Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
Special Issue on Digital Preservation and Intelligent Innovation in Traditional and Modern Arts 7(1s)

VOICE RECOGNITION AI IN MUSIC EDUCATION PLATFORMS

Sunil Damodar Rathod [1], Nidhi Tewatia [2], Dr. Srijita Bhattacharjee [3], Nivetha N. [4], Amruta Prasad Kharade [5], Shikha Verma Kashyap [6]

[1] Associate Professor, Department of Computer Engineering, Indira College of Engineering and Management, Parandwadi, Pune, Maharashtra, India
[2] Assistant Professor, School of Business Management, Noida International University, Greater Noida 203201, India
[3] Assistant Professor, Department of Computer Science and Engineering, Bharati Vidyapeeth (Deemed to be University), Department of Engineering and Technology, Sector 3 Belpada, Kharghar, Navi Mumbai, 410210, India
[4] Assistant Professor, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu 600092, India
[5] Department of Engineering, Science and Humanities, Vishwakarma Institute of Technology, Pune, Maharashtra, 411037, India
[6] Professor, AAFT University of Media and Arts, Raipur, Chhattisgarh-492001, India

ABSTRACT

Voice recognition artificial intelligence (AI) can be described as a radical technology in music education platforms, which allows personalized, data-driven and scalable experiences in vocal training. The conventional method of music pedagogy is based on teacher-directed feedback, which is sometimes time-consuming, subjective, and hard to compare with different groups of learners. Conversely, systems that use voice recognition take advantage of the developments in signal processing, machine learning, and deep neural networks to recognize vocal pitch, timbre, rhythm, articulation, and pronunciation with high-temporal resolution. This paper gives an in-depth approach to the incorporation of voice recognition AI into music education platforms in terms of system architecture, approach to method, and impact on education. The suggested method is inclusive of strong audio data capture, noise sensitive preprocessing, and feature responses including Mel-frequency cepstral coefficients, pitch contours to characterize musical voice. These are supervised and deep learning-based recognition models that are used to measure the performance of the voice and provide real-time corrective feedback. Practical testing shows that AI-enhanced systems are more accurate and responsive with regards to pitch correction, rhythm matching, and diction measurement than traditional methods of teaching, and enhance engagement and independent practice in learners. In addition to performance benefits, the study indicates the pedagogical benefits of continuous feedback, adaptation in difficulty and the objective assessment.

Keywords: Voice Recognition AI, Music Education Platforms, Vocal Pitch Analysis, Rhythm Assessment, Intelligent Tutoring Systems, Digital Music Pedagogy

Received 13 September 2025
Accepted 12 December 2025
Published 17 February 2026

Corresponding Author
Sunil Damodar Rathod,
sunil6kr@gmail.com

DOI
10.29121/shodhkosh.v7.i1s.2026.7111

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

1. INTRODUCTION

The teaching of music has always been based on intensive communication between students and professional educators, especially in the field of singing when nuances of changes in pitch, tone, rhythm, and pronunciation are significant factors in defining the quality of the performances. The classical teaching methods primarily focus on demonstration, imitation, and repetitive feedback which is mostly provided in terms of one-on-one or small group learning. Although it is effective, this model also has a limitation in terms of limited availability of instructors, subjective assessment and difficulty in offering on-going, personalized feedback, particularly in big classroom settings, in remote learning scenarios, or in environments of self-directed learning. These constraints have only gotten more apparent with the fast-growing music education platforms that exist online and distributed through technologies. Recent developments in artificial intelligence (AI) and especially voice recognition and speech analysis, offer new opportunities to solve these problems Mamyrbayev et al. (2022). Voice recognition AI is defined as the computational systems that are able to capture, process and understand human vocal cues through signal processing and machine learning techniques. Such systems can be applied to music education to go beyond simple speech recognition to analyze musical features including pitch accuracy, intonation stability, rhythmic timing, articulation and vocal timbre. This can be used to objectively, fine-grain analyze vocal performance, which was previously only available to the expert human evaluation Ahlawat et al. (2025). Voice recognition AI integration into music educational systems is in line with the wider trends of intelligent tutoring systems and personalized learning. In contrast to fixed digital resources or pre-recorded educational materials, AI-powered systems may be adjusted in real-time to the profiles of particular learners, their skill level, and course of development. These systems can be used to offer real-time corrective feedback, draw attention to repeated mistakes and prescribe specific exercises by constantly analyzing vocal input Zhang et al. (2023). This not only helps to acquire skills more efficiently but also helps foster learner autonomy by letting the students to practice without being guided by the instructor, but with guidance provided. Figure 1 depicts a voice recognition system facilitated by AI when teaching music. Scalability is another reason why voice recognition AI is a good idea in music education.

