COMBINING HAND POSE ESTIMATION WITH AUDIO FOR NOISE-ROBUST PIANO SCORE FOLLOWING

Kaitlin PET (kaitlinpet@gmail.com) [1] and **Akira MAEZAWA** (akira.maezawa@music.yamaha.com) [2]

[1] *Luddy School of Informatics, Computing, and Engineering*, **Indiana University Bloomington**, IN USA
[2] **Yamaha Corpration**, Hamamatsu, Japan

ABSTRACT

Tracking the musical score position from a user's piano performance is difficult in noisy environments and other cases when the audio signal is unreliable. We present a multi-modal score following method that seamlessly fuses audio and hand tracking from top-down video of a pianist's hands on the keyboard, using a probabilistic model and a feature-weighing mechanism. Experiments demonstrate our method's robustness under noisy environments and when pieces contain mistakes, compared to a score follower based only on audio input.

1. INTRODUCTION

Online score following allows a given input music performance to be continually temporally matched to its corresponding symbolic score [1, 2]. Score followers typically use performance audio as the input. They are often formulated as a Hidden Markov Model (HMM) whose states, emission probability, and transition probabilities are derived from the pitches and rhythms specified in the musical score [3–5].

Audio-based score followers become unreliable when the performance audio becomes unintelligible or inaudible, or the user plays unexpected notes that cannot be reasonably captured by the score follower's acoustic model. When audio cannot be relied on for score following, fusing complementary modalities can be helpful. For instance, during loud orchestral sections of a piano concerto, conductors might rely on visual cues from the pianist's body language. In a group music lesson where multiple students play the same piece on the keyboard in the same room simultaneously, a teacher might look at each student's hands and fingers to get an idea of what a student is playing. Audio-based score followers may also perform poorly if the pitches and rhythms being played differ from the underlying score, for example, in cases where the musician makes mistakes. When human ensemble members encounter a collaborator who makes mistakes, they may use a complementary source of information, like knowledge of common errors, to help interpret the performer's intention.

Previous research has shown that multi-modal score followers can mitigate tracking failures under low Signal-to-Noise Ratio (SNR) conditions like the situations described above. But, it is non-trivial to fuse the multiple streams seamlessly. One line of work incorporates additional inputs at a few carefully chosen locations [3, 6] such as before the beginning of a piece or after a *fermata*. Another line of work uses visual data as a complementary time series and chooses a media stream to track based on an estimate of reliability. For example, IMuSE automatic accompaniment system [7] uses multiple score followers operating on different modalities, including audio and the center of mass position of a pianist's left/right hands. During a performance, it uses tracking information from whichever follower has the smoothest series of inferred note onsets.

In general, the advent of deep learning-based pose frameworks such as MediaPipe [8] has made it easier than ever to accurately incorporate gestural information from a video feed into tasks in music informatics, such as multi-modal piano transcription [9–11]. Given such progress, it is increasingly important to explore its potential in score following and formulate a principled and simple mechanism for media fusion.

We present a multi-modal score follower that combines video and audio inputs for classical piano score following, leveraging hand and individual finger positions obtained from deep learning-based pose estimation alongside audio data. The fused score follower learns the expected audio and visual trajectory from a combination of the musical score and a "rehearsal" video of a pianist playing the music, then can track the pianist's hand and finger positions in a second, "performance" take. To integrate the audio and video streams seamlessly, we introduce a feature weighting mechanism to adjust the importance of each modality dynamically, enabling multi-modal real-time score following without the need for manual selection of modality. We show in two different environments that the fused score follower (1) can be tuned to largely retain the accuracy of an audio-only score follower when the audio signal is clear, (2) maintains a reasonable accuracy even when audio background noise levels are raised to the point that the piano playing is inaudible to the human ear, (3) outperforms an audio-only score follower when the pianist makes mistakes.

