# OPTIMIZING THE MAPPING FROM A SYMBOLIC TO AN AUDIO REPRESENTATION FOR MUSIC-TO-SCORE ALIGNMENT

*Cyril Joder, Slim Essid, Gaël Richard*

Institut Telecom, Telecom ParisTech, CNRS-LTCI
37-39 rue Dareau — 75014 Paris, FRANCE
firstname.lastname@telecom-paristech.fr

## ABSTRACT

A key processing step in music-to-score alignment systems is the estimation of the intantaneous match between an audio observation and the score. We here propose a general formulation of this matching measure, using a linear transformation from the symbolic domain to any time-frequency representation of the audio. We investigate the learning of the mapping for several common audio representations, based on a best-fit criterion.

We evaluate the effectiveness of our mapping approach with two different alignment systems, on a large database of popular and classical polyphonic music. The results show that the learning procedure significantly improves the precision of the alignments, compared to common heuristic templates used in the literature.

## 1. INTRODUCTION

In many automatic music analysis tasks, such as audio-to-score alignment [1], chord recognition [2] or automatic transcription [3], the audio information (or a low-level representation directly extracted from it) has to be matched with a symbolic description of the music. To this aim, one calculates a *matching measure* between the audio observations and the possible symbolic events. In a probabilistic model, the matching measure is given by the conditional probabilities of the observations. This can then be combined with temporal constraints (or prior model) in order to favor or penalize certain progressions.

For the audio-to-score alignment problem, all the events of the symbolic representation are already known. Since this information provides strong constraints on the possible alignment paths, many alignment works use a straightforward matching measure and focus on the efficient exploitation of this structural information [4]. Indeed, even in a probabilistic framework, an estimation of the observation distributions has seldom been attempted. To our knowledge, only [5] and [6] describe a learning of these conditional probabilities in the context of audio-to-score alignment. However, these works deal with instrument-specific learning of monophonic data. In the polyphonic multi-instrumental case, most of the proposed probabilistic models exploit heuristic forms for the observation model. These strategies often boil down to a template-based approach, where the score elements are mapped into the acoustic descriptor domain and then compared to the audio observations thanks to some distance function [4].

A similar mapping is studied by İzmirli and Dannenberg in [7] in the case where both the pitch vector and the audio representation are projected into a 12-dimension space. They show that the canonical chroma mapping is not the most effective one for the task of discriminating aligned and non-aligned frames. However, they focus on "chroma-like" audio representations and the evaluation is performed on a classification task. In the present paper, we generalize this idea to any time-frequency representation of the audio, thanks to a formulation of the matching measure as a linear transformation of a "piano-roll-like" representation of the score. We then propose a best-fit strategy for the learning of the mapping matrix using several common audio descriptors. Experiments conducted on a large database of popular and classical polyphonic music show that this learning can improve the accuracy of an alignment system.

The rest of this paper is organized as follows. The formulation of the matching measure is presented in Section 2 and the heuristic mappings used for the common audio representation of the literature are detailed in 3. The learning strategy for the learning of the mapping from the symbolic to the pitch observation domain is exposed in the following section. Finally, we evaluate the influence of these mappings on the alignment accuracy of two alignment systems in Section 5, before suggesting some conclusions.

## 2. THE MATCHING MEASURE

In this work, we are interested in measuring the match between an audio recording and a score on the basis of the pitches played. We define the *concurrencies* of the score as the largest temporal units of constant (multi-) pitched content [8]. Hence, each audio frame will be compared to each score concurrency. Assuming that the range of a musical piece does not exceed the range of the grand piano (from A0 to C8), we define the "pitch vector" representation of a musical score by numbering the possible pitches from 1 to 88, following the chromatic scale. Each component of the pitch vector $h_c$, associated to a concurrency $c$, is then the number of notes of the corresponding pitch in the concurrency. In order to take into account the portions of the signal where no note is played (in the case of silence or unpitched sounds), we introduce an additional component in the pitch vector, which is equal to 1 iff all the other notes are inactive. Thus, the dimension of a pitch vector is $J = 89$.

