Vocal Learner Ability Profiles Integrating Graph Neural Networks

Li Tian
Department of Music, Xi'an Shiyou University, Xi'an 710065, Shaanxi, China
Corresponding author: Li Tian (e-mail: tianli0728@126.com)

ABSTRACT Existing competency assessments often evaluate individual skills independently and fail to represent the structural dependencies between professional knowledge and practical performance. This paper applies Graph Neural Networks to construct dynamic ability profiles for vocal learners based on knowledge-structure modeling. Practice records and curriculum data are integrated into a heterogeneous graph containing learners, skills, and knowledge points. An R-GCN first generates initial embeddings through relation-specific message passing, preserving the semantic effects of different edge types. A Temporal GAT then applies temporal encoding to dynamic graph sequences to capture the evolution of learner abilities over time. The final time-series ability embedding is mapped to five dimensions, including pitch, rhythm, breath, resonance, and emotional expression. Experimental results show that structural modeling achieves average node reconstruction accuracy of at least 0.81 and relationship prediction F1 of at least 0.80. A dynamic embedding similarity slope of 0.82 and ability-change direction cosine of 0.85 indicate high temporal consistency in tracking ability evolution. The proposed framework supports structured inference of learner competence and provides a technical basis for personalized vocal-training intervention.

INDEX TERMS graph neural network, ability profiles, heterogeneous graph, r-gcn, temporal graph attention

I. INTRODUCTION

VOCAL learning involves multiple couplings of physiological regulation, auditory feedback, and artistic expression. Its ability development is rooted in the deep connection between knowledge points and the dynamic accumulation of long-term training [1,2]. To precisely portray the comprehensive state of learners in the dimensions of pitch, rhythm, breath, resonance, and emotion, it is necessary to break through the limitations of single indicator evaluation and turn to structured cognitive modeling [3]. In recent years, educational artificial intelligence has focused on the computable representation of implicit abilities.

Especially in the field of complex skills, such as those manufacturing, there is an urgent need for an analytical framework that integrates domain knowledge structure and multimodal behavioral data (e.g., design sketches, machinery operation logs, and material testing results) to accurately assess professional competence [4-6]. Constructing a vocal learner portrait with semantic interpretability and temporal evolution capabilities not only provides a basis for personalized feedback, but also gives intelligent teaching systems cognitive reasoning capabilities [7,8], which is conducive to promoting the in-depth development of music education in the direction of data-driven and theoretical integration.

This study aims to construct a dynamic profiling system for vocal learners, overcoming the dual bottlenecks of existing methods in structural relevance and temporal sensitivity. The core contributions are reflected in three aspects. First, a multi-source heterogeneous graph construction strategy is proposed to uniformly encode learner practice behavior, acoustic parameter analysis results, and the syllabus knowledge system.

This strategy clarifies the multiple connections between knowledge points, skill elements, and repertoire, forming a graph structure with educational semantics. Second, a hierarchical graph neural network architecture is designed, combined with R-GCN to implement relationship-based message filtering and aggregation. This ensures that the impact of different edge types on node embeddings is differentiated, improving the semantic accuracy of the initial representation. Third, a TGAT network is applied to model learning sequences for time perception. Using relative time encoding and attention mechanisms, the system captures the heterogeneous evolution of ability development and identifies key turning points and skill transition intervals. Finally, a learnable projection layer maps the temporal embeddings to five professional dimensions: pitch, rhythm, breath, resonance, and emotion, generating a clinically interpretable ability profile. This framework not only achieves a paradigm shift from isolated skill assessment to structured ability network inference—but also provides an operational cognitive basis for personalized vocal teaching interv ention.

By offering a deep, structured understanding of competence, it promotes the intelligent music education system to move towards a deeper level of intelligent understanding and intervention.

II. RELATED WORK

In the field of learner modeling, some studies have attempted to infer implicit ability states through the Bayesian knowledge tracking method [9,10]. Ma Z's research, through theoretical analysis and empirical investigation, found that a blended learning model based on deep learning can effectively improve vocational college students' self-learning ability and mental health literacy, thus providing strong support for optimizing current mainstream teaching methods [11]. For music learning scenarios, some scholars used Mel-frequency cepstral coefficients and pitch contours to extract singing features and combined them with machine learning for automatic speech emotion recognition [12,13].

The rise of graph neural networks has provided a new paradigm for modeling structured relationships in complex educational systems. Research has shown that heterogeneous graph structures can effectively integrate multi-type interactions between learners, knowledge points, and resources. Relationship-aware models such as R-GCN can distinguish different types of message passing paths by edge types, significantly outperforming traditional methods in knowledge tracking tasks and have been widely used in anomaly detection and graph access [14,15]. However, existing applications still suffer from three shortcomings: first, most graph structure designs fail to fully consider the nonlinear skill coupling unique to arts disciplines; second, temporal modeling often focuses on changes in event frequency, ignoring the phased nature of qualitative change in ability; third, the output space often remains fixed in a binary mastery/non-mastery judgment, lacking a mapping mechanism for a multidimensional continuous ability space. To address these issues, this paper proposes a joint modeling framework that integrates heterogeneous graph construction, relation-specific aggregation, and temporal attention. This framework aims to reconstruct the evolutionary trajectory of vocal learners' abilities from the three dimensions of structure, semantics, and time.

III. CONSTRUCTION AND EMBEDDING LEARNING OF HETEROGENEOUS VOCAL KNOWLEDGE GRAPHS

Figure 1 shows a dynamic profiling of vocal learners' abilities based on a graph neural network. A heterogeneous graph is constructed by integrating four entities: learner, repertoire, skills, and knowledge points. Relation-specific information aggregation is performed using R-GCN to form a structured initial embedding. During the temporal modeling phase, Fourier encoding of time intervals is applied, and TGAT is used to capture the evolution of abilities in heterogeneous practice sequences. Finally, the timing is embedded and projected into five dimensions: pitch, rhythm, breath, resonance and emotion, generating an interpretable multi-dimensional ability development trajectory and achieving complete modeling from structural association to dynamic representation.

A. MULTI-SOURCE DATA FUSION AND HETEROGENEOUS GRAPH NODE DEFINITION

1) Heterogeneous Graph Node Type Definition and Multi-Source Information Encoding

To achieve structured representation of multimodal and multi-level information in the vocal learning process, this paper constructs a heterogeneous graph structure consisting of four entity nodes: learners, knowledge points, skills, and repertoire. Each node type is semantically modeled based on domain knowledge and empirical data to ensure that its attribute vector carries interpretable educational significance. Learner nodes are individual entities. Their initial attribute vector is composed of statistical features of historical practice behavior and basic physiological parameters, including average pitch deviation, rhythm stability variance, practice frequency density, and a weighted aggregation of teacher evaluations. This creates a low-dimensional dense vector that represents the learner's baseline ability at the beginning of training.