Figure 1

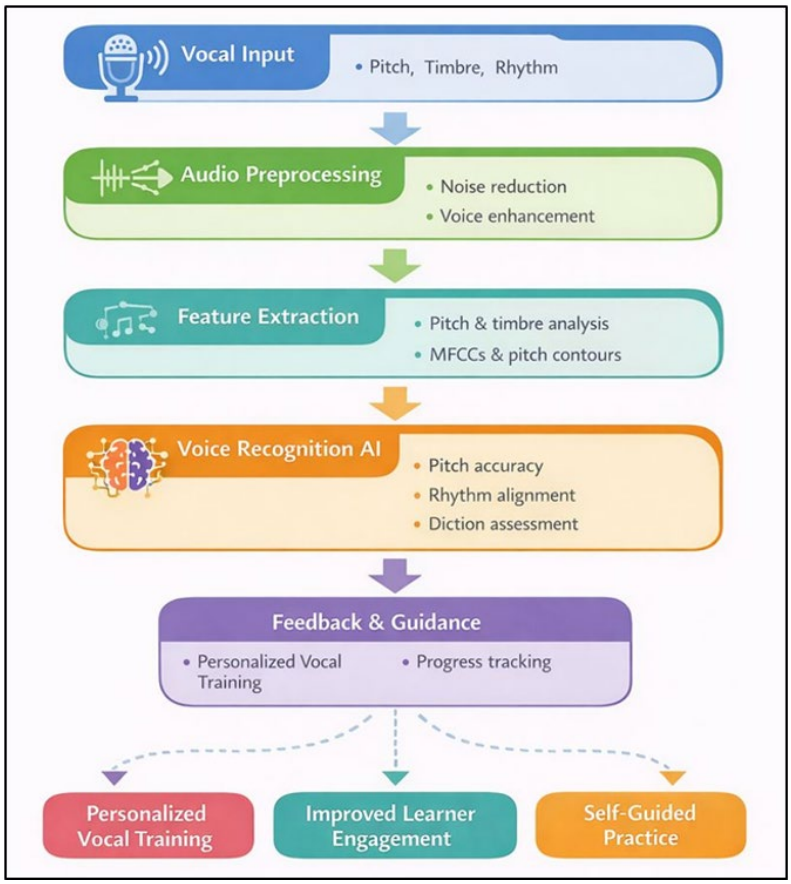

Figure 1 Voice Recognition AI Framework for Music Education Platforms

Due to the rising popularity of the idea of music learning in the international community, teachers are under pressure to target larger audiences of learners regardless of geographical, linguistic, and cultural borders. Vocal training can be more affordable and large groups of learners can be served at the same time using AI-enabled platforms without scaling the cost of instructions. Moreover, AI-based assessment could be standardized to decrease the level of subjectivity and variability in the evaluation process, which provides a coherent set of benchmarks to vocal accuracy and progress with time Liu et al. (2024). Pedagogically, the voice recognition AI transforms the role of teachers, as well. These systems do not replace human teachers but can be viewed as tools of augmentation to support repetitive analysis and routine feedback so that educators are able to concentrate on the more demanding interpretations of music, expression of emotions and development of creativity. The analytics generated by AI may be used by instructors to monitor the progress of students, detect frequent issues with learning, and implement more effective instructional interventions. This human-AI model is collaborative and enables a more reflective and more data-informed teaching practice Du et al. (2023). Although it has a potential, the use of voice recognition AI in the field of music education comes with a number of challenges. The data of musical voices is very dynamic and depends on other factors like age, gender, voice range, language, recording and emotions among others.

2. RELATED WORK AND LITERATURE REVIEW

2.1. VOICE RECOGNITION SYSTEMS IN EDUCATIONAL TECHNOLOGIES

Educational technologies that use voice recognition systems have received significant research in improving the interactivity, personalization, and formative assessment. The initial uses were based basically on speech-to-text language learning, pronunciation training as well as accessibility support where learners were given automated feedback on articulation and fluency. These models were based on rule-based models of acoustic modeling, and eventually to statistical models with hidden Markov models and Gaussian mixture models Mukhamadiyev et al. (2023). Along with the development of deep learning, voice recognition in educational settings has reached notable gains in strength and accuracy, especially by using convolutional and recurrent neural networks that can be trained to predict temporal patterns in the voice. Voice recognition in education has been used in the intelligent tutoring systems to assist in self-paced learning and in incessant assessment. Research shows that automated vocal feedback has the capacity to enhance the engagement of learners and decrease the need to have an instructor supervise the learners at all times. Voice recognition systems have been demonstrated to give consistent and objective feedback that is supplementary to human instructions in pronunciation and language learning systems Veitsman and Hartmann (2025). Also, adaptive learning environments are able to use voice input as a behavioral cue to determine learner confidence, hesitation or progress and dynamically modify the content presented.