For a time-frequency representation of the audio recording (such as power spectrogram or chromagram), let $v_n$ be the vector extracted from frame $n$ of the recording. The value of the matching measure $f(v_n, c)$ between observation $v_n$ and concurrency $c$ has the form:

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

where $D(\cdot\,,\cdot)$ is some divergence function and $\mathbf{W}$ is a $I \times J$ matrix, $I$ being the dimension of the observation vectors. In order to be robust to level dynamics, the observations and the pitch vectors are normalized so that their sum is unitary. The matrix $\mathbf{W}$ operates

This work has been partly supported by the Quaero Program, funded by OSEO, French State agency for innovation.

978-1-4577-0693-6/11/$26.00 ©2011 IEEE

| Acronym | Meaning |
|---|---|
| PS | Power Spectrum |
| FBSG | FilterBank Semigram |
| CQTSG | CQT Semigram |
| MPCP | Müller's PCP (from filterbank) |
| ZPCP | Zhu's PCP (from CQT) |

Table 1: Summary of the pitch representations tested.

as a linear mapping from the pitch vector domain to the observation domain and each column of $\mathbf{W}$ can be seen as a pitch template.

Note that in practice, all the considered values are non-negative. Thus, this formulation is very close to the Non-negative Matrix Factorization (NMF) problem [9]. However, in the alignment case, the matrix $\mathbf{W}$ is fixed and the set of possible pitch vectors $h_c$ is finite.

Any distance function can be used in (1). However, in the present work, we only use the symmetric Kullback-Leibler divergence, whose expression is

$$D\left(v,u\right)=\sum_{i=1}^{I}v(i)\log\left(\frac{v(i)}{u(i)}\right)+u(i)\log\left(\frac{u(i)}{v(i)}\right). \quad (2)$$

Other kinds of distance functions have been considered in some preliminary tests, including the Itakura-Saito divergence and the cosine distance. However, they did not prove more efficient than the KullBack-Leibler divergence in terms of alignment accuracy.

## 3. HEURISTIC MAPPINGS FOR COMMON AUDIO REPRESENTATIONS

The present section details the heuristic mapping used with each of the audio representations (or acoustic descriptors) considered in this work. Table 1 sums up these representations.

### 3.1. Power Spectrum

For the power spectrum representation, the mapping can be constructed as in [4]. A pitch is represented as a Gaussian mixture in the spectral domain, whose components correspond to the first $K$ harmonics. In our experiments, the weights of the partial components are proportional to $1/k^2$ where $k$ is the harmonic index and the variances are set to 30 cents.

The Fourier transform is 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 4 kHz.

### 3.2. Semigram Representation

The semigram representation [7] is a spectrum representation with logarithmically spaced frequency bins corresponding to the semitones of the musical scale (12 bins per octave). Two methods for calculating this representation are tested here. The first one, called FilterBank SemiGram (FBSG) is composed of the short-term energy at the output of elliptic filters as in [10]. We also use the magnitude of a constant Q transform (CQT). In this case, in order to maintain a good temporal precision, only frequencies over 100 Hz are considered. We also limit the highest frequency bin to 4 kHz. This representation is referred to as CQT Semigram (CQTSG).

As in [11], the mapping is obtained by associating to each pitch a binary template, where the positive elements are the bins of the harmonics. In this work, we only consider the first two harmonics, as is visible on Fig. 1 (left).

### 3.3. Chromagram Representations

The chromagram (also called Pitch Class Profile) is a 12-component vector representation corresponding to the spectral energies of the 12 musical pitch classes (A, A#,...). We use here two different algorithms to obtain them. The first one, proposed by Müller [10] is the integration of the *FBSG* features over the different octaves. The second chroma representation is calculated according to Zhu's method [12]. These representations are denoted respectively by MPCP (for Müller's Pitch Class Profile) and ZPCP (Zhu's).

The canonical chroma template of a pitch is a binary vector whose only positive component is the chromatic class of the pitch.

### 3.4. Noise Template

In practice, a uniform component is superposed to the presented templates to model background noise. The proportion of noise used is dependent on the representation, and has been set according to preliminary experiments. The last column of $\mathbf{W}$, representing the absence of pitched sound, is assigned this noise vector.

## 4. LEARNING OF THE MAPPING MATRIX

Although the heuristic mappings presented in the previous section are reported to yield good performances [1, 4], one might wonder if they could be improved by some learning from real musical data. Hence, we explore a best-fit strategy for the learning of $\mathbf{W}$, using the Minimum Divergence (MD) criterion.