Knowledge point nodes are derived from the vocal theory system in the syllabus, covering topics such as phonation principles, breathing mechanisms, and vowel transition rules. Each knowledge point is assigned structured attributes, including a cognitive difficulty level (labeled by experts on a scale of 1–5), the knowledge module to which it belongs, and a label for typical error patterns. The initial embedding for this node is generated by mapping discrete attributes to a continuous space. For example, a learnable embedding lookup table is used to encode the difficulty level, which is then concatenated with the one-hot encoded module categories and normalized using a linear transformation.

2) Initialization and Structured Fusion Mechanism of Node Attribute Vectors

Based on the node definition, to ensure that the initial embedding can not only retain the physical meaning of the original data but also adapt to the nonlinear transformation requirements of the deep model, this paper designs a set of hierarchical attribute fusion strategies. For attributes with clear numerical indicators, they are directly normalized and used as embedding components; for categorical or hierarchical variables, they are mapped into continuous vector representations through a trainable embedding matrix. Finally, the initial embedding of each node is generated by its multi-source attributes through a gated fusion mechanism:

$$h_v^{(0)} = \sigma\left(W_g[e_{\text{num}}, E_{\text{cat}}]\right) \odot \tanh\left(W_l[e_{\text{num}}, E_{\text{cat}}]\right) \quad (1)$$

$h_v^{(0)}$ represents the initial embedding of node v; $[e_{\text{num}}, E_{\text{cat}}]$ is the original attribute vector after splicing

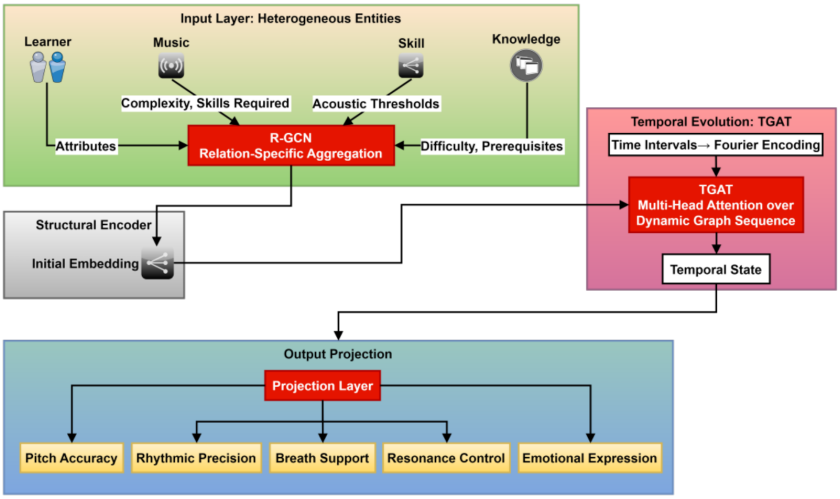

FIGURE 1. Dynamic profiling of vocal learners' ability based on graph neural networks

(numerical part and categorical embedding); W_g and W_l are learnable parameter matrices; σ is the sigmoid function; $\odot$ represents the element-by-element product. This gated structure allows the model to adaptively adjust the contribution strength of different attributes in the initial representation, avoiding the subjective bias caused by artificial weighting. In particular, for skill nodes, the tolerance range of acoustic indicators is further applied as a soft constraint to construct a regularization boundary in the embedding space, so that the subsequent message transmission process can respond more sensitively to abnormal performance exceeding the threshold.

B. RELATIONSHIP-SPECIFIC INFORMATION AGGREGATION BASED ON R-GCN

1) Multi-Type Relationship Modeling and Heterogeneous Graph Edge Structure Construction

To precisely depict the complex interactions between entities in the vocal learning process, this paper explicitly defines five types of directed edges with educational semantics in a heterogeneous graph: "learner-practice-repertoire", "repertoire-includes-skill", "skill-dependency-knowledge point", "knowledge-prerequisites-knowledge", and "learner-mastery-skill". Each type of edge carries a specific pedagogical logical relationship, avoiding the conflation of relationships of different natures. For example, the "repertoire-includes-skill" edge represents the specific technical elements required for performing a particular repertoire. Its weight is determined by the skill's frequency of occurrence and difficulty in the repertoire. The "skill-dependency-knowledge" edge reflects the degree to which a specific skill relies on theoretical knowledge. Given that this graph contains five semantically distinct directed relationships, standard GCN or GAT methods—which ignore edge types—fail to distinguish information flows across different relational paths, leading to semantic confusion. In contrast, R-GCN explicitly models the propagation mechanism of each edge type through relation-specific transformation matrices, making it the essential choice for processing such heterogeneous educational graphs.

Figure 2 visualizes a heterogeneous graph network structure tailored for vocal learning and its message passing process. The horizontal and vertical axes of the graph are two-dimensional projected coordinates generated using a force-directed algorithm to clearly visualize the complex topology. The numerical values themselves have no specific physical meaning. The graph contains four types of nodes: circles represent learners; squares represent repertoires; triangles represent skills; diamonds represent knowledge points.

Data is represented by the lines connecting the nodes, with five semantic relationships distinguished by different colors: blue solid lines represent "Practices" (connecting learners and repertoires); red solid lines represent "Contains" (connecting repertoires and skills); purple solid lines represent "Depends-On" (connecting skills and knowledge points); green solid lines represent "Prerequisites" (connecting knowledge points); orange solid lines represent "Masters" (connecting learners and skills). The thickness of the lines intuitively reflects the strength of the relationship. The figure specifically illustrates a complete message chain, highlighted in bold red, starting from the learner node 1, passing through the repertoire node 5, the skill node 11 contained in the repertoire, and finally reaching the knowledge node 18 on which the skill depends.

