ALIGNING UNSYNCHRONIZED PART RECORDINGS TO A FULL MIX USING ITERATIVE SUBTRACTIVE ALIGNMENT

Daniel Yang **Kevin Ji** **TJ Tsai**
Harvey Mudd College
dhyang,kji,ttsai@hmc.edu

ABSTRACT

This paper explores an application that would enable a group of musicians in quarantine to produce a performance of a chamber work by recording each part in isolation in a completely unsynchronized manner, and then generating a synchronized performance by aligning, time scale modifying, and mixing the individual part recordings. We focus on the main technical challenge of aligning the individual part recordings against a reference "full mix" recording containing a performance of the work. We propose an iterative subtractive alignment approach, in which each part recording is aligned against the full mix recording and then subtracted from it. We also explore different feature representations and cost metrics to handle the asymmetrical nature of the part–full mix comparison. We evaluate our proposed approach on two different datasets: one that is a modification of the URMP dataset that presents an idealized setting, and another that contains a small set of piano trio data collected from musicians during the pandemic specifically for this study. Compared to a standard pairwise alignment approach, we find that the proposed approach has strong performance on the URMP dataset and mixed success on the more realistic piano trio data.

1. INTRODUCTION

This paper explores an application that would enable a group of musicians in quarantine to produce a performance of a piece of chamber music through asynchronous musical collaboration. Asynchronous musical collaboration is usually done with musicians performing their parts synchronized to a reference "click" track. This paradigm works well for many genres of music where the tempo is relatively constant (e.g. pop music) or the musicians are expected to follow a conductor (e.g. choral music). However, this paradigm does not work well with genres of music where musicians are constantly adapting to and influencing one another. As a representative example of the latter, we focus in this paper on the genre of piano trio music, which is ill-suited to a click track paradigm for several reasons: the tempo is constantly changing and may vary widely across a single movement, the main melodic line is carried by different instruments at different times and may be shared by two or more instruments, parts may contain extended periods of silence, and each instrument is given considerable latitude for individual musical expression. Our goal is to allow each musician the freedom to perform their part as they wish, while still allowing for asynchronous musical collaboration.

Our approach to this problem is to allow the recording of individual parts to be unsynchronized, and to use MIR tools to achieve synchronization post-recording. Figure 1 shows a high-level overview of this paradigm for a piano trio. The primary inputs to the system are three recordings: a recording of the piano part only, a recording of the cello part only, and a recording of the violin part only. These will be referred to as "part" recordings, since they only contain the performance of a single part. The first step (bottommost block) is to determine the alignment between the part recordings. Because the part recordings are not directly comparable to one another (i.e. they may be playing different notes), we can provide a reference "full mix" recording (e.g. by finding a YouTube video of the piece) that contains all parts played in synchrony, and then use the full mix as additional information to assist our estimate of the joint alignment among the three part recordings. Once we have estimated the alignment among the part recordings, we can then use time scale modification (TSM) to adjust the tempos in each part to produce time scale modified, synchronized part recordings. [1] These synchronized part recordings can be mixed together to produce the final performance. TSM is a well-studied problem [1], and there are effective approaches based on phase vocoding and various overlap-add methods (e.g. [2–4]). The main technical challenge in Figure 1, therefore, is the joint alignment problem among the part recordings and the full mix. We will focus on this technical problem in the remainder of this paper.

Alignment tasks have long been a topic of interest to the MIR community due to their applications in score following, retrieval, and synchronization of various forms of music data. An exhaustive survey of alignment research is beyond the scope of this paper, but here we simply point

[1] Note that the reference recording is only used to assist in the joint alignment estimation problem. Once we have estimated the joint alignment, we can time scale modify the part recordings however we wish (e.g. modifying two part recordings to match the third part recording).

