Digital technology empowers innovation in vocal teaching research on performance skill evaluation and personalized teaching path optimization based on multimodal data

Jiakai Cao
The Philippine Women's University, Manila 1004, Philippines
Corresponding author: Jiakai Cao (e-mail: caojiakai12345@163.com)

ABSTRACT Against the dual backdrop of digital transformation in education and high-quality development of art education, traditional vocal teaching has long faced bottlenecks such as strong subjectivity, vague standards, lagging feedback, difficulty in quantification, and insufficient personalization, making it difficult to meet the refined needs of cultivating high-level artistic talents in the new era. Digital technology, multimodal perception, and big data analysis provide a new path for the scientific transformation of vocal teaching The inclusion of electromyographic and respiratory data makes the approach suitable for wearable bioelectrical sensing in vocal-training assessment.

INDEX TERMS vocal teaching,multimodal data,skill evaluation,personalized teaching,bioelectrical sensing

I. INTRODUCTION

A. RESEARCH BACKGROUND

WITH the deepening implementation of China's education digitization strategy, precise teaching and personalized learning have become the core directions for higher education reform and the algorithmic control of smart weaving and high-performance fiber production. Vocal music, as a performing art highly dependent on auditory perception, physiological control, and emotional expression, has strong experiential, implicit, subjective, and individualized characteristics in its teaching process. Traditional vocal teaching mainly relies on one-on-one small lessons, relying on the teacher's auditory experience, physical demonstration, and oral description, while students rely on imitation, perception, and repeated practice to form singing skills. This model has irreplaceable value in artistic inheritance and aesthetic cultivation, but its inherent limitations are becoming increasingly prominent in the context of expanding the scale of modern talent cultivation, improving teaching efficiency requirements, and increasingly standardized skill standards[1-3].

Firstly, teaching feedback mainly relies on qualitative descriptions, making it difficult to achieve precise quantification. Expressions such as' too shallow voice ',' insufficient breath ',' high throat position ', and' incorrect position 'are ambiguous, making it difficult for students to translate language descriptions into clear physiological movements, resulting in low learning efficiency[4].

Secondly, the learning process lacks data recording and dynamic tracking, resulting in unclear growth trajectories. Traditional teaching is difficult to preserve long-term, continuous, and multidimensional learning data, and cannot scientifically analyze progress, weak links, and stability. Teaching decisions lack basis[5-6].

B. RESEARCH SIGNIFICANCE

1) Theoretical significance

Promote vocal evaluation from subjective qualitative to scientific quantitative Through multidimensional data fusion, the computable feature expressions of core elements such as pitch accuracy, rhythm, breath, resonance, enunciation, and emotion are clarified. A quantitative evaluation model for vocal techniques with high reliability and efficiency is established to promote the development of vocal evaluation theory towards objectivity, standardization, and refinement.

Improve the theoretical foundation of personalized art teaching Based on cognitive load theory, constructivist learning theory, mastery learning theory, and adaptive learning theory, this paper proposes a hierarchical progressive, targeted intervention, and dynamically adapted teaching path theory for the formation of vocal skills, providing a new theoretical perspective for the cultivation of personalized artistic talents.

2) Practical significance

Improve vocal teaching efficiency and training accuracy Through real-time data feedback, students can intuitively perceive their own problems, quickly establish the connection between "sound physiology data", shorten the blind practice cycle, and improve the efficiency of skill formation.

Reduce subjective evaluation bias of teachers and improve evaluation consistency Multimodal quantification results can provide objective references for teachers, reduce evaluation fluctuations caused by experience differences, and improve the fairness and stability of teaching evaluation. Realize personalized cultivation in large-scale teaching In the context of the expansion of vocal enrollment in universities and the increasing popularity of collective and group courses, using intelligent systems to achieve large-scale, precise, and sustainable personalized guidance can alleviate teacher pressure.

Provide engineering basis for the development of intelligent vocal teaching system The data collection plan, feature index system, evaluation model, and teaching strategy library formed by the research can provide core algorithms and functional design basis for intelligent teaching platforms, portable training devices, and online vocal learning systems.

C. CURRENT RESEARCH STATUS AT HOME AND ABROAD

In recent years, digital research on vocal music in China has developed rapidly, mainly focusing on three directions:

One is vocal acoustic analysis, exploring the relationship between acoustic characteristics such as resonance peaks, pitch accuracy, and breath and singing level;

The second is the application of digital teaching, exploring the auxiliary role of online teaching, virtual teaching, and multimedia resources in vocal teaching;

The third is artificial intelligence and vocal evaluation, attempting to construct an audio based singing rating model[7].

However, there are still significant shortcomings in existing research:

Mostly single modal research, lacking multidimensional data fusion. Most only use audio without integrating physiological, posture, airflow, video and other information, making it difficult to fully reflect singing skills. Emphasis is placed on technical implementation, lacking deep embedding of teaching methods. Technical research is disconnected from actual teaching objectives, training steps, and error correction logic, making it difficult to implement. Personalized teaching remains at the conceptual level, without a path-oriented and operable system. Lack of a closed-loop design of "diagnosis intervention training reassessment".

