EFFECT-TIME CUE SCHEDULING FOR DISTRIBUTED GAZE-BASED MUSICAL PERFORMANCE

Doohyun JUNG (gazecomposer@gmail.com) [1]

[1] **Independent Researcher**, Seoul, South Korea

ABSTRACT

We present a distributed gaze-based musical performance system designed as a prototype for remote gaze-driven co-performance. The system coordinates visual targets with pre-cues and scheduled effect times, while audible MIDI events remain controlled locally through gaze-based grid selection. Three real-world conductor–performer sessions were analyzed through the same cloud-routed architecture: a co-located Korea–Korea trained-use reference session and two first-time remote sessions in India and South Africa. Cue-error p95 (95th percentile), defined here as participant-side cue-lifecycle timing error rather than audio-onset error, stayed near 10 ms while round-trip time (RTT) ranged from 164 ms to 357 ms. The main failures appeared in other layers. Gaze acquisition was lower in the first-time sessions and appeared sensitive to calibration familiarity, while reconnect conditions produced missing and duplicate cue-fire reports. The sessions suggest that distributed gaze performance should be evaluated by separating scheduled cue activation, gaze-driven sound generation, onboarding, and session-state management.

1. INTRODUCTION

Gaze is a compelling musical input: it is continuous, embodied, and closely linked to attention. It is also fragile. Calibration drift, visual context, head position, and familiarity all affect whether a gaze-driven interface can be played reliably. Once a gaze-based interaction is distributed across distance, network delay and session-state consistency become additional concerns.

The system described in this paper addresses a specific musical problem: how can performers in different places play together through gaze? In the present project, performers generate local MIDI events by selecting grid cells through gaze while responding to conductor-selected visual targets. Effect-time cue scheduling governs when the main visual cue becomes active and when the participant client reports the corresponding cue-lifecycle event; it does not itself generate a MIDI note. This separation allows geographically distant performers to receive visual pre-cues, prepare their gaze orientation, and share scheduled formal references without executing cue-state changes immediately on network-message arrival.

Copyright: © 2026. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Networked music performance research has long treated delay as a central obstacle to remote collaboration [1, 2]. Recent telematic systems have also shown that musical systems can remain useful by designing around delay instead of only attempting to eliminate it [3, 4]. In parallel, gaze-based musical interfaces have established gaze as a viable means of musical control [5–7, 9]. Ensemble-oriented systems have further used gaze and body motion to support awareness between co-performers [8].

The intersection of these two lines remains underexplored: gaze-driven performance deployed across real networks, with performers who differ in calibration familiarity. This paper evaluates effect-time cue scheduling as the timing layer of a distributed gaze-based musical instrument. The guiding question is which parts of the architecture remain stable, and which break down, when a gaze-based musical interface moves from trained local experimentation into first-use telematic deployment.

2. RELATED WORK

2.1 Networked Music Performance

Chafe et al. demonstrated that even modest temporal separation can disrupt rhythmic synchronization in paired performance [1]. Rottondi et al. surveyed networked music performance technologies and identified latency, jitter, synchronization, and transport design as recurring architectural constraints [2]. These studies frame delay as a primary engineering and musical challenge.

L2Ork Tweeter and Global Hyperorgan represent a complementary design philosophy. L2Ork Tweeter uses a control-data-based approach to support tightly timed telematic and crowdsourced musicking under latency, sync, and bandwidth constraints [3]. Global Hyperorgan frames geographically distributed organ performance as a creative telematic space, not as a delayed substitute for co-located performance [4]. These projects show that telematic music can succeed through architectures that absorb or reframe delay.

2.2 Gaze-Based Musical Interaction

EyeMusic provided an early demonstration that eye movement could serve as musical input [5]. EyeHarp turned gaze into a structured digital musical instrument with a strong accessibility orientation [6]. Focal extended eye tracking toward expressive parameter control in live performance [7]. LuminUs used gaze and body motion feedback to support awareness between co-performers [8]. DJ-

eye foregrounded accessibility in gaze-based musical design by enabling DJing operations through gaze and winking interaction [9].

These systems demonstrate gaze as a local interface, an expressive controller, or an accessibility-oriented workflow. The present work differs by treating gaze as one layer in a distributed cue-scheduling system. It examines how cue timing, reconnect behavior, and cue-state consistency behave under cloud-routed deployment conditions.

3. SYSTEM AND PERFORMANCE SETUP

The system consists of a performer-side gaze instrument, a conductor interface, and a cloud-routed coordination server (Fig. 1). The performer interface estimates point of regard from a webcam-based gaze tracker and displays the target grid used for gaze selection. The conductor interface issues target cues in real time. The server distributes visual cues, maintains timing references, and records cue creation, participant-side cue-fire reports, participant status, gaze traces, and gaze-target entries.

