IMPROVING ACCURACY OF POLYPHONIC MUSIC-TO-SCORE ALIGNMENT

Bernhard Niedermayer
Department for Computational Perception
Johannes Kepler University Linz, Austria
bernhard.niedermayer@jku.at

ABSTRACT

This paper presents a new method to refine music-to-score alignments. The proposed system works offline in two passes, where in the first step a state-of-the art alignment based on chroma vectors and dynamic time warping is performed. In the second step a non-negative matrix factorization is calculated within a small search window around each predicted note onset, using pretrained tone models of only those pitches which are expected to be played within that window. Note onsets are then reset according to the pitch activation patterns yielded by the matrix factorization. In doing so, we are able to resolve individual notes within a chord. We show that this method is feasible of increasing the accuracy of aligned note's onsets which are already aligned relatively near to the real note attack. However it is so far not suitable for the detection and correction of outliers which are displaced by a large timespan. We also compared our system to a reference method showing that it outperforms bandpass filtering based onset detection in the refinement step.

1. INTRODUCTION

Opposed to blind audio analysis there are several applications where the recording of an already known piece of music has to be analysed. These applications range from computational musicology, especially performance analysis, and pedagogical systems to augmented audio players and editors as well as special query engines. Knowing that a huge number of symbolic transcriptions of classical as well as modern pieces are publicly available, this leads to the task of automatic music-to-score alignment.

Most current approaches are based on a local distance measure – mainly chroma vectors or features derived from chroma vectors – to compare the similarity between one time frame of the audio and one time frame of the score representation. These distances are then used by a global optimization algorithm, usually Dynamic Time Warping (DTW) or Hidden Markov Models (HMM), which finds the best matching alignment between these two feature sequences.

Recently much attention has been drawn on online algorithms for audio-to-score alignment, also known as score following, like described in [1]. However less work has focused on improvements of the accuracy of offline algorithms. In this paper we present ongoing work towards accurate measurement of individual notes' parameters. The calculation of accurate alignments is not only of use for the above mentioned applications but can also provide training and test data for less informed tasks like blind audio transcription [2].

We propose a two-pass system where in the first step a standard alignment routine based on chroma vectors and DTW is performed. In the second step this alignment is refined using a non-negative matrix factorization (NMF) approach. For each note a search window is set around the estimated note onset. With each of theses windows an NMF using pretrained tone models of only those notes excepted to occur within the respective audio segment plus a noise component is performed. In doing so, the system is able to resolve individual note onsets within whole chords.

We will show that this method provides a good means of refining the estimated onset times of notes that are relatively well detected by standard alignment. However in hard cases where the alignment deviates considerably from the ground truth the method shown here is prone to errors as well.

Section 2 is a brief overview of related work. In Sections 3 and 4 we explain the first alignment step and the NMF-based refinement respectively. Section 5 contains a description of the evaluation method used as well as the experimental results before we conclude our work in Section 6.

2. RELATED WORK

Much work, including [2–5], has focused on audio-to-score alignment based on acoustic features and Dynamic Time Warping (DTW). In [6] chroma vectors, Pitch Histograms, and two Mel-Frequency Cepstrum Coefficient (MFCC) related features have been compared in the context of DTW based audio matching and alignment. It was shown that chroma vectors perform significantly better than the other features.

Since DTW applied on two sequences of length n is of complexity $O(n^2)$ in time as well as in space the resolution of the features used is limited by runtime as well as memory constraints. One way of refining audio alignments is to increase this resolution while keeping computational costs within reasonable bounds. This is done by multi-scale approaches like described in [5] or [7] where the resolutions are increased iteratively but on the other hand search paths are constrained by tentative solutions found so far.

The resolution based refinement does not overcome an important side effect of alignments based on dynamic time warping. Notes that are struck together in the score, like it is the case for chords, can not be treated independently. This is a major drawback in applications like performance analysis, where the accurate timing of individual chord notes is an important expressive characteristic. [8] and [9] use pitch specific energy levels in order to estimate the timings of individual notes.

