

PREDICTING FLUTIST ONSET TIMING IN DUET PERFORMANCE: A MULTIMODAL ANALYSIS OF GESTURE AND BREATH CUES

Jaeran Choi Taegyun Kwon Juhan Nam
Graduate School of Culture Technology, KAIST, South Korea
{jaeran.choi, ilcobo2, juhan.nam}@kaist.ac.kr

ABSTRACT

In ensemble performances, musicians use gesture and breath cues to synchronize their initial notes at the beginning of a piece, but the precise relationship between these cues and onset timing remains under-explored. This study investigates how flutists' gesture and breath cues encode the timing information for the initial note onset. This research consists of four components: (1) Collection of a cue dataset containing synchronized video and audio recordings of flute-piano duets, (2) Identification of cue candidate points through facial movement curves and breath onset-offset analysis, (3) Verification of predicted onset accuracy using linear regression on these cues compared to human onset asynchronies and (4) Introduction and exploration of a 'trigger' concept, defined as immediate, clearly perceivable gestures (such as stopping or raising the head) indicating the precise moment of onset. Our findings suggest a dual-cue system: preparatory cues broadly predict onset timing, while precise triggers refine the exact onset. We compared the time difference between the predicted and piano onsets with the flute-piano asynchronies and verified the concepts of cue and trigger through expert interviews. This research contributes to a deeper understanding of the complex phenomena of musical cues during performance through multimodal analysis. This paper provides an open-access cue dataset, which can be found on the accompanying website.¹

1. INTRODUCTION

Music performance is inherently multimodal, combining sound and motion. Although these elements primarily convey musical expression to audiences [1], they also play a critical role in ensemble synchronization among performers [2]. Performers often employ specific gestures and breathing sounds as musical cues to synchronize their note onsets, particularly at the beginning or during critical moments in a performance. The convention of cueing approx-

imately one beat before the musical onset has been extensively documented in previous studies [3–5]. Head nodding gestures, in particular, are commonly used as intuitive musical cues, and previous studies have utilized such gestures to define cue timings, even extending their application to interactions with robotic musicians [5–7]. Therefore, understanding musical cues not only deepens synchronization knowledge but is also crucial for designing interactive performance systems.

Previous studies have examined the role of gestures and breathing in performer synchronization. Bishop et al. showed that performers use visual and auditory cues to synchronize after silence or rests [8]. Additionally, gestures at the beginning of a piece were found to correlate with tempo, particularly through falling acceleration curves that typically reach their midpoint approximately one beat before the onset [9]. However, this midpoint timing was often not precisely one beat ahead, and exact onset prediction based on these curves was not explored. Vera et al. demonstrated increased onset asynchrony when performers restarted together after rests without visual contact, and identified relationships between breath onset-offset timings and rest durations, yet did not clarify how gestures or breathing specifically encode onset timing [10]. Although these studies highlight the significance of gestures and breath cues in synchronization, they have not thoroughly investigated how combined visual and auditory cues precisely predict intended onset timing, especially at the beginning of a piece.

One reason for the limited quantitative analyses in previous studies is the difficulty of accurately tracking gestures. Bishop et al. utilized Kinect sensors and accelerometers to measure motion curves [9], while Timmers et al. employed infrared markers to measure bow velocity in a string quartet setting [11]. In contrast, our approach uses sensorless image processing techniques, applying optical flow methods [12] to quantitatively track facial gesture movements.

This paper investigates how a flutist's intended onset timing is encoded through gesture and breath cues, comprising four main components: (1) Collection of synchronized video and audio data from flute-piano duet performances involving 20 flutists (Total 1,320 trials), (2) Extraction and annotation of gesture cue features using face movement curves analyzed by optical flow, alongside manual annotation of breath onset and offset timings, (3) Verification of predictive accuracy for cue-based onset pre-