2) Relationship-Aware Message Delivery and Hierarchical Aggregation Mechanism

Based on the constructed heterogeneous graph, the R-GCN architecture is used for multi-hop information aggregation. The update process of each layer strictly follows the edge type for separated message passing to ensure that the impact of different types of relationship paths on the target node is independently modeled. Specifically, in the l-th layer, the embedding update of node v depends on its neighboring

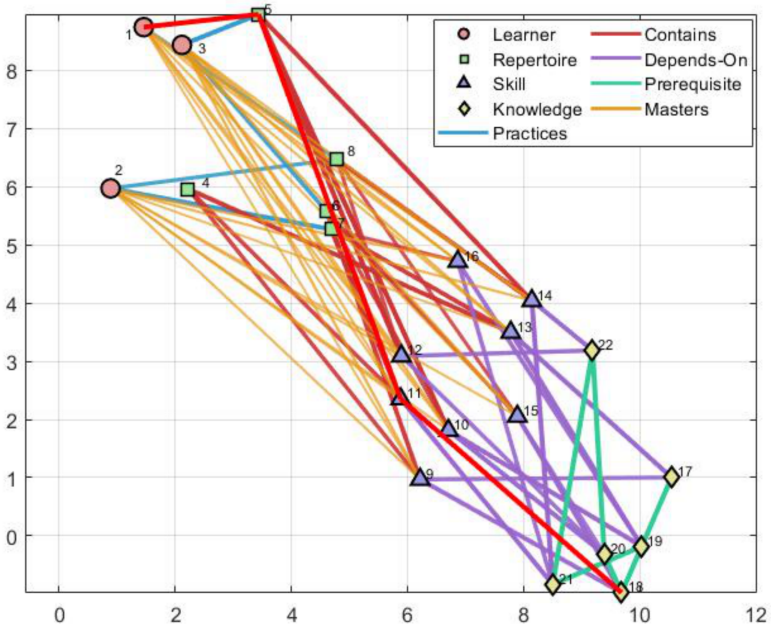

FIGURE 2. Heterogeneous graph structure and message passing

node set N_v^r under each type of relationship r, that is, it is achieved by aggregating the information of all source nodes connected to v through r-type edges. In the message generation stage, the current embedding $h_u^{(l)}$ of each source node u is linearly mapped through the transformation matrix of the corresponding relationship r to form a relationship-adapted message vector:

$$m_{v,r}^{(l)} = \sum_{u \in N_v^r} W_r^{(l)} h_u^{(l)} \tag{2}$$

$m_{v,r}^{(l)}$ represents the total message transmitted from the neighborhood through the relationship r, and $W_r^{(l)}$ is the learnable parameter matrix of the relationship r in the l-th layer, which is used to adjust the information transformation method of this type of relationship. This design allows the information of different semantic paths such as "skill-dependent knowledge points" and "repertoire containing skills" to undergo different linear transformations during the transmission process, thereby retaining the uniqueness of their educational meaning. The messages of all relationship channels are concatenated at the target node and then weighted and summed. The weights are controlled by the learnable relationship importance vector α_r to achieve adaptive enhancement of key relationship paths.

After completing the aggregation of multi-relational messages, the new embedding of the target node v is updated through a nonlinear activation function and residual connection:

$$h_v^{(l+1)} = \sigma\left(\sum_r m_{v,r}^{(l)} + b^{(l)}\right) + h_v^{(l)} \tag{3}$$

$b^{(l)}$ is the bias term. The residual connection alleviates the gradient degradation problem caused by deep stacking. Through multiple layers of iteration, the node embedding gradually integrates high-order neighborhood structure information. For example, the learner node not only absorbs the skill features of the repertoire it directly practices, but also indirectly perceives the knowledge point system that these repertoires rely on, thereby forming a deep representation that integrates teaching structure and individual behavior. The final output node embedding serves as the initial state of subsequent temporal modeling, and has structural sensitivity and semantic consistency to the vocal ability network.

Figure 3(a) shows the evolution of the relative contribution of relation-specific messages at different R-GCN layers as the number of layers increases: with increasing layers, the contribution of Skill-Depends-On-Knowledge increases (0.15 to 0.30), while the direct contribution of Learner-Masters-Skill decreases relatively (0.25 to 0.15). This trend indicates that multi-hop propagation and relation-specific projection matrices enable the model to rely more heavily on knowledge dependency paths to update learner representations at deeper layers, rather than relying solely on single skill observations. Figure 3(b) shows the evolution of node embeddings after principal component dimensionality reduction, which includes four clusters: Learner, Skill, Knowledge, and Repertoire. It can be seen that the clusters gradually converge from a scattered distribution at the shallow level to semantic alignment (Layer 1 to Layer 3). Relationship-aware aggregation achieves semantic alignment of knowledge-skill coupling, helping to reveal the knowledge support structure behind skill mastery.

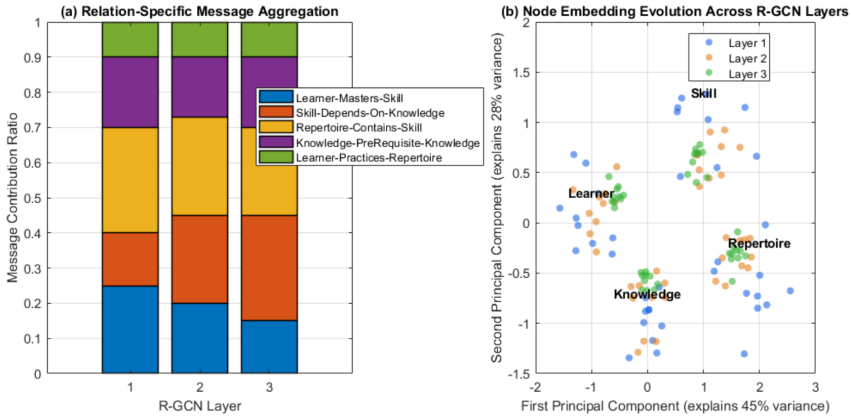

FIGURE 3. Message passing contribution and embedding evolution

C. SEQUENCE MODELING OF TIME-AWARE GRAPH ATTENTION NETWORKS

1) Dynamic Graph Sequence Construction and Time-Aware Message Enhancement

To capture the non-uniform evolution of vocal ability during training, this paper models the learner's behavior trajectory as a series of timestamp-ordered dynamic graph snapshots, each corresponding to the moment of a complete practice event. The nodes in the graph structure remain consistent across time, but the existence and attributes of edges are dynamically updated with the practice content. The node initial embedding used here is the 128-dimensional structured representation output by R-GCN, which contains knowledge-skills coupling semantics obtained by relationship-specific aggregation. For example, when a learner completes practice of a new piece, the corresponding "learner-practice-piece" edge is added to the subgraph at the current moment, carrying the acoustic assessment score and duration of this practice. This event-driven graph construction avoids the information dilution or misalignment caused by fixed time windows, ensuring that the graph sequence precisely aligns with the actual training rhythm. To address the problem of highly irregular time intervals between adjacent graph snapshots (for example, two exercises may be separated by hours or days), a time interval-based Fourier encoding mechanism is applied to map the time difference τ into a periodic feature vector to capture the cyclical laws in ability development (such as weekly training patterns) and long-term attenuation effects.