Insufficient empirical research, lacking large sample, long-term, and strictly controlled teaching experiments. Overall, systematic research with multimodal data as the core, teaching innovation as the goal, and personalized paths as the outcome is still scarce, which is the breakthrough space of this study.

D. RESEARCH IDEAS, CONTENT, AND METHODS

1) Research Approach

Guided by the pain points of vocal teaching, supported by multimodal data technology, with precise evaluation of singing skills as the core and personalized teaching path optimization as the goal, according to: Theoretical construction → System design → Indicator construction → Model establishment → Path generation → Empirical verification → System improvement Conduct research on the technical roadmap.

2) Research Content

The dilemma and transformation logic of vocal teaching in the digital age.

Theoretical basis and technical framework for applying multimodal data to vocal teaching.

Multimodal data collection scheme, feature index system, and quantitative evaluation model for singing skills. Personalized teaching path generation mechanism and strategy library based on defect diagnosis. Design and Implementation of a Digital Vocal Teaching Closed Loop System.

Teaching experiment verification and effect analysis.

Promotion and guarantee mechanism of digital vocal teaching and future prospects.

3) Research Methods

Literature research method: Sort out literature on vocal teaching methods, multimodal perception, AI education, acoustic analysis, etc.

Expert interview method: Interview vocal educators, sound scientists, and technical researchers.

Multimodal data acquisition method: Construct a synchronous acquisition system to obtain audio, physiological, posture, airflow, and other data.

Machine learning modeling method: Construct singing skill evaluation, defect diagnosis, and level grading models.

Quasi experimental research method: Set up an experimental group (multimodal personalized teaching) and a control group (traditional teaching) to compare long-term teaching.

Statistical analysis method: SPSS was used for significance testing, correlation analysis, and effect size calculation.

II. CORE CONCEPTS AND THEORETICAL FOUNDATIONS

A. DEFINITION OF CORE CONCEPTS

1) Digital Technology

The digital technology in this study refers to a series of technologies that can be applied to vocal teaching, including multimodal perception, signal processing, computer vision, physiological sensing, machine learning, big data analysis,

real-time feedback, and adaptive teaching. Its function is to transform invisible physiological movements, unmeasurable sound features, and unquantifiable artistic expressions into data, achieving objectivity, precision, and personalization in the teaching process[8-10].

2) Multimodal Data

Multimodal data refers to multidimensional data collected synchronously and reflecting the singing state:

Audio modes: pitch, loudness, spectrum, resonance peak, jitter, vibrato, etc;

Physiological modes: respiratory electromyography, chest and abdominal movements, laryngeal acceleration, airflow pressure, etc;

Visual modalities: head posture, facial muscles, mouth shape, body posture, etc;

Temporal modes: rhythm, duration, onset speed, coherence of musical phrases, etc.

The core value of multimodal fusion lies in complementarity: audio reflects results, physiological mechanisms, posture reflects habits, and achieves a complete explanation from "sound expression" to "physiological mechanisms"[11-12].

3) Evaluation of Singing Techniques

Performance skill evaluation is a quantitative and qualitative assessment of dimensions such as pitch accuracy, rhythm, breath, resonance, throat stability, enunciation, vibrato, coherence, and emotional expression. Traditionally, subjectivity is the main focus, but this study achieves a scientific evaluation that integrates subjectivity and objectivity, supported by objective data[13].

4) Personalized Teaching Path

Personalized teaching path refers to the dynamic generation of exclusive training plans, target sequences, feedback methods, error correction strategies, and difficulty gradients based on students' basic level, physiological conditions, weak links, cognitive styles, and learning progress. This approach facilitates a highly granular and adaptive pedagogical framework where training interventions are dynamically recalibrated to match specific learner deficiencies and longitudinal progress metrics[14].

B. THEORETICAL BASIS

1) Constructivist Learning Theory

Singing skills are not passively accepted, but actively constructed by students through feedback, adjustment, repetition, and correction. Multimodal data provides students with external references, enabling them to independently establish connections between sound and physiology, accelerating skill construction[15].

2) Adaptive Learning Theory

Adaptive learning centers around learner models, domain models, and instructional models, dynamically adjusting content and strategies based on real-time conditions. This study constructs a personalized teaching path generation engine based on this[16].

3) Vocal Art Performance Theory

Vocal music is the unity of science and art. Digitization does not replace artistic expression, but liberates artistic expression: when physiological skills are stabilized and automated, students can focus more on emotional, aesthetic, and style creation.

