Towards Polyphony Reconstruction Using Multidimensional Multiple Sequence Alignment

Dimitrios Bountouridis[1(✉)], Frans Wiering[1], Dan Brown[2], and Remco C. Veltkamp[1]

[1] Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands
d.bountouridis@uu.nl
[2] David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada

Abstract. The digitization of printed music scores through the process of optical music recognition is imperfect. In polyphonic scores, with two or more simultaneous voices, errors of duration or position can lead to badly aligned and inharmonious digital transcriptions. We adapt biological sequence analysis tools as a post-processing step to correct the alignment of voices. Our multiple sequence alignment approach works on multiple musical dimensions and we investigate the contribution of each dimension to the correct alignment. Structural information, such musical phrase boundaries, is of major importance; therefore, we propose the use of the popular bioinformatics aligner MAFFT which can incorporate such information while being robust to temporal noise. Our experiments show that a harmony-aware MAFFT outperforms sophisticated, multidimensional alignment approaches and can achieve near-perfect polyphony reconstruction.

1 Introduction

Optical music recognition (OMR), has been one of the earliest applications of optical character recognition dating back to the late 1960s. The goal is to parse a printed music sheet, typically through scanning, and convert its elements (e.g. notes, clefs) to a digital format. From there, one can visualize, process or play back the digitized score. The OMR process typically comprises image recognition and machine learning components; however, despite the technology advancements, OMR has been imperfect with frequent pitch and temporal errors.

Recognition errors arise due to low quality printing, ambiguous music notation and written music's higher complexity than traditional text e.g. merged staff symbols. Temporal errors are those arising from predicting the wrong position (onset) or duration of a note. For example, incorrectly recognizing an eighth note as a quarter will result in all the following notes being shifted by an eighth. Multiple errors of this type tend to accumulate. The problem becomes more pronounced when dealing with polyphonic scores, since temporal note shifts can

J. Correia et al. (Eds.): EvoMUSART 2017, LNCS 10198, pp. 33–48, 2017.
DOI: 10.1007/978-3-319-55750-2_3

result to an alignment of voices that besides being incorrect, can sound *inharmonious* as well. Interestingly, incorrect alignment of voices can occur even if the OMR is perfect, i.e. in printed musical sources from the 16^{th}–18^{th} centuries. The music from that period is generally typeset using a font consisting of a limited set of musical symbols. Because of this property, it is possible to attain recall rates between 85% and 100% on good-quality prints [19]. However, there are several complicating factors, two of which are particularly important for this research. One is that music was generally not printed in score format but in separate voices, each in its own "partbook". The other is that barlines were only introduced around 1600, so that an important mechanism for coordination of voices that might counterbalance rhythmical errors in the OMR, is missing.

Polyphony reconstruction can be defined as the task of restoring a polyphonic piece to its original temporal formation-arrangement. In the same manner, *polyphony construction* can be defined as the prediction of the temporal formation of a set of unaligned voices. Understanding the temporal aspect of polyphony, a requirement for both tasks, can find application to automatic music generation and musicological analysis as well. Surprisingly, there is a lack of related research, which can be partially attributed to the high complexity of the problem: cognitively, aligning music notes with each other to form "meaningful" polyphonic pieces, is a process involving many musical dimensions such as harmony, durations and structure (e.g. segments, phrases, repetitions).

Interestingly, the synchronization of music voices has many parallels to the well studied multiple sequence alignment problem in bioinformatics [28]. In biological sequence analysis, measuring the similarity of more than two sequences is performed by examining possible multiple sequence alignments (MSA) in order to find an optimum, given a "meaningful" distance measure [5]. Traditionally, MSA algorithms are applied on unidimensional sequences; however, similar to music, some biological sequences have multiple dimensions and consequently MSA approaches that can deal with such information have been investigated through the years; for example, MSA methods that are aided by proteins' secondary structure (an abstraction of the three-dimensional form of local segments) [18].

We adapt tools used in biological sequence analysis towards understanding the temporal aspect of polyphony and towards solving polyphony reconstruction in particular. We employ a multidimensional MSA approach that allows us to identify the contribution of each musical dimension to the correct reconstruction. Our first round of experiments shows that, besides harmonic relations, structural information is a crucial for the task. However, structural information is rarely available, and since most algorithms for automatic structure analysis rely on temporal information (e.g. onset positions, durations), their predictions can be highly unreliable when temporal corruption is present (e.g. due to OMR errors). To accommodate for such corruption, we propose the use of the MSA algorithm MAFFT (Multiple Alignment using Fast Fourier Transform) [14], which can guide the alignment by structural segments computed from non-temporal information, such as pitch. We show that a harmony-aware MAFFT can almost perfectly reconstruct a artificial dataset of temporally-corrupted polyphonic pieces; a fundamental step towards real-life applications.

The remainder is organised as follows. Section 2 presents a brief overview of the related literature. Section 3 defines the problem of polyphony reconstruction and explains the simplifications we make in order to reduce its inherent complexity. Section 4 investigates the importance of musical dimensions to polyphony reconstruction. Section 5 introduces MAFFT and proposes its use for the task. Discussion and conclusions are presented in Sect. 6.

2 Related Work

OMR is an area of active scientific research and as consequence, various solutions have been proposed through the years. However, the task is merely secondary to the scope of this paper. The reader is referred to [20] which provides a complete overview of the algorithms and related literature regarding OMR. We note that only few of the proposed systems address historical forms of music notation, such as mensural notation from the Renaissance or lute tablature. The best results on Renaissance polyphony (the notation that inspired this study) is attained with Aruspix[1].

