the Euclidean distance yields better results than other possible measures, such as the cosine distance.

Based on the cost function, a similarity matrix SM can be constructed. The rows of this matrix represent the time frames of the audio recording, whereas the columns represent the time frames of the score. Hence, the value of each cell SM_{ij} contains the cost of aligning the i^{th} frame of the audio signal to the j^{th} frame of the score. Any continuous, monotonic path through this matrix that begins and ends at the two end-points of the main diagonal represents a valid alignment between the two sequences. The objective is to minimize the global alignment cost, i.e., the sum of all local costs SM_{ij} along the path through the similarity matrix.

Using DTW, the optimal alignment is calculated in two steps. The forward step starts at $[0, 0]$ and the corresponding cost $SM_{0,0}$. Then, all other optimal partial alignments ending with the i^{th} frame of the audio recording aligned to the j^{th} frame of the score are obtained by recursively building a second matrix $Accu$ according to

$$Accu(i,j) = \min \begin{cases} Accu(i-1, j-1) + SM_{ij} \\ Accu(i-1, j) + SM_{ij} \\ Accu(i, j-1) + SM_{ij} \end{cases} \qquad (2)$$

In the forward step, another matrix stores which of these three options has been used in order to advance to the next cell. As soon as the end point $[N-1, M-1]$ has been reached, this information is utilized to reconstruct the path, i.e., the optimal alignment. A more detailed description of the DTW algorithm is given in [10].

[1] http://timidity.sourceforge.net

3.2 NMF and Anchor Selection

The global alignment resulting from the DTW is robust. However, local inaccuracies are inherent to the algorithm. Therefore, an additional feature based on non-negative matrix factorization (NMF) is used to reestimate the onset of each individual note.

NMF is the decomposition of one matrix V of size $m \times n$ into two output matrices W and H of sizes $m \times r$ and $r \times n$, respectively, such that all elements of W and H are non-negative and

$$V \approx WH \tag{3}$$

Applied to audio processing, such a decomposition of a spectrogram results in a dictionary W of r weighted frequency groups and the corresponding activation energies H of these frequency groups over time.

Here we use a modification, as described in [12] and [9], in which W is set to a pretrained set of tone models. These models are computed from audio recordings of single tones played on a piano by, in essence, taking each bin's weighted average energy over the time span where the tone is sustained. The weight of a frame is the inverse of the amplitude envelope to compensate for different loudnesses.

Assuming a fixed W, only H is estimated. Since the pitch described by an individual tone model is known, the i^{th} column of H is a feature representing the activation energy of each pitch in time frame i.

To improve the onset time estimates, a search window of length l is centered around the onset time t_{dtw} obtained by the DTW algorithm. Within this search window a factorization is performed using a dictionary W consisting of tone models of all those pitches that are expected to be played within that time span and an additional white noise component in which the energies are spread uniformly over all frequency bins. A new onset time candidate t_{nmf} is then obtained by choosing the time frame with the largest increase in energy of the pitch under consideration. In contrast to t_{dtw}, t_{nmf} can deviate from other notes with the same score time.

When thinking of repeated notes or of fast passages in which a certain pitch is played several times within the search window, it becomes obvious that this method is too simple to yield meaningful results. However, estimating the onsets of repeated notes is a relatively hard problem in itself. Spectral energy of a sustained note weakens the indicators for the onset of a new note if they have the same pitch. Under these circumstances, algorithms are likely to get mislead by onsets of other notes with overlapping harmonics. This fact makes such notes ineligible to be anchor notes, as a high confidence in the exact estimation of the onset time is essential. Thus, all notes which are played twice or even more often within the time span of the search window, as determined from the score, are discarded from the anchor candidates.

Likewise, all notes are dropped from the list of anchor candidates, for which the initial onset estimate t_{dtw} and the estimate given by the factorization-based feature t_{nmf} differ by more than a certain time span which could have plausibly been caused by an arpeggio or a simple asynchrony. This is justified because such a conflict decreases the confidence in the onset estimation. Moreover, there is no safe way to give either t_{dtw} or t_{nmf} a preference over the other. On the one hand, t_{dtw} is supposed to be more robust, since much more context information is incorporated. On the other hand, t_{nmf} is not bound by the constraints inherent to the DTW algorithm, and therefore able to yield more accurate results [9].