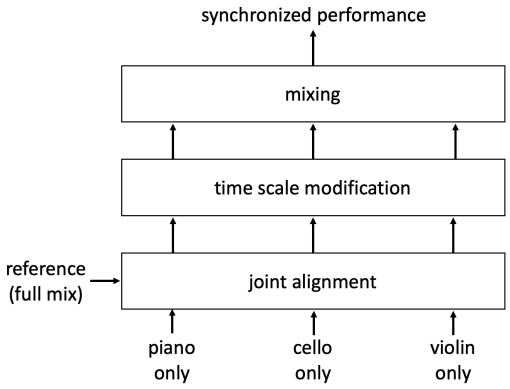

Figure 1. Overview of asynchronous musical collaboration with unsynchronized part recordings. Each musician records their part in isolation in a completely unsynchronized manner, and the part recordings are modified and mixed to produce a synchronized performance.

out current trends in the MIR alignment literature in order to situate our current work in proper context. Recent works on alignment tasks in the MIR literature tend to fall into one of three groups. The first group focuses on alignment across different modalities of music data. The main challenge here is to find a feature representation that enables a direct comparison of similarity between different modalities. Some recent examples of this include aligning lyrics and audio [5, 6], aligning sheet music images and audio [7–9], and aligning sheet music images and MIDI [10, 11]. The second group focuses on performing alignment under a different set of assumptions or conditions than traditional dynamic time warping (DTW). Some examples include handling discontinuities due to jumps or repeats [12–14], handling completely unconstrained jumps such as might be observed in a practice session [15, 16], or aligning two audio mixtures containing a non-disjoint subset of parts from a common piece of ensemble music [17]. The third group focuses on issues of scalability and efficiency of alignment techniques. DTW has quadratic cost in memory and computation, which limits its utility in dealing with long sequences. Previous works have proposed alignment approaches that operate at multiple scales [18, 19], and recent works explore ways to reduce the memory cost [20, 21] or total runtime through parallelization [22].

The alignment problem shown in Figure 1 falls into the second group above — it frames the alignment problem with a different set of assumptions and context. There are at least three significant differences between our proposed scenario and a typical audio–audio alignment scenario. First, we are aligning each part recording against a full mix containing all three parts, so the alignment must be estimated in the presence of other significant sound sources. If we assume that each part has equal volume and interpret the other parts as highly correlated additive noise, we are effectively estimating an alignment in the regime of $10 \log_{10} \frac{1}{2} = -3$ dB SNR. In a typical audio–audio alignment scenario, we might align two full mix recordings of the same piece, which corresponds to a high SNR regime. [2] Second, we are estimating the alignment between a *set* of part recordings and a full mix recording, rather than considering a single isolated pairwise alignment. Because we know that the full mix is a mixture of all three parts, the knowledge of one part–full mix alignment is relevant to our estimate of the other alignments. Third, the part recordings may be sparse — they may contain extended periods of silence where the instrument is not playing. In typical audio–audio alignment scenarios, both recordings are usually assumed to be "dense" recordings that contain musical information at all times. Because of the characteristics above, we will refer to the problem in Figure 1 as the part–full mix joint alignment problem.

Our approach to the part–full mix joint alignment problem has two distinct characteristics. First, we adopt an iterative subtractive approach, in which we align each part recording to the full mix, time scale modify the part recording to match the full mix, and then subtract the part recording from the full mix. Second, we explore different cost metrics that have the desired asymmetric behavior: we do not want to penalize the full mix for having spectral peaks that are not present in the part recording (since these peaks may come from the other parts), but we do want to reward the part recording for "explaining" spectral peaks that are observed in the full mix.

This paper has two main contributions. First, we propose and motivate a part–full mix joint alignment problem that would allow musicians to produce chamber music performances without any synchronization or communication, and we present two different datasets to enable its systematic study. One dataset is a modification of the URMP dataset [23], which contains ensemble works of various instrumentation. The other dataset is a small set of real world data that serves as a case study for our application of interest. It was collected during the pandemic specifically for this study and contains multiple performances of a single piano trio work. Second, we propose an iterative subtractive alignment approach, in which each individual part recording is aligned against a reference full mix recording and then subtracted from it. We explore several different cost metrics to account for the asymmetrical nature of the part–full mix comparison. We find that the proposed approach has strong performance on the URMP benchmark and mixed success on the more realistic piano trio data. We present experimental results on both datasets to provide more insight into the performance of our proposed approach. [3]

2. SYSTEM DESCRIPTION

Figure 2 shows an overview of our proposed iterative subtractive alignment approach. There are two key steps which are repeated multiple times: aligning a single part to the full mix and subtracting the part from the full mix.