Few relevant works are tangentially related to polyphony reconstruction. For example, Boulanger-Lewandowski *et al.* [2] use recurrent neural networks (RNN) to model temporal dependencies between polyphonic voices for the purposes of music generation and music transcription. Similarly, Lyu *et al.* [17] propose the fusion of a Long Short-Term Memory RNN and restricted Boltzmann machines (RBM) for the purpose of music generation. However, none of the approaches provide any insights regarding the cognitive process of polyphony construction.

A number of researches outside bioinformatics have adopted multidimensional multiple sequence alignment (MDMSA) through the years. For example, Joh *et al.* [13] use MDMSA to compute the similarity between activity patterns. Sanguansat [22] uses a multidimensional version of the Dynamic Time Warping (DTW) algorithm for the task of query-by-humming. Closer to our work, van Kranenburg [25] uses a multidimensional extension of the DTW-based pairwise alignment. In his work, the scoring function incorporates heuristics to accommodate for more dimensions and is applied on melody classification.

3 Problem Definition and Polyphony Representation

Monophonic melodies can be considered as sequences of three-dimensional objects known as notes. Each note n_i can be represented by its pitch, duration and onset components: $n_i = (p_i, d_i, o_i)$. Perceptually, these dimensions never appear in isolation but constantly interact with each other. For example, onsets and durations allow us to perceive the *rhythm* dimension. Pitch patterns and rhythm create melodic segments (e.g. phrases). In polyphonic pieces, where multiple melodic sequences (voices) sound in parallel, the dimension interaction is higher. For example, the polyphonic temporal organization of notes creates more

[1] www.aruspix.net.

complex rhythm patterns. In addition, notes of different pitches from different voices happening in similar onset times allow us to perceive *harmony*.

Pitch errors aside, errors in the duration and onset dimensions, such as those due to incorrect OMR, lead to both incorrect harmony and rhythm. The original piece can be temporally reconstructed, as soon as both components are corrected from errors. In this paper we are solely interested in reconstructing the original harmony as a first step towards the complete polyphony reconstruction.

The representation of a polyphonic piece as a tractable harmony structure is of major importance for the task. The sequential nature of certain music documents (e.g. melodies, chord progressions) has allowed for their representation as sequences of symbols in a wide range of pattern recognition and Music Information Retrieval tasks (MIR) tasks. It is therefore logical to apply this successful scheme to each voice in a polyphony. Our interest in harmony reconstruction solely, allows us to use a representation that does not consider the duration and onset dimensions. Each voice is represented as a sequence of pitch values folded into one octave and mapped into an alphabet.

The question now is how to encode the harmonic relations between the sequences. Multiple sequence alignment (MSA) is the arrangement of sequences (via the introduction of gaps "-") so that they have they have the same length, while keeping related symbols aligned. MSA seems like a perfect fit to encode harmony: gaps "-", that can be interpreted as rests, can be introduced to the pitch sequences such that original pitch alignments are retained (see Fig. 1).

The MSA representation of harmony allows us to reformulate the harmony reconstruction task: assume a polyphonic score S and a function $h : S \rightarrow A_{harm}$ that maps the score's harmony into a multiple sequence alignment (A_{harm}), for example by removing durations and focusing on simultaneous pitches. Given S^*, a corrupted version of S in terms of note duration and onsets, our goal is to find

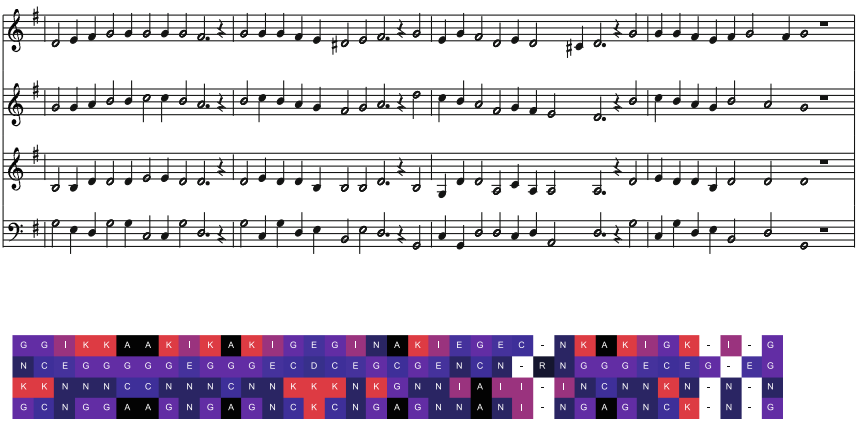

Fig. 1. A polyphonic score represented as a duration-onset-agnostic MSA (bottom). Colors are used for visualization purposes. Only the harmonic relations are retained from the original score. (Color figure online)

a score correction function f such that $h(f(S^*)) = h(S)$. To simplify, given a corrupted score in terms of onsets and durations, our goal is to realign the pitch sequences (voices) so that the original harmony is reconstructed.

4 Importance of Musical Dimensions

Our musical intuition suggests aligning music voices is a multidimensional process. Obviously, the relation between note pitches should meet the stylistic requirements for consonance and dissonance. However, one could posit that for example, it is more likely for notes of similar duration to be aligned together. This section aims to investigate to which extent various musical dimensions contribute to the correct harmony reconstruction.

