Journal of New Music Research
2011, Vol. 40, No. 1, pp. 27–41

Routledge
Taylor & Francis Group

Evaluation of Features for Audio-to-Audio Alignment

Holger Kirchhoff[1] and Alexander Lerch[2]

[1]Queen Mary University of London, UK; [2]zplane.development, Berlin, Germany

Abstract

Audio-to-audio alignment is the task of synchronizing two audio sequences with similar musical content in time. We investigated a large set of audio features for this task. The features were chosen to represent four different content-dependent similarity categories: the envelope, the timbre, note-onsets and the pitch. The features were subjected to two processing stages. First, a feature subset was selected by evaluating the alignment performance of each individual feature. Second, the selected features were combined and subjected to an automatic weighting algorithm.

A new method for the objective evaluation of audio-to-audio alignment systems is proposed that enables the use of arbitrary kinds of music as ground truth data. We evaluated our algorithm by this method as well as on a data set of real recordings of solo piano music. The results showed that the feature weighting algorithm could improve the alignment accuracies compared to the results of the individual features.

1. Introduction

Audio-to-audio alignment describes the process of retrieving corresponding points in time in between two audio signals with the same or a similar content. It requires an analysis of the audio files that enables a mapping of points in time in one signal to points in time in the other signal.

The knowledge of those synchronization points enables a variety of different use cases.

- From a musicological point of view, the information could be used to analyse several recordings of the same piece of music and compare it to a given reference in order to investigate the tempo variations.
- In the same context, alignments are used to enable quick browsing for certain parts in recordings in order to easily compare parts auditorily (Dixon & Widmer, 2005; Müller, Mattes, & Kurth, 2006).
- In connection with a dynamic time-stretching algorithm, the knowledge of corresponding points in time can be used to adjust the timing of one recording to that of a second. This has a practical use especially in a music production environment. The different voices of a homophonic arrangement (e.g. the backing vocals in a pop song) can for example be automatically synchronized to the lead voice. The same applies for different instruments playing lines in unison or at least in the same rhythm.
- In film productions the audio track occasionally has to be re-recorded in case of unwanted noise and distortions in the original track. An automatic synchronization could aid matching the studio recording to the original track.

As these examples illustrate, the signals that are to be aligned can range from speech over monophonic music signals to complete mixes. Therefore, the criteria on which an alignment of the signals is based can be diverse.

In this work, we focus on four different types of inter-signal similarities and investigate the performance of various audio features in combination with the standard dynamic-time-warping algorithm (Rabiner & Juang, 1993). Most of these features are well established and can be assumed to give a meaningful representation of certain aspects of the content of audio signals.

The remainder of the paper is structured as follows: in the following section, previous work on audio-to-audio alignment is shortly summarized. In Section 3 we

Correspondence: Holger Kirchhoff, Centre for Digital Music, Queen Mary University of London, London, E1 4NS, UK.
E-mail: holger.kirchhoff@eecs.qmul.ac.uk

DOI: 10.1080/09298215.2010.529917

introduce the investigated features and describe the criteria by which they were chosen. An explanation of the evaluation procedure is given in Section 4, where we describe the construction of the ground truth as well as the evaluation metrics. The results of the single feature evaluation and the subset selection are presented in Section 5. Section 6 pursues the question of how to combine the features and introduces a new method for this task. This method is evaluated in Section 7. Section 8 finally concludes this work with a summary and an outlook.

2. Related work

Previous work can be found in the context of *audio matching* where different performances of the same piece of music are compared to allow for switching the playback between them. There are also closely related publications in the context of *audio-to-score alignment*; here, some approaches align a synthesized version of the score to the audio track and thus also perform audio-to-audio alignment.

Dixon and Widmer (2005) proposed an alignment-system called 'MATCH'. To measure the similarity between the audio frames of the recordings a single onset-related feature was used. For each audio track, the half-wave-rectified difference spectrum of consecutive non-linearly warped FFT spectra was calculated and the Euclidean distance was used to measure the cost between the frames of the different tracks. For the computation of the alignment path, the standard dynamic time warping (DTW) algorithm was modified to allow for a forward calculation of the path. The ground-truth data for the evaluation consisted of three different data sets: the first set included several piano recordings of the same piece of music played on a special grand piano that enabled the recording of the exact onset times of the notes. For the second data set, different commercially available piano recordings were labelled by a beat-tracking system. The third set consisted of non-piano music and was solely evaluated informally.

The system of Müller et al. (2006) mainly focused on reducing the computing time of the DTW to enable the analysis of long pieces such as whole movements of classical music. The proposed method was called 'multi-scale DTW' (MsDTW), and worked by iteratively increasing the resolution of the cost matrix in three stages. In stages two and three the previously computed path was used to define the constraint region for the current path calculation. As the only feature a filter-bank based pitch chroma (see below) was used, which was calculated at different frame sizes and frame rates, the smallest being 200 ms and 10 Hz. An evaluation of the alignment accuracy itself did not take place, only the deviation of the MsDTW path from the optimal unconstrained DTW path at the highest resolution was calculated.

In the work of Hu and Dannenberg (2003), several features including MFCCs, pitch histogram and pitch chroma representations (see below) were evaluated. The authors report that the pitch chroma proved to be most appropriate for the task. A DTW algorithm without locality constraints was used. The accuracy of the algorithm was evaluated by manually annotating five points in three different pieces of music and calculating the deviation of the DTW path at these specific positions.

Turetsky and Ellis (2003) investigated a larger set of features: a reduced version of the STFT-magnitude spectrum was raised to several powers and the discrete derivatives in both time and frequency domain were calculated. The alignment results were not subjected to an objective evaluation but simply compared by simultaneously listening to the audio and an aligned resynthesis of the score.

3. Feature set

It has already been mentioned in the introductory section that different use cases may require different similarity measures for the alignment procedure. More specifically, the similarities between the signals can be based on various musical parameters such as rhythm, pitch, timbre and dynamics. The voices of a homophonic arrangement for example differ in pitch, but may be aligned by rhythm or timbre information. Different instruments playing in unison may be synchronized by pitch and rhythm, but timbre information will be useless. Thus, we chose features that provide useful low-level information for the higher level musical parameters mentioned above. The features were grouped into the four categories:

- envelope features,
- timbre features,
- pitch features and
- onset features.

The features were extracted frame by frame. Both the length of the frame and the temporal distance between consecutive frames have influence on the attainable precision of the synchronization. As long as nothing else is specified in the following sections, a frame size of 23 ms was used for time domain features and a frame size of 46 ms for features in frequency domain. The step size (hop size) of 23 ms was the same for all features assuring that the features were calculated at the same points in time. Before the feature extraction process, the audio files were downmixed to mono and normalized so that the maximum absolute amplitude was set to 1.

In the following sections, $X(k, b)$ denotes the FFT spectrum of the b-th audio frame at the kth frequency bin index. The sample rate is denoted by f_s.

3.1 Envelope features

By visually examining the waveform of audio files in an arbitrary audio editor, an intuitive identification of similar parts is frequently possible. Especially when the signals were recorded in the same room and with the same musicians, the signal envelope enables the identification of corresponding parts. The envelope feature group includes on the one hand simple objective measures for the envelope itself and on the other hand perceptually motivated loudness measures. The following features were implemented:

- *Power (P)*: power level of each frame;
- *Maximum value (Max)*: level of the maximum absolute sample of each frame;
- *Loudness DIN 45631/ISO 532 B (L_{DIN}) and Loudness ITU-R BS.1387 (L_{ITU})*: these two loudness measures are based on the work of Zwicker and Fastl (1999) and utilize models for the outer ear function, the calculation of so-called excitation patterns to consider masking effects, and finally the overall loudness as the integral over the specific loudness values per critical band. The implementation stems from an open source project called 'FEAPI'.[1] Both features use a framesize of 0.74 s.