[2] In this case, the primary distortion (apart from the time warping) comes from differences in the instrument, the performer's interpretation & articulation, and recording conditions. These distortions are also present in our scenario.

[3] Code and data can be found at `https://github.com/HMC-MIR/PianoTrioAlignment`

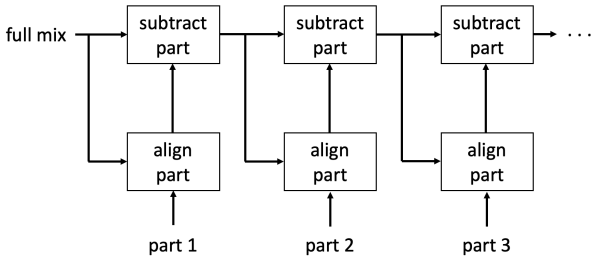

Figure 2. Overview of the subtractive alignment approach to solve the part–full mix joint alignment problem.

These two steps will be described in the next two subsections.

2.1 Aligning a Single Part

The first key step is to align a single part to the full mix. If we were aligning two full mix recordings, we could simply use standard chroma features combined with a cosine distance metric. In our case, however, the full mix contains a mixture of parts, only one of which matches the part recording. In computing a pairwise cost $C[i, j]$ between part frame i and full mix frame j, we do not want to penalize the full mix for containing energy at frequencies not found in frame i of the part recording. This motivates the need for a different feature representation and cost metric. We explore two different methods to account for this asymmetry.

The first method is based on a constant Q transform (CQT). We first compute a CQT on the full mix and part recording using 12 bins per octave between C1 to C10. Let $x_i \in \mathbb{R}^{109}$ be the CQT values for the i^{th} frame in the part recording and let $y_j \in \mathbb{R}^{109}$ be the CQT values for the j^{th} frame in the full mix. We compute the pairwise cost between x_i and y_j as

$$C[i, j] = -sum(min(x_i, y_j))/sum(x_i)$$

where the min operator computes the elementwise minimum between two vectors and the sum operator sums the elements in a vector to produce a single scalar. The numerator term $sum(min(x_i, y_j))$ is a single scalar that indicates how much of the CQT energy in y_j is "explained" by x_i. The denominator term $sum(x_i)$ normalizes this value by the total amount of energy in x_i.[4] Therefore, this cost metric rewards x_i for explaining the spectral peaks in y_j, but does not penalize y_j for having more energy than x_i in a frequency bin. This exhibits the type of asymmetrical behavior we desire in our scenario. Note that $C[i, j]$ will always be in the range $[-1, 0]$, where -1 indicates a strong agreement between x_i and y_j.

The second method is based on a binarized constant Q transform (BCQT). We binarize the part CQT and the full mix CQT by applying a hard threshold γ_k to the k^{th} CQT frequency bin. The threshold γ_k is determined by considering 6 bins above and below the k^{th} frequency bin, treating the resulting $13 \times N$ matrix as a grayscale image (where N is the number of frames in the CQT), and applying the triangle binarization algorithm to determine a threshold [24]. Let $x_i \in \{0, 1\}^{109}$ be the BCQT values for the i^{th} frame in the part recording and let $y_j \in \{0, 1\}^{109}$ be the BCQT values for the j^{th} frame in the full mix. We compute the pairwise cost between x_i and y_j as the negative normalized inner product $C[i, j] = -x_i^T y_j / sum(x_i)$ where the sum operator sums the elements of a vector to produce a scalar. Again, $C[i, j]$ will be in the range $[-1, 0]$, where -1 indicates a strong agreement between x_i and y_j. Note that the first method weights the importance of each frequency bin according to the amount of energy in it, whereas the binarized approach gives all frequency bins equal importance.

