

LearnTune: An Exploratory Pilot Study of Multimodal Closed-Loop Music Biofeedback During Word Learning

Zimo Fan¹ and Ting Jiang^{2*}

¹ Suzhou North America High School, China
zimofan_ethan@126.com

² Faculty of Psychology, Beijing Normal University, China
psytingjiang@gmail.com

Abstract. Arousal-related variability is associated with attention and memory, but whether this association can be translated into intervention benefits during learning requires testing with specific measures and tasks. This exploratory pilot study presents the implementation of individually calibrated multimodal closed-loop music biofeedback during word learning and examines its operational feasibility and the measurement limitations of the composite state index. The system combines a frontal EEG power ratio, camera-derived heart rate, and facial arousal, continuously adjusting the music according to deviations of the state from an individual reference center without requiring learners to actively self-regulate. A within-participant experiment compared closed-loop music feedback with music presented using static parameters; the primary state analysis included 24 participants, and the behavioral analysis included 29. Over the full feedback period, which included a 25 s adaptation period, the standard deviation of the composite index was 12.8% lower under closed-loop feedback (paired t -test $BH\ q = 0.0554$; Wilcoxon $q = 0.0231$). No significant difference in the mean was detected, and differences in the variability of individual channels were observed mainly for facial arousal. No improvements were detected in recognition discriminability or reaction time for hits. The contribution lies in system implementation and assessment of the composite index’s measurement limitations; the present results do not demonstrate a learning benefit.

Keywords: Closed-loop music biofeedback · Individual calibration · Multimodal fusion · Arousal variability · Word learning

1 Introduction

Mean level and temporal variability capture different aspects of a learner’s state: similar mean levels can correspond to different temporal trajectories. Arousal-biased competition theory emphasizes that arousal enhances the competitive advantage of high-priority information during processing, with effects that depend on information priority [12]. Pupillometry studies have further found that variability in pupillary responses is associated with working memory capacity, attention control, and task disengagement [19], and that pupillary changes are related to fluctuations in sustained attention [20]. These

* Corresponding author.

findings motivate the study of state changes during learning: in addition to examining the level of a state, it is necessary to consider how that state changes over time.

This question is directly motivated by memory research. In word-learning and free-recall tasks, Robison et al. found that participants with a larger coefficient of variation in baseline pupil diameter across word lists had poorer recall performance [16]. However, this individual-differences association does not establish whether changing state variability through feedback during an individual’s learning would produce memory benefits. The measure underlying the association may reflect several processes that vary together, and its timescale may not match that of the states that an intervention can change. These uncertainties motivate an initial evaluation of system implementation and measurement interpretation during word learning, with recognition performance examined as an exploratory task outcome.

Physiological computing and passive brain–computer interfaces provide a technical foundation for automatic state feedback during tasks [4, 23]. Previous pupil-based biofeedback research has used training to support active regulation of arousal levels and observed behavioral changes in an auditory oddball task [13], illustrating that state regulation and its effects on task performance can be examined through intervention studies. Music can also convey feedback continuously while accompanying a visual learning task: previous studies have used heart-rate-driven music to support active arousal regulation [2], or mapped EEG-based emotion estimates to generated music so that users could explore regulation strategies [3]. Leslie et al. also investigated music driven by physiological information during sustained-attention tasks without requiring active attention to breathing or music [11]. These studies provide precedents for regulation; the present study focuses on implementing automatic music feedback during word learning and assessing its operational feasibility and measurement limitations.

This study constructs an operational composite index from a frontal EEG power ratio, camera-derived heart rate, and facial arousal. Individual calibration establishes a reference center and scale, and multiple inputs are integrated according to their real-time quality. The controller adjusts the music according to deviations of the state from the reference center, while the within-block standard deviation is used to evaluate the resulting temporal variability; the reference center was not optimized for memory performance. The pilot comparison examines state and task outcomes under the full feedback strategy based on an individual reference center, without independently manipulating variance. The fusion values used for analysis and the filtered estimates used for control are recorded separately to clarify the processing stage to which changes in the index correspond.

This pilot study focuses on system implementation and measurement interpretation. It uses a within-participant comparison of closed-loop music and music with static parameters produced by the same engine to address three research questions: RQ1, what differences in the variability and mean level of the composite index are observed between conditions? RQ2, what differences in recognition discriminability and correct-response speed are observed between the closed-loop and static conditions? RQ3, how do single-channel changes and input quality constrain the interpretation of changes in the composite index? Following reporting guidelines for neurofeedback research

[17], intervention implementation, state indices, and behavioral outcomes are presented separately.

The main contributions of this exploratory pilot study are the implementation of an individually calibrated multimodal closed-loop music biofeedback system during word learning and an assessment of the measurement limitations and interpretive scope of its composite state index. The evaluation considers system operation, input quality, and the distinction between signals used for analysis and control. State-index and recognition outcomes provide exploratory observations to inform subsequent validation.