III. THE REALISTIC DILEMMA OF TRADITIONAL VOCAL TEACHING AND THE LOGIC OF DIGITAL TRANSFORMATION

A. THE CORE REALITY DILEMMA OF TRADITIONAL VOCAL TEACHING

1) Feedback information is highly qualitative and vague, making it difficult to translate into specific physiological actions

Vocal music is a highly refined performing art of physiological control: breathing relies on the coordinated cooperation of the diaphragm, intercostal muscles, and abdominal muscles; Vocalization depends on the degree of closure and vibration pattern of the vocal cords; Resonance relies on the dynamic regulation of the pharyngeal cavity, oral cavity, nasal cavity, and other cavities. These internal movements cannot be directly observed and can only be inferred through external sounds[17].

In traditional teaching, teachers commonly use qualitative, experiential, and metaphorical language for feedback, such as:

The breath needs to sink down

The voice should stand up and hang up

Keep your throat steady and avoid getting stuck

The resonance position is incorrect, too shallow/too dark

The voice is not smooth enough and lacks support

This type of expression has value in terms of artistic inspiration, but there are obvious limitations in terms of skill correction:

Firstly, there is a lack of clear correspondence between language description and physiological movements, making it difficult for students to directly locate muscle groups, control strength and movement timing through language;

Secondly, the level of abstraction is too high, and the same vocabulary forms completely different understandings in the minds of different students;

Thirdly, there is a lack of quantitative scale in feedback, and students do not know "to what extent it is considered qualified" or "to what extent it is stable is considered standard";

Fourthly, it is difficult to form reproducible and verifiable training objectives, and the learning process highly relies on "comprehension" and "trial and error"[18].

The result is that a large number of students are in a state of vague practice, inefficient repetition, and blind adjustment for a long time, and the skill formation cycle is

greatly prolonged, making it difficult to improve teaching efficiency[19-20].

2) Physiological regulation processes are invisible and unmeasurable, resulting in low teaching error correction efficiency

The essence of vocal skills is a stable, automated, and coordinated physiological acoustic system. There is a strict temporal and intensity matching relationship between breath, vocal cords, resonance, and posture. For example, insufficient breath support can lead to pitch drift; Excessive tension of the extralaryngeal muscles can cause compression and jamming; Body stiffness can disrupt respiratory coordination.

In traditional teaching, these internal physiological processes are all in a black box state:

Teachers are unable to directly observe students' diaphragm movement status;

Unable to accurately determine the amplitude of fluctuations in throat position;

Unable to quantify the ratio of chest and abdominal expansion and ventilation rate;

Unable to distinguish between "insufficient breath" and "imbalanced breath confrontation";

Unable to objectively distinguish between "tone preference" and "vocal error".

Teachers can only infer internal mechanisms through reverse inference of the final sound, which involves a lot of uncertainty: the sound may be weak, possibly due to issues with breath, vocal cords, resonance, or posture. In traditional teaching, teachers often need to make multiple attempts and adjustments to identify the true cause, which relies heavily on experience and incurs high trial and error costs for students.

3) Homogenization of teaching modes, difficult to adapt to individual differences among students

Vocal learners have natural differences in physiological structure, vocal conditions, vocal range, basic level, learning habits, cognitive style, and muscle control ability. Some people have good breath but weak pitch, some have stable pitch but poor resonance, some are prone to nervousness and stiffness, and some are too loose.

However, in practical teaching, whether it is vocal classes in universities, collective courses in teacher training majors, or courses in primary and secondary schools and training institutions, a homogeneous model of unified textbooks, unified repertoire, unified training steps, and unified progress is commonly adopted. It is difficult for teachers to provide long-term, detailed, and personalized diagnosis and planning for each student within limited class hours.

The result is:

Students with a good foundation may not have enough to eat;

Students with weak foundations cannot keep up;

Students with specific defects cannot receive targeted training;

The advantageous conditions cannot be strengthened and developed.

The contradiction between unified teaching and individualized needs has become a key structural contradiction that restricts the quality of vocal talent cultivation.

4) Lack of data retention in the learning process, making it impossible to achieve dynamic tracking and scientific evaluation

Traditional vocal teaching almost lacks the ability to record data, track long-term performance, analyze growth curves, and analyze defect evolution. Teachers' judgments of students mainly rely on classroom memory and subjective impressions; Students' perception of their own progress is often vague.

This model brings three limitations:

Firstly, it is difficult to objectively measure the extent of progress and distinguish between "true progress" and "good state";

Secondly, it is difficult to track whether the defects have truly stabilized and improved, which can lead to the repeated phenomenon of "this class is right, the next class goes back";

Thirdly, it cannot provide a basis for teaching adjustments, and teachers find it difficult to determine whether it is a problem with training methods, training volume, or physiological habits.

In the context of high-quality education development, measurability, traceability, comparability, and verifiability have become the basic requirements for talent cultivation, while traditional vocal teaching lags behind in terms of data, process, and visualization.

B. THE INTERNAL TRANSFORMATION LOGIC OF DIGITAL TECHNOLOGY EMPOWERING VOCAL TEACHING

Digital technology is not meant to replace teachers, artistic insights, or oral instruction, but to solve the "black box problems, fuzzy problems, deviation problems, and efficiency problems" that traditional teaching cannot solve through objective, visual, quantitative, and personalized means, achieving a new teaching paradigm that unifies science and art, integrates technology and humanities, and balances precision and individuality.

