OFF-LINE REFINEMENT OF AUDIO-TO-SCORE ALIGNMENT BY OBSERVATION TEMPLATE ADAPTATION

Cyril Joder, Björn Schuller

Machine Intelligence & Signal Processing Group, MMK
Technische Universität München, Germany

ABSTRACT

Audio-to-score alignment aims at matching a symbolic representation (the score) to a musical recording. A key problem in this application is the great variability of audio observations which can be explained by a single symbolic element. Whereas most previous works deal with this problem by training or heuristic design of a generic observation model, we propose the adaptation of this model to each musical piece. We exploit a template-based formulation of the observation model and we investigate two strategies for the adaptation of the templates using a Hidden Markov Model for the alignment.

Experiments run on a large dataset of popular and classical piano music show that such an approach can lead to a significant improvement of the alignment accuracy compared to the use of a single generic model, even if the latter is trained on real data.

Index Terms— music processing, audio-to-score alignment, model adaptation

1. INTRODUCTION

Audio-to-score alignment aims at the synchronization of a musical recording and its corresponding score. This task can lead to multiple interesting applications, such as multi-modal browsing of musical pieces [1] or score-informed source separation [2, 3].

The alignment process usually uses an instantaneous matching measure evaluating the match between each element of the audio and the score. These measures are then combined with possible structural constraints, to obtain the aligned sequences. Regarding the design of the matching measure, the main difficulty is the diversity of timbres that can correspond to a single element of the score [4].

To overcome this problem, many studies exploit audio representations designed to be robust to timbre changes, such as chroma representations [5, 6]. Other types of information have also been exploited, such as note onsets [7] or tempo information [8, 9].

Some works use a prior learning of the observation model, with statistical [10–12] or template-based approaches [13, 14]. In the latter approach, a template is built for each symbolic element, as the superposition of single-note templates. The drawback of this solution is the need for training data for all the encountered instruments. Hence, many systems resort to heuristic forms [5, 15]. A learning of generic templates has recently been investigated [16, 17], operating a trade-off between the timbres of the training database. However, it could be beneficial to use templates adapted to the very recording which is processed, so as to precisely model the instruments involved. To our knowledge, this approach has only been explored in [18], where the latent variables of a Hidden Markov Model (HMM) are estimated on each musical piece using a *nonparametric Bayesian method*. However in such a method, the number of unknown parameters, and thus the complexity, is very high.

In the present paper, we investigate strategies for the adaptation of the observation model, in the framework of a template-based matching measure. Hence, this method can benefit from a generic learning approach and further refine the matching measure on each particular recording with a limited complexity. We evaluate the effectiveness of our approaches with two different alignment systems, on a large database of popular and classical polyphonic music. The results show that the adaptation procedure improves the precision of the alignments, compared to the initial generic templates.

The rest of this paper is organized as follows: the template-based model is presented in Section 2 and two adaptation approaches are detailed out in Section 3. We evaluate the proposed approaches in Section 4, before suggesting some conclusions and discussing the relation with prior work.

2. OBSERVATION MODEL

As mentioned above, audio-to-score alignment requires a quantitative measure of the match between each element of the score and the recording. As in [17, 19], the score elements are the *concurrencies*, defined as the largest units of constant content (in terms of notes) in the score. The audio representation consists of some time-frequency representation such as chromagram or semigram [16]. With these representations, one can assume that the feature vectors extracted from the superposition of several notes are (approximately) equal to the superposition of the corresponding single-note vectors.

Hence, one can build a template vector corresponding to any concurrency of a musical score, from a set of single-note template vectors. Let $\mathbf{W}$ be the matrix whose columns are the single-note templates of all possible pitches. Since it can be considered as realizing a mapping from the symbolic domain to the observation domain, this matrix is called *mapping matrix*. For a concurrency c, we can define h_c as the vector whose components correspond to the number of notes of each pitch. We can then define the concurrency template $u_c = \mathbf{W}h_c$.