¹ https://github.com/jaeranchoi/flutist_cue_dataset



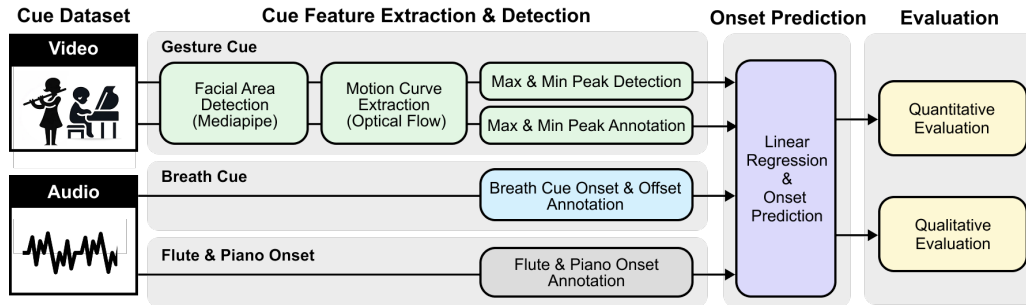


Figure 1: The overall framework for the musical cue detection and onset timing prediction

dictions by comparing linear regression–derived timings to observed human onset asynchronies and (4) Exploration of a trigger concept, which involves three types of immediately clearly perceivable movements: quickly raising the head, stopping head movement after the cue or slowly raising the head. The first type just before fully raising the head, the second triggers onset immediately after stopping, and the third provides a less clear signal. These trigger movements allow performers to precisely detect the onset moment for accurate synchronization.

Our analysis demonstrates clear relationships between the lengths of gesture and breath cues and onset timings. Gesture cues showed a linear relationship between position, velocity and acceleration peaks in vertical facial movements and subsequent onset timing. Similarly, breath cues exhibited a correlation between breath duration and the timing interval to the note onset, with variations observed across different tempos. We further verified our hypotheses through expert interviews. Additionally, we conducted a case study to address related phenomena, including adaptation effects—reduced discrepancies through repeated rehearsal—and instances of cue execution failures, where significant differences between cue-based predicted onset and actual onset were observed. Based on these observations, we conclude that musicians utilize two primary synchronization strategies: rough timing indication through gesture and breath cues, and precise, immediate signals through triggers. Furthermore, exact synchronization is refined through repetitive rehearsals.

2. RELATED WORKS

2.1 Musical Cue for Synchronization

Musical cues, essential for performer communication, include visual gestures and non-musical elements like breathing. They are particularly valuable for precise coordination, such as in pieces with abrupt tempo changes [13], or synchronization after rests and tempo variations [8]. The beginning of a musical piece is challenging for coordination due to the absence of preceding audio cues. Bishop et al. examined visual cues at piece initiation, finding that the peak of the acceleration curve in nodding gestures indicated beat positions, while gesture duration and periodicity conveyed tempo information [9]. However, they did not investigate the predictive capability of gesture cues for onset timing prediction.

Most studies on synchronization between gestures and

musical rhythm have emphasized velocity peaks rather than spatial positions as primary features. Su [14] demonstrated this using minimal laboratory setups with bouncing point-light and auditory stimuli. Similar findings emerged from string quartet studies linking bow speed to tempo cues [11], and conductor studies emphasizing baton velocity and acceleration [15]. Vera et al. further highlighted breath cues’ significance in synchronization, particularly when visual contact is limited [10]. These studies underline the role of gesture and breath cues in synchronization, suggesting velocity, acceleration, and cue length influence timing. However, few studies have simultaneously examined those cues in wind instruments to assess their impact on timing. Our study extends this by quantitatively examining correlations between gestures, breath cues, and onset timings through multimodal analysis of flute-piano duets.

2.2 Gesture Analysis of Performance

Motion tracking methods utilizing sensors or optical flow-based video tracking [16] are common for gesture analysis. Previous studies by Bishop et al. and Timmers et al. used attachable sensors or markers to measure movement and acceleration [9, 11], whereas Bochen et al. applied audio-visual analysis with optical flow to examine vibrato patterns from string players’ hand movements [17, 18]. Maezawa et al. proposed MuEns, a multimodal score-following system employing optical flow to track gestures for automated piano accompaniment. However, in that particular work, the authors utilized arbitrarily defined gesture cues instead of systematically analyzing how performers naturally encode onset timings [5].

3. DATASET

3.1 Musical Cue Dataset

We created a multimodal cue dataset containing video and audio recordings of flute-piano duet performances to analyze musical cues from gestures and breathing sounds. Following previous studies [9, 14], we identified peaks in position, velocity, and acceleration curves as potential gesture cue points. Breath cue points were defined by the breath onset and offset timings. We termed the interval between paired cue points as ‘cue length’ and the duration from cue initiation to flute onset as ‘cue-onset length’.