2 System Design and Implementation

2.1 System Architecture and Measurement Apparatus

The system comprises a custom-built frontal EEG headband, a standard USB camera, a computer, and headphones (Fig. 1). The headband is built around a KS1092 dual-channel biopotential analog front end and an MS88SF3 module with an integrated nRF52840. Fp1 and Fp2 acquire frontal signals, REF provides the reference, and RLD is connected to the bias drive circuit; the gains of both channels are set to 720, and raw waveforms are delivered to the software at 250 Hz [9, 10]. Camera images are used separately for heart rate estimation by remote photoplethysmography (rPPG) and for facial arousal estimation.

Features from the three channels are updated asynchronously. Every 0.5 s, the fusion and control thread reads the latest results from each branch and computes the composite index and music control variable according to their quality status. Threads exchange the latest state data under lock protection. The audio thread independently schedules continuous audio and does not perform video inference or disk writes, to reduce interference from other computations with audio continuity. Task events, physiological features, and control logs are recorded using a common time base. After putting on the equipment and completing individual calibration, participants only need to perform the word task, without actively regulating their physiological state or operating a feedback interface.

2.2 Feature Extraction and Individual Calibration

Each EEG channel maintains a 16 s buffer. Approximately every 0.4 s, the mean is removed and the power spectrum is estimated using Welch’s method, with approximately 2 s segments, a Hann window, and 50% overlap [22]. Power in the θ , α , and β bands is extracted to calculate an engagement power ratio [15]. The rPPG branch uses the POS method to process skin color changes on the forehead and both cheeks [21], with a 10 s window and recalculation every 1 s. Output is produced after at least 4 s and 64 frames of valid data have accumulated. Spectral peaks are sought within 45–180 bpm, with low-frequency priority and harmonic correction rules used to reduce misidentification of the second harmonic. The facial model provides continuous arousal and valence outputs, and the arousal component enters the fusion. The three state variables are defined as follows:

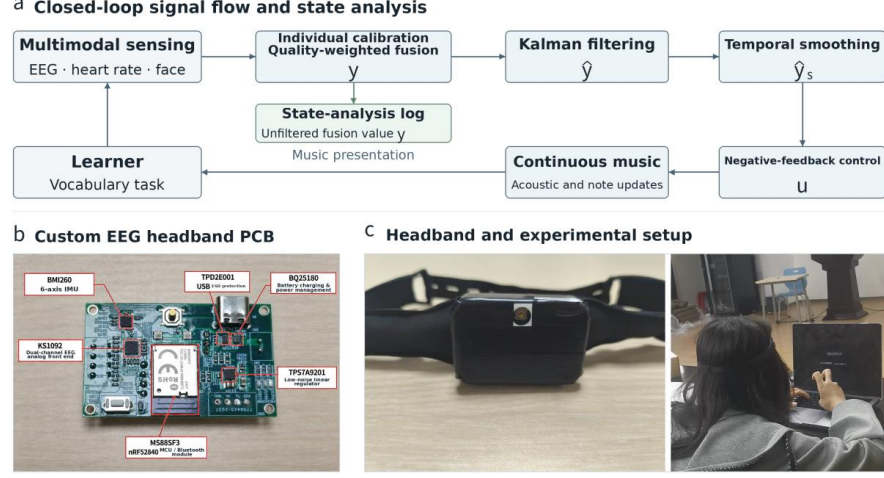


Fig. 1. System architecture and measurement apparatus. The fusion value y is recorded separately and used for outcome analyses; the control branch generates the control variable u after Kalman filtering and exponential smoothing. The PCB, headband exterior, and an example of the device being worn show the actual apparatus. Arrows indicate the signal and feedback paths specified in the design.

$$x_e = \frac{P_\beta}{P_\alpha + P_\theta + 10^{-9}}, \quad x_r = 0.40 + 0.70 \frac{HR - HR_{\text{rest}}}{50}, \quad (1)$$

$$x_f = 0.5 + 0.5\tilde{a}_f.$$

Here, HR is the median of the three most recent heart rate estimates, HR_{rest} is the median heart rate during the 20 s silent period before the current block, and $\tilde{a}_f$ is the facial arousal output in the range $[-1, 1]$. The three features have different temporal supports. The most recent values read by the fusion process are not treated as new independent sensor samples, and no additional interpolation is applied to missing values. The longer EEG analysis window mixes activity before and after a state change; frequent recalculation therefore does not imply equally rapid physiological responses.

Before the formal blocks, participants completed 20 s of quiet relaxation, 30 s of serial subtraction of 7, and 45 s of relaxation with their eyes closed, in that order. For each channel, the mean μ_i and standard deviation σ_i are estimated from samples meeting the quality criteria, requiring at least 8 samples and $\sigma_i > 10^{-6}$. Standardized scores are clipped at ± 2.5 to limit the influence of extreme values. The control center c and scale s are obtained from the filtered fusion series $\hat{y}_{\text{cal}}$ during calibration:

$$z_i = \text{clip}\left(\frac{x_i - \mu_i}{\sigma_i}, -2.5, 2.5\right), \quad (2)$$

$$c = \text{median}(\hat{y}_{\text{cal}}), \quad s = \max\left(\frac{P_{90} - P_{10}}{2}, \frac{\text{IQR}}{1.349}, 0.12\right).$$

P_{90} and P_{10} denote the corresponding percentiles of the series, and IQR denotes its interquartile range. μ_i , σ_i , c , and s remain fixed across all four blocks, whereas HR_{rest} is updated for each block. This calibration establishes a within-individual reference range; it does not identify an optimal level based on memory performance. Clipping standardized scores is also not equivalent to artifact rejection.