3.1 Musical Interaction

The reported sessions used a sparse, conductor-led cue-following performance protocol. The performer calibrated gaze using a 4×4 calibration procedure and then performed on a 5×10 gaze target grid. Each grid region was mapped to a discrete MIDI piano pitch. The grid functioned as a simple pitch plane: horizontal movement selected adjacent scale steps, while vertical movement shifted the register by larger pitch intervals within the available MIDI range. A note was rendered locally through a MIDI piano timbre when gaze-based cell selection completed; cue activation alone did not sound the note.

The conductor selected sparse target cues in real time instead of replaying a fixed automated sequence. The performer's role was primarily responsive: to acquire the indicated target through gaze and thereby select its mapped pitch. Musical agency was distributed between the conductor's real-time target selection and the performer's gaze-based execution, with limited freedom to select other grid cells. The sessions were designed to make remote visual cue-following audible as a musical process. The simple MIDI piano sound kept the sonic layer transparent, allowing the study to focus on gaze orientation, scheduled cue activation, and session-state reliability.

3.2 Cue Scheduling Model

The key design choice is effect-time cue scheduling. A cue packet is not activated when it arrives. The conductor sends three time-related values with each target: a pre-cue onset, a future effect time, and an expiry time. The pre-cue gives the performer a preparation window, the effect time marks main-cue activation and the participant-side cue-fire report, and the expiry time ends the cue association. These timing events are separate from sound generation. The gaze-grid runtime produces a MIDI note when gaze-based cell selection satisfies its selection and local timing conditions; no note is generated solely because effect time or expiry is reached. A note may therefore occur before or after effect time depending on when cell selection completes. In the configuration used by the system, cues used a 1500 ms pre-cue window and expired 5000 ms after the scheduled effect time.

Each cue follows a defined visual and logging lifecycle: created by the conductor, received by the performer, displayed as a pre-cue, activated and reported at effect time, and discarded after expiry. In this paper, a *cue-fired* record denotes that participant-side lifecycle report, not an audible MIDI note onset. This distinction supports analysis of cues with missing or repeated reports and of stale state after reconnect.

3.3 Timing and Routing

Performer and conductor communicate through the same cloud-routed server instead of through a purely local path. Measured RTT therefore reflects routing architecture as well as geography. The Korea–Korea session illustrates this point: although performer and conductor were physically co-located, median RTT still reached 164 ms because all traffic passed through a cloud server. The Korea–Korea session is therefore a trained-use reference sharing the same architectural bottleneck as the remote sessions.

Scheduled cue times were generated on the server and distributed as shared timing references to the participant-side system. The present architecture relies on server-timed scheduling instead of independently synchronized local clocks. Reported cue-error values therefore reflect the observed difference between a server-issued effect time and a corresponding locally logged cue-fired lifecycle timestamp under this shared timing model. They do not measure audible MIDI onset timing.

4. METHOD

We analyze three real-world sessions using the same deployed architecture. The analysis is a system and case-study evaluation, not a controlled user study. Each session involved one conductor and one performer. Participant familiarity and geographic location covary across sessions, so gaze-acquisition differences are treated as descriptive observations, not controlled evidence of geography-driven performance differences.

4.1 Participants and Familiarity

The Korea–Korea session serves as a trained-use reference. The performer had completed at least ten prior sessions and was familiar with calibration and gaze-based performance. This session shows the system in a comparatively clean state and provides a trained-user gaze reference.

The India and South Africa sessions were first-time remote sessions. In both cases, performers had no prior experience with the calibration procedure or with performing through gaze; both informally reported that calibration and gaze-based interaction were difficult. These sessions reflect a realistic onboarding scenario for new users.

