Face the Music: Summarizing Unscripted Music Practice from Audio

Christopner Raphael
Indiana University, Bloomington, U.S.A.

Keywords: Music Practice, Dynamic Programming, Visualization, Switching Kalman Filter, Summarization.

Abstract: We present ongoing work in developing a system to support instrumental practice in which a students plays from a score but can move freely within the score, as is typical of score-based music practice. Our system develops a correspondence between the practice audio and the score, partitioning the audio into a collection of score-aligned excerpts using dynamic programming. We examine several offline approaches to help interpret or summarize the practice audio. One is a tool that allows score-driven browsing of the audio. We also look at several score-based visualization tools that highlight aspects of the practice data. Finally we develop a technique that assembles an "optimal" audio performance from the score-aligned fragments, seeking an assembly that is rhythmically most plausible according to a simple probabilistic model for musical timing.

1 INTRODUCTION

Music instruction systems provide support for instrumental music practice by identifying errors or inaccuracies, providing guidance, and, perhaps, even offering a kind of companionship during music practice. These systems hold promise for making the rewarding, lifelong activity of music-making more broadly available, increasing the individual attention received by music students.

These systems can sense the student's actions with a variety of kinds of data, such as audio, video, MIDI, haptic, etc., though we use audio due to our focus on traditional musical instruments, such as strings, woodwinds, brass, and piano, where other data sources are not available.

Such instruction systems are often organized around a *call and response* paradigm: the practicing student is presented with a series of short score passages which are played by the student and, in turn, evaluated the system. The commercial system *Yousician* (Yousician, 2022) as well as (Fober et al., 2004), (Dannenberg et al., 1990), and (Zhang et al., 2019) are all examples of such *call and response* approaches. Of course, this paradigm is familiar, and often effective, in the larger computer supported education space as well.

The evaluation of the passage allows the instructional system to determine if the presented challenge has been met, thus influencing the choice of the next passage — perhaps a repeat of the most recent one. Thus the overall loop of call, response, and evaluation becomes the basic structure of the system. The *call and response* paradigm simplifies the evaluation process since it is much easier to judge accuracy when the system "knows" what the student intended to play. However, it also decreases the agency of the student, requiring the student to follow the practice regimen presented by the system, rather than allowing a student to direct the practice session. Evidence suggests that taking "ownership" of the the practice session, as well as the learning process more generally, is important for long term success (Coutts, 2019).

Rather than requiring a music student to fit into what is easiest for the computer, we explore an approach that works with the way music students naturally practice. In typical score-based music practice a student plays from a score, jumping around in the score at will. Often short sections are practiced repeatedly, gradually moving forward through the score in an attempt to build fluidity through a larger section. Though, occasionally, the student may skip from one section of the score to another. Sometimes a student may even depart from the score temporarily to practice an exercise derived from a particular passage. It is this "unscripted" score-based practice we address here, assuming only that most of the audio played during the session comes from the score, as is common in a broad range of practice scenarios. We develop an audio recognition technique, a variant of traditional audio score alignment, that maps the audio onto the music score as a sequence of "excerpts" —

Raphael, C.
Face the Music: Summarizing Unscripted Music Practice from Audio.
DOI: 10.5220/0013493800003932

In *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025) - Volume 1*, pages 685-691
ISBN: 978-989-758-746-7; ISSN: 2184-5026

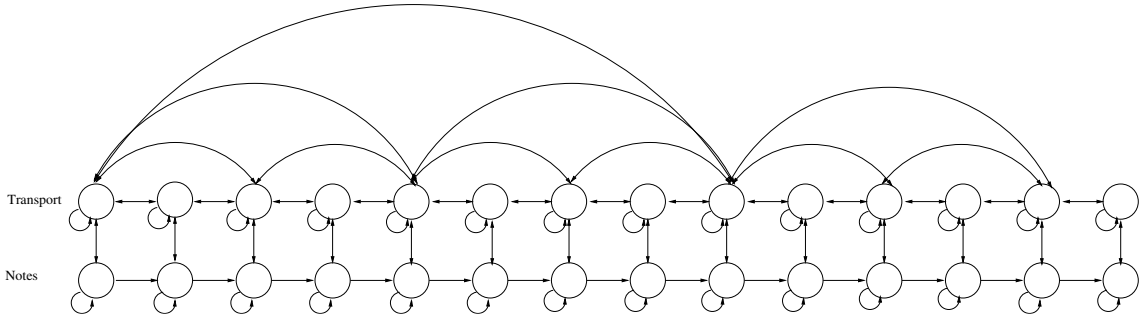

Figure 1: We associated the frame sequence with a path in the score graph, above, thus mapping the audio onto the score.

time-labeled sequences of score notes.

