Virtual Ensemble Assembly: Musicality in Separation

Kaitlin Pet
Indiana University
kpet@iu.edu

Christopher Raphael
Indiana University
craphael@indiana.edu

ABSTRACT

Remotely assembled music today depends on a predetermined "reference track". While reference tracks ensure all parts sound 'in sync' when combined, they restrict a musician's ability pick their own tempo or perform other timing-related musical gestures. Working with expert ensembles, we present experiments that address this problem. We first created a remote assembly pipeline allowing artists to have temporal freedom during the recording process. We then develop three algorithms that preserved players' original timing intentions while ensuring relative synchrony between parts: 1) direct modeling of chamber music expertise, 2) optimization of desirable performance qualities, and 3) performance simulation based on competing goals. Though the resulting music we assembled did not capture the full expressive range of an in-person performance, our work both increases the quality and ease of making remote recording and introduces exciting new applications for aiding synchronous rehearsal and performance.

1. INTRODUCTION

Historically, classical music has been performed synchronously and in the same space — real time interaction and communication between players are considered key to a cohesive performance. The COVID-19 pandemic severely curtailed in-person ensemble playing, thus bringing increased interest to methods of remote music-making. Though networked music technologies (e.g. SoundJack [5])offer the possibility of playing semi-synchronously over the Web, they have inherent network latency issues. Asynchronous remote recording was thus a more widely utilized option. In 2020, orchestras such as the New York Philharmonic [15], the Baltimore Symphony Orchestra [2], and online chamber festivals such as zFestival[25] asynchronously created fully remote orchestral and chamber performances in response to pandemic-related restrictions. Most of such performances required each player to record their part with a standard reference recording. This reference usually takes one of three forms: 1) a "click track", i.e. an audio track containing a series of percussive hits marking the intended time of each beat in the final recording [19], 2) a fixed recording of an accompaniment part, e.g. the piano accompaniment for a choir[23], or 3) a "conducting video" showing a conductor's gestures that players can follow [23]. Since every individual of the ensemble records their part while listening to the reference track, parts can now be assembled together to give the impression of a synchronously recorded performance.

Remote performance technology has many exciting possibilities beyond just mitigating COVID-specific social distancing issues. It has the potential to create giant ensembles where physical location or nationality is irrelevant. Eric Whitacre's Virtual Choir was one of the few instances of pre-COVID remote performance and united hundreds of singers from around the world. Each singer separately recorded choral parts of Whitacre's works using an accompaniment track and conducting video [23]. Remote recording could also serve as a feasible alternative for amateur or professional musicians whose schedules or geographic locations preclude the possibility of all members being in the same room at the same time for normal, synchronous performance. Lastly, remote recording could be a cheaper and easier alternative to musical modalities already recorded with click track. A prime example is film music. Many movie soundtracks are recorded synchronously with a full orchestra and conductor. However, unlike a traditional concert, a film ensemble's playing needs to line up with film visuals. Thus, instead of deciding their own tempo, the conductor and/or players may wear headphones piping in a click track representation of the film visuals' timing(see example in [22]). Remote assembly provides a cheaper alternative that does not require booking a hall or other expensive recording infrastructure.

However, the current practice of recording parts asynchronously via click track or reference recording has several limitations. Tuning and balance issues resulting from musicians not hearing each other can be fixed by recording engineers in "post", either by hand or through automatic means [6, 10, 17]. However, there are inherent downsides when musicians base their musical timing off a preset reference instead of making their own timing choices. Timing is a key aspect of musical expression [4]. Within the reference track framework, players no longer have the autonomy to choose their own tempo or selectively slow down or accelerate their playing for expressive purposes. Though musicians can express musicality through other avenues like dynamics and timbre, an unyielding reference track can curtail their expressive range and make the performance sound "robotic".

The lack of real-time feedback from other musicians can also result in more practical concerns. In isolation, perform-

Web Audio Conference WAC-2022, December 6–8, 2022, Cannes, France.

ers could interpret the same musical gestures differently, misread rhythms, or miscount resting sections. These mistakes are easy to catch in a synchronous setting. A performer could notice they are diverging from the group, or others could point out their mistake. In "post", however, timing errors are time-consuming to correct. For example, while many existing digital audio workstation (DAW) technologies allow users to manually time-stretch tracks, capacity for automatic misalignment correction is very limited [1]. Fixing the larger-scale timing inaccuracies described above is thus cumbersome and time-consuming.

If players do not record to a standard reference, how can one ensure the final recording is "in sync"? Our work demonstrates an alternative to referenced-based assembly that is more conducive to eliciting and retaining performers' musically expressive timing. We first designed a pipeline that can variably stretch each player's audio to an arbitrary score-based grid of times, exactly specifying the temporal location of each note played. We tested this system with two professional chamber ensembles and one semi-professional chamber ensemble. Next, we developed algorithms to generate a score-based grid. These algorithms balance competing objectives to keep parts in sync, preserve the performers' musical intentions, and reflect the music's stylistic requirements. Our algorithms can be divided into three categories:

1. **Direct modeling of chamber music expertise**

   Novice chamber groups will often have coaches who give them advice on how to play as a group. We chose two pieces of procedural advice: 1) "Follow the Leader" and 2) "Focus on Big Beats", and operationalized this chamber music advice via simple algorithms. For example, "Follow the Leader" translates to setting one part as the leader and warping all other parts to fit into the leader's timing scheme.

2. **Optimization of Desirable Performance Qualities**

   Chamber performances are judged by technical qualities like "togetherness" as well as musical qualities like interpretation and stylistic integrity. In this approach, we translate each quality into a separate loss term and calculate a global timing scheme which minimizes the sum of all loss terms. Weighting different terms allows us to place more or less emphasis on different qualities.