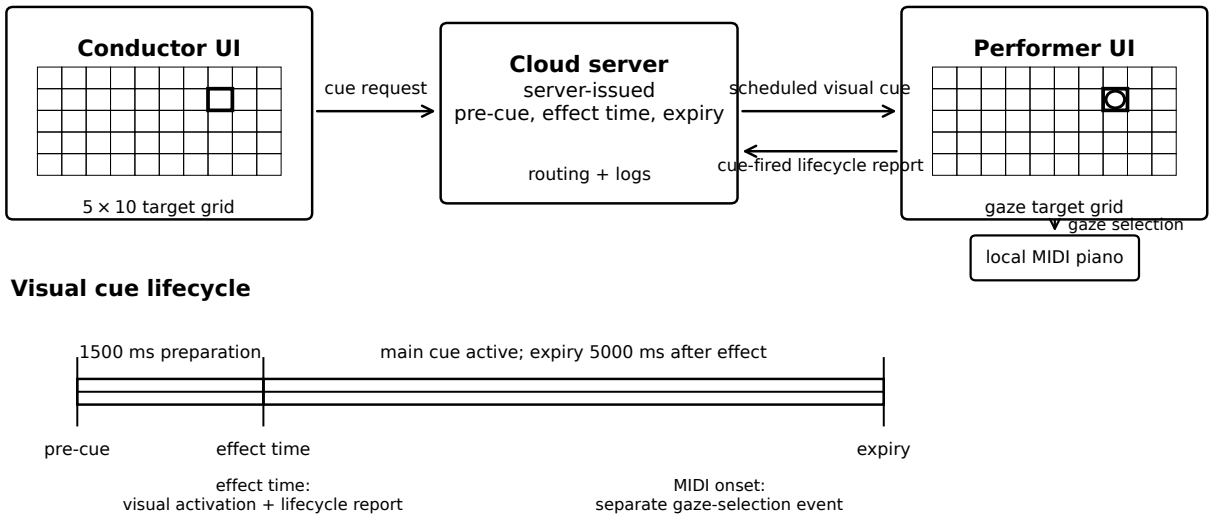


Figure 1. Performance setup and cue-following interaction. The conductor selects sparse target cues in real time, and the server routes each cue with a pre-cue, scheduled effect time, and expiry time. At effect time, the main visual cue becomes active and the participant client reports a cue-lifecycle event. Audible MIDI notes are generated separately when gaze-based grid selection completes.

4.2 Protocol

Each session followed the same broad structure. The conductor-side interface issued a sequence of sparse grid targets while the performer used the gaze interface to orient toward and select them. The server logged scheduled cues, participant-side cue-fire reports, connection state, and gaze traces throughout. The protocol was intentionally lightweight, preserving instability that a purely local validation would miss. Cue counts differed across sessions (39, 45, and 60), so the reported sessions should be interpreted as field deployments rather than normalized repetitions of a single fixed-length trial.

4.3 Metrics

The reported summary metrics are RTT p50 (median), cue-error p95, missing and duplicate cue-fire reports, disconnect count, and gaze hit rate. Cue error is computed in milliseconds as the cue-fired lifecycle timestamp, expressed in the server-timed model, minus the scheduled effect time of the corresponding cue. It is a scheduler/lifecycle metric, not an audio-onset metric. Cue error can be computed only for observed cue-fired records; a cue with no report therefore contributes to the missing-report count but not to the cue-error distribution. Duplicate reports are additionally counted as session-state indicators. The notation p95 denotes the 95th percentile: the value below which 95% of observed samples fall. A cue was counted as a gaze hit if at least one gaze sample entered the corresponding target region during the cue-linked analysis window used for the session summary. Because the retained session artifacts do not permit reconstruction of the exact window boundaries, gaze-hit rates are interpreted only as descriptive interaction indicators. One South Africa cue-error record of 2532 ms occurred during the initial warm-up state before steady cue-following and was excluded from steady-state p95 summaries; the exclusion did not change cue coverage or missing/duplicate-report counts.

We distinguish three levels of analysis. Cue-error measures participant-side lifecycle timing; missing and duplicate cue-fire reports and disconnects indicate session-state handling; gaze hit rate describes interaction with the cued target. Audible MIDI onset is produced by gaze-grid selection and is not measured by cue error. This separation is central to the paper's argument: stable scheduled activation can coexist with fragile cue state and gaze interaction.

5. RESULTS

5.1 Stable Layer: Scheduled Cue-Lifecycle Timing

Participant-side cue-lifecycle timing did not drift as RTT increased. RTT rose from 164 ms in the Korea–Korea reference session to 253 ms in India and 357 ms in South Africa. Cue-error p95 stayed close: 9.5 ms, 10.4 ms, and 11.1 ms, respectively (Fig. 2 and Table 1). The South Africa value summarizes steady-state lifecycle behavior after excluding the single initial warm-up outlier described in the Method section.

This result is expected under an effect-time design: lifecycle timing error should remain relatively stable when cue messages arrive before their scheduled effect times. The observation is nevertheless useful because it confirms that scheduled visual-cue activation and reporting behaved as intended under the network conditions observed here, including RTT above 350 ms.