Once we compute the pairwise cost matrix, we use DTW to estimate the alignment between the part recording and the full mix. Because some parts may not be active at the beginning of the piece, we use subsequence DTW to estimate the alignment. Subsequence DTW is a variant of DTW that finds the best alignment between a short query sequence and any subsequence in a longer reference sequence. This allows the alignment path to begin and end anywhere in the full mix, rather than assuming that the full mix and part recording both begin and end at the same time. We allow for (part, full mix) transitions of $(1, 1)$, $(1, 2)$, and $(2, 1)$ with multiplicative transition weights of 1, 1, and 2, respectively.

At the end of the first key step, we have an estimated alignment between a single part and the full mix recording. This estimated alignment is passed to the subtraction block, which we describe in the next subsection.

2.2 Subtracting a Single Part

The second key step is to subtract the part recording from the full mix. This is done on the CQT representation through spectral subtraction. This process consists of four substeps, which are described in the next four paragraphs.

The first sub-step is to time warp the part recording to match the timing of the full mix. This can be done very easily by using the estimated alignment between the part recording and full mix. For example, if frame $y_k \in \mathbb{R}^{109}$ in the full mix CQT is aligned to frame $x_i \in \mathbb{R}^{109}$ in the part CQT, then frame $\tilde{x}_k$ in the *time-warped* part CQT will be $\tilde{x}_k = x_i$. When there are $(1, 2)$ transitions in the estimated alignment, we can estimate "missing" frames through interpolation. In order to handle frames that fall outside the estimated alignment (e.g. the part recording begins matching the full mix 20 seconds into the performance), we simply pad additional frames with zeros at the beginning and end as needed. At the end of this first substep, we have a time-warped part CQT $\tilde{X}$ which has the same dimensions as the full mix CQT Y.

At this point, we *could* simply subtract the time-warped part CQT from the full mix CQT. However, this does not account for volume differences between the two recordings. For example, a violin part recording containing a single solo violin player may be much louder than the violin signal in the full mix recording. These differences may

[4] We do not square the elements in order to avoid too heavily penalizing large differences.

be global differences due to microphone recording levels or local differences due to musical interpretation (e.g. the recording levels are the same, but one violinist prefers to play a particular section more softly than the other violinist). In order to account for these volume differences, we break the part recording into segments and estimate a volume gain factor for each segment.

The second substep, then, is to break the part recording into segments. We do this by performing silence detection on the part recording, and then consider contiguous regions of silence or non-silence as segments. Because the part recording only contains a single instrument, the silence detection is relatively straightforward, and we use a simple energy-based approach. We first compute the amount of energy in windows of length 0.75 seconds across the entire recording. We then model the distribution of log energy (within a single window) with a Gaussian mixture model with 3 mixtures. We interpret the Gaussian with the smallest mean as a model for silence in the recording. We compute the probability that a frame is silence $P(silence|log\ energy)$ using Bayes' rule and threshold at 0.5. We find this energy-based silence detection approach to be sufficiently robust for our application. Because the piano part is playing throughout the entire piece, we use a very simple scheme for segmenting the piano recording: we simply break it up into non-overlapping 5 second segments.

The third substep is to estimate a volume gain factor for each (time-warped) segment. We estimate the optimal volume gain factor α^* in the following manner. Let $\tilde{X}_s \in \mathbb{R}^{109\times L}$ be the CQT representation of a single time-warped segment of length L frames, and let $Y_s \in \mathbb{R}^{109\times L}$ be the CQT representation for the corresponding section of the full mix. We calculate the optimal gain factor α^* as

$$\alpha^* = \arg\max_{\alpha} sum(min(\tilde{X}_s\alpha, Y_s) - max(\tilde{X}_s\alpha - Y_s, 0))$$

where the min and max operators are performed elementwise between two matrices and the sum operator sums all elements in a matrix to produce a single scalar. The $min(\tilde{X}_s\alpha, Y_{seg})$ term indicates how much of the CQT energy in the full mix segment is explained by the volume-scaled & time-warped part recording. The $max(\tilde{X}_s\alpha - Y_{seg}, 0)$ term is a penalty for overestimating the energy in the full mix CQT. To approximately solve this optimization problem, we simply consider a range of values for α between 0.1 and 100 and use the value that maximizes the objective function. After this third substep, we have a volume gain factor α^* for every segment in the part recording.