Formally, we are interested in $\arg\min[d(h(f_D(S^*)), h(S))]$ where d is a distance function between two MSAs and D is a set of musical dimensions that can be incorporated in a function f. To achieve this we first need to have a reference set of correctly aligned polyphonic pieces, the S component (see Sect. 4.1). Secondly, we need an aligner of multiple sequences that can incorporate more than one dimensions, the f function (Sect. 4.2). We also need to establish a meaningful distance measure so that we can compare the polyphonic ground truth to the reconstruction created by the multidimensional MSA, the d function (Sect. 4.3). Finally, after we discuss the musical dimensions we consider in our work, the D component (Sect. 4.4), we put them to the test (Sect. 4.5) (Fig. 2).

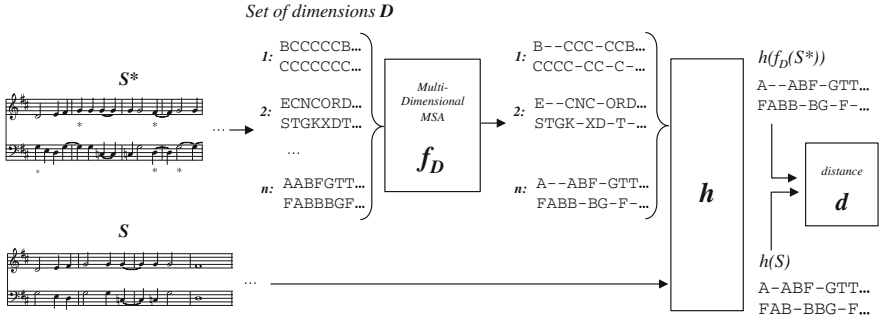

Fig. 2. The pipeline and components (S, S^*, D, f, h, d) for discovering the importance of musical dimensions in harmony reconstruction.

4.1 Dataset

Our dataset, called Hymns[2], comprises of 153, sixteenth century 4-voice religious pieces. We picked this dataset due to its particular properties. First, all

[2] www.genevanpsalter.com/music-a-lyrics/2-complete-collections/181-midi-collections.

pieces are transposed to the same key which allows us to learn a global harmony model. Secondly, the average length of the pieces is small enough to run different experiments with a reasonable time complexity. Finally, the number of notes over the length of each voice is almost the same, which avoids alignment ambiguities. For example, the sequence `ABC` can be aligned to both the beginning or end of `ABCXXXABC`, so any algorithm would have trouble distinguishing which alignment is better. This is an undesirable property in the context of this experiment. For more information regarding this problem, the reader is referred to [12].

4.2 Multidimensional Multiple Sequence Alignment (MDMSA)

We are interested in a function f that can incorporate more than one dimension to align the multiple voices of a polyphonic score. We use a multidimensional extension of the popular progressive alignment (PA) approach which was originally used for aligning multiple unidimensional sequences [9]. We selected PA due to its simplicity, both in terms of concept and development, and ease of adaptation to more than one dimension. In this section, we first present the concepts of unidimensional MSA and PA before their multidimensional extensions.

The unidimensional multiple sequence alignment is the output of a process that introduces gaps "-" to sequences of symbols so that they have the same length. Formally, given k sequences $s_1, s_2, ..., s_k$ over an alphabet $\mathcal{A}$, a gap symbol "-" $\notin \mathcal{A}$ and let $g : (\{-\} \cup \mathcal{A})^* \rightarrow \mathcal{A}^*$ a mapping that removes all gaps from a sequence containing gaps. A multiple sequence alignment A consists of k sequences $s'_1, s'_2, ..., s'_k$ over $\{-\} \cup \mathcal{A}$ such that $g(s'_i) = s_i$ for all i, $(s'_{1,p}, s'_{2,p}, .., s'_{k,p}) \neq (-, ..., -)$ for all p; and $|s'_i|$ is the same for all i.

There is a great number of possible MSAs for a single input of sequences [8]. We typically want to pick the most "meaningful" considering our task at hand. More formally: given a scoring function $c : A \rightarrow \mathbb{R}$ that maps each alignment to a real number, we are interested in $A' = \arg\max(c(A))$. The most widely used such function is the weighted sum-of-pairs (WSOP) [24]:

$$c(A) = \sum_{p=1}^{L} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} w_{i,j} v(s_{i,p}, s_{j,p}) \tag{1}$$

where L is the length of the MSA, $w_{i,j}$ is a weight of the pair of sequences i, j and $v(a, b)$ is a "relatedness" score between two symbols $a, b \in \{-\} \cup \mathcal{A}$. The scores are typically stored in a matrix format called the substitution matrix. Literature suggests that A' would be "meaningful" as long as the substitution matrix captures "meaningful" relationships between symbols [8]. WSOP can also be extended to take into consideration affine gap scores (different scores for gap insertions and gap extensions).

The exact computation of A' is NP-hard [27], so it cannot be used in practice. Instead, the focus is on heuristic approaches that give good alignments not guaranteed to be optimal. The most popular approach is progressive alignment (PA) [11], which comprises three fundamental steps. At first, all pairwise alignments

between sequences are computed to determine the WSOP similarity between each pair. In the second step, a similarity tree (guide tree) is constructed using a hierarchical clustering method. Finally, working from the leaves of the tree to the root, one aligns alignments, until reaching the root of the tree, where a single MSA is built. The drawback of PA is that incorrect gaps are retained throughout the process from the moment they are first inserted.