In summary, the two times t_{dtw} and t_{nmf} are calculated by the DTW algorithm and finding the maximum slope within the factorization-based pitch activation. A note is then selected as an anchor if the following two criteria are met:

1. $|t_{dtw} - t_{nmf}| <$ threshold
2. there are no other notes of the same pitch within $t_{dtw} \pm l/2$

In our experiments, we used an STFT with window and hop sizes of 4095 and 1024 frames, respectively, to compute the chroma features from the audio signal. In order to extract chroma vectors from the score, window and hop sizes had been scaled such that the overall number of frames and the overlap ratio remained unchanged relative to the audio representation. Since the DTW misplaces only a negligible fraction of all notes by more than a second, we chose 2.0 seconds for the size l of the search window. Within this search window the hop size was decreased to 256 frames. The maximum difference $|t_{dtw} - t_{nmf}|$ allowed between the two onset estimates was set to 20 frames, i.e., a little more than a tenth of a second.

An evaluation of the extraction of anchor notes is presented in Section 5.

4. NOTE REFINEMENT

After extracting the anchor notes, the remaining notes must be revised. For each of them (with the exception of notes played before the first or after the last anchor notes) the span of time during which it can be played is clearly constrained by the preceding and the successive anchor.

4.1 Beta distribution

In addition to a new search window, bounded by the nearest anchors, rhythmic information in the score can be used to make even more detailed predictions on where to look for an onset. Therefore, the numbers or fractions of beats between the anchor notes and the note under consideration are extracted and their relation is transferred onto the timescale of the audio recording. To account for inexactnesses of the anchor extraction and expressive tempo changes, the "expectation strength" of the onset occurring at time t is modeled by a beta distribution [2] . The beta distribution is defined continuously on the interval $[0, 1]$ and

[2] The beta distribution was chosen for pragmatic reasons (the flexibility of its shape and its restriction to a fixed interval) rather than for precise probability-theoretic reasons.

zero outside this range. Depending on the values of its parameters α and β, the density function can take several forms, for example, that of a uniform distribution, it can be strictly increasing or decreasing, U-shaped, or – as in our case – it is unimodal ($\alpha > 1$ and $\beta > 1$). Its density function is defined as

$$f(x)_{\alpha,\beta} = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{4}$$

where B is the beta function

$$B(\alpha,\beta) = 2\int_0^{\pi/2} \cos^{2\alpha-1}\theta \sin^{2\beta-1}\theta \, \mathrm{d}\theta \tag{5}$$

Mode $\hat{x}$ and variance σ^2 of the distribution are therefore given by

$$\hat{x} = \frac{\alpha-1}{\alpha+\beta-2} \tag{6}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{7}$$

In our application, we set the parameters α and β by fixing a mode $\hat{x}$ and a variance σ^2. The former is assumed to be at the onset time we expect according to score and anchor notes. Since the density function is only defined on $[0,1]$, we use a linear projection to convert between the domain of the beta distribution and the score time.

The variance is chosen such that it allows for expressive variations and inexactnesses of the anchor extraction, but prevents notes from being placed at rhythmically unreasonable timings. Experiments showed that the value $\min(\hat{x}, 1-\hat{x})/20$ results in plausible expectation strengths.

Two such functions are depicted in Figure 1. The upper plot shows the onset likelihood for the onset time of the third note, assuming that the first and fifth note are anchors. The time span between the anchor comprises three beats. Since the note should be played after the first out of these three beat-to-beat intervals, the function is clearly skewed. This is desirable because a musician's freedom of expressive timing is greater when the score calls for longer inter-onset intervals. The second function is the likelihood of the fourth note's onset time given notes number one and six as anchors. The function is now symmetric, since the onset time given by the score is exactly half the time span (two out of four beat-to-beat intervals).

In order to transfer these expectation strength functions from the score into the audio domain, another linear projection is applied.