1) Objectification Logic: Using Multimodal Data to Reduce Subjective Evaluation Bias

Traditional teaching relies on "person person" feedback, while digitalization introduces third-party references of "person data person". Multimodal data such as audio, physiology, posture, and airflow can provide stable, consistent, and reproducible measurements of pitch, rhythm, breath, throat position, resonance, tension, etc., providing objective basis for teachers and stable standards for students, thereby significantly improving evaluation reliability.

2) Visualization Logic: Transforming Implicit Skills into Visual Information

Multimodal data can transform invisible physiological movements, sound structures, and respiratory patterns into curves, graphs, numerical values, and dynamic images, allowing students to intuitively see:

The fluctuation amplitude of one's own throat position;

Is the breath curve stable;

The distribution of pitch deviation on the time axis;

Is the body posture tense and stiff;

Is the ventilation point reasonable.

Visualization significantly reduces cognitive load, enabling abstract skills to quickly translate into concrete actions.

3) Precision Logic: Achieving Targeted Defect Diagnosis and Precise Intervention

Multimodal data can achieve precise positioning, attribution, and training:

Is pitch drift a problem with breath, vocal cords, or hearing?

Is the sound jamming caused by tension in the extralaryngeal muscles, insufficient respiratory support, or resonance regulation errors?

Is the lack of coherence in musical phrases due to unreasonable ventilation, unstable rhythm, or weak control? Traditional teaching requires a lot of trial and error, while digitalization allows for one-time collection, multidimensional analysis, direct localization, and targeted solutions, significantly improving training efficiency.

4) Personalized Logic: Implementing One Person One Path Based on Learner Model

By using multimodal data to construct student ability profiles, defect maps, learning styles, and progress speed models, the system can automatically generate:

Personalized training objectives;

Personalized training content;

Personalized difficulty gradient;

Personalized feedback method;

Personalized tracks and practice suggestions.

Truly realizing the modern transformation of "teaching students according to their aptitude".

5) Closed loop Logic: Building a Teaching Loop of "Evaluation Diagnosis Training Re evaluation"

Traditional teaching is mostly an open-loop model of "classroom feedback - after class exercises - next time back to class", and the intermediate process is uncontrollable and unpredictable. Digital teaching forms a complete closed loop of data collection, quantitative evaluation, defect diagnosis, teaching intervention, training feedback, and effect comparison, achieving a continuous optimization, steady improvement, and stable solidification of the skill formation path.

IV. CONSTRUCTION OF SINGING SKILL EVALUATION SYSTEM BASED ON MULTIMODAL DATA

A. DESIGN OF MULTIMODAL DATA SYNCHRONIZATION ACQUISITION SYSTEM

1) Acquisition Modes and Equipment Selection

Audio modal acquisition Audio is the core modality of singing evaluation, used to extract acoustic features such as pitch, loudness, rhythm, spectrum, resonance peaks, jitter, Shimmer, noise spectrum ratio, vibrato frequency and amplitude.

Equipment: Directional condenser microphone, frequency response 20Hz-20kHz, signal-to-noise ratio $\geq$ 90dB;

Sampling rate: 48kHz, 16 bit quantization;

Layout: 30cm in front, with a height level with the mouth;

Noise reduction: using directional pickup and environmental noise suppression algorithm, suitable for piano room scenes.

Physiological modal acquisition

Physiological signals are used to reveal the control mechanisms behind sound, avoiding misjudgments based solely on sound.

Chest and abdominal respiratory sensor: records the expansion curves of the abdomen and chest, reflecting respiratory patterns, breath support, ventilation rate, and stability;

Throat acceleration sensor: pasted on the outer side of the thyroid cartilage, non-invasive collection of laryngeal displacement and fluctuation amplitude, to determine laryngeal stability;

Surface electromyography (EMG) sensor: collects the tension of the sternocleidomastoid and anterior cervical muscle groups to determine whether there is compression or compensatory tension.

Attitude modal acquisition Posture directly affects respiratory support and body coordination, and has long been overlooked but extremely critical in vocal teaching.

Monocular RGB camera, 1080P/30fps;

Implementing human keypoint detection based on MediaPipe, extracting:

Shoulder inclination;

Spinal curvature;

Head forward tilt angle;

Mouth opening and closing degree and timing;

Facial tension level.

Airflow modal acquisition The airflow signal directly reflects exhalation control, breath pressure, and coherence of musical phrases, and is a key objective indicator for determining breath support.

Miniature airflow sensor, placed 5cm in front of the mouth;

Record: Peak expiratory flow rate, expiratory duration, inspiratory duration, ventilation ratio, and airflow stability.