3.1.1 Participants

A total of 20 professional flutists and 3 pianists participated in the experiment, all holding bachelor’s or master’s

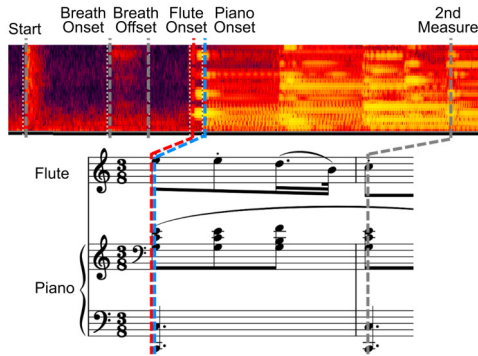


Figure 2: Breath Cue and Onset Annotation Breath cue is defined by the breath onset and offset annotated on the mel spectrogram. Six markers are annotated per trial. Detailed information can be found in Section 3.2

173 degrees in performance. Each flutist-pianist pair had not
174 previously performed together.

175 3.1.2 Placement and Equipment

176 The setup mimicked a concert stage, with flutists posi-
177 tioned facing away from the pianists. The pianists could
178 observe the flutists, while the flutists were instructed to
179 give cues without turning or looking at the pianists. A
180 camera recorded flutists at 60 fps², and microphones sepa-
181 rately captured audio from each instrument at 44.1 kHz.
182 A neutral-colored screen behind flutists minimized back-
183 ground interference for accurate facial movement tracking.

184 3.1.3 Musical Piece

185 This dataset comprises simultaneously starting flute-piano
186 duets. Part 1 included a C major scale and Pachelbel's
187 Canon performed at slow (50 BPM), medium (100 BPM)
188 and fast (150 BPM) tempos, each repeated twice as warm-
189 up exercises (12 trials). Part 2 consisted of 18 classical
190 pieces simplified for piano and arranged for simultaneous
191 starts. Each piece was assigned a specific tempo (50, 100,
192 or 150 BPM), and the entire set of 18 pieces was repeated
193 three times, resulting in 54 trials. Each duet performed a
194 total of 66 trials, resulting in 1,320 trials overall. Sheet
195 music and sample audio were provided in advance.

196 3.1.4 Procedure

197 The recording procedure for each piece included: (1) An
198 experimenter's clap signaling start, followed by a measure
199 of clicks matching the given tempo; (2) The flutist giving a
200 cue after clicks ended; (3) The duet beginning in response
201 to the cue.

202 3.1.5 Post-session Interview

203 The interviews collected the insights of the participants
204 on cue strategies. Participants reported providing cues ap-
205 proximately one beat (or half or two beats, depending on
206 the piece) ahead, using body movements or breath. Some
207 participants mentioned that in typical performance situa-
208 tions, they adjust their cue timing based on the accompa-
209 nying instrument and ensemble context.

² Some videos were recorded at 30 fps due to camera overheating

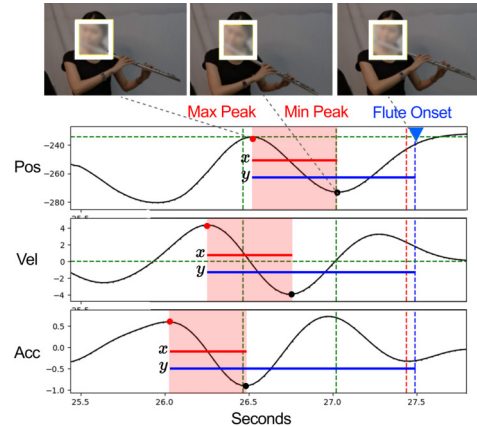


Figure 3: Gesture Cue Example Gesture cue is defined with the maximum (red) and minimum (black) peaks. The interval between these peaks, called 'cue length' (x , red line), and the duration between the maximum peak to the flute onset, called 'cue-onset length' (y , blue line).

210 3.2 Annotation and Preprocessing

211 Video and audio data synchronization was achieved
212 through an experimenter's clap at the start. Using the spec-
213 trogram viewer in Adobe Audition, we manually annotated
214 six markers per trial on the mel spectrogram (Figure 2): the
215 experimenter's clap (*Start*), breath sound onset and offset
216 (*Breath Onset*, *Breath Offset*), initial note onsets of flute
217 and piano (*Flute Onset*, *Piano Onset*), and the flute's sec-
218 ond measure onset (*2nd Measure*).