3. **Performance simulation based on competing goals**

   Our last method aims to directly model the process of interactively making music as a group. Each player is assigned to an 'agent' imbued with musical goals a player might care about. For example, a player's goals to "exert my musical idea", "be together with the group", and "maintain a steady pulse" can be translated into factors determining where the corresponding agent chooses to place their next notes.

Of course, asynchronous recordings can never replace the emotional depth of live performance with real-time interaction and communication between musicians. However, we find that striving for the best possible asynchronous recording can pave the way to exciting new applications and practices. Our pipeline and algorithms create the possibility of creating highly expressive asynchronous recordings that can be adapted to any existing tempo framework; for example, a film soundtrack where players perform the parts as they choose, then all parts are manipulated to align with the film visuals. In addition, asynchronous recordings could provide new insight into the traditional synchronous performance process. We lastly hope our work inspires readers to use timing modification in new and creative ways that go beyond what we present.

2. BACKGROUND AND RELATED WORK

In most existing asynchronous recordings, tools like click tracks, conducting videos, or pre-recorded accompaniment prescribe a set global timing that players must follow. With "stems" created in this manner, assembly is simple. A recording engineer can place all tracks in a DAW, then arrange them so each track's reference aligns. If the references are in the same temporal position, the stems should, by extension, align as well. The DAW will then automatically add the aligned signals when "bouncing" the final recording [1].

Assembly becomes more complicated if players are allowed to play their parts without an external beat, speeding up and slowing down as they choose. First, one can use score-alignment methods to determine when each note of the score is played in each recorded part[16, 8, 21, 7, 14]. Every note can be stretched or compressed without changing the pitch or timbre of the audio using phase vocoding [10] or PSOLA [6], and shifted to an arbitrary location. Ideally, the new note locations would result in parts being score-synchronous; i.e. notes from different parts that appear at the same score location sound at the same time. The simplest example of this is a "metronomic" grid, where the played time (e.g. in seconds) of each note is proportional to the musical time (e.g. in beats). For example, one could create a grid where, regardless of context, each quarter note in the score is translated to a played length of 1 second, each half note in the score is translated to a length of 2 seconds, etc. The result is a "metronomic" recording without any expressive timing.

For many songs within the rock and pop music genre, this metronomic-style timing is the norm and completely acceptable, even desirable. Within classical music, timing specification tends to be more complicated. For example, certain pieces encourage a "wiggly" tempo where players speed up and slow down to emphasize musical ideas, while others expect a relatively constant tempo to remain within the bounds of "good taste". Additionally, classical musicians will sometimes lengthen certain notes to bring out their tension, or speed up slightly to convey a more agitated mood. These gestures are important in capturing the nuance and expressive intention in a classical music performance.

Previous literature has modeled the relationship between classical musical scores and expressive performance. These "expressive synthesis" models link information from the score to time series of expression-related variables like note onset times and volume over time. Using expressive synthesis models, one can render a score in a way that captures expressive tendencies characteristic of the music's genre or period. Some models are "expert system"-based and seek to distill musician knowledge into algorithmic form [3, 20]; others are data-driven and generate models based on aggregating labeled performance data [18, 13]. Besides synthesizing expressive performances, such models can also be used to better understanding how humans approach the act of performing music [4]. Other models such as our previous work

[12] "smooth" or create altered tempo schemes, which improve the timing of existing performances by seeking simple parameterizations of observed performance data.

Our work is related to expressive synthesis in that we wish to obtain an "expressive" series of note onset times. However, in addition to obtaining information from the score, we are taking into account the players' interpretation in their asynchronous part recordings. We want players to see their expressive intentions manifested in the final timing grid, ideally recognizing the expressive timing gestures in the final recording as something they could have done themselves in a synchronous setting.

Though our low quantity of data makes a pure machine learning approach more of a challenge, we are still able to take the expert system approach by hand-crafting features which can be translated into heuristics or explicit loss functions. In addition, our optimization-based approaches use folk knowledge about desirable characteristics of synchronous recordings to determine the terms of a minimizable loss function.

Another area of related work focuses on the study of how musicians approach ensemble playing, as well as works that explicitly model group power dynamics and players' moment-to-moment tempo choices[24, 9, 11]. Wing et al. use a "phase-matching" model to describe how two string quartets approach an excerpt of Haydn's Op. 74 No. 1 string quartet [24]. Their model shows differences between players who lead vs. players who are more dependent on the group's timing — leaders are less willing to correct their tempo to match that of the group. We seek to recreate this leader/follower dynamic in several of our approaches. We both use a heuristic approach to explicitly set leaders as those who do not modify their tempo, and use a simulation-based approach to model more complex interactions between group members.

3. METHODS

3.1 Participants

We collaborated with three ensembles: 1) an undergraduate piano/clarinet/cello trio from the Indiana University Jacobs School of Music (JSoM), 2) an octet containing both the authors and faculty/students from the Yale School of Music (YSM), and 3) an octet consisting of faculty from the Universität der Künste(UdK) in Berlin. We chose to work with groups consisting of pre-professional and professional musicians, including those considered experts in their respective fields. Because of their high musical standards,their input was invaluable to refining our algorithmic approaches. See **Future Directions** for potential remote ensemble assembly applications aimed at amateur or even beginner musicians.

The JSoM undergraduate trio played the third movement of Johannes Brahms' *Clarinet Trio, Op. 114*. They were an established ensemble before the pandemic began, and had previously rehearsed the trio in person.

The UdK group chose the first movement of Felix Mendelssohn's *String Octet in E-flat Major*. Unlike the JSoM group, the UdK group had not played their piece together in person before lockdown. However, all members knew the piece well and had performed it previously with different groups.