In the specific implementation, the time interval τ is transformed by high-frequency basis functions to generate a time encoding vector $TE(\tau)_i$:

$$TE(\tau)_i = \begin{cases} \sin(\omega_i \tau), & \text{if } i \text{ even}, \\ \cos(\omega_i \tau), & \text{if } i \text{ odd}, \end{cases} \quad \omega_i = \frac{1}{10000^{2i/d_t}} \tag{4}$$

This section employs a simplified Fourier fundamental frequency setting suitable for modeling non-uniform time intervals, which, unlike the absolute position encoding of the standard Transformer, is designed to effectively capture relative time dynamics. The time coding here is not based on absolute calendar time, but on the time difference between adjacent exercise events, to accommodate non-uniform training rhythms and capture the relative timing effects of ability decay and consolidation.

2) Temporal Dependency Modeling and State Update Based on Multi-head Attention

On the graph structure of each time step, the time-aware graph attention mechanism is used to selectively aggregate the historical graph state to achieve fine-grained tracking of the ability evolution path. For the target node v at the current time t, its update process not only depends on the neighborhood structure in the current graph, but also needs to retrieve the high-order dependency path with semantic association from the historical graph sequence. To this end, a cross-temporal attention function is defined to calculate the attention weight between the current node v and similar nodes u in all historical moments. This weight is determined by three parts: node semantic similarity (measured by the dot product of the embedding vector), the time decay factor (an exponential decay term based on the time interval τ), and the consistency of the relationship path.

The attention mechanism adopts a multi-head structure for parallel calculation. Each attention head h independently learns different dependency patterns and finally weights them together:

$$\alpha_{v,u}^{(h)} = \frac{\exp\left(a_h^\top \cdot g(W_h h_u^{(t')}, W_h h_v^{(t)}, TE(\tau))\right)}{\sum_{u' \in H_v} \exp\left(a_h^\top \cdot g(W_h h_{u'}^{(t')}, W_h h_v^{(t)}, TE(\tau))\right)} \tag{5}$$

$\alpha_{v,u}^{(h)}$ represents the influence weight of the historical node u on the current node v under the h-th attention head; g is the nonlinear activation function LeakyReLU (Leaky Rectified Linear Unit); W_h is the head-specific projection matrix; $a_h^\top$ is the learnable attention vector; H_v is the set of candidate historical nodes. Through the multi-head mechanism, the model can simultaneously capture multiple

evolutionary modes such as "skill consolidation", "ability degradation", and "migration enhancement". Finally, the outputs of each head are spliced and linearly transformed to generate the time-series updated embedding of the current node, which is then fused with the R-GCN output in the current graph to complete the state iteration.

Table 1 lists the key hyperparameter configurations used in the R-GCN model training process. This setup aims to balance model expressiveness and training stability. The number of GNN (graph neural network) layers is set to 3 to avoid oversmoothing caused by deep stacking. The hidden dimension is set to 128, balancing expressiveness and computational efficiency. The Adam optimizer is used with a learning rate of 0.001 to accelerate convergence, and dropout (0.3) and L2 regularization (5e-4) are applied to prevent overfitting. The batch size is 64 to ensure the stability of gradient estimation. Training is performed for a maximum of 200 epochs, with an early stopping mechanism of 20 epochs and quasi-dynamic termination based on the validation set loss to improve generalization performance. These parameters work together to ensure the effectiveness and robustness of relation-specific information aggregation.

TABLE 1. R-GCN model training hyperparameter configuration

Hyperparameter	Value	Description
Number of GNN Layers	3	Controls the number of message passing hops
Hidden Dimension	128	Feature dimension of node embeddings
Learning Rate	0.001	Step size for parameter updates
Dropout Rate	0.3	Proportion of units randomly dropped
Batch Size	64	Number of samples per training iteration
Optimizer	Adam	Adaptive moment estimation algorithm
L2 Regularization Coefficient	5e-4	Weight decay strength
Maximum Training Epochs	200	Upper limit of training iterations
Early Stopping Patience	20	Maximum epochs to wait for loss improvement

D. MAPPING AND PROFILE GENERATION OF FIVE-DIMENSIONAL ABILITY SPACE

1) Learnable Projection Mechanism of Multi-Dimensional Ability Space

To achieve semantic alignment from high-dimensional temporal embedding to interpretable vocal ability dimensions, this paper designs a dimensionality reduction structure based on learnable linear mapping, which precisely projects the temporal learner embedding vector output by the model into a continuous ability space composed of five professional dimensions: pitch, rhythm, breath, resonance, and emotion. These five dimensions are derived from the consensus competency framework in the field of vocal education, rather than the result of automatic clustering or dimension reduction of the model, aiming to ensure that the output results have teaching interpretability and intervention guidance value. This projection process does not use a linear combination of fixed weights, but instead applies five sets of independent trainable projection vectors, corresponding to the semantic directions of the five major ability dimensions. At each time step t, the final embedding h_t of the learner node is subjected to the inner product operation with each projection vector to generate the potential ability score of each dimension at that moment:

$$s_k(t) = h_t^\top p_k + b_k, \quad k \in \{\text{Pitch}, \text{rhythm}, \text{breath}, \text{resonance}, \text{emotion}\} \tag{6}$$

$s_k(t)$ represents the learner's original output score on the ability dimension k at time t; p_k is the semantic projection direction of the dimension; b_k is the bias term used to adjust the baseline level of each dimension. This design enables the model to adaptively extract the most relevant feature combinations for a specific ability from complex structured embeddings, rather than relying on manually set rule mappings.

For example, the projection vector of the "pitch" dimension may focus more on the graph propagation path response related to fundamental frequency stability and interval jump accuracy, while the "emotion" dimension tends to activate the historical attention weights related to the semantic association of lyrics and the dynamic strength change pattern.

All raw scores $s_k(t)$ are then normalized to the [0, 1] interval through an S-shaped nonlinear transformation and converted into ability scores with teaching interpretation significance:

$$c_k(t) = \frac{1}{1 + \exp(-\gamma s_k(t))} \tag{7}$$

$c_k(t)$ is the final output ability curve value, and the scaling factor γ is a parameter for independent learning of each ability dimension. That is, each dimension has its own scaling factor. This parameter is shared among all learners and does not change with individuals, so as to ensure a uniform scale of the ability space and cross-learner comparability. This nonlinear mapping not only ensures the continuity and smoothness of the output, but also simulates the nonlinear features of vocal ability development - obvious progress in the early stage and slow improvement in the later stage, which is in line with the actual training rules.

2) Generation and Structured Representation of Dynamic Competency Profiles

Based on the above mapping results, the system generates five parallel competency development trajectories for each learner. Each curve uses time as the horizontal axis and competency score as the vertical axis, forming the core output of the multi-dimensional dynamic profile. To enhance temporal coherence, the score sequence output at discrete time points undergoes local weighted smoothing,

using a kernel smoothing method based on time-density adaptiveness. This ensures that detailed fluctuations are preserved during periods of intensive practice, and trend stability is maintained during periods of sparse practice. This process does not alter the directionality of the original competency changes, but eliminates transient noise caused by single abnormal performance events, making the profile more realistic.