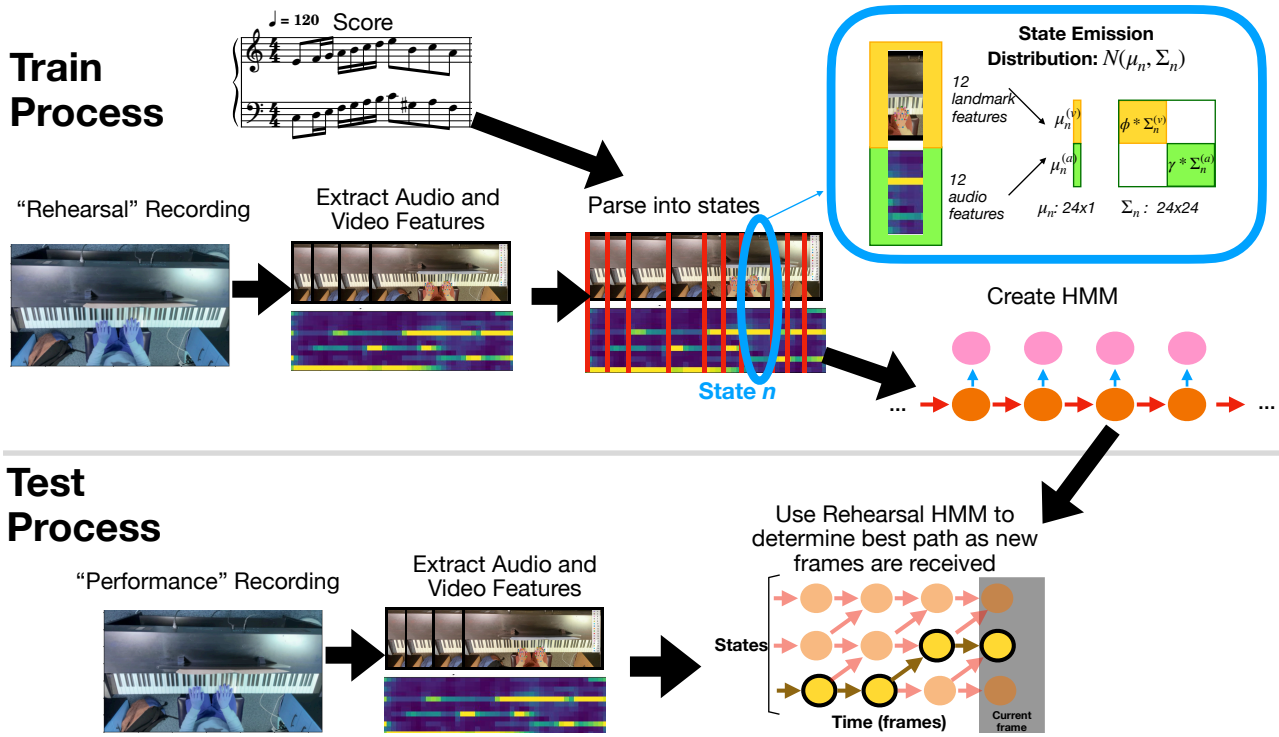

Figure 1. Overview of our method's training process (top), and test process (bottom).

2. METHODS

Our method matches a reference ("rehearsal") piano performance to a second performance of the same piece in real-time, as seen in Fig. 1. Tracking can rely on pitch-based features and hand landmark-based features. First, the "rehearsal take" of a piece is used to build the states of an HMM and set its emission and transition probabilities. To track the second "performance take" in an online manner, we use the rehearsal HMM to decode a state sequence, where state x_t in the decoded sequence is the state with the highest filtered probability at time t. Note that this method of online state assignment does not ensure a linear progression through the score – in the decoded state sequence, states are not guaranteed to appear in score-order and may jump around. This could make our state decoding method less-than-ideal for a real-world score following application. However, we find this method is sufficient for comparative evaluation of different score following strategies.

2.1 Hand landmark feature extraction

From the video stream, we extract a 12-dimensional position feature corresponding to the locations of 1) fingertips and 2) left hand/right hand center of mass, along the axis of the span of the keyboard. MediaPipe's Hand Landmarker [8] is first used to identify 21 "landmark locations" on each hand in every frame of the input video. Fig. 2 shows examples of labeled landmarks. In cases where a landmark is not detected, or a hand is absent from the frame, the missing landmark positions are set to their last observed positions.

The raw landmark positions are then adjusted using an affine transform to align with a pre-selected piano template. Before proceeding with further processing, alignment is visually confirmed for each recording. The transform is initialized based on the assumption that the middle finger of each hand would be positioned near the border between black and white keys. This initial transform is subsequently refined to maximize the enhanced correlation coefficient of the two images. This process ensures transformed coordinates are reliable and consistent for both takes, particularly along the horizontal axis, assuming that the middle fingers in the reference image were in the expected locations.

The Hand Landmarker model provides x, y, and z coordinates for each landmark point. Following the warping process, x denotes the horizontal position on the piano keyboard, y denotes the vertical position of the finger on the keyboard, and z denotes the finger depth with the wrist at $z = 0$, representing the extent to which the finger was raised or depressing a key. Preliminary analysis found that the trajectory of the x-coordinate exhibited the highest similarity across different takes of the same piece. This finding is logical since the horizontal position of the fingers directly corresponds to the notes being played. Although the trajectories of y coordinates showed moderate similarity, technical challenges during preprocessing led us to focus solely on the x coordinates for our landmark-based score follower.