Another method to iteratively refine audio alignments is a bootstrap approach as described by [4]. There an audio segmenter is trained on an initial alignment. This segmenter can produce a refined alignment which is then used for a repeated training step. This method allows for the application of supervised machine learning techniques without the need for external training data.

Non-negative matrix factorization, as used here, was first applied to audio alignment in [10]. There, the combination of NMF and Hidden Markov Models was able to create alignments for polyphonic instruments in realtime.

3. BASIC ALIGNMENT

3.1 Chroma Feature

In the first pass the proposed system performs a state-of-the-art audio-to-midi alignment based on chroma vectors and Dynamic Time Warping. Chroma vectors have 12 elements representing the single pitch classes (i.e. C, C#, D, D#,...). The values are calculated based on a short time Fourier transform. Each frequency bin is then related to the index i of a pitch class by

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

where f_k is the center frequency of the k^{th} bin. The tuning frequency is supposed to be 440 Hz but can easily be changed to any other value. The summand 9 shifts the vector such that the pitch class C has index 0. The individual values are then obtained by summing up the energies of all bins corresponding to a certain pitch class.

A similar feature that yields comparable results has been suggested by [11] which on the one hand takes only bins containing energy peaks into account but on the other hand also considers harmonics. At the extraction of the so called Harmonic Pitch Class Profile the energy of a frequency bin k does not only contribute to the pitch class best matching the center frequency f_k but also to those pitch classes best matching f_k/h with $h = 2, 3, 4 \ldots$ This accommodates for the assumption that the energy in bin k can also represent the h^{th} harmonic of a pitch. Since the energy of a partial decreases with the order of the harmonic, an additional weighting factor of $w_{harm} = d^{h-1}$ with $0 < d \leq 1$ is introduced.

The calculation of the chroma representation based on a MIDI file instead of audio data is straightforward since each MIDI event can be directly assigned to the corresponding pitch class. However when using the Harmonic Pitch Class Profile, errors are made when letting the energy of the actual f_0 contribute to the pitch classes corresponding to $f_0/3$, $f_0/5$,... This inexactness has to be reproduced in order to obtain equivalent representations of audio and score. Likewise when using default chroma vectors, contributions of a note to other pitch classes than the one corresponding to the f_0 caused by harmonics can be considered as well.

Preliminary experiments have shown that chroma vectors and Harmonic Pitch Class Profiles yield comparable results. Therefore chroma vectors have been used for the remainder of this work due to computational advantages.

3.2 Dynamic Time Warping

Based on this chroma representation a globally optimal alignment is calculated. Therefore a sequence of chroma vectors for the audio file as well as for the score representation is calculated. In doing so the score MIDI is divided into time frames such that the overall number of frames and the overlap ratio between frames is the same as of the STFT applied on the audio data. The Euclidean distance is used to compute a similarity matrix SM comparing each frame of one feature sequence to each frame of the other sequence, after all feature vectors have been normalized. Mapping corresponding frames to each other is the same as finding a minimal cost path through this similarity matrix. A path through SM_{ij} is then equivalent to the alignment of frame i of the score feature sequence to frame j of the performance feature sequence. Dynamic time warping (DTW) is a well-established dynamic programming based algorithm that finds such optimal paths. A detailed tutorial can be found in [12].

In order to get meaningful results an alignment path has to meet several constraints.

Continuity The constraint of continuity forces a path to proceed through adjacent cells within the similarity matrix. Jumps would be equal to skipping frames without considering the costs of this operation.

Monotonicity The constraint of monotonicity in both dimensions guarantees that the alignment has the same temporal order of events as the reference sequence.

End-point constraint The end-point constraint forces the ends of the path to be the diagonal corners of the similarity matrix. In doing so it is assured that the alignment covers the whole sequences.

The optimal path according to DTW is calculated in two steps. The forward step starts a partial path at the point

$[0,0]$ and rates it with the cost SM_{ij}. Then it calculates the minimum path costs for all other partial alignments ending with frame i of the score being aligned to frame j of the recorded performance in a recursive manner according to equation 2.

