# A weakly-supervised discriminative model for audio-to-score alignment

Rémi Lajugie, Piotr Bojanowski, Philippe Cuvillier, Sylvain Arlot, Francis Bach

# A WEAKLY-SUPERVISED DISCRIMINATIVE MODEL FOR AUDIO-TO-SCORE ALIGNMENT

*Rémi Lajugie*[⋆] *Piotr Bojanowski*[†] *Philippe Cuvillier*[*] *Sylvain Arlot*[⋆] *Francis Bach*[⋆]

[⋆] Sierra project team, INRIA/École Normale Supérieure, Paris, France
[†]Willow project team, INRIA/École Normale Supérieure, Paris, France
[*] Mutant project team, INRIA/UMR STMS IRCAM-CNRS-UPMC, Paris, France

## ABSTRACT

In this paper, we consider a new discriminative approach to the problem of audio-to-score alignment. We consider the two distinct informations provided by the music scores: (i) an *exact* ordered list of musical events and (ii) an *approximate* prior information about relative duration of events. We extend the basic dynamic time warping algorithm to a convex problem that learns optimal classifiers for all events while jointly aligning files, using this weak supervision only. We show that the relative duration between events can be easily used as a penalization of our cost function and allows us to drastically improve performances of our approach. We demonstrate the validity of our approach on a large and realistic dataset. **Keywords**: weakly supervised learning, score-following, audio-to-score

## 1. INTRODUCTION

This paper deals with aligning a temporal signal to its associated sequence of symbolic events. Given an audio recording of a musical piece and its music score, the goal is to retrieve the actual duration of each musical event, which may differs from the one provided by the score.

Beyond the interest in itself of tracking live performances, it is also a front-end for many applications in music such as automatic accompaniment [5], audio editing [2], and automatic turning of score pages [4]. This task also called score-following [6, 7] when performed in real-time.

Many state-of-the-art alignment algorithms are elaborations of the standard dynamic time warping (DTW) procedure [8, 9, 4]. Alignment algorithms use the duration information provided by scores together with models for each event, that are pre-designed before running alignments. This step usually involves some ad hoc knowledge, like acoustical models [8, 9, 10, 11], and/or supervised training on *fully-labeled* databases which are expensive to gather. For instance, [12] trains a conditional random field, [13] builds classifiers for each possible musical event using a support vector machine. Our work cannot be compared directly to those, as (i) it only relies on *weakly-supervised* data, *i.e.*, pairs of one audio recording and its score; (ii) it performs both learning and alignment steps *simultaneously*. To do so, we propose to learn an optimal alignment function by minimizing a discriminative square loss criterion.

This work shares deep links with *discriminative clustering* methods [14, 15]. This has recently attracted interest for further applicative domains beyond music, *e.g.*, action localization [16], image co-segmentation [17], video co-localization [18], named entity classification [19], or video-to-text alignment [20]. Discriminative cost functions are usually prone to degenerate solutions. To get rid of them, [20] arbitrarily suggests two penalizations, whereas we motivate their use with the prior information encoded in music scores. We show that these priors can be seamlessly expressed using a proper representation of alignments.

**Contributions** The contributions of the paper are four-fold: **(i)** We cast the set of alignments on a matrix space $\mathcal{Y}$, for which we interpret the dynamic time warping (DTW) algorithm as a linear program solver. **(ii)** We propose a discriminative approach to the alignment problem. It learns an optimal DTW-based alignment function while jointly aligning the inputs. We relax the obtained problem into a convex program, and solve it efficiently with the Frank-Wolfe algorithm. **(iii)** We cast the information about relative duration of events provided by music scores as two different priors. **(iv)** We evaluate our model on a monophonic dataset, prove the benefits of the priors on performances, and show the discriminative approach is robust to intense white noise.