219 4. METHODS

220 4.1 Gesture Cue Detection

221 4.1.1 Motion Detection

222 To detect gesture cues from flutists, we used MediaPipe's
223 face landmark detection³ [19] to reliably identify face re-
224 gions, even when partially obscured by the flute. Subse-
225 quently, optical flow methods [17, 18] were applied to track
226 facial motion. A pilot study indicated optimal face land-
227 mark detection accuracy when the face occupied at least
228 50% of the video frame height; videos were accordingly
229 resized.

230 4.1.2 Motion Feature Extraction

231 We analyzed facial gestures by extracting position, veloc-
232 ity, and acceleration magnitude curves from averaged y-
233 axis optical flow values. Due to quantized pixel positions
234 causing discrete velocity curves, we applied zero-phase fil-
235 tering⁴ to smooth the curve while preserving peak posi-
236 tions.

237 4.1.3 Motion Peak Picking

238 Figure 3 illustrates a typical gesture cue pattern. Within a
239 one-measure window preceding the flute onset ('cue win-
240 dow'), we identified maximum and minimum peaks on
241 position, velocity, and acceleration curves using the *find-
242 peaks*⁵ algorithm. This approach automatically detected

³ available at: <https://developers.google.com/mediapipe>

⁴ [scipy.signal.filtfilt](https://docs.scipy.org/doc/scipy/reference/signal.html#scipy.signal.filtfilt)

⁵ [scipy.signal.find_peaks](https://docs.scipy.org/doc/scipy/reference/signal.html#scipy.signal.find_peaks)

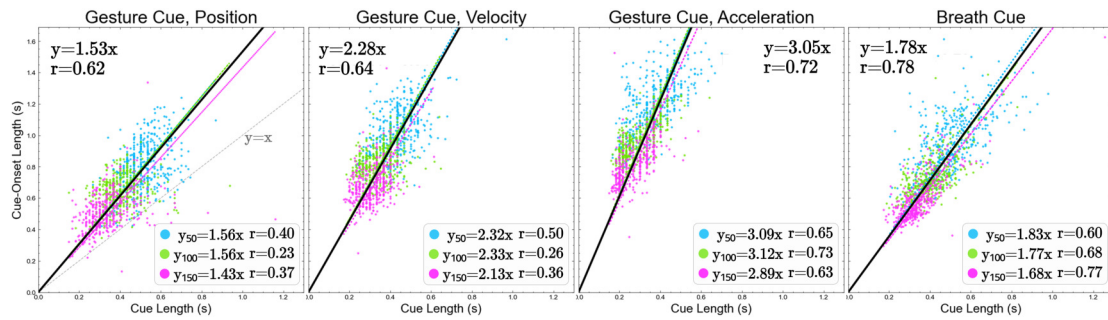


Figure 4: Relationship between Cue Length and Cue-Onset Length for Gesture Position, Velocity, Acceleration, and Breath Cue Black line: overall regression; blue, green, pink: 50, 100, 150 BPM, respectively. Slopes ($y = ax$) and correlation (r) are shown for each curve.

848 peaks, with 394 manually annotated. Trials without clear peak patterns or failed tracking were excluded, resulting in 1,242 usable trials out of 1,320. Gesture curves from cases where peak tracking failed are also available on the accompanying webpage.

4.2 Breath Cue Detection

To detect breath cues, we annotated breath onset and offset within the same one-measure ‘cue window’ preceding flute onset, based on mel spectrograms (Figure 2). Analyses were conducted exclusively on the 1,242 trials that were verified to contain valid gesture cues.

4.3 Onset Timing Prediction

We examined four cues: gesture position, velocity, acceleration, and breath. To explore the relationship between these cues and onset timing, we applied a simple linear regression model. Each cue length was set as the independent variable x , while the cue-onset length (the interval from cue start positions—maximum peak or breath onset—to the actual onset) was the dependent variable y . The regression model, without bias, is defined by $y=ax$, as illustrated in Figure 3. The slope a derived from the regression indicates the ratio between cue length and onset timing, enabling onset prediction. Additionally, we analyzed the trials separately by tempo to assess differences in cue-onset relationships, examining both the regression slope and Pearson correlation. A high correlation would confirm the cue’s validity for predicting onset timing.