In the current work we limit our attention to the "offline" version of unscripted score based practice in which we analyze a recorded practice session *after* it has been recorded. Offline analysis is the basis for our computer-supported review of music practice, as well as practice summary. In contrast, "online" score following seeks to understand the practice audio *as it is generated*, as would be necessary to offer real-time support. As we seek to build actual systems and get them into the hands of practicing music students, we are interested in both offline and online analysis, however the online aspect is beyond the scope of the current discussion.

We describe the essential methodology behind the offline score alignment in Section 2. Once we have the alignment we seek meaningful ways to browse the practice audio, and summarize what is important for the student to know, presenting our results in accessible visualizations. Given the centrality of the score to much of music practice, and to our statement of score alignment problem, it makes sense to leverage the music score in our student feedback. Section 3 sketches a score-based practice browser that allows the student to explore the practice in an easy and intuitive self-directed manner, as well as several score-based visual summaries of the practice. Section 4 considers the problem of assembling an "optimal" complete performance from the practice session. Section 5 provides some larger context for the current work and discusses developing ideas.

2 OFFLINE SCORE ALIGNMENT

Here we present briefly our approach to audio-score alignment for unscripted score-based music practice.

Score following (online) and score alignment (offline) were simultaneously proposed in 1984 by Dannenberg (Dannenberg, 1984) and Vercoe (Vercoe, 1984). Many variations on score alignment have been proposed, including adaptations for polyphonic music (Hu et al., 2003), audio alignment with images of printed scores (Dorfer et al., 2018), the simultaneous identification tempo or other latent variables (Raphael, 2006), and audio-to-audio matching (Müller et al., 2005). A common methodological theme runs through nearly all of this work: the performance is represented as a path through a state graph in which nodes represent score positions, while dynamic programming (DP) is used to find the best interpretation given the audio data. We use the term "dynamic programming" in a general sense to include both most likely path and filtering computations (as with an HMM). An overview of the score alignment and following can be found in (Dannenberg and Raphael, 2006), while a thorough history of the methodological ideas is given in (Cuvillier, 2016).

Our particular focus is on score alignment that allows the player to "jump around" in the score, as is typical in most realistic practice scenarios. Restrictive versions of this idea were first proposed in (Pardo and Birmingham, 2005) and (Fremery et al., 2010) considering possible jumps to and from important structural boundaries. The scenario involving arbitrary jumps was first proposed by Nakamura and Sagayama (Nakamura et al., 2015), while our modeling framework is similar while (Jiang et al., 2019) compares and contrasts the approaches.

We begin with an audio recording of a music practice session, typically about 20 minutes in length in our experiments. The recording is divided into a sequence of "frames" of about 30 ms. each, $y_1, \ldots, y_T$. Our approach to score alignment views the music score as a sequence of notes, without regard for the notated lengths of these notes. If we are treating a polyphonic instrument, such as the piano, then the score can be regarded as a sequence of chords, with a new chord appearing when any note of any voice changes. We relate the audio to the score by modeling the $y_1, \ldots, y_T$ as a path through the state graph of Figure 1, $x_1, \ldots, x_T$ — one state for each frame of

audio. The graph explicitly models the possibility of skips in the score.

The lower level of the graph, labeled "Notes" in Figure 1, depicts the notes (or chords) of the score, connected in left-to-right order, as indicated by the right arrows that connect this level. Since the number of frames the player spends in each note is unknown, each of these states, and all other states in the graph, contains a self-loop — we can remain in a state for any number of frames. In reality each of the states in the note layer are actually sub-models involving several states, however this is omitted from Figure 1 for simplicity's sake.

The upper layer of Figure 1, labeled "Transport," models the player's score skips: from any note in the score one may move to the Transport layer, moving in this layer to the next score note to be played. The curved arcs in Figure 1 allow transitions in the Transport layer that move several, or many, states to the left or right, thus allowing long score skips spanning many notes in a single frame. A probability distribution over the arcs that exit each state model our preference for transitions, favoring linear motion through the Notes layer over skips, local skips over distant ones, and backward skips over forward skips. Thus, $x_1,\ldots,x_T$ is modeled as a Markov chain.