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

The three options correspond to partial paths ending with a diagonal step, an upwards step, and step to the right within the similarity matrix SM. In addition to the actual local distances, weights w_d and w_s are needed to yield reasonable path costs. If there were no such weights, diagonal paths would be strongly favored over straight ones which are twice as long. Experiments have shown that the values 1.4 and 1.0 (still giving diagonal steps a preference over straight ones) perform well. In our implementation we do this cost calculation in place, i.e. overwriting the values SM_{ij} by $Accu(i,j)$ in order to save memory space.

The backtracking step of DTW starts as soon as all values $Accu(i,j)$ have been calculated. $Accu(N-1,M-1)$ is the minimal cost of a complete alignment between the two feature sequences. Therefore the optimal path is reconstructed starting from $[N-1,M-1]$ going back to $[0,0]$. In order to be able to do so, a second matrix is built during the forward step, memorizing whether the last step leading to a point $[i,j]$ was diagonal, upwards, or to the right.

4. NMF-BASED REFINEMENT

4.1 Non-negative Matrix Factorization

Within the last few years non-negative matrix factorization (NMF) has become of increasing interest in the domain of blind audio transcription. The basic idea is that an input matrix V of size $m \times n$ is decomposed into two output matrices W and H of size $m \times r$ and $r \times n$ respectively where the elements of all these matrices are strictly non-negative and

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

Assuming that V represents real-world data such factorizations will most likely not be perfect. The reconstruction error caused by any deviation of WH from V can be measured by a cost for which the Euclidean distance or the I-divergence are common choices. In minimizing this cost function, W and H are learned as an initially determined number r of basis vectors and their activation patterns over time respectively.

Performing such a decomposition on a spectrogram, as obtained by a short time Fourier transform, will result in a dictionary W of weighted frequency groups and their occurrence H over time. According to the input V and the parameter r, the base components in W will, in the ideal case, represent models of single pitches or chords played on a certain instrument. But due to the unsupervised nature of the method, elements of W might as well correspond to special frequency patterns during the attack, sustain, or decay phase of a note, single partial or just noise.

However, as soon as the piece and its score are known, as it is the case in the context of audio alignment, the instrument(s) used to perform the piece are most probably known as well. So there is no need to learn a set of base components. Instead a number r of tone models can be trained in advance which overcomes the above mentioned uncertainty of unsupervised learning. Also the number and kind of tone models can be adjusted to the respective piece.

With only H being left unknown Equation 3 can be rewritten as

$$v \approx \overline{\overline{W}} \cdot h \tag{4}$$

where $\overline{\overline{W}}$ is the fixed dictionary of tone models. v and h are single column vectors of V and H that can now be processed independently, which leads to a much simpler decomposition task [13]. The vectors h are very sparse in nature and represent an f_0 estimation for the corresponding frame.

Throughout this work the mean square criterion given as

$$c_{err} = \frac{1}{2} \parallel \overline{\overline{W}}h - v \parallel_2^2 \tag{5}$$

is used as cost measure for factorization errors since computationally efficient algorithms for its optimization are available [14].

4.2 Tone Model Training

In order to get meaningful factorizations at least one tone model per possible pitch has to be contained in $\overline{\overline{W}}$. Given a set of training samples, such tone models can be trained in advance using the same method as described above. In the ideal case those training samples are audio recordings of single pitches played on a certain instrument. Starting from Equation 3 again, W and H become vectors w and h since there is only one basis component present ($r = 1$). h can further be approximated by the amplitude envelope, leaving only w to be unknown. The actual computation is then done by the same implementation as used during the performing step of the algorithm.

Throughout this work we use an additional basis component representing white noise. Experiments have shown that such a noise model significantly improves the alignment results.

4.3 Local Refinement

In the first stage of the proposed system a music-to-score alignment has already been performed. The advantage of this alignment is that it is globally optimized and very robust. However independent from all parameters that can be set, accuracy is limited by the fact, that such an alignment algorithm can never differentiate between notes that are struck together in the score.

To overcome this limitation and still preserve high robustness we define a search window of length l around the initially estimated onset time. Within this local context the refinement step tries to find the exact temporal position of each individual (chord-)note. The parameter l has been chosen to be 2 seconds since preliminary evaluation of the first alignment step has shown that only a marginal number of outliers deviates from the ground truth by more than a second.