5. RESULTS AND DISCUSSION

5.1 Patterns in Gesture Curves

The position curve did not consistently show the same shape in all trials, but in most cases it remained static initially and then displayed a clear downward-up-down motion pattern that served as a signal (Figure 3). Even when the position curve deviated from the typical pattern, an up-down motion immediately preceding onset was consistently present (1242 out of 1320 trials, see Section 4.1.3). These movements were often periodic and sinusoidal, resulting in velocity and acceleration curves that mirrored the position curve, but phase-shifted by approximately a quarter cycle. Further investigation could examine how variations in curve characteristics, such as the degree of sinusoidal shape and periodicity consistency, influence per-

formers’ interpretation of cues and their subsequent synchronization accuracy. Moreover, analyzing deviations from typical sinusoidal patterns might uncover additional insights into performer-specific gesture strategies. Due to challenges in quantifying these curve characteristics precisely, this topic remains open for future research.

5.2 Results of the Linear Regression

Figure 4 illustrates the relationships between cue length and cue-onset length for gesture and breath cues across all participants. Both types of cues showed linear relationships with cue-onset durations. For breath cues, eight outliers were identified due to ambiguous annotations; these were excluded from subsequent regression analyses. Linear regression indicated slopes of 1.53 (gesture position), 2.28 (gesture velocity), 3.05 (gesture acceleration), and 1.78 (breath cue). Interestingly, none of the cues yielded a slope close to 2, suggesting that counting the cue length and subsequent duration as equal units (similar to counting two beats) is not consistently applicable. The velocity and breath cues had slopes closest to 2, but the interval from the velocity cue to onset was slightly longer (1.28 times cue length), while for the breath cue it was slightly shorter (0.78 times cue length). Gesture acceleration showed a strong correlation (0.72) with cue-onset durations, but breath cues exhibited an even stronger correlation (0.78), highlighting their superior reliability for predicting onsets. The higher correlation with velocity and acceleration compared to position aligns with previous studies, suggesting performers primarily perceive velocity or acceleration peaks as cue indicators rather than positional points. However, acceleration peaks closely align with previous position peaks, indicating a potential two-peak encoding pattern in position curves. Precisely quantifying this relationship is challenging and thus remains a topic for future research.

Additionally, slopes generally decreased slightly with increased tempo, indicating flute onsets occurred sooner than expected based on proportionally shortened cue lengths. Furthermore, correlation coefficients for gesture cues decreased at higher tempos, whereas breath cues exhibited higher correlations at faster tempos, reinforcing the effectiveness of breath cues for synchronization.

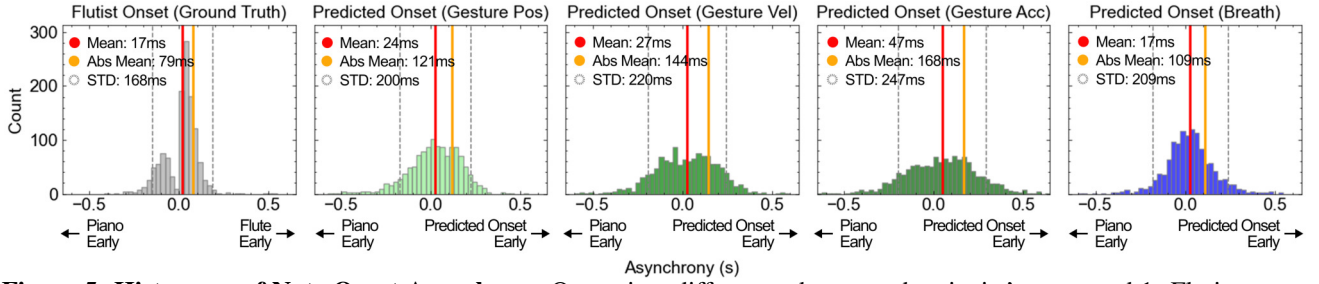


Figure 5: Histogram of Note Onset Asynchrony Onset time differences between the pianist’s onset and 1. Flutist onset (Ground truth) 2-4. Predicted onset from gesture position, velocity, acceleration cue 5. Predicted onset from breath cue.

ID	Group	Gesture Clarity	Onset Timing	Breath Clarity	Onset Timing	Async	Abs Async	STD
A	late	4.5	3.3	3.9	3.4	71	71	39
	good	4.0	3.6	4.8	3.0	54	54	31
	fast	3.8	2.4	4.4	2.6	-54	125	66
B	late	4.4	3.1	3.9	3.1	71	71	39
	good	3.1	3.4	3.3	3.4	54	54	31
	fast	3.9	3.9	4.5	3.9	-54	125	66

Table 1: Result of the Expert Evaluation Expert ratings of cue clarity and execution timing (1 = late, 3 = on-time, 5 = early) on a 5-point scale, with asynchrony metrics (ms) indicating average group asynchrony.

