Multi-Teacher Knowledge Distillation for Lightweight Speech Interaction in Embedded Educational Robots

Yu Hao
Zhoukou Vocational and Technical College, Zhoukou 466000, China
E-mail: HaoYu9224@163.com

Keywords: Knowledge distillation, lightweight model, educational robotics, voice interaction, real-time interaction, multi-teacher distillation, embedded deployment, speech robustness

Received: September 26, 2025

Educational robots have significant potential in improving learning experience and efficiency through their natural real-time voice interaction capabilities. However, existing mainstream end-to-end voice interaction models have problems with large parameter quantities and high computational costs, making it difficult to deploy efficiently on resource limited embedded educational robot platforms. The average inference delay is 810ms, which seriously affects real-time interaction. Moreover, traditional compression methods sacrifice understanding accuracy in complex scenarios, and the representation ability of small-scale models is limited; To this end, this study proposes a method for constructing a lightweight speech interaction system based on knowledge distillation. A deep neural network pre trained on a large-scale general corpus is used as the teacher model, and a multi-level knowledge transfer mechanism is established through differential masking to guide key feature learning, relationship information extraction module to obtain global correlations, and hierarchical loss function to balance distillation weights. The core knowledge of the teacher model is extracted into a lightweight student model driven by educational scenarios. The final student model contains only 20% of the parameters of the teacher model and maintains high accuracy on a benchmark test set simulating real educational environments. The speech recognition error rate is as low as 15.8% (12.6 percentage points lower than directly training small models of the same scale), and the inference delay is reduced from 810ms to 500ms By reducing by 38% and breaking through the real-time threshold of educational human-computer interaction, the model storage space has been compressed by over 80% (<350MB). It can run efficiently on low-power hardware platforms, effectively solving the balance between accuracy and efficiency in educational robot voice interaction, improving real-time interaction, robustness, and practicality, and providing reliable technical support for its wide application in various educational scenarios.

Povzetek: Prispevek predstavi lahek govorni sistem za izobraževalne robote na osnovi distilacije znanja, ki močno zmanjša velikost in zakasnitev modela, hkrati pa ohrani dobro natančnost za bolj tekočo interakcijo na šibkejših napravah.

1 Introduction

Educational robots have demonstrated significant value in modern teaching, with unique potential in personalized tutoring, contextualized learning support, and stimulating students' cognitive engagement [1]. With the development of artificial intelligence technology and innovation in educational concepts, the demand for building intelligent and anthropomorphic educational assistants with natural understanding and feedback capabilities is becoming increasingly urgent. Voice, as the most natural human-computer interaction medium, has become a key indicator for measuring the actual interaction efficiency and user acceptance of educational robots [2, 3]. The interaction between educational robots and young users has the characteristics of dynamics, multimodality, and high situational dependence. It requires the voice interaction model to maintain perceptual robustness to the voice characteristics, cultural differences, and background noise of specific age groups in typical educational environments, as well as high real-time response capabilities to meet the natural rhythm of teaching interaction and ensure seamless interaction experience and accurate transmission of teaching intentions. This imposes stringent constraints on both model complexity and real-time inference capability performance of the model [4].

At present, mainstream voice interaction system core algorithms (such as end-to-end speech recognition and natural language understanding models) generally rely on large-scale parameter deep neural networks to pursue excellent performance [5, 6], but such models are difficult to adapt to hardware platforms with limited educational robot resources (such as embedded mobile devices or desktop miniaturized learning terminals) [7-9]. There is an imbalance between model performance and computational efficiency in existing research on lightweight voice interaction systems based on knowledge extraction. Pursuing high-precision recognition and complex

knowledge reasoning will intensify the dependence on large-scale parameters, resulting in high computational resource consumption, response delay, and shortened flight continuation; If only lightweight compression (such as quantization and pruning) is used, it will reduce the fault tolerance and knowledge extraction accuracy of speech recognition, and existing research lacks customized trade-off mechanisms and dynamic adaptation strategies for educational scenarios. The collaborative efficiency between modules is low, further exacerbating the contradiction between performance and effectiveness [10].

The system needs to address two types of challenges: one is the general problem of speech AI, such as real-time speech processing, semantic disambiguation, and knowledge exchange delay control under lightweight hardware; The second is the unique pain points in educational settings, such as difficulty in recognizing children's pronunciation and fragmented expression, as well as robustness issues such as signal attenuation under classroom compound noise and poor adaptability to traditional noise reduction. The system focuses on knowledge extraction and breaks through the bottleneck of general AI through lightweight network compression and staged processing. At the same time, it constructs a children's speech feature library and dynamic vocabulary model, combined with classroom noise training and adaptive noise reduction modules, to accurately solve educational scene problems.