4.2 Onset estimation

To extract revised onset estimates for non-anchor notes, we calculate the constant Q spectrogram over the time span in which the onset likelihood as described above is greater than zero. The parameters of the constant Q spectrogram are chosen such that each energy bin corresponds to a specific pitch. The hop size is set to 256 frames, resulting in a very high overlap ratio at the lower bins.

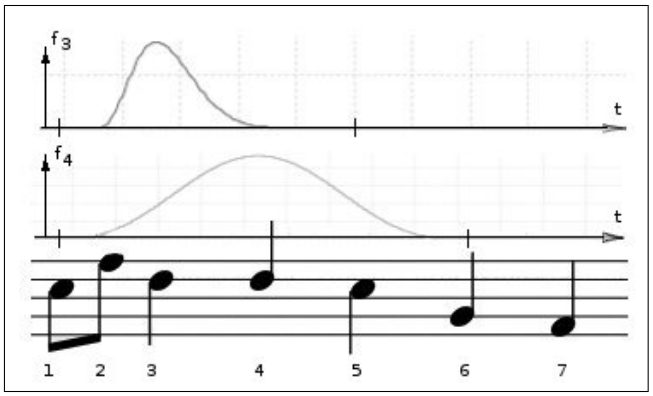

Figure 1. Onset expectation strength for the 3^{rd} and 4^{th} note.

For the purpose of onset detection, energy changes are calculated and half-wave rectified. In order to incorporate the score information, the result is then weighted by the expectation strength. The final onset is set to the time corresponding to the maximum of this detection function.

5. EXPERIMENTAL RESULTS

5.1 Evaluation Method

Since this work was done in the context of musical performance and style analysis, we used classical (polyphonic) piano music to evaluate our system. The test data consisted of the first movements of 11 Mozart sonatas played by a professional pianist. The overall performance time amounted to more than one hour, comprising more than 30.000 notes. The instrument used for the performance was a computer-controlled Bösendorfer SE290 grand piano, which enables exact logging of all events such as keys being hit or released and changes in pedal pressure.

The evaluation was done using mechanical scores represented in MIDI format and the real audio recording from the performances as input data. The data recorded by the Bösendorfer SE290 served as ground truth. The main evaluation criterion was the absolute timing displacement between aligned notes and the ground truth.

On the one hand, robustness and a high overall accuracy are important issues. On the other hand, our work is directed towards providing methods for semi-automatic audio annotation. One objective of such a system must be to minimize human input. In post-processing, the user must correct the onset time as soon as there is a noticeable error. Therefore, we investigated not only the median and percentile errors, but also how many of the notes were detected well enough for a human to accept it.

In [4], listening tests showed that the human hearing system does not detect timing variations of up to 10 ms in sequences of short notes, and even greater displacements in sequences of very long notes. Therefore, our evaluation criteria were the proportions of notes aligned with a displacement of less than 10 ms and 50 ms respectively. The 50 ms tolerance was included because it is a common margin in onset detection.

piece	duration	# notes	# anchors	$50\% < x$[ms]		$75\% < x$[ms]		$95\% < x$[ms]		max [ms]	
				non.a.	anch.	non.a.	anch.	non.a.	anch.	non.a.	anch.
K.279-1	4:55	2803	1136	15.2	5.5	29	11	138	37	879	494
K.280-1	4:48	2491	1257	23.2	5.4	45	11	165	46	687	664
K.281-1	4:29	2648	1235	23.7	6.1	48	12	176	48	993	442
K.282-1	7:35	1907	705	23.8	6.9	60	13	439	72	4805	3008
K.283-1	5:22	3304	1130	16.2	7.9	28	13	75	34	673	467
K.284-1	5:17	3700	1223	15.2	6.1	31	14	120	71	1000	502
K.330-1	6:14	3160	1176	16.3	5.6	30	10	179	35	960	835
K.332-1	6:02	3470	1017	23.2	11.8	42	19	171	82	857	632
K.333-1	6:44	3774	1471	17.8	7.5	31	13	132	38	941	404
K.457-1	6:15	2993	1086	22.0	8.9	42	16	317	62	1773	1787
K.475-1	4:58	1284	483	38.4	16.3	98	24	304	115	4471	2663

