ALIGNMENT KERNELS FOR AUDIO CLASSIFICATION WITH APPLICATION TO MUSIC INSTRUMENT RECOGNITION
Cyril Joder, Slim Essid, Gaël Richard

To cite this version:
Cyril Joder, Slim Essid, Gaël Richard. ALIGNMENT KERNELS FOR AUDIO CLASSIFICATION WITH APPLICATION TO MUSIC INSTRUMENT RECOGNITION. 16th European Signal Processing Conference, Aug 2008, Lausanne, Switzerland. hal-02943674

HAL Id: hal-02943674
https://telecom-paris.hal.science/hal-02943674
Submitted on 20 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ALIGNMENT KERNELS FOR AUDIO CLASSIFICATION WITH APPLICATION TO MUSIC INSTRUMENT RECOGNITION

Cyril Joder, Slim Essid, and Gaël Richard

TELECOM ParisTech/Institut TELECOM, LTCI/CNRS
46, rue Barrault, 75013 Paris, France
{joder, essid, grichard}@telecom-paristech.fr

ABSTRACT
In this paper we study the efficiency of support vector machines (SVM) with alignment kernels in audio classification. The classification task chosen is music instrument recognition. The alignment kernels have the advantage of handling sequential data, without assuming a model for the probability density of the features as in the case of Gaussian Mixture Model-based Hidden Markov Models (HMM). These classifiers are compared to several reference systems, namely Gaussian Mixture Model, HMM classifiers and SVMs with “static” kernels. Using a higher-level representation of the feature sequence, which we call summary sequence, we show that the use of alignment kernels can significantly improve the classification scores in comparison to the reference systems.

1. INTRODUCTION
Owing to the large volume of audio data now available to the general public, there has been a growing interest in the research community for automatic tools to index and describe the content of audio recordings. Many audio indexing tasks can be handled with a common classification approach. First, an intermediate description of the signal is obtained thanks to features, which capture specific properties of the given signal. These features are generally extracted over short temporal analysis windows, over which the signal can be considered as stationary. Hence, the signal is represented by the series of these features. Then, a statistical classifier is used to determine the most probable class for the observed features. In a supervised strategy, the classifier is trained using a labelled database containing various examples of each class.

In order to classify sequences of features, Hidden Markov Models (HMM) have been extensively used, especially in speech recognition systems, and also to a smaller extent in other domains such as music instruments recognition [4]. One of their main advantages is their ability to model the temporal evolution of the features. However, in many audio classification systems, the features temporal properties are often not taken into account. Indeed, these systems generally lie on the assumption that the observations of the features in different frames are statistically independent. In other words, they suppose that the evolution of these parameters over time is not informative about the class membership of a given sound. Thus, a decision is made for each frame independently of the others. See for example [8].

Some new ways to take into account the temporal evolution of the features have been proposed for audio classification, including the use of a filterbank to summarize the periodogram of each feature [9] or the exploitation of autoregressive models to approximate the feature dynamics [10].

In this work, we explore the use of recently proposed alignment kernels [1, 16, 2] with support vector machine (SVM) classifiers and compare their performance to several state-of-the-art classifiers, namely GMM, HMM classifiers and SVM with Gaussian kernel, on a musical instrument recognition task. Experiments are run to classify two kinds of feature sequences exploiting a segmentation into so-called sonic units. We show that the combined use of alignment kernels and a higher-level feature sequence representation outperform the reference systems.

In the following section, we briefly recall the SVM classifier principle and define the alignment kernels used. The reference systems and the results of the experiments on a music instrument recognition task are presented in Section 3. Some conclusion are finally suggested in Section 4.

2. CLASSIFICATION WITH SUPPORT VECTOR MACHINES

2.1 SVM Classifiers
Support vector machines (SVM) are powerful classifiers that have proven to be efficient for various classification tasks, such as face recognition, speaker identification and also instrument recognition [5]. These classifiers are known for their good generalization property, even in high dimension. SVMs also have the advantage of being discriminative, as opposed to generative approaches, i.e. they do not assume any particular form of the data probability density.