Our data model computes the probability of a frame of audio given a particular model state, $P(y_t|x_t)$ with details presented in (Jiang et al., 2019). Implicit in the notation is the assumption that the tth frame of audio, y_t, depends only on the current state, x_t. While we omit the details here, if x_t is a note or chord in the Notes layer the probability model depends on the associated pitches of the note or chord. If x_t is in the Transport layer we use a rest model for the audio, (we assume no notes are currently sounding).

The result of these assumptions is a hidden Markov model. We interpret the audio by identifying the most likely sequence of states given the audio data

$$\begin{aligned}\hat{x}_1^T &= \arg\max_{x_1^T} P(x_1^T|y_1^T)\\ &= \arg\max_{x_1^T} P(x_1^T)P(y_1^T|x_1^T)\\ &= \arg\max_{x_1,\ldots,x_T} P(x_1)\prod_{t=2}^{T}P(x_t|x_{t-1})\\ &\quad \prod_{t=1}^{T}P(y_t|x_t)\end{aligned}$$

The dynamic programming computation of the most likely path, $\hat{x}_1^T$ is well known, e.g. (Rabiner, 1989), and is omitted here. The most likely path can be partitioned into intervals, $\hat{x}_{\mathrm{lo}(e)}^{\mathrm{hi}(e)}$, for $e = 1\ldots,E$ that lie completely in the Notes layer separated by intervals that lie completely in the Transport layer. Each $\hat{x}_{\mathrm{lo}(e)}^{\mathrm{hi}(e)}$, becomes an excerpt in our interpretation of the audio, identifying a sequence of score notes that were played as well as the onset frame for each note in the sequence.

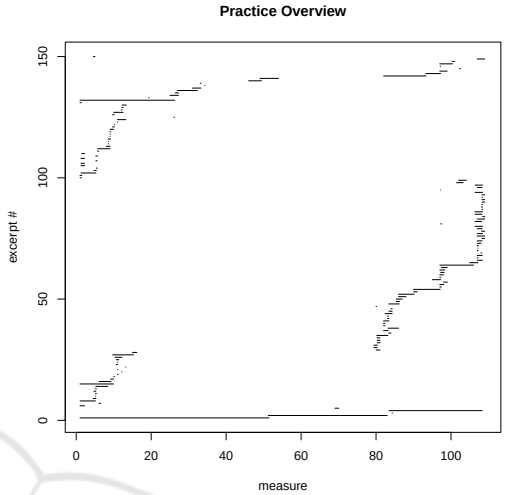

Figure 2: Each horizontal segment shows the range of notes covered by an excerpt for the first 150 excerpts of a practice session. Aspects of the practice strategy are evident in this simple overview.

Figure 2 gives an example of this distillation of the audio into excerpts, showing the first 150 excerpts in a practice session. From this simple analysis one gets an overview of the practice strategy employed by this particular student, beginning by playing through the entire etude in long sections over the first several excerpts, followed by a focus on shorter sections with lots of repetition, gradually moving forward through the current section.

(Jiang et al., 2019) evaluates the accuracy of this approach on a small sample, showing promising results on the test set. However, there is a great deal of variation in practice data, while the most straightforward ways of collecting labeled data are prohibitively time-consuming. Using synthetic data may offer some clarity, though this approach must make assumptions concerning the generation mechanism which are bound to be simplistic. Thus there is more work to be done in validating the accuracy of our proposed method.

3 REVIEWING PRACTICE

A former teacher impressed on his students the importance of "facing the music," by which he meant