2.3 Quality-Weighted Fusion and Negative Feedback Control

Channel quality enters the weights through availability a_i . When EEG contact is valid, quality is good or average, and α and β power are valid, $a_e = 1$; otherwise, it is 0. When face detection and the model output are valid, $a_f = 1$; otherwise, it is 0. Heart rate quality q is the ratio of energy around the fundamental frequency and its second harmonic to the total energy in the heart rate band, and is converted to a_r through the continuous gating function in Table 1. The weights also account for the ratio of individual calibration variability to the reference scale and are normalized to obtain the composite index:

$$w_i = a_i \text{clip}\left(\frac{\sigma_i}{\sigma_{\text{ref},i}}, 0.1, 1\right), \quad (3)$$

$$y = \text{clip}\left(0.5 + 0.22 \frac{\sum_i w_i z_i}{\sum_i w_i}, 0, 1\right).$$

σ_{ref} is set to 0.08, 0.04, and 0.25 for heart rate, facial arousal, and EEG, respectively, based on typical fluctuation magnitudes observed in self-tests during development. When the weights sum to 0, the fused standardized score is set to 0. State comparisons in this study use y , whereas the controller uses the estimate $\hat{y}$ obtained with a one-dimensional random-walk Kalman filter [8]. Observation noise is adjusted according to the number of valid channels, with parameters listed in Table 1; $\hat{y}$ is initialized at 0.40 in each block. Online availability and offline quality screening of the raw waveforms are calculated separately (Sect. 3.3).

The controller further applies exponential smoothing to $\hat{y}$ with a time constant of 2 s to obtain $\hat{y}_s$, and then calculates the target control variable from the deviation relative to the individual center:

$$z_c = \text{clip}\left(\frac{\hat{y}_s - c}{s}, -2, 2\right), \quad (4)$$

$$u^* = \text{clip}(0.5 - 0.45z_c, 0, 1).$$

When the state is above the center, u is reduced, making the music darker, quieter, and sparser; the adjustment is reversed when the state is below the center. When facial valence $v_f \leq -0.25$, the target is shrunk toward the neutral value according to $u^* \leftarrow 0.5 + 0.5(u^* - 0.5)$. The rate of change of the actual output is limited to 0.18/s. When

Table 1. Key processing parameters

Module	Parameters and settings
EEG	250 Hz; $\theta/\alpha/\beta$: 4–8/8–13/13–30 Hz; 16 s window, recalculated approximately every 0.4 s
Heart rate	1280 × 720 video at 15 fps; POS; 10 s window, recalculated every 1 s; $a_r = \text{clip}[(q - 0.40)/0.20, 0, 1]$
Fusion and calibration	0.5 s cycle; σ_{ref} (heart rate/facial arousal/EEG) = 0.08/0.04/0.25; calibration: 20/30/45 s
State filtering	$Q = 6 \times 10^{-4}$, $R_0 = 1.2 \times 10^{-3}$; $\text{conf} = \text{clip}(0.25 + 0.40n_{\text{live}}, 0, 1)$, $R = R_0/\text{conf}$
Control and fallback	Smoothing: 2 s; rate limit: 0.18/s; all channels missing: hold for ≤ 2 s, return toward the neutral value with a 5 s time constant after > 2 s
Music execution	64 BPM, 4/4; 44.1 kHz; 1024 samples/block, smoothing coefficient: 0.10; scheduled 4.5 s ahead, sound card buffer: approximately 1 s

n_{live} is the current number of valid channels; clip restricts a value to the specified lower and upper bounds. The table lists software and hardware configuration values.

all channels are unavailable for no more than 2 s, the most recent target is held; after 2 s, it returns to 0.5 with a time constant of 5 s.

2.4 Continuous Music Feedback

Music is generated continuously by FluidSynth according to predefined rules, without lyrics, at a fixed tempo of 64 BPM and a 4/4 meter. The composition layer reads u at bar boundaries to change part dynamics, melodic register, rests, and arpeggiated note events. The acoustic layer reads u once every 1024 samples and smooths it with a coefficient of 0.10 before adjusting the low-pass cutoff frequency, high-shelf gain, and amplitude envelope. These three acoustic parameters are mapped as $350 + 7000u_a$ Hz, $-10 + 20u_a$ dB, and $-12 + 12u_a$ dB, respectively, where u_a is the control variable after smoothing in the acoustic layer. The composition layer changes only bars that have not yet been scheduled, preserving the phrase structure. The acoustic layer applies continuous modulation across audio blocks, allowing feedback without switching between music tracks.

The static condition uses the same engine, with u fixed at 0.5, corresponding to 3850 Hz, 0 dB, and -6 dB. The same random seed settings and basic generation rules are used across blocks, but the control variable changes the actual pitches and event sequences; a fixed tempo also does not imply a fixed density of note events. Music is sampled at 44.1 kHz, notes are scheduled 4.5 s in advance, and the sound card buffer is approximately 1 s. These are software settings; end-to-end audible latency was not independently measured by recording the audio output.