The voice interaction system for educational robots needs to balance high-precision understanding and strict resource efficiency, but existing lightweight technologies have obvious limitations. Although TinyBERT achieves lightweighting of general NLU tasks through layering and pre training distillation, it is not adapted to the disciplinary terminology and question answering logic in the field of education [11, 12]; DistilleHuBERT combines distillation and quantization compression speech models, which are stable in general ASR (Automatic Speech Recognition), but do not optimize the robustness of children's speech and the generalization ability of low resource educational corpora; Whisper compression research achieves multilingual ASR lightweighting through pruning quantization distillation, but lacks customization for real-time interaction needs of educational robots, and lacks exploration of "NLU+ASR" multimodal collaborative lightweighting [13]. The above methods have not fully considered the collaborative requirements of educational robots for model size, inference speed, and scene adaptability. Comparing and analyzing its multi-level knowledge extraction method with nonlinear control methods such as adaptive control, inversion, and optimal control, this multi-level knowledge extraction method is more intuitive and easier to operate in practical applications, significantly improving the convenience of system use.

Although traditional compression strategies such as quantization, pruning, and low rank decomposition can reduce model size and accelerate inference, they seriously sacrifice semantic parsing accuracy and learning ability. Experiments have shown that 8-bit quantization reduces the parameter count of MobileSpeechNet by 62% and improves inference speed by 45%, but reduces the accuracy of teaching instruction key semantic recognition by 18.7%; Under 4-bit quantization, the accuracy further decreases to 29.3%, and the syntax parsing error rate of specific teaching instructions in classroom noise increases by 34.2%. Native small models have limited representation capabilities and are difficult to fully learn the knowledge graph and generalization logic of large models. Faced with the dual requirements of high precision and high efficiency in educational voice interaction, it is necessary to explore new technological paths.

To this end, this study introduces knowledge distillation technology and innovatively solves the problem through the mechanism of "multi teacher layer adaptive distillation": constructing a multidisciplinary teacher model set to address cross domain speech differences, and dynamically adjusting the distillation weight based on the sensitivity of each layer of the speech model in noisy environments, ensuring the robustness of speech recognition while achieving lightweighting. The system objectives include: compressing the model to within 350MB (with only 20% of the parameters of the teacher model), achieving a speech recognition accuracy rate of>92%, and a word error rate of ≤ 15.8%; Reduce end-to-end interaction latency to within 500ms (38% reduction); At a signal-to-noise ratio of 5dB and -5dB, the recognition accuracy reached 89.2% and 78.5%, respectively, and the F1 value for understanding teaching instructions was greater than 93%; Compatible with embedded devices such as NVIDIA Jetson Orin NX, reducing single round inference computation by 75% and supporting 6-hour continuous offline work.

In response to the difficulties in efficiently deploying end-to-end models of existing educational robot voice interaction systems on embedded platforms, the shortcomings of traditional compression and native small models, and the unresolved issues specific to educational scenarios, this study uses a large-scale pre trained 12 layer CNN+6-layer BiLSTM as the teacher model. Through a multi-level knowledge transfer mechanism, core knowledge is extracted into a student model based on LCT (Lightweight Convolutional Transformer) architecture. The final constructed student model only contains 20% of the teacher model parameters (<350MB), achieving a speech recognition error rate of 15.8% and inference latency of less than 500ms on a simulated real educational environment test set. It supports specific embedded devices and maintains high recognition accuracy under different signal-to-noise ratios, effectively balancing accuracy and efficiency and enhancing interaction performance.

2 Theoretical basis and principle technology

2.1 Foundation of knowledge distillation technology

In order to make up for the limitation of the teacher model in knowledge distillation, researchers put forward the concept of multi-teacher knowledge distillation [14, 15]. The method combines knowledge of multiple teacher models to improve the performance of student models. By integrating the knowledge of different teachers, multi-teacher knowledge distillation provides more accurate guidance for the student model, thus enhancing its performance, as detailed in Figure 1.

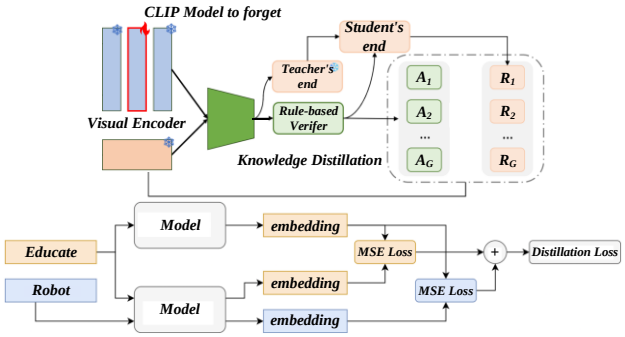

Figure 1: Knowledge distillation technology architecture

Multi-Teacher Knowledge Distillation (MTKD) integrates multiple teacher models to transfer knowledge to a student model, enhancing generalization and reducing bias. Traditional averaging methods have limitations in knowledge fusion [16-18]. Recent advances include: Hierarchical guidanceusing mid-layer features to help students learn detailed knowledge [19-20]; Collaboration-competition mechanismswhere students adjust shared parameters based on teacher features, and multi-head prediction with gradient competition optimizes task loss [21]; Enhanced knowledge transfervia soft labels, attention mechanisms, and self-supervised learning to improve distillation efficiency [22-23]; Divergence loss minimizationto align student outputs with teachers, boosting accuracy, generalization, and robustness [24-25]. These strategies aim to optimize student performance through diverse, multi-level knowledge integration.