Subsequently, we refine the landmark features by computing each hand's center-of-mass and retaining only the locations of the fingertips, yielding a total of 12 hand position features across both hands. These features were chosen because fingertips are directly involved in producing piano sounds, and hand center-of-mass was shown to be

Figure 2. Frames of top-down performance videos with MediaPipe landmarks colored. The top frame is from *no-mistake* dataset; the bottom frame is from *mistake* dataset

an effective tracking feature in Ritter et al. [7].

2.2 Multi-modal HMM for Score Following

The purpose of this study is to evaluate whether a fused audio-visual HMM-based score follower performs better than an audio-only HMM-based score follower. Thus, we opt for a simple, prescriptive approach in assigning states and transition probabilities. We focus more on the method to train the different emission probabilities, as these differ for visual and audio features.

We represent the score by a sequence of n hidden states, $\{x_1 \dots x_n\}$. These hidden states are matched to m frames of observed features $\{y_1 \dots y_m\}$. The features of y could consist of **(1) audio features** only, **(2) hand landmark features** only, or **(3) a combination** of the two.

States of the HMM are created by partitioning the piece's score by the union of note onsets and running sextuplets (1/6 beat units). For example, say a measure in the score has notes of length [2 beats, 1 beat, 1/12 beat, 1/12 beat]. This measure would translate to 20 total states, with 18 states corresponding to a 1/6 beat unit, and 2 states corresponding to a 1/12 beat unit.

Transitions from the current to the next state depend on the ratio of their score-based lengths. Formally, the probability of staying in state x_n is $\frac{l_n}{l_n+l_{n+1}}$, and the probability of moving from state x_n to state x_{n+1} is $\frac{l_{n+1}}{l_n+l_{n+1}}$, where l_n is the length (in beats) corresponding to the note value used to create state x_n. In the above example, l_n would correspond to 1/6 or 1/12. Backward transitions or skips have a zero probability.

Emission probability for each state x_n is a multivariate normal distribution constructed from 1) x_n's audio emission distribution, $N(\mu_n^{(a)}, \Sigma_n^{(a)})$, and 2) x_n's landmark emission distribution, $N(\mu_n^{(v)}, \Sigma_n^{(v)})$. We assume audio and landmark features are independent. Therefore, we can form μ_n by concatenating $\mu_n^{(a)}$ and $\mu_n^{(v)}$. We can form Σ_n by combining $\Sigma_n^{(a)}$ and $\Sigma_n^{(v)}$ in a blocked diagonal structure. To adjust the audio and visual features' relative contribution to the combined score follower, we introduce a video variance penalty parameter ϕ and an audio variance penalty parameter γ such that Σ_n is given as $\text{diag}([\gamma\Sigma_n^{(a)}, \phi\Sigma_n^{(v)}])$. See Figure 1 for a visual representation of this fusion method.

To create the audio-based emission distribution $N(\mu_n^{(a)}, \Sigma_n^{(a)})$, a "score reference audio" is first generated by rendering the symbolic score with the "Steinway Piano" virtual instrument in Logic Pro X [12]. The audio is transformed into a 12-feature constant-Q chromagram measuring the observed amount of each pitch class in the Western scale. Audio is then normalized such that each frame sums to one. MIDI from the rehearsal take is similarly translated to "rehearsal audio" in Logic Pro X, then transformed into a normalized chromagram. Next, the score reference chromagram and the rehearsal take chromagram must be parsed into states. Since our state definition is score-based, this means we first need a score match to the audio. In the score reference audio, state boundaries can be automatically computed by considering the tempo used to render the audio – since the score is translated literally, the "onset time" of each state is proportional to its position in the score. For the rehearsal audio, we first perform an offline score match using the *Orchestra* program [13]. Rehearsal audio may contain wrong notes, omitted notes, or additional notes (e.g. the player could play a short sequence of notes several times until it was error-free). These cases require manual correction to the alignment. If a note p is played at the wrong pitch, p's onset time is still set to the beginning of the intended note. If p is omitted, a note onset location for p is extrapolated based on the parsed note onset locations of notes $p+1$ and $p-1$. If a sequence of k notes corresponding to score positions $\{p...p+k\}$ is repeated, the note onsets times of $\{p...p+k\}$ are assigned to either the first or last repetition. After obtaining a score match, state onset times are determined by further segmenting the score via running sextuplets, as described above. For each state, $\mu_n^{(a)}$ is set to the sample mean of the state from the score reference chromagram – this way, expected pitch content would not be "polluted" by wrong, omitted, or repeated notes found in the rehearsal take. $\Sigma_n^{(a)}$ was computed by obtaining the sample covariance of frames in both the score reference chromagram and the rehearsal chromagram. Based on our state partitioning method, certain states may have insufficient unique frames to compute sample variance. We thus manually select a minimum covariance. This minimum covariance is set as the emission covariance should the state be too short for covariance computation; otherwise, this minimum covariance is added to the computed covariance values.