The YSM-based octet played the *Adagio* movement from Wolfgang Amadeus Mozart's *Serenade no. 11 for Winds in E-flat Major*. Out of the three groups, the YSM group had the least "prior information" going into recording. They had previously never played together as group, and the *Serenade* was new to several players (though all musicians had played other pieces by Mozart in the past). Moreover, they did not discuss any aspect of the performance before recording, which was shown by Whitacre as a method of getting players "on the same page" in regards to interpretation in asynchronous ensembles [23].

We asked all participants to record their part alone, either in their homes or in a practice room. Some musicians used professional recording equipment, while others used recording apps on their phones.

3.2 Audio Assembly Process

We used Hidden Markov Model(HMM)-based score-alignment[16] to automatically identify the note onset times (in seconds) of every note in each part. As a control on the process, we edited the alignments using an interactive tool that gives a user access to both sound and a visual representation of that sound when adjusting the onsets of the notes. The changes made through this process were, for the most part, minor. We hope to consider fully automatic assemblies in the future.

Since players were unable to match their pitch and tune to each other through remote recording, our first assemblies suffered from distracting and systematic intonation problems. To correct this we performed a fully automatic tuning adjustment for each part at the note level. In this process we estimated a single representative frequency for each note by interpolating the spectral energy in the neighborhood of a prominent note harmonic for each frame in the note, averaging these frame-level frequencies. We then adjusted the audio data by resampling each note separately, resulting in an average frequency consistent with equal-tempered tuning centered around A442. Our collaborators were also bothered by the fact that the relative volume of different parts in the final recording (balance) did not reflect how the piece was suppose to sound, so we adjusted levels and reverb by hand to appropriately balance and spacialize the mix.

3.3 Determining a Global Timing Scheme

We denote the estimated onset times from our score alignment process as $\{t^0_{i,j}\}$ where i indexes the P independent musical parts and j indexes the notes in the part. As a convenience we interpolated these times to the level of the *composite* rhythm — the union of all score onset times. Thus the score time (in measures) associated with $t^0_{i,j}$, $s_{i,j}$, satisfies $s_{i,j} = s_{i',j}$ for $i, i' = 1, \ldots, P$. This construct allows every part to use the same set of indices. See Figure 1 for more detail.

Our overarching goal is a mapping from the players' original timing schemes, $\{t^0_{i,j}\}$, to the final timing scheme, $\{t_{i,j}\}$, where, for now, score-synchronous notes from different players are rendered at the same time. Once we have created this "global" timing scheme we can warp the audio of each part to conform to the scheme using phase-vocoding, where the phase vocoder rate is $\frac{t^0_{i,j+1} - t^0_{i,j}}{t_{i,j+1} - t_{i,j}}$ for the jth note of the ith part.

3.4 Most Basic Implementation

The simplest version of this idea would use "metronomic"

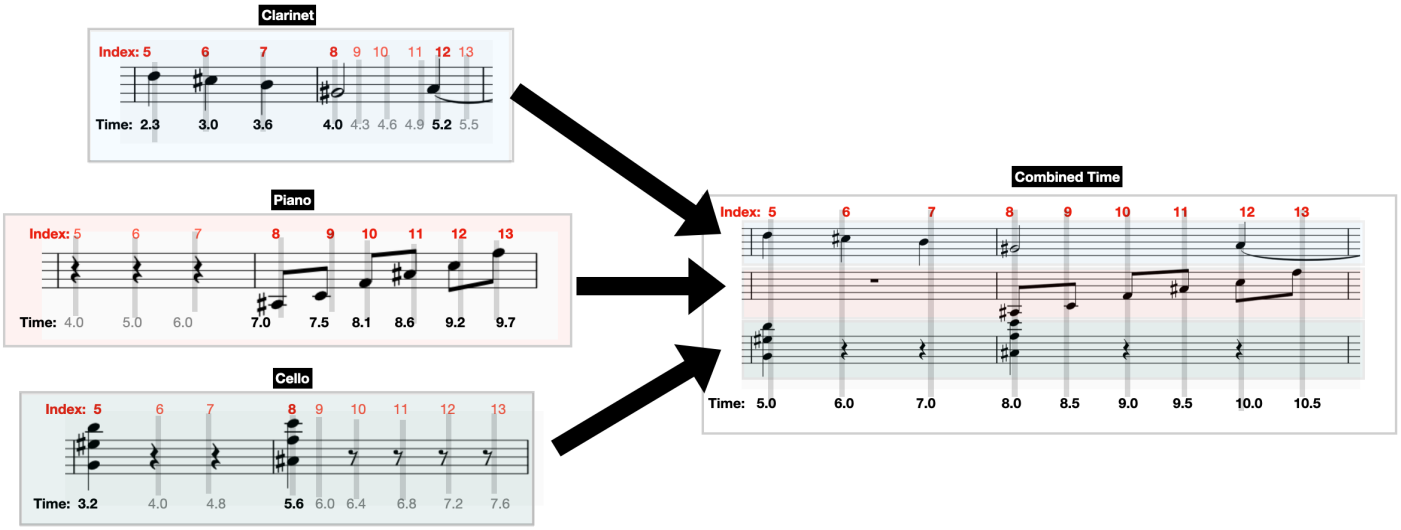

Figure 1: A graphical depiction of how parts recorded at different tempos are stretched or compressed to align. This example shows the combined rhythm indices $j = 5$ to $j = 13$ of the Brahms Trio. Note that combined rhythm points for any given part may occur during rests or held notes. In this example, clarinet, cello, and piano place their notes at different offsets and move through the score at different rates. By aligning the composite rhythm indices, the parts can be made score-synchronous.

timing for the global scheme, thus $t_{i,j} = Cs_{i,j}$ where C is the tempo, expressed in seconds per measure. A "metronomic" rendition of Mozart Serenade can be **heard below**[1] Unsurprisingly, this rendition was considered unsatisfactory by players, with one noting *" It was as if everything that was me was extracted from those notes....This version has succeeded in achieving uniformity by leaching out our individual personalities, and I am not happy about that."*

Such early comments prompted us to explore methods that sought to preserve the players' musical intent — and to resolve conflicts when their intents were not mutually compatible. We explored three approaches to generating global timing schemes, focusing on directly implementing musical folk knowledge. Our first approach uses explicit synchrony transforms, which require that score-simultaneous notes are played at the same time in seconds. Our second approach finds the global timing agenda through an optimization, where we seek a timing configuration that best fits desirable qualities of a finished recording of the piece. Finally, we sought to model the forces at work in chamber music performance by simulating the ways our participants might interact with each other in a real-time synchronous setting.