3 Experimental Methods

3.1 Experimental Design and Participants

This exploratory pilot study used a within-participant design to compare closed-loop music driven by the participant’s own real-time state (Self) with music using static parameters (Static). Each participant completed four blocks, two per condition. In the static condition, the control variable was fixed at 0.5; music settings for both conditions are described in Sect. 2.4.

Data were collected in 34 participant sessions, of which 4 did not yield complete formal records for technical reasons. Among the 30 participants with complete records, 1 was excluded according to the rule that the proportion of mind-wandering responses to attention probes reached or exceeded 40%, leaving 29 participants for behavioral and subjective analyses. Applying the additional criterion of raw EEG availability $\geq 60\%$ across the entire session retained 24 participants for the primary state analyses. This sample was aged 18–25 years and included 12 female participants, 11 male participants, and 1 in the other category. Behavioral data from the remaining 5 participants whose EEG did not meet the criterion were retained.

The four sequences actually used were TTSS, TSST, SSTT, and STTS, where T denotes static and S denotes closed-loop. Among the 30 participants, the corresponding counts were 12, 9, 6, and 3. Sequence allocation was unequal, with closed-loop blocks tending to occur later overall; sensitivity models therefore adjusted for block position and stimulus set. It cannot currently be confirmed that the analysis windows and exclusion thresholds were all determined before condition differences were inspected, so the corresponding results are reported as exploratory evidence. This study received ethics approval from the Faculty of Psychology at Beijing Normal University (BNU 202608240315), and participants provided written informed consent.

3.2 Materials and Task Procedure

The materials comprised 96 English word families, each consisting of one target word and one orthographically or phonologically similar lure, divided into four nonoverlapping word lists. Each block involved encoding 24 target words, followed by a recognition test comprising these 24 old words, 24 corresponding lures, and 12 unrelated new words; 8 additional nonoverlapping practice words were provided. During encoding, participants judged whether the referent of each word was animate, to promote semantic processing and record online responses [14].

Participants first completed the equipment self-check, individual calibration, and practice. Each formal block began with a 20 s silent period to establish the heart rate baseline. After the instructions, a keypress started the music, followed by a 25 s adaptation period. Each word was then presented in a sequence consisting of a 0.5 s fixation point, a 1.8 s word display, and a 1.0 s interval. Three attention-state probes, each lasting 4 s, were interspersed. Music was played throughout the adaptation and encoding periods.

Encoding was followed by a 30 s two-digit mental arithmetic filler task and a 60-item recognition test, both performed in silence. Recognition used old/new responses on a

1–4 scale, with a Remember/Know judgment requested after an old response [18]. After each block, participants rated the perceived responsiveness of the music, subjective arousal, pleasure, interference, effort, and familiarity on 1–9 scales, with higher scores indicating a greater degree of the corresponding attribute. The responsiveness item was “The music seemed to change along with me.” Participants also completed the NASA-TLX workload questionnaire [5]. Breaks between blocks lasted at least 90 s; including fitting and preparation, each session took approximately 50 min.

3.3 Measures and Statistical Analysis

For RQ1, the primary state outcome was the within-block standard deviation of the unfiltered fusion value y during the full feedback period (the 25 s adaptation period plus encoding). Analyses were also conducted for encoding alone and for different sample inclusions. The fusion value y was kept separate from the controller-side filtered estimate $\hat{y}$, and the mean and variability were evaluated separately. Block means and population standard deviations (ddof = 0) were first calculated from the time series and then averaged with equal weights across the two blocks of each condition within each participant. Supplementary measures included the mean absolute deviation from the fixed reference of 0.55 and the proportion of logged points within [0.52, 0.58]; corresponding comparisons relative to the individual calibration center were also conducted as supplementary analyses. For RQ3, single-channel means and variability, valid coverage, and actual weights were considered when interpreting changes in the composite index.

Offline EEG was linearly detrended in nonoverlapping segments of 500 samples (2 s) and screened using all of the following criteria: standard deviation $\geq 3 \mu\text{V}$, robust peak-to-peak amplitude (the difference between the 97.5th and 2.5th percentiles) $\leq 600 \mu\text{V}$, high-frequency power proportion ≤ 0.30 , and low-frequency power proportion ≥ 0.20 . Spectra were calculated using an FFT with a Hann window. High and low frequencies were defined as $20 \leq f < 45 \text{ Hz}$ and $1 \leq f < 13 \text{ Hz}$, respectively, with total power over $1 \leq f < 45 \text{ Hz}$ plus 10^{-9} as the denominator. The proportion of valid segments across the entire session was calculated separately for the two channels, and the higher value was used; participants with a value $\geq 60\%$ entered the primary state analyses. This rule determined sample inclusion and does not imply that the same segment-wise rejection was implemented during online fusion. Heart rate quality sensitivity analyses used $q > 0.4$ or $q \geq 0.6$, requiring at least 30 qualifying logged points per block and two qualifying blocks in each condition.