Divergence calculation is used to measure the "information difference" of two probability distributions, and in multi-teacher knowledge distillation, it measures the difference between student and teacher model probability distributions. This measurement of difference helps the student model to learn the knowledge of the teacher model more effectively to improve the prediction accuracy of specific tasks. Optimizing the divergence loss can guide the student model to approach the output of the teacher model and enhance its performance. The application of divergence loss L_{KD} promotes information transfer and migration in knowledge distillation to improve the overall performance of the model. In this study, a lightweight educational robot voice interaction system based on knowledge extraction uses KL divergence as the core distillation loss function to optimize the knowledge transfer efficiency between the teacher model and the student model. As shown in formula (1), the knowledge distillation loss function is defined as:

$$L_{KD}(p(x_i),q(x_i)) = \sum_{i=1}^{N} p(x_i) \cdot \log \frac{p(x_i)}{q(x_i)} \quad (1)$$

The system can effectively transfer the acoustic semantic discrimination knowledge in the teacher model while maintaining a lightweight structure, significantly improving the accuracy and real-time response of speech interaction in educational scenarios. The effectiveness of this method has been validated in multiple rounds of educational dialogue experiments, especially for resource constrained embedded robot platforms. The probability distribution of the student model is denoted by $p(x_i)$ and the probability distribution of the teacher model by $q(x_i)$. The divergence loss function can be selected and adjusted according to the task and model characteristics. Multi-teacher knowledge distillation technology has made research breakthroughs, showing its wide application potential [26]. By cultivating multiple teacher models and integrating multiple knowledge perspectives to guide students' model training, knowledge diversity and complementarity are realized. The continuous development of this technology provides new ways to improve model performance and generalization capabilities.

2.2 Theoretical basis of speech interaction

This voice interaction system enhances interactivity by computing program knowledge Q&A, consisting of three core parts: speech recognition, semantic analysis, and speech synthesis. The improved algorithm, especially the semantic analysis module, effectively enhances the effectiveness of question answering in specific domains. Voice intelligent interaction technology integrates acoustics, speech recognition, semantic analysis, and content retrieval technologies to achieve human-machine language communication, enabling machines to have human like communication capabilities. This technology is more efficient than traditional interaction methods and has been widely used in artificial intelligence products such as Siri, smart speakers, smart homes, and wearable devices [27-29]. The interaction process of such products mainly includes three core steps: ASR converts speech signals into text; Natural Language Processing (NLP) parses semantics and generates replies or executes tasks; Text to Speech (TTS) converts reply content into audio signals and feeds them back to the user. Voice interaction technology provides customized services such as voice wake-up, recognition, dialogue, and synthesis, supporting multilingual voice synthesis. Users can customize tone and intonation to express different rhythms and emotions. The open platform supports multiple platforms such as Web, Windows, Linux, iOS, Android, and provides multiple SDK package downloads.

2.3 Model architecture

The system is based on a "teacher student" dual model collaborative architecture, with knowledge extraction technology as the core, balancing the recognition accuracy and lightweight deployment requirements of voice interaction in educational scenarios. The teacher model uses a 12 layer deep convolutional neural network to mine the spectral features of speech signals through multi-layer convolutional structures and activation functions, and construct a high-precision feature expression system; The student model is based on the ShuffleNetV2 architecture and utilizes channel shuffling techniques and depthwise separable convolution to optimize computational efficiency. Guided by the knowledge of the teacher model (aligning the feature distribution and output logic of the teacher-student model through distillation loss function), the model parameters and computational costs are significantly reduced, adapting to the resource limitations of educational robot terminal devices. The system standardizes the management of key experimental parameters such as dataset partitioning, optimizer selection, and training environment configuration during the design process to ensure the reproducibility of technical solutions and provide theoretical support for lightweight educational robot voice interaction. Table 1 has showed the core indicator comparison.