3.5 Direct Synchrony Transforms

In this approach, we developed algorithms based on musical advice that would put constraints on $\{t_{i,j}\}$. Specifically, $t_{i,j} = t_{i',j}$ for $i, i' = 1 \ldots, P$. In other words, the final timing scheme requires all parts to place their j^{th} rhythm index at the same time.

We wanted algorithms that had the above constraint while simultaneously maintaining the individuals' interpretation and the appropriate musical style. We developed two algorithms: one that allows expressive tempo changes to work with music from the Romantic period, and one that imposes a steady tempo to work with music from the Classical period.

3.5.1 *Romantic algorithm: Follow the Leader*

Different members of a chamber ensemble can play different roles at different times. For example, one leading part could control the tempo flow. Other players would listen carefully to the leading part to fit within the leader's timing choices, and the person playing the dominant part would use body language and other musical techniques to demonstrate a tempo to follow. Additionally, Wing et al.[24] empirically observed that players who serve the role of a leader tend to be more rigid in their expressive tempo actions, expecting others to conform to their interpretation. Our first algorithm takes this idea to its logical extreme by assigning a lead part $L(j) \in \{1 \ldots P\}$ for all j values of the combined rhythm. From an expressive standpoint, the leader is the only player with agency at any given point. When a part is in the lead role, its timing scheme is left as is, and all other parts are warped to fit. Each point in the combined rhythm can be computed by:

$$t_{i,j+1} = t_{i,j} + (t^0_{L(j+1),j+1} - t^0_{L(j),j})$$

This construction is well-suited to the Brahms *Trio* and the Mendelssohn *Octet*, but poorly suited to the Mozart Serenade. When leaders 'switch off', i.e. $L(j-1) \neq L(j)$, the tempo of the piece will change to that of the new leader. In Romantic pieces, a slight tempo change resulting from the changing leadership role gave the intended effect of another musician in the group "taking the lead". In contrast, many pieces by Mozart are traditionally performed in a more reserved way, maintaining a steady tempo. As a result, we only used this algorithm to assemble the Brahms *Trio*

[1] https://drive.google.com/file/d/1IRMHRGhm6VPyEOW92wzq_Dy_IpUhYnTd/view?usp=sharing

and Mendelssohn *Octet*, developing others to work with the Mozart *Serenade*.

3.5.2 Classical Algorithm: Focus on the Big Beats

One of our Mozart Serenade collaborators, YSM clarinet professor David Shifrin, suggested we approach timing by aligning phrases and larger musical units, i.e. "big beats" like measures. In a synchronous setting, this mindset prevents the performance from becoming "bogged down" and helps it maintain a sense of flow.

This observation prompted us to relax the requirement of $t_{i,j} = t_{i',j}$ for all beats in the combined rhythm. Instead, we only enforced synchrony at the subset of the combined rhythm indices which corresponded to big beat alignment points. Instead of stretching each note to a final grid as described above, we now stretch each big beat to its corresponding position in the final grid, using the same phase vocoding techniques.

To implement this in algorithmic form, we defined our big beats as the j indices which corresponded to the beginnings of measures. The style of Mozart calls for a relatively constant overall tempo, so we chose a final timing scheme enforcing the same length for each measure. We then stretched each measure in the recorded parts to the same length and placed every measure at their respective location in the score.

Our first implementation was reasonably score-synchronous for most of the piece, but a few sections were perceptibly out of alignment. To correct the distracting sections, we explored shrinking the size of the "big beat" from a whole measure to smaller units of time (e.g. half measures, quarter notes, etc.) in cases where alignment was poor. See assembled audio in **Results**.

3.6 Optimizing Desirable Performance Qualities

Chamber performances are judged by technical qualities like "togetherness" as well as musical qualities like interpretation and stylistic integrity. In this approach, we translated each quality into a separate loss term and calculated the global timing scheme which minimized the sum of all loss terms. Unlike the approaches described above, score-synchrony is no longer an explicit requirement. Instead, it is only one of many competing qualities.

By weighting the different loss terms, we could place more or less emphasis on different qualities. One can thus express the total loss as:

$$L = \sum_h \lambda_h L_h$$

where h indexes the functions and $\sum \lambda_h = 1$.

We still consider the most basic quality of a chamber performance as being "together", i.e. a performance where parts are score-synchronous. We translate this desire for a synchronous performance into "ensemble loss", defined below:

$$L_e = \sum_j \sum_{i,i',i\neq i'} (t_{i,j} - t_{i',j})^2$$

Simply stated, ensemble loss penalizes instances where different parts fail to play the same beat at the same time. The more a beat diverges, the higher the loss.

Incorporating the performer's original stylistic choices necessitates another loss term. We created a "stretch loss" term that penalizes situations where ratio of lengths between adjacent notes in the originally played parts differs from adjacent note ratios in the new version. Because we are using a ratio instead of absolute length, there is no penalty if the entire part becomes faster or slower. This allows parts recorded at different tempos to converge at the same final tempo. Stretch loss is defined mathematically below:

$$\mathrm{L}_s = \sum_i \sum_j (\frac{t_{i,j+1} - t_{i,j}}{t^0_{i,j+1} - t^0_{i,j}} - \frac{t_{i,j} - t_{i,j-1}}{t^0_{i,j} - t^0_{i,j-1}})^2$$

Simply said, L_s penalizes when the phase vocoding playback rate does not varies smoothly.