To create the landmark-based emission distribution $N(\mu_n^{(v)}, \Sigma_n^{(v)})$, we first obtain landmark trajectories for both hands' centers of mass and all fingertips using the pro-

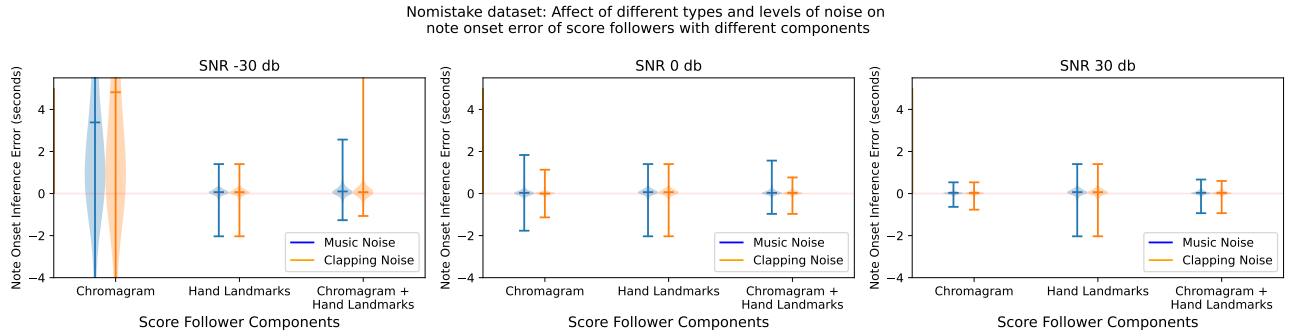

Figure 3. Violin plots showing the median, range and distribution of raw note onset error among three types of score follower (audio-only, landmark-only, and combined audio+landmark) on the *no-mistake* test set, given different noise types and SNR. Positive error values indicate the inferred onset time is late, negative error values indicate the inferred onset time is early. Violin plots show the full range of onset errors, as well as highlighting the median. Plots show behavior at chromagram weight $\gamma = 10$

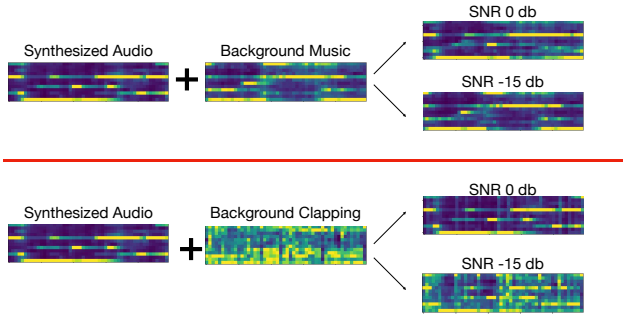

Figure 4. Two types of noise, music (a recording of Beethoven's *Symphony No. 5*) and clapping, are added in varying amounts to the synthesized piano. At higher SNR (*e.g.*, 0 dB), the original signal is still clearly visible, while at lower SNR (*e.g.*, -15 dB), the original signal is much harder to see and hear.

cedure described in Section 2.1. The trajectory sequences are then parsed into states using the state boundaries derived from the rehearsal audio. $\mu_n^{(v)}$ is set to the sample mean of landmark features in state n, and $\Sigma_n^{(v)}$ is based on sample covariance. In this computation, we assume three groups of independent features: the right hand's center of mass, the left hand's center of mass, and fingertips. A similar minimum covariance construct is used to ensure $\Sigma_n^{(v)}$ can is defined for all states.

3. EXPERIMENTS

We evaluated the effect of incorporating audio and visual features in our score follower, depending on the audio variance penalty γ. We also tested the performance of audio-only, landmark-only, and fused score followers when performances contained mistakes or varying amounts of background audio noise.