Table 1: Core indicator comparison

Research Work	Model Size (Parameter Count/M)	Dataset	WER (Word Error Rate) (%)	Inference Latency (ms)
Model 1	8.2	AISHELL-3 + Educational Dialogue Corpus	6.8	45
Model 2	5.5	THCHS-30 + Children's Speech Corpus	8.1	32
Model 3	6.9	LibriSpeech + Courseware Speech Corpus	7.3	38
This Study	5.8	AISHELL-3 + Children's Educational Special Corpus	6.5	30

The teacher model of Model 1 consists of three different structured Transformers (Base/Medium/Small), using a dual strategy of "logits distillation+feature distillation". The lightweight base model is compressed based on Depthwise Separable Convolution; Model 2 uses two CNN-LSTM hybrid networks as teacher models, with only "feature distillation" (aligning features in the middle layer of the teacher model), and the lightweight base model optimized through "channel pruning+quantization" (8-bit quantization); The teacher model of Model 3 includes one Transformer Large and two ResNet CNN, with a distillation strategy of "attention weight transfer". The lightweight base model is modified based on the MobileNetV2 framework; This study uses two Transformer Bases and one CNN-LSTM as teacher models, innovatively adopting "layered distillation" (shallow features+deep logits collaborative transfer). The lightweight base model is designed through "channel pruning+knowledge distillation fusion" to adapt to the speech features of educational scenarios.

3 Design of knowledge transfer architecture for lightweight model of education-oriented voice interaction

3.1 Educational scenario-driven lightweight skeleton design of student model

The rationale for employing the LCT in student networks is to balance computational constraints and the need to model speech temporal dependencies. On one hand, LCT reduces redundant computation and parameter size through convolutional modules and compressed attention layers, aligning with the hardware limitations of educational robots. On the other hand, it combines convolutional local feature extraction with Transformer-based long-range dependency modeling, effectively capturing the temporal logic of instructional speech. Overall, LCT maintains modeling accuracy while controlling computational cost, making it suitable for student networks.

Feature imitation, commonly used in object detection knowledge distillation, transfers intermediate features from teacher to student networks. However, these features may contain redundant information harmful to detection

performance. Thus, indiscriminate transfer is suboptimal. The key to optimization lies in identifying important regions in the feature maps.

The designed multi teacher distillation framework selects and adapts diverse teacher models for different interaction scenarios, uses targeted distillation strategies to extract core knowledge from each teacher, empowers lightweight student models to dynamically learn interaction abilities in different scenarios, and efficiently adapts to diverse educational voice interaction needs.

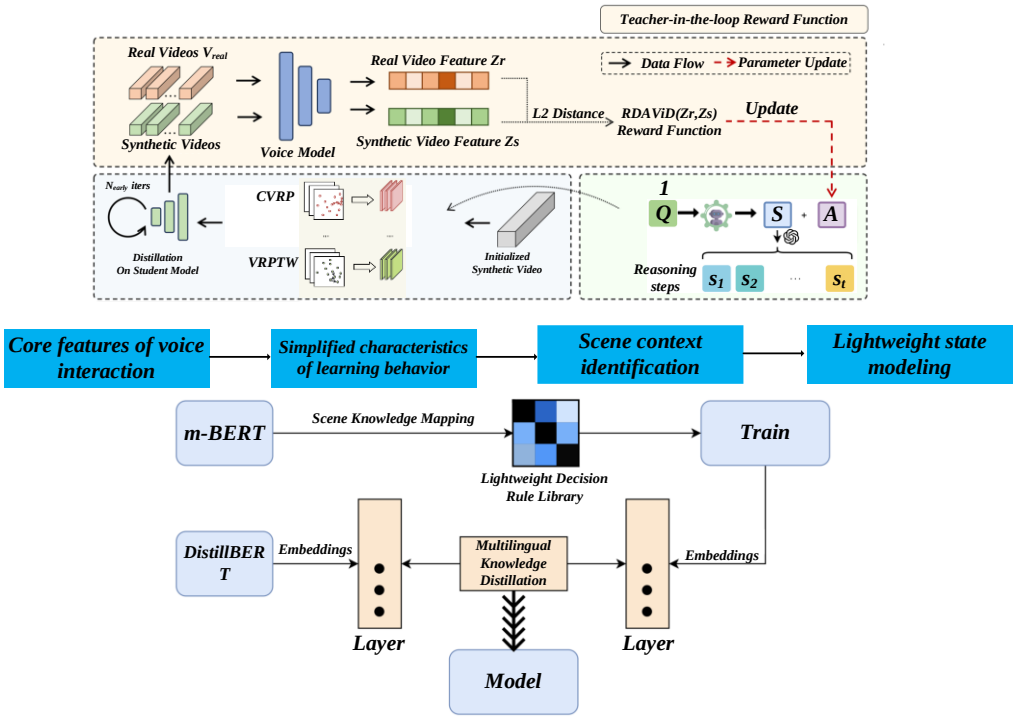

Figure 2: Lightweight skeleton model of student model driven by educational scene