IV. QUANTITATIVE COMPARISON OF DYNAMIC ABILITY PROFILES

A. EXPERIMENTAL DATA

The experimental data, derived from a conservatory's vocal teaching platform, encompasses 16 weeks of practice records from 128 learners of varying skill levels. This study was approved by the Ethics Committee of Xi'an Shiyou University, and was performed in accordance with the principles of the Declaration of Helsinki.

All eligible participants signed an informed consent form. The data includes: complete audio files and acoustic analysis reports for 7,424 singing practice sessions; a structured knowledge system based on the syllabus, encompassing 157 knowledge points and 43 core skills; technical annotations for each practice piece (associated knowledge points and skills); the teacher's multi-dimensional ratings and textual comments for each practice session. All data is anonymized and aligned by timestamp to form a multi-source sequence that supports dynamic graph modeling. The dataset is divided according to learner level and randomly allocated to the training set, validation set and test set in a ratio of 7:2:1. This ensures that all practice events of the same learner appear in only one subset, thereby avoiding information leakage in time series modeling.

Learners are divided into beginner, intermediate, and advanced groups based on their skill levels to ensure representativeness across the ability distribution. This dataset covers a wide range of training behaviors and skill levels, providing a robust and reliable foundation for validating the model's effectiveness in structural dependency modeling, tracking skill evolution, and generating multidimensional profiles.

B. ACOUSTIC FEATURES AND ATTRIBUTE FUSION MECHANISM

Based on real acoustic analysis results and expert annotations, the original singing performance data corresponding to different skill nodes and their preset tolerance interval thresholds are extracted. The scoring was completed independently by three vocal teachers with more than 10 years of teaching experience. The five ability dimensions were scored using a 5-point Likert scale. The intragroup correlation coefficient was 0.86, indicating good inter-rater reliability. The tolerance range defined by experts is based on vocal teaching experience and is calibrated by combining the pitch deviation distribution of historical high-scoring samples (top 10%). It belongs to the empirical-statistical hybrid threshold. Subsequently, the multimodal attributes of each node (learner, skill, knowledge, and repertoire) are fed into a gated fusion module for forward propagation. The learnable parameters within this module are trained and converged using historical data during the heterogeneous graph construction phase. Finally, the gating weights for each node type for different attribute types are calculated and output. This result directly reflects the effectiveness and rationality of the model's adaptive fusion strategy.

Figure 4 provides a detailed analysis of the acoustic features of skill nodes and the effectiveness of the attribute fusion mechanism. Figure 4(a) shows the pitch deviation distributions and expert-defined tolerance thresholds for three skill nodes (stable pitch, wide vibrato, and jump interval). Data for stable pitches is tightly clustered within the $\pm$15-cent range, indicating their requirement for precise physiological control. Data for wide vibrato is evenly distributed within the $\pm$20-cent range, reflecting their controlled fluctuations. The data for jump intervals, however, is shifted and dispersed, revealing widespread learners' difficulty executing jump intervals and systematic pitch deviations. These distributional differences stem from the varying physiological mechanisms and cognitive complexity of each skill, demonstrating the importance of using tolerance intervals as soft constraints in the embedding space: they enable the model to quantify the degree of deviation rather than simply judge correctness. The heatmap in Figure 4(b) reveals the adaptive adjustment of the weights of different node attributes by the gating mechanism. The "skill" node assigns high weights to acoustic properties (Pitch Deviation, 0.7) and Knowledge Module (0.8) because their performance directly depends on acoustic metrics and is associated with specific vocal theories. Practice frequency is given a low weight (0.1) because training volume contributes little to the definition of the skill's essence. The "learner" node prioritizes numerical behavioral data (Pitch Deviation, 0.9; Exercise Frequency, 0.8).

C. COMPARISON OF AGGREGATION EFFICIENCY IN GRAPH NEURAL NETWORKS

Based on the time series embedding vectors generated by R-GCN and the baseline GCN, a classifier is constructed after partitioning the training and test sets. After each training epoch, the model parameters are fixed, and the embedding representations of all nodes are extracted. During training, performance is monitored only using the validation set to determine early stopping; the test set is used only for a one-time evaluation of the final model. The embedding quality is evaluated on the test set using a linear support vector machine (SVM) classifier. Accuracy, measured on the vertical axis, reflects the model's ability to capture the semantic information of the graph structure.

Figure 5 demonstrates the advantages of relation-specific aggregation over general isotropic aggregation through quantitative comparative experiments. The horizontal axis represents the number of model training cycles, and the ver-

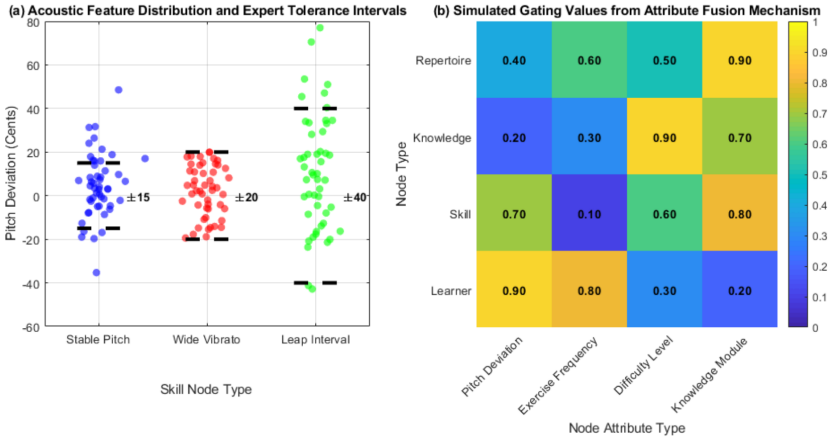

FIGURE 4. Acoustic features and attribute fusion mechanism

tical axis represents the accuracy on the node classification task, a key downstream metric for measuring the quality of generated embeddings. The two curves represent the performance evolution of the two models: the blue solid line represents the performance of the baseline model (isotropic GCN), which ignores relation type and aggregates all neighbor information equally; the red solid line represents the performance of the R-GCN model used in this paper. The data shows that in the early stages of training (Epochs 1-5), the performance of both models increases rapidly, but the rate of increase for R-GCN is significantly higher than that of the baseline. As training progresses, R-GCN converges rapidly to a higher performance plateau, while the baseline model converges more slowly and performs lower. After training stabilizes, the final accuracy of R-GCN (0.949) is significantly and consistently higher than that of the baseline model (0.893).

D. VERIFICATION OF STRUCTURAL DEPENDENCY MODELING CAPABILITIES

Node reconstruction accuracy (NRA) and relationship prediction F1 score are used as core validation metrics.