The unidimenional multiple sequence alignment can be extended to accommodate for multiple MSAs that we call "dimensions". More formally: a multidimensional multiple sequence alignment (MDMSA) consists of N multiple alignments $A_1, A_2, ..., A_N$. Each A_n consists of k sequences $s_1^n, s_2^n, ..., s_k^n$ over an alphabet $\{-\} \cup \mathcal{A}^n$ such that $|g(s_m^n)|$ is the same for all n and if $s_{m,p}^z$ is a gap at a dimension z, then $s_{m,p}^n$ is also a gap for all n. Figure 3 presents examples of simple MSA and MDMSA. In the same manner, WSOP can be extended to MDMSAs by summing over all dimensions:

$$c(A) = \sum_{n=1}^{N} W_n \sum_{p=1}^{L} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} w_{i,j} v_n(s_{i,p}^n, s_{j,p}^n) \tag{2}$$

where W_n and v_n are the weight and scoring function of the n^{th} dimension respectively. Extending the progressive alignment algorithm to accommodate MDMSAs is similarly straightforward; we extend pairwise alignment to multiple dimensions (through the multidimensional WSOP score).

Multiple Sequence Alignment

Input:	Output:
s_1: ABCABC	s'_1: ABCABC
s_2: ACC	s'_2: A-C--C
s_3: BCAC	s'_3: -BCA-C

Multidimensional Multiple Sequence Alignment

Input:	Output:
$s^1{}_1$: ABCABC	$s^1{}_1$: ABCABC
$s^1{}_2$: ACC	$s^1{}_2$: AC--C-
$s^1{}_3$: BCAC	$s^1{}_3$: BC-AC-
$s^2{}_1$: ZGGZGZ	$s^2{}_1$: ZGGZGZ
$s^2{}_2$: GGG	$s^2{}_2$: GG--G-
$s^2{}_3$: ZZZG	$s^2{}_3$: ZZ-ZG-

Fig. 3. Examples of: a multiple sequence alignment of three sequences (left), a multidimensional multiple sequence alignment of three sequences and two dimensions (right). Note that gaps in both dimensions are at the same positions.

4.3 Distance Between Two Multiple Sequence Alignments

Although the polyphonic scores S and S^* are multidimensional, both $h(f_D(S^*))$ and $h(S)$ are unidimensional MSAs reduced to only the pitch dimension. Therefore, we need to define a meaningful distance measure between two MSAs of the same sequences; the smaller the distance the more similar two alignments of the same voices should sound. The previous definition implies that any distance

measure we devise should correlate with the perceived distance between two harmony alignments A and B. For the sake of convenience, we make the following intuitive assumption: the larger the portion of the voices that are misaligned, the higher the perceived distance; any misalignment of voices leads to a bad sounding polyphony, despite the fact that it might sound "nice" by pure chance.

Based on this definition we generate a synthetic set of corrupted Hymn polyphonic pieces: each note $n_i = (p_i, d_i, o_i)$ in a voice s_j is modified (doubled or halved in duration) with a probability for modification $P = (l^2 \times o_i)/|s_j|$, where l the misalignment degree. Every note with onset value larger than o_i is consequently shifted resulting to misalignment. Notes at the end of the piece (larger o_i) have higher chance to be modified, since altering initial notes would result to larger misaligned portions. We generate misalignments at different degrees $l = 0.1, 0.2, ..., 0.8$. Figure 4 presents an example of two voices from the score of Fig. 1 corrupted at two different degrees l.

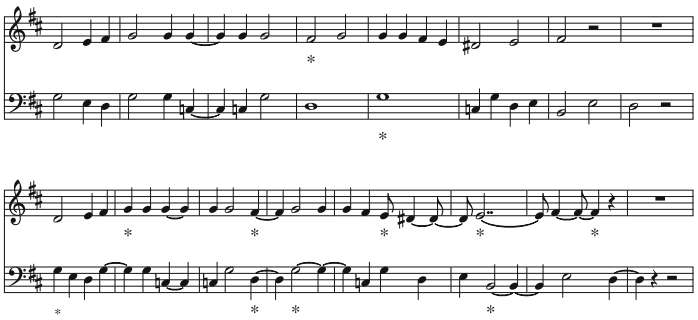

Fig. 4. Two (out of four) voices from the polyphonic score in Fig. 1 corrupted at different degrees of misalignment (top, bottom). Star signs "*" represent which notes were modified in terms of duration (halved or doubled) to generate the misalignments.

Now we identify a good measure for two MSAs. A widely used distance measure is based on the sum-of-pairs similarity, recoded as a dissimilarity: d_{SP} represents the ratio of aligned symbols in A that could not be found in B over $|A|$. Blackburne and Whelan [1] argue that d_{SP} is not a real metric because it violates the core principles of symmetry and triangle inequality. They presented four alternatives that differ in the way they treat gaps: *(a)* the "Symmetrized SP", or d_{SSP} which aims to be a correction of the d_{SP} score by ignoring all gaps *(b)* d_{seq} which incorporates raw gap information meaning that each gap is simply recoded as G_i, indicating it occurred in sequence i, *(c)* d_{pos} which in addition includes the position where gaps occur in a sequence, and *(d)* a metric that incorporates information from the phylogenetic tree of the MSA; omitted in our work since phylogeny information for our dataset is absent.