For each such search window the contained notes and their pitches are determined in order to define the tonal context of the note under consideration. This information is used to build a dictionary $\overline{W}_{local}$ made up by tone models describing only those pitches that are present within the local context plus an additional (white) noise component. The resulting activation patterns H are smoothed using a median filter and used in order to extract following features for each time frame.

Activation energy Since activation patterns H are very sparse in nature (even when sparsity is not enforced), activation energies greater than zero are strong indicators for note positions.

Energy slopes The first derivative of the activation energy corresponds to energy changes. Positive slopes as they occur at note onsets are filtered by half wave rectification.

Relative energy slopes Since transients at note onsets are characterized by energy burst across the whole spectrum, other pitches – especially ones with shared harmonics – might show low activation energies during such phases as well. Therefore the increases in energy of the pitch under consideration in relation to the overall frame energy is also taken into account.

Experiments have shown that the maxima of the derivatives are good predictors for note attacks while the maximal activation energy itself has turned out to be less significant. Comparing the slope of the absolute energy to the one of the relative energy revealed a slight advantage of the relative energy derivative which was therefore chosen as onset detection criterion.

5. EXPERIMENTAL RESULTS

5.1 Evaluation Method

We limit our evaluation to classical piano music using a database consisting of the first movements of 11 Mozart sonatas played by a professional pianist. The performance was done on a computer monitored Bösendorfer SE290 grand piano, producing an automatic MIDI transcription of the exact ground truth of played notes as well as pedal events. Aligning a single movement instead of a whole sonata at a time is a valid simplification since individual movements are per default separate tracks on audio CDs. Nevertheless the overall performance time of this test set is still about one hour containing more than 30.000 notes.

The tone models used for the NMF-based refinement have been learned from single tones played on the same grand piano. Since such a recording was not available for each pitch, the missing models have been acquired by simple interpolation.

For evaluation purpose we calculated an alignment for each piece using the audio recording of the expressive performance and a mechanical score representation in MIDI format. We compared the resulting onset times to our given ground truth data and took the absolute displacement as evaluation criterion. This evaluation was done for the initial alignment step only as well as for the whole system including the refinement.

Initial alignments were done using a short time Fourier transform (STFT) with a window length of 4096 samples and a hop size of 441 samples, which corresponds to a time resolution of 100 frames per second. For the refinement step a search window of radius one second was used and the STFT hop size was reduced to 256 samples, resulting in time frames of a length of 5.8 ms.

First experiments with this setup have shown that although the calculation of the factorization base feature is narrowed down to a small search window as well as a small pitch range, it is still not as robust as expected. About 10% of the notes have not been detected by the factorization step and therefore left unchanged during refinement.

Concerning the remaining notes it turned out to be the best strategy to only modify those notes where the initial alignment position and the timing resulting from refinement are approximately consistent. This is the case for about half of the overall number of notes. In situations where these two onset candidates differ by more than 20 frames (i.e. 116 ms) a conflict is detected – although its resolution has been left to future work. One cause for such conflicts are repeated notes which cannot be handled by the simple detection mechanism as described above.

5.2 Evaluation Results

In Table 1 the limits of the quartiles as well as the 95^{th} percentile are given. Within the first three quartiles the refinement has improved results for each individual piece. However concerning notes that are displaced by more than 100 ms in the initial alignment tend to be displaced even further by the refinement step.

For most applications a transcription is good as soon as a human listener can not distinguish it from the original. This implies that in the context of music-to-score alignment a note can be counted as correctly aligned if its deviation from the ground truth is less than the just noticeable difference of the human perception. In an experimental environment, where listeners were asked to adjust the timing of one tone within a series, such that the inter-onset intervals became perfectly regular, this just noticeable difference was investigated [15]. It was found to be around 10 ms for notes shorter than 250 ms and about 5% of the note duration for longer ones.