## 2. DISCRIMINATIVE APPROACH

### 2.1. Alignment task

**Notations.** Let us consider an audio recording $X$ that is sampled in $T$ *timestamps*, thus $X \in \mathbb{R}^{T\times p}$. We assume that $X$ is given with its score. Every score consists of an ordered list of $E$ events in a dictionary of individual notes or chords (superposition of notes). Assuming there are $K$ base notes, an event is a subset of $\{1,\dots,K\}$. We represent each event $e$ by a binary indicator vector $\Phi_e \in \{0,1\}^K$, that we concate-

nate in a matrix $\Phi \in \{0,1\}^{E\times K}$. Such a matrix is called a *template*. If the template sums to one along rows, it corresponds to a monophonic score. Otherwise it corresponds to a polyphonic one, but please note that we don't impose any restriction. In this paper, we call *alignment* the task of classifying timestamps of a series of features $X$ on its template $\Phi$. The goal is to find an *alignment mapping* (or *path*) $m$ from the timestamps $\{1,\dots,T\}$ to its list of events $\{1,\dots,E\}$.

**Parametrization of alignments.** In audio-to-score alignment, we assume that all events occur in order and no event is skipped; so the path constraints are as follows: $m$ is a *non-decreasing* mapping from $\{1,\dots,T\}$ to $\{1,\dots E\}$ such that (i) $m(1)=1$, (ii) $m(T)=E$, (iii) $m(t+1)=m(t)+1$ or $m(t+1)=m(t)$. An alignment mapping $m$ can be represented through an *alignment matrix* $Y$ of dimension $T\times E$, such that $Y_{t,e}$ is equal to 1 if $m(t)=e$ and 0 otherwise. The set of all these alignment matrices between an input of length $T$ and a template of length $E$ is denoted by $\mathcal{Y}(T,E)$, or simply $\mathcal{Y}$ in the sequel.

**DTW algorithm as a LP solver.** An alignment procedure usually starts by building a local cost matrix $A\in\mathbb{R}^{T\times E}$ whose elements $A_{t,e}$ measures the dissimilarity between each pair of features $X_{t,.}$ – the $t$-th row of $X$ – and event $e$. Then, the cost of an alignment $m$ is defined as the sum of local costs along the path: $\sum_{t=1}^{T} A_{t,m(t)} = \mathrm{Tr}(Y^\top A)$. Given any affinity matrix $A$, the dynamic time warping (DTW) algorithm [21] uses dynamic programming to find the minimum cost path $\mathrm{argmin}_{Y\in\mathcal{Y}}\mathrm{Tr}(Y^\top A)$ in $O(TE)$ operations. Thus, DTW is an efficient linear program (LP) solver over $\mathcal{Y}$.

**DTW cost function with Euclidian local distance.** A common choice [11] of local distance $A_{t,e}$ is the squared Euclidian norm between some transform of the input features $\Psi(X_{t,.})\in\mathbb{R}^K$ and some template features $\tau_e\in\mathbb{R}^K$ that represent the event $e$: $A_{t,e}=\|\tau_e-\Psi(X_{t,.})\|_2^2$. With our notations, $\tau_e$ is the $e$-th column of $\Phi$, so an alignment cost equals:

$$\sum_{t=1}^{T} A_{t,m(t)} = \sum_{t=1}^{T} \|\mathbf{e}_{m(t)}^\top\Phi - \Psi(X_{t,.})\|_2^2,$$

where $\mathbf{e}_k$ denotes the $k$-th standard basis vector of $\mathbb{R}^E$. Let $\|.\|_F$ denote the Frobenius norm. If we define $\psi : \mathbb{R}^{T\times p}\mapsto\mathbb{R}^{T\times K}$ such that $\psi(X)$ is the concatenation of vectors $\Psi(X_{t,.})$, the DTW alignment cost reads:

$$\min_{Y\in\mathcal{Y}} \|Y\Phi-\psi(X)\|_F^2. \tag{1}$$

### 2.2. Weakly-supervised discriminative learning

Before any alignment is performed, templates $\Phi$ are usually designed with prior knowledge such as a synthesized signal [4], or learned with supervision and annotated data [13]. We rather want to perform alignment while optimizing our cost function (1) *without supervision*. To do so, we follow the DIFFRAC framework [15] and learn an optimal linear transform[1] of the input features $\psi(X)=WX$, where $W\in\mathbb{R}^{K\times p}$, while keeping the templates $\Phi$ fixed. The criterion we choose on $W$ is to minimize the DTW cost function (1) plus a Tikhonov regularization with some $\lambda\geq 0$. So the joint estimation of $Y$ and $W$ reads:

$$\min_{W\in\mathbb{R}^{K\times p}} \min_{Y\in\mathcal{Y}} \|Y\Phi - XW\|_F^2 + \frac{\lambda}{2}\|W\|_F^2. \tag{2}$$

This problem leads to tractable convex relaxation thanks to the *joint convexity* in $W$ and $Y$ of the objective function. Beforehand, the unconstrained optimization in $W$ is solved using first order condition. Following [15, 17], this yields the explicit expression: $W=(X^\top X+T\lambda\mathrm{Id}_T)^{-1}X^\top Y\Phi$. Plugging it back in Eq. (2) provides the following minimization problem: $\min_{Y\in\mathcal{Y}}\mathrm{Tr}(\Phi^\top Y^\top BY\Phi)$ where $B=\mathrm{Id}_T - X(X^\top X+T\lambda\mathrm{Id}_p)^{-1}X^\top$. This objective function is still convex in $Y$ but the set $\mathcal{Y}$ is discrete. To make the obtained problem convex, we relax $\mathcal{Y}$ into its convex hull $\overline{\mathcal{Y}}$, and get a quadratic program (QP):

$$\min_{Y\in\overline{\mathcal{Y}}} \mathrm{Tr}(\Phi^\top Y^\top BY\Phi). \tag{3}$$

This relaxation is attracted to two kind of degenerate solutions: the constant solution, which is a minimizer of any convex relaxation invariant by column permutation [15, 14]; solutions $Y$ that assign all timestamps to the same class, as noted by [16]. In our case, the constraints on $\mathcal{Y}$ linked to the sequential structure get rid of some degenerate solutions. However, as shown in the experimental section, as the number of events $E$ grows, the supervision gets weaker and Eq. (3) gets plagued by solutions that are almost equal to the trivial ones. To overcome this drawback, one needs to get rid of the symmetries of our objective function. We propose to do so by plugging the prior knowledge given by the score into the cost function.

## 3. USING ADDITIONAL PRIOR KNOWLEDGE

**Expected alignment.** A music score induces a prior about relative duration of each event $e$. Such information can be encoded through an *expected alignment* $\bar{Y}\in\mathcal{Y}$, that would be obtained if the actual duration of every event was equal to the duration in the score.

**Global prior.** We penalize the distance between a candidate alignment $Y$ and the expected one $\overline{Y}$, using the squared sum of absolute differences between the start times (*onsets*) of corresponding events. This distance turns out to be used as an evaluation metric of music-to-score alignment [7]. One can show it equals $\|YL-\overline{Y}L\|_F^2$, with $L$ the strictly lower triangular matrix of size $E\times E$ with ones, it is a version of the area

[1] The extension to affine transforms $\psi(X)=WX+\mathbf{1}\mathbf{b}^\top$ where $\mathbf{b}\in\mathbb{R}^K$, is straightforward and has been used in experiments.

loss introduced by [22], which turns out to be exactly the area between the two warpings seen as binary matrices. We call this term the *global prior*, as its promotes alignments where the local distortions compensate themselves and the actual interpretation has globally the same shape as the score (*rubato* in musical terminology).