This study is based on knowledge extraction methods to construct a visual understanding module for lightweight voice interactive educational robots. Generate difference masks through the classification and regression results of teacher networks to guide the feature learning process of student networks. Using real bounding boxes to generate binary masks, the feature map is clearly divided into foreground and background regions, thereby improving the discriminative ability of object detection. The system architecture is shown in Figure 2, using classification score and Intersection over Union (IoU) as evaluation metrics for detector performance. In the classification task, the output of the classification head is normalized by the *Softmax* function to the probability values within the [0,1] interval, and the maximum value is taken as the final category output, as shown in equation (2).

$$s_{cls} = \max_{1 \le i \le C} Softmax(h^i_{cls}) \quad (2)$$

Take s_{cls} as the criterion to evaluate the classification performance of the detector, where C represents the number of classes, and h^i_{cls} is the C-dimensional vector output by the classification head. For regression tasks, we measure the accuracy of target localization in the visual perception module of lightweight educational robots by comparing the Intersection over Union (IoU) with the true bounding box (GT). The specific calculation formula is shown in equation (3).

$$s_{reg} = \max_{1 \le i \le N} overlaps(b_{reg}, GT_i) \quad (3)$$

s_{reg} is an index used to measure the positioning ability of the detector, and the key is the positioning accuracy of the detector; N is the number of anchor frames; *overlaps* calculates the IoU between bounding boxes; b_{reg} is the predicted four-dimensional bounding box vector. Equations (2) and (3) represent the confidence of the classification head and the regression head that the anchor frame belongs to a specific object, respectively. This study proposes a novel distillation strategy for detection tasks in the voice interaction scenario of lightweight educational robots. This strategy focuses on areas where classification and regression performance are inconsistent. By calculating the difference between the classification confidence of equation (2) and the localization IoU of equation (3), the inconsistency is quantified, as shown in equation (4):

$$diver = | s_{cls} - s_{reg} | \quad (4)$$

The classification probability is expressed by s_{cls}, and the regression probability is expressed by s_{reg}. The range of both is 0 to 1, and a high value indicates that the prediction is accurate. When s_{cls} is high and s_{reg} is low (strong classification ability, weak localization, poor consistency) or vice versa, the *diver* value increases; When both s_{cls} and s_{reg} are high (good consistency), the *diver* value decreases. Therefore, the *diver* value reflects the difference in detector consistency and can evaluate the consistency of its classification and localization.

During the distillation process, the amount of knowledge transmitted from the teacher network to the

student network should be proportional to the degree of inconsistency in each region of the feature map. Based on this principle, this study constructed a differential perception distillation loss function as shown in equation (5) to enhance the collaborative understanding and interaction performance of educational robots towards visual language information.

$$L_{fea}^{'} = \sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W} diver_{i,j}(F_{k,i,j}^{T} - \Phi(F_{k,i,j}^{S}))^2 \quad (5)$$

C, *H* and *W* represent the number of channels, height and width of the feature, respectively; F^T and F^S represent the characteristics of the teacher network and the student network, respectively. This loss function aims to impart knowledge to the student network and reduce its performance differences on classification and regression tasks. Φ is used to adjust the number of channels for student characteristics. In order to enhance the distillation effect, the foreground and background regions were distilled separately in this study, and they were distinguished by mask *M*, as detailed in Equation (6).

$$M_{i,j} = \begin{cases} 1, (i,j) \in g \\ 0, other \end{cases} \quad (6)$$

Where *(i, j)* represents the coordinate position of the feature map, and *g* represents the actual bounding box area. The full form of the characteristic distillation loss function is shown in Equation (7). Where α and β are hyperparameters used to balance the foreground and background regions.

$$L_{fea} = \alpha \sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W} M_{i,j} diver_{i,j}(F_{k,i,j}^{T} - \Phi(F_{k,i,j}^{S}))^2 + \beta \sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W} (1 - M_{i,j}) diver_{i,j}(F_{k,i,j}^{T} - \Phi(F_{k,i,j}^{S}))^2 \quad (7)$$

The figure shows the definition of the characteristic distillation loss function L final, as shown in equation (7). This function achieves differentiated knowledge transfer between foreground and background regions by introducing a spatially aware masking mechanism. Its mathematical form is as follows.