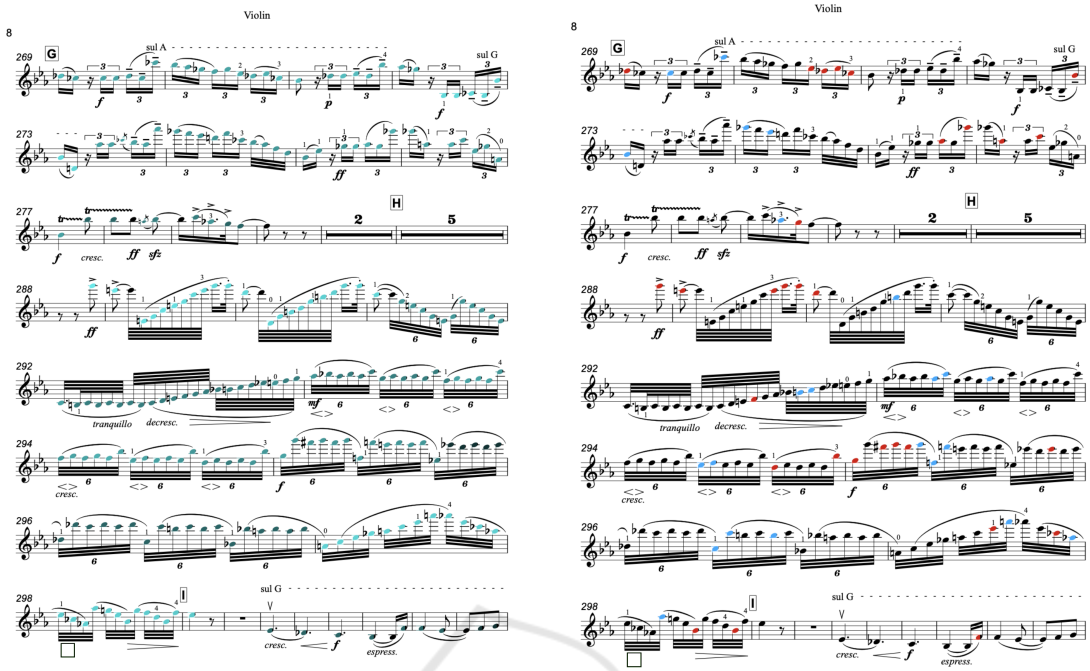

Figure 3: **Left:** The coverage of the rehearsal is shown by tinting the note heads. Brighter notes were played more frequently. **Right:** Tuning expressed by coloring note heads. Red notes are sharp while blue notes are flat in relation to equal A=440 tempered tuning.

listening to recordings of our playing. Such listening is important because we often don't hear ourselves objectively in real-time, (Silveira J. M., 2016). Perhaps this is because listening is a matter of habit — we hear what we listen for, especially when so much cognitive bandwidth is directed toward the mechanics of playing. However, listening to a recording *offline* — *facing the music* — has a way of breaking this attention habit, making clear what we otherwise miss. Our user interface, discussed here, seeks to facilitate this directed listening.

The music score is usually the focal point during score-based practice, thus we also orient our directed listening around the score. Our interface allows one to navigate through the practice session at the *note* level, either using arrow keys to move forward/backward while highlighting the current score note with a box around the note — our "cursor." Playback can be initiated from the cursor position at any time. Alternatively, the user can click on a score note to jump to the associated score and playback position in the recorded audio. In contrast, a traditional audio browser requires the user to search for musical events in order to hear them.

Simply holding down the forward arrow key in our interface advances through the score at the keyboard repeat rate, providing a movie-like summary of the sequence of notes visited during practice, analogous to Figure 2, as shown **here**. A more fine-grained review process is shown **here**, where the user interactively explores the practice clicking on notes that deserve further review. One can imagine a productive exchange between a teacher and student oriented around such an interactive tool, allowing the teacher to observe and explore the student's practice itself, rather than just the final result.

The score also works well as a basis for static visualization, much as a geographical map is an effective reference for spatial data (Hogräfer et al., 2020). The left panel of Figure 3 shows a score page where each note is colored according to how many times it was visited in the practice session. In the image a black note would mean the note was never played while a bright blue was played the most. One can see both regions that were virtually untouched, as well as those receiving special attention.

The right panel of Figure 3 shows an analogous depiction of tuning, where a blue note is one that is flat while a red note is sharp. We make these determinations by estimating the frequency difference between the A=440 equal-tempered target frequency and the average frequency for the note, measured in cents. We highlight the note when the discrepancy is greater than a user-adjustable threshold. The reference tuning level of A=440 can also be adjusted. In creating such tuning maps one can either consider aggregate tuning, averaged over all of the excerpts in the rehearsal, or a more targeted choice of tuning — say

the most recently played notes.

Of course there are other aspects of the rehearsal that can be visualized in this way, such as rhythmic accuracy, while one can combine the note tinting with interactive browsing.