B. MULTIMODAL TIME SYNCHRONIZATION SCHEME

The core value of multimodal data lies in temporal alignment: sound changes, throat movements, breathing curves, and posture adjustments must be presented on the same

TABLE 1. Multimodal Data Acquisition Modalities, Equipment and Technical Parameters in Vocal Teaching

Modality	Sensor/Device	Sampling Rate	Key Measurement Index	Application Purpose
Audio	Condenser microphone	48 kHz, 16-bit	Pitch, rhythm, formant, jitter, shimmer, spectral centroid	Intonation accuracy, timbre, vocal stability
Respiration	Abdomen-thorax strain sensor	1 kHz	Expansion ratio, breath stability, inhalation duration	Breath support pattern, air supply control
Laryngeal Position	3-axis accelerometer	1 kHz	Displacement amplitude, fluctuation variance	Laryngeal stability, excessive elevation
Muscle Tension	Surface EMG sensor	1 kHz	Amplitude, activation duration	Cervical muscle tension, vocal squeeze
Posture	RGB camera + MediaPipe	30 fps	Shoulder symmetry, spinal curvature, head tilt	Physical relaxation, body support
Airflow	Airflow transducer	1 kHz	Peak flow, expiration smoothness	Phrase coherence, breath control

timeline to achieve accurate attribution. This system adopts a combination of hardware timestamp and software synchronization:

All sensors are based on a unified clock, and the sampling points carry timestamps; Audio and physiological signals are triggered by hard synchronization, with a synchronization error of $\leq$ 10ms; The posture video is aligned through frame timestamps utilizing a linear interpolation algorithm to map 30 fps visual coordinates onto the 1 kHz physiological timeline. To mitigate jitter and potential frame loss, the system employs a global NTP (Network Time Protocol) clock for initial anchoring, followed by a zero-order hold (ZOH) compensation method to ensure temporal alignment across heterogeneous sampling rates.

After synchronization is completed, a four-dimensional fusion temporal data segment is formed, which can provide a comprehensive interpretation of any singing segment in terms of "sound results, physiological mechanisms, posture habits, and airflow control".

C. CONSTRUCTION OF MULTIMODAL CHARACTERISTIC INDEX SYSTEM FOR SINGING SKILLS

Based on vocal teaching methods, acoustic principles, vocal medicine, and sports science, this study constructs a singing technique evaluation index system that includes 4 primary indicators, 12 secondary indicators, and 36 tertiary quantitative features, all of which are computable, extractable, and comparable.

1) Pitch Accuracy and Rhythm Characteristics (audio modality)

Pitch deviation value (Cent): Calculate real-time deviation based on standard pitch;

Pitch stability: the amplitude of pitch fluctuations within a musical phrase;

Rhythm error: the difference between onset time and standard rhythm;

Accuracy of timing: Differences in note length compared to standard templates;

Starting speed: defined as the Vocal Attack Time (VAT), calculated as the duration from the initiation of the physiological respiratory pulse to the stabilization of the fundamental frequency (f_0). Mathematically, it is expressed as the time $\Delta t = t_{\text{steady}} - t_{\text{onset}}$, where t_{onset} is the point where the amplitude envelope exceeds 10% of the peak energy. This determines the explosive power and precision of vocal onset.

2) Characteristics of Breath and Airflow Control (Physiological+Airflow Modes)

Chest and abdominal expansion ratio: Determine whether it is chest breathing, abdominal breathing, or a combination of chest and abdominal breathing;

Breath stability: standard deviation of airflow curve during exhalation stage;

Ventilation rate: proportion of inhalation time;

Duration of breath support: Maximum stable gas supply duration for musical phrases;

Respiratory electromyography intensity: the level of activation of respiratory muscle groups.

3) Vocal and Laryngeal Stability Characteristics (audio+physiological modality)

Throat fluctuation amplitude: integrated displacement of acceleration signal;

Throat stability coefficient: variance of fluctuations within musical phrases;

Jitter/Shimmer: Vocal fold vibration stability;

Noise spectrum ratio: roundness of sound;

Neck muscle electromyography amplitude: determining the degree of compression.

4) Resonance, Articulation, and Posture Features (Audio+Posture Modes)

Resonance peak F1/F2 trajectory: vowel standard degree;

Spectral center of gravity: quantization of tone brightness and darkness;

Time sequence of mouth opening and closing: clarity of bite;

Head posture deviation: Reasonable body support;

Shoulder symmetry: degree of tension and stiffness This system achieves a complete mapping from subjective descriptors to quantitative features:

Unstable breath "$\rightarrow$ Low airflow stability and large fluctuations in chest and abdominal curves;

'Throat lifting' $\rightarrow$ Throat displacement amplitude exceeds the standard;