Table 1. Comparison between accuracy (median, 75^{th} percentile, 95^{th} percentile, and maximum) of the anchor notes (anch.) and the non-anchor notes (n.a.)

piece	# non-anchors	$50\% < x$[ms]	$75\% < x$[ms]	$95\% < x$[ms]	max [ms]
K.279-1	1667	9.1	28	127	879
K.280-1	1234	9.2	24	147	706
K.281-1	1413	11.2	31	187	1035
K.282-1	1202	15.9	42	432	4822
K.283-1	2174	12.0	21	92	464
K.284-1	2477	9.0	26	125	1004
K.330-1	1983	9.6	21	134	835
K.332-1	2453	18.0	30	175	781
K.333-1	2303	12.1	22	93	1000
K.457-1	1907	16.5	37	246	1790
K.475-1	812	24.1	49	398	4377

Table 2. Accuracy of non-anchor notes after the refinement step (median, 75^{th} percentile, 95^{th} percentile, and maximum)

5.2 Evaluation Results

Table 1 presents the results of the anchor detection step. About a third of the overall notes were chosen as anchors. Although this seems to be a very large fraction, it is justified by the high accuracy of the selected notes. For half of the pieces, the 95th percentile still met the 50 ms criterion used for the evaluation of onset detection algorithms.

However, for each piece a small number of outliers were picked as well. Some of them are due to our trade-off between a small search window at the NMF calculation and computational costs. Notes for which the initial alignment deviates from the real onset by more than a second are post-processed using a time frame that does not even contain the correct onset.

Table 2 shows that, in comparison to Table 1, a majority of non-anchor notes were improved by the refinement step. Both the median deviation and the 75th percentile improved for all the pieces. Only the accuracy of the outliers decreased further in some cases. This might be due to poor anchor notes, which mislead onset detection.

The overall result as given by Table 3 shows the potential of the proposed method. It clearly outperformed the reference algorithm from [9] in which the initial alignment and the factorization-based post-processing were done in a similar way but without using score information to refine critical notes. Especially the proportion of note onsets identified with a deviation of less than 10 ms – i.e., the threshold of human perception, according to [4] – was increased significantly from 40.0% to 49.8%. This is important for the construction of data acquisition tools which are able to extract descriptions of musical expression from audio recordings semi-automatically.

6. CONCLUSION AND FUTURE WORK

We have proposed a multi-pass method for the accurate alignment of musical scores to corresponding audio recordings. The main contribution is the introduction of an expectation strength function modeling the expected onset time of a note between two anchors. Although results are encouraging, there are specific circumstances where the algorithm fails, i.e., temporal displacement of notes is large.

One class of such errors are poor alignments at a piece's ending. There, two disadvantageous factors coincide. On the one hand, there is no additional subsequent note which could serve as hint for the alignment or as anchor in the post-processing. On the other hand, a high degree of polyphony in combination with long and soft notes is to be expected at endings. Such passages are inherently difficult to handle from a signal processing point of view.

An interesting example of such an error can be found in the sonata K.282, in which one note was even wrongly picked as an anchor although it was out of place by more

piece	# notes	$50\% < x$[ms]		$75\% < x$[ms]		$95\% < x$[ms]		$x < 10\,ms$		$x < 50\,ms$	
		ref.	new	ref.	new	ref.	new	ref.	new	ref.	new
K.279-1	2803	12	7.2	27	18	101	103	43.2%	61.7%	88.4%	90.2%
K.280-1	2491	14	7.1	34	16	127	93	42.5%	63.1%	85.0%	90.8%
K.281-1	2648	15	8.5	36	19	112	114	38.5%	56.8%	83.4%	89.9%
K.282-1	1907	15	11.8	44	27	380	378	39.2%	43.5%	76.8%	83.2%
K.283-1	3304	12	10.2	26	18	65	70	44.2%	49.1%	92.2%	92.4%
K.284-1	3700	13	8.0	29	21	98	110	41.7%	58.2%	87.2%	87.7%
K.330-1	3160	11	7.6	24	15	124	103	46.7%	61.0%	89.7%	91.2%
K.332-1	3470	18	16.0	37	27	147	148	32.5%	29.7%	82.7%	87.9%
K.333-1	3774	13	9.9	20	18	80	68	42.2%	50.5%	90.1%	92.8%
K.457-1	2993	15	13.4	35	26	257	183	35.9%	38.2%	83.2%	84.8%
K.475-1	1284	24	20.1	75	37	393	376	23.6%	22.5%	66.8%	78.6%
all	31534	14	10.1	32	21	137	121	40.0%	49.8%	85.6%	88.9%