4 ASSEMBLING AN OPTIMAL PERFORMANCE

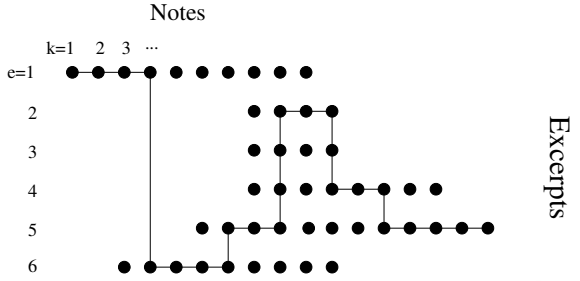

Figure 4: A schematic view of the assembly problem. We seek to link together portions of the excerpts to make the best overall performance.

In addition to having visual summaries of practice data, audio summaries are also helpful. In this section we discuss a method for generating a single "optimal" performance produced by assembling the audio fragments generated during practice. Such a recording could be shared with a teacher as a representative example of the level attained by the student (at her best). Or it may be useful to the student as a distillation of the entire practice session (or several sessions), reducing the data to a manageable quantity.

Our approach for performance assembly is based on rhythm analysis, so, unlike Section 2, we need to consider notated rhythm. We assume that our score is composed of K notes with notated lengths $l_1,\ldots,l_K$ in whole note units — that is, if note k is a quarter note then $l_k = 1/4$, regardless of the time signature. The analysis of this section applies equally well to polyphonic scores, such as with piano practice, in which l_k describes the length of the kth chord, taking a homophonic (sequence of pitch simultaneities) view of the score.

We begin with a collection of excerpts, as in Figure 2, from which we wish to assemble an "ideal" complete performance by piecing together note sequences, schematically depicted in Figure 4. At present we consider only rhythmic fluidity in assembling this performance, though pitch accuracy or other performance elements could easily be included into our formulation. Thus, in short, we seek the path through the excerpt notes of Figure 4, that is most rhythmically fluid, linking these note sequences together to form the optimal performance.

Our measure for rhythmic fluidity is based on a probabilistic model for musical timing that uses a latent tempo process $t_1,\ldots,t_K$ where t_k is the local tempo at note k measured in seconds per whole note. The onset times for the complete performance we will construct are given by $o_1,\ldots,o_K$, measured in seconds. The joint structure of these variables is modeled by a joint Gaussian distribution, modelling the initial tempo, t_1, as $t_1 \sim N(\mu_t,\sigma_1^2)$ where μ_t is the tempo given in the score, expressed in secs per whole note, and letting the initial onset, o_1, have $o_1 \sim N(0,\tau_1^2)$ where τ_1^2 is a nearly infinite variance — we do not care when the sequence begins. The process then evolves according to

$$\begin{aligned} t_{k+1} &= t_k + \varepsilon_k^t \\ o_{k+1} &= o_k + l_k t_k + \varepsilon_k^o \end{aligned}$$

for $k = 1,\ldots,K-1$ where the $\{\varepsilon_k^t\}$ are $N(0,l_k^2\sigma^2)$ variables, the $\{\varepsilon_k^o\}$ are $N(0,l_k^2\tau^2)$, with the variables $\{\varepsilon_k^t,\varepsilon_k^o\}_{k=1}^{K-1}, t_1, o_1$ assumed mutually independent.

The 0-mean and small variances of the $\{\varepsilon_k^t\}$ lead to a smoothly varying tempo process, which is reasonable since we want the tempo to be stable within and between our assembled fragments. In a situation where a known tempo change occurs in the score, we could easily allow a "reset" of the tempo process. The kth note has length $o_{k+1} - o_k$, which, according to our model, has mean length $l_k t_k$ — this is the length that would be predicted purely based on the tempo. However the model allows additional variation in the actual note lengths though the $\{\varepsilon_k^o\}$ variables.

The model defines a joint Gaussian density on all model variables, $P(t_1^K, o_1^K)$. Thus, the plausibility of a sequence of onset times, o_1^K, for the assembled performance could be measured by $\max_{t_1^K} P(t_1^K, o_1^K)$.