In certain cases, it is beneficial to impose explicit temporal qualities on the final recording. For example, like stated earlier, the Mozart *Serenade* needs to maintain a steady macro-level tempo. In other applications like film music, certain beats in the piece may need to align with specific visual cues. Thus, a last loss term, L_{ref}, was introduced to penalizes differences from a pre-determined final timing scheme. It is defined as:

$$L_{ref} = \sum_i \sum_j (t_{i,j} - t^{ref}_j)^2$$

Where t^{ref}_j is the position of a reference time scheme at combined rhythm index j.

3.7 Performance Simulation based on Competing Goals

Simulation-based timing generation attempts to capture the rehearsal process by modeling each part as an "agent" with competing desires. On one hand, each agent wants to respect the player's interpretation by preserving the original part's temporal pattern. On the other hand, each agent wants to stay relatively in sync with the rest of the group. In certain situations, agents may also want to maintain other goals, like a steady tempo.

All agents start by playing the first two notes at predetermined time points. Then, the location of the next note is calculated by weighting trajectories based on their different goals. For a given goal, g,

$$\hat{t}^g_{i,j+1} = G_g(\vec{t_1} \dots \vec{t_j})$$

where

- $\hat{t}^g_{i,j+1}$ is the location of potential future time point $j+1$ of part i if fulfilling goal g.
- $\vec{t_j}$ is the vector of times corresponding to all parts $1 \dots P$ at time point j.
- $G_g(\vec{t_1} \dots \vec{t_j})$ is the function computing $\hat{t}^g_{i,j+1}$, which can depend on all the histories of all parts.

Multiple goals can be expressed by a weighted sum:

$$t_{i,j+1} = \sum_g w_{g,i,j} \hat{t}^g_{i,j+1}$$

where $w_{g,i,j}$, is the weight of goal g at time point j for part i, and $\sum_g w_{g,i,j} = 1$.

We define three types of goals:

1. G_{tog} aims to remain with the rest of the group by considering all member's tempo trajectories:

$$\hat{t}^{tog}_{i,j+1} = \frac{1}{n_p}\sum_{i=1}^{n_p} t_{i,j} + (s_{i,j+1} - s_{i,j})\frac{1}{n_p}\sum_{i=1}^{n_p}\frac{t_{i,j} - t_{i,j-1}}{s_{i,j} - s_{i,j-1}}$$

   Where s_j is the score location (in measures) corresponding to time point j.

2. G_{self} aims to preserve the player's original tempo pattern by replicating the original recording's tempo trajectory:

$$\hat{t}^{self}_{i,j+1} = t_{i,j} + \frac{t^0_{i,j+1} - t^0_{i,j}}{t^0_{i,j} - t^0_{i,j-1}}(t_{i,j} - t_{i,j-1})$$

3. G_{tempo} aims to maintain a steady tempo R:

$$\hat{t}^{tempo}_{i,j+1} = t_{i,j} + (s_{i,j+1} - s_{i,j})R$$

   where R is the desired tempo in seconds per measure.

4. RESULTS

Music is fundamentally subjective. Even among our collaborators, different members of a group had different views on the appropriate method to assemble parts with conflicting interpretations. We have no objective metrics to determine whether an assembled piece is good or bad. Instead, we will present what we qualitatively consider the "best" assembled recordings, providing both our own analysis and comments from the performers. We will also mention other experiments that yielded less ideal audio results but shed light on characteristics of the algorithms.

4.1 Direct Synchrony Transform

Two of our methods used the idea of a Direct Synchrony Transform, where chamber music advice was used to fix the times of the composite rhythm so score-synchronous events happen simultaneously. "Follow the Leader" and "Focus on the Big Beats" were unique among our approaches in that they 1) explicitly required score-synchrony and 2) the final version directly replicated the temporal patterns in the player's original recordings. However, "Follow the Leader" and "Focus on the Big Beats" differed from each other in two key ways. First, different musical units were aligned and stretched; "Follow the Leader" aligned every note, while "Focus on the Big Beats" aligned a larger musical chunk, such as a measure. Second, the final timing grid in "Follow the Leader" was derived from the player with the highest note density, while "Focus on the Big Beat" had a constant grid with the length of each big beat proportional to a constant tempo C. In our opinion, these differences made "Follow the Leader" more suitable for the Brahms and the Mendelssohn, and "Focus on the Big Beats" more suitable for the Mozart.

Linked Below [2] is a recording of the JSoM Brahms trio assembled using the "Follow the Leader" algorithm. The video part was created by adjusting the video frame playback rate to mirror the audio warping.

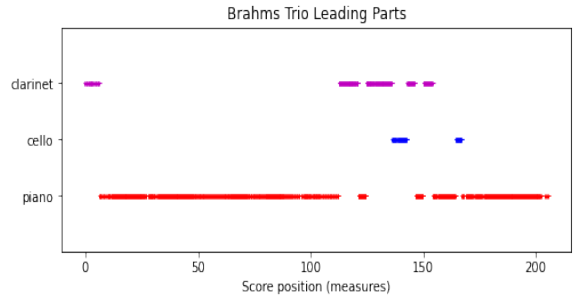

Figure 2: In the "follow the leader" algorithm, a leader is chosen based on which part has the highest note density. In the Brahms Trio, piano is the most frequent leader, while cello leads very little.

Theoretically, we could have assigned the lead part for each beat $L(j)$ by hand based off a musical analysis. However for these experiments, we instead made an simplifying assumption to automatically determine which part was was leading at a given beat. We calculated $L(j)$ by optimizing an objective function that rewarded high note density in the lead part but penalized changes in leader. See Figure 2 for a graphical representation of which part was leading at which time in the Brahms.