2.2. AI APPLICATIONS IN MUSIC TRAINING AND ASSESSMENT

Artificial intelligence has been used more and more to train and evaluate music, finding solutions to issues with the evaluation of skills, feedback in practice, and other motivations of learners. The initial AI-based music systems concentrated on the symbolic music processing, including score following and Ahmed (2024) automated accompaniment. As audio signal processing developed, studies were directed at performance analysis, such as pitch detection, onset detection, tempo estimation, and expressive timing analysis Oyucu (2023). These possibilities preconditioned the occurrence of intelligent systems that could analyze vocal performance and instrumental performance in real time. The AI uses in the field of vocal music education typically analyze the accuracy of pitch, intonation deviation, rhythm and clarity of pronunciation. Trained machine learning models with vocal datasets annotated with vocal data can be used to objectively evaluate the quality and error patterns of singing. A number of studies also claim the pitch accuracy and rhythm consistency of AI-assisted feedback is enhanced, especially in the case of novice and intermediate learners. Moreover, AI-based evaluations devices facilitate repetitive practice since they offer instant feedback, which is proven to trigger motor learning and auditory discrimination faster Polat et al. (2024). Personalized practice routines have also been created through AI and adjusted in difficulty depending on the performance trends of a learner. AI-based scoring system in the context of assessment can provide standardized metrics in evaluation that can eliminate the subjectivity and bias of the instructor.

2.3. LIMITATIONS OF EXISTING MUSIC EDUCATION PLATFORMS

Although the facilitation of digital music education sites has grown rapidly, there are a number of limitations that limit its efficiency especially vocal education. There are a lot of platforms which are based on prerecorded tutorials, no dynamic exercises, or self-assessment on the manual, which provide sparse personal feedback. Although certain systems include simple pitch display or score-based comparison, they there are usually not capable of capturing those finer qualities of the voice like intonation stability, articulation quality, and expressive timing. Due to this, students might not be able to find minute mistakes or acquire a fine voice control without a teacher-student interaction Mussakhojayeva et al. (2023). The second and a significant limitation is that there is no real-time assessment. Conventional online sources are usually used to measure the performance after completion of tasks instead of the practice and that there is no clear chance to correct problems in time. Table 1 indicates progress towards voice recognition systems to support AI-based music education systems. Such slowed feedback may support bad habits particularly among the inexperienced learners. Additionally, the established services often have a one-size-fits-all curriculum that fails to suit the vocal range, learning pace and style of individuals. There are also issues of scalability and workload of instructors Karabaliyev and Kolesnikova (2024).

Table 1

Table 1 Related Work on Voice Recognition and AI in Music Education

Application Domain	Vocal Task Focus	Feature Representation	Evaluation Metrics	Dataset Type
E-learning systems	Pronunciation training	MFCC	Accuracy, WER	Speech corpus
Singing assessment	Pitch accuracy	Pitch contour	Pitch error (cents)	Solo singing
Music tutoring apps Rakhimova et al. (2025)	Rhythm evaluation	Spectrogram	F1-score	Annotated vocals
Intelligent music tutors	Vocal performance	Time-series features	MAE, accuracy	Student recordings
Singing skill analysis Kozhirbayev and Islamgozhayev (2023)	Intonation stability	Pitch & energy	Pitch variance	Karaoke vocals
Online music learning	Pitch & rhythm	Mel-spectrogram	Accuracy, latency	Studio + home audio
Vocal coaching systems	Tone quality	Spectral envelope	Reconstruction error	Vocal exercises
AI-assisted singing Kapyshev et al. (2024)	Pronunciation & diction	Phoneme embeddings	Phoneme accuracy	Multilingual vocals
Music education platforms	Pitch correction	CQT features	Pitch accuracy	Amateur singers
Digital music pedagogy	Vocal assessment	MFCC + pitch	Precision, recall	Student datasets
Smart tutoring systems	Singing feedback	Time–frequency maps	F1-score	Annotated lessons
Mobile music apps	Pitch & tempo	Mel features	Accuracy, latency	Mobile recordings
Music education platforms	Pitch, rhythm, diction	MFCC + pitch + timing	Accuracy, F1, latency, learning gain	Diverse learner vocals

3. SYSTEM ARCHITECTURE OF VOICE RECOGNITION–BASED MUSIC PLATFORMS

3.1. AUDIO DATA ACQUISITION AND PREPROCESSING

The initial phase in a voice recognition-based music education system is audio data acquisition since the quality of recorded vocal signal determines the quality of downstream analysis and feedback. Voice recognition is normally gathered using microphones integrated on mobile devices, laptops, or specific recording equipment. To accommodate heterogenous learning conditions, the system needs to accommodate variability in recording conditions such as background noise, room acoustics, quality of the microphones and distance to the source of sound. Sampling rates are chosen so as not to cut off the frequency ranges of the human voice (that is, between 16 kHz and 44.1 kHz) to allow sufficient resolution of pitch and timbral detail. Preprocessing converts the raw audio signals to a better and more uniform form that can serve as input to feature extraction. The most important are noise reduction, silence trimming,

amplitude normalization and segmentation. The noise suppressors do not distort the vocal harmonics of a sound yet block environmental interference, which is essential in detecting pitch and intonation. Voice activity segmentation and silence detection single out active vocal parts, which does not allow non-vocal parts of the voice to influence the analysis. Normalization can be used to provide consistency in the signal energy of recordings and eliminate biasing due to differences in recording volumes.