In bi-class problems, SVMs aim to find the hyperplane that separate the feature vectors of the two classes with the maximum margin. Formally, the algorithm searches for the hyperplane w · x + b = 0 that separates the training samples x1, . . . , xn which are assigned labels y1, . . . , yn, with yi ∈ {−1, 1}, so that

\[ \forall i, \; y_i (x_i \cdot w + b) \geq 1 \]  

under the constraint that the margin \( \frac{2}{|w|} \) be maximal. Feature vectors for which the equality in (1) holds are called support vectors. A vector x is then classified with respect to the sign of the function:

\[ f(x) = x \cdot w + b = \sum_{i=1}^{n_s} \alpha_i y_i s_i \cdot x + b \]

where \( s_i \) are the support vectors, \( \alpha_i \) are Lagrange multipliers and \( n_s \) is the number of support vectors. For the case of non-separable classes, slack variables are introduced as a mechanism for tolerating outliers, controlled with a penalty coefficient denoted by C. See [15] for more details.

In order to perform multi-class classification, we adopt a “one versus one” strategy and use Platt’s approach [13] which derives posterior class probabilities after the two-class SVMs.
2.2 Kernels

In order to enable non linear decision surfaces, it is possible to map the input vectors to a higher dimension space where the two classes can become linearly separable. The dot product in this vector space is given by a function called a kernel. Let \( k \) be this kernel function, a vector \( x \) is classified according to the sign of:

\[
f(x) = \sum_{i=1}^{n_x} \alpha_i y_i k(s_i, x) + b.
\]

Note that neither the corresponding space nor the mapping function are necessary to express the classification function. The knowledge of the kernel alone is sufficient and this function can be seen as a similarity measure between vectors.

Given such a function and a family of vectors \( x_1, \ldots, x_m \), the Gram Matrix of \( k \) with respect to \( x_1, \ldots, x_m \) is the \( m \times m \) matrix \( G \) defined by \( G_{i,j} = k(x_i, x_j) \). A sufficient condition for \( k \) to be a proper kernel function, i.e. to represent the dot product between vectors “mapped” in a Hilbert space, is that for all \( m \in \mathbb{N} \) and all vectors \( x_1, \ldots, x_m \), the corresponding Gram matrix be positive definite. Such a kernel is then called a positive definite kernel. In this case, the Hilbert space into which the feature vectors are mapped can be explicitly constructed (it is then called Reproducing Kernel Hilbert Space). Note that kernels which are not positive definite may obtain in good classification results in practice, although there is no theoretical proof that their use is well justified.

As our reference kernel, we choose the Gaussian radial basis function (RBF) kernel, denoted by \( k_0 \), because it achieves the best classification performance for our instrument recognition problem, among the kernels tested in [5]. The following form is used

\[
k_0(x, y) = \exp \left( -\frac{||x - y||^2}{d \sigma^2} \right)
\]

where \( d \) is the dimension of the vectors and \( \sigma^2 \) is a parameter of the kernel.

We also use another kernel defined by

\[
\xi(x, y) = \frac{1}{2} k_0(x, y) \frac{1 - \frac{1}{2} k_0(x, y)}{1 - \frac{1}{2} k_0(x, y)}.
\]

This kernel, which is numerically similar to the Gaussian kernel, has been introduced so that \( \xi \) be positive definite, as will be explained in Section 2.3. It can be proven that this kernel satisfies Mercer’s condition using the form: \( \xi(x, y) = \sum_{j=1}^{\infty} 2^{-j} k_0(x, y)^j \).

2.3 Alignment Kernels

Let \( \mathbf{x} = (x_1, \ldots, x_n) \) be a finite series of feature vectors which we want to classify. In a “static” strategy, each vector \( x_i \) is compared (thanks to the kernel function) to every support vector, and then classified according to \( f(x_i) \) as in (2.2). However, it is not always relevant to classify a vector independently of all the others. As the temporal structure of music is important, it may be more meaningful to compare whole sequences of vectors.

Alignment kernels allow for the comparison of trajectories of feature vectors, instead of operating on single observations. Thus, a sequence \( \mathbf{x} \) can be classified “as a whole”, according to a decision function:

\[
f(\mathbf{x}) = \sum_{i=1}^{n_x} \alpha_i y_i k(s_i, \mathbf{x}) + b
\]

where \( s_i \) are sequences instead of isolated support vectors.

In order to cope with the problems of feature sequences synchronization, the comparison is made after a temporal alignment of the sequences, which may be of different lengths. We briefly describe the alignment algorithm before presenting the alignment kernels used.