The JSoM group had a very positive reaction to this rendition, with members commenting that the final recording *"sounds awesome"*. Note that these were young, "tech-positive" musicians.

The Mendelssohn *Octet* recording, **linked below** [3], had a less positive reception. One musician noted *"the first violin is not leading anymore...because of the 'following' 1st violin, the piece loses the music phrases too often. It hurts my heart when I am listening to it.."*

This comment highlights challenges in the "Follow the Leader" framework. "Follow the Leader" will translate a *single* human musicians' expressive choices without any interference or adjustment. If the assigned leader plays expressively, the assembled piece will be rendered in an expressive manner. On the flip side, if the assigned leader does not play expressively, the interpretation will sound more flat. Our implementation designated the leader as the member with highest note density for long stretches of time. In the case of Mendelssohn, this usually corresponded to a part playing the 16^{th} note accompaniment figure. As a result, the resulting interpretation is more subdued, and actively counters our collaborator's musical expectation that 1st violin should lead. In contrast, the Brahms Trio was for the most part lead by piano, which makes sense in a musical context. This could be part of the reason the Brahms recording ended up with more "character".

While "Follow the Leader" translates a single human musician's expressive choices at a time, "Focus on the Big Beats" allows all musicians' expressive choices to be realized simultaneously. **Linked below**[4] is our final rendition of "Focus on

[2] https://drive.google.com/file/d/1JbDaHAtMaMRZVZGHp65wEFBK4OETLZ2f/view?usp=sharing

[3] https://drive.google.com/file/d/1Pk-5pZwteXTcaDJYOJApAdHxfa8Kdw9C/view?usp=sharing

[4] https://drive.google.com/file/d/

the Big Beats". This rendition uses measure-long big beats for most of the piece, but uses reduced-length chunks to improve synchrony in certain sections. One collaborator noted that *"Some parts sound very good, very natural"*. However at the same time, she noted the interpretation was not *"idiomatic"*— certain sections did not feel like they were played by a professional ensemble. This amateur quality could manifest through parts *"luxuriating in the lines a little too much"*. Ultimately, *" it sounds like people are trying to do musicality who have never heard Mozart before"*.

Some of these issues of "unnatural" musicality could be a result of the asynchronous nature of recording. Though musicians will often imagine the other parts of an ensemble while playing their part in isolation, the lack of real audio feedback may lead to more extreme musical gestures. But in general, the authors thought the implementation successfully achieved our initial goals. The final recording closely reconstructed the original parts' tempo patterns while ensuring parts were together. However, as our collaborator noted, this was not enough to give the ensemble a sense of cohesion — it seems like every player is sticking to their own interpretation, ignoring the expressive gestures their fellow players are trying to convey. While the resulting recording sounded like a realistic performance, it did not accurately reflect our group's expertise and level of playing.

4.2 Optimization of Desirable Performance Qualities

Optimization allowed us to capture multiple desirable qualities of a final recording at a same time. Instead of mandating synchrony, optimization pushes the final timing scheme toward synchrony with L_e. While players' temporal patterns are not explicitly copied, the L_s term encourages those temporal patterns to be retained in the final result. Lastly, L_{ref} provides a way to impose the steady overall tempo required in the Mozart *Serenade*.

Linked below[5] is a version of Mozart generated with $\lambda_e = 1 * 10^{-6}$, $\lambda_s = 0.99999$, and $\lambda_{ref} = 9 * 10^{-6}$. This version uses our original metronomic grid as the reference time scheme $\{t_j^{ref}\}$. In general, our collaborators liked this version best out of all the assembled recordings. One player liked how the recording sounded "natural", but qualified this by noting *"Overall, there's also a metronomic feeling without much artistic individuality...it doesn't sound like we are blending or on the same page musically"* . Despite these shortfalls, the optimized version was still her favorite. She explained *"it has a musical shape, it doesn't have the most personality but it's a nice rendition. In other renditions, there were standout parts but also parts that bombed. For this kind of piece, you want consistency instead of bombing in one part and sounding really good in another part"*.

In this implementation, we have three controllable parameters corresponding to the weight of each loss term. Care must be taken in choosing these parameters for realistic results. Notably, when ensemble loss is weighted higher than .00001, the result is very smooth, sounding similar to our original metronomic rendition. Another interesting variation considers a λ_e of 0, resulting in a situation that pits the desire of staying with a metronomic grid against the desire to preserve the original interpretations. This arrangement yields a result that is identifiably score-synchronous for some of the piece, but goes out of sync when players have different interpretations.

4.3 Performance Simulation Based on Competing Goals

The simulation method allowed us to convey more complex inter-musician interactions. For example, it has the potential of representing a more realistic version of "leading" than our previously described "Follow the Leader" algorithm, incorporating Wing et al.'s notion of a leader as a player who adjusts *less* than others, instead of one who does not adjust at all [24]. This flexibility leads to a very high number specifiable parameters. Our previous "Follow the Leader" needs lead part $L(j) \in \{1...P\}$ defined for every index of the combined rhythm — though we used a simplifying assumption to avoid specifying values manually, a person could in theory have the power to designate each value by hand or with a sequencer-like graphical user interface. Our current simulation method needs a continuous weight value, $w_{g,i,j}$, defined for every goal at every beat for every instrument. This could give a potential user a lot more freedom to specify how leadership roles can be divided among the parts and in different sections of the piece.