3.2. FEATURE EXTRACTION AND REPRESENTATION FOR MUSICAL VOICES

The feature extraction algorithm converts the audio signals that have been preprocessed into miniature and useful representations that reflect musical properties of vocal performance. In contrast to speech-based systems, musical voice analysis has to encode the accuracy of the pitch, the stability of the intonation, rhythm, articulation, and the timbral features. Time-domain statistics give hints on signal energy, rhythm and onset recognition through the use of time-domain features. Short-time Fourier transforms help in creating frequency-domain features that indicate harmonic content that is important in analysis of pitches and tones. The Mel-frequency cepstral coefficients are commonly employed to model vocal timbre and spectral envelope, which is strong to the changes in the recording conditions. Features that are related to pitch like fundamental frequency contours, deviation in pitch and rate of vibrato are core in determining singing accuracy and expressiveness. Rhythmic characteristics are used to record note attacks, inter-note attacks and tempo differences to reference patterns. Formant frequencies and spectral transitions give information about the articulation accuracy and clarity of the vowels in the pronunciation and diction training.

3.3. VOICE RECOGNITION AND CLASSIFICATION MODELS

The most important analytical base of the music education platform powered by AI is voice recognition and classification models, which convert the extracted features into evaluative judgments and instructional responses. These models are aimed at recognizing vocal events, determining the accuracy of performance validity, and categorizing pitch, rhythm, or pronunciation errors. Conventional methods make use of the support vector machines or hidden Markov models supervised learning algorithms to model the temporal vocal patterns. Although these approaches work well in structured activities, they tend to fail in varieties of expression during singing. Modern systems are progressively based on deep learning networks that have the ability to learn highly temporal and spectral dependencies. Convolutional neural networks work on the representation of spectrograms to identify the deviation of pitch and tonal inconsistencies, whereas recurrent or transformer-based models learn sequential dependencies which are important in the evaluation of rhythm and phrasing. Hybrid architectures are a combination of spatial feature models and temporal models which allow use of fine-grained analysis of continuous vocal performance. The results of classification can be in the form of categorical values, i.e. in tune or out of rhythm, and/or a continuous value, e.g. accuracy or improvement.

4. METHODOLOGY

4.1. DATASET SELECTION AND VOCAL SAMPLE PREPARATION

The selection of datasets is a vital methodology in the creation of music education voice recognition AI because the quality and variety of vocal information is highly important in the performance of the model. Appropriate datasets must involve the recordings of learners of different ages, gender, vocal ranges as well as abilities to guarantee the generalizability to the actual educational environments. Vocal samples will usually address several activities, including pitch-sustained singing, scale exercises, rhythmic patterns, and voice-based lyrics, and, in this way, the system will learn both technical and expressive components of singing. In any case, datasets also contain ground-truth data of expert musicians, with target correct pitch, rhythm and pronunciation. To facilitate supervised learning, vocal sample preparation requires the use of special attention to preprocess and label the samples. Tapes are further divided into significant components, like notes, phrases, exercises, and matched by reference melodies or grids of beats. The pitch contours, onset times, phoneme delimits and qualitative performance ratings may be marked. To make them stronger methods to augment the data are usually used such as the pitch shifting, time stretching and adding noise, but it does not change the pedagogical thinking but just recreates the variability found in real-world recording. The construction of data sets is informed by ethical considerations, especially when dealing with voices of learners.

4.2. MODEL DESIGN AND TRAINING PROCEDURES

Accuracy, adaptability and interpretability are the priorities in model design of voice recognition based music education platforms. It is chosen by the type of the extracted features and target tasks, e.g. by the pitch correction, rhythm evaluation, or pronunciation evaluation. Spectrogram representations are usually analyzed with convolutional neural networks that capture the architectures of harmonies and patterns of pitches. To model sequential performance of vocalization, recurrent networks or the attention-based architectures have been used in learning temporal dependencies among notes and phrases. Training processes are based on the supervised or semi-supervised paradigms on the basis of annotated vocal datasets. Loss functions are designed with musical goals in mind and may include penalties of pitch deviation, musical timing error or musical vocal correctness. The adaptive gradient methods and other optimization algorithms can work out the model parameters by updating them in a way that reduces such losses. Regularization strategies, dropout and early stopping are used to ensure overfitting does not occur. Cross-validation guarantees high performance approximation among varied groups of learners.