5.3 Note Onset Asynchrony

5.3.1 Asynchrony Comparison

Figure 5 shows histograms of time differences between the pianist’s onset and five conditions: actual flutist onset (gray) and four predicted onsets from the linear regression model (Section 5.2). This error represents the discrepancy expected if the flutist perfectly followed our linear model and executed the cues precisely. A notable feature in the pianist-flutist asynchrony histogram is the minimal occurrence of trials just before zero milliseconds. This suggests a tendency for the leader (flutist) to start slightly earlier than the follower (pianist), aligning with observations confirmed by expert interviews (Section 5.5.2). Additionally, we consider auditory reaction (performers starting in response to hearing the partner’s onset) unlikely in most cases, as the observed asynchronies are typically smaller than the average auditory reaction time (150 ms) [20].

The ‘Pianist-Flutist asynchrony’ had the narrowest spread (Absolute(Abs) mean = 79ms, Standard Deviation (STD) = 168ms), consistent with previous research reporting the first onset asynchronies slightly above 80ms [9]. Gesture-based predictions showed absolute mean errors between 121–168ms, with acceleration predictions exhibiting higher variability (STD = 247ms). Breath predictions were more consistent (Abs mean = 109 ms, STD = 209 ms). Although the acceleration cue demonstrated the highest Pearson correlation, its greater variability and longer cue length contributed to larger errors. These findings suggest that while cue-based linear predictions are slightly less precise than human synchronization, breath cues provide more reliable predictions than gesture-based cues. The precision of predictions compared with actual flutist onsets exhibited similar patterns.

5.3.2 Reduction of Asynchrony Through Repetition

We also investigated whether synchronization accuracy improves as pianists and flutists adapt to each other’s cues.

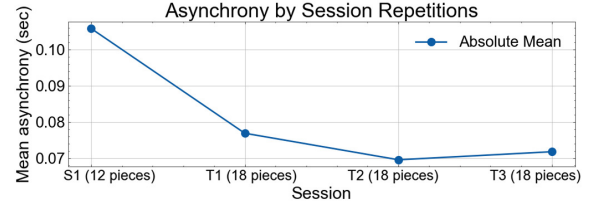


Figure 6: Human onset asynchrony across session repetition.

As described in Section 3.1.3, we observed asynchrony changes through an initial warm-up session of 12 trials (S1) and three repeated sets of 18 pieces (T1–T3), illustrated in Figure 6. Asynchrony notably decreased from the initial warm-up session (S1) through the second repetition session (T2) but stabilized thereafter. We interpret this as indicating adaptation effects, where performers quickly improved synchronization by familiarizing themselves with each other’s cues. However, ongoing piece variation and unresolved consensus about triggers (Section 5.4) likely prevented further reduction in asynchrony beyond a certain threshold.

5.4 Triggers

Despite the considerable accuracy of linear predictions (Section 5.2), questions remained regarding precise cue recognition, particularly for velocity and acceleration peaks. Given potential perception errors in identifying cue points, we investigated additional factors musicians might use to ensure precise synchronization. We observed characteristic gesture patterns immediately following cue movements near onset timings. After the downward-upward-downward cue motion, flutists typically executed one of three distinct trigger patterns (Figure 7): (A) initiating onset just before raising the head again, typically aligned with the previous upward cue movement, (B) briefly pausing with the head lowered and starting immediately afterward, or (C) ambiguously initiating onset while slowly raising the head. Patterns (A) and (B) provided clear, immediate signals suitable as precise triggers. Pattern (C), however, represented ambiguous or absent triggers. Without prior agreement, these ambiguous gestures could cause confusion—for example, a flutist intending pattern (A) might be misinterpreted by the pianist as pattern (B), resulting in significant timing discrepancies. Such cases were indeed observed, and the validity of these trigger patterns was further confirmed through expert interviews (Section 5.5). Thus, we propose that musicians initially encode approximate timing through cues and then

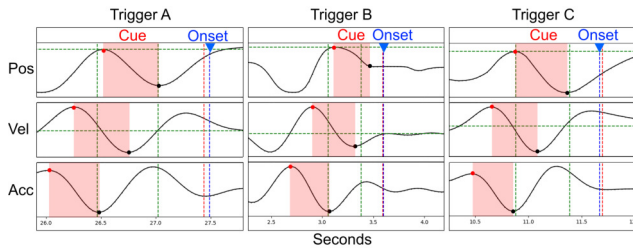


Figure 7: Examples of three trigger types (A, B, C) Green dashed lines indicate breath onset and offset, the blue line marks flute onset, and the red line piano onset. The red shaded area represents the cue interval.

achieve precise synchronization through clearly defined trigger gestures.