For RQ2, recognition analyses excluded timed-out or invalid responses. Lures and unrelated new words were pooled to calculate the false alarm rate. The calculation of d' used the Hautus log-linear correction [6], adding 0.5 to the hit and false alarm counts and 1 to the corresponding numbers of valid trials. Reaction times were restricted to valid, correct old/new judgments within 0.2–5 s. For each participant and condition, the two blocks were pooled and the median was calculated. Comparisons were conducted separately for hits and for correct rejections of all new words, lures, and unrelated new words.

All paired differences were defined as closed-loop minus static, and 95% confidence intervals based on participant-level data were reported. Six subjective ratings, three

composite-index measures (standard deviation, deviation from the fixed reference, and proportion within the fixed range), three recognition measures (d' , hit rate minus false alarm rate, and the proportion of Remember responses among hits), and the proportion of mind wandering formed a set of 13 outcomes. Both two-sided paired t -tests and Wilcoxon signed-rank tests were applied, with BH correction performed separately for each test type [1]; the state measures used 24 participants and the remaining measures used 29. The six NASA-TLX subscales underwent a separate BH correction. Holm correction [7] was applied separately to four mean measures (the composite index and three channels), three single-channel variability measures, four correct-response reaction time measures, and 15 online quality measures. Results were also examined with the three channel means and three variability measures combined into a family of six. Block-level sensitivity models included a participant random intercept and adjusted for block position and stimulus set.

4 Results

4.1 System Operation and Data Quality

Mean offline EEG availability among the 24 participants in the primary state analysis was 89%, with a range of 60%–99%. Online EEG availability was close to 100%, but its mechanism of retrieving the latest feature value differs from screening the raw waveforms; the two cannot be treated as the same quality measure. All three channels were simultaneously available online during 26.79% and 29.36% of cycles in the closed-loop and static conditions, respectively. The mean normalized weights were approximately 49% for facial arousal, 44% for EEG, and 6%–7% for heart rate. The composite index therefore relied mainly on facial and EEG inputs for most of the time.

In the closed-loop logs from the 24 participants in the primary sample, cycle-by-cycle recalculation of the target control variable yielded errors no greater than 0.006, and recalculation of the actual control variable after rate limiting yielded errors no greater than 0.002. The values received by the audio thread matched the controller output from the preceding cycle in 100% of cases. The within-block standard deviation of the closed-loop control variable was 0.091 ± 0.042 , whereas the static control variable remained at 0.5; neither the hold state nor the protection state was triggered during the experimental recordings. No condition differences were detected in any of the 15 online quality measures after Holm correction.

4.2 Composite Index Variability

During the full feedback period, the within-block standard deviation of the composite index was lower in the closed-loop condition than in the static condition, representing a 12.8% reduction relative to the static condition mean; 17 of the 24 participants showed the same direction of difference (Fig. 2, Table 2; $d_z = -0.65$). Both paired tests indicated the same direction of effect. After BH correction across 13 comparisons, the Wilcoxon test yielded $q = 0.0231$, whereas the paired t -test yielded $q = 0.0554$, which did not meet the 0.05 threshold. The condition coefficient in the mixed model including

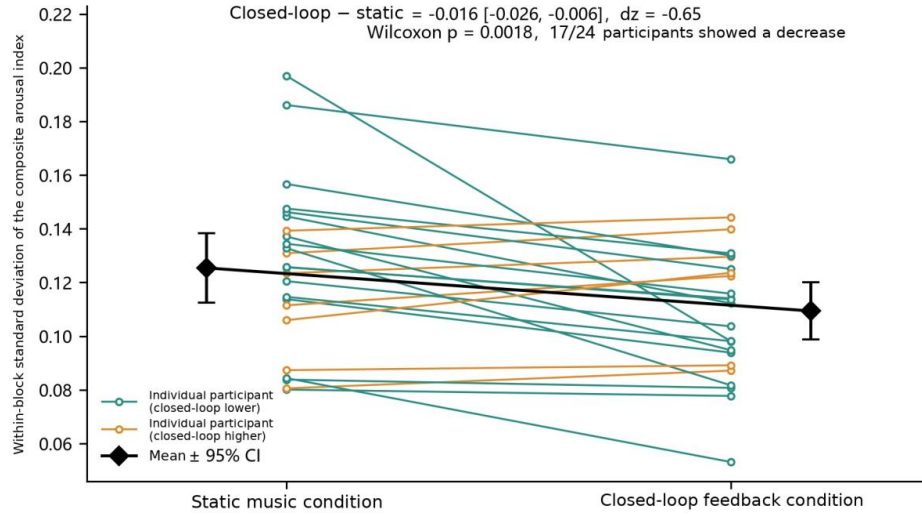


Fig. 2. Paired within-participant comparison of the within-block standard deviation of the composite index during the full feedback period ($N = 24$). Each point represents the equally weighted average of a participant's two blocks in the same condition, and lines connect observations from the same participant. Black diamonds and error bars indicate group means and 95% CIs. The analysis used fusion values y before controller-side filtering; after BH correction across 13 comparisons, the paired t -test yielded $q = 0.0554$ and the Wilcoxon test yielded $q = 0.0231$. Full numerical and test results are provided in Table 2.

block position and stimulus set was -0.017 , $p = 0.0003$. The primary result was based on the unfiltered fusion value y and cannot simply be attributed to smoothing by the controller-side filter.