Therefore an evaluation based on this criterion was done as well. In Table 2 the amount of notes with a time dis-

piece	# notes	duration	$25\% < x$		$50\% < x$		$75\% < x$		$95\% < x$	
			bas.	ref.	bas.	ref.	bas.	ref.	bas.	ref.
kv279-1	2803	4:55	7 ms	5 ms	16 ms	12 ms	30 ms	27 ms	103 ms	101 ms
kv280-1	2491	4:48	11 ms	5 ms	23 ms	14 ms	42 ms	34 ms	126 ms	127 ms
kv281-1	2648	4:29	12 ms	6 ms	24 ms	15 ms	42 ms	36 ms	114 ms	112 ms
kv282-1	1907	7:35	10 ms	6 ms	23 ms	15 ms	53 ms	44 ms	337 ms	380 ms
kv283-1	3304	5:22	7 ms	5 ms	15 ms	12 ms	27 ms	26 ms	62 ms	65 ms
kv284-1	3700	5:17	7 ms	6 ms	15 ms	13 ms	31 ms	29 ms	97 ms	98 ms
kv330-1	3160	6:14	7 ms	5 ms	15 ms	11 ms	28 ms	24 ms	118 ms	124 ms
kv332-1	3470	6:02	9 ms	7 ms	20 ms	18 ms	39 ms	37 ms	138 ms	147 ms
kv333-1	3774	6:44	8 ms	5 ms	16 ms	13 ms	29 ms	20 ms	79 ms	80 ms
kv457-1	2993	6:15	10 ms	6 ms	19 ms	15 ms	37 ms	35 ms	214 ms	257 ms
kv475-1	1284	4:58	13 ms	11 ms	30 ms	24 ms	78 ms	75 ms	360 ms	393 ms
all	31534	1:02:39	8.3 ms	5.6 ms	18 ms	14 ms	35 ms	32 ms	132 ms	137 ms

Table 1. Comparison between accuracy after the basic alignment step (bas.) and the additional refinement (ref.)

piece	$x < 10\,\text{ms}$		$x < 50\,\text{ms}$	
	bas.	ref.	bas.	ref.
kv279-1	33.8%	43.2%	88.2%	88.4%
kv280-1	22.4%	42.5%	81.5%	85.0%
kv281-1	20.1%	38.5%	80.4%	83.4%
kv282-1	25.3%	39.2%	73.7%	76.8%
kv283-1	36.2%	44.2%	92.6%	92.2%
kv284-1	34.6%	41.7%	86.9%	87.2%
kv330-1	35.5%	46.7%	89.9%	89.7%
kv332-1	27.1%	32.5%	83.0%	82.7%
kv333-1	31.5%	42.2%	90.1%	90.1%
kv457-1	27.3%	35.9%	82.5%	83.2%
kv475-1	20.0%	23.6%	63.9%	66.8%
all	29.6%	40.0%	84.8%	85.6%

Table 2. Comparison between accuracy after the basic alignment step (bas.) and the additional refinement (ref.)

placement less than 10 ms is shown for the initial and the refined alignment. According to the chosen STFT time resolution this corresponds to a deviation of one frame at maximum. In addition the number of notes having a displacement error less than 50 ms is given as well since this is a common evaluation criterion in onset detection.

Again it is shown that the refinement improves those notes already aligned relatively close to their real onset. The amount of notes with displacement errors less than 10 ms was increased from about 30% to 40% while the number of notes with errors below 50 ms was only moderately changed from 84.8% to 85.6%.

5.3 Feature comparison

From the list of related work presented in section 2, [8] is the one that presents the approach which is most similar to the system proposed here. There onset detection by selective bandpass filtering is described in the context of score supported audio transcription. According to this method a note is found by summing up the energy in all frequency bands corresponding to the f_0 as well as the harmonics of a pitch and then finding a maximum in the derivative of this indication function. In order to avoid the influence of other pitches with overlapping harmonics, partials that collide with those of an other note struck at the same time are neglected.

	fact.	s.b.f.
$25\% < x$	5.6 ms	10.0 ms
$50\% < x$	14 ms	20 ms
$75\% < x$	32 ms	40 ms
$95\% < x$	137 ms	128 ms
$x < 10\,\text{ms}$	40.0%	24.9%
$x < 50\,\text{ms}$	85.6%	81.3%