5.5 Expert Evaluation and Interview

After the main analysis, we conducted a two-hour interview with two expert flutists, each with over 20 years of ensemble and teaching experience. The session involved two main tasks: evaluating trials categorized as fast, good, or late based on onset predictions from velocity cues, and assessing our trigger hypothesis. Due to time constraints, only velocity cues were used, as they offered the most interpretable and perceptually reliable basis for cue classification.

5.5.1 Evaluation of Linear Model Predictions

Experts reviewed 8 trials from each group (fast, good, late) to evaluate the clarity of the cues (accuracy) and whether the actual onset timing matched the timing implied by the cue (execution timing). The experts evaluated 24 randomly ordered trials without being informed of the group labels. After agreeing with the assumption that head-nodding gestures and breathing serve as cues, experts rated cue accuracy on a scale from 1 (no recognizable cue) to 5 (clearly recognizable cue) and execution timing from 1 (very late) to 5 (very early), with 3 representing precise timing. Results are summarized in Table 1.

Surprisingly, the experts differed notably in their evaluations of breath cue execution timing, which also did not align clearly with actual pianist-flutist asynchrony. Expert A rated the ‘good’ group’s gesture cues as slightly late but breath cues as most accurate. Expert B rated the groups in descending order (late-good-fast) of perceived lateness but still considered the ‘late’ group relatively early (average 3.1). Actual asynchrony partly aligned with our predictions; the ‘good’ group exhibited the smallest asynchrony, consistent with the cue hypothesis. However, unexpected discrepancies emerged, such as the ‘fast’ group showed later flute onsets than the piano. These inconsistencies likely reflect individual variations in interpreting and executing cues.

5.5.2 In-depth Interview

After the video evaluation, we conducted detailed discussions about the linear prediction model and the trigger concept. Experts acknowledged the general practice of cueing approximately one beat ahead, noting that specific

movements were intuitively executed rather than explicitly planned. Both emphasized tempo-related influences on gesture and breath cues, highlighting gesture periodicity as crucial for clear cue delivery.

Regarding triggers, experts initially did not consciously recognize different trigger types but agreed with the proposed classifications after reviewing examples. Both agreed type (A) was optimal, while type (B) was considered challenging due to the flute’s physical constraints. Opinions on type (C) diverged: Expert A considered it inherently prone to higher errors, whereas Expert B believed it could be viable when accompanied by precise breath cues and rehearsal. Experts noted that triggers may vary depending on musical context (phrasing, emphasis). They also observed potential triggers, including the flute’s endpoint position, lip shape, and finger movements. Finally, both proposed that in flute-piano duets, slight delays in piano onset might cognitively benefit synchronization, potentially explaining the scarcity of trials where piano onset preceded flute onset, as observed in Figure 5.

5.6 Limitation

Although this study contributes to understanding cue-based synchronization, several limitations remain. First, our trigger analysis was not fully quantitative; future research should develop precise methods for defining triggers from gesture curves and consider additional signals such as lip shape, finger movements, and horizontal flute actions. Second, The decision-making process for selecting triggers was not investigated. Exploring how musicians choose and agree on triggers could further clarify synchronization strategies. Additionally, individual variability in cue and trigger preferences was not quantitatively examined; future research could explore personalized synchronization strategies. Lastly, our findings apply specifically to flute-piano duets and piece initiation. Generalizing these methods to other instrument combinations and ongoing musical contexts remains necessary.

6. CONCLUSION

This study investigated how gesture and breath cues used by flutists in flute-piano duets encode note onset timing at the initiation of musical pieces. To quantitatively analyze cues, we collected a cue dataset, identifying cue points via motion tracking and breath sounds. Our multimodal analysis revealed linear relationships between cue lengths and onset timings, confirming that gesture and breath cues reliably predict onset timing. In addition, we introduced the concept of triggers, defined as immediate gestures that indicate precise onset moments, which we validated through expert interviews. Future research could further develop the analysis of triggers to achieve a more systematic understanding. Additionally, developing regression models that jointly use gesture and breath cues, or exploring sophisticated predictive models, would be valuable. It could further extend our cue analysis methods to other instruments, ensemble configurations, and within-performance synchronization, providing deeper insights into the complexity of ensemble coordination.