3.1 Data Collection

We obtained rehearsal and performance takes of multimodal piano performance data (two takes of each piece), capturing performance MIDI and top-down video. Data collection was performed with two pianists at different institutions (Yamaha and Indiana University Bloomington). See Figure 2 for sample frames from the two recording sessions. For the Yamaha dataset, Burgmüller's 25 etudes for piano were recorded by a late-intermediate to early-advanced pianist on a digital piano (Yamaha P125). The camera was suspended above the piano to give a top-down view covering most of the piano's width. As can be seen in Figure 2, a few of the leftmost keys were outside the camera's view. The player was not given specific performance instructions and was allowed to make mistakes. For example, when the pianist did not play a phrase to her satisfaction, she would sometimes repeat the phrase one or more times before continuing on with the performance. Of the 25 etudes, twelve were retained for analysis [1]. Six etudes were used for parameter tuning as described in Section 3.2, and six were used a validation set. **We will henceforth refer to the validation set from this recording session as the *mistake* dataset.**

The second dataset was collected at Indiana University Bloomington on a Roland electric piano (study done under the approval of Indiana University IRB #22017). This dataset was recorded to create an error-free version of pieces from the *mistake* dataset. In these recordings, the musician was explicitly asked to 1) not make mistakes and 2) play both rehearsal and performance takes with the same fingerings and hand distribution. We also positioned the camera in the second dataset to capture a top-down view. However, the camera in the second dataset was placed slightly higher so that all piano keys would be in-frame. **We will henceforth refer to data recorded in this session as the *no-mistake* dataset.**

For all performance takes, audio was synthesized from performance MIDI using the "Steinway Piano" virtual in-

[1] Two etudes contained too many mistakes, two contained pickup notes that were incompatible with the pipeline, and nine could not be effectively pre-processed for other reasons.

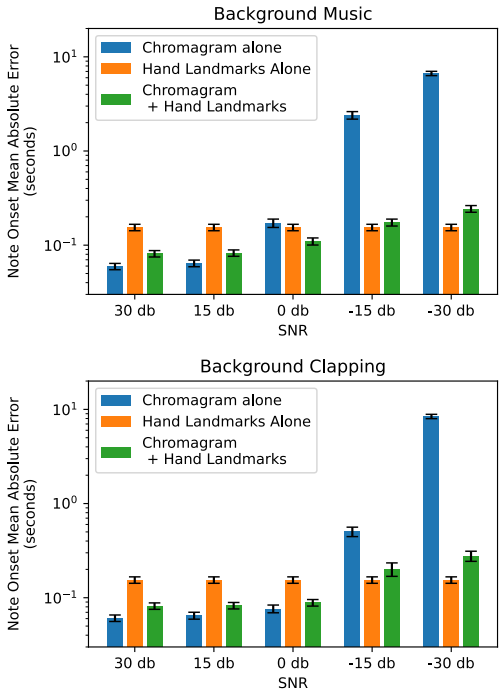

Figure 5. Highlighting behavior at the chromagram penalty level (γ) of 10. These graphs show the mean absolute note onset error of different score followers across a range of SNRs. Error bars represent +/- two standard errors.

strument in Logic Pro X [12]. To obtain ground truth onset labels, performance data (rendered audio and video recordings) were score-matched – *i.e.*, for each etude, all notes in its symbolic score were matched to a unique timestamp in the performance. This offline score match was achieved with a combination of custom code, the *Orchestra* program [13], and hand corrections, yielding a labeling resolution of 33 milliseconds. The *mistake* dataset lacked a one-to-one relationship between note onsets specified in the score and actual onsets in the performance, thus ground truth note onset labeling was performed using the onset labeling method described in Section 2.2.

3.2 Parameter Tuning

Six of the twelve etudes from the *mistake* dataset were used as a "tuning" dataset to hand-tune (1) ϕ, the penalty associated with the hand landmarks; (2) the minimum covariance for each data source; and (3) a good score-to-state partitioning strategy (See Section 2.2 for details). These parameters were used in creating the score following HMM trained by each rehearsal take. The same rehearsal-training pipeline described in Section 2.2 was used for both the *mistake* and *no-mistake* datasets.

Note that in the tuning set, the performer would sometimes play wrong notes, omit notes, or repeat a sequence of notes several times until she was ready to move on. These types of mistakes involve pressing keys with low horizontal distance to their correct counterparts (e.g. a wrong note can happen as the result of accidentally hitting the adjacent black or white key). Thus, picking $\phi > 1$ could allow the emission probability of a physically close mistaken note on a correct state to still be fairly high.