For the encoding period alone, the paired difference in standard deviation was -0.0102 , 95% CI $[-0.0205, 0.0002]$, with a paired t -test $p = 0.054$; after adjustment for block position and stimulus set, the model coefficient was -0.012 , $p = 0.019$. Compared with the full feedback period, the effect estimate was smaller and the conclusion was more sensitive to the analysis model. When the sample was expanded to the 29 participants included in the behavioral analysis, the paired difference was approximately -0.011 , 95% CI $[-0.021, -0.0005]$, with a t -test $p = 0.041$. Across all 30 complete recordings, the paired difference was approximately -0.008 , 95% CI $[-0.019, 0.003]$, $p = 0.139$. The results had the same direction across samples, but the strength of the evidence decreased.

After real-time quality weights were replaced with fixed participant-specific weights, the paired difference in the 24 participants became smaller in magnitude, with an estimate of -0.0094 , 95% CI $[-0.0219, 0.0032]$, $p = 0.1356$. This exploratory recalculation retained low-quality heart rate values and does not constitute a more reliable physiological reference; the results suggest that the definition of the index can influence differences in variability.

Table 2. Main state and task results

Measure and sample	Closed-loop $M \pm SD$	Static $M \pm SD$	Closed-loop minus static Difference [95% CI]	Paired tests and adjustment for multiple comparisons
Composite index standard deviation $N = 24$	0.1094 ± 0.0253	0.1254 ± 0.0305	-0.0160 [-0.0265, -0.0056]	$t p = 0.0043$, BH $q = 0.0554$; $W p = 0.0018$, BH $q = 0.0231$
Composite index mean $N = 24$	0.5053 ± 0.1475	0.5325 ± 0.1549	-0.0272 [-0.0604, 0.0060]	$t p = 0.1033$; Holm-adjusted $p = 0.3098$
Facial arousal standard deviation $N = 24$	0.0327 ± 0.0167	0.0391 ± 0.0196	-0.0065 [-0.0113, -0.0016]	$t p = 0.0115$, Holm-adjusted $p = 0.0344$; $W p = 0.0048$
Recognition discriminability d' $N = 29$	2.520 ± 0.756	2.593 ± 0.493	-0.073 [-0.313, 0.166]	$t p = 0.5357$, BH $q = 0.7890$; $W p = 0.2748$, BH $q = 0.5104$
Reaction time for hits (ms) $N = 29$	1255.8 ± 307.9	1301.6 ± 338.5	-45.8 [-116.0, 24.3]	$t p = 0.1918$; Holm-adjusted $p = 0.7671$
Perceived responsiveness of the music $N = 29$	3.621 ± 1.860	3.103 ± 1.961	+0.517 [0.070, 0.965]	$t p = 0.0250$, BH $q = 0.1627$; $W p = 0.0326$, BH $q = 0.2121$

The closed-loop and static columns report participant-level $M \pm SD$. Within-block standard deviation is an outcome measure; the value following “ $\pm$ ” in these rows still describes between-participant dispersion. For reaction time, M is the group mean of participants’ within-condition medians. All differences are closed-loop minus static, and brackets contain the 95% CI for the paired difference. W denotes the Wilcoxon test. Each family of tests was adjusted separately as specified in the Methods; for facial arousal variability, the Holm-adjusted p was 0.0689 when the six comparisons of means and variability were adjusted together.

4.3 Mean Levels and Results for Individual Channels

The difference in the mean level of the composite index did not reach statistical significance, and the confidence interval still allowed for a lower mean in the closed-loop condition (Table 2); the mean level therefore cannot be considered unchanged. No condition differences were detected in the mean absolute deviations from the fixed reference or the individually calibrated center, or in the proportions within the corresponding ranges. A reduction in standard deviation alone is therefore insufficient to infer improved target tracking.

The within-block standard deviation of facial arousal was lower in the closed-loop condition, with a Holm-adjusted $p = 0.0344$ across the three comparisons of variability in individual channels. When the mean and variability comparisons across the three channels were combined into a six-comparison family, the adjusted p was 0.0689, indicating that the strength of the evidence depended on the definition of the test family. Heart rate standard deviation was 18.868 beats/min in the closed-loop condition and 18.277 beats/min in the static condition ($p = 0.2268$); the standard deviation of the EEG power ratio was 0.2928 and 0.2944, respectively ($p = 0.9808$). On the original scales and in the full primary sample, variability differences in individual channels were mainly observed in the facial model output.

No significant condition differences were detected in the mean levels of the three individual channels. This conclusion was unchanged after Holm correction across the four comparisons of mean heart rate, EEG power ratio, facial arousal, and composite index.

When heart rate estimates were screened using $q > 0.4$ and two valid blocks per condition were required, only 13 participants met the criteria, and no differences were detected in either their mean heart rate or its standard deviation. At $q \geq 0.6$, only 3 participants remained, which was insufficient to support reliable population-level inference. The standard deviation here describes variability in the quality-screened rPPG heart rate estimates, rather than heart rate variability calculated from interbeat intervals.