For our experiments with Mozart, we chose to specify these weights in broad strokes. Our first version assumed all agents had the same parameters for all notes, in this case $w_{i,j}^{tog} = .36$, $w_{i,j}^{self} = 0.54$, and $w_{i,j}^{tempo} = .1$. In the recording **linked below** [6], one can see that while certain sections are cohesive, others have parts which become out of sync. This divergence occurs when two parts have highly conflicting interpretations - the G_{tempo} and G_{avg} terms are insufficient to bring the parts into unison. One collaborator noted *"It has moments that stand out as good, but overall there are glaring timing issues and some parts that completely unravel."* She also noted this recording *"had more character than the [optimization version] despite its alignment issues – many sections that were more rigid in the [optimized version] flowed more naturally in the [simulation version] before it starts unravelling."* To counter this "unraveling", we identified sections with especially bad synchrony (e.g. timestamps 1:42-2:10) and adjusted the ratio of w_{avg} to w_{self} such that w_{avg} played a more prominent role. Though agents had less "personality" in these sections, the ensemble was significantly better. See a recording of the adjusted version **linked below** [7].

Notably, certain weight combinations gave poor results that reflect real issues musicians face during ensemble playing. If we give no weight to the "internal pulse" parameter G_{tempo}, the resulting recording gets slower and slower over time. This is because the simulation gets caught in a "feedback loop" whenever instruments take extra time to be expressive. Musicians will sometimes add expression by lengthening certain notes then returning to their original tempo. Experienced chamber ensembles will recognize these longer notes as micro-level expressive gestures — though they may adjust their playing to stay aligned, they will con-

1JKpyP5T2CAzRJIPbkm8Cn32P1eyySU1Z/view?usp=sharing

[5] https://drive.google.com/file/d/1nMu56vJGgfY_u9-hjIybR9wS4u8RxdWq/view?usp=sharing

[6] https://drive.google.com/file/d/1IGkFHHYf_QII25pmJuOsYk9t47cQdPzX/view?usp=sharing

[7] https://drive.google.com/file/d/17Pe_9OwP2UbP4b7t_YMIn7mqiyTXU9ef/view?usp=sharing

tinue the piece at the same tempo. But if the simulation has no G_{tempo}, the other agents will naturally "perceive" lengthened notes as the intention to slow down, causing the piece to get slower and slower. Interestingly, a similar phenomenon can be observed in beginner ensembles who focus too much on 'reactively' listening to each other instead of 'proactively' maintaining the correct tempo. It is possible this simulation-based method can be used as tool to better understand how synchronous ensemble playing works.

5. DISCUSSION AND FUTURE DIRECTIONS

Our overarching goal was to produce characteristic and expressive performances of assembled pieces by combining the musicians' individual raw performance data with "chamber music knowledge" translated into algorithmic form. We investigated three groups of methods: those that modeled musical expertise, those that optimize desirable performance qualities, and those that simulate the process of performing. Each of these methods reflected different aspects of synchronous performance.

In some ways, the inherent nature of our "record once then assemble" process can never reach the quality and experience of performing together in person. Players do not have the opportunity to adjust their original interpretation — what they start with is what they are stuck with. In a real chamber setting, a group of advanced and expert players like those in the YSM group will often adjust their interpretation over time, coalescing into an interpretation that communicates a cohesive musical idea.

On a deeper level, the process of remote assembly cannot replace the experiencial aspect of playing in person — recording alone cannot replicate the joy of making music together with other people. This can be seen in current remote recording trends. Now in 2022 with many social distancing restrictions lifted, we see far fewer new remote recordings are being created compared to the height of the pandemic.

Our vision for this technology is not a replacement for in-person music-making. We present two future paths for this technology. The first highlights how our improved asynchronous music production framework could serve a middle ground between synchronous recordings and MIDI-based music synthesis. The second focuses on how asynchronous assembly could be used as a tool to improve synchronous rehearsals and performance.

5.1 "MIDI+" Music Synthesis

Though our collaborators had mixed responses to the musical aspects of our assembled recordings, many were impressed by the overall quality given the relatively easy recording process. Our collaborators just had to set up their recorders and play through their parts, without worrying about sticking to an external reference. Some players explicitly chose to not count extended rests between sections of playing. If we had used a reference track, miscounting rests between sections would be considered a major timing error that a recording engineer would need to manually correct. Since our framework is based on note onset identification given a known score, the score matching algorithm was able to automatically account for shortened resting sections. The ability to catch rhythmic irregularities automatically gives our framework a big advantage over the reference track framework.

In addition, two of our algorithms — "Focus on the Big Beats" and "Optimization of Desirable Performance Qualities" — allow for avenues of imposing rhythmic structure onto the generated music while still maintaining synchrony and musicality. In our experiments, we chose a steady framework to match Mozart's stylistic requirements. However, a user-specified temporal framework could also be used to satisfy other goals. For example, imagine the case where a film score requires certain musical gestures to align with visual cues. In this case, the timing of the final reference grid could be based on the timing of the visual cues. This allows the players to record their parts without worrying about synchronization-related details. Moreover, if the visuals are edited, the audio can be easily re-synthesized to match the new visual timing. Currently, this level of freedom over audio content is only possible with fully synthesized audio, which tends to sound worse than music played by live musicians. Our technology introduces the possibility of "MIDI+" music recordings that retain some musical qualities of live player but are much cheaper and easier to produce.

5.2 Assembly as a Rehearsal Tool

In addition to producing finished recordings, our assembled ensembles could have the potential to provide valuable insight in synchronous rehearsals.

When musicians are playing a technically challenging chamber music piece, they may have to focus more attention on staying in sync with the other players. This extra attention on lining up complex rhythmic patterns can come at the expense of playing in a stylistically appropriate or emotive way — often, music gets stilted or "bogged down" because musicians are thinking about the placement of their next note instead of the overall line. This problem is especially acute for pieces where a melody is placed over an complex rhythmic backdrop. Ensemble assembly gives players an opportunity to hear what they could sound like if not under rhythmic duress. When recording their parts alone, players could focus on shaping their musical lines rather than aligning with the group. Small rhythmic errors can be "parsed out" by the score alignment process instead of leading to the ensemble falling apart. After assembly, players could hear a version of themselves perform tricky sections with better flow.