ϕ was chosen such that a score follower using only the landmark features worked reasonably well for every piece in a tuning dataset with mistakes. $\phi = 100$ was used in all experiments below, while γ, the audio variance penalty, varied depending on the experiment.

3.3 Experimental Conditions

Score followers were created to track three types of input data: (1) audio alone, (2) hand landmarks alone, and (3) combined audio and hand features. For each piece in the validation set, the rehearsal take was used to train three HMM-based score followers corresponding to these three types of input data. Each score follower was tested by feeding it audio and/or video of the performance take in 30-millisecond frames. After receiving each new frame, the score follower HMM was used to compute the state with the maximum probability in the current forward posterior distribution.

To evaluate the effectiveness of each score follower, we computed note onset error as the difference between the observed note onset times in the performance take and the inferred note onset times from the HMM. Given our online state determination method, certain states corresponding with note onsets may be skipped or repeated. We thus computed the observed time of each note onset state x_o as

$$\text{obs}(x_o) = \max\{\max(l(x < x_o)), \min(l(x \geq x_o))\}$$

where $l(x)$ is the frame in the test sequence where state x is observed. This definition ensures an onset time exists for every note. Below, we describe the effect on different performance conditions and score followers on onset error.

3.3.1 Effect of Audio Variance Penalty

As can be seen in Figure 6, our γ penalty was successful in weighing the behavior of the audio vs. video input streams in our fused score follower. We can see that at low γ values, the fused score follower more closely tracked the audio-only score follower's behavior, while at larger γ values, the combined score follower's behavior was more similar to the landmark-only one. Note that aside from weighing the audio and landmarks' relative strengths, γ also directly affected the emission variance of states in the audio-only score follower, and by extension, the audio-only score follower's note onset accuracy. When $\gamma < 1$, the audio-feature variance shrinks, potentially overfitting to the expected audio content and reducing accuracy when the performance differed from the rehearsal. If γ is increased too much, it can become difficult to distinguish between emission distributions of different score-based states. Notably, at the γ where audio score follower error is lowest, the fused score follower tends to also be at or near its lowest value. In Figure 6 we highlight a potential γ "sweet spot," $5 \leq \gamma \leq 20$, where the fused score follower's behavior is close to or better than whichever single-source score follower has higher accuracy.