Let \( y = (y_1, \ldots, y_m) \) be another finite feature series. An alignment path \( \pi \) of length \( p \) between \( \mathbf{x} \) and \( y \) is a function from \( \{1, \ldots, p\} \) to \( \{1, \ldots, n\} \times \{1, \ldots, m\} \) such that, with the notation \( \pi(i) = (\pi_1(i), \pi_2(i)) \), where \( i \in \{1, \ldots, p\} \):
1. the functions \( i \mapsto \pi_1(i) \) and \( i \mapsto \pi_2(i) \) are increasing,
2. the function \( \pi \) is injective.

The set of all possible alignment paths between \( \mathbf{x} \) and \( y \) is denoted by \( \mathcal{A} \). The DTW distance \( D(\mathbf{x}, y) \) is a distance measure between the two aligned sequences along the optimal path according to the following criterion:

\[
D(\mathbf{x}, y)^2 = \min_{\pi \in \mathcal{A}} \frac{1}{M_x} \sum_{i=1}^{p} m_x(i)||x_{\pi_1(i)} - y_{\pi_2(i)}||^2
\]

where \( m_x(i) \) are non-negative weighting coefficients and \( M_x = \sum_{i=1}^{p} m_x(i) \) is the normalization factor. The value of the weighting coefficients is a function of the increment \( \pi(i) - \pi(i-1) \). This function influences the optimality criterion, hence favoring or penalizing certain kinds of paths.

The Gaussian Dynamic Time-Warping kernel (GDTW) introduced in [1] uses this alignment between two sequences. The idea is to exploit the DTW distance instead of the Euclidian distance in the calculation of a Gaussian kernel. Thus, the resulting value of this kernel is the geometric mean of the Gaussian kernel values along the optimal alignment path. The GDTW kernel is then defined as:

\[
K_{\text{GDTW}}(\mathbf{x}, y) = \exp \left( -\frac{1}{d \sigma^2} D(\mathbf{x}, y)^2 \right) = \prod_{i=1}^{p} k_0(x_{\pi_1(i)}, y_{\pi_2(i)}) \frac{m_x(i)}{M_x}.
\]

The Dynamic Time-Alignment Kernel (DTAK) proposed in [16] calculates another similarity measure between two sequences by considering the arithmetic mean of the kernel values along the alignment path. Applying this idea to the Gaussian kernel, the DTAK kernel is then defined as:

\[
K_{\text{DTAK}}(\mathbf{x}, y) = \max_{\pi \in \mathcal{A}} \frac{1}{M_x} \sum_{i=1}^{p} m_x(i) k_0(x_{\pi_1(i)}, y_{\pi_2(i)}).
\]

Note that the optimal alignment paths considered by these two kernels may be different. Indeed, the local similarity used for the GDTW kernel is the Euclidian distance, whereas the one used in (3) is the Gaussian kernel value.

Cuturi et al [2] emphasize the fact that these alignment kernels have not been proven positive definite. In the same work, they introduce another alignment kernel type which is positive definite under a certain assumption. It is similar to the GDTW kernel but considers the values obtained with all the possible alignments. Given a “static” kernel \( \kappa \), a corresponding alignment kernel \( K_{\kappa} \) can be defined as:

\[
K_{\kappa}(\mathbf{x}, y) = \sum_{\pi \in \mathcal{A}} \prod_{i=1}^{p} \kappa(x_{\pi_1(i)}, y_{\pi_2(i)})
\]
Here all the weighting coefficients are equal to 1 and there is no normalization. The authors prove that if \( \kappa \) is such that \( \frac{1}{1 + \kappa} \) is positive definite, then \( K_\kappa \) is also positive definite.

The similarity measure induced by this kernel is different from the previous alignment kernels. Indeed, the sum in (4) takes advantage of every possible alignments instead of only the optimal one. Thus, two sequences are similar in the sense of \( K_\kappa \), not only if they have an alignment which results in a small DTW distance, but also share numerous suitable alignments.

We consider two instances of (4). The first alignment kernel is the application of this framework with the “static” kernel \( \xi \frac{k_0(\mathbf{x}_1, \mathbf{y}_1)}{1 - \frac{1}{2}k_0(\mathbf{x}_1, \mathbf{y}_1)} \). Thus, the alignment kernel obtained is positive definite. Its formulation is

\[
K_\xi(\mathbf{x}, \mathbf{y}) = \sum_{\sigma \in \mathcal{A}} \prod_{i=1}^{p} \frac{1}{1 - \frac{1}{2}k_0(\mathbf{x}_{\sigma_1(i)}, \mathbf{y}_{\sigma_2(i)})}.
\]