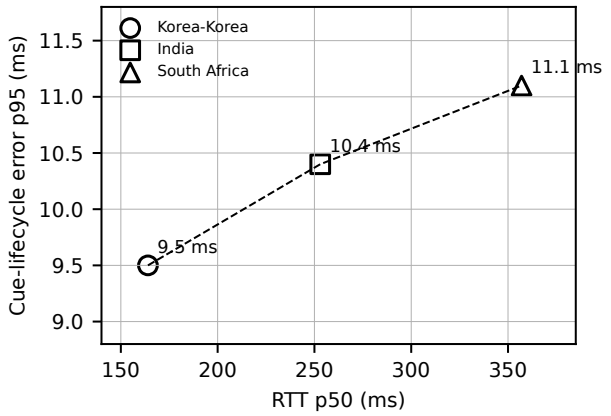


Figure 2. RTT p50 versus cue-error p95 for participant-side cue-lifecycle reports. Despite RTT more than doubling from Korea–Korea to South Africa, cue-error p95 remained near 10 ms. These values do not measure MIDI note onset; South Africa excludes one 2532 ms warm-up outlier.

Table 1. Session comparison. South Africa p95 reflects steady-state cue-lifecycle behavior excluding one 2532 ms warm-up outlier. RTT p50 and cue-error p95 (CE p95) are in ms. Miss/Dup denotes missing and duplicate cue-fire reports, not MIDI note onsets. Gaze hit rates are descriptive only.

Session	Fam.	RTT p50	CE p95	Miss/Dup	Disc.	Hit
Korea	Exp.	164	9.5	0 / 0	2	59%
India	1st	253	10.4	8 / 0	8	38%
S. Africa	1st	357	11.1	0 / 53	6	28%

5.2 Broken Layer: Cue-State Reporting

Cue-state reporting failed in different ways across the first-time remote sessions. The Korea–Korea reference run had no missing or duplicate cue-fire reports, two disconnects, and a cue-error p95 of 9.5 ms. India kept stable lifecycle timing but produced eight missing reports and eight disconnects. South Africa again kept stable steady-state lifecycle timing, yet 53 duplicate reports appeared despite full cue coverage (Table 2).

These are distinct report-level failure modes. In India, eight scheduled cues had no corresponding participant-side cue-fire report. The eight disconnects suggest that connection interruptions contributed to missing reports, although the present dataset is not sufficient to isolate a single causal mechanism. In South Africa, all 60 cues had at least one report, with 53 additional duplicate records yielding 113 total cue-fire records. The reconnect-associated pattern is more consistent with cue-state management failure than with scheduler timing failure. Because cue-fire reports are distinct from gaze-generated note onsets, these counts do not establish missing or duplicated MIDI notes.

5.3 Interaction Layer: Gaze Acquisition

The interaction data reinforce the same picture. Gaze hit rate fell from 59% in Korea–Korea to 38% in India and 28% in South Africa. These values should be read descriptively, since familiarity, calibration quality, selection thresholds, and reconnect-related data loss may contribute to the differences. A gaze hit indicates target-region acquisition in the analysis trace; it is not identical to a logged MIDI note onset.

Table 2. Cue-fire report coverage by session. India and South Africa maintained stable lifecycle timing but exhibited opposite report-level failure modes.

Session	Cues	Unique reports	Dominant failure
Korea	39	39	None
India	45	37	Missing reports
S. Africa	60	60	Duplicate reports

6. DISCUSSION

This study evaluates effect-time cue scheduling as the visual-cue timing layer of a distributed gaze-based musical instrument under three real deployment conditions. The main contribution is the separation of scheduled cue activation, gaze-driven sound generation, and session-state reporting. Cue-lifecycle timing remained stable in all three sessions, while gaze acquisition and reconnect-related cue state remained fragile.

The temporal coordination result supports future-timed visual cueing as a coordination mechanism. Even with RTT differences exceeding 190 ms, cue-error p95 stayed near 10 ms. The scheduler therefore provided a stable reference for main-cue activation and shared target changes across distance. This result concerns cue-lifecycle timing, not the onset precision of gaze-generated notes.

The musical implications depend on the sparse cue-following design. The sessions were not intended as dense free improvisations. The conductor shaped sparse visual targets, and the performer made them audible by completing gaze-based cell selection. The effect-time scheduler coordinated visual cue activation but did not sound the pitch. Missing and duplicate cue-fire reports therefore indicate cue-state inconsistency rather than directly observed missing or duplicated notes. Connecting cue records, gaze selection, and MIDI note-on events is necessary to establish their precise musical consequences.

The co-located Korea–Korea reference also shows that infrastructure becomes part of the musical condition. Physical co-location did not imply low-latency communication because traffic still passed through a cloud server. For telematic music systems, server topology shapes the performance condition alongside geography and performer skill.