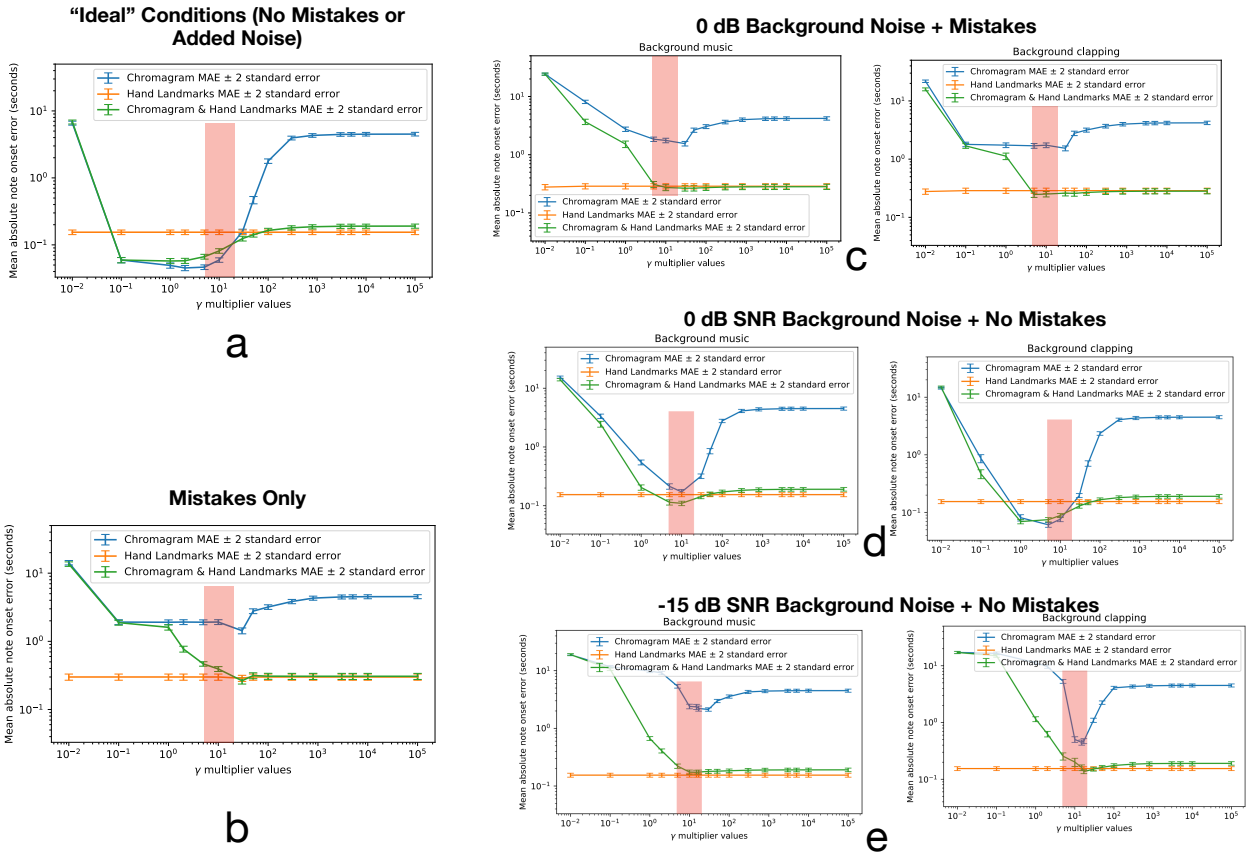

Figure 6. Highlighting behavior of varying γ in the audio-only and fused score followers, under different noise and mistake conditions. Error bars are +- 2 standard errors of the mean. In general, the combined score follower (green) exhibits more chromagram-like behavior at low penalty values and more landmark-like behavior at high penalty values. A potential "sweet spot" for γ is highlighted in red – using this range of γ values, the fused score follower always behaves closer to the better single-source score follower or outperforms both single-source score followers, despite differing amounts of background noise and the presence/absence of performer mistakes.

3.3.2 Effect of Noise on Score Following

We evaluated the score followers as different noises were added at different levels to the piano performance audio. We prepared two kinds of noise signals: (1) background music audio [14] and (2) background clapping audio. Both noises were tested at a signal-to-noise ratio (SNR) of -30 dB to 30 dB in 15 dB increments. Clapping audio was meant to simulate clapping during a live show, while music audio was intended to simulate a "backstage" or "practice room" environment where alternate music sources cannot be isolated from the input audio. Figure 4 visualizes the effect of different types and amounts of background noise with piano performance audio.

A violin plot of absolute error pieces in the *no-mistake* dataset with $\gamma = 10$ is shown in Figure 3, and the mean absolute onset error in Figure 5. Figure 6d and Figure 6e highlight behavior at 0 dB and 15 dB SNR across different γ values. Figure 5 shows how while the audio-only score follower had the best performance at ≥ 15 dB SNR, it became significantly worse than the other score followers when audio quality degraded to ≤ -15 dB SNR. As SNR decreases, performance with background music degraded more quickly than performance with background clapping. This is expected because another music signal is more likely to confuse the pitch-based audio score follower by giving a high emission probability in incorrect score states. The landmark-only score follower had an (invariant) mean absolute error of $\sim$150 milliseconds.

These results show that when the noise level varies or is unexpected, it is important to fuse the streams to take advantage of the more useful modality. In most cases, the fused score follower had a note onset error between the audio-only and landmark-only score followers. At $\gamma = 10$, it tended to perform more similarly to the source with lower error. In high-noise situations, the fused score follower performed more similarly to the landmark-only score follower, while in low-noise situations, the fused score follower more closely tracked the behavior of the audio score follower. Occasionally, the fused score follower performed better than either the audio or landmark score followers alone (for example, 0 dB SNR background music).

3.3.3 Mistake vs. No Mistake Conditions

The behavior of the different score followers on the *mistake* versus *no-mistake* datasets can be observed by comparing Figure 6b with Figure 6a. The presence of mistakes severely degrades the performance of the audio-only score follower. Without mistakes, the audio score follower was able to achieve mean absolute error of below 50 milliseconds, a level of error potentially tolerated within certain real-world score following applications. With mistakes, the average error never fell below 10 seconds – this is excessive error for any real-world score following application. Our landmark-only score follower and combined score follower showed much better resistance to performer error than the audio-only score follower. The landmark-only score follower performed comparably with and without performer mistakes – without mistakes, the average error was $\sim$150 milliseconds, while with mistakes, the average error was only raised to $\sim$290 milliseconds. The mistake-resistance shown by the landmark-only score follower is likely a result of how we approached determining ϕ from the tuning set as described in Section 3.2. Figure 6c shows how this mistake-resistance is retained even in the presence of background music or background clapping.