The second one uses the kernel \( \chi = e^{k_0} \). However, Cuturi et al. found that the Gram matrices obtained with this alignment kernel were exceedingly diagonally dominant, that is the diagonal values of these matrices are many orders of magnitude larger than the other kernel values. Thus, the different vectors are almost orthogonal in the reproducing space and it has been observed in practice that the SVMs do not perform well in such situations [17]. The authors suggest using the logarithm of these values, arguing that although it does not conserve positive definiteness, it achieves good classification performances. This kernel is thus defined as

\[
K_\chi(\mathbf{x}, \mathbf{y}) = \log \left( \sum_{\sigma \in \mathcal{A}} \prod_{i=1}^{p} e^{k_0(\mathbf{x}_{\sigma_1(i)}, \mathbf{y}_{\sigma_2(i)})} \right).
\]

As pointed out by the authors, this kernel calculates in fact the soft-max of scores of all possible alignments, rather than the simple maximum as for the first two alignment kernels.

### 3.2 Experimental Setup

#### 3.2.1 Features Extraction

Our tests have been performed using Essid’s database [5]. This database is composed of solo musical phrases of several instruments, mainly drawn from commercial Compact Disc recordings, with the aim to assess the classification system’s generalization ability. Thus, it allows for the classification of “real world performances”, as opposed to isolated notes which are the object of most of the publicly available databases. The few solos which are present in the RWG base have also been used. In this work, eight instruments were considered, corresponding to about 3h35’ of audio data. The database was split in two approximately equal sets: a training set and a test set. It was made sure that sounds of different sets are extracted from different sources.

The sound files were downsampled to a 32-kHz sampling rate and normalized.

A set of 40 acoustic features of various types are used, including cepstral coefficients, zero-crossing rates or wavelet transform coefficients. They are obtained by automatic feature selection based on Linear Discriminant Analysis [3] from a wide set of common audio features [12]. These features are computed over 32-ms analysis frames, with 16-ms overlap.

#### 3.2.2 Segmentation

One of the advantages of the classifiers using alignment kernels is their ability to compare feature trajectories instead of isolated vectors. In order to exploit this ability, the choice of the segments to be compared has to be addressed. We believe that the feature trajectories are “meaningful” and distinctive of an instrument over the musical notes. However, as the segmentation of an audio file into musical notes is a very complex task, we choose to perform an automatic segmentation into so-called sonic units. The sonic units are time segments which are supposed to be very close to musical notes.

We proceed as follows: First, silence frames are detected as in [5] and are removed from the data. Then an onset...
detector is run, using an algorithm proposed by Leveau and
Daudet [7]. This algorithm detects transients, which are
supposed to be note attacks. Finally, the sonic units are
defined as the intervals between two successive onsets, as
represented in Figure 2. In order to cope with the onset
detector errors, the length of the sonic units is forced to
be between 5 frames (0.1 s) and 125 frames (2 s), thus forbidding
too short or too long segments.

The inputs of the classification systems are then the fea-
ture vector sequences corresponding to these sonic units, as
represented in Figure 1. Therefore, a classification decision
is taken for each sonic unit.

3.2.3 Classification
The SVM classification software used for this study is based
on the “SVM-Light” implementation by Joachims [6], which
was extended in order to include the alignment kernels. The
parameter C is set to 1, based on the results of previous work
[5]. We test 9 values of the parameters $\sigma^2$ for each kernel,
between 0.25 and 32.

For the GMMs and HMMs, we use Murphy’s MATLAB tool-
box [11]. We model the notes by a left-right 3-state HMMs,
with an 8-component GMM in each state.

The training of the HMMs is performed with the EM
algorithm over the separated sonic units. The states are
supposed to capture respectively the attack, sustain and re-
lease of each sonic unit. In order to have the same number of
Gaussian components as for the HMM classifier, we tested a
24-component GMM system, but its performance was lower
than a 8-component GMM. Therefore, we only present the
results of the latter system.

The alignment kernels presented in Section 2.3 use a form
of the DTW algorithm. The chosen weighting coefficients
$m_k(t)$ are equal to 1 for a “horizontal” or “vertical” step and
2 for a unitary “diagonal” step, so that the normalization
coefficient $M_k$ is independent of the alignment path.

Since in our database, the amount of data differs from
one instrument to another, the score which we use to com-
pare the classification systems is the average recognition rate,
i.e. the average of all the classes recognition rates. With the
figures (all expressed in percents), we also specify the radius
of the largest 95% confidence interval (corresponding to the
worst case), which will be referred to as confidence.