Three design implications follow. First, effect-time scheduling is a suitable layer for coordinating visual targets; future-timed activation should be retained without conflating it with sound onset. Second, session-state handling deserves the same design attention as timing. Fire-once guarantees, reconnect cleanup, acknowledgments, and logs that link cue reports to gaze selection and MIDI note-on events are priorities. Third, onboarding should be treated as part of the instrument. In a gaze-based system, calibration is part of the playable interface rather than only

a setup step. The gap between trained and first-time gaze hit rates suggests that calibration fluency may have a larger practical effect on interaction quality than network delay within the observed range.

6.1 Scope and Limitations

The study includes only three sessions with one performer in each session, all using a conductor–performer dyad format. The experienced/first-time distinction covaries with geography, so gaze-acquisition differences cannot be attributed to any single factor. Multi-performer configurations, trained and untrained users across multiple server routes, and repeated trials would strengthen the analysis.

The duplicate-report pattern in South Africa and missing reports in India reveal that session-state handling remains an engineering weakness. Stronger fire-once guarantees, explicit cue-received acknowledgment, and more reliable reconnect cleanup are priorities for future iterations. The cue-error values should be interpreted within the present server-timed lifecycle scheme: they neither independently estimate cross-device clock offset nor measure MIDI note onset. The available report-level analysis also does not determine whether duplicate cue-fire records produced duplicate sounds. Finally, the study includes no formal perceptual or musical-quality evaluation. Whether the visual coordination and gaze-generated note timing are musically sufficient remains an open question for future work.

7. CONCLUSION

This paper presented a distributed gaze-based musical instrument prototype built on effect-time cue scheduling and evaluated through three conductor–performer sessions across Korea, India, and South Africa. Participant-side cue-lifecycle error p95 remained near 10 ms despite RTT variation from 164 ms to 357 ms, an expected scheduling outcome observed under real deployment conditions. This metric does not measure audible MIDI onset: sound generation remained gaze-driven through local grid selection. Gaze acquisition and cue-state reporting were less stable and appeared sensitive to calibration familiarity, reconnect behavior, and session-state robustness.

Distributed gaze performance involves more than latency. It requires clear separation between visual scheduling, gaze-based sound generation, onboarding, and architectural robustness. The next step is to retain the visual timing architecture while improving calibration, performer-facing monitoring, pre-cue lifecycle clarity, reconnect-safe fire-once state management, and event logs that align cues, gaze selections, and note onsets before scaling toward larger multi-user sessions.

Acknowledgments

The author thanks the participants in the Korea, India, and South Africa sessions for their time, patience, and feedback during system testing. Their participation made it possible to examine the system under both trained use and first-use remote conditions.

8. REFERENCES

[1] C. Chafe, J.-P. Cáceres, and M. Gurevich, "Effect of temporal separation on synchronization in rhythmic performance," *Perception*, vol. 39, no. 7, pp. 982–992, 2010.

[2] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An overview on networked music performance technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016.

[3] I. I. Bukvic, "Latency-, sync-, and bandwidth-agnostic tightly-timed telematic and crowdsourced musicking made possible using L2Ork Tweeter," in *Proc. Int. Conf. New Interfaces for Musical Expression (NIME)*, 2022.

[4] R. Harlow, M. Petersson, R. Ek, F. Visi, and S. Östersjö, "Global Hyperorgan: A platform for telematic musicking and research," in *Proc. Int. Conf. New Interfaces for Musical Expression (NIME)*, 2021.

[5] A. J. Hornof and L. Sato, "EyeMusic: Making music with the eyes," in *Proc. Int. Conf. New Interfaces for Musical Expression (NIME)*, pp. 185–188, Hamamatsu, Japan, 2004.

[6] Z. Vamvakousis and R. Ramírez, "The EyeHarp: A gaze-controlled digital musical instrument," *Frontiers in Psychology*, vol. 7, art. 906, 2016.

[7] S. Greenhill and C. Travers, "Focal: An eye-tracking musical expression controller," in *Proc. Int. Conf. New Interfaces for Musical Expression (NIME)*, pp. 230–235, 2016.

[8] E. Morgan, H. Gunes, and N. Bryan-Kinns, "The LuminUs: Providing musicians with visual feedback on the gaze and body motion of their co-performers," in *Human–Computer Interaction – INTERACT 2015*, pp. 47–54, 2015.

[9] F. Bottarelli, N. Davanzo, G. Presti, and F. Avanzini, "DJeye: Towards an accessible gaze-based musical interface for quadriplegic DJs," in *Proc. Sound and Music Computing Conf.*, pp. 117–124, Stockholm, Sweden, 2023.