Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.25987

Audiovisual Action Aligned Adaptive Feedback System for Interactive Virtual Piano Performance Tutor

Xinyu Cao

Xi'an Traffic Engineering Institute, Xi'an, China
470957522@qq.com

Abstract. With the rapid advancement of computer vision, audio processing, and artificial intelligence technologies, virtual music education systems have increasingly demonstrated significant advantages in piano instruction by overcoming temporal and spatial constraints, enabling personalized guidance, and promoting equitable access to resources. However, most existing virtual piano tutoring systems remain focused primarily on capturing and analyzing audio information, neglecting the crucial role of the performer's physical movements in technical execution and musical expression. This leads to feedback that is often limited and unidimensional. Furthermore, the lack of efficient multimodal alignment mechanisms hinders accurate identification of learners' overall performance. This study proposes an adaptive feedback system based on audiovisual action alignment, aiming to integrate visual and auditory information through standard performance templates and real-time representation analysis to deliver multidimensional and precise feedback on learners' piano playing behaviors. Experimental results show that, compared with traditional video-based instruction, learners using this system exhibited significant improvements in rhythm accuracy and fingering correctness, with all measured indicators reaching statistical significance. The research not only offers a novel technical solution for virtual music education by addressing the limitations of conventional systems in guiding technical actions, but also provides an important theoretical foundation for the design and optimization of intelligent tutoring systems.

Keywords: Virtual Piano Tutor, Audiovisual Alignment, Adaptive Feedback, Computer Vision, Interactive Music Learning

1. Introduction

As the computer vision, audio processing, and machine intelligence technologies expanded at a rapid pace, the virtual music education systems have increasingly shown potential in traditional music teaching. More especially in the domain regarding teaching the piano, they have shown salient advantages in surmounting spatiotemporal limitations, providing personalized guidance, and fairness in resources [1]. At the present time, most of the virtual systems in teaching the piano still focus on the recording and processing of audio data immensely but neglect the crucial role of the movement of the performer in technical play and expression in music [2]. As such, the resultant feedback is normally solitary and narrow. Furthermore, the inefficacy in the multimodal alignment

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.25987

mechanism renders the systems incapable in accurately identifying the performance of the learners in general and thus limit the specificity and usability of the feedback [3]. To address the above shortcomings, the current study creates an adaptive system for audiovisual action alignment feedback. It aims to integrate visual and auditory information, construct standard performance templates, and conduct real-time representation analysis to achieve multidimensional and precise feedback on learners' piano performance behaviors. The system design fully considers the requirements of teaching interactivity and technical feasibility, explores the integration path of audiovisual collaboration in virtual piano instruction, and expands the application scope of intelligent music education.

2. Literature review

2.1. Virtual music tutors and piano pedagogy

The development of virtual music tutors has promoted the transition of piano education from traditional face-to-face instruction to digital platforms, yet most systems remain at the level of symbolic recognition and basic rhythm correction, lacking deep perception of performance actions and expressive details [4]. Mainstream platforms such as Simply Piano and Yousician mostly evaluate performance quality through pitch and rhythm comparison, although they possess certain real-time feedback capabilities, they overlook essential performance elements such as finger positioning, posture transitions, and key touch control [5]. This feedback mechanism, which heavily relies on audio signals, is insufficient for guiding technical execution and undermines the overall instructional integrity of the system [6]. In piano education, the accuracy of technical movements directly determines the controllability and developmental potential of musical expression. Therefore, teaching systems that rely solely on audio analysis cannot meet the needs of systematic instruction.

2.2. Audiovisual synchronization in music performance

Audiovisual synchronization technology provides a multimodal and cross-dimensional assessment path for music performance analysis. In performance learning scenarios, the temporal alignment between playing actions and sounds reflects not only learners' understanding of rhythm but also their technical coordination and expressive control [7]. Studies have shown that applying computer vision to identify gesture trajectories and matching them with audio time sequences can effectively detect potential rhythm errors and fingering mistakes, enhancing the timeliness and relevance of feedback. Especially in piano performance, the synchronization of both hands, the timing of key strikes, and the variations in touch strength constitute key dynamic features of sound [8]. If action and audio can be jointly analyzed, a more comprehensive learning assessment model can be constructed. Some studies have applied dynamic time warping and optical flow analysis to achieve preliminary action-audio alignment, but most methods remain at the level of low-level feature matching.

2.3. Adaptive feedback models in intelligent tutoring systems

Adaptive feedback models analyze and model learners' real-time behavioral data to generate personalized learning paths and adjust difficulty levels. In music instruction, higher demands are placed on the real-time and expressive nature of feedback, particularly in perceiving the nonlinear and dynamic processes of performance. Existing models often struggle to accurately detect performance deviations and provide flexible corrections. Some studies have attempted to introduce

DOI: 10.54254/2755-2721/2025.25987

reinforcement learning, Bayesian networks, and rule-based engines into music tutoring systems to enhance the diversity and adaptability of feedback strategies, but they still fall short in processing multimodal inputs, identifying ambiguous performance errors, and generating semantically clear feedback [9]. Furthermore, the expressive and technical dimensions of music learning are closely intertwined. If adaptive feedback models cannot simultaneously recognize pitch and rhythm deviations along with motion errors, their instructional function in practice will be limited [10].

3. Methodology

3.1. System architecture design

This study presents a multimodal virtual piano tutoring system that integrates image capture, audio input, and adaptive feedback to comprehensively sense and guide learners' performance. The visual module applies OpenPose for real-time gesture recognition, accurately tracking hand movements, while the audio module uses MFCC to extract pitch and rhythm features. At the core, a dynamic time warping-based alignment module synchronizes multimodal data to detect timing and technique errors. The adaptive feedback engine generates personalized responses through visual markers and voice prompts, providing real-time, interactive guidance. The modular architecture balances responsiveness and scalability, offering a robust foundation for intelligent music education systems. As shown in Figure 1.

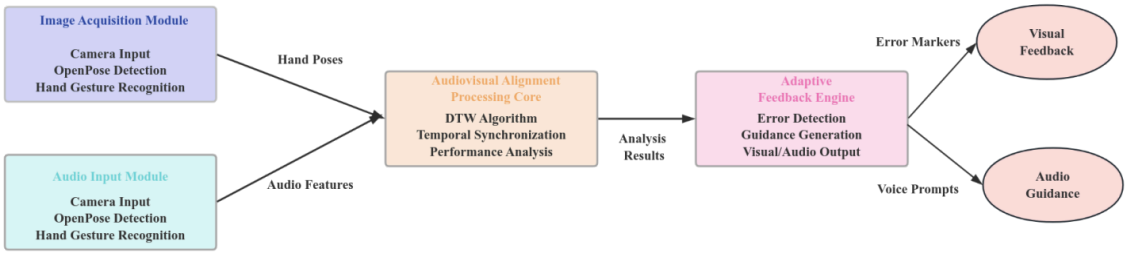


Figure 1. Irtual piano tutor system architecture

3.2. Motion and audio alignment algorithm

The motion and audio alignment algorithm serves as the core technical component for realizing multimodal instructional feedback [11]. This study applies an improved dynamic time warping algorithm to process the temporal synchronization problem during the performance process. The distance computation formula for dynamic time warping is defined as follows:

$$D(i,j)=d(x_i,y_j)+\min\{D(i-1,j),D(i,j-1),D(i-1,j-1)\} \quad (1)$$

Where x_i represents the feature vector at the i-th time point in the gesture sequence, y_j represents the MFCC coefficient at the j-th time point in the audio feature sequence, and $d(x_i,y_j)$ denotes the Euclidean distance function between two feature vectors.

To improve the alignment accuracy, the system further introduces a weighted time annotation mechanism to establish a semantic association model for audio and video features by analyzing the note start and stop times in the standard performance template with the timestamps of the corresponding gesture actions. The weighted alignment loss function is denoted as:

DOI: 10.54254/2755-2721/2025.25987

$$L_{align}=\sum_{t=1}^{T} w_t \cdot \|f_{visual}(t)-f_{audio}(t+\delta_t)\|^2 \quad (2)$$

Where w_t is the importance weight of time point t, $f_{visual}(t)$ and $f_{audio}(t)$ represent the visual and audio feature vectors at time ttt, and δ_t denotes the time offset calculated by the system.

3.3. Adaptive feedback generation strategy

The adaptive feedback generation strategy is the key mechanism for transforming the results of audiovisual alignment analysis into effective instructional guidance. The system dynamically evaluates the learner's performance level and intelligently adjusts the granularity and presentation of the feedback content. The adaptive feedback weight calculation formula is defined as follows:

$$W_{feedback}=\alpha \cdot E_{rhythm}+\beta \cdot E_{gesture}+\gamma \cdot E_{pitch} \quad (3)$$

Where E_{rhythm}, $E_{gesture}$, and E_{pitch} represent the quantified severity metrics of errors in rhythm, gesture, and pitch respectively. The weighting coefficients α, β, and γ are dynamically adjusted based on the learner's historical performance data to ensure that the feedback content matches individual learning characteristics and current skill levels.

The decision function for feedback generation adopts a piecewise linear model:

$$F(x)=\begin{cases} F_{basic}(x) & \text{if } x<\theta_1 \\ F_{intermediate}(x) & \text{if } \theta_1 \le x<\theta_2 \\ F_{advanced}(x) & \text{if } x \ge \theta_2 \end{cases} \quad (4)$$

Where x denotes the comprehensive performance score, θ_1 and θ_2 are preset skill level thresholds, and F_{basic}, $F_{intermediate}$, and $F_{advanced}$ correspond to feedback strategy functions of different difficulty levels. The system continuously monitors the learner's performance progress and error distribution patterns, dynamically adjusting the frequency and form of feedback presentation. It highlights incorrect gesture positions in the visual interface and provides targeted improvement suggestions using speech synthesis technology.

4. Results

4.1. Evaluation metrics and setup

The study adopted a randomized controlled trial design to evaluate the instructional effectiveness of the virtual piano tutoring system. Thirty volunteers were recruited from the School of Music at a local normal university. All participants were rigorously screened to ensure they had no prior piano training and had not received more than two hours of keyboard instruction. A balanced randomization procedure was used to assign participants into two groups, with 15 in the experimental group and 15 in the control group. Each participant practiced on the same model of digital piano, the Yamaha P-125. Visual data were captured using a Logitech C920 Pro HD camera, mounted 1.2 meters directly above the keyboard to ensure full visibility of both hands during performance. Audio was recorded using an Audio-Technica AT2020 professional microphone, with a sampling rate of 44.1 kHz and placed 20 centimeters from the keys to minimize background noise interference.

DOI: 10.54254/2755-2721/2025.25987

The practice material was standardized to a selected segment from Bach's Minuet in G Major, BWV Anh. 114, with a fixed practice duration of 20 minutes. The experimental group used the audiovisual action aligned adaptive feedback system developed in this study, while the control group followed traditional piano instructional videos. A double-blind evaluation procedure was employed, where three experienced piano instructors independently rated the performance outcomes using a unified scoring rubric to ensure objectivity and fairness.

4.2. Precision improvement effect

The analysis of experimental data indicates that learners using the audiovisual action aligned adaptive feedback system achieved significant improvements in key skill indicators, as shown in Table 1. Statistical analysis was conducted using independent samples t-tests with a significance level set at $\alpha = 0.05$. All data were processed using SPSS 26.0 software.

Table 1. Comparison of performance skill indicators between the two groups of learners

Evaluation Metric	Experimental Group (n=15)	Control Group (n=15)	Improvement	t Value	p Value	Effect Size (Cohen's d)
Rhythm Accuracy (%)	78.6 ± 8.4	55.2 ± 9.1	23.4	4.23	< 0.001 ***	2.68
Fingering Accuracy (%)	84.9 ± 6.7	65.3 ± 8.5	19.6	3.87	< 0.001 ***	2.54
Note Recognition Rate (%)	82.4 ± 7.2	63.8 ± 9.4	18.6	3.52	< 0.01 **	2.21
Overall Fluency Score	7.3 ± 1.1	5.8 ± 1.3	1.5	3.14	< 0.01 **	1.98
Practice Focus Duration (min)	18.7 ± 1.2	16.4 ± 2.1	2.3	2.89	< 0.01 **	1.85

In addition, in-depth analysis revealed that the most notable improvement in the experimental group was in rhythm control. The accuracy rate for quarter note beats reached 85.3 percent, and for eighth notes 74.2 percent, both significantly higher than the control group's 61.8 percent and 48.7 percent, respectively. In terms of fingering technique, the experimental group demonstrated a finger independence error rate of only 12.4 percent, a substantial reduction compared to 29.7 percent in the control group. This improvement was particularly evident in the coordination of the ring and little fingers. The enhancement in note recognition accuracy was concentrated in the dimensions of pitch identification and duration judgment, with the experimental group achieving 84.1 percent and 80.6 percent, respectively, significantly outperforming the control group's 64.2 percent and 63.4 percent.

5. Discussion

The experimental results demonstrate that the audiovisual action aligned adaptive feedback system significantly enhances instructional effectiveness in virtual piano education, validating the application value of multimodal fusion technologies in the field of music education. Compared with traditional audio-only analysis systems, the proposed system integrates visual gesture recognition and audio feature extraction to provide a more comprehensive understanding of learners' performance behaviors, thereby enabling more precise technical guidance. These findings for better rhythm accuracy and fingering appropriateness demonstrate the motion-audio synchronization mechanism in real-time effectively addresses the shortcomings in the guidance in technical

DOI: 10.54254/2755-2721/2025.25987

performance by existing systems. The research has some caveats, though. The samples were very small in size, and the practice was short in duration. A subsequent study should consider the stability and transferability in large-scale and long-duration implementations.

6. Conclusion

This study successfully developed a virtual teaching system for the piano through audiovisual action alignment by integrating computer vision, audio processing, and machine learning technologies in order to achieve multidimensional perception and adaptive feedback on students' piano performance behaviors. The core innovations of the system lie in the improved dynamic time warping algorithm and the hierarchical feedback generation strategy, which effectively address the limitations and singularity of feedback in traditional virtual instruction systems. The research outcomes provide not only a novel technical solution for the field of virtual music education but also an important reference for the design and optimization of intelligent tutoring systems.

References

[1] Wilson, Katie, and Phillip E. Pfeiffer. "Feedback in augmented and virtual reality piano tutoring systems: a mini review." Frontiers in Virtual Reality 4 (2023): 1207397.
[2] Simonetta, Federico, Stavros Ntalampiras, and Federico Avanzini. "Acoustics-specific Piano Velocity Estimation." 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2022.
[3] Amm, Vivienne, et al. "Mixed reality strategies for piano education." Frontiers in Virtual Reality 5 (2024): 1397154.
[4] Kato, Norihiro, et al. "Computational Analysis of Audio Recordings of Piano Performance for Automatic Evaluation." European Conference on Technology Enhanced Learning. Cham: Springer Nature Switzerland, 2023.
[5] Liu, Lei. "Research on Piano Program Programming and Creative Optimization Methods Based on Big Data Technology." International Journal of High Speed Electronics and Systems 34.02 (2025): 2440090.
[6] Nusseck, Manfred, et al. "Effects of audio feedback interventions with the Disklavier on the performance of piano students." Frontiers in Psychology 16 (2025): 1568021.
[7] Gopal D. Upadhye, Satpalsing Rajput, Shalini V. Wankhade, Ajinkya Walunj, Akash Bhandari, Aneesh Dighe and Anish Sagri. "Virtual Piano using Assembly Language Programming." Communications on Applied Nonlinear Analysis (2024).
[8] Lange-Nawka, Dominik, Burkhard C. Wünsche, and Samuel ER Thompson. "Auditory temporal hints in AR piano tutoring." 2023 38th International Conference on Image and Vision Computing New Zealand (IVCNZ). IEEE, 2023.
[9] Liu, Yuru. "BP Neural Network and Computer Virtual Reality Technology in Piano Performance Practice." 2023 2nd International Conference on Artificial Intelligence and Computer Information Technology (AICIT). IEEE, 2023.
[10] Zhang, Yuting, et al. "Computer Simulation of Piano Tones and Design of Virtual Piano System." Applied Mathematics and Nonlinear Sciences 9.1 (2024).
[11] Liu, Ruofan, et al. "Synchronized Hand Difference Visualization for Piano Learning." ACM SIGGRAPH 2022 Posters. 2022. 1-2.