The fourth substep is to perform the spectral subtraction. For each segment, we subtract the volume-scaled & time-warped part CQT from the full mix CQT as

$$Y_{s,mod} = max(Y_s - \tilde{X}_s\alpha^*, 0)$$

where the max operator is performed elementwise. The max operator ensures that the modified full mix CQT remains a non-negative matrix. The output of the subtraction block in Figure 2 is the modified full mix CQT with the part recording subtracted out.

Recording Type	# Recordings		Duration	
	Train	Test	Train	Test
Piano only	2	2	21.8m	22.4m
Violin only	2	2	17.5m	18.9m
Cello only	2	2	19.0m	18.9m
Full mix (YouTube)	2	2	19.1m	21.2m
Total	8	8	77.3m	81.3m

Table 1. Summary of the Mendelssohn piano trio data collected from musicians in quarantine. All possible combinations of the recordings are considered, resulting in 16 training episodes and 16 test episodes.

2.3 Aligning Multiple Parts

Once the first part has been aligned and subtracted out from the full mix CQT, we repeat the entire process with the next part recording. At each iteration of this process, we always use the most updated version of the full mix CQT with previous parts subtracted out. The final output of the system is an estimated alignment between each part recording and the full mix recording.

3. EXPERIMENTAL SETUP

Doing a rigorous empirical study of our proposed application of interest presents several significant challenges. As with constructing any alignment dataset, annotating beat timestamps is a very time-consuming task. But even beyond that, simply *collecting* suitable audio data for our application is an even more significant challenge. Because musicians don't have any incentive to record themselves playing a single part of a chamber work, such data is not readily available in the wild. Because of the challenges of getting realistic data, we opted for a two-pronged approach: we collected a small amount of data that is specifically tailored to our application to serve as a case study, and we also modified an existing dataset to study the same alignment problem in an idealized setting.

The application-specific data was collected in the following manner. In the midst of the pandemic, we recruited three musicians to participate in our study: a cellist and a violinist in the Claremont Colleges Orchestra and a pianist who has studied privately in college. Based on the expressed preferences of all three musicians, they agreed to learn the first movement of the Mendelssohn piano trio no. 1 in D minor. The musicians were asked to learn their parts and then record themselves playing their part in isolation. During the period when the musicians were learning their parts, they had no joint practices together and did not have any communication about their interpretation of the piece (e.g. at what tempo to play the piece). Each musician used a cell phone to record themselves playing their part four times from beginning to end, including counting out measures of rest. In addition to the individual part recordings, we also found four different performances of the Mendelssohn piano trio on YouTube. These YouTube performances serve as the reference full mix recording shown

Piece Type	# Pieces		Duration	
	Train	Test	Original	Modified
Duos	6	5	20.8m	215.0m
Trios	6	6	17.7m	181.8m
Quartets	6	8	27.5m	285.8m
Quintets	2	5	14.2m	149.0m
Total	20	24	80.2m	831.6m

Table 2. Summary of the URMP data used as an additional evaluation benchmark. The full mix recordings were randomly modified in tempo to generate 10 episodes per piece, resulting in 200 training episodes and 240 test episodes.

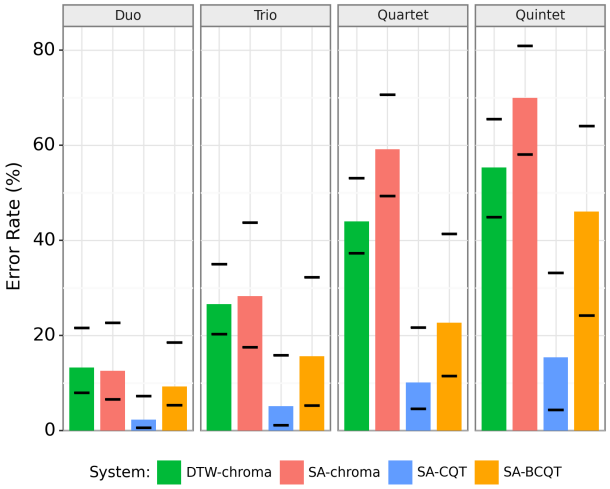

Figure 3. Results on the URMP dataset. Colored bars indicate the error rate at an error tolerance of 200 ms, and the horizontal bars above and below each colored bar indicate the error rate at error tolerances of 100 ms and 400 ms, respectively. Results are separated by piece type.

at left in Figure 2. Put together, the piano trio data contains a total of 16 recordings and 159 minutes of data.