3.2 Timbre features

The timbre of an instrument describes its sound quality. Besides pitch and loudness, timbre is considered as 'the third attribute of the subjective experience of musical tones' (Rasch & Plomb, 1982). Unlike loudness and pitch—which are unidimensional properties as sounds with different loudness and pitch can be ordered on a single scale from quiet to loud and low to high, respectively—timbre is a multidimensional property. Although timbre is nowadays understood as a phenomenon that takes into account both temporal and spectral patterns (Moore, Glasberg, & Bear, 1997) the features presented below describe spectral shape only.

- *Spectral centroid (SC)*: the spectral centroid is defined as the centre of gravity of the power density spectrum. Our implementation follows the definition of the MPEG-7 standard (ISO/IEC, 2002). Listening test results indicate that the spectral centroid is well correlated to the perception of the brightness of a sound (v. Bismarck, 1974).
- *Spectral spread (SS)*: the spectral spread measures how far the spectral power is spread around the centroid. The exact definition can be found in the MPEG-7 standard (ISO/IEC, 2002).
- *Spectral rolloff (SR)*: the spectral rolloff is a measure for the extent of the spectrum. It computes the frequency below which 85% of the accumulated magnitude is concentrated (Scheirer & Slaney, 1997).
- *Spectral flatness (SF)*: the spectral flatness estimates the similarity of a given spectrum to the spectrum of white noise. It is defined as the ratio of the geometric and the arithmetic mean of the power spectrum (Jayant & Noll, 1984).
- *Mel frequency cepstral coefficients (MFCC)*: originally introduced to the speech processing domain, MFCCs have proven to be adequate to also describe similarities of music signals (Logan, 2000). The calculation we used here stems from Slaney (1998). It divides the magnitude spectrum into 40 mel bands. A discrete cosine transform is applied to the logarithmized mel spectrum. We used only the first five coefficients as features.
- *Mono strength (MS)*: this custom-designed feature provides information about whether the given spectrum contains a single note or several notes. It estimates the most salient fundamental frequency as the maximum of the *harmonic product spectrum (HPS)* (Schroeder, 1968) and relates the energy of the first harmonics $E_h(b)$ to the total energy of the spectrum $E_{total}(b)$. The HPS is computed by

$$HPS(k,b) = \prod_{i=1}^{N_{h,1}} |X(i \cdot k, b)|,$$

with

$$N_{h,1} = \min\left\{5, \text{floor}\left(\frac{f_s}{2 \cdot f(k)}\right)\right\}.$$

The energy of a tone with a fundamental frequency at index k_0 is computed as:

$$E_h(b) = \sum_{i=1}^{N_{h,2}} \max\{|X(i \cdot k_0 - 1, b)|^2,\ |X(i \cdot k_0, b)|^2,\ |X(i \cdot k_0 + 1, b)|^2\},$$

with

$$N_{h,2} = \text{floor}\left(\frac{f_s}{2 \cdot f_0}\right)$$

as the total number of harmonics in the spectrum. The result is finally calculated by

$$\text{MS}(b) = \frac{E_h(b)}{E_{total}(b)}.$$

[1] http://feapi.sourceforge.net/

3.3 Pitch features

The pitch-chroma is a well-established feature for the low-level representation of tonal content in polyphonic music signals. It is an octave-independent measure for the intensity of each pitch class (Bartsch & Wakefield, 2001).

- *Pitch chroma (PC) 1 and 2*: the first two pitch chroma computations apply weights to the magnitudes of the spectrum and sum up all magnitudes at the bins belonging to the same pitch class. Pitch chroma 2 uses triangular weighting filters, centred at the semitones of the equal-tempered scale and assuming a tuning frequency of 440 Hz (see lower plot in Figure 1). Pitch chroma 1 considers the case that the tuning frequency may not equal 440 Hz and that the pitches do not exactly match the equal-tempered scale. Therefore, the weighting filters have a flat frequency response in the range of 30 cent around the semitone's mid frequency and take the form of a trapezoid (see upper plot in Figure 1). For both pitch chromas, the following calculation rule is used:

$$\mathrm{PC}_i(l,b) = \sqrt{\sum_{\forall k \in K_l} (|X(k,b)| \cdot F_i(k))^2},$$

  where i takes the values 1 and 2, $F_i(k)$ denotes the weighting function and K_l is the set of frequency indexes of the pitch class l.
- *Pitch chroma 3*: instead of applying weights and accumulating the magnitudes, the third calculation rule uses the maximum amplitude of each semitone band. If a detected maximum lies at the edge of the band, it is checked whether the magnitudes at the adjacent frequency indexes are smaller than the maximum. Thus, only real local maxima are taken into account. The individual pitch classes are computed in the same way as with pitch chromas 1 and 2 except that only the maxima are summed.
- *Pitch chroma 4*: pitch chroma 4 searches for local maxima in the same way as pitch chroma 3. Additionally local maxima with a level smaller than 60 dB below the maximum amplitude are discarded.

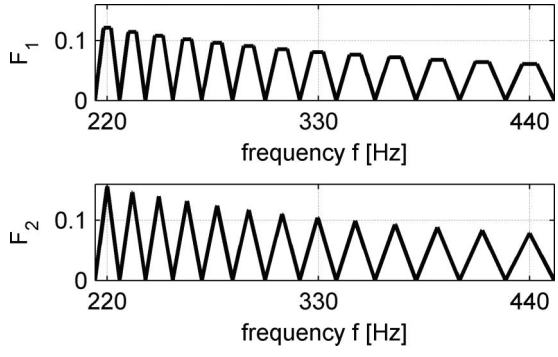

Fig. 1. Trapezoidal and triangular weighting of the semitone intervals.

3.4 Onset features

Especially the alignment accuracy of note onset times appears to be of utter importance for the alignment of musical sequences when considering the effort musicians put into playing synchronously. However, the alignment of two series of onset times seems to be prone to errors given the reduced amount of information in these series and the likelihood of onset detection errors. Therefore, instead of a series of discrete onset times we used a novelty function that indicates the likelihood of an onset occurrence for each analysis block.

- *Spectral flux (SX)*: one of the simplest measures for onset detection is the spectral flux, which is calculated as the Euclidean distance of consecutive short-time-magnitude-spectra (e.g. Scheirer & Slaney, 1997). Since the flux does not consider the direction of the magnitude variations, it is a measure for both onsets and offsets.
- *Detection function after Goto (OG)*: the detection function proposed by Masataka Goto (2001) computes a distance between the power spectra of consecutive audio frames. For every frame and for every frequency index a decision is made whether the power density in the preceding frame at the same index and the two adjacent indexes is less than the power density at the considered index of the current and the next frame. The distance is only increased if both the current and the successive frame have greater values than all of the three indexes of the preceding spectrum.
- Closely related to pitch chroma 1, the *difference pitch chroma 1 (DPC1)* weights the spectral differences with the trapezoid filters shown in Figure 1. This can be expressed by

$$X_{\mathrm{diff}}(k,b) = (|X(k,b)| - |X(k,\, b-1)|) \cdot F_1(k).$$

For each pitch class the squared differences are summed taking into account the signs of the differences:

$$X_{\mathrm{sum}}(l,b) = \sum_{\forall k \in K_l} \mathrm{sign}(X_{\mathrm{diff}}(k,b)) \cdot X^2_{\mathrm{diff}}(k,b).$$

Finally, the *DPC1* is then computed with

$$DPC1(l,b) = \mathrm{sign}(X_{\mathrm{sum}}(l,b)) \cdot \sqrt{|X_{\mathrm{sum}}(l,b)|}.$$

The symbols used here were explained for pitch chromas 1 and 2 in Section 3.3.