The matching measure can be calculated as a simple distance between the concurrency template and the actual observation vector. The value of this measure $f(v_n, c)$, comparing an observation v_n to a concurrency c is then defined as:

$$f(v_n, c; \mathbf{W}) = D\left(v_n, \mathbf{W}h_c\right), \qquad (1)$$

where $D(\cdot, \cdot)$ is some distance of dissimilarity function.

The formulation of (1) can also be found by assuming a generative model similar to [20]. In this model, we assume that components $v(i)$ of the observation vector, given that a note of pitch j is played, follow independent Poisson distributions, whose parameters are the elements $\mathbf{W}_{i,j}$ of the mapping matrix. Furthermore, we assume an additive model, where the observation vector is the sum of independent random vectors corresponding to each played single note.

Then, the conditional probability of an observation, given that the concurrency c is played, can be written as

$$P(V|c;\mathbf{W}) \propto e^{-f(c,V;\mathbf{W})}, \tag{2}$$

where f is here a particular case of the matching measure of (1) using the generalized Kullback-Leibler (KL) divergence as the dissimilarity function. This divergence is defined as

$$D_{\mathrm{KL}}(x\|y) = \sum_k x_k \log\left(\frac{x_k}{y_k}\right) - x_k + y_k. \tag{3}$$

3. ADAPTATION OF THE MAPPING MATRIX

Before detailing the approaches for the adaptation of the templates to each musical recording, we first define the alignment model used in the adaptation process. We exploit an HMM model, where the hidden states represent the concurrencies played at each time frame. For complexity reasons, we choose a prior model similar to the Markovian model of [8], which only constrains the concurrency sequence to follow the same structure as in the score. We assume that the concurrencies are numbered in the order in which they appear in the score. The transition probabilities from concurrency c to concurrency c' is defined as:

$$p_{c,c'} = \begin{cases} \frac{1}{2} \text{ if } c' = c \\ \frac{1}{2} \text{ if } c' = c+1 \\ 0 \text{ otherwise,} \end{cases} \tag{4}$$

and the initial probabilities force the initial state to be the first concurrency of the score (representing an initial silence). The observation model used is defined in (2).

3.1. EM-Based Adaptation

The first adaptation strategy aims at estimating the optimal mapping matrix $\mathbf{W}$ according to the Maximum Likelihood (ML) criterion. Let $V_{1:N} = V_1, \ldots, V_N$ be the sequence of observations. The optimization problem is:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}}\, P(V_{1:N};\mathbf{W}) \tag{5}$$

where $P(V_{1:N};\mathbf{W})$ is the marginal probability of the observation of the HMM. However, since this criterion is not convex, we exploit the well-known Expectation-Maximization (EM) algorithm [21]. This corresponds to an iterative procedure, in which the mapping matrix is updated according to

$$\mathbf{W} \leftarrow \underset{\mathbf{W}'}{\operatorname{argmin}} \sum_{n=1}^{N} E\Big[f(C_n, V_n; \mathbf{W}')\Big| V_{1:N};\mathbf{W}\Big] \tag{6}$$

where C_n is the random variable representing the concurrency at frame n and $E[\cdot|\cdot;\mathbf{W}]$ denotes the conditional expectation calculated by the HMM model using the parameter $\mathbf{W}$. The auxiliary function to be minimized can be easily calculated using the forward-backward algorithm. Furthermore, the auxiliary function is convex if the dissimilarity function D of (1) is convex, which is the case with the KL divergence. As the gradient and Hessian matrix can be easily calculated, we use a Newton-based optimization strategy, which locally minimizes the quadratic Taylor approximation of the cost function.

The whole procedure can be iterated until convergence. This algorithm is proven to make the criterion (5) decrease. However, it is known to be sensitive to initialization. Thus, we initialize $\mathbf{W}$ with a generic mapping matrix, estimated by the supervised learning strategy of [17].