TABLE 2. Multimodal Quantitative Evaluation Index System for Singing Skills

First-level Index	Second-level Index	Quantitative Feature	Unit	Standard Range
Intonation & Rhythm	Pitch accuracy	Pitch deviation	Cent	$\leq$ 15 cent
Intonation & Rhythm	Rhythm consistency	Onset time error	ms	$\leq$ 50 ms
Intonation & Rhythm	Tone stability	Pitch fluctuation variance	—	$\leq$ 0.02
Breath Control	Breath smoothness	Airflow curve SD	—	$\leq$ 0.10
Breath Control	Breath ratio	Inspiratory/expiratory ratio	%	10%–20%
Breath Control	Thoracic-abdomen coordination	Expansion correlation coefficient	—	$\geq$ 0.75
Vocalization	Laryngeal stability	Displacement SD	mm	$\leq$ 0.50
Vocalization	Vocal fold stability	Jitter, Shimmer	%	$\leq$ 0.5%
Vocalization	Tone clarity	Noise-to-harmonic ratio	dB	$\geq$ 15 dB
Resonance & Posture	Resonance feature	Formant F1/F2 trajectory	—	Standard vowel range
Resonance & Posture	Articulation	Mouth opening amplitude	px	Normal range
Resonance & Posture	Body posture	Shoulder inclination angle	°	$\leq 3°$

'Sound squeezing' $\rightarrow$ high cervical muscle electromyography+high Jitter;

'Pitch drift' $\rightarrow$ pitch deviation Cent exceeds the standard+insufficient breath support;

Body stiffness $\rightarrow$ Shoulder tilt+abnormal spinal curvature.

D. PERFORMANCE SKILL EVALUATION MODEL BASED ON MULTIMODAL FUSION

1) Model Architecture

This study constructs a dual branch fusion evaluation model:

Acoustic branch: Extracting audio temporal features based on CNN+LSTM;

Physiology posture branch: Extracting physiological and posture statistical features based on MLP;

Attention Fusion Layer: Automatically assigns weights to different segments and highlights key phrases;

Output layer: comprehensive score, scores for each dimension, and defect probability.

The output results of the model include:

Comprehensive singing skill score (0-100);

Six dimensional radar chart of pitch accuracy, rhythm, breath, vocalization, resonance, and posture; Probability distribution of defect types;

Characteristic difference curve from standard sample.

2) Model Training and Validation

The dataset includes:

Professional singing samples (vocal teachers, professional actors);

Intermediate level sample (college students);

Beginner level sample (for beginners);

Typical error samples (insufficient breath, upward movement of the throat, stuttering, pitch deviation, etc.). Through expert scoring and model output fitting training, the model accuracy reached 92.3%, with a correlation coefficient of r=0.891 with subjective expert evaluation. It has high reliability and validity and can be used for teaching evaluation, process monitoring, and personalized diagnosis.

3) Automatic Diagnosis Engine for Singing Defects

The diagnostic engine adopts a dual-mode of rule-based reasoning and machine learning classification to automatically recognize 36 common singing defects:

Breath type: shallow breath, rapid ventilation, uneven exhalation, insufficient support, and imbalance in resistance;

Pronunciation category: overall low, overall high, tail drift, inaccurate beginning;

Vocal category: high throat position, compression, incomplete closure of vocal cords, weak voice, tight voice;

Rhythm categories: snatch, drag, inaccurate timing, and broken musical phrases;

Resonance type: excessive nasal sound, excessive laryngeal resonance, inability to open the mouth, and position biased forward/backward;

Posture category: stiff body, forward leaning head, shrugging shoulders, breathing disconnected from posture. The diagnostic output includes:

Defect name;

Severity of defects (mild/moderate/severe);

Time period and position of musical phrases;

Possible physiological reasons;

Recommend training direction.

This engine shifts vocal teaching from empirical judgment to data diagnosis, providing a direct basis for personalized teaching paths in the future.

E. SUBJECTIVE OBJECTIVE FUSION EVALUATION MECHANISM

To balance scientificity and artistry, this study establishes a subjective objective fusion evaluation mechanism:

Objective score (60%): multimodal model output;

Subjective score (40%): Teachers rate based on dimensions such as artistic expression, emotional processing, timbre aesthetics, and stage performance.

Ensuring the objectivity of skill assessment while retaining the teacher's leading role in artistic aesthetics, realizing the teaching evaluation concept of science as the foundation and art as the soul.

V. TEACHING EXPERIMENT AND EFFECTIVENESS VERIFICATION

TABLE 3. Typical Singing Defects and Corresponding Personalized Training Strategies

Defect Category	Typical Defect	Multimodal Characteristics	Targeted Training Strategy
Breath Defect	Shallow breathing, insufficient support	Low expansion ratio, high airflow SD	Slow inhalation, prolonged exhalation, visual breathing curve training
Breath Defect	Frequent inhalation, uneven airflow	Short expiration, high breath fluctuation	Phrase breathing design, smoothness training
Vocal Defect	Laryngeal elevation, squeezing	Large laryngeal displacement, high EMG	Laryngeal stabilization, light attack, neck relaxation
Vocal Defect	Hoarse voice, insufficient closure	High jitter, low NHR	Semi-occluded vocal tract exercises
Intonation Defect	Pitch drift, unstable tail	High pitch deviation, poor breath support	Pitch calibration, long tone holding, auditory feedback
Posture Defect	Stiff body, shoulder lifting	Asymmetric shoulders, forward head	Posture correction, relaxation training, skeleton visualization