- *Difference pitch chroma 2 (DPC2)*: analogous to pitch chroma 3, this feature detects the maxima and minima of the spectral differences in each semitone band and chooses the one with the largest absolute value. The extrema are combined in the same manner as for pitch chroma 3.
- *Difference pitch chroma 3 (DPC3)*: the third variant equals DPC1 except that in the difference spectrum only magnitude increases are considered. This process is commonly denoted as half-wave rectification.

Note onsets and offsets may not only be described by magnitude increases in the frequency domain. Many other parameters may indicate note-on and note-off events. A transient for instance will in most cases exhibit a noisy spectrum while the subsequent quasiperiodic part will show a harmonic spectrum. In this case, the onset is characterized by a change in timbre. Onsets of percussive sounds on the other hand are often easily identifiable in the waveform. In this case, the time domain envelope can be used to characterize the onset. These examples make clear that an onset can be identified by any change of the features referred above. We thus additionally used the difference of consecutive feature values of the features described in the preceding sections as onset features.

3.5 Low pass filtering

It might be possible to increase the robustness of the synchronization even in the case of relatively dissimilar signals by applying a low pass filter to the features. The idea was to eliminate quick changes in the feature values so that the DTW path will follow the direction of the true path coarsely. Thus, additional features can be generated by applying a single pole low pass filter with a time constant of 250 ms to a large part of the features. In order to avoid a delay, the filter was used in both forward and reverse direction.

3.6 Feature summary and naming convention

In Sections 3.1–3.4, 23 different features were reviewed. For all features except for the two loudness measures, a low pass filtered equivalent was computed. Those features are labelled with an appended 'LP'. Additionally, the differences of consecutive feature values were computed for the features of all groups except for the group of onset features, which is denoted by a prepended 'Dev.'. Accordingly, a total number of 62 features were evaluated individually for their alignment accuracy (see Section 5.1). The onset feature group is the largest group with a total number of 28 features, followed by the group of timbre features with 20 features. The group of the pitch features contains only eight features—each 12-dimensional feature vector is treated as a single feature—and the envelope feature group contains only six features.

3.7 Feature postprocessing and evaluation

The features described in the preceding sections represent different properties of the audio signals and have different output ranges. When combining the features in a feature vector and calculating distances between those vectors, it is necessary to adjust their ranges in order to ensure that all features have the same influence on the distance calculation. But not only the range of the feature values is important, they should also possess the same or at least a similar distribution. For example, even if two features are designed to lie within the same range, the distribution of the first may have its maximum near the lower bound of the range and the distribution of the second near the upper bound. In this case, the distance calculation will be dominated by the feature that is concentrated near the upper bound.

3.7.1 Feature distributions

In order to match the feature distributions, a target distribution has to be chosen to which all the feature distributions will be transformed if necessary. In principle, any arbitrary distribution can act as target, however, in most cases the Gaussian or normal distribution is used because it is observable in many natural processes.

Various approaches exist to transform a given distribution into a normal distribution. Most common is the power transform resp. Box–Cox (1964) transform that requires the estimation of a parameter λ. However, the Box–Cox transform only considers a limited class of transformation functions and thus does not guarantee that any arbitrary distribution can be approximated to a normal distribution. More recently a numerical method has been proposed that exactly fits arbitrary distributions to the normal distribution (van Albada & Robinson, 2007). The drawback of this method, however, is that the transformation function for every feature has to be stored numerically.

To keep it simple, we did not apply any of these transformation methods to the features. Only those features that exhibited a nearly exponential distribution were transformed by the natural logarithm. All features that showed a single apex and an evident symmetry were left unmodified.

To test a given distribution for normality, various procedures exist (e.g. the Kolmogorov–Smirnov test, Lilliefors test, Shapiro–Wilk test, D'Agostino–Pearson omnibus test) (Thode Jr., 2002), but according to experience they fail for most audio feature distributions, since a very large number of observations is used. Instead

of using one of the above-mentioned normality tests, we simply calculated the *skewness* and the *kurtosis* of the features to measure the similarity to the normal distribution. The skewness measures the symmetry of the distribution; fully symmetrical distributions exhibit a skewness of 0. As a rule of thumb, distributions with values smaller than 2 are not significantly skewed (Miles & Shevlin, 2001). This rule applied to 90% of the unidimensional features described above, 65% even showed a skewness of less than 1. The kurtosis provides information about the steepness of the distribution, a normal distribution has a kurtosis of 0. A proportion of 85% of the investigated features had positive kurtosis values meaning that the distribution is more peaked than the normal distribution.

3.7.2 Normalization

After having ensured that the distributions are sufficiently similar, the features were standardized. In the case of a normal distribution this is done by subtracting the mean from all feature values and dividing them by the standard deviation. Since the distributions are slightly skew, the median is better suited to match the maximum of the distribution than the mean. Moreover, we used the root mean squared deviation from the median instead of the standard deviation as divisor. The standardization is given by

$$f_{i,\mathrm{std}}(b) = \frac{f_i(b) - m_i}{s_i}.$$

$f_i(b)$ denotes the i-th feature, m_i the median of the feature distribution and s_i the root mean squared deviation of the distribution which is calculated by

$$s_i = \sqrt{\frac{1}{B}\sum_{b=0}^{B-1}(f_i(b) - m_i)^2},$$

with B being the total number of observations of the feature.

The normalization of multidimensional features such as the pitch chroma requires special consideration. In the particular case of the chroma, the pitch information should be level-independent. Thus, each single vector has to be normalized. Usually the Manhattan or the Euclidean norm are applied. With the former, the vectors are mapped to a hyperplane in the 12-dimensional space, with the latter, the vectors are mapped to a hypersphere. This is illustrated for the two-dimensional case in Figure 2. We decided to apply the Manhattan norm to the chroma-vectors.

3.7.3 Principal component analysis

In order to get an impression of the correlation of the features, each of the four feature groups was subjected to a principal component analysis (PCA). The PCA gives information about the true dimensionality of the feature space. In the onset-feature group, we omitted the difference-pitch-chroma calculations, because the correlation of features with different dimensions cannot be computed. The correlation of the pitch chroma vectors was obtained by concatenating the observed chroma vectors to one vector, assuming that all pitch classes are equally likely to appear. The eigenvalue spectra of the PCA are displayed in Figure 3.

The results show that the correlation among the envelope features and the correlation among the pitch features is high. For the remaining two feature groups, the true dimensionality cannot be specified clearly because the eigenvalues decrease continuously and there is no clearly identifiable step in the eigenvalue spectrum. If we assume the dimensionality to be defined by the number of eigenvalues greater than 1—indicated by the dashed line in Figure 3—in both cases a dimensionality of 5 can be observed.

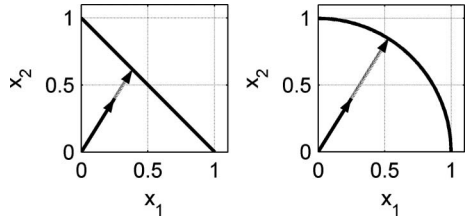

Fig. 2. Standardization of two-dimensional features. Left: Manhattan norm; right: Euclidean norm.