3.2. Viterbi-Based Adaptation

The second adaptation approach consists first in decoding the HMM in order to find the optimal concurrency sequence, knowing the observation sequence. Then, a supervised learning approach can be employed, using the decoded concurrencies as 'ground-truth'. The decoding of the model is performed using the *Maximum A Posteriori* (MAP) criterion, employing the initial value of the mapping matrix $\mathbf{W}$. The optimal sequence $\hat{C}_{1:N}$ is then defined as:

$$\hat{C}_{1:N} = \underset{C_{1:N}}{\operatorname{argmax}}\, P(C_{1:N}|V_{1:N};\mathbf{W}) \tag{7}$$

and can be computed by the Viterbi algorithm [21]. The mapping matrix is then updated using the following rule:

$$\mathbf{W} \leftarrow \underset{\mathbf{W}'}{\operatorname{argmin}} \sum_{n=1}^{N} f(\hat{C}_n, V_n; \mathbf{W}'). \tag{8}$$

This update procedure is similar to the one used in [22] for score-informed source separation. As in the previous case, the objective function is convex if the dissimilarity function is convex. The same optimization algorithm as in this case is used. Note that this adaptation procedure can also be seen as an approximate version of the EM algorithm, where the states of the decoded concurrency sequence are assumed to concentrate all the posterior probability. However, there is no guarantee that the criterion (5) increases.

4. EXPERIMENTS

4.1. Experimental Settings

We test our adaptation approaches on three types of audio representations. The first is the Power Spectrogram (PS), derived from a short-time Fourier transform calculated over 100 ms windows. In order to reduce noise due to percussion in the high and low frequencies, we only exploit the frequencies between 100 Hz and 3.6 kHz [17].

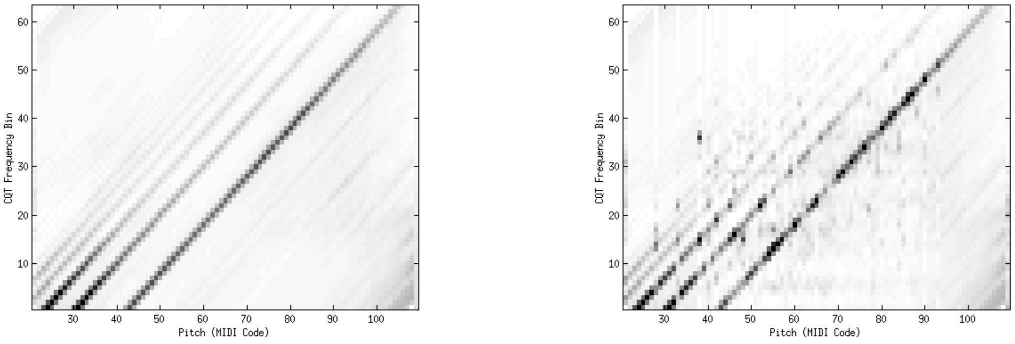
The second representation, called *semigram* (SG) [16], is a spectrum representation with a logarithmic frequency scale corresponding to semitones (12 bins per octave). We calculate this spectrum as magnitude of a Constant Q Transform (CQT) [23]. Similarly to above, only frequencies between 100 Hz and 3.6 kHz are considered.

Finally, the third representation is the chromagram, or Pitch Class Profile (PCP). It is a 12-component vector representation corresponding to the spectral energies of the 12 musical pitch classes. We compute the chromagram according to Zhu's method [24].

In our tests, we compared two dissimilarity functions: The first is the generalized KL divergence of (3) and the second is the symmetric version of this divergence, defined as $D_{\mathrm{KLs}}(x,y) = D_{\mathrm{KL}}(x\|y) + D_{\mathrm{KL}}(y\|x)$. Although the latter function loses the generative interpretation of Section 2, we found that it led to a better accuracy. Thus, the results are here presented for the symmetric dissimilarity function.