We now turn to the problem of constructing the ideal performance. Our analysis of the practice session results in E excerpts, indexed by $e = 1,\ldots,E$, where the eth excerpt covers the range of notes $\mathrm{lo}(e)\ldots,\mathrm{hi}(e)$ as in Figure 2 or Figure 4. For excerpt e we denote the onset time of the kth by $o_{k,e}$, where $\mathrm{lo}(e) \le k \le \mathrm{hi}(e)$. For each note we want to choose one of the possibly many examples observed in the practice session. As notation we let e_k be the excerpt from which we take the kth note. Once we have chosen e_1^K, we can construct a sequence of onset times, o_1^K according to

$$o_1 = o_{1,e_1} \quad (1)$$

$$o_k = o_{k-1} + (o_{k,e_k} - o_{k-1,e_{k-1}}) \quad (2)$$

$k = 1,\ldots,K-1$. For this construction to make sense both o_{k,e_k} and $o_{k-1,e_{k-1}}$ must come from the same excerpt so that their difference measures the length of

the kth note. Thus we require that for each note, k, with $k \neq 1$ we have $\mathrm{lo}(e_k) < k$.

Now, emphasizing the dependence of o_1^K on e_1^K by explicitly writing $o_1^K(e_1^K)$, we could view the most plausible assembly by choosing the e_1^K that maximizes $\max_{t_1^K} P(t_1^K, o_1^K(e_1^K))$. However, we also want to reduce the amount of skipping around between excerpts, so we add a penalty for each such excerpt switch, $L(e_1^K) = C^{|\{k:e_{k+1} \neq e_k\}|}$ for some positive constant C. Then we define our optimal e_1^k, $\hat{e}_1^K$, by

$$\hat{e}_1^K = \arg\max_{e_1^K} L(e_1^K) \max_{t_1^K} P(t_1^K, o_1^K(e_1^K)) \qquad (3)$$

The construction of $\hat{e}_1^K$ poses some interesting methodological challenges that are beyond the scope of our current effort. However, (Raphael, 2006) describes a method for computing the globally optimal sequence of hidden states in a switching state space model containing both 1-dimensional Gaussian and discrete hidden variables. This situation is very close to ours, so the methodology easily adapts to the situation at hand.

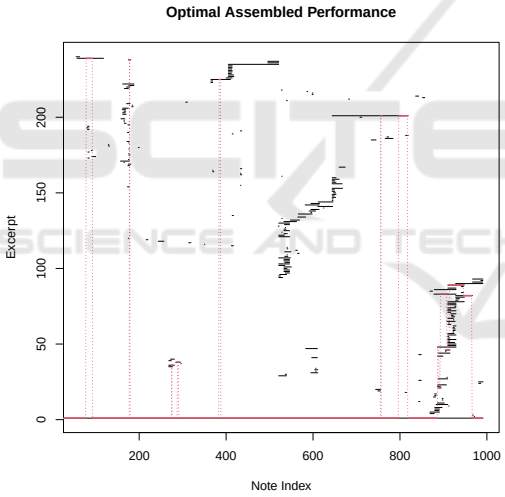

Figure 5: Assembly of a complete performance of the Bruch Violin Concerto No. 1, Mvmt 1 from a practice session.

Figure 5 shows the analog of Figure 4 constructed from a real practice session of the Bruch Violin Concerto No. 1, first movement. The figure shows that the practice session began with a complete run through of the movement, with later practice focusing on specific sections, so a significant portion of the movement was played only once. This is reflected in the optimal assembly, shown in red, that draws heavily from the 1st excerpt, as it must.

A score-aligned video of the result is given **here**. In constructing the audio we have made no effort to cover up the splice points, which can be heard as clicks in the audio or places where the microphone placement seems to suddenly change. While a user-facing result might blend the audio over the splice points, it seems better to leave them in the raw state for purposes of our demonstration.

5 FUTURE WORK

The discussion above gives an overview of the ideas we are developing for audio-based instrumental practice support systems. It is safe to say that the challenges are significant and open-ended, while many remain unmentioned in our discussion. We have presented a number of ideas for practice visualization based on the mapping we construct from the practiced notes to the score. Given the centrality of the score for many practicing students, such score-based visualization seems an obvious and essential component of such a system.

It is worth mentioning, however, that the idea is more general than what we have sketched. Our score alignment technique of Section 2 establishes a many-to-one map between the practice audio and something familiar to the student, the score. However, having mapped the notes to the score, we could reduce further to say, the chromatic pitches that are playable on the instrument. Such a mapping could display, for instance, average tuning for the different notes or registers. This could be particularly useful for instruments, such as woodwinds, that have particular notes or registers that tend to be out of tune. Or, rather than reducing the mapping to a smaller range, one could expand it, looking, for example, at particular *sequences* of pitches. From this analysis we may ask, for instance, if a particular pattern of pitches tend to lack rhythmic fluidity in the practiced audio? We mention these examples to stress that there is a large and largely-unexplored space that may contribute significantly to the challenge at hand.