Table 3. Overall accuracy of the proposed anchor-based method (new) compared to the reference method as described in [9] (ref.)

than three seconds. The explanation is, that the last two chords of this piece differ by only one single note (b♭-a♭-d-f and e♭-a♭-d-f, respectively). The algorithm was not able to distinguish the two chords. As a consequence, the notes of the last chord were aligned to the onset of the preceding chord as well. The resulting temporal displacement of about three seconds is slightly shorter than the duration of the first of these chords.

This clearly leads further work towards the issues of improved mechanisms for anchor detection and the handling of inherently "difficult" passages, such as the endings. An approach that could benefit both fields is the introduction of a more sophisticated local confidence or fitness measure for arbitrary sections of an alignment.

Another aspect which has not been considered yet is the detection of deviations from the score, such as when the pianists adds ornamentations or playing errors occur.

7. ACKNOWLEDGEMENTS

This research is supported by the Austrian Federal Ministry for Transport, Innovation and Technology, and the Austrian Science Fund (FWF) under project numbers TRP 109-N23, P19349-N15, and Z159.

8. REFERENCES

[1] A. Cont: "Realtime Audio to Score Alignment for Polyphonic Music Instruments Using Sparse Non-negative Constraints and Hierarchical HMMs", *Proceedings of the IEEE International Conference in Acoustics and Speech Signal Processing (ICASSP)*, Toulouse, 2006.

[2] A. Cont: "A coupled duration-focused architecture for realtime music to score alignment", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 99(1), 2009.

[3] S. Ewert and M. Müller: "Refinement Strategies for Music Synchronization", *Proceedings of the 5th International Symposium on Computer Music Modeling and Retrieval (CMMR 2008)*, Copenhagen, 2008.

[4] A. Friberg and J. Sundberg: "Perception of just noticeable time displacement of a tone presented in a metrical sequence at different tempos", *Proceedings of the Stockholm Music Acoustics Conference*, pp. 39–43, Stockholm, 1993.

[5] N. Hu, R. B. Dannenberg, and G. Tzanetakis: "Polyphonic Audio Matching and Alignment for Music Retrieval", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, 2003.

[6] N. Hu and R. B. Dannenberg: "Bootstrap Learning for Accurate Onset Detection", *Machine Learning*, Vol. 65, No. 2, pp. 457–471, 2006.

[7] Y. Meron and K. Hirose: "Automatic alignment of a musical score to performed music", *Acoustical Science and Technology*, Vol. 22, No. 3, pp. 189–198, 2001.

[8] M. Müller, H. Mattes, and F. Kurth: "An Efficient Multiscale Approach to Audio Synchronization", *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*, Victoria, 2006.

[9] B. Niedermayer: "Improving Accuracy of PolyphonicMusic-to-Score Alignment", *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)*, Kobe, 2009.

[10] Rabiner, L. R. and Juang, B.-H. "Fundamentals of speech recognition". Prentice Hall, Englewood Cliffs, NJ, 1993.

[11] C. Raphael: "Aligning Music Audio with Symbolic Scores Using a Hybrid Graphical Model", *Machine Learning*, Vol. 65 (2-3), pp. 389–409, 2006.

[12] F. Sha and L. Saul: "Real-time pitch determination of one or more voices by nonnegative matrix factorization", *Advances in Neural Information Processing Systems 17*, K. Saul, Y. Weiss, and L. Bottou (eds.), MIT Press, Cambridge, MA, 2005.