The database used in this work consists of 59 classical piano pieces (about 4 h 15 min) from the MAPS database [25, 26] and 90 songs (about 6 h) from the RWC popular music database [27, 28]. The ground-truth is given by aligned MIDI files. 50 pieces (20 from MAPS and 30 from the RWC corpus, 220 min) are used for the learning of the generic mapping matrix. The test set is composed of the remaining pieces, which are to be aligned with tempo-modified versions of the annotation scores[1]. The alignment accuracy is measured

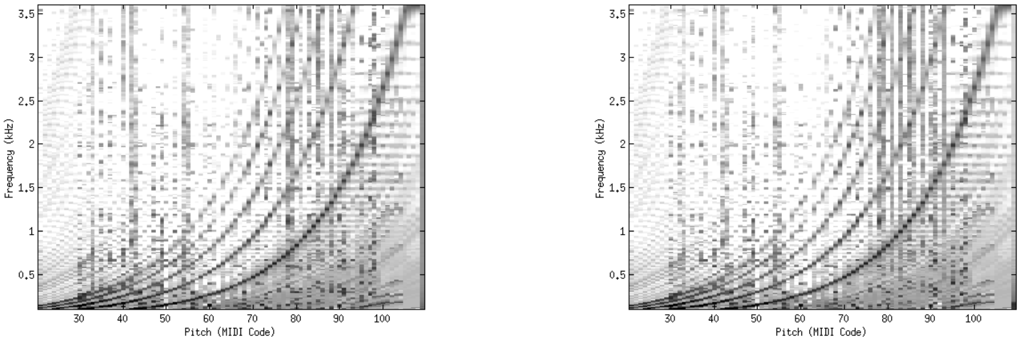
[1] The list of the pieces in training and test sets can be found on `http://www.openaudio.eu/Joder13-OLR.zip`



(a) Original

(b) After one EM iteration

Fig. 1. Effect of the adaptation on the mapping matrix for the semigram (SG) representation. The grayscale is the same for both matrices. Here, the adaptation is performed on a the piano piece `MAPS_MUS-grieg_butterfly_ENSTDkAm` from the MAPS dataset.



(a) EM Adaptation

(b) Viterbi Adaptation

Fig. 2. Comparison of the mapping matrices obtained after one update iteration on the song no. 54 of the RWC-pop dataset, for the Power Spectrogram (PS) representation. The grayscale is the same for both matrices.

by the *alignment rate*, defined as the proportion of concurrency onsets which are detected less than a threshold θ away from their real onset time. The test set contains about 120 000 onsets. We choose $\theta = 100$ ms for a precise evaluation of the alignments.

4.2. Adapted Mapping Matrices

Figure 1 displays an example of adapted mapping matrix for the SG representation. The learning process of the original matrix involves a smoothing procedure over the 'neighboring' pitches, for a better generalization property. However, in our case generalization is not necessary and the estimated templates match the observations of the specific recording. In particular, only the notes appearing in the piece are updated, which explains the 'uneven' look of the adapted mapping matrix. An 'artifact' of the adaptation is visible on the template of pitch 38 (D1), which exhibits a strong energy on a higher frequency bin, which corresponds to the fundamental frequency of pitch 78 (F#4). A closer look reveals that the note D1 only appears twice in the pieces, and both times in conjunction with the note F#4. Hence, although the updated template does not capture the real spectral shape of the isolated note, it complies with the actual audio content.

In Figure 2, the results of both adaptation strategies on the mapping matrix of the power spectrogram representation are compared. In this example, the matrices are adapted to a pop song containing a strong percussion part. This is visible in both matrices, where many updated note templates capture wide-band spectral shapes. One can see some differences between both estimated mapping matrices, especially around pitches 80 to 90 where the Viterbi adaptation seems to be more affected by noise-like percussive spectra. While this is hard to interpret, it could be explained by the 'hard decisions' taken by the Viterbi algorithm as opposed to the calculation of state probabilities in the EM algorithm. Hence, the latter estimation takes into account several possible positions for the update of each note and thus it is less sensitive to these local 'noisy events'.