The node reconstruction task refers to the process where, after masking some knowledge point nodes, the model needs to predict their node type (i.e., identity) and their connection relationship with other nodes in the graph, rather than restoring the original feature values. During training, some knowledge point nodes and their associated edges are randomly masked. The model is required to reconstruct the identities and connectivity patterns of the removed nodes based on the contextual graph structure and learner behavior sequences to test its ability to implicitly encode the inherent logic of the knowledge system. Furthermore, in the relationship prediction task, hidden edges are sampled from key semantic paths to evaluate the model's balance between precision and recall in recovering true relationships under multi-hop reasoning. The proposed GNN is compared with the widely used GAT (Graph Attention Network), GraphSAGE (Graph SAmple and aggreGatE), TransE, and LightGCN.

Figure 6 shows an evaluation of the proposed model's modeling accuracy for vocal knowledge structure at different learning stages, using metrics such as node reconstruction accuracy (NRA) and relationship prediction F1. Figure 6's horizontal axis compares the performance of the proposed GNN framework with GAT, GraphSAGE, TransE, and LightGCN at the beginner, intermediate, and advanced stages, the error bars on the bars represent standard deviations. The results show that the proposed method achieves optimal performance at all stages, with average node reconstruction accuracy exceeding 0.81 and average relationship prediction F1 exceeding 0.80, demonstrating its superior ability to capture complex semantic dependencies between skills and knowledge points. As learning progresses, the stability of structural recovery varies across models: while the compared methods show limited improvement at the advanced stage, the proposed model maintains a continuous improvement trend, attributed to its fine-grained modeling of heterogeneous relationships and dynamic perception of temporal evolution paths.

E. TEMPORAL CONSISTENCY IN CAPTURING ABILITY EVOLUTION TRENDS

The Dynamic Embedding Similarity Slope (DESS) and Capability Change Direction Cosine (CCDC) are used to measure the consistency of the model output embedding sequence with the actual ability improvement trend. DESS measures the decay trend of cosine similarity between adjacent time-step embedding vectors, where values closer to 1 indicate smoother and more stable evolutionary trajectories. CCDC calculates the cosine similarity between the model's predicted five-dimensional ability increment vector (the difference between embedded projections at two consecutive time steps) and the teacher-labeled ability improvement direction vector. A value closer to 1 indicates that the model's predicted change direction aligns with expert observations. On the data of each learner's 8 consecutive weeks of practice, the trend fitting ability of the GNN in

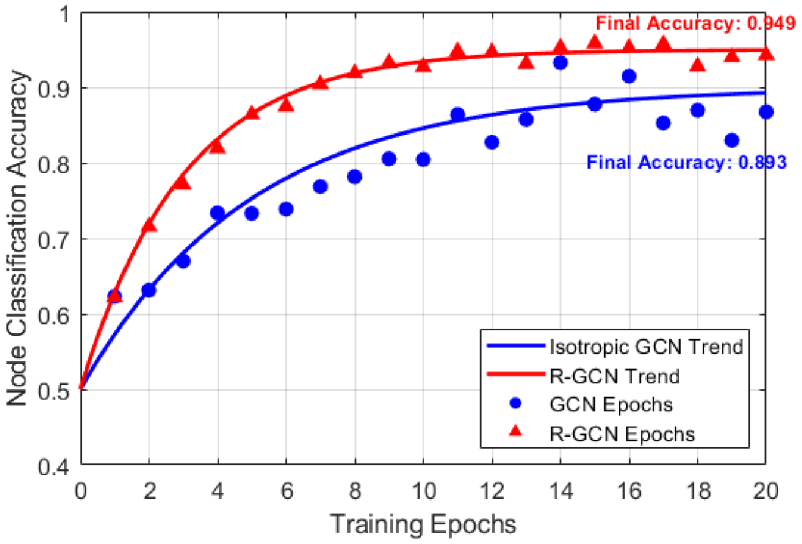

FIGURE 5. Comparison of aggregation efficiency in graph neural networks

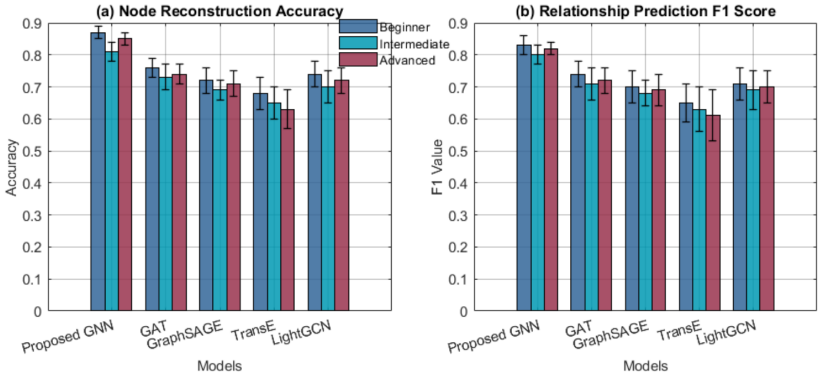

FIGURE 6. Node reconstruction accuracy and relationship prediction F1

this paper is compared with that of GAT, GraphSAGE, TransE, and LightGCN under different practice frequencies (low frequency $<$ 2 times/week, medium frequency 2-5 times/week, high frequency $\geq$ 5 times/week).

Figure 7 shows the temporal consistency modeling performance of different models for vocal ability evolution in a multi-frequency practice scenario. The overall trend indicates that the fitting accuracy of all models improves with increasing practice frequency, reflecting that denser behavioral data provides a more continuous and reliable signal foundation for ability trajectory modeling. Notably, the proposed GNN framework significantly outperforms the comparison models on both metrics, particularly in the high-frequency practice group, demonstrating that its time-aware architecture effectively captures subtle patterns of ability evolution in high-density sequences. In the low-frequency mode, the dynamic embedding similarity slope is 0.82, and the direction cosine of ability change is 0.85. The DESS results demonstrate the ability of each model to maintain monotonicity in ability progression, while CCDC emphasizes the directional alignment of coordinated changes in multidimensional abilities, which is more instructive for instructional interventions. The significant advantage of the proposed method over CCDC is attributed to its joint modeling of dynamic dependencies between skills and non-uniform time intervals, resulting in predicted directions of change that are closer to the actual cognitive pathways assessed by experts. In contrast, traditional graph models lack time-sensitive mechanisms, leading to information overload or response lag in high-frequency scenarios.

F. VALIDATION OF DECOUPLED REPRESENTATION OF MULTIDIMENSIONAL ABILITIES

To validate the model's ability to achieve decoupled representation in a five-dimensional ability space, based on data from a specialized intervention experiment, a group of learners are selected for three consecutive weeks of practice before and after systematic breathing training. This training explicitly focuses on improving breath control, with theoretical expectations that only the "breath" dimension shows significant changes.