After this data had been collected, we asked each musician to annotate the downbeats in all 16 audio recordings using SonicVisualizer. Because there are many extended periods of silence in the cello and violin part recordings, the musicians were asked to selectively annotate only those downbeats that they felt could be reasonably inferred. We collected all $16 \times 3 = 48$ annotation files ($23,274$ total beat annotations) and merged them by taking the average of annotated timestamps for all downbeats containing three annotations. Downbeats with less than three annotations were discarded. The result of this merging process is a set of 16 ground truth annotation files for the 16 audio recordings.

In summary, the application-specific data consists of 16 audio recordings containing 4 violin part recordings, 4 cello part recordings, 4 piano part recordings, and 4 full mix recordings. All 16 recordings are unsynchronized, and each recording has ground truth annotations of downbeats. We will refer to this dataset as the piano trio data. Table 1 shows an overview of this dataset. As discussed above, it has the benefit of being real world data that is collected in the exact application scenario of interest, but has the drawback of being very limited in diversity (only one piece and one set of performers). The results on this dataset should therefore be seen as a case study.

We evaluated system performance in the following manner. We set apart two violin part recordings, two cello part recordings, two piano part recordings, and two full mix recordings for training, and we consider all possible combinations of training recordings, resulting in $2 \times 2 \times 2 \times 2 = 16$ different training episodes. For test evaluation, we likewise consider all possible combinations of test recordings, resulting in 16 test episodes. For each episode, we evaluate the accuracy of the predicted alignment between each part recording and the full mix recording. We calculate the percentage of annotated downbeats whose predicted alignment error is greater than a maximum allowable error tolerance, and we calculate this error rate for several different error tolerances. Because the alignment accuracy of different instruments varied widely, we report results for each instrument separately (i.e. piano–full mix, cello–full mix, and violin–full mix).

In parallel with the piano trio data, we also constructed a benchmark using the URMP dataset [23] to study the same alignment problem, albeit in an idealized setting. The URMP dataset contains recordings of 44 different ensemble works of various instrumentation and ranging from duos up to quintets. For each piece, the dataset contains recordings of each individual part played in isolation, as well as a full mix recording containing all parts mixed together. When recording the data, the musicians listened to a reference track on earphones, so all part recordings are synchronized. The full mix recording was generated synthetically by mixing the individual part recordings together with appropriate offset. The dataset contains a symbolic score for each piece, annotations of F0 trajectories, and timestamps for note onsets.

We modify the URMP dataset to study the part–full mix joint alignment problem. Because each part recording is already synchronized with the full mix recording, the alignment problem is trivial. To make the problem non-trivial, we modify the full mix recording in the following way. First, we split the full mix recording into three segments of equal length. For each segment, we use time scale modification to change the tempo by a constant, random factor while preserving the pitches of all parts. We use the phase vocoder implementation in [25] to perform the time scale modification. The tempo is uniformly sampled between $0.66 r_{nom}$ and $1.5 r_{nom}$ on a log scale, where r_{nom} indicates the nominal tempo of the full mix segment. Because this process is probabilistic, we generate 10 different modified full mix recordings for each piece in URMP, which provides us with 10 episodes per piece.

We evaluate performance on our modified URMP data in the same fashion as for the piano trio data. We set apart 20 pieces for training and 24 pieces for testing. Because

ground truth timestamps are provided for note onsets, we evaluate alignment accuracy at each note onset (rather than at downbeats). As before, we compute the error rate at several different error tolerances. The alignment accuracy varied widely based on the number of parts in the piece, so we report results for duos, trios, quartets, and quintets separately. Table 2 provides an overview of the URMP data used in our experiments.