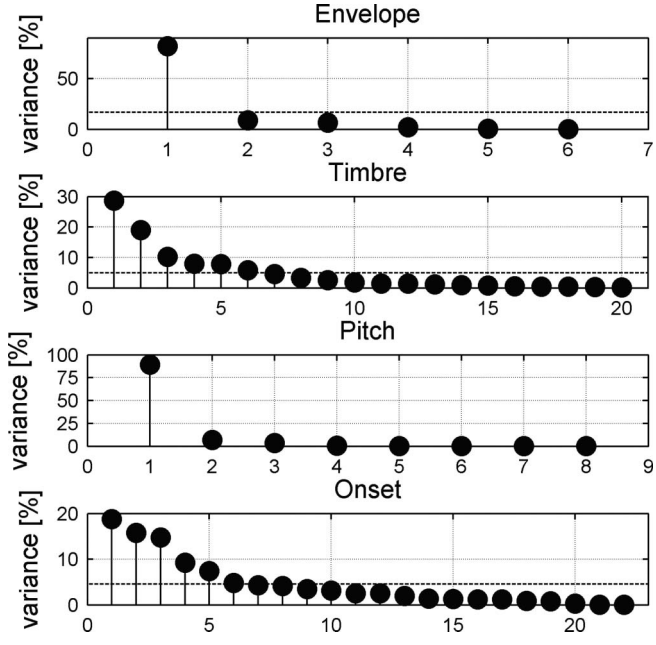

Fig. 3. Variance of principal components of the feature groups.

In the case of the pitch features, the high correlation is not surprising, since the calculation rules for the different chroma-features differ only slightly. The high correlation of the envelope-features is explained by the unidimensionality of the envelope itself. As mentioned before (Sections 3.2 and 3.4), both the timbre of a sound and the onset of musical events are multidimensional properties without a clearly defined dimensionality.

4. Data sets and evaluation metrics

In order to evaluate the accuracy of an estimated alignment path, a set of pairs of audio files with clearly defined synchronization points is necessary. As described in Section 2, earlier studies generated the ground truth data either by manually or automatically labelling certain corresponding points in time or by measuring onset times during the recording by using a computer-monitored grand piano. Manually annotating points in time is a rather arduous process and thus generally only few points can be labelled. The drawback of using the piano-generated data is its restriction to solo piano music; furthermore, confining the evaluation to note-onsets might be sufficient for the case of solo piano music, however, other kinds of music may also require the synchronization of the time in between onsets.

We propose another method to produce the ground truth data in Section 4.1 which enables the evaluation of alignment algorithms with high accuracy and on a wide range of musical styles and instrumentations. This dataset was used for the individual evaluation of the features (see Section 5.1) as well as for the training and evaluation of our proposed feature weighting algorithm (Section 6.1). This data set is based on artificially modified pairs of audio signals; in order to evaluate the algorithm on real recordings, we used a second data set which is described in Section 4.2.

4.1 Time-stretching data set

4.1.1 Description

The pairs of audio files for this data set were generated by subjecting a number of audio files to a widely used commercially available dynamic time-stretching algorithm[2] and subsequently applying timbre and pitch modifications to these signals. The audio files were chosen to represent typical use cases for the synchronization task and comprised several monophonic and polyphonic music signals as well as speech signals. All files were trimmed to a length of 30 s.

[2]élastique Pro by zplane.development

Two different tempo curves were applied to the audio files: for the first one, the time stretch factor was modulated by a triangular waveform at a frequency of $\frac{1}{3}$ Hz. The resulting stretch factors had the range $\left[\frac{3}{4};\ \frac{4}{3}\right]$. A study of different performances of a Beethoven string quartet indicates that the amount of this variation is a reasonable assumption (Lerch, 2008, pp. 115–116). The second tempo curve had a monotonically increasing stretch factor in order to ensure that a completely diagonal path—the path preferred by standard DTW if the similarity matrix entries are very similar to each other—will not result in unreasonably good evaluation results.

To simulate certain real-world use cases, further editing was applied to these signals: to simulate different singers and speakers, formant shifting was applied to some of the signals. Different voices of a homophonic arrangement were obtained by the use of an intelligent harmonizing plug-in and finally varying timbres of musical instruments were modelled by synthesizing MIDI-files with different sounds. Since not all of these processing steps are appropriate for any kind of audio signals, only those modifications were applied that seemed suitable for the chosen signals.

Since it is not reasonable to use pitch features to align audio signals with different pitches or to use timbre features to synchronize music played by different instruments, each test signal pair was assigned to those feature groups (see Section 3) that can in principle be used for the alignment process. In order to restrict the amount of data generated by the feature extraction stage, for every feature group 20 test signal pairs were chosen. These 80 pairs of approx. 30 s length contain a total of more that 200,000 observations and the alignment of each pair of audio files includes more than 1,500,000 comparisons between audio frames.

As an example, Table 1 shows the test set for the onset features. It contains different types of audio signals, ranging from solo instruments and voices to ensemble music up to male and female speakers. One half of the signals was processed with the first tempo curve (see above), the other half with the second. The additional modifications to the signals are listed in the right column. The amount of modified MIDI files was kept at a minimum: two of the signals of the envelope and the onset data set, five signals of the pitch data set and none of the signals of the timbre data set were synthesized MIDI files.

4.1.2 Metrics

With the ground truth described in the preceding section, it is possible to calculate the error of any given alignment path. Figure 4 exemplifies the calculation: the true path is shown in grey, the estimated path in black. δ_b denotes the deviation between the point in time to which frame

Table 1. Test set for onset features. Each test signal was subjected to a dynamic time-stretching algorithm, applying one of the two tempo curves. Further modifications of the stretched signal are shown in the right column.

No.	Content	Tempo curve	Modification
1	female singer	1	formant shift
2	female singer	2	upper part
3	male singer	1	formant shift
4	male singer	2	upper part
5	solo violin	1	upper part
6	solo saxophone	2	upper part
7	solo clarinet	1	upper part
8	solo trombone	2	different timbre
9	solo violoncello	1	different timbre
10	orchestra + choir	2	none
11	string quartet	1	none
12	wind quintet	2	none
13	voice and piano	1	none
14	chamber orchestra	2	none
15	male voice 1	1	formant and pitch shift
16	male voice 2	2	formant and pitch shift
17	male voice 3	1	formant and pitch shift
18	female voice 1	2	formant and pitch shift
19	female voice 2	1	formant and pitch shift
20	female voice 3	2	formant and pitch shift

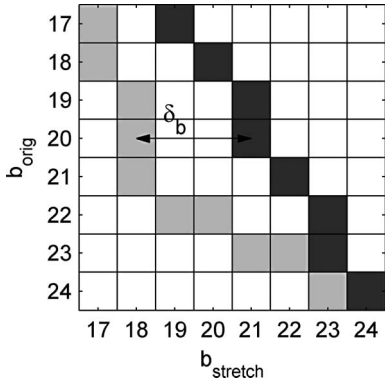

Fig. 4. Calculation of the alignment accuracy. b_{orig} and b_{stretch} indicate analysis blocks of the audio signals of a test signal pair; the reference path is shown in light grey, the calculated alignment path in dark grey.

number 20 of the original signal is mapped by the calculated path and its actual position in the time-stretched signal. Given the deviations at all frames, several overall error measures can be defined:

- the *mean deviation*:

$$\delta_{\text{mean}} = \frac{1}{B_{orig}} \sum_{b=0}^{B_{orig}-1} \delta_b,$$

with B_{orig} denoting the total number of frames in the reference audio file. The mean deviation gives information about a possible bias of the path;
- the *mean absolute deviation* measures the precision of the alignment and is thus the most important criterion:

$$\delta_{\text{abs}} = \frac{1}{B_{orig}} \sum_{b=0}^{B_{orig}-1} |\delta_b|;$$

- the *maximum deviation* is a measure for the robustness of the alignment:

$$\delta_{\max} = \max_b (|\delta_b|);$$

- another error measure that does not make use of the single deviations δ_b is the *relative number of matching path points* which calculates the ratio of the number of path points matching the ground truth to the total number of reference path points.

Since the mean deviation can be very small even if the alignment path shows large deviations and the relative number of matching path points may be small even if the alignment is relatively precise, mainly the mean absolute deviation δ_{abs} and the maximum deviation $\delta_{\max}$ were used for the evaluations in Sections 5 and 7.