TABLE 4. Comparison of Teaching Effects Between Experimental Group and Control Group

Index	Experimental Group (Mean ± SD)	Control Group (Mean ± SD)	t-value	p-value	Improvement
Intonation Accuracy (%)	92.57 ± 3.62	78.43 ± 6.15	10.82	<0.001	+14.14%
Breath Stability	0.86 ± 0.07	0.62 ± 0.09	11.36	<0.001	+38.71%
Laryngeal Stability (mm)	0.42 ± 0.11	0.78 ± 0.15	10.61	<0.001	-46.15%
Rhythm Error (ms)	38.64 ± 8.52	69.35 ± 11.27	11.94	<0.001	-44.28%
Teacher Overall Score	87.32 ± 4.26	76.51 ± 5.83	8.27	<0.001	+14.13%

A. EXPERIMENTAL DESIGN

Using quasi experimental research method, set up:

Experimental group: Personalized teaching based on multimodal data;

Control group: Traditional vocal teaching.

Control variables: teacher, textbook, class hours, and track consistency.

Cycle: 16 weeks.

B. EXPERIMENTAL SUBJECTS

60 students in vocal courses at universities were randomly divided into two groups of 30 each, with no significant difference in initial level between groups ($p>0.05$).

This study was approved by the Ethics Committee of The Philippine Women's University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

C. EVALUATION INDICATORS

Objective indicators: pitch accuracy, breath stability, throat position stability, rhythm error;

Subjective indicators: teacher professional rating, artistic performance rating;

Learning efficiency: number of skill mastery class hours, training speed to meet standards;

Learning satisfaction: learning confidence, training interest, and level of understanding.

D. EXPERIMENTAL RESULTS

Objective skill indicators The experimental group demonstrated superior performance across all primary metrics ($p<0.01$). Preliminary ablation analysis indicates that the integration of physiological and airflow modalities contributed to 24% of the diagnostic accuracy improvement, particularly in identifying latent respiratory bottlenecks that audio-only models failed to detect. While the current effect sizes are high, we acknowledge that the controlled laboratory setting may amplify performance; future iterations will evaluate model robustness in noisier, non-idealized acoustic environments.

Subjective evaluation by teachers The average score of the experimental group was 87.3, while that of the control group was 76.5, with a significant difference ($p<0.001$).

Learning efficiency The experimental group saved an average of 38% of class hours by mastering core skills.

Learning satisfaction 93% of the experimental group students believe that they have a clearer understanding of their own problems, while 53% of the control group.

VI. CONCLUSION

The evaluation of singing skills and personalized teaching paths based on multimodal data can significantly improve the effectiveness of vocal instruction.

VII. RESEARCH CONCLUSIONS, INNOVATIVE POINTS, AND PROSPECTS

A. RESEARCH CONCLUSION

Traditional vocal teaching faces five major challenges: subjective bias, vague feedback, physiological black box, insufficient personalization, and lack of data tracking.

Multimodal data can achieve objective, visual, and quantitative evaluation of singing skills, significantly improving evaluation reliability.

Based on multimodal diagnosis, personalized teaching path can achieve targeted training, precise intervention, dynamic iteration, and greatly improve efficiency.

The deep integration of digital technology and vocal teaching can build a new teaching paradigm that combines science and art, precision and individuality.

B. INNOVATION POINTS

Build a multimodal data acquisition and synchronization system for vocal teaching, achieving four-dimensional in-

tegration of audio, physiology, posture, and airflow.

Establish quantifiable evaluation indicators and models for singing skills that are computable, interpretable, and trainable.

Design a personalized teaching engine that includes defect diagnosis, path generation, and dynamic iteration to form a complete closed loop.

Propose a vocal evaluation mechanism that integrates subjectivity and objectivity, balancing scientific and artistic aspects.

C. LIMITATIONS

Large equipment is difficult to be completely portable;

The quantification of emotional expression still needs to be deepened;

While the current study focuses on a cohort of 60 university students specializing in Bel Canto and National styles, the model's transferability to highly stylized genres such as Opera or Pop remains to be fully validated. The variance in formant positioning (F1 , F2) and laryngeal stability across diverse vocal traditions necessitates the expansion of the training dataset to include heterogeneous populations and varied aesthetic standards.

AUTHOR CONTRIBUTIONS

Jiakai Cao designed, collected and analyzed the data, and drafted the manuscript. Jiakai Cao conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Jiakai Cao participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

HUMAN RESEARCH SUBJECTS

This study was approved by the Ethics Committee of The Philippine Women's University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] A. Yuhong and A. Wuyun, "A Method for Evaluating and Improving Vocal Pronunciation Quality of Vocal Music Students Based on Artificial Intelligence Technology," International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), vol. 21, no. 1, pp. 1–22, 2026, doi: 10.4018/IJWLTT.402021.

[2] C. Lei, "An Intelligent and Adaptive Vocal Teaching System Based on Enhanced Human–Computer Interaction and A2MARL," International Journal on Artificial Intelligence Tools, to be published, 2026, doi: 10.1142/S021821302550023X.