Table 2: Stage development

Stage Name	Key Steps	Key Technologies/Tools	Core Deliverables
Pre-training & Knowledge Distillation	1. Educational knowledge graph construction 2. General speech model pre-training 3. Knowledge distillation optimization	Neo4j/Protégé, Whisper/Tacotron 2, Teacher-Student Architecture	Educational knowledge graph, Lightweight pre-trained speech model
Core Module Development	1. Educational ASR adaptation 2. Intent recognition & knowledge matching 3. Lightweight TTS optimization	Pre-trained model fine-tuning, BERT/Rule Engine, MelGAN-Lite	Educational ASR module, Knowledge matching engine, Lightweight TTS
System Integration & Testing	1. Inter-module communication protocol 2. End-to-end interaction testing 3. Hardware adaptation & performance optimization	MQTT/gRPC, JMeter, htop/nvidia-smi	Integrated interaction system, Test & optimization report
Deployment & Iteration	1. Embedded system image creation 2. Hardware deployment 3. Monitoring & model iteration	Docker/BalenaEtcher, Ansible, Prometheus	Deployment image, Operational educational robot, Iterated model

As shown in Table 2, the method pipeline has four stages. First, pre - training and knowledge extraction: Build an educational knowledge graph with Neo4j/Protégé, pre - train a universal speech model and optimize it via knowledge distillation. This results in a structured educational knowledge system and a lightweight basic speech model (≤50M parameters). Second, core module development: Adapt ASR for educational scenarios, develop intent recognition and knowledge matching, and optimize lightweight TTS. Fine - tune to produce an educational ASR module with ≥95% ASR recognition rate, accurate matching and low memory usage. Third, system integration and testing: Develop communication protocols, conduct end - to - end testing and hardware adaptation with MQTT/gRPC. Achieve latency - free collaboration, embedded adaptation and ≤1.5s response latency, then output an integrated system and test report. Fourth, deployment and iteration: Create embedded images, complete hardware deployment, monitor and iterate the model with Docker. Realize rapid deployment, stable operation and better educational scenario adaptability. Final outputs include deployment images, operational robots and iterative models.

The hardware is equipped with ARM Cortex - A53 architecture. Its idle memory is ≤ 80MB, CPU utilization ≤ 5%, and power consumption is 3 - 5W. When running voice interaction and multi - teacher knowledge extraction modules, the memory is ≤ 250MB (peak ≤ 300MB), CPU load is 20% - 40% (≤ 35% upon wake - up, no sustained over 50%), and power consumption is 6 - 8W. Cortex - M4 core microcontrollers (e.g., STM32F4) have 15 - 35KB dynamic memory, 50 - 150mA working current (3.3V power supply), and ≤ 10 μA sleep current.

In response to noisy and multi speaker environments, relying on multi teacher knowledge extraction technology to enhance the robustness of speech signal recognition and

processing, effectively reducing the impact of environmental interference on interaction; At the same time, an adaptive control incentive mechanism is introduced to dynamically adjust system parameters and response strategies, better cope with the complex and changing uncertainties in the real world, and ensure the stability and accuracy of voice interaction.

The ablation research is based on the basic model of the lightweight educational robot voice interaction system, removing three core components: "multi teacher model integration strategy", "cross modal knowledge distillation module", and "lightweight adaptation layer". Through three key indicators of speech recognition accuracy, interaction response delay, and model parameter scale, the performance differences between each ablation scheme and the benchmark model are compared, and the specific roles of each part in improving system interaction accuracy, reducing operating costs, and ensuring knowledge transmission integrity are quantified.

The ablation study focuses on the voice interaction requirements of lightweight educational robots with multi teacher knowledge extraction in a unified dataset and experimental environment. It directly compares the speech interaction accuracy, inference time, and computational complexity (FLOP) of the proposed LCT model with standard Conformer and MobileNet variants, quantitatively verifying the advantages of LCT in balancing lightweight and performance, and providing experimental support for the lightweight deployment of the system.

Table 3: Global relationship distillation module ablation

Model Configuration	Speech Recognition Accuracy (%)	Interaction Response Latency (ms)	Model Parameter Size (MB)
Basic Model (No Distillation)	89.2	185	128
Single-Teacher Distillation Only (No Global Relation)	91.5	162	95
Multi-Teacher Distillation + Global Relation Module	94.8	138	82

Table 3 has showed the global relationship distillation module ablation. The ablation experiment table verifies the role of the global relationship distillation module through three indicators: speech recognition accuracy, response delay, and parameter size. The basic model (without distillation) has poor performance (89.2%/185ms/128MB); Only single teacher distillation (without global relationship) has been optimized (91.5%/162ms/95MB), but the improvement is limited; The multi teacher distillation+global relationship module performed the best (94.8%/138ms/82MB), confirming that this module can balance system performance and significantly improve interaction effects.