4.2 Chopin data set

4.2.1 Description

The Chopin data set (Goebl, 2001) consists of recordings of two excerpts of solo piano music by F. Chopin (Etude in E major, op. 10 No. 3, bars 1–21 and Ballade in F major, op. 38, bars 1–45) played by 22 pianists on a computer-monitored grand piano. It is the same data set that was used by Dixon and Widmer (2005) (cf. Section 2). The length of the performances of the Ballade ranged from 1:52 to 2:31 min and the Etude recordings were between 1:10 and 1:34 min in length. For each recording the onset time of each played note is available, which enables the comparison between the corresponding note onset times of two performances and an alignment path.

4.2.2 Metrics

For this data set, the same evaluation metrics as described by Dixon and Widmer (2005) were applied. The authors define a *score event* as a set of simultaneously played notes according to the score. Each of these score events is assigned a unique onset time by averaging the slightly varying onset times of simultaneously played notes of the performance. This enables a unique mapping of the onset times of all score events of two distinct performances. Each mapping of a score

event can be marked as a single point within the distance matrix and the deviation to the alignment path can be computed as the Manhattan distance between each score event and the nearest point of the alignment path. Given all pointwise distances, the average and the maximum deviation can be computed.

Additionally the authors propose the computation of the percentage of deviations less than or equal to 0, 1, 2, 3, 5, 10, 25 and 50 frames.

5. Feature performance and subset selection

5.1 Individual feature performance

All features described in Section 3 were individually tested for their suitability for the synchronization task using the test set and the metrics introduced in Section 4.1. For every individual feature, a distance matrix between every test signal pair was computed. The entries of the distance matrix were calculated using the absolute difference of the unidimensional features, for multi-dimensional features—such as the pitch chromas—the Euclidean distance between the feature vectors was used. The path through the distance matrix was detected using the DTW algorithm. We used the standard DTW algorithm that allows single steps for the alignment path in horizontal, vertical and diagonal directions. No penalty was applied to the diagonal direction. The results are shown in Figure 5. The features are grouped into the four groups introduced in Section 3 and are sorted by their mean absolute error. The mean absolute error is displayed by the black bars, the grey bars show the maximum error. Since the error measures have different ranges, the domains are displayed at the left and right vertical axes.

When comparing the different feature groups, it is important to bear in mind that the results were obtained by means of different test signal sets (see Section 4.1). That means that depending on the use case and the given pair of audio signals, a pitch feature for instance may not yield a better synchronization result than an envelope feature. Comparisons between features of different feature categories are only possible provided that the signals are principally synchronizable by those categories.

Under this precondition the pitch features perform better than the remaining feature groups regarding both accuracy and robustness. The four pitch chromas nearly yield the same result, *PC3* achieves the best accuracy with a mean absolute deviation of 16.9 ms. The fact that the low pass filtered chromas achieve a higher mean absolute deviation is not surprising since fast variations of the features are eliminated. But even the worst feature (*PC2 LP*) still achieves a comparably low δ_{abs} of 40.7 ms. The maximum deviation, however, slightly decreases when using the low pass filtered chroma vectors, the lowest maximum error of 0.56 s is achieved by *PC3 LP*.

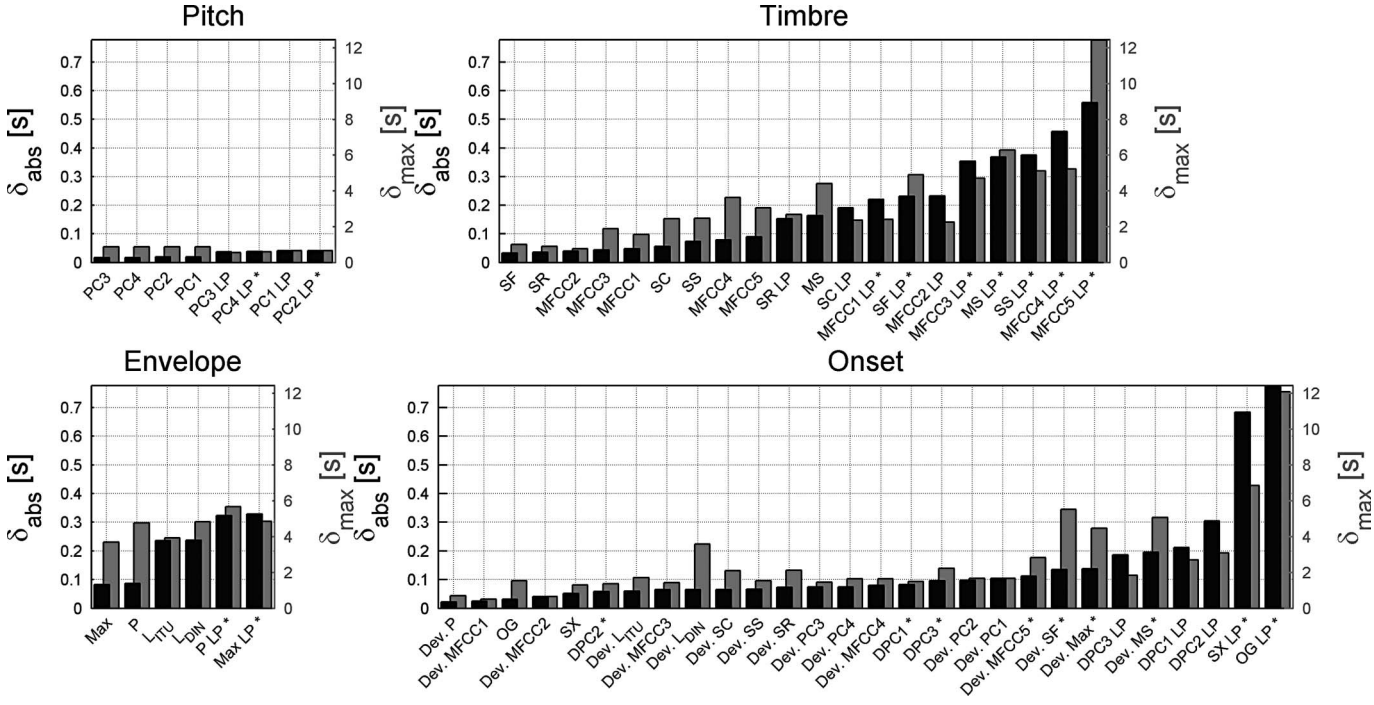

Fig. 5. Synchronization results of individual features: the left ordinate displays the mean absolute deviation of each feature (black bars), the right ordinate displays the maximum deviation (grey bars).

The envelope, on the contrary, seems to be a weak property to synchronize audio files, since even the best envelope feature (*Max*) yields a maximum deviation of approx. 3.7 s. Furthermore, the mean absolute deviation of the best envelope features is higher than that of many other features and amounts to 83.2 ms.

The results of the single features are significantly more spread in the remaining feature groups. While there are multiple features that perform well and exhibit mean absolute deviations of less then 100 ms, there are others that do not work at all for the task. Maximum deviations of more than 12 s at a file length of approx. 30 s indicate a complete failure of the alignment and thus a completely useless feature. Among the timbre features, the spectral flatness (*SF*) achieves the highest accuracy with a δ_{abs} of 33.5 ms, however, *MFCC2* yields a smaller δ_{max} of 0.77 s. In the onset feature group, the use of the first difference of the power feature (*Dev. P*) results in a mean absolute deviation of 22.7 ms, the smallest maximum deviation of 0.51 s is achieved by *Dev. MFCC1*.

5.2 Subset selection

To reduce dimensionality and to decrease computational cost, the worst features were eliminated from the feature set. The subset selection was accomplished only based on the results of the single feature analysis according to the following rule: for each group, those features that belonged to the 40% worst in terms of the mean absolute deviation were discarded when they at the same time belonged to the 40% worst with respect to the mean deviation or the maximum deviation. All discarded features are marked by an asterisk in Figure 5.