### 4.1. Formulation

The chosen approach consists in finding the mapping matrix which optimizes the matching measure, defined in (1), on the training data. Let $v_1^s \dots v_{N_s}^s$ and $h_1^s \dots h_{N_s}^s$ be respectively the audio observations and the ground-truth pitch vectors of the $s$-th training sequence (of length $N_s$). The optimal matrix $\hat{\mathbf{W}}^{\text{MD}}$ is defined by:

$$\hat{\mathbf{W}}^{\text{MD}}=\underset{\mathbf{W}}{\operatorname{argmin}}\sum_{s}\sum_{n=1}^{N_s}D\left(v_n^s,\mathbf{W}h_n^s\right). \quad (3)$$

We also add a non-negativity constraint on matrix $\mathbf{W}$. This both prevents the matching measure of (2) from diverging and maintains an intuitive interpretation of this matrix.

It can be easily shown that, with the symmetric Kullback-Leibler divergence, the cost function of (3) is convex w.r.t. $\mathbf{W}$. We use a trust-region optimization strategy, which locally minimizes the quadratic Taylor approximation of the objective function. The solution of the quadratic minimization problem is then approximated thanks to the method exposed in [13].

### 4.2. Database

The database used in this work comprises 59 classical piano pieces (about 4h15 of audio data), from the MAPS database [3] and 90 pop songs (about 6h) from the RWC database [14]. The ground-truth annotation is given by aligned MIDI files. The training database is composed of 50 randomly selected pieces (220 min), 20 from MAPS and 30 from the RWC corpus. In order to reduce overfitting

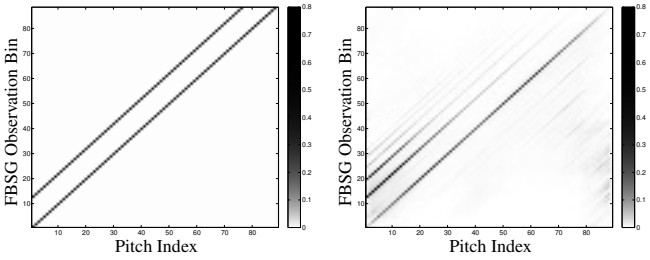



Figure 1: Comparison of two mapping matrices, for the FBSG representation. Left: heuristic matrix; right: learned matrix.

to specific pitches or keys, 12 versions of each piece are used in the training process, by transposing both the audio observations and the pitch vectors up to $-6$ and $+5$ semitones. The remaining pieces of both datasets are used for the evaluation. In this case, target scores are tempo-modified versions of the ground-truth scores.

### 4.3. Obtained Mapping Matrices

The learned mapping matrix for the FBSG representation is represented in Fig. 1 and compared with the corresponding heuristic mapping. In this example, it is visible that the one-note templates (i.e. the columns of the matrix) select not only the fundamental frequency and the first harmonic as the heuristic templates do, but also a number of higher partials. Moreover, the weights given to these partials are not uniform since they depend on the note. Hence, the learning process actually captures a kind of average partial energy distribution over the different pitches.

## 5. INFLUENCE ON THE ALIGNMENT ACCURACY

Two different alignment approaches are tested, in order to evaluate the influence of the mapping in an alignment task.

### 5.1. Alignment with no Duration Constraint

The first system corresponds to the simplest model in [8]: given the observation vector sequence $\underline{\mathrm{v}}_{1:N}$ (of length $N$) extracted from the audio recording, the alignment is performed by searching the optimal concurrency sequence $\hat{\underline{\mathrm{c}}}_{1:N}$, defined by:

$$\hat{\underline{\mathrm{c}}}_{1:N} = \underset{\underline{\mathrm{c}}_{1:N} \in \mathcal{C}}{\operatorname{argmin}} \sum_{n=1}^{N} f\left(c_n, v_n\right), \tag{4}$$

where $\mathcal{C}$ is the set containing all the possible concurrency sequences, that is the sequences where the concurrencies follow the same order as in the score. This method can be seen as a Dynamic Time Warping (DTW) alignment approach [1], where every concurrency lasts a single frame in the "score sequence". We call it the *unconstrained* alignment strategy, since no duration constraint is taken into account. It is not expected to provide very accurate alignments, but rather to emphasize the differences between the matching measures.