4.4 Recognition Performance and Subjective Experience

Among the 29 participants, recognition discriminability d' did not show a closed-loop benefit; the confidence interval for the paired difference encompassed both poorer and better performance. Although reaction time for hits was numerically shorter, the difference was not significant either before or after Holm correction across the four comparisons (Table 2). No differences were detected in the other three types of correct-rejection reaction time. Hit rate, the false-alarm rate for lures, and the false-alarm rate for unrelated new words also showed no significant condition differences. The current results therefore do not support improved discriminability or faster correct responses.

Immediate ratings of perceived responsiveness of the music ("The music is following me") were higher in the closed-loop condition, but the BH q values were 0.1627 for the t -test and 0.2121 for the Wilcoxon test, neither of which reached significance. The remaining subjective ratings, the Remember proportion, hit rate minus false-alarm rate, and the mind-wandering proportion also did not reach significance after this correction.

Among the 13 comparisons, only the Wilcoxon test of the composite index standard deviation reached $q < 0.05$. No differences were detected in any of the six NASA-TLX subscales after BH correction within the respective test families; thus, no clear evidence of reduced subjective workload was obtained.

5 Discussion

5.1 System Implementation and Exploratory State-Index Findings

For RQ1, the primary state-analysis sample of 24 participants showed lower composite index variability under closed-loop feedback over the full feedback period. The two tests yielded results in the same direction, but only the Wilcoxon test reached significance after BH correction. These findings provide exploratory evidence of a condition-related difference in the variability of the implemented system’s composite index during word learning. Previous word-memory research observed an association between pupil variability across word lists and recall performance [16]; the present study compares feedback conditions to characterize the composite index while the system operates during word learning. Both address temporal change as a dimension of interest, but between-participant associations and within-participant interventions answer different questions; the present results do not demonstrate that the association is causal.

The individual reference center allows feedback to operate according to each participant’s calibration distribution, but its functional role in learning remains to be tested. The controller responds to deviations from the center, whereas the standard deviation describes the resulting distribution; these are not equivalent. The confidence interval for the difference in the mean composite index still allows a decrease, and neither deviation from the center nor the proportion of values within the range showed a clear improvement. The current results therefore cannot confirm regulation of variability while maintaining an unchanged mean, or more accurate target tracking. These findings suggest that evaluation of a strategy that regulates states around an individual center requires examination of both the task relevance of the reference state and the resulting changes in the distribution.

5.2 Measurement Interpretation of State Variability

RQ3 concerns the theoretical interpretations that a variability measure can support. This study evaluates outcomes using the fusion value y before controller-side filtering, but feature windows, standardization, clipping, and quality weights still contribute to its construction. The EEG power ratio is a candidate engagement index [15], rPPG heart rate is an estimate of a peripheral physiological variable, and facial arousal is the output of an appearance-based model; the three are not equivalent readouts of the same latent process. Previous pupillometry research [16, 19, 20] provides a cognitive rationale for examining variability, but cannot substitute for validation of the current composite index. The coefficient of variation in pupil diameter across word lists and the within-block standard deviation of fusion values also have different timescales and statistical meanings.

Differences in single-channel variability were observed mainly for facial arousal, while valid-data coverage and the actual weight of heart rate were low; the estimated composite effect was smaller when recalculated using fixed weights. These results indicate that changes in y must be interpreted in terms of both the input signals and the composition of the inputs. Previous music research reported slower breathing together with increased variability in inter-breath intervals [11], further suggesting that the functional meaning of “variability” must be evaluated separately for different measures. The current evidence supports continued investigation of changes in the defined index, but is insufficient to establish stabilization across multiple physiological systems or to interpret lower variability directly as a better learning state.

5.3 Exploratory Recognition Outcomes

For RQ2, the behavioral comparison in 29 participants did not detect improvements in recognition discriminability or reaction time for hits, and the confidence intervals still allow some improvement or deterioration. This study therefore did not obtain evidence of learning benefits from the current feedback strategy. Preliminary evidence of a change in index variability and evidence supporting a learning benefit must be evaluated separately. These exploratory recognition findings constrain the interpretation of the index difference; direct evidence is still needed to establish its functional value for learning.

Future validation should prioritize the key uncertainties in measurement and intervention interpretation. The current comparison of closed-loop and static music involves both temporal variation in the music and the link between the music and the participant’s own state, and has not separated their contributions. The primary state results are sensitive to the sample analyzed and the analysis window, and condition order was not fully counterbalanced. Independent studies should prespecify quality rules, analysis windows, and statistical plans, use dynamic yoked controls or randomized control probes, and simultaneously record independent physiological reference measures and the actual audio. Future work should retain complete multimodal sequences from calibration and additionally validate sensing accuracy, the headband’s performance as a complete device, and fallback behavior under forced input loss. These tests can clarify the state changes and effects of feedback, after which learning value can be evaluated using task-appropriate reference states and delayed tests.

6 Conclusion

This exploratory pilot study implemented an individually calibrated multimodal closed-loop music system and conducted a word-learning and recognition experiment to examine its operational feasibility and measurement limitations. The primary state-analysis sample of 24 participants provided preliminary evidence of reduced composite index variability over the full feedback period, with different conclusions from the two tests after correction for multiple comparisons; no recognition benefit was detected in the behavioral sample of 29 participants. The study’s contribution is system implementation and assessment of the composite index’s measurement limitations, using mean levels,

individual channels, and quality results to constrain interpretation. The present findings do not establish a learning benefit. Further validation using independent measurements and cognitive outcomes is required to assess the index's value as a target for learning interventions.