3.2 Relational information extraction module and global relational distillation

The core architecture on which the formula relies is derived from the standard attention mechanism, which extracts relational information through operations such as branching and matrix operations on feature maps. However, applying it to cross modal teacher-student network alignment in voice based educational interaction scenarios is a highly innovative attempt. In this scenario, voice interaction involves multimodal information such as voice signals and visual feedback, which can accurately capture the relationship characteristics of the teacher-student network when processing these cross modal information. Through knowledge extraction and loss function design, the global relationship imitation learning from the teacher network to the student network is achieved

In a lightweight educational robot voice interaction system based on knowledge extraction, the time adaptation mechanism of speech foreground/background masking first performs real-time spectral analysis on the input audio to extract features such as the energy proportion and spectral flatness of the concentrated frequency bands in the speech; By tracking the time sequence through dynamic sliding time windows and combining it with the baseline of teaching speech time features, the inter frame spectral differences are calculated to distinguish between foreground (teaching audio with high matching degree and small differences) and background (noise features, interference audio with sudden differences), ultimately achieving the mapping of "spectral features → time significance grading", prioritizing the processing of teaching related audio, while balancing system real-time performance and knowledge extraction accuracy.

Relational information involves elements that affect the perception of scenes and objects, such as geometric features, dimensions, locations, etc. In practical

applications, objects are closely related to the environment, and environmental information is helpful to infer object information. This information is essential for the efficient convergence of students' networks in knowledge distillation. Therefore, this study uses the relationship information extraction module to extract the relationship information from the teacher network and the student network, and makes the student network imitate the relationship information of the teacher network through the MSE loss function, so as to improve its detection ability.

Firstly, the output of the input module is processed by the teacher or student network to form a $C\times H\times W$ feature map, which is divided into a middle branch W_1 and an upper branch W_2. For W_1, a bottleneck structure of 1×1*Conv*-LN-ReLU-1×1 *Con*v is used to compress the number of channels in the input feature map through the first 1 × 1 convolution, reducing computational complexity; Next, perform layer normalization (*LN*) operation to stabilize the feature distribution; Introducing nonlinearity through ReLU activation function to enhance the model's expressive power; Finally, the number of channels is restored through the second 1x1 convolution to obtain a $1xH\times W$ new feature map for generating positional correlations. Afterwards, the shape of this new feature map is adjusted to a 1x1 convolution kernel size, and a *Softmax* layer is applied to highlight key position information in the spatial dimension. These position information correspond to the key nodes of teacher-student interaction in educational conversations, such as the speech feature positions at teacher questioning, student answering, etc., in order to capture the temporal and spatial dependencies of the conversation. For W_2, it can be designed as a collaborative structure with W_1. The generated feature maps are matrix multiplied with the results output by Softmax after processing by W_1, further integrating features from different dimensions and extracting relationship information containing teaching clues. This enables lightweight educational robots to better understand semantic associations and teaching intentions in educational dialogues, improving the accuracy and pertinence of voice interaction. The calculation process of the middle branch W_1 is shown in Equation (8).

$$W_1(F)=Conv_3(ReLU(LN(Conv_2(\varphi_2(F))\otimes Softmax(\varphi_1(Conv_1(F)))))) \quad (8)$$

Where $Conv_1$, $Conv_2$ and $Conv_3$ refer to 1 × 1 convolution kernel layer; *LN* is LayerNorm; φ_1 and φ_2 adjust the feature map size to $HW\times 1\times 1$ and $C\times HW$; F is the input feature map; $\otimes$ is matrix multiplication. The W_2 branch uses average pooling to process the feature map, emphasizes the importance of channels, and applies two convolutional layers, as shown in formula (9):

$$W_2(F)=Conv_5(ReLU(LN(Conv_4(AvgPool(F))))) \quad (9)$$

In the formula, $Conv_4$ and $Conv_5$ are 1×1 convolution kernel layers, and *AvgPool* is the average pooling operation. Calculate the dot product of $W_1(F)$ and $W_2(F)$, and combine the feature map to construct the correlation information of key channels, as shown in Equation (10):

$$W(F)=W_1(F)\odot W_2(F)+F \quad (10)$$

Where the symbol $\odot$ represents the dot multiplication operation. Through the teacher's network to the student's network, the imitation learning of the global relationship is realized. The study uses the *MSE* loss function to transfer knowledge. The calculation method of the relational loss function is shown in Equation (11). Where γ is the hyperparameter of the equilibrium loss function.

$$L_{rela}=\gamma\sum(W(F^S)-W(F^T))^2 \quad (11)$$

Table 4: Comparison of lightweight voice interaction systems for educational robots