6. Feature space transformation

The alignment results can be improved further by using multiple features to represent the audio. Provided that each feature contributes new information about the audio signals, the combination of features may lead to better results than using just one single feature. The features can be combined in a feature vector $\mathbf{f}(b)$ per block. This vector weights all features equally. However, it can be assumed that the optimal alignment accuracy requires unequal feature weights because individual features may be of different importance for the computation of the alignment path.

In the following subsection we propose a supervised learning procedure that automatically finds the weights for the features. Section 6.2 introduces different choices of the training data for the learning algorithm and in Section 6.3 different ways of handling silence within the audio files are discussed. The proposed algorithm is finally evaluated in Section 7.

6.1 Feature weighting procedure

Finding the feature weights that lead to an optimal alignment can be regarded as a classical optimization problem. The cost function can be any one of the error measures from Section 4.1.2. Standard approaches such as the gradient descent cannot be applied because it is not possible to describe the iterative procedure of the DTW as a mathematical function that can be differentiated. Furthermore, other iterative optimization methods that require many iteration steps are not easily applicable due to the high computation time of the time warping.

The approach we used in this study is based on the fact that the path of the ground truth will be found when the distances on this path will be small compared to the distance values in the rest of the distance matrix or at least in the neighbourhood of the ground truth path. Figure 6 illustrates the optimal case, in which all distances on the ground truth path are 0 and all other distances of the matrix are 1.

It will of course not be possible to find weights so that all non-path values of the distance matrix will be set to 1 because large content-based similarities can also occur at points not belonging to the ground truth path. The aim is rather to minimize the distances on the path and to maximize those in the vicinity of the path.

We denote the difference vector between two arbitrary feature vectors $\mathbf{f}_{\text{orig}}(b_i)$ and $\mathbf{f}_{\text{stretch}}(b_j)$ as $\Delta\mathbf{f}(b_i, b_j)$. It is calculated by

$$\Delta\mathbf{f}(b_i, b_j) = \left|\mathbf{f}_{\text{orig}}(b_i) - \mathbf{f}_{\text{stretch}}(b_j)\right|.$$

As explained above, we tried to map all difference vectors on the true path to a small value—in the optimal case to 0—and to map the distances near the path to a high value, defined as 1. This can be seen as a two-class classification that assigns a given difference vector to one of the classes '*on the ground truth path*' (C_1) or '*not on the*

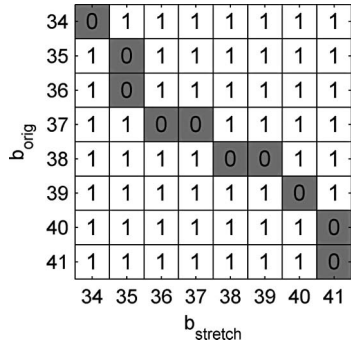

Fig. 6. Distance matrix for an optimal alignment result: all distances on the ground truth path amount to 0, all other distances amount to 1.

ground truth path' (C_2). Accordingly, the probability $P(C_2|\Delta\mathbf{f}(b_i, b_j))$ can be used as a distance measure for the distance matrix. Linear Discriminant Analysis (LDA) as a simple but robust classifier was chosen. LDA separates the data by a hyperplane that minimizes the classification error. An introduction to LDA can be found in Duda, Hart, and Stork (2000).

6.2 Training data

The training data for the classifier can be generated by means of the test set introduced in Section 4.1. The difference vectors on the ground truth path were used as the training data for class C_1. For the selection of the training data of class C_2, four different possible choices were investigated (see Figure 7): points in the immediate vicinity of the ground truth path (set 1), points with a distance of one point (set 2) and two points (set 3) in the diagonal direction from the ground truth path, and finally points randomly chosen from the distance matrix on one side of the path (set 4).

Training set 1 has the potential of increasing the contrast between the distances on the ground truth path and distances very close to the path. A high contrast between those directly adjacent points all along the ground truth path will cause the DTW path to accurately follow the ground truth path. However, it may not be easily possible to increase the contrast between these two groups of points, because distance vectors at adjacent points are likely to take similar values and thus may not be well separable by LDA. Hence, if the contrast is not high enough, the DTW path may deviate from the ground truth path and thus reduce the alignment accuracy.

Training set 4 on the other hand aims at coarsely increasing the distance values of the points off the ground truth path with no special emphasis on points in the vicinity of the ground truth path. This may not enhance the accuracy of the alignment path but may lead to more robust results. The remaining training sets represent a compromise between these two approaches by choosing points at greater distances from the ground truth path than training set 1.

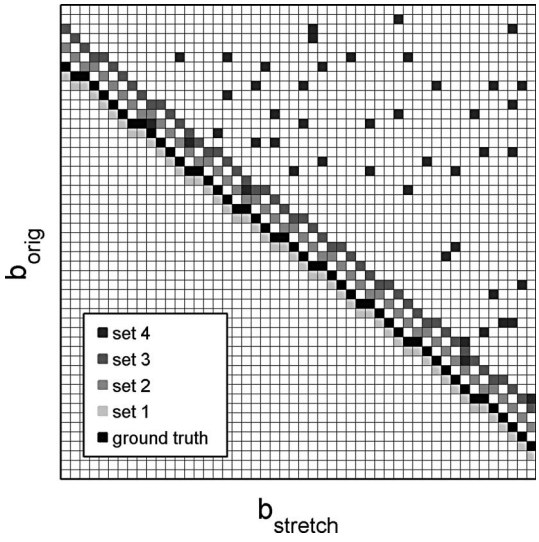

Fig. 7. Different choices for the training data. Points on the ground truth path represent class C_1, points for class C_2 are chosen in four different ways: in direct vicinity to the ground truth, in a distance of one and two points in diagonal direction, and randomly from one side of the ground truth path.

6.3 Silence

Pauses within the audio files pose a special problem for a content based alignment algorithm because there is no information by which the alignment could be accomplished. Furthermore, within pauses there are no points in time which actually correspond to each other. Thus, the best way for the alignment path within the distance matrix would be to go straight from the preceding to the subsequent non-pause content. This would require another processing step to detect the beginning and the end of a pause.

In this work, we do not propose methods for the detection of start and end points of musical pause sections. However, the influence of pause frames for the training process of the classifier was investigated. A pause frame was defined by a very simple criterion: audio frames of the normalized audio signals with a power of less than -60 dBFS were denoted as pause frames.

7. Evaluation

The evaluation was carried out on the two different data sets introduced in Section 4. The time-streching data set was used to learn the feature weights by the procedure described in Section 6. We report the results on this data set in Section 7.1. In order to verify that similar results can be obtained on real recordings, we evaluated the weighted features on the Chopin data set (cf. Section 4.2). These results are discussed in Section 7.2.

7.1 Time-stretching data set

The evaluation of the procedure described in Section 6 was accomplished in three steps. As a first evaluation step, the influence of different training data sets (see Section 6.2) on the alignment results was investigated. Subsequently, we surveyed the consideration of pause frames (see Section 6.3) during the training process using the best training data set. In a third step, we compared the results of each feature group to the results when using the standardized feature vectors with equal weights.

As opposed to Section 5, we calculated the error measures for each audio signal pair of the test set individually. Ten-fold cross-validation was used to compute the error measures: the test set for every feature group was randomly divided into 10 equal folds, nine folds were used for the training and the algorithm was tested on the tenth fold.

For the evaluation of the different training data sets and the pause frames, the results of all test signal pairs of all feature groups were considered, whereas for the third evaluation step the results were evaluated for each feature group individually.