In addition to the algorithmic, visualization, and modeling challenges, such as those discussed within, we are also interested in building an actual working practice support system, making this available for general use by music students on the familiar app stores. This goal is motivated by the promise shared by nearly all music instruction systems — to help more people appreciate and enjoy music-making, especially those who cannot afford traditional one-on-one music instruction, or do not have access to it.

In addition to the inherent value of such a system, wide distribution would provide a means for collecting score-aligned music practice data from willing contributors, thus supporting a large scale, empirical view of music practice. Such data could be used to

track a student's progress over a large period of time, or could be used as a tool for studying the effectiveness of various practice strategies. The benefits in transforming our approach from a single practice session to a large corpus of practice data could make a lasting contribution to music pedagogy.

REFERENCES

Coutts, L. (2019). Empowering students to take ownership of their learning: Lessons from one piano teacher's experiences with transformative pedagogy. *International Journal of Music Education*, 37(3):493–507.

Cuvillier, P. (2016). *On temporal coherency of probabilistic modesl for audio-to-score alignment*. PhD thesis, Université Pierre-et-Marie-Curie.

Dannenberg, R. and Raphael, C. (2006). Music score alignment and computer accompaniment. *Commun. ACM*, 49:38–43.

Dannenberg, R. B. (1984). An on-line algorithm for real-time accompaniment. In *Proceedings of the 1984 International Computer Music Conference, ICMC 1984, Paris, France, October 19-23, 1984*. Michigan Publishing.

Dannenberg, R. B., Sanchez, M., Joseph, A., Capell, P., Joseph, R., and Saul, R. (1990). An expert system for teaching piano to novices. In *Proceedings of the 1990 International Computer Music Conference, ICMC 1990, Glasgow, Scotland, September 10-15, 1990*. Michigan Publishing.

Dorfer, M., jr., J. H., Arzt, A., Frostel, H., and Widmer, G. (2018). Learning audio–sheet music correspondences for cross-modal retrieval and piece identification. *Transactions of the International Society for Music Information Retrieval*, 1(1):22.

Fober, D., Letz, S., Orlarey, Y., Askenfelt, A., Falkenberg-Hansen, K., and Schoonderwaldt, E. (2004). IMUTUS - an Interactive Music Tuition System. In IRCAM, editor, *Sound and Music Computing Conference*, pages 97–103, Paris, France.

Fremery, C., Müller, M., and Clausen, M. (2010). Handling repeats and jumps in score performance synchronization. In *Proc. of the International Conference on Music Information Retrieval (ISMIR)*.

Hogräfer, M., Heitzler, M., and Schulz, H.-J. (2020). The state of the art in map-like visualization. In *Computer Graphics Forum*, volume 39, pages 647–674. Wiley Online Library.

Hu, N., Dannenberg, R. B., and Tzanetakis, G. (2003). Polyphonic audio matching and alignment for music retrieval. In *in Proc. IEEE WASPAA*, pages 185–188.

Jiang, Y., Ryan, F., Cartledge, D., and Raphael, C. (2019). Offline score alignment for realistic music practice. In *Proc. of the Sound Music and Computing Conference (SMC)*.

Müller, M., Kurth, F., and Clausen, M. (2005). Audio matching via chroma-based statistical features. pages 288–295.

Nakamura, T., Nakamura, E., and Sagayama, S. (2015). Real-time audio-to-score alignment of music performances containing errors and arbitrary repeats and skips. *CoRR*, abs/1512.07748.

Pardo, B. and Birmingham, W. P. (2005). Modeling form for on-line following of musical performances. In *AAAI*.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Raphael, C. (2006). Aligning music audio with symbolic scores using a hybrid graphical model. *Machine Learning*, 65:389–409.

Silveira J. M., . G. R. (2016). The effect of audio recording and playback on self-assessment among middle school instrumental music students. *Psychology of Music*, 44(4):880–892.

Vercoe, B. (1984). The synthetic performer in the context of live performance. In *Proceedings of the 1984 International Computer Music Conference, ICMC 1984, Paris, France, October 19-23, 1984*. Michigan Publishing.

Yousician (2022). Yousician. http://yousician.com.

Zhang, Y., Li, Y., Chin, D., and Xia, G. (2019). Adaptive multimodal music learning via interactive-haptic instrument. *arXiv preprint arXiv:1906.01197*.