References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* **57**(1), 289–300 (1995). <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
2. Bergstrom, I., Seinfeld, S., Arroyo-Palacios, J., Slater, M., Sanchez-Vives, M.V.: Using music as a signal for biofeedback. *International Journal of Psychophysiology* **93**(1), 140–149 (2014). <https://doi.org/10.1016/j.ijpsycho.2013.04.013>
3. Ehrlich, S.K., Agres, K.R., Guan, C., Cheng, G.: A closed-loop, music-based brain-computer interface for emotion mediation. *PLoS ONE* **14**(3), e0213516 (2019). <https://doi.org/10.1371/journal.pone.0213516>
4. Fairclough, S.H.: Fundamentals of physiological computing. *Interacting with Computers* **21**(1–2), 133–145 (2009). <https://doi.org/10.1016/j.intcom.2008.10.011>
5. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): Results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (eds.) *Human Mental Workload*, pp. 139–183. North-Holland, Amsterdam (1988). [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
6. Hautus, M.J.: Corrections for extreme proportions and their biasing effects on estimated values of d' . *Behavior Research Methods, Instruments, & Computers* **27**(1), 46–51 (1995). <https://doi.org/10.3758/BF03203619>
7. Holm, S.: A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**(2), 65–70 (1979)
8. Kalman, R.E.: A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* **82**(1), 35–45 (1960). <https://doi.org/10.1115/1.3662552>
9. Kingsense Electronics: KS108X and KS109X Series Chip Register Configuration Guide (2023), version 1.2
10. Kingsense Electronics: KS1091/KS1092: Low-Noise 1-, 2-Channel Analog Front-End for EEG and Biopotential Measurements (2024), data Sheet, Rev. 1.5
11. Leslie, G., Ghandeharioun, A., Zhou, D.Y., Picard, R.W.: Engineering music to slow breathing and invite relaxed physiology. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII) (2019), <https://arxiv.org/abs/1907.08844>, arXiv:1907.08844
12. Mather, M., Sutherland, M.R.: Arousal-biased competition in perception and memory. *Perspectives on Psychological Science* **6**(2), 114–133 (2011). <https://doi.org/10.1177/1745691611400234>
13. Meissner, S.N., Bächinger, M., Kikkert, S., et al.: Self-regulating arousal via pupil-based biofeedback. *Nature Human Behaviour* **8**(1), 43–62 (2024). <https://doi.org/10.1038/s41562-023-01729-z>
14. Nairne, J.S., VanArsdall, J.E., Pandeirada, J.N.S., Cogdill, M., LeBreton, J.M.: Adaptive memory: The mnemonic value of animacy. *Psychological Science* **24**(10), 2099–2105 (2013). <https://doi.org/10.1177/0956797613480803>
15. Pope, A.T., Bogart, E.H., Bartolome, D.S.: Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology* **40**(1–2), 187–195 (1995). [https://doi.org/10.1016/0301-0511\(95\)05116-3](https://doi.org/10.1016/0301-0511(95)05116-3)

16. Robison, M.K., Trost, J.M., Schor, D., Gibson, B.S., Healey, M.K.: Pupillary correlates of individual differences in long-term memory. *Psychonomic Bulletin & Review* **29**(4), 1355–1366 (2022). <https://doi.org/10.3758/s13423-022-02081-5>
17. Ros, T., Enriquez-Geppert, S., Zotev, V., et al.: Consensus on the reporting and experimental design of clinical and cognitive-behavioural neurofeedback studies (CRED-nf checklist). *Brain* **143**(6), 1674–1685 (2020). <https://doi.org/10.1093/brain/awaa009>
18. Tulving, E.: Memory and consciousness. *Canadian Psychology* **26**(1), 1–12 (1985). <https://doi.org/10.1037/h0080017>
19. Unsworth, N., Robison, M.K.: The importance of arousal for variation in working memory capacity and attention control: A latent variable pupillometry study. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **43**(12), 1962–1987 (2017). <https://doi.org/10.1037/xlm0000421>
20. Unsworth, N., Robison, M.K., Miller, A.L.: Pupillary correlates of fluctuations in sustained attention. *Journal of Cognitive Neuroscience* **30**(9), 1241–1253 (2018). https://doi.org/10.1162/jocn_a_01251
21. Wang, W., den Brinker, A.C., Stuijk, S., de Haan, G.: Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering* **64**(7), 1479–1491 (2017). <https://doi.org/10.1109/TBME.2016.2609282>
22. Welch, P.D.: The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics* **15**(2), 70–73 (1967). <https://doi.org/10.1109/TAU.1967.1161901>
23. Zander, T.O., Kothe, C.: Towards passive brain–computer interfaces: Applying brain–computer interface technology to human–machine systems in general. *Journal of Neural Engineering* **8**(2), 025005 (2011). <https://doi.org/10.1088/1741-2560/8/2/025005>