The relationship between the misalignment degree l and a ideal distance measure should be monotonic and linear; since as stated before, any distance measure, should correlate with the perceived distance between two polyphonic alignments. We measure these by calculating the Spearman and Pearson correlation coefficients respectively between l and the MSA pairwise distance. We also compute the coefficient of determination R^2. Table 1 presents those figures for all four distance measures considered in our work (all significance p values are smaller than 10^{70} and are omitted). The simple d_{SP} shows the highest correlation to the misalignment degree l, therefore it will be used from now on whenever we refer to a distance between two MSAs.

Table 1. The Spearman coefficient, the Pearson coefficient and the coefficient of determination R^2 for the d_{SP}, d_{SSP}, d_{pos} and d_{seq} measures.

	Spearman	Pearson	R^2
d_{SP}	0.826	0.801	0.641
d_{SSP}	0.800	0.765	0.586
d_{pos}	0.825	0.799	0.639
d_{seq}	0.817	0.796	0.633

4.4 Dimensions and Sequence Representation

We now explain the different dimensions of a polyphonic score that we consider in our work, and how they were represented as sequences, which is a prerequisite of the multidimensional PA algorithm.

Pitch. Pitch information is probably the most important dimension when it comes to harmony reconstruction. The representation of pitches into sequences is achieved by folding the pitch values into one octave and mapping them into an 12-sized alphabet of symbols.

Duration. We have also hypothesized that notes of similar duration might have higher chance to be sounded together in a polyphonic piece. It is also interesting to investigate to which extent duration corruption affects harmony reconstruction. We represent the duration dimension as a sequence by assigning an alphabetic symbol to each note value (e.g. thirty-second to "A", sixteenth to "B", eighth to "C" and so on).

Segment Boundaries. In musicology, "meaningful" units of notes are referred as "phrases", "segments", "sections" and so on, although the distinctions between them are vague. Music psychologists consider segmentation a fundamental listening function in terms of how humans perceive and structure music [16]. As such, information regarding segments has been frequently employed in MIR applications [21]. Given the reasonable assumption that humans generate a segment structure mentally as they listen to music, we hypothesize that the

segment beginnings (or ends) have higher chance to be aligned between different polyphonic voices, i.e. segments boundaries are more likely to sound in parallel.

Our dataset does not include segment boundary information, therefore we use three automatic segmentation algorithms (applied on each voice seperately): $seg_{gestalt}$ by Tenney and Polansky [23] which is based on Gestalt principles, seg_{markov} which is based on Markov probabilities of segment boundaries derived from the Essen collection [15] and seg_{LBD} which is based on the Local Boundary Detection Model by Cambouropoulos [4]. A number of segmentation algorithms exist beyond the ones considered [21], however those three should be sufficient for our task: understanding the importance of the segmentation dimension in harmony reconstruction. All three are based on onset information so any score corruption might have a major effect on their output. We represent the segmentation dimension of each voice as a sequence by binning the space of values into 26 bins so that each note is assigned an alphabetic character corresponding to its bin index. For example, if the segmentation output $\in [0, 1]$ for a melody consisting of ten notes is $1.0, 0.2, 0.0, 0.0, 0.9, 0.2, 0.2, 0.5, 0.2, 1.0$, then the sequence representation would be `ZFAAWFFMFZ`.

Metric Weights. The importance of a note in the temporal domain can be represented by its metric weight (not to be confused with distance metrics). We use the Inner Metric Analysis (IMA) [26] to compute the metric weights based on the note onsets. Two different variations of the IMA algorithm are computed: $IMA_{spectral}$ and $IMA_{metrical}$. We represent both IMA dimensions into sequences by binning the space of values into 26 bins. Each note is assigned an alphabetic character corresponding to its bin index.

Figure 5 presents an example of four MSAs corresponding to different dimensions of a polyphonic score.

4.5 Experiment

Settings. We aim to find which dimension(s) are most important for the reconstruction of the harmony of a polyphonic piece S after it has been corrupted to S^*, i.e. the set D that minimizes $d(h(f_D(S^*)), h(S))$. Besides the distance of the reconstruction to the ground truth, we are also interested in the difference in distance before and after reconstruction $\delta = d(h(f_D(S^*)), h(S)) - d(h(S^*), h(S))$. We perform the experiment on the Hymns dataset corrupted with misalignments at different degrees l (see Sect. 4.3).

Regarding the substitution matrix v_n for each dimension, we express the probabilities of symbols appearing in pairs in the so called log-odds scores. This means the substitution matrix for each dimension is learned from the ground truth in a similar manner as in [3,7,10]. Particularly for the pitch dimension, the substitution matrix can be considered a rough *harmony model*, since it encodes which pairs of pitch values are frequently sounded together. All substitution matrices are normalised to have zero mean and unit variance. All dimensions are assigned equal weights W_i for the sake of simplicity, although different weight settings may result to differences in performance. All pairs of sequences are