As in [8] the alignments are evaluated using the *recognition rate*, defined as the proportion of concurrency onsets which are detected less than a threshold $\theta$ away from their real onset time. We choose $\theta = 100$ ms, for a precise evaluation of the alignments.

The obtained recognition rates are displayed in Table 2. Note that the 95% confidence intervals are smaller than 0.3%. It is clear from these results that learning the mapping matrix does improve the alignment accuracy. Indeed, for all the tested representations, a significant increase of the recognition rates (between +1.3% and +3.6%) is observed, compared to the heuristic templates.

| Mapping | PS | FBSG | CQTSG | MPCP | ZPCP |
|---|---|---|---|---|---|
| Heuristic | 66.3 | 60.4 | 64.9 | 52.4 | 56.9 |
| Learned | 69.9 | 61.7 | 68.2 | 54.6 | 58.6 |

Table 2: Recognition rates (in %) of the unconstrained alignment system with all the tested representations.

| Mapping | PS | FBSG | CQTSG | MPCP | ZPCP |
|---|---|---|---|---|---|
| Heuristic | 79.0 | 75.3 | 75.7 | 65.1 | 70.6 |
| Learned | 79.9 | 75.6 | 78.1 | 68.0 | 73.2 |

Table 3: Recognition rates obtained with a DTW alignment strategy.

We can also observe the relative efficiencies of the representations. The best result is obtained by the power spectrum representation (69.9%). Then the semigrams induce a higher accuracy than the chromagrams. This can be explained by the reduction of dimensionality in the latter representations, which leads to a loss of information. Nevertheless, the chroma representations remain interesting, since they have the potentiality of an improved robustness, especially to octave errors. Finally, the representations based on a Constant Q Transform (CQTSG and ZPCP) outperform the filterbank-based representations (FBSG and MPCP). This can be explained by the smaller bandwidth of the used filters, which can overly penalize pitch imprecisions. Another reason to this is the level of noise in the low frequencies, which can be very high when a bass drum is present. Thus, a good solution is sometimes to simply discard very low frequencies, which is the case in our CQT.

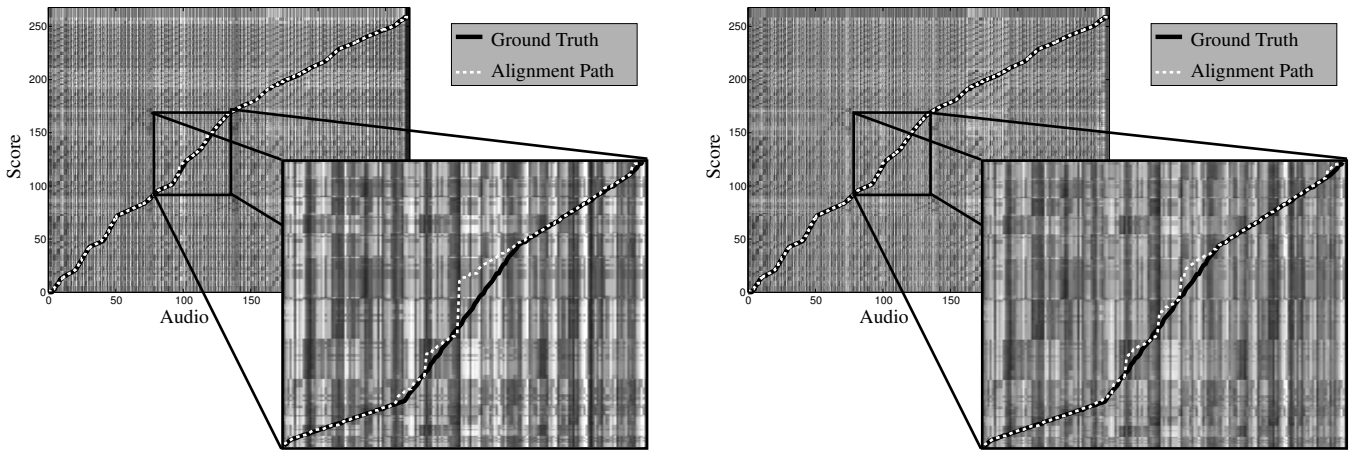
### 5.2. Dynamic Time Warping (DTW) Alignment Strategy

For an evaluation on a more realistic setting, we run another alignment experiment using the Dynamic Time Warping (DTW) algorithm, such as in [1]. The score is first converted into a template sequence, which takes into account the score durations (contrary to the previous approach). Then we compute the "similarity matrix", containing the matching measures between the score templates and the audio observations. Finally, the alignment is performed by calculating the continuous path in this matrix which optimizes the cumulative matching measure.