4.3. Alignment with a Hidden Markov Model

We now evaluate the benefit of the adaptation on the alignment accuracy. In this experiment, the model used for the alignment is the same HMM as in the adaptation step. For each piece, the alignment is performed by decoding the HMM with the Viterbi algorithm in the same manner as in Subsection 3.2. Note that this system is not expected to provide very accurate alignments, since no duration constraint is taken into account. It is rather intended to emphasize on the differences between the observation models.

The influence of the adaptation on the alignment rate is represented in Figure 3, as a function of the number of iterations. We limit the experiments to three iterations of the adaptation algorithms, since

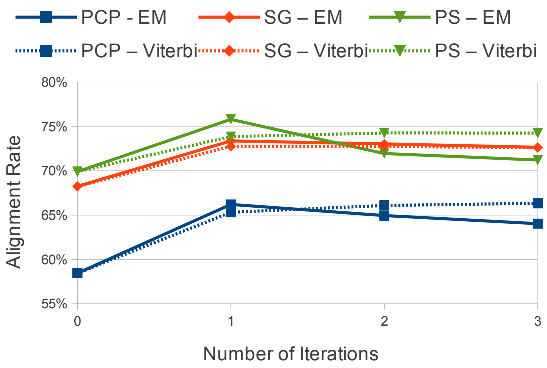


Fig. 3. Alignment with the Hidden Markov Model: alignment rates as a function of the number of adaptation iterations.

Algorithm	HMM		DTW	
Adaptation	none	EM (1 it.)	none	EM (1 it.)
PCP	58.4%	66.2%	68.2%	70.5%
SG	68.2%	73.4%	74.0%	76.6%
PS	69.9%	75.8%	75.1%	77.7%

Table 1. Influence of the adaptation (after 1 EM iteration) on the alignment rates. The 95% confidence intervals (computed for i.i.d. Bernoulli samples) are smaller than 0.3%.

no real improvement is obtained after that. With both adaptation methods, the accuracy increases after the first iteration. This indicates that adaptation is an efficient approach to enhance the alignment quality, whatever the audio representation.

While the performance remains constant or even slightly improves with further iterations of the Viterbi adaptation, the accuracy degrades with the EM algorithm. This is explained by the 'soft decisions' mentioned in the previous subsection. Indeed, the EM algorithm updates all the note templates using each audio observation. Even if the contributions are weighted by the state probabilities, this performs a smoothing of the estimated templates. Thus, increasing the likelihood of the generative model with several EM iterations can actually degrade the discriminative power of the observation model. However, this algorithm proves useful since for all the tested representations, the best results are obtained with a single EM iteration.

4.4. Alignment with Dynamic Time Warping

For an evaluation in a more realistic setting, we perform another set of experiments, using the same alignment strategy as in [5]. The score is first converted into a template sequence, and the alignment is then calculated by the Dynamic Time Warping (DTW) algorithm. Similarly to the previous model, the local cost function employed in the DTW is given by the symmetric KL divergence. This method is expected to show better results than the HMM, since it takes into account the concurrency durations indicated in the score for the creation of the template sequence.

Table 1 summarizes the influence of the adaptation process on the alignment accuracy. As expected, for every considered setting, the present system outperforms the HMM model. As in the previous case, the adapted templates lead to significantly higher scores than the baseline model, indicating that this improvement does not depend on the alignment strategy.

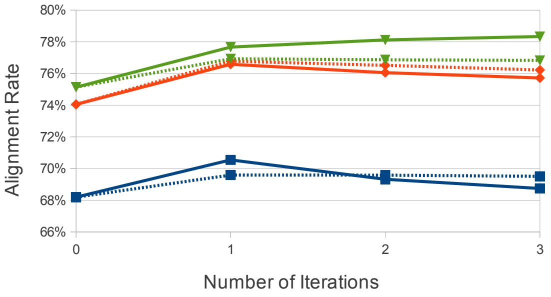


Fig. 4. Alignment with the Dynamic Time Warping algorithm: alignment rates as a function of the number of adaptation iterations.