Table 3. Comparison between refinement based on factorization (fact.) and based on selective bandpass filtering (s.b.f.) [8]

We have compared our system to an own implementation of this approach. In doing so, we used the same computational framework and only exchanged the factorization feature in the refinement step by this onset detector based on selective bandpass filtering. The accumulated results on the whole test set are shown in Table 3. It demonstrates that bandpass filtering yields results less accurate than those produced by NMF, and mostly even less accurate than those achieved by the alignment based on chroma vectors. A possible reason is that the STFT based version of selective bandpass filtering relies on just a few frequency bins while NMF takes the whole spectrogram into account.

6. CONCLUSION AND FUTURE WORK

We have introduced a new method to increase accuracy of music-to-score alignments by a two-pass system. Whereas the first step consists of a state-of-the-art alignment using chroma features and dynamic time warping the second step is a refinement based on non-negative matrix factorization.

We have shown that this refinement step performs very well on notes which have already been detected relatively close to their real onset time by the alignment step. The number of notes placed with a time deviation below the just noticeable difference according to [15] of 10 ms has been increased from about 30% to 40%. This is remarkable since so far only those notes without any conflicting features have been modified.

However the method does not bring any improvements for notes where the deviation of the initial alignment from the ground truth is large. On one hand the refinement step only works within a search window which should be kept as small as possible. Notes that are misaligned such that the actual onset is out of this window can never be corrected by the method described here. On the other hand chroma features as well as factorization based pitch separation rely on prominent energy peaks in the spectrogram. If the spectrogram is blurred due to heavy use of pedal or very rich polyphony both approaches are prone to errors.

This clearly dictates future work to concentrate on the problem of detecting and handling possible outliers and 'hard' regions. The most obvious approach is to develop a method of handling conflicting features as this is the case for about 40% of all notes. We think that introducing a tempo model and enforcing reasonable inter-onset intervals entails the potential of further improvements.

Also the 10% of notes that have not been covered by the factorization based feature are worth being reconsidered. Standard STFT favors the detection of higher pitches due to its linear frequency scale. Additional spectral transformations like multi-rate filterbanks or a constant-Q transform could help to enhance the note detection, especially within low pitch ranges.

7. ACKNOWLEDGMENTS

The research presented in this paper is supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung (FWF) under project number P19349-N15.

8. REFERENCES

[1] S. Dixon: "Live Tracking of Musical Performances Using On-Line Time Warping", *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx)*, Madrid, 2005.

[2] R. J. Turetsky and D. P. W. Ellis: "Ground-Truth Transcriptions of Real Music from Force-Aligned MIDI Syntheses", *Proceedings of the 4th International Symposium of Music Information Retrieval (ISMIR)* Baltimore, MD, 2003.

[3] 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.

[4] N. Hu and R. B. Dannenberg: "A Bootstrap Method for Training an Accurate Audio Segmenter", *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, London, 2005.

[5] M. Müller, F. Kurth, and T. Röder: "Towards an efficient algorithm for automatic score-to-audio synchronization", *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*, Victoria, 2006.

[6] 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.

[7] N. Adams, D. Marquez, and G. Wakefield : "Iterative Deepening for Melody Alignment and Retrieval", *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, London, 2005).

[8] E. D. Scheirer: "Using Musical Knowledge to Extract Expressive Performance Information from Audio Recordings", *Readings in Computational Auditory Scene Analysis*, H. G. Okuno and D. F. Rosenthal (eds.), Lawrence Erlbaum Publication, Mahweh, NJ, 1997.

[9] M. Müller, F. Kurth, and M. Clausen: "Audio Matching via Chroma-based Statistical Features", *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)*, Barcelona, 2005.

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

[11] E. Gómez and P. Herrera: "Automatic Extraction of Tonal Metadata from Polyphonic Audio Recordings", *Proceedings of 25th International AES Conference*, London, 2004.

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

[13] 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.

[14] Lawson, C. L. and Hanson, R. J. "Solving least squares problems", *Prentice Hall*, Lebanon, Indiana, 1974.

[15] 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.