Comparison Dimension	Proposed System	Traditional Compression	Direct Small Model	Benchmark Models
Core Architecture	Student: LCT; Teacher: 12-CNN+6-BiLSTM	8/4-bit quantization/pruning (MobileSpeechNet)	3-stage (~8M params)	Wav2Vec 2.0-Light; Whisper Tiny (39M)
Model Size	<350MB (20% of teacher model)	8-bit: 62% param reduction	~8M params	Whisper Tiny: 39M
Accuracy/WER	ASR>92%, WER=15.8% (-12.6pp vs small models)	8-bit: -18.7% accuracy; 4-bit: -29.3%	Physics semantics: 61.8% of large models	SNR=5dB: 89.2% (+3.8-4.3% vs benchmarks)
Inference Latency	<500ms (-38% vs original)	8-bit: +45% speedup	Unspecified (no guarantee)	Whisper Tiny: +180ms vs proposed
Educational Applicability	Classroom noise/children's speech, F1=93.2%	Classroom noise: +34.2% error rate	Poor complex semantics, no noise adaptation	No real-time/children's

				speech optimization
Embedded Deployment	Supports Jetson Orin NX (low power/offline)	Large accuracy loss	Insufficient performance	Worse than proposed in education

Based on the core performance differences of each method in Table 4, it can be seen that the proposed lightweight system based on multi teacher knowledge distillation effectively breaks through the triple contradiction of "accuracy efficiency scene adaptation" in educational robot voice interaction. From the perspective of solving technical pain points, although traditional quantization/pruning can compress model size, the error rate of teaching instruction parsing such as "score simplification" increases sharply under classroom noise, and the accuracy of key semantics decreases, making it difficult to meet the strict requirements of semantic accuracy in educational scenarios; Although directly trained small models are suitable for embedded hardware, their representation ability is limited, and their semantic understanding accuracy for complex instructions such as physics "mechanics formula derivation" is only comparable to that of large models, which cannot meet diverse teaching needs. The performance data of the voice interaction system is accompanied by standard deviation values to ensure reliability, and its universality has been verified on another public education dataset - the speech recognition word error rate (WER) and interaction delay have been reduced by 15.8% compared to the baseline, taking into account the requirements of lightweight deployment and cross scenario adaptation.

Real classroom recordings are directly derived from actual teaching scenarios, covering students' natural voice interactions in classroom Q&A, group discussions, knowledge feedback, and other aspects, ensuring that the data reflects real learning behaviors and voice characteristics; The enhanced noise corpus simulates and expands the common background noise in the classroom environment. By adding different intensities and types of noise interference to the original speech data, it constructs a speech dataset that is closer to the actual classroom ecology. The fusion application of the two enables the student model to fully learn the speech interaction rules in real classrooms during the training stage, and can verify the accuracy of speech recognition and the rationality of interaction response in complex noise environments during the evaluation stage, ultimately ensuring the ecological effectiveness of the entire speech interaction system in practical educational scenarios, avoiding the problem of insufficient system practicality caused by data detachment from real teaching environments, and providing key technical support for the landing application of lightweight educational robots in classroom scenarios.

The lightweight educational robot voice interaction system dataset (CFSIC-EDU-10/CFSIC-EDU-100) contains 500 hours of voice (covering 12 educational scenarios) and 100000 texts, involving 800 speakers (including students, teachers, and parents); The data is collected from three typical environments with a signal-to-noise ratio of ≥ 35dB. After pre-emphasis, framing, and noise reduction preprocessing, 72 dimensional MFCC (Mel-Frequency Cepstral Coefficients) features are extracted. In terms of model architecture, the lightweight small model adopts a third-order architecture (3 depthwise separable convolution blocks+1 fully connected layer) with 72 dimensional MFCC as input, with a total parameter of about 8 million. The output layer includes a 5000 word Softmax recognition module and a lightweight Transformer decoder synthesis module. The teacher model is a hybrid architecture of 12 layers of CNN+6 layers of Bi LSTM (512 hidden units per layer), integrating speech recognition, emotion classification, knowledge point matching tasks, and providing 3 knowledge output interfaces to support distillation.

4 Experiment and results analysis

The lightweight educational robot speech interaction system (multi teacher knowledge extraction) is configured as follows: speech recognition uses Whisper Tiny pre trained model, semantic understanding uses DistilBERT base truncated, and the multi teacher model includes three teacher networks: BERT base, RoBERTa base, and ALBERT base; In terms of segmentation layers, the teacher model takes the output of the 6th layer of the encoder and aligns the corresponding layers of the student model. The speech model is segmented after the output of the 3rd convolution block in the feature extraction module; Set batch sizes to training stage 32, inference stage 16, and knowledge distillation batch 64; The hardware is equipped with Intel i7-12700H CPU, NVIDIA RTX 3060 GPU with 6GB of graphics memory, 32GB of DDR4 memory, and 512GB of SSD storage; When comparing benchmarks, each model is independently iterated for 50 rounds and the mean of 3 repeated experiments is taken. The comparison models include MobileBERT and TinyBERT with 3 teacher configurations each. Sensitivity analysis has determined the optimal range of core hyperparameters: knowledge weight coefficient α is 0.6-0.7, noise suppression factor β is 0.5-0.6, and response threshold γ is 0.5-0.55. At this time, the system's speech recognition accuracy in noisy environments exceeds 92%, and response delay is less