**Local prior.** Another idea is to penalize individually the discrepancy between the actual duration and the expected duration of each event: $\sum_{e=1}^{E}(\mathbf{1}_T^\top Y_{\cdot,e} - \mathbf{1}_T^\top \overline{Y}_{\cdot,e})^2 = \|\mathbf{1}_T^\top Y - \mathbf{1}_T^\top \overline{Y}\|_2^2$ where $\mathbf{1}_T \in \mathbb{R}^T$ is the vector with ones. This loss could be interpreted as a Gaussian prior on individual duration. This local penalization promotes alignments where the relative durations are correct for almost all events, except for a few ones. In musicology, such events are called *fermata* [23], where the interpret can unpredictably wait a very long time.

**Complete cost function.** Adding these two priors to the cost function of Eq. (3) yields the following relaxed problem on $\overline{\mathcal{Y}}$ ($\mu, \nu > 0$ are arbitrary):

$$\min_{Y \in \mathcal{Y}} \operatorname{Tr}(\Phi Y^\top B Y \Phi) + \frac{\lambda}{2}\|W\|_2^2 + \mu\|(Y - \overline{Y})L\|_F^2 + \nu\|\mathbf{1}_T^\top (Y - \overline{Y})\|_2^2. \tag{4}$$

Note that the priors do not increase the complexity as Eq. (4) is still a QP. Such a problem cannot be solved in closed form. But the DTW algorithm provides an efficient LP solver $\max_{Y \in \mathcal{Y}} \operatorname{Tr}(AY)$ on the set $\mathcal{Y}$, hence on its hull $\overline{\mathcal{Y}}$. Consequently, the QP can be efficiently solved with the Frank-Wolfe algorithm [24, 25], a.k.a. conditional gradient descent – refer to [16, Algorithm 1] for all implementation details. As the problem is relaxed into $\overline{\mathcal{Y}}$, its solution $Y^*$ might not be a valid alignment in $\mathcal{Y}$. From $Y^*$, we can always deduce the optimal classifiers $W^*$ (see equation above). But rounding the solution $Y^*$ is needed to get a valid alignment $Y$.

**Rounding with DTW.** The first way to round, is to perform a DTW alignment with the optimal classifiers $W^*$. This consists in solving the DTW problem of Eq. (1), which boils down to the following LP: $\max_{Y \in \mathcal{Y}} \operatorname{Tr}(\Phi^\top Y^\top X W^*)$.

**Rounding in $\Phi Y$.** Following [16], a natural idea is to round $\Phi Y^*$ to the closest assignment matrix $\Phi Y$, in the sense of the Euclidean norm. It amounts to solve: $\min_{Y \in \mathcal{Y}} \|Y^*\Phi - Y\Phi\|_F^2$. Expanding the squared norm proves it also boils down to an LP, in both monophonic and polyphonic settings.

## 4. EXPERIMENTS

**Dataset and features.** Our experiments are run on the Finnish folk song dataset [26]. It is made of $\sim$ 48 hours of music available in MIDI format (8614 songs). Our $K$ classes are the 44 notes appearing in the dataset, plus an additional "silence" class. Audio files are synthesized from the available MIDI files, after having randomly modified the local tempo, using the MIR toolbox [27]. That way, we know the exact groundtruth alignment. A MIDI file encodes the score and provides the expected alignment $\overline{Y}$, which is different from the groundtruth $Y_{gt}$. We consider four different setups: **(1)** Non-stretched data: $\overline{Y} = Y_{gt}$. **(2)** Rubato: tempo is alternatively sped up and slowed down. **(3)** Fermata: most notes are played with the original duration except for a few whose duration is increased. **(4)** A combination of (2) and (3). We compute a 1200-dimensional spectrogram of the audio signal using half-overlapping windows of length 160ms. Then, we bin it in 40 dimensions using the mel-scale filterbank, as implemented in the MIR toolbox. Note that our method similar results on the full spectrogram. Songs are split into a train, validation and test sets. We use between 100 and 300 songs for training and between 200 and 500 songs for both validation and testing.