The obtained recognition rates are displayed in Table 3. As a comparison, the DTW system of [15] obtains a recognition rate of 67.1%. All of our settings, except for the heuristic mapping with the MPCP representation, outperform this system. The best results are as high as 79.9%, which show the efficiency of the presented framework, even with heuristic mappings.

The relative results of the audio representations are the same as in the previous experiments. Besides, the learning also allow the DTW systems to increase the alignment accuracy, up to a 3.5% improvement for MPCP. This indicates that the observed differences between the settings are not dependent on the alignment strategy. Thus, optimizing the mapping from symbolic to audio representations does have the potential to improve any alignment system.

Fig. 2 displays an example of the obtained alignments on a pop song from RWC. On this example, the similarity matrices are quite




**Heuristic Mapping** **Learned Mapping**

Figure 2: Example of similarity matrices and alignment results on a RWC pop song for both mappings of the power spectrum representation.

"noisy". This is due to the presence of percussion in the audio, as well as the high variability in the mixing levels of the different instruments. Nevertheless, although the differences between both similarity matrices are somewhat subtle, one can observe that the learned mapping induces a smoother alignment path. Indeed, the matching measure is less affected by the variability of the audio (corresponding to the vertical structures of the matrix).

## 6. CONCLUSION AND FUTURE WORK

In this paper, we have described a general template-based approach for the matching of an audio sequence with a score, thanks to a linear mapping from the symbolic domain to a time-frequency representation of the audio. We have taken advantage of the form of this transformation to learn the mapping using a best-fit criterion. The evaluations, performed on a large database of polyphonic music, show that this learning leads to a significant increase of the alignment accuracy. Furthermore, we have compared the usefulness of several representations of the audio in this alignment task. Our results indicate that both the spectrogram and the CQT-based semigram representations provide very precise alignments.

Many perspectives can be imagined for the continuation of this work, including the use of other distance functions. Other learning methods such as the discriminative strategy of [7], or more elaborate, non-linear mappings could also be investigated. Finally, the mapping could be made instrument-dependent. Hence the learning process could capture some timbral characteristics of the instruments, which would constitute an additional clue for the alignment.

## 7. REFERENCES

[1] N. Hu, R. B. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proc. IEEE WASPAA*, 2003, pp. 185–188.

[2] L. Oudre, C. Fevotte, and Y. Grenier, "Probabilistic template-based chord recognition," *IEEE Trans. Audio, Speech, Language Processing*, vol. PP, no. 99, p. 1, 2010.

[3] 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.

[4] A. Cont, "A coupled Duration-Focused architecture for Real-Time Music-to-Score alignment," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 6, pp. 974–987, June 2010.

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

[6] A. Cont, D. Schwarz, and N. Schnell, "Training ircam's score follower," in *Proc. IEEE ICASSP*, vol. 3, 2005, pp. 253–256.

[7] O. İzmirli and R. Dannenberg, "Understanding features and distance functions for music sequence alignment," in *Proc. ISMIR*, 2010, pp. 411–416.

[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 (in press)*, vol. PP, 2011.

[9] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE WASPAA*, Oct. 2003, pp. 177 – 180.

[10] M. Müller, *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

[11] M. Müller, F. Kurth, and T. Röder, "Towards an efficient algorithm for automatic score-to-audio synchronization," in *Proc. ISMIR*, 2004, pp. 365–372.

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

[13] M. A. Branch, T. F. Coleman, and Y. Li, "A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems," *SIAM Journal on Scientific Computing*, vol. 21, no. 1, pp. 1–23, 1999.

[14] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. ISMIR*, 2002, pp. 287–288.

[15] D. P. W. Ellis, "Aligning midi scores to music audio," 2008. [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/alignmidiwav/