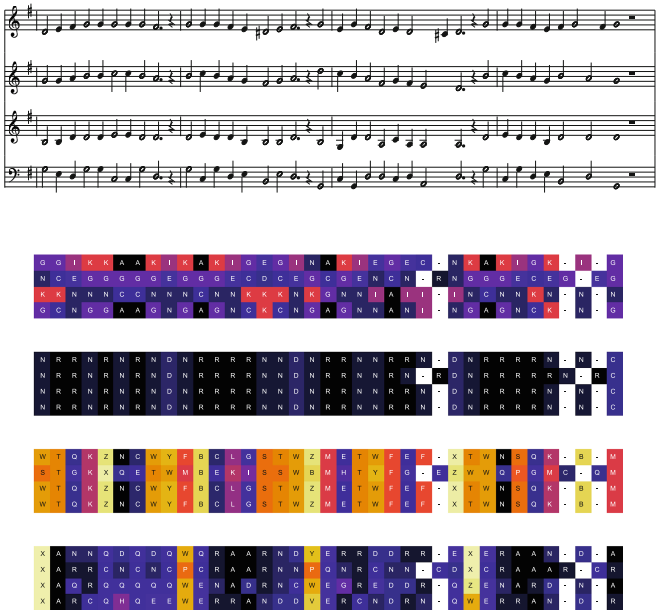

Fig. 5. Four properly aligned MSAs corresponding to the following dimensions of a polyphonic score (top): pitch, durations, $IMA_{spectral}$ and seg_{LBD}. Colors are used for visualization purposes. (Color figure online)

assigned equal weights $w_{i,j}$. Gap open and gap extend scores are set to -0.8 and -0.2 respectively. Although gap settings have great effect on alignments in general [6], preliminary results have showed that the core findings of our experiments are not affected.

In theory, given seven dimensions, we need to investigate the performance of $2^7 = 128$ different D set combinations. In practice, knowing that the pitch dimension is essential we can reduce that number to $2^6 = 64$, which is still impractical considering the time complexity of the PA. We therefore decided to combine dimensions empirically, starting from fewer dimensionalities to more.

Results. We start by combining the pitch dimension with any of the remaining six, i.e. $D = \{\text{pitches},y\}\ \forall y \in \{$ durations, $seg_{gestalt}$, seg_{markov}, seg_{LBD}, $IMA_{spectral}, IMA_{metrical}\}$. Figure 6 presents the $d(h(f_D(S^*)), h(S))$ and δ values achieved at different misalignment degrees. Considering the $d(h(f_D(S^*)), h(S))$ figure, three observations become immediately obvious: First, any addition to the pitch dimension makes the reconstruction more accurate. Second, duration is the dimension contributing the most. Third, all dimensions' positive contribution is weakened as the misalignment degree increases. It seems that a harmony model (pitch dimension) by itself is not sufficient to reconstruct the original harmony but the incorporation of more dimensions leads to a better reconstruction in comparison. However, all dimensions besides pitch are onset, duration-based and their reliability weakens as the degree of misalignment increases.

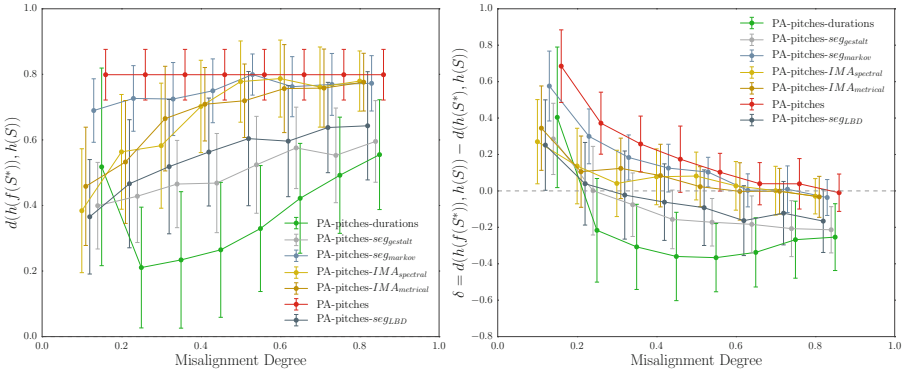

Fig. 6. The results for two dimensions, $|D| = 2$. Left: the distance of the reconstructed alignment (y axis), after corrupted at different misalignment degrees (x axis) to the ground truth. Right: the difference in distance (δ) between the reconstructed and the corrupted version to the ground truth. δ values above 0 mean that the method results in a worse harmony reconstruction compared to the input corrupted score. The results for the unidimensional MSA using only pitch information (PA-pitches) are also plotted as a baseline.

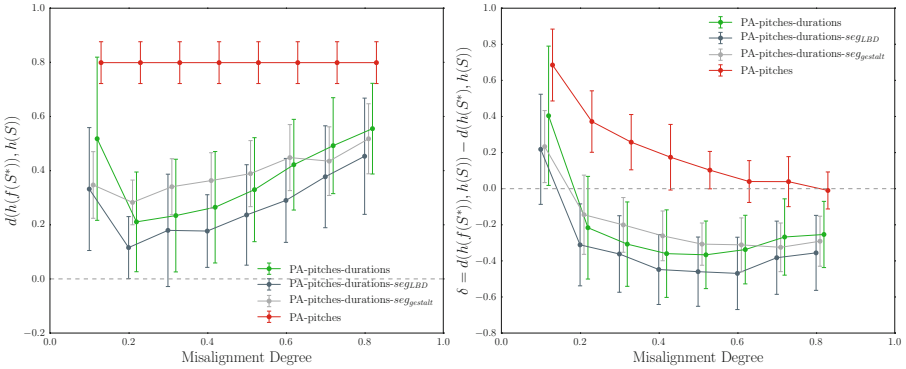

Fig. 7. The results for three dimensions, $|D| = 3$. The results for the one-dimensional MSA using only pitch information (PA-pitches) and the two dimensional PA using pitch and duration (PA-pitches-duration) are also plotted as baselines.