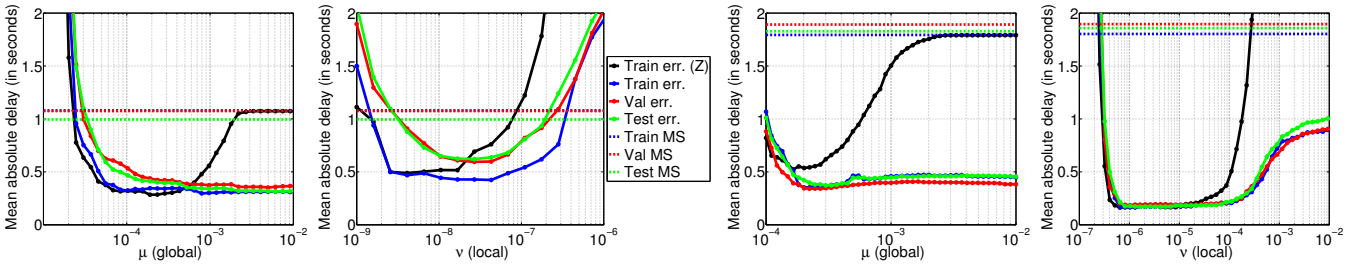
| $E$ | 10 | 15 | 20 | 30 | 50 |
|---|---|---|---|---|---|
| mean length | 64 | 93 | 124 | 174 | 224 |
| delay (s) | 0.42 | 0.59 | 0.68 | 0.93 | 1.09 |

**Table 1**: Average onset delay for different song lengths $k$, without priors ($\mu = \nu = 0$).

**Performance measure.** The quality of a audio-to-score alignment is usually quantified using the mean delay between onsets [7], used in Sec. 3: $\frac{1}{E}\|YL - Y_{\mathrm{gt}}L\|_F$, where $E$ is the number of events.

**Need for priors.** We first run experiments with no duration prior and no tempo stretching like in [16], by setting $\mu = \nu = 0$. We use 300 full songs that we split into shorter songs of a fixed length $k$. Alignment results for various values of $k$ are presented in Table 1. Performances clearly deteriorate as the length of series increases. Indeed, the set $\mathcal{Y}$ has more and more symmetries: the objective is attracted by degenerate solutions [16]. This situation calls for our additional priors so as to work on typical real-world scores.

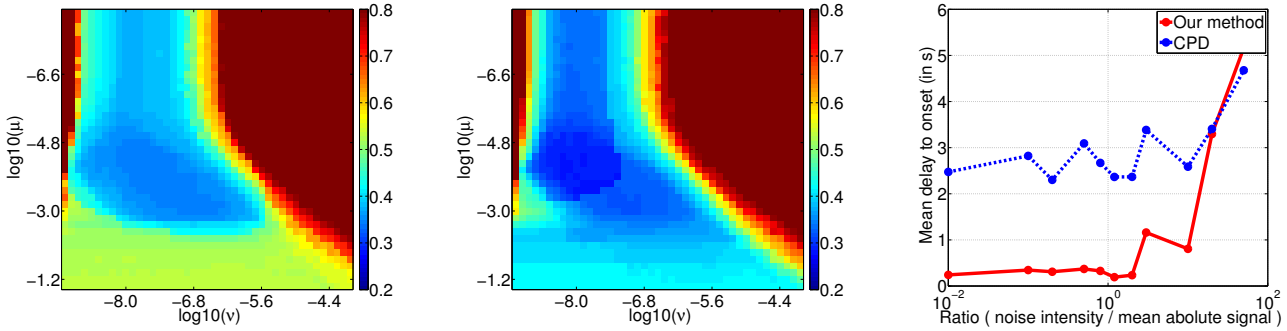
**Rubato performance.** We compare the separate effect of the global and the local prior in Eq. (4) in the Rubato setting. We consider 100 songs for training, 200 for testing and 200 for validation. Dashed curves (train MS, val MS and test MS) depict the onset loss $\|Y_{gt} - \overline{Y}\|$. In Fig. 1a (left), the rounding in $Z$ has the same performance as the dashed curve for large enough $\mu$. In this case, Eq. (4) consists in minimizing a well-conditioned quadratic form; so $Y^*$ is already in $\mathcal{Y}$ and the rounding procedure has little effect. On the contrary, Fig. 1a (right) shows that for large $\nu$, our method gets above the dashed baseline; indeed, the quadratic form it minimizes is ill-conditioned (low rank). In Fig. 1a (left), for large $\mu$, the optimal $Y^*$ is almost equal to the score $\overline{Y}$, as explained above. In that case, the DTW rounding procedure solves the alignment problem with classifiers $W$ learned on $\overline{Y}$. In this rubato setting, this method works better. We recall that rubato