Table 2 shows the results of the different training data sets. For each training data set, it displays the minimum, median and maximum value of the δ_{abs} and the δ_{max} per test signal pair of all pairs. That means that e.g. for set 1 the smallest δ_{abs} that could be achieved by one of the test signal pairs amounted to 2.82 ms and the maximum δ_{abs} of a different pair to 179.1 ms.

Except for the maximum δ_{abs} and δ_{max} among the test signal pairs, the results showed only minor differences. Set 4 yielded the lowest precision. For this set, the minimum and the median of δ_{abs} were the highest among the four sets. However, the maximum δ_{abs} and δ_{max} lay considerably below those of sets 1 and 2. Set 3 on the other hand, the training set with a distance of two data points from the ground truth path, showed the best results. All of the displayed error values of this set exhibited smaller amounts than any of the three other sets.

Table 2. Results for the four different training data sets: minimum, median and maximum of the δ_{abs} and δ_{max} of the test signal pairs. For comparison the last row displays the results when using equally weighted features.

	δ_{abs}			δ_{max}		
set	min	median	max	min	median	max
1	2.82 ms	9.18 ms	179.1 ms	23.2 ms	75.5 ms	1.44 s
2	2.77 ms	8.34 ms	262.6 ms	23.2 ms	69.7 ms	2.07 s
3	2.73 ms	8.18 ms	91.5 ms	23.2 ms	69.7 ms	0.88 s
4	3.19 ms	9.32 ms	108.7 ms	23.2 ms	110.3 ms	0.95 s
eq. w.	2.80 ms	8.26 ms	454.38 ms	23.2 ms	92.9 ms	3.87 s

Table 3. Comparison of the training with and without pauses: minimum, median and maximum of the δ_{abs} and δ_{max} of the test signal pairs. The last row shows the results when using equally weighted features.

	δ_{abs}			δ_{max}		
pauses	min	median	max	min	median	max
with	2.73 ms	8.18 ms	91.5 ms	23.2 ms	69.7 ms	0.88 s
without	2.73 ms	8.61 ms	98.2 ms	23.2 ms	69.7 ms	0.88 s
eq. w.	2.80 ms	8.26 ms	454.38 ms	23.2 ms	92.9 ms	3.87 s

Based on the usage of training set 3, Table 3 displays the results of the training with and without pauses. The results differed only slightly and δ_{max} showed no difference at all. The minimum δ_{max} of 23.2 ms corresponds to a deviation of only one audio frame at the applied step size and sample rate. When comparing the results of δ_{abs} only the median and the maximum are slightly lower when pauses were considered during the training. A possible reason for this might be that the audio signal pairs of the test set originated from the same file, so that the noise still had a similar quality and a similar envelope so that an alignment might have been possible even within pauses.

Figure 8 displays the alignment results of the proposed feature combination procedure using training data set 3 and considering pauses during the training process. The box and whisker plots enable comparing the results to those that are obtained when vectors of equally weighted features are used for the distance calculation. The upper row shows the mean absolute errors, the lower row the maximum errors, both with logarithmically scaled vertical axes.

The results of the envelope feature group were improved most noticeably by using the proposed method. The median of mean absolute deviation δ_{abs} was lowered from 50 ms to approximately 20 ms and the maximum value of δ_{abs} was reduced from 454 to 91 ms. The same applies to the δ_{max} of this feature group: it was decreased from 3.8 to 0.7 s. A right-tailed two-sample t-test with a significance level of 5% showed that the mean of both error measures is significantly lower when using the proposed weighting method.

For the timbre features, smaller deviations from the reference path can be observed. The maximum δ_{max} was decreased from 0.24 to 0.18 s and the already small median of δ_{abs} was also slightly lowered to a value of 4.5 ms. However, for this feature group, the improvements are insignificant at a significance level of 5%. The mean of the maximum deviation is significantly lower when choosing a significance level above 6.2%.

For the remaining two feature groups the improvements are less obvious. While the median values nearly stay the same, there seems to be a tendency of slightly lower quartile boundaries. The changes for these two groups, however, are statistically insignificant.

7.2 Chopin data set

For the Chopin data set alignments were computed for all pairwise combinations of audio files. Given the 22 recordings for each of the music excerpts, a total number of 462 pairs was considered. We evaluated all feature

groups individually as well as all combinations of two feature groups on all audio signal pairs.

The results of this evaluation are shown in Table 4. The table displays the percentages of deviations less than 0, 1, 2, 3, 5, 10, 25 and 50 frames as well as the average error and the maximum error for each feature group and each of the two excerpts. In particular the average error and the maximum error enable a comparison between the results of the time-stretching data set and the Chopin data set.

Although generally a little bit higher than the results of the time-stretching data set, a similar trend of the four feature groups is recognizable: the onset and pitch feature groups yield significantly more accurate alignment results than the envelope features. While the average error of the onset and pitch features is in the

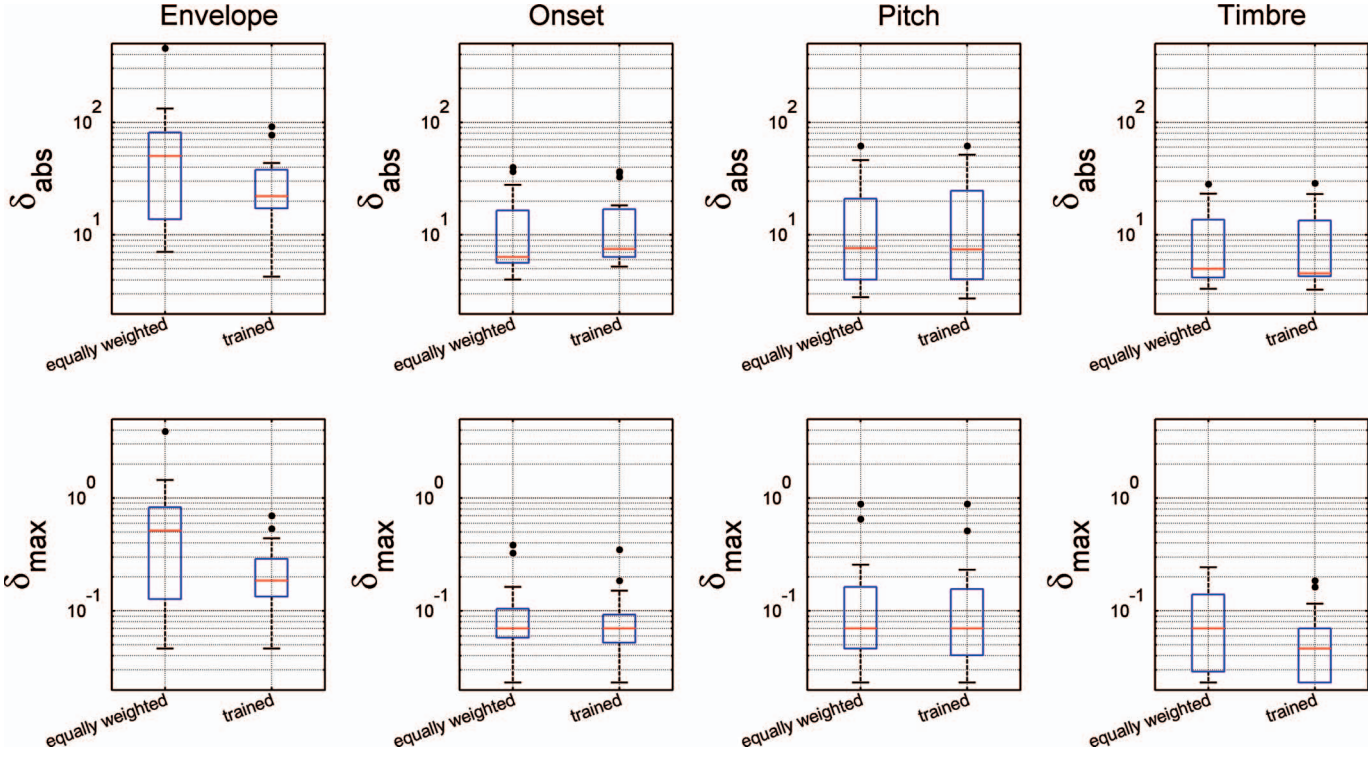

Fig. 8. Comparison of the alignment accuracy between the proposed procedure and vectors of equally weighted features.