Considering the δ figure (right part of Fig. 6), it become obvious that only the duration, $seg_{gestalt}$ and seg_{LBD} dimensions result in an improvement on the reconstruction (any δ value above 0 means we make the alignment worse). Also the improvement happens only when the misalignment degree is above 0.2. In other words, even though the fusion of the pitch with any other dimension results in a better harmony reconstruction than the pitch dimension solely, only a couple of dimension combinations actually have a positive effect. Also, given an almost perfect polyphony we are more likely to make it worse than fix it.

In general, the results show that the pitch dimension by itself is inadequate for harmony reconstruction as is the incorporation of only one extra dimension. However, there is a strong indication that the incorporation of more dimensions will lead to a better reconstruction. Based on the previous findings we proceed into combining three dimensions: $D = \{\text{pitches,durations,}y\}$ $\forall y \in \{ seg_{gestalt}, seg_{LBD}\}$. Results are presented in Fig. 7. We can make the following observations. First, the pitches-durations-seg_{LBD} performs better than just the pitches-durations-$seg_{gestalt}$ and the pitches-durations approaches. Second, similar to the previous experiment, given a near-perfect initial score (misalignment degree less than 0.2), any combination of dimensions will result to a worse reconstruction.

5 Harmony Reconstruction Using MAFFT

The previous experiments revealed that the duration and segmentation dimensions (beside pitches) are the most important for harmony reconstruction: notes of similar duration and notes at segment boundaries are more likely to be sounded together in a polyphonic piece. Unfortunately, as soon as the note durations are altered drastically (highly corrupted), both dimensions become unreliable to be used for harmony reconstruction via multidimensional MSA. Segmentation particularly, degrades since the algorithms used in our work rely on musical heuristics applied on onset and duration information. It becomes clear that harmony reconstruction requires a segmentation technique that is impervious to duration-onset errors and according to our knowledge, such an algorithm for single voices, does not exist. It is also clear that the only reliable information for our task is pitches. Consequently, the question becomes whether useful structural information can be extracted from multiple pitch sequences corresponding to the different voices of a polyphonic piece.

Interestingly, locating very similar sub-regions (segments) between large sequences has been in important task in bioinformatics. Such segments can efficiently reduce MSA runtimes and as a consequence, MSA solutions that incorporate segmentation, such as DIALIGN [18] and MAFTT [14], have found successful application. MAFFT in particular, is a unidimensional progressive alignment method at its core, but uses the fast Fourier transform to identify short subregions that are high-scoring matches between the sequences in the alignment.

We hypothesize that MAFFT's pipeline can be a viable solution to the harmony reconstruction problem. Therefore, we apply MAFFT solely on the pitch dimension and we incorporate a harmony model similarly to the PA approaches, i.e. a learned log-odds substitution matrix from the pitch dimension of the ground truth. Figure 8 presents the results for MAFFT compared to the best performing PA-pitches-durations-seg_{LBD} approach from the previous experiment. For the sake of completeness we also include a five-dimensional approach PA-pitches-durations-seg_{LBD}-$seg_{gestalt}$, although we know in advance its performance will not be robust to high misalignments degrees. For this dataset, the results show that MAFFT achieves almost perfect harmony reconstruction and performs better than any multidimensional PA approach. In addition, given an

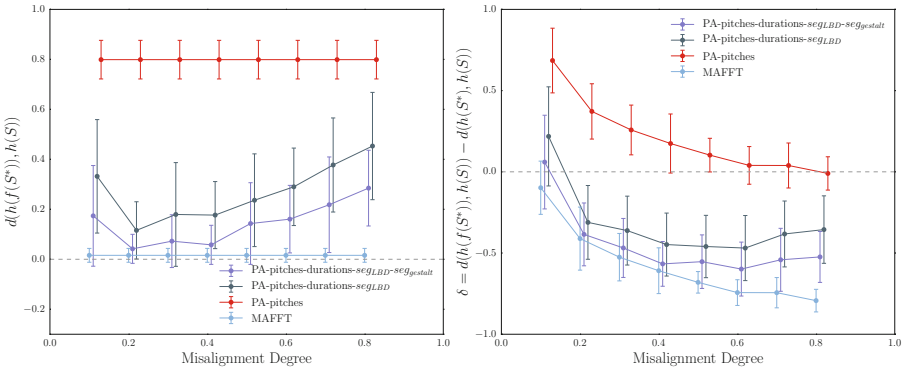

Fig. 8. The results for various dimensionalities and MAFFT.

almost perfect polyphonic score, MAFFT is more likely to fix it than worsen it. More importantly, MAFFT's performance is stable and invariant to any duration and onset noise. Consequently, given our particular task (harmony reconstruction), a harmony-aware MAFFT approach is the most reliable solution.

6 Conclusions

In this paper we introduced the problem of polyphony reconstruction and tackled its harmonic component: how to align pitches from different voices after they have been corrupted in terms of durations and onsets. By using a multidimensional version of MSA we showed that structural information, namely segment boundaries, are essential for the correct polyphony reconstruction. Since most segmentation algorithms are based on duration and onset information, we proposed the use of the bionformatics MSA aligner MAFFT extended with a harmony model. We additionally showed its superiority and perfect fit for the task.