4. RESULTS

Figure 3 shows the performance of four different systems on the URMP benchmark. The first system ('DTW-chroma') simply performs an independent pairwise alignment between each part recording and the full mix using chroma features and cosine distance metric. This is can be considered our baseline, since it is a default choice in many audio–audio alignment applications. The second and third systems use the subtractive alignment approach based on the CQT representation ('SA-CQT') and BCQT representation ('SA-BCQT'). We also include a fourth system that uses the subtractive alignment approach with standard chroma features and cosine distance metric ('SA-chroma') as a way to tease apart the effect of the feature representation and the iterative subtraction approach. The colored bars indicate the error rate at an error tolerance of 200 ms, and the horizontal bars above and below each colored bar indicate the error rate at error tolerances of 100 ms and 400 ms, respectively. The results are shown separately for duos, trios, quartets, and quintets.

There are three things to notice about Figure 3. First, the performance of all systems gets progressively worse as we move from duos to trios to quartets to quintets. This is to be expected, since the problem becomes progressively more challenging as more "noise" sources are added and the effective SNR becomes lower. Second, the SA-CQT approach has the best performance by a wide margin, far outperforming the baseline DTW-chroma system. For example, the DTW-chroma baseline has a 44.0% error rate with 200 ms tolerance on quartet pieces, while the SA-CQT approach achieves 10.1% error rate. Third, even with a subtractive approach, the SA-chroma and SA-BCQT approaches scale poorly with the number of instruments.

Figure 4 shows the results of the same four systems on the piano trio data. Results are shown separately for each instrument's alignments against the full mix. For the subtractive approaches, we use a piano–cello–violin ordering for spectral subtraction, which we found to work best on the training data. There are three things to notice about these results. First, the performance depends a lot on the instrument: the piano alignments are best by far, and the cello and violin alignments are much worse. This is perhaps not too surprising, since the piano part has a much more distinctive spectral profile due to being a polyphonic instrument. Second, the subtractive approach potentially provides a significant improvement for the cello alignment, but at best only marginal improvements for violin. Third, SA-CQT is no longer a clear winner as it was on the URMP data. Instead, we can see that different approaches seem to work best for different instruments: SA-CQT and DTW-chroma work best for piano, SA-chroma works best for cello, and SA-BCQT works best for violin. One area that might be interesting to explore in the future is using different feature representations and cost metrics for different part recordings in order to exploit the unique characteristics of each instrument.

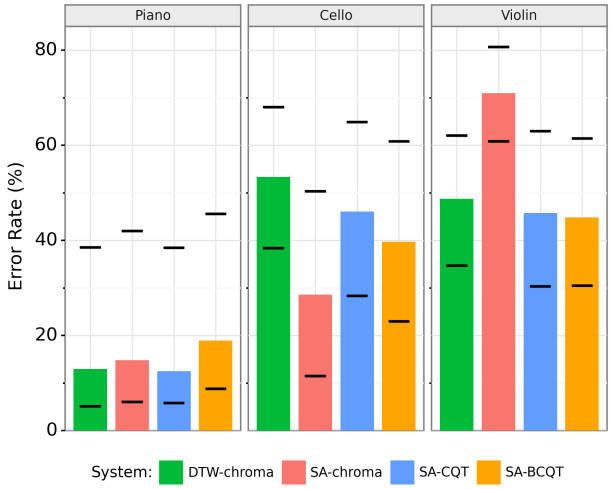

Figure 4. Results on the piano trio dataset. Results are shown separately for each instrument's alignments against the full mix. All subtractive approaches use a piano–cello–violin ordering when performing spectral subtraction.

5. CONCLUSION

This paper envisions an application in which a group of musicians in quarantine can generate a performance of a chamber work by recording each part in isolation in a completely unsynchronized manner, and then aligning, time scale modifying, and mixing the recordings. We focus on the main technical challenge of aligning the individual part recordings. Our approach is to use an auxiliary "full mix" recording of the piece as a reference, and to align each part recording against the full mix. We explore an iterative subtractive alignment approach in which each part is aligned against the full mix and then subtracted from it. We characterize the performance of several variants of this approach on two different datasets: one derived from the URMP dataset that contains ensemble works of various instrumentation, and the other consisting of multiple recordings of a piano trio collected from musicians in quarantine during the pandemic. We find that the subtractive alignment approach works reasonably well on the URMP data, but has mixed success on the piano trio data. We present experimental analysis and suggest directions for future improvement.