7. ETHICS STATEMENT

This study was approved by the Institutional Review Board (IRB), and all participants consented to video recording and data sharing. Each flutist's session lasted 1.5 hours, pianists had 30-minute breaks between sessions, with up to four sessions per day. Participants received appropriate compensation. To protect privacy, the video data will remain confidential.

8. ACKNOWLEDGMENTS

This work has been supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) under Grant RS-2023-NR077289 and Grant RS-2024-00358448.

9. REFERENCES

- [1] C.-J. Tsay, "Sight over sound in the judgment of music performance," *Proceedings of the National Academy of Sciences*, vol. 110, no. 36, pp. 14 580–14 585, 2013.
- [2] L. Bishop, C. Cancino-Chacón, and W. Goebel, "Moving to communicate, moving to interact: Patterns of body motion in musical duo performance," *Music Perception: An Interdisciplinary Journal*, vol. 37, no. 1, pp. 1–25, 2019.
- [3] F. K. Hukporti, *Your Guide to Basic Conducting*. Acra: Noyam Publishers, 2023.
- [4] R. Page-Shipp, D. Joseph, and C. van Niekerk, "Conductorless singing group: a particular kind of self-managed team?" *Team Performance Management: An International Journal*, vol. 24, no. 5/6, pp. 331–346, 2018.
- [5] A. Maezawa and K. Yamamoto, "MuEns: A multi-modal human-machine music ensemble for live concert performance," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 4290–4301.
- [6] X. Gao, A. Rogel, R. Sankaranarayanan, B. Dowling, and G. Weinberg, "Music, body, and machine: gesture-based synchronization in human-robot musical interaction," *Frontiers in Robotics and AI*, vol. 11, 2024.
- [7] A. Lim, T. Mizumoto, L.-K. Cahier, T. Otsuka, T. Takahashi, K. Komatani, T. Ogata, and H. G. Okuno, "Robot musical accompaniment: integrating audio and visual cues for real-time synchronization with a human flutist," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 1964–1969.
- [8] L. Bishop and W. Goebel, "When they listen and when they watch: Pianists' use of nonverbal audio and visual cues during duet performance," *Musicae Scientiae*, vol. 19, no. 1, pp. 84–110, 2015.
- [9] —, "Beating time: How ensemble musicians' cueing gestures communicate beat position and tempo," *Psychology of Music*, vol. 46, no. 1, pp. 84–106, 2018.
- [10] B. Vera, E. Chew, and P. G. Healey, "A study of ensemble synchronisation under restricted line of sight," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2013, pp. 293–298.
- [11] R. Timmers, S. Endo, A. Bradbury, and A. M. Wing, "Synchronization and leadership in string quartet performance: a case study of auditory and visual cues," *Frontiers in Psychology*, vol. 5, p. 645, 2014.
- [12] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 522–535, 2018.
- [13] S. Kawase, "Gazing behavior and coordination during piano duo performance," *Attention, Perception, & Psychophysics*, vol. 76, pp. 527–540, 2014.
- [14] Y.-H. Su, "Audiovisual beat induction in complex auditory rhythms: Point-light figure movement as an effective visual beat," *Acta Psychologica*, vol. 151, pp. 40–50, 2014.
- [15] G. Luck and J. A. Sloboda, "Spatio-temporal cues for visually mediated synchronization," *Music Perception*, vol. 26, no. 5, pp. 465–473, 2009.
- [16] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.
- [17] B. Li, K. Dinesh, Z. Duan, and G. Sharma, "See and listen: Score-informed association of sound tracks to players in chamber music performance videos," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2906–2910.
- [18] B. Li, C. Xu, and Z. Duan, "Audiovisual source association for string ensembles through multi-modal vibrato analysis," *Proc. Sound and Music Computing (SMC)*, pp. 159–166, 2017.
- [19] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," *arXiv preprint arXiv:2006.10204*, 2020.
- [20] R. J. Kosinski, "A literature review on reaction time," *Clemson University*, vol. 10, no. 1, pp. 337–344, 2008.