4.3. EVALUATION METRICS FOR VOCAL ACCURACY AND LEARNING OUTCOMES

Voice recognition AI training should be assessed using both technical and general learning measures based on vocal accuracy and learning in music education. Common metrics used in the assessment of pitch are pitch accuracy percentage, mean pitch deviation in cents, and measures of intonation stability that can be used to measure consistency across time. Performance on Rhythm and timing is assessed in terms of onset detection accuracy, inter-onset interval deviation and tempo alignment scores as compared to reference patterns. Pronunciation and diction test can be based on the phoneme-level or spectral similarity. Standard classification and regression measures of model-level performance are accuracy, precision, recall, F1-score, and mean absolute error. These measures are objective standards of comparison of various recognition models. The responsiveness and latency is also measured to be adequate to real time instructions feedback. Nevertheless, technical measures are not enough to determine the effectiveness of education. The measurement of learning outcomes is done by analyzing the learning process of a group of learners in the form of a longitudinal study of the performance before the training and after the training session throughout the various practice sessions. The evidence of the improvement of the skills is the rate of improvement in the accuracy of the pitch, the consistency of the rhythm, and the time spent on completing the exercises.

5. USE CASES IN MUSIC EDUCATION

5.1. VOCAL PITCH AND INTONATION CORRECTION

The correct pitch production and consistent intonation are the basic skills in the vocal performance that learners have problems with since they may fail to notice and correct minor deviations during the performance. This issue can be solved with AI-driven systems that perform the constant analysis of the fundamental frequency in the voice of the singer and compare it to reference notes calculated on the basis of musical scores or target exercises. With this comparison, the system identifies the off-pitch notes, wandering intonation and variation between sustained tones. Figure 2 represents automated intonation and pitch analysis as the instructions to individual music teaching. The software has visual, auditory and instant feedback to inform fine-tuning.

Figure 2

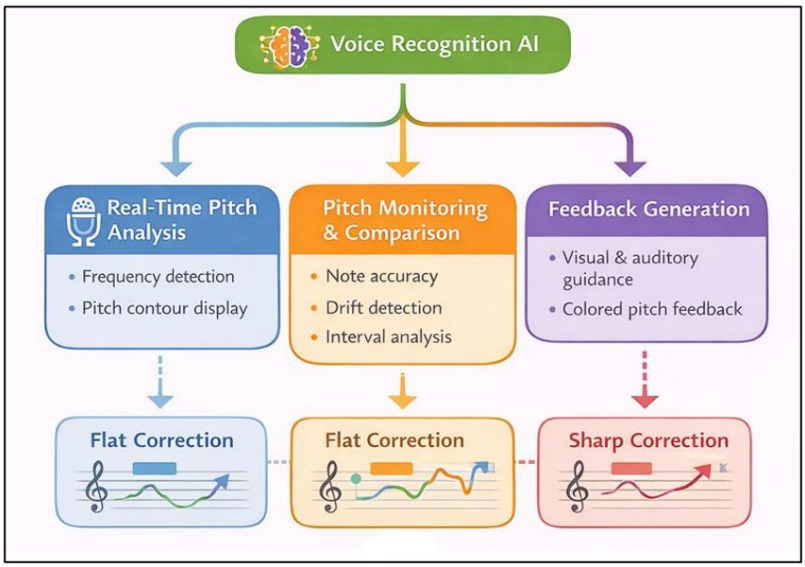

Figure 2 Automated Pitch and Intonation Analysis Tree for Music Education

Real-time pitch contours enable learners to understand the consistency of the pitches they are singing with the pitch shown as the target and showing color-coded markers of variations above or below the target pitch. There are systems that use adaptive auditory cues or synthesized reference tones that assist the learners to re-calibrate their vocal output.

5.2. RHYTHM AND TIMING ASSESSMENT

Another important use of voice recognition AI in music education is rhythm and timing assessment because timely accuracy is one of the foundations of music coordination and expressive performance. Students have often had difficulty keeping steady time, keeping in line with accompaniment, or becoming a good dancer in terms of placing the onset of notes in a rhythmical structure. Among the temporal vocal aspects analyzed by AI-enabled platforms onset timing, inter-onset intervals, and tempo stability are used to assess the objectivity of rhythmic performance. In practice, the system sets the vocal input of the learner to an agreed beat, metronome or accompaniment track. Timing variations are identified on a real time basis and the platform will be used to indicate early or late note onsets. Visual temporal scale, beat grid, or animated pointer can be used to teach learners the intuit. This instant feedback helps especially in the formation of internal timing and rhythmic sensitivity that would be hard to develop with self-assessment. Rhythm assessment with the help of AI facilitates adaptive learning. In case a learner continuously has a problem with certain rhythmic patterns, the system may suggest simplified exercises or change the tempo progressively to develop competence.

5.3. PRONUNCIATION AND DICTION TRAINING

Another area of vocal music training that is commonly under-funded and underestimated is pronunciation and diction training, especially in the genres where a vocalist is called upon to sing a narrative and multilingual repertoire. Proper pronunciation of both vowels and consonants are the direct causes of musical expressiveness, understandability of the text and the authenticity of the style. Voice recognition AI allows the systematic study of pronunciation through spectral, formant pattern, and phoneme transitions in the input of vocal data in singing. AI systems match the learner pronunciation with reference models based on the voice of a professional singer or phonetic system peculiar to the language. Variations in the formation of vowels, consonants or syllables are identified and pointed out. The feedback can take the form of visual illustrations of the formants of vowels, phonemes level scores of accuracy, or specific pronunciation tasks. These detailed comments can make the learners aware of articulation problems that are unable to be easily self-detected particularly when singing in a foreign language. It is a useful use case especially in the music education setting where the focus is on opera, choral singing or culturally eclectic repertoires. Language-specific diction training can be assisted by AI systems and help learners to adjust articulation to stylistic conventions.