This paper designs Dimension Separation (DS) and

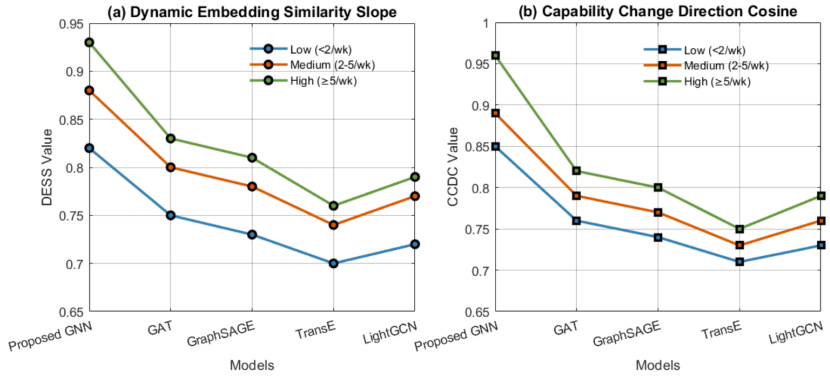

FIGURE 7. Temporal consistency in modeling ability trajectories

Cross-Skill Interference Ratio (CIR) as primary quantitative metrics, and also calculates the average non-target fluctuation and the magnitude of change in the breath dimension. The DS measures the prominence of the target dimension's response intensity relative to other dimensions by calculating the standard deviation ratio of the change in scores for each dimension before and after training. The CIR measures the proportion of falsely significant changes in non-target dimensions, reflecting the degree of semantic leakage during the mapping process. The amplitude of change in the breath dimension represents the average improvement in the breath dimension. The average non-target fluctuation is the mean standard deviation of the scores for the four non-target dimensions of pitch, rhythm, resonance, and emotion, used to measure the degree of bypass activation during the mapping process.

Table 2 systematically evaluates the decoupling performance of different models in the multidimensional ability space under targeted breathing training intervention, revealing essential differences among the methods in semantic separability and change localization accuracy. The proposed GNN framework exhibits optimal dimensionality independence. Its high dimensionality separation (0.89 ± 0.03) and low cross-skill interference rate (0.12) demonstrate that the model precisely anchors training-induced ability changes to the "breath" dimension while minimizing the impact on non-target dimensions such as pitch, rhythm, and emotion. In contrast, general-purpose graph models such as GAT and GraphSAGE exhibit significant side effects in non-target dimensions, reflecting that their mapping mechanisms fail to fully decouple physiologically related but functionally independent skill modules. TransE and LightGCN, due to their lack of explicit modeling of educational semantic structure, result in diffuse propagation of ability change signals in the embedding space, generating extensive cross-dimensional interference. The advantage of the proposed method stems from its dual constraint mechanism, which integrates a knowledge dependency graph and a learnable projection. On the one hand, the heterogeneous graph structure guides information propagation along real physiological and cognitive pathways; on the other hand, the targeted projection into the five-dimensional ability space suppresses co-activation of irrelevant dimensions. This feature enables the model to not only identify the presence of ability changes but also accurately attribute them to the domains in which they occur, providing a clinically interpretable analytical tool for evaluating the effectiveness of personalized interventions.

G. FINE-GRAINED DISCRIMINATION TEST

Learner Discriminability Entropy (LDE) and Minimum Discriminable Change Threshold (MDCT) are applied to evaluate the model's ability to discriminate between learners of similar abilities. In a group of learners with similar total scores but different skill combinations (such as "strong intonation - weak breath" vs. "weak intonation - strong breath"), the profiling effects of the proposed GNN are compared with those of GAT, GraphSAGE, TransE, and LightGCN at different pitch change frequencies (divided into low frequency (<2) and high frequency (>6) according to the average number of note pitch changes per measure).

Table 3 shows the performance of different models in a fine-grained learner discrimination task, focusing on the ability to discriminate individuals with heterogeneous skill configurations but similar overall proficiency.

The results show that the proposed GNN exhibits optimal discrimination performance under both types of pitch variation density. At low frequencies, it achieves the highest learner discrimination entropy (0.87 ± 0.03) and the lowest minimum discernible difference threshold (0.11 ± 0.02), indicating that the generated ability profiles are more individual-specific and sensitive. The model's advantage is further enhanced in the context of high-frequency pitch variation, reflecting its ability to deeply model the dynamic coupling of multiple skills in complex vocal tasks. In contrast, general graph models such as GAT and GraphSAGE lack explicit constraints on educational semantic paths, making it difficult to effectively separate highly correlated skill dimensions, resulting in a fuzzy ability profile. TransE and LightGCN, due to their neglect of relational heterogeneity and temporal evolution, tend to have smoothed embedding spaces and are less responsive to subtle differences in ability. By integrating heterogeneous graph structures with

TABLE 2. Validation of the decoupled expression of multidimensional abilities

Model	Dimension Separation	Breath Dimension Change	Average Non-target Fluctuation	Cross-skill Interference Ratio	Primary Interference Dimensions
Proposed GNN	0.89 ± 0.03	0.32	0.08	0.12	Resonance (weak coupling)
GAT	0.76 ± 0.05	0.27	0.14	0.28	Resonance, Pitch
GraphSAGE	0.73 ± 0.06	0.25	0.16	0.31	Rhythm, Resonance
TransE	0.65 ± 0.07	0.21	0.19	0.42	Pitch, Rhythm, Emotion
LightGCN	0.70 ± 0.05	0.23	0.17	0.36	Emotion, Peripheral Breath Sub-dimensions

a time-aware attention mechanism, the proposed method constructs a more disentangled ability representation in a high-dimensional space. This fine-grained resolution not only improves the cognitive accuracy of the profiling system but also provides a reliable basis for personalized teaching interventions, achieving a paradigm shift from analyzing "group average trends" to "individual ability topology".