Table 4. Results for the evaluation of the weighted features on the Chopin data set. The percentage of frames less or equal than 0, 1, 2, 3, 5, 10, 25 and 50 audio frames is displayed for the four feature groups and the combination of onset and pitch features for the two music excerpts. The last rows show the average and maximum errors.

	Cumulative percentage									
	Envelope		Timbre		Onset		Pitch		Onset + Pitch	
Error ≤ Frames	Etude	Ballade	Etude	Ballade	Etude	Ballade	Etude	Ballade	Etude	Ballade
0	21.6	23.5	27.3	25.1	36.6	32.0	43.4	36.8	42.8	36.9
1	47.8	55.8	57.9	58.5	76.9	71.4	85.0	80.3	85.2	80.3
2	58.5	71.4	68.3	73.6	89.4	86.5	94.1	94.4	95.0	94.3
3	64.0	79.4	73.2	80.4	93.6	92.1	96.1	97.7	97.3	97.8
5	69.9	87.8	78.4	86.8	97.2	95.9	97.1	98.7	98.7	99.1
10	75.8	95.2	83.6	92.2	99.1	98.4	97.7	98.9	99.4	99.5
25	90.7	97.9	91.5	95.4	99.8	99.7	99.2	99.2	99.9	99.7
50	96.9	99.2	96.2	97.8	100.0	99.9	99.8	99.6	100.0	99.9
Average error [ms]	202.5	77.6	182.8	122.1	32.5	38.9	35.9	35.1	24.9	28.2
Maximum error [s]	4.36	5.71	7.72	7.04	3.08	4.54	2.80	5.85	1.87	3.66

order of 35 ms, the envelope feature group exhibits average errors of more than 77 ms. The maximum error of all these three groups ranges from 2.8 to 5.85 s. This value seems to be dependent on the audio content, as the maximum error is generally lower for the Etude compared to the Ballade. Somewhat surprising are the results of the timbre feature group which seem to be significantly worse than those of the other groups regarding both average and maximum error. Here we found that the algorithm specifically had problems with certain parts in the recordings where musical motifs were repeated several times. The algorithm aligned different versions of these motifs resulting in comparably large deviations from the ground truth.

We also evaluated the algorithm on all combinations of two feature groups. This was accomplished by summing the entries of the distance matrices of the feature groups and computing the alignment path from this combined matrix. The best result was obtained by the combination of onset and pitch features which is displayed in the two rightmost columns of Table 4. These results are of the same magnitude as the results reported by Dixon and Widmer (2005) who also use a feature that takes into account both onset and pitch information (cf. Section 2). Compared to this study, our procedure could improve the average error of the Ballade by approx. 7 to 28.2 ms and the maximum error of the Etude by approx. 0.5 to 1.87 s.

8. Conclusion

In this paper, it was illustrated why the choice of suitable features has to depend on the use case at hand.

We presented an approach to objectively measure the quality of an alignment by constructing a ground truth data set and defining several error measures. This method enabled the evaluation and comparison of different features and feature combinations.

It was demonstrated that various features beyond the feature set of previous studies are suitable for the task of audio-to-audio alignment. Precise alignments can occasionally be achieved even by using one individual feature. However, the use of multiple features makes the algorithm more accurate and robust in general.

Furthermore, it was shown that applying different weights to the individual features can in some cases improve the alignment results significantly.

The results for the alignment accuracy are generally quite satisfactory and seem to be sufficient for most of the real world applications. However, in future work we aim to investigate how the feature groups can be combined for certain specific real world use cases. Furthermore, we plan to evaluate our proposed method on a data set of real recordings of various kinds of music and with different instrumentations.

References

Bartsch, M.A., & Wakefield, G.H. (2001). To catch a chorus: Using chroma-based representations for audio thumbnailing. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, pp. 15–18.

Box, G.E.P., & Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252.

Dixon, S., & Widmer, G. (2005). MATCH: A music alignment tool chest. In *6th International Conference on Music Information Retrieval (ISMIR 2005)*, London, UK, pp. 492–497.

Duda, R.O., Hart, P.E., & Stork, D.G. (2000). *Pattern Classification* (2nd ed.). New York: Wiley.

Goebl, W. (2001). Melody lead in piano performance: Expressive device or artifact? *Journal of the Acoustical Society of America*, *110*(1), 563–572.

Goto, M. (2001). An audio-based real-time beat tracking system for music with or without drum sounds. *Journal of New Music Research*, *30*(2), 159–171.

Hu, N., & Dannenberg, R. (2003). Polyphonic audio matching for score following and intelligent audio editors. In *Proceedings of the International Computer Music Conference*, Singapore, pp. 27–34.

ISO/IEC. (2002). Information technology – multimedia content description interface. ISO/IEC 15938. Retrieved from http://standards.iso.org/iso/

Jayant, N.S., & Noll, P. (1984). *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice Hall.

Lerch, A. (2008). *Software-based extraction of objective parameters from music performances* (PhD thesis). Technical University Berlin, Germany.

Logan, B. (2000, 23–25 October). Mel frequency cepstral coefficients for music modeling. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR 2000)*, Plymouth, MA, USA.

Miles, J., & Shevlin, M. (2001). *Applying Regression and Correlation: A Guide for Students and Researchers*. Thousand Oaks, CA: SAGE Publications.

Moore, B.C.J., Glasberg, B.R., & Bear, T. (1997). A model for the prediction of thresholds, loudness and partial loudness. *Journal of the Audio Engineering Society (JAES)*, *45*(4), 224–240.

Müller, M., Mattes, H., & Kurth, F. (2006). An efficient multiscale approach to audio synchronization. In *Proceedings of the 7th International Symposium on Music Information Retrieval (ISMIR 2006)*, Victoria, Canada, pp. 192–197.

Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice Hall.

Rasch, R.A., & Plomb, R. (1982). The perception of musical tones. In D. Deutsch (Ed.), *The Psychology of Music* (2nd ed.) (pp. 89–109). New York: Academic Press.

Scheirer, E., & Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1997*, Munich, Germany, pp. 1331–1334.

Schroeder, M.R. (1968). Period histogram and product spectrum: New methods for fundamental-frequency measurement. *Journal of the Acoustical Society of America*, *43*(4), 829–834.

Slaney, M. (1998). *Auditory toolbox – a MATLAB toolbox for auditory modeling work* (Technical report). Interval Research Corporation, Palo Alto, CA.

Thode Jr., H.C. (2002). *Testing for Normality*. New York: Marcel Dekker.

Turetsky, R.J., & Ellis, D.P. (2003). Ground-truth transcriptions of real music from force-aligned MIDI-syntheses. In *4th International Symposium on Music Information Retrieval (ISMIR) 2003*, Baltimore, MD, USA, pp. 135–141.

v. Bismarck, G. (1974). Sharpness as an attribute of the timbre of steady sounds. *Acustica*, *30*(3), 159–172.

van Albada, S.J., & Robinson, P.A. (2007). Transformation of arbitrary distributions to the normal distribution with application to EEG testretest reliability. *Journal of Neuroscience Methods*, *161*(2), 205–211.

Zwicker, E., & Fastl, H. (1999). *Psychoacoustics: Facts and Models* (2nd ed.). Berlin: Springer.