The alignment rates are displayed in Figure 4 as a function of the number of iterations. These experiments exhibit the same tendencies as previously. The only difference in the relative behavior of the two systems is observed for the EM adaptation of the Power Spectrogram representation. In this case, the alignment rate continues to improve after the first iteration. This is explained by the implicit temporal constraints introduced in the DTW algorithm. Indeed, the DTW favors alignment paths which correspond to the temporal indications of the score and the loss of discriminative power can then be compensated by the temporal priors.

5. CONCLUSION

In this paper, we described an approach for the refinement of audio-to-score alignment by adapting the observation model to each particular musical piece to be processed. We exploited a template-based model, which allowed for a simple formulation of the adaptation procedure as the estimation of a mapping matrix. Two criteria were proposed for this estimation and we evaluated their influence on the alignment precision on a large database of popular and classical polyphonic music. These experiments exhibited a significant improvement of the accuracy with the adapted model compared to the initial generic templates. Furthermore, these improvements were observed on all the tested settings, which indicates that it is independent of the audio representation used, as well as the alignment strategy used.

This work opens several perspectives for improvement. First, one could imagine different criteria for the adaptation of the mapping matrix, for example by constraining the modifications of the estimated templates in order to limit the effect of the percussion. The mapping could also be made instrument-dependent, so as to capture the timbral characteristics of each instrument. Finally, one could imagine to 'adapt the score' to the recording. For example, the intensities and actual lengths of the notes indicated by the score could be estimated, for the analysis of a musical performance.

6. RELATION TO PRIOR WORK

The work presented here has focused on the learning of the observation model for audio-to-score alignment. Most previous studies rely on a prior training of this model, which can be instrument-specific [10–14] or generic [16, 17]. However, these methods requires relevant training data, which are not always available. Maezawa *et al.* [18] jointly estimate the alignment and a parametric observation model on each recording. While we adopt a similar approach, the present work capitalizes on a significantly less complex template-based model.

7. REFERENCES

[1] M. Müller, M. Clausen, V. Konz, S. Ewert, and C. Fremerey, "A multimodal way of experiencing and exploring music," *Interdisciplinary Science Reviews*, vol. 35, no. 2, pp. 138–153, 2010.

[2] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE J. Select. Topics Signal Processing*, vol. 5, no. 6, pp. 1205–1215, Oct. 2011.

[3] C. Joder and B. Schuller, "Score-informed leading voice separation from monaural audio," in *Proc. of Inter. Soc. for Music Information Retrieval Conf. (ISMIR)*, Porto, Portugal, Oct. 2012, pp. 277–282.

[4] M. Müller and S. Ewert, "Towards timbre-invariant audio features for harmony-based music," *IEEE Trans. Audio, Speech, Language Processing*, vol. 18, no. 3, pp. 649–662, Mar. 2010.

[5] N. Hu, R. B. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proc. of IEEE Workshop Applicat. of Signal Processing to Audio and Acoust (WASPAA)*, New Paltz, NY, USA, Oct. 2003, pp. 185–188.

[6] M. Müller, F. Kurth, and M. Clausen, "Audio matching via chroma-based statistical features," in *Proc. of Inter. Soc. for Music Information Retrieval Conf. (ISMIR)*, London, UK, Sept. 2005, pp. 288–295.

[7] S. Ewert, M.Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proc. of IEEE Inter. Conf. Acoust. Speech, Signal Processing (ICASSP)*, 2009, pp. 1869–1872.

[8] C. Joder, S. Essid, and G. Richard, "A conditional random field framework for robust and scalable audio-to-score matching," *IEEE Trans. Audio, Speech, Language Processing*, vol. 19, no. 8, pp. 2385–2397, Nov. 2011.

[9] Bernhard Niedermayer and Gerhard Widmer, "A multi-pass algorithm for accurate audio-to-score alignment," in *Proc. of Inter. Soc. for Music Information Retrieval Conf. (ISMIR)*, 2010, pp. 417–422.