4. CONCLUSION

Most score followers currently rely on audio as their sole input source. Our work highlights the fragility of this strategy. In "good" audio conditions, *i.e.*, when background noise is low or the performer executes a take without extensive errors, the performance of an audio-only score follower is hard to beat. However, disruptions like background noise and mistakes can quickly derail a score follower or cause catastrophic failure if they are not specifically accounted for in the model.

We presented a multi-modal score following system that fuses deep learning-based pose features with audio. We demonstrated how this system mitigated the effect of noise and mistakes, and could be configured to harnesses the low error the audio-based score follower under "good" audio conditions and the noise/mistake robustness of a hand landmark-based score follower under "bad" audio conditions. We further showed that our hand tracking method is robust enough to be used in different video conditions – the combination of a generic hand-landmarker model and rehearsal-based training allowed successful tracking in recordings made in separate spaces, with different performers and pianos. We believe our work expands the applicability of piano score following in noisy environments, such as practice rooms or music genres for which cheering or clapping during a performance is the norm.

In the future, we hope to continue investigating how this system performs in less ideal lighting conditions that more closely simulate a live concert scenario. We also hope to investigate methods that allow for training the landmark-based features via piano fingering annotation [15], instead of a rehearsal take.

Acknowledgments

We thank Christopher Raphael for his helpful insights and letting us use his *Orchestra* program for note onset labeling.

5. REFERENCES

[1] R. Dannenberg, "An on-line algorithm for real-time accompaniment," in *Proceedings of the International Conference on Computer Music (ICMC)*, 1984, pp. 193–198.

[2] B. Vercoe, "The synthetic performer in the context of live performance," in *Proceedings of the International Conference on Computer Music (ICMC)*, 1984, pp. 199–200.

[3] A. Maezawa and K. Yamamoto, "MuEns: A Multimodal Human-Machine Music Ensemble for Live Concert Performance," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Denver Colorado USA: ACM, May 2017, pp. 4290–4301. [Online]. Available: https://dl.acm.org/doi/10.1145/3025453.3025505

[4] A. Cont, "ANTESCOFO: Anticipatory Synchronization and Control of Interactive Parameters in Computer Music." in *International Computer Music Conference (ICMC)*, Belfast, Ireland, Aug. 2008, pp. 33–40. [Online]. Available: https://hal.inria.fr/hal-00694803

[5] C. Raphael, "Music Plus One and Machine Learning," in *Proceedings of the Twenty-Seventh International Conference on Machine Learning*, 2010, pp. 21–28.

[6] T. Mizumoto, A. Lim, T. Otsuka, K. Nakadai, T. Takahashi, T. Ogata, and H. Okuno, "Integration of flutist gesture recognition and beat tracking for human-robot ensemble," *Proc of IEEE/RSJ-2010 Workshop on Robots and Musical Expression*, 11 2010.

[7] M. Ritter, K. Hamel, and B. Pritchard, "Integrated multimodal score-following environment," in *International Computer Music Conference*, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:799716

[8] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," 2019.

[9] Y. L. Wan, Z. G. Wu, R. H. Zhou, and Y. H. Yan, "Automatic transcription of piano music using audio-vision fusion," in *Measurement Technology and Engineering Researches in Industry*, ser. Applied Mechanics and Materials, vol. 333. Trans Tech Publications Ltd, 9 2013, pp. 742–748.

[10] X. Wang, W. Xu, J. Liu, W. Yang, and W. Cheng, "An audio-visual fusion piano transcription approach based

on strategy," in *2021 24th International Conference on Digital Audio Effects (DAFx)*, 2021, pp. 308–315.

[11] J. Lee, B. Doosti, Y. Gu, D. Cartledge, D. Crandall, and C. Raphael, "Observing pianist accuracy and form with computer vision," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 1505–1513.

[12] "Logic Pro X." [Online]. Available: https://support.apple.com/guide/logicpro/welcome/mac

[13] C. Raphael, "A hybrid graphical model for aligning polyphonic audio with musical scores," in *International Society for Music Information Retrieval Conference*, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:1949303

[14] L. v. Beethoven, "Symphony No. 5 in C Minor," IMSLP, 2002, Fulda Symphonic Orchestra.

[15] W. Gao, S. Zhang, N. Zhang, X. Xiong, Z. Shi, and K. Sun, "Generating fingerings for piano music with model-based reinforcement learning," *Applied Sciences*, vol. 13, no. 20, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/20/11321