3.3 First Results
The results of our first experiments are given in Table 1. In
these tests, we classify the feature vectors sequences corre-
sponding to the sonic units, with 6 systems. The calculation
of the SVM solutions using the alignment kernels $K_\xi$ and $K_\chi$
is not presented since the executions did not terminate
in acceptable running times. An explanation of this phe-
nomenon may be the absence of normalization in the com-
putation of these kernels (see Eq. 4). Consequently, the
kernel values depend on the sequences length and the ob-
tained gram matrix may be ill-conditioned, resulting in a
bad convergence of the optimization algorithm.

We observe that the HMM system shows significantly
better results than the GMM classifiers (+1.6% improve-
ment whereas the 95% confidence interval radius is 1%), which
suggests that taking into account the temporal dependency
of the feature vectors does improve the classification perfor-
mance. However, the “static” SVM systems perform better
than both GMM and HMM classifiers. We believe that the
reason lies in the fact that the latter classifiers are generative,
assuming mixture of Gaussian densities which may not be
appropriate for this problem, whereas SVMs are discrimina-
tive classifiers that model the classes “boundaries” without
supposing any special form of the probability densities.

However, we see that the SVM classifiers using the align-
ment kernel GDTW and DTAK turn out to be ineffective.
Indeed, they achieve the lowest classification scores among
the systems we test. As these kernels appear not to be ap-
propriate for our problem when used on feature vectors se-
quencies over the entire sonic units, we adopt another stra-
tegy where the classifiers are run on so-called summary se-
quencies.

3.4 Summary Sequences Classification
In this new approach, whose principle is represented in Fig-
ure 2, the sonic units are split in a small number of subseg-
ments of the same length. Then, the mean of the feature
vectors is computed over each of the subsegments. Finally,
the summary sequences (the sequences of these means), are
classified so that a decision is made over every sonic unit.
We run two sets of experiments, using 3 and 5 subsequences
for each sonic unit. We choose these small numbers because
the sonic units can be as short as 5-frames length.

Since the number of subsegments is constant, the length
of these subsequences depends on the sonic unit total length.
This choice may seem odd, as the summary sequence con-
struction is in fact a kind of linear time-warping transfor-
mation, which may interfere with the alignment algorithm
embedded in the kernels. However, it is a way to overcome
the normalization problem explained in Section 3.3 and thus
use the kernels $K_\xi$ and $K_\chi$.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Segmentation</th>
<th>Feature Vector Sequences</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio Files</td>
<td></td>
<td></td>
<td>Decision</td>
</tr>
<tr>
<td>Features Extraction and Selection</td>
<td>Feature Vectors</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Recognition Rates: Classification of Sonic Units. Confidence: 1%. We have chosen the best parameter $\sigma$ for each SVM kernel: $\sigma = 1$ for the RBF kernel, $\sigma = 2$ for the kernel $\xi$ and $\sigma = 0.5$ for the two alignment kernels.

<table>
<thead>
<tr>
<th>System</th>
<th>GMM+DF</th>
<th>HMM</th>
<th>RBF-SVM+DF</th>
<th>$\xi$-SVM+DF</th>
<th>DTAK-SVM</th>
<th>DTAK-SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 subsegments</td>
<td>72.7</td>
<td>72.5</td>
<td>76.7</td>
<td>76.9</td>
<td>75.8</td>
<td>75.4</td>
</tr>
<tr>
<td>5 subsegments</td>
<td>73.1</td>
<td>70.5</td>
<td>75.5</td>
<td>75.5</td>
<td>73.7</td>
<td>77.2</td>
</tr>
</tbody>
</table>

Table 2: Recognition Rates: Classification of Summary Sequences. Confidence: 0.9%. In boldface are the best scores of the reference systems and of the alignment kernels.

The results of these experiments are presented in Ta-
ble 2. The scores of the SVM classifiers are obtained with
The main drawback of the alignment kernels which have been found efficient is the complexity of the resulting SVMs. This prevents us from running them on whole sonic units sequences. Thus, a useful development would be to explore ways to overcome this increase of complexity, e.g. by a subsampling or by reducing the number of kernel evaluations.

Another perspective for future work is the study of the influence of the weighting coefficients $m_s$ of the alignment algorithm (see Section 2.3), as they are potentially important kernel parameters.

**REFERENCES**