TABLE 3. Fine-grained discrimination test

Model	Test Condition	Learner Discriminability Entropy	Minimum Discriminable Change Threshold
Proposed GNN	Low Frequency	0.87 ± 0.03	0.11 ± 0.02
Proposed GNN	High Frequency	0.93 ± 0.02	0.07 ± 0.01
GAT	Low Frequency	0.72 ± 0.05	0.18 ± 0.03
GAT	High Frequency	0.76 ± 0.04	0.16 ± 0.03
GraphSAGE	Low Frequency	0.69 ± 0.06	0.20 ± 0.04
GraphSAGE	High Frequency	0.73 ± 0.05	0.19 ± 0.03
TransE	Low Frequency	0.61 ± 0.07	0.25 ± 0.05
TransE	High Frequency	0.64 ± 0.06	0.23 ± 0.04
LightGCN	Low Frequency	0.66 ± 0.05	0.22 ± 0.04
LightGCN	High Frequency	0.70 ± 0.05	0.20 ± 0.03

V. CONCLUSION

This paper proposes a vocal learner ability profiling framework that integrates graph neural networks and dynamic semantic modeling, an approach that is equally valuable in complex vocational fields like . By breaking through the limitations of isolated dimension analysis and static representation in traditional assessments, the framework can provide a more holistic and dynamic understanding of competency. By constructing a heterogeneous graph structure encompassing learners, knowledge points, skills, and repertoire, and combining it with R-GCN to aggregate relationship-specific information, it effectively captures the deep dependencies between vocal knowledge. TGAT is applied to perform time-aware modeling of non-uniform practice sequences, precisely capturing the phased and nonlinear features of ability evolution. Finally, through learnable projection, the time series embedding is mapped into the five-dimensional space of pitch, rhythm, breath, resonance, and emotion, generating a dynamic ability profile with pedagogical explanatory power. Experimental verification demonstrates that the average node reconstruction accuracy is no less than 0.81; the average relationship prediction F1 value is no less than 0.80; the dynamic embedding similarity slope is as low as 0.82; the direction cosine of ability change is as low as 0.85. This model significantly outperforms mainstream baseline methods in terms of structural recovery and trend consistency, demonstrating enhanced robustness and discriminative accuracy in sparse data and high-confusion scenarios. This research not only advances the capability modeling paradigm from behavioral description to cognitive mechanism inference, but also provides an interpretable, interventional, and traceable cognitive foundation for intelligent vocal teaching systems.

Ethics Approval and Consent to Participate This study was approved by the Ethics Committee of Xi'an Shiyou University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

AUTHOR CONTRIBUTIONS

Tian Li designed, collected and analyzed the data, and drafted the manuscript. Tian Li conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Tian Li 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 was supported by the major research project of social sciences of Shaanxi Province:

"Composition and Performance of New Vocal Music Works Based on Classical Chinese Poetry", grant number 2022HZ1659.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] J. N. Audet, M. Couture, and E. D. Jarvis, "Songbird species that display more-complex vocal learning are better problem-solvers and have larger brains," Science, vol. 381, no. 6663, pp. 1170-1175, 2023, doi: 10.1126/science.adh3428.

[2] A. Rybner, E. T. Jessen, M. D. Mortensen, S. N. Larsen, R. Grossman, N. Bilenberg, et al., "Vocal markers of autism: Assessing the generalizability of machine learning models," Autism Research, vol. 15, no. 6, pp. 1018-1030, 2022, doi: 10.1002/aur.2721.
[3] Z. You, B. Han, Z. Shi, M. Zhao, S. Du, J. Yan, et al., "Vocal cord leukoplakia classification using deep learning models in white light and narrow band imaging endoscopy images," Head & Neck, vol. 45, no. 12, pp. 3129-3145, 2023, doi: 10.1002/hed.27543.
[4] T. O'Rourke, P. T. Martins, R. Asano, R. O. Tachibana, K. Okanoya, and C. Boeckx, "Capturing the effects of domestication on vocal learning complexity," Trends in Cognitive Sciences, vol. 25, no. 6, pp. 462-474, 2021, doi: 10.1016/j.tics.2021.05.002.
[5] Q. Zhao, Y. He, Y. Wu, D. Huang, Y. Wang, C. Sun, et al., "Vocal cord lesions classification based on deep convolutional neural network and transfer learning," Medical Physics, vol. 49, no. 1, pp. 432-442, 2022, doi: 10.1002/mp.15371.
[6] J. A. Cahill, J. Armstrong, A. Deran, C. J. Khoury, B. Paten, D. Haussler, et al., "Positive selection in noncoding genomic regions of vocal learning birds is associated with genes implicated in vocal learning and speech functions in humans," Genome Research, vol. 31, no. 11, pp. 2035-2049, 2021, doi: 10.1101/gr.275989.121.
[7] C. W. Espinola, J. C. Gomes, J. M. S. Pereira, and W. P. dos Santos, "Vocal acoustic analysis and machine learning for the identification of schizophrenia," Research on Biomedical Engineering, vol. 37, no. 1, pp. 33-46, 2021, doi: 10.1007/s42600-020-00097-1.
[8] Y. Shi, "The use of mobile internet platforms and applications in vocal training: Synergy of technological and pedagogical solutions," Interactive Learning Environments, vol. 31, no. 6, pp. 3780-3791, 2023, doi: 10.1080/10494820.2021.1943456.
[9] É. R. Santana, L. Lopes, and R. M. de Moraes, "Recognition of the effect of vocal exercises by fuzzy triangular naive Bayes, a machine learning classifier: A preliminary analysis," Journal of Voice, vol. 39, no. 2, pp. 560.e21-560.e30, 2025, doi: 10.1016/j.jvoice.2022.10.001.
[10] Y. Wu, E. D. Jarvis, and A. Sarkar, "Bayesian semiparametric Markov renewal mixed models for vocalization syntax," Biostatistics, vol. 25, no. 3, pp. 648-665, 2024, doi: 10.1093/biostatistics/kxac050.
[11] Z. Ma, P. Ruannakarn, and A. Jansaeng, "The role of blended instructional models based on deep learning theory in enhancing students' autonomous learning ability in Chinese Higher Vocational Colleges," Journal of Statistics Applications and Probability, vol. 13, no. 3, pp. 947-959, 2024, doi: 10.18576/jsap/130309.
[12] M. D. Pawar and R. D. Kokate, "Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients," Multimedia Tools and Applications, vol. 80, no. 10, pp. 15563-15587, 2021, doi: 10.1007/s11042-020-10329-2.
[13] S. Mustafa, A. Khan, S. Hussain, M. Z. Jhandir, R. Kazmi, and I. S. Bajwa, "Automatic speech emotion recognition using Mel frequency cepstrum co-efficient and machine learning technique," Pakistan Journal of Engineering and Technology, vol. 4, no. 1, pp. 124-130, 2021, doi: 10.51846/vol4iss1pp124-130.
[14] Y. Fang, X. Li, R. Ye, X. Tan, P. Zhao, and M. Wang, "Relation-aware graph convolutional networks for multirelational network alignment," ACM Transactions on Intelligent Systems and Technology, vol. 14, no. 2, pp. 1-23, 2023, doi: 10.1145/3579827.
[15] M. Zhang, H. Liu, C. Chen, G. Gao, H. Li, and J. Zhao, "AccessFixer: Enhancing GUI accessibility for low vision users with R-GCN model," IEEE Transactions on Software Engineering, vol. 50, no. 2, pp. 173-189, 2023, doi: 10.1109/TSE.2023.3337421.