(a) Effects of local and global prior in the rubato setting.

(b) Effects of local and global priors in the fermata setting.

**Fig. 1**: Evaluation of separate priors in the Rubato and Fermata settings.




(a) Validation error

(b) Test error

(c) Noise robustness

**Fig. 2**: Combined rubato and fermata setting. **(a-b)** Perf. as function of $\mu$ and $\nu$ on val. and test sets. **(c)** Robustness to noise and comparison to change-point detection.

means tempo fluctuations around its average. As a result, the average delay between the score and the input is low. Therefore, learning a model on the slightly faulty alignment $\overline{Y}$ provides a good classifier. On the contrary, the local prior is not very robust against rubato – as expected.

**Fermata performance.** In the fermata setting, the performances of our approach are depicted in Fig. 1b. It features the same baseline as Fig. 1a. We have also run experiments with respectively $\mu = 0$ and $\nu = 0$ but these are off the charts. For the $\nu$ parameter, a clear trade-off appears, and when properly adjusted our method outperforms other baseline. Contrary to the rubato experiment, the global prior does not help. The best performance is for large $\mu$, for which the predicted alignment $Y$ sticks to the expected one $\overline{Y}$.

**Combined performance.** In the mixed setting depicted in Fig. 2(a-b), we observe a trade-off between the local and global priors. When properly adjusted, our method with the joint priors outperforms all other baselines.

**Robustness to noise.** We consider data that have been stretched both with Fermata and Rubato, and assess the robustness of our approach against a white noise, up to a very intense level. We add to the synthesized signals white noise whose intensity is controlled through the ratio between the standard deviation of the noise $\sigma$ and the mean absolute value of the signal. We use our approach on 300 training songs, validate $\mu$ and $\nu$ on 250 and test on 250 others. Figure 2(c) compares the performance of our method with a change-point detection baseline (CPD). This basic algorithm detects changes in the mean of a homoscedastic Gaussian process – refer to [28] for details; it knows the number $E$ of events but is unaware of the redundancy of notes. Removing the class information in our algorithm makes it equivalent to this CPD. So this baseline shows our algorithm do benefits from class redundancy.

## 5. CONCLUSION

This paper describes a discriminative and weakly-supervised approach for audio-to-score alignment. Our method relies on the estimation of individual classifiers for each of the possible notes, and corresponds to the optimization of the DTW cost function. This step is achieved by the minimization of a convex and quadratic objective function that can be solved efficiently using a conditional gradient algorithm. The experiments run in the mono-instrument monophononic setting are very promising and show the robustness of the method to tempo deformation as well as white noise. Our method can be used in the polyphonic setting with no modifications. It could also be extended to the polyphonic polyinstrumental setting by considering separate classifiers for each instrument.

## 6. REFERENCES


[1] A. Ghias, J. Logan, D. Chamberlin, and B. Smith, "Query by humming: musical information retrieval in an audio database," in *Proc. ICM*, 1995.