[10] Christopher Raphael, "Automatic segmentation of acoustic musical signals using hidden Markov models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 360–370, 1999.

[11] Arshia Cont, Diemo Schwarz, and Norbert Schnell, "Training IRCAM's score follower," in *Proc. of IEEE Inter. Conf. Acoust. Speech, Signal Processing (ICASSP)*, Philadelphia, PA, USA, Mar. 2005, vol. 3, pp. 253–256.

[12] Christopher Raphael, "Aligning music audio with symbolic scores using a hybrid graphical model," *Machine Learning Journal*, vol. 65, pp. 389–409, 2006.

[13] Arshia Cont, "Realtime audio to score alignment for polyphonic music instruments using sparse non-negative constraints and hierarchical HMMs," in *Proc. of IEEE Inter. Conf. Acoust. Speech, Signal Processing (ICASSP)*, Toulouse, France, May 2006, pp. 245–248.

[14] Bernhard Niedermayer, "Improving accuracy of polyphonic music to score alignment," in *Proc. of Inter. Soc. for Music Information Retrieval Conf. (ISMIR)*, Kobe, Japan, Oct. 2009, pp. 585–590.

[15] Nicola Montecchio and Arshia Cont, "A unified approach to real time autio-to-score and audio-to-audio alignment using sequential monte-carlo inference techniques," in *Proc. of IEEE Inter. Conf. Acoust. Speech, Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 193–196.

[16] Özgür İzmirli and Roger B. Dannenberg, "Understanding features and distance functions for music sequence alignment," in *Proc. of Inter. Soc. for Music Information Retrieval Conf. (ISMIR)*, Utrecht, the Netherlands, Aug. 2010, pp. 411–416.

[17] C. Joder, S. Essid, and G. Richard, "Optimizing the mapping from a symbolic to an audio representation for music-to-score alignment," in *Proc. of IEEE Workshop Applicat. of Signal Processing to Audio and Acoust (WASPAA)*, New Paltz, NY, USA, Oct. 2011, pp. 121–124.

[18] A. Maezawa, H. G. Okuno, T. Ogata, and M. Goto, "Polyphonic audio-to-score alignment based on bayesian latent harmonic allocation hidden markov model," in *Proc. of IEEE Inter. Conf. Acoust. Speech, Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 185–188.

[19] W. P. Birmingham, R. B. Dannenberg, G. H. Wakefield, M. A. Bartsch, D. Mazzoni, C. Meek, M. Mellody, and W. Rand, "Musart: Music retrieval via aural queries," in *Proc. of Inter. Soc. for Music Information Retrieval Conf. (ISMIR)*, Bloomington, IN, USA, Oct. 2001, pp. 73–81.

[20] T. Virtanen, A. T.Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. of IEEE Inter. Conf. Acoust. Speech, Signal Processing (ICASSP)*, Las Vegas, NV, USA, Apr. 2008, pp. 1825 –1828.

[21] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[22] R. Hennequin, B. David and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proc. of IEEE Inter. Conf. Acoust. Speech, Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 45–48.

[23] J. C. Brown, "Calculation of a constant q spectral transform," *Journal Acoust. Soc. Am.*, vol. 89, no. 1, pp. 425–434, 1991.

[24] Y. Zhu and M. S. Kankanhalli, "Precise pitch profile feature extraction from musical audio for key detection," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 575–584, June 2006.

[25] V. Emiya, N. Bertin, B. David, and R. Badeau, "MAPS - A piano database for multipitch estimation and automatic transcription of music," Technical report, METISS - INRIA - IRISA , Laboratoire traitement et communication de l'information - LTCI, July 2010.

[26] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. Audio, Speech, Language Processing*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.

[27] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. of Inter. Soc. for Music Information Retrieval Conf. (ISMIR)*, Paris, France, Oct. 2002, pp. 287–288.

[28] M. Goto, "Development of the RWC music database," in *Proc. of Inter. Congress on Acoust.*, Kyoto, Japan, Apr. 2004, pp. 553–556.