6. REFERENCES

[1] J. Driedger and M. Müller, "A review of time-scale modification of music signals," *Applied Sciences*, vol. 6, no. 2, p. 57, 2016.

[2] J. Driedger, M. Müller, and S. Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 105–109, 2013.

[3] A. Moinet and T. Dutoit, "PVSOLA: a phase vocoder with synchronized overlap-add," in *Proc. of the 14th Int. Conference on Digital Audio Effects (DAFx)*, 2011, pp. 269–275.

[4] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

[5] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 181–185.

[6] A. Vaglio, R. Hennequin, M. Moussallam, G. Richard, and F. D'alché-Buc, "Multilingual lyrics-to-audio alignment," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2020, pp. 512–519.

[7] F. Henkel, R. Kelz, and G. Widmer, "Learning to read and follow music in complete score sheet images," in *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, 2020, pp. 780–787.

[8] M. Dorfer, J. Hajič, A. Arzt, H. Frostel, and G. Widmer, "Learning audio-sheet music correspondences for cross-modal retrieval and piece identification," *Trans. of the International Society for Music Information Retrieval*, vol. 1, no. 1, pp. 22–33, 2018.

[9] M. Dorfer, F. Henkel, and G. Widmer, "Learning to listen, read, and follow: Score following as a reinforcement learning game," in *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, 2018, pp. 784–791.

[10] D. Yang, T. Tanprasert, T. Jenrungrot, M. Shan, and T. Tsai, "Midi passage retrieval using cell phone pictures of sheet music," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2019, pp. 916–923.

[11] T. Tsai, D. Yang, M. Shan, T. Tanprasert, and T. Jenrungrot, "Using cell phone pictures of sheet music to retrieve midi passages," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3115–3127, 2020.

[12] C. Fremerey, M. Müller, and M. Clausen, "Handling repeats and jumps in score-performance synchronization," in *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, 2010, pp. 243–248.

[13] M. Grachten, M. Gasser, A. Arzt, and G. Widmer, "Automatic alignment of music performances with structural differences," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2013, pp. 607–612.

[14] M. Shan and T. Tsai, "Improved handling of repeats and jumps in audio-sheet image synchronization," in *Proc. of the International Society for Music Information Retrieval Conference*, 2020, pp. 62–69.

[15] T. Nakamura, E. Nakamura, and S. Sagayama, "Real-time audio-to-score alignment of music performances containing errors and arbitrary repeats and skips," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 329–339, 2015.

[16] Y. Jiang, F. Ryan, D. Cartledge, and C. Raphael, "Offline score alignment for realistic music practice," in *Sound and Music Computing Conference*, 2019.

[17] A. Maezawa and H. G. Okuno, "Audio part mixture alignment based on hierarchical nonparametric bayesian model of musical audio sequence collection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5212–5216.

[18] M. Müller, H. Mattes, and F. Kurth, "An efficient multiscale approach to audio synchronization," in *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, 2006, pp. 192–197.

[19] S. Salvador and P. Chan, "FastDTW: Toward accurate dynamic time warping in linear time and space," in *Proc. of the KDD Workshop on Mining Temporal and Sequential Data*, 2004.

[20] C. Tralie and E. Dempsey, "Exact, parallelizable dynamic time warping alignment with linear memory," in *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, 2020, pp. 462–469.

[21] T. Prätzlich, J. Driedger, and M. Müller, "Memory-restricted multiscale dynamic time warping," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (CASSP)*, 2016, pp. 569–573.

[22] T. Tsai, "Segmental dtw: A parallelizable alternative to dynamic time warping," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[23] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 522–535, 2018.

[24] G. W. Zack, W. E. Rogers, and S. A. Latt, "Automatic measurement of sister chromatid exchange frequency," *Journal of Histochemistry & Cytochemistry*, vol. 25, no. 7, pp. 741–753, 1977.

[25] S. Yong, S. Choi, and J. Nam, "PyTSMod: A python implementation of time-scale modification algorithms," in *Late-Breaking Demo Session at the International Society for Music Information Retrieval Conference (ISMIR)*, 2020.