[2] R. Dannenberg, "An intelligent multi-track audio editor," in *Proc. ICMC*, 2007.

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

[4] A. Arzt, G. Widmer, and S. Dixon, "Automatic page turning for musicians via real-time machine listening," in *Proc. ECAI*, 2008.

[5] Christopher Raphael, "A Bayesian network for real-time musical accompaniment," in *Adv. NIPS*, 2001.

[6] N. Orio, S. Lemouton, and D. Schwarz, "Score following: State of the art and new developments," in *Proc. NIME*, 2003.

[7] A. Cont, D. Schwarz, N. Schnell, and C. Raphael, "Evaluation of real-time audio-to-score alignment," in *Proc. ISMIR*, 2007.

[8] Meinard Müller, Frank Kurth, and Michael Clausen, "Audio matching via chroma-based statistical features.," in *ISMIR*, 2005, vol. 2005, p. 6th.

[9] S. Dixon and G. Widmer, "MATCH: A music alignment tool chest.," in *Proc. ISMIR*, 2005.

[10] I. Özgür and R. Dannenberg, "Understanding features and distance functions for music sequence alignment.," in *Proc. ISMIR*, 2010.

[11] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proc. ICASSP*, 2009, pp. 1869–1872.

[12] C. Joder, S. Essid, and G. Richard, "Learning optimal features for polyphonic audio-to-score alignment," *IEEE Trans. on ASLP*, vol. 21, no. 10, pp. 2118–2128, 2013.

[13] J. Keshet, S. Shalev-Shwartz, Y. Singer, and D. Chazan, "A large margin algorithm for speech-to-phoneme and music-to-score alignment," *IEEE Trans. on ASLP*, vol. 15, no. 8, pp. 2373–2382, 2007.

[14] Y. Guo and D. Schuurmans, "Convex relaxations of latent variable training," in *Adv. NIPS*, 2007.

[15] F. Bach and Z. Harchaoui, "DIFFRAC: a discriminative and flexible framework for clustering," in *Adv. NIPS*, 2008.

[16] P. Bojanowski, R. Lajugie, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic, "Weakly supervised action labeling in videos under ordering constraints," in *Proc. ECCV*, 2014.

[17] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. CVPR*, 2010.

[18] A. Joulin, K. Tang, and L. Fei-Fei, "Efficient image and video co-localization with Frank-Wolfe algorithm," in *Proc. ECCV*, 2014.

[19] E. Grave, "A convex relaxation for weakly supervised relation extraction," in *Proc. EMNLP*, 2014.

[20] P. Bojanowski, R. Lajugie, E. Grave, F. Bach, I. Laptev, J. Ponce, and C. Schmid, "Weakly-supervised alignment of video with text," *arXiv preprint arXiv:1505.06027*, 2015.

[21] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on ASLP*, vol. 26, no. 1, pp. 43–49, 1978.

[22] R. Lajugie, D. Garreau, S. Arlot, and F. Bach, "Metric learning for temporal sequence alignment," in *Adv. NIPS*, 2014.

[23] M. Kennedy, *The Oxford dictionary of music*, Oxford University Press, 1994.

[24] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval research logistics quarterly*, vol. 3, no. 1-2, pp. 95–110, 1956.

[25] M. Jaggi, "Revisiting Frank-Wolfe: Projection-free sparse convex optimization," in *Proc. ICML*, 2013.

[26] T. Eerola and P. Toiviainen, "Finnish folk song database," *http://esavelmat.jyu.fi/*, 2004.

[27] O. Lartillot and P. Toiviainen, "A Matlab toolbox for musical feature extraction from audio," in *Proc. DAFx*, 2007.

[28] G. Rigaill, "A pruned dynamic programming algorithm to recover the best segmentations with $1$ to $k_{max}$ change-points," *ArXiv e-prints*, Apr. 2010.