However, polyphony reconstruction is far from being considered solved, while we cannot claim that we now understand the cognitive process behind aligning voices. Besides excluding the crucial rhythm component, our work made a set of simplifications; first, we employed progressive alignment which is heuristic rather than an exact MSA algorithm. Secondly, each musical dimension was considered equally important, although the literature contradicts this. Thirdly, we have not yet investigated how pitch errors could be dealt with. And finally, the repertoire we have chosen is rhythmically simple in comparison to most polyphony from the 16^{th} century. Despite those facts, our work set strong foundations for understanding polyphony and towards a complete solution to the polyphony reconstruction problem.

Aknowledgments. The authors would like to thank Meinard Müller and Hendrik Vincent Koops for comments that greatly improved the manuscript.

References

1. Blackburne, B.P., Whelan, S.: Measuring the distance between multiple sequence alignments. Bioinformatics **28**(4), 495–502 (2012)
2. Boulanger-Lewandowski, N., Bengio, Y., Vincent, P.: Modeling temporal dependencies in high-dimensional sequences: application to polyphonic music generation and transcription. arXiv preprint arXiv:1206.6392 (2012)
3. Bountouridis, D., Koops, H.V., Wiering, F., Veltkamp, R.C.: A data-driven approach to chord similarity and chord mutability. In: International Conference on Multimedia Big Data, pp. 275–278 (2016)
4. Cambouropoulos, E.: The local boundary detection model (LBDM) and its application in the study of expressive timing. In: International Computer Music Conference, pp. 17–22 (2001)
5. Carrillo, H., Lipman, D.: The multiple sequence alignment problem in biology. SIAM J. Appl. Math. **48**(5), 1073–1082 (1988)
6. Carroll, H., Clement, M.J., Ridge, P., Snell, Q.O.: Effects of gap open and gap extension penalties. In: Biotechnology and Bioinformatics Symposium, pp. 19–23 (2006)
7. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C.: 22 a model of evolutionary change in proteins. In: Atlas of Protein Sequence and Structure, vol. 5, pp. 345–352. National Biomedical Research Foundation Silver Spring, MD (1978)
8. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge (1998)
9. Feng, D.-F., Doolittle, R.F.: Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. J. Mol. Evol. **25**(4), 351–360 (1987)
10. Henikoff, S., Henikoff, J.G.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. **89**(22), 10915–10919 (1992)
11. Hogeweg, P., Hesper, B.: The alignment of sets of sequences and the construction of phyletic trees: an integrated method. J. Mol. Evol. **20**(2), 175–186 (1984)
12. Hudek, A.K.: Improvements in the accuracy of pairwise genomic alignment (2010)
13. Joh, C.-H., Arentze, T., Hofman, F., Timmermans, H.: Activity pattern similarity: a multidimensional sequence alignment method. Transp. Res. Part B Methodol. **36**(5), 385–403 (2002)
14. Katoh, K., Misawa, K., Kuma, K., Miyata, T.: Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res. **30**(14), 3059–3066 (2002)
15. Lartillot, O., Toiviainen, P., Eerola, T.: A matlab toolbox for music information retrieval. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) Data Analysis, Machine Learning and Applications, pp. 261–268. Springer, Heidelberg (2008)
16. Lerdahl, F., Jackendoff, R.: An overview of hierarchical structure in music. Music Percept. Interdisc. J. **1**(2), 229–252 (1983)
17. Lyu, Q., Wu, Z., Zhu, J., Meng, H.: Modelling high-dimensional sequences with LSTM-RTRBM: application to polyphonic music generation. In: International Conference on Artificial Intelligence, pp. 4138–4139. AAAI Press (2015)
18. Morgenstern, B., Dress, A., Werner, T.: Multiple dna and protein sequence alignment based on segment-to-segment comparison. Proc. Natl. Acad. Sci. **93**(22), 12098–12103 (1996)

19. Pugin, L., Crawford, T.: Evaluating omr on the early music online collection. In: International Society on Music, Information Retrieval, pp. 439–444 (2013)
20. Rebelo, A., Fujinaga, I., Paszkiewicz, F., Marcal, A.R.S., Guedes, C., Cardoso, J.S.: Optical music recognition: state-of-the-art and open issues. Int. J. Multimedia Inf. Retrieval **1**(3), 173–190 (2012)
21. Rodríguez López, M.E.: Automatic Melody Segmentation. Ph.D. thesis, Utrecht University (2016)
22. Sanguansat, P.: Multiple multidimensional sequence alignment using generalized dynamic time warping. WSEAS Trans. Math. **11**(8), 668–678 (2012)
23. Tenney, J., Polansky, L.: Temporal gestalt perception in music. J. Music Theory **24**(2), 205–241 (1980)
24. Thompson, J.D., Higgins, D.G., Gibson, T.J.: Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**(22), 4673–4680 (1994)
25. van Kranenburg, P.: A computational approach to content-based retrieval of folk song melodies. Ph.D. thesis (2010)
26. Volk, A., Garbers, J., Van Kranenburg, P., Wiering, F., Veltkamp, R.C., Grijp, L.P.: Applying rhythmic similarity based on inner metric analysis to folksong research. In: International Society on Music Information Retrieval, pp. 293–296 (2007)
27. Wang, L., Jiang, T.: On the complexity of multiple sequence alignment. J. Comput. Biol. **1**(4), 337–348 (1994)
28. Wang, S., Ewert, S., Dixon, S.: Robust joint alignment of multiple versions of a piece of music, pp. 83–88 (2014)