[3] B. Dhiman, "Sonic Authority and Actor Training: Amrish Puri and the Call for a Multicultural Vocal Pedagogy," Voice and Speech Review, vol. 20, no. 1, pp. 138–142, 2026, doi: 10.1080/23268263.2025.2566604.

[4] T. Fu, "A Preliminary Study on the Diversified Application of Adult Vocal Teaching Mode in the New Century," Journal of Sociology and Education, pp. 1(10), 2025, doi: 10.63887/JSE.2025.1.10.33.

[5] L. Jiang and D. Li, "A Study on Personalized Identification and Evaluation of Higher Vocational Teaching Based on Multimodal Data," Proceedings of the 2024 9th International Conference on Cyber Security and Information Engineering, pp. 1003–1007, 2024, doi: 10.1145/3689236.3696745.

[6] W. Luyao, "Application Effect Evaluation of Digital Teaching Tools in Vocal Music Teaching for Music Majors in Colleges and Universities," Journal of Sociology and Education, pp. 1(9), 2025, doi: 10.63887/JSE.2025.1.9.19.

[7] H. Ji, "Exploration on the Application of Artistic Voice Medicine in Vocal Performance Practice," Arts Studies and Criticism, pp. 6(5), 2025, doi: 10.32629/ASC.V6I5.4605.

[8] F. Su and Q. Guo, "The optimization of vocal music teaching by integrating the STEAM concept with the intelligent recommendation system," Scientific reports, vol. 16, no. 1, pp. 739–739, 2025, doi: 10.1038/S41598-025-30288-8.

[9] C. Chen, "Integrating Musical Form and Vocal Pedagogy: An Analysis of Susanna's Aria Giunse alfin il momento," Highlights in Art and Design, vol. 12, no. 2, pp. 1–10, 2025, doi: 10.54097/PE1ZP076.

[10] Y. Gao, "Digital Tools in Instrumental and Vocal Pedagogy: A Pre- and Post-Pandemic Analysis at the University Level," Journal of Education and Educational Research, vol. 15, no. 3, pp. 35–39, 2025, doi: 10.54097/ZBCK7A58.

[11] K. Wang and Y. Lei, "The Influence of Intelligent Speech Recognition on the Interactivity and Learning Efficiency of Vocal Music Teaching," Journal of Higher Education Teaching, pp. 2(5), 2025, doi: 10.62517/JHET.202515502.

[12] L. Pan, Q. Chen, L. Lou, et al., "Re-validation of the Kenny Music Performance Anxiety Inventory-revised among Chinese university students majoring in vocal music," Frontiers in Psychology, pp. 161667404–1667404, 2025, doi: 10.3389/ FPSYG.2025.1667404.

[13] M. Xiang, "Music Vocal Spectrum Classification Model based on Improved SVM Algorithm," 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), pp. 1244–1247, 2022, doi: 10.1109/ icaiss55157.2022.10010883.

[14] J. Yi and Y. Zhang, "Exploration of Innovative Approaches in Applying Information Technology to Vocal Music Teaching in University," Journal of Contemporary Educational Research, vol. 9, no. 8, pp. 198–204, 2025, doi: 10.26689/JCER.V9I8.11342.

[15] L. Chai, "Research on the Application of "Three-Wheel Cycle" Progressive Vocal Music Basic Training Mode under the Background of Digital Intelligence," Contemporary Education Frontiers, vol. 3, no. 7, pp. 55–60, 2025, doi: 10.18063/ CEF.V3I7.773.

[16] P. Qi, "Mode Construction of Red Music Culture Integrating into Vocal Music Teaching in Colleges and Universities from the Perspective of "Ideological and Political Curriculum"," Contemporary Education Frontiers, vol. 3, no. 7, pp. 49–54, 2025, doi: 10.18063/CEF.V3I7.772.

[17] Q. Ying and M. Dingyuan, "Research on Vocal Music Identification and Classification Based on Energy Entropy Ratio and AlexNet Model," International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), vol. 19, no. 1, pp. 1–19, 2025, doi: 10.4018/IJCINI.386839.

[18] Z. Pan, "Analysis on the Integration of National Singing and Bel Canto in Vocal Music Teaching," International Education Forum, vol. 3, no. 7, pp. 65–70, 2025, doi: 10.26689/IEF.V3I7.11534.

[19] J. Wang, "Preserving and Innovating: Strategies for Integrating Traditional Chinese Singing Techniques into Contemporary Vocal Pedagogy in Higher Education," Research and Commentary on Humanities and Arts, pp. 3(7), 2025, doi: 10.70711/RCHA.V3I7.7828.

[20] Y. Cui, "The Application of Low Frequency Ultrasonic Signal Denoising Processing Technology in Vocal Music Teaching," Journal of The Institution of Engineers (India): Series C, vol. 106, no. 4, pp. 1–14, 2025, doi: 10.1007/S40032-025-01210-Y.