6. ACKNOWLEDGMENTS

All our musical collaborators were invaluable to this project, both in their beautiful playing and their comments on our recordings. We would like to thank David Shifrin, Nikki Pet, Frank Morelli, Eleni Katz, William Purvis, and Olivia Martinez from the Yale School of music; Kaden Larsen, Shinae Ra, and Elle Crowhurst from the Jacobs School of Music; and Yuta Nishiyama and colleagues from the Universitäte der Künste.

7. REFERENCES

[1] Apple. Logic Pro X User Guide, May 2022. URL: https://support.apple.com/guide/logicpro/welcome/mac.

[2] Baltimore Symphony Orchestra. BSO Virtually Performs Powerful Ending of Mahler's Third Symphony, 03 2020. URL:

https://www.youtube.com/watch?v=YOy_JkmGX6s&ab_channel=BaltimoreSymphonyOrchestra.

[3] R. Bresin and A. Friberg. Synthesis and decoding of emotionally expressive music performance. In *Proceedings of the IEEE 1999 Systems, Man and Cybernetics Conference - SMC'99*, volume 4, pages 317–322, 1999. QCR 20180918. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=812420.

[4] C. E. Cancino-Chacón, M. Grachten, W. Goebl, and G. Widmer. Computational models of expressive music performance: A comprehensive and critical review. *Frontiers in Digital Humanities*, 5, 2018. URL: https://www.frontiersin.org/article/10.3389/fdigh.2018.00025, doi:10.3389/fdigh.2018.00025.

[5] A. Carôt, M. Dohler, simon Saunders, F. Sardis, R. Cornock, and N. Uniyal. The world's first interactive 5G music concert: Professional quality networked music over a commodity network infrastructure. In *Proceedings of the 17th Sound and Music Computing Conference*, June 2020. URL: https://zenodo.org/record/3898918, doi:10.5281/zenodo.3898918.

[6] F. Charpentier and M. Stella. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 2015–2018, 1986. doi:10.1109/ICASSP.1986.1168657.

[7] P. Cuvillier. *On temporal coherency of probabilistic models for audio-to-score alignment*. PhD thesis, Paris, France, 2016.

[8] R. Dannenberg and C. Raphael. Music score alignment and computer accompaniment. *Commun. ACM*, 49:38–43, 08 2006. doi:10.1145/1145287.1145311.

[9] J. Davidson and J. Good. Social and musical coordination between members of a string quartet: An exploratory study. *Psychology of Music*, 30:186–201, 10 2002. doi:10.1177/0305735602302005.

[10] M. Dolson. The phase vocoder: A tutorial. *Computer Music Journal*, 10(4):14–27, 1986. URL: http://www.jstor.org/stable/3680093.

[11] W. Goebl and C. Palmer. Synchronization of timing and motion among performing musicians. *Music Perception - MUSIC PERCEPT*, 26:427–438, 06 2009. doi:10.1525/mp.2009.26.5.427.

[12] Y. Gu and C. Raphael. Modeling piano interpretation using switching Kalman filter. pages 145–150, 01 2012.

[13] D. Jeong, T. Kwon, Y. Kim, and J. Nam. Graph neural network for music score data and modeling expressive piano performance. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3060–3070. PMLR, 09–15 Jun 2019. URL: https://proceedings.mlr.press/v97/jeong19a.html.

[14] M. Müller, F. Kurth, and M. Clausen. Audio matching via chroma-based statistical features. pages 288–295, 01 2005.

[15] New York Philharmonic. A Boléro from New York: NY Philharmonic Musicians Send Musical Tribute to Healthcare Workers. URL: https://www.youtube.com/watch?v=D3UW218_zPo&ab_channel=NewYorkPhilharmonic.

[16] C. Raphael. Music Plus One and Machine Learning. In *Proceedings of the Twenty-Seventh International Conference on Machine Learning*, pages 21–28, 2010.

[17] J. Scott, M. Prockup, E. Schmidt, and Y. Kim. Automatic multi-track mixing using linear dynamical systems. In *Proceedings of the 8th Sound and Music Computing Conference*, 01 2011.

[18] Z. Shi. Computational analysis and modeling of expressive timing in Chopin's Mazurkas. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, pages 650–656, Online, Nov. 2021. ISMIR. doi:10.5281/zenodo.5624515.

[19] S. Steven. *Music Editing for Film and Television : The Art and the Process*. Routledge.

[20] J. Sundberg, A. Friberg, and L. Fryden. Threshold and Preference Quantities of Rules for Music Performance. *Music Perception*, 9(1):71–91, 10 1991. arXiv:https://online.ucpress.edu/mp/article-pdf/9/1/71/191885/40286159.pdf, doi:10.2307/40286159.

[21] V. Thomas, C. Fremerey, M. Müller, and M. Clausen. Linking Sheet Music and Audio – Challenges and New Approaches. *Multimodal music processing*, 3, 01 2012.

[22] Walt Disney Animation Studios. The Magic of Orchestration Clip | Into the Unknown: Making Frozen 2 | Disney+. URL: https://www.youtube.com/watch?v=3uXhyHzuGMY.

[23] E. Whitacre. The Virtual Choir: How we did it, Nov 2020. URL: https://ericwhitacre.com/blog/the-virtual-choir-how-we-did-it.

[24] B. A. Wing Alan M., Endo Satoshi and V. Dirk. Optimal feedback correction in string quartet synchronization. *J. R. Soc. Interface*, 11:186–201, 2014. doi:20131125.20131125.

[25] zFestival. zfestival: A virtual new music and art festival, 2020. URL: https://zfestival.wordpress.com/.