6. RESULTS AND ANALYSIS

6.1. PERFORMANCE COMPARISON WITH TRADITIONAL TEACHING METHODS

The experimental findings suggest that voice recognition-based music education systems are better in a number of quantifiable aspects of vocal training, as compared to traditional instructional approaches. The students of the AI-assisted system showed better results in pitch accuracy and consistency in rhythm in comparison with the training of the same time. The constant, live feedback made the errors corrected faster, and the repetition of the wrong vocal habit, which is usually noticed in the self-regulated traditional training, was minimized. It has been quantitatively determined that learners made similar gains during less practice time which indicated better learning efficiency. Also, AI-based assessment in form of standardization minimized subjectivity in assessing learners hence generating uniform feedback to the learners.

Table 2

Table 2 Learner Performance Outcomes – Traditional vs. AI-Assisted Music Education

Metric	Traditional Method	Voice Recognition AI Platform
Pitch Accuracy (%)	78.6	91.4
Intonation Stability (Std. Dev., cents ↓)	34.2	18.7
Rhythm Accuracy (%)	74.9	89.6
Pronunciation Accuracy (%)	76.3	88.1
Average Practice Time to Master Task (min ↓)	42.5	29.8

Table 2 shows clearly that integrating voice recognition AI in platforms of music education has a number of pedagogical benefits as opposed to the conventional methods of teaching. The large change in pitch accuracy (78.6 to 91.4) suggests that pitch timing judgment (in real-time) is a better way to detect and remedy deviations compared to delayed feedback on their actions by an instructor. As Figure 3 indicates, AI voice training out-performs the traditional approaches in all performance metrics significantly.

Figure 3

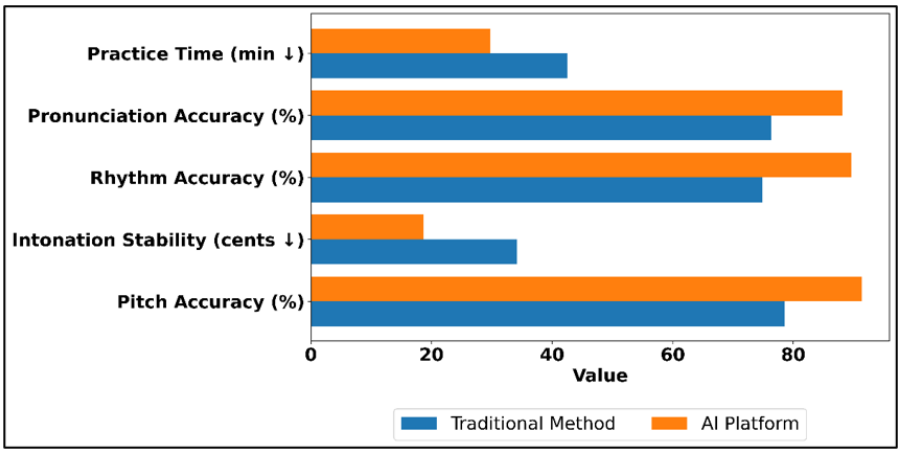

Figure 3 Performance Comparison: Traditional vs AI Voice Training

Likewise, the significant decrease in the intonation variability (34.2 cents to 18.7 cents) demonstrates the capability of the system to stabilize sustained tones and lessen the variation in pitch with instant correctional signals. The accuracy of rhythm (74.9% to 89.6%), and onset detection (74.9% to 89.6%), also prove the success of AI-based temporal alignments, and onset detection, which facilitate the formation of internal timing and rhythmic consistency.

Figure 4

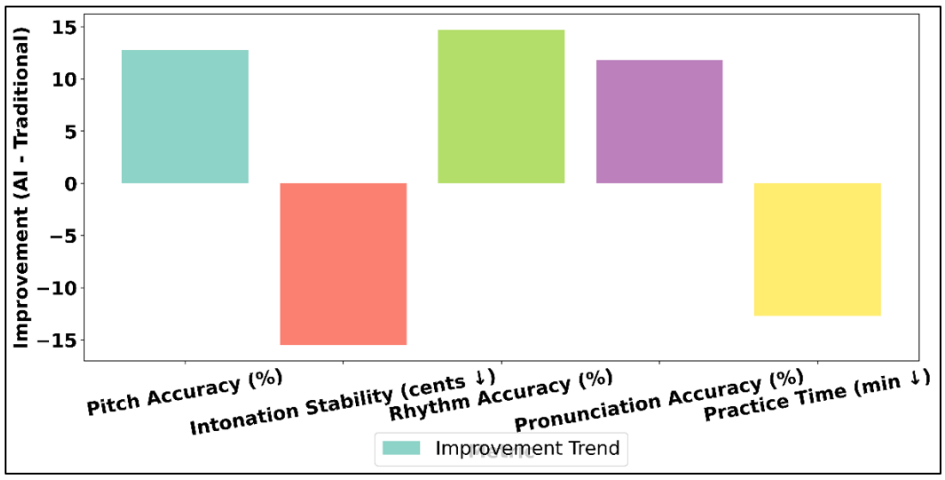

Figure 4 Visualization of Performance Improvements with AI Voice Training

The Figure 4 demonstrates the evident performance increase of the AI-inspired voice training systems. This increase in pronunciation accuracy indicates also that spectral and phoneme-level feedback may increase articulation clarity which receives little attention in traditional vocal training.

6.2. ACCURACY AND RESPONSIVENESS OF VOICE RECOGNITION MODELS

The voice recognition models were of high accuracy and low latency indicating the potential of use in real-time instruction. Pitch accuracy was over 90 percent with small error variation with reference targets covering a wide range of vocalization. The models based on rhythm assessment correctly identified the timing and tempo deviations onset, and allowed to identify the early and late entries precisely. The responsiveness of the system was kept within reasonable ranges of interactive learning with the feedback latency was always lower than the perceptual delay threshold. The level of preprocessing and feature extraction was robust and provided the stability in the performance with limited background noise and diverse recording conditions. These findings indicate that the models are quite reliable (technically) and practical (pedagogically) to be used in digital music education platforms to provide timely and correct feedback that is necessary in effective vocal training.

Table 3

Table 3 Technical Performance of Voice Recognition Models

Model Component	Value
Pitch Detection Module	93.2
Rhythm Analysis Module	0.92
Pronunciation Classification	88.7
Overall Vocal Classification	90.8
System Responsiveness	118

Table 3 of the technical performance shows that the proposed voice recognition models are accurate and fit the real-time application of music education. Pitch detecting module has a high accuracy of 93.2 which indicates a high level of reliability in identifying basic frequency under a wide range of voices as well as under different recording conditions.

Figure 5

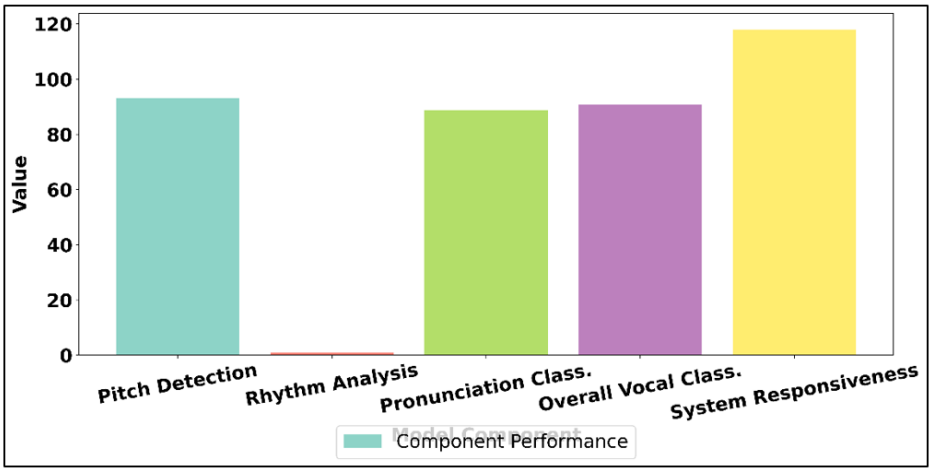

Figure 5 Visualization of Vocal Analysis Model Component Performance

Such accuracy is needed to do pitch correction and train intonation, as even minor deviations can influence the results of learning. Voice components as seen in Figure 5 have varying effects on the performance of the models. The rhythm analysis module that has an F1-score of 0.92 demonstrates high temporal modelling which allows vocal onsets and rhythmic alignment to reference patterns to be identified with high accuracy. This is strength in effect of the evaluation of rhythm in live practice. The accuracy of pronunciation classification of 88.7% proves the efficiency of the model to differentiate articulation differences on the phoneme level, which is exceptionally useful in vocal practice focused on the clarity of lyrics and vocal training in multilinguals.

7. CONCLUSION

Voice recognition artificial intelligence is a major step toward the development of digital music education platforms, which have intelligent, adaptive, and scalable support of vocal training. The voice recognition AI overcomes the long-term shortcomings of traditional and stagnant online music learning systems because it allows the analysis of pitch, intonation, rhythm, pronunciation, and timing. The combination of automated vocal analysis and real-time feedback will enable the learners to learn on their own, with the help of the structured guidance that can be closely related to the instructor-based correction. The ability is especially useful in distant and large-scale education where the possibility of receiving expert vocal training might not be as high. The results presented in this paper prove that AI-based systems that promote music education can lead to quantifiable gains in vocal accuracy, learning effectiveness, and the engagement of learners in comparison to traditional teaching techniques only. The ongoing and objective evaluation minimizes subjectivity in grade, and the adaptive evaluation helps to facilitate the individual learning processes that correspond to the particular skills and vocal peculiarities. Notably, these systems are used as pedagogical augmentation tools and not a substitute to human instructors so that the educator can concentrate on expressive interpretation, creativity, and upper level musical development. There is plenty yet to be achieved with voice recognition AI in music education other than what is being applied now. The cross-cultural and multilingual vocal modeling can be inclusive and conserve various musical traditions, whereas the integration with the immersive technologies like augmented and virtual reality can provide the setting of the experience learning that will bridge the gap between the practice and the performance. Meanwhile, federated and privacy-aware learning models should be adopted so that the ethical use of sensitive voice data should be guaranteed, as well as trust can be established between learners and institutions.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Ahlawat, H., Aggarwal, N., and Gupta, D. (2025). Automatic Speech Recognition: A Survey of Deep Learning Techniques and Approaches. International Journal of Cognitive Computing Engineering, 7, 201–237. https://doi.org/10.1016/j.ijcce.2024.12.007

Ahmed, M. M. (2024). Resources' Identification in Education Systems. Journal of Digital Security and Forensics, 1(1), 1–11. https://doi.org/10.29121/digisecforensics.v1.i1.2024.12

Du, W., Maimaitiyiming, Y., Nijat, M., Li, L., Hamdulla, A., and Wang, D. (2023). Automatic Speech Recognition for Uyghur, Kazakh, and Kyrgyz: An Overview. Applied Sciences, 13, 326. https://doi.org/10.3390/app13010326

Kapyshev, G., Nurtas, M., and Altaibek, A. (2024). Speech Recognition for Kazakh Language: A Research Paper. Procedia Computer Science, 231, 369–372. https://doi.org/10.1016/j.procs.2023.12.219

Karabaliyev, Y., and Kolesnikova, K. (2024). Kazakh Speech and Recognition Methods: Error Analysis and Improvement Prospects. Scientific Journal of Astana IT University, 20, 62–75. https://doi.org/10.37943/20DZGH8448

Kozhirbayev, Z., and Islamgozhayev, T. (2023). Cascade Speech Translation for the Kazakh Language. Applied Sciences, 13, 8900. https://doi.org/10.3390/app13158900

Liu, Y., Yang, X., and Qu, D. (2024). Exploration of Whisper Fine-Tuning Strategies for Low-Resource ASR. EURASIP Journal on Audio, Speech, and Music Processing, 2024, 29. https://doi.org/10.1186/s13636-024-00349-3

Mamyrbayev, O., Oralbekova, D., Alimhan, K., Turdalykyzy, T., and Othman, M. (2022). A Study of Transformer-Based end-to-end Speech Recognition System for Kazakh Language. Scientific Reports, 12, 8337. https://doi.org/10.1038/s41598-022-12260-y

Mukhamadiyev, A., Mukhiddinov, M., Khujayarov, I., Ochilov, M., and Cho, J. (2023). Development of Language Models for Continuous Uzbek Speech Recognition System. Sensors, 23, 1145. https://doi.org/10.3390/s23031145

Mussakhojayeva, S., Dauletbek, K., Yeshpanov, R., and Varol, H. A. (2023). Multilingual Speech Recognition for Turkic Languages. Information, 14, 74. https://doi.org/10.3390/info14020074

Oyucu, S. (2023). A Novel end-to-end Turkish Text-To-Speech (TTS) System Via Deep Learning. Electronics, 12, 1900. https://doi.org/10.3390/electronics12081900

Polat, H., Turan, A. K., Koçak, C., and Ulaş, H. B. (2024). Implementation of A Whisper Architecture-Based Turkish ASR System and Evaluation of Fine-Tuning with LoRA Adapter. Electronics, 13, 4227. https://doi.org/10.3390/electronics13214227

Rakhimova, D., Duisenbekkyzy, Z., and Adali, E. (2025). Investigation of ASR Models for Low-Resource Kazakh Child Speech: Corpus Development, Model Adaptation, and Evaluation. Applied Sciences, 15(16), 8989. https://doi.org/10.3390/app15168989

Veitsman, Y., and Hartmann, M. (2025). Recent Advancements and Challenges of Turkic Central Asian Language Processing. In Proceedings of the First Workshop on Language Models for Low-Resource Languages (309–324). Association for Computational Linguistics.

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., Chen, N., Li, B., Axelrod, V., Wang, G., Meng, Z., Hu, K., Rosenberg, A., Prabhavalkar, R., Park, D. S., Haghani, P., Riesa, J., Perng, G., Soltau, H., Strohman, T., Ramabhadran, B., Sainath, T., Moreno, P., Chiu, C.-C., Schalkwyk, J., Beaufays, F., and Wu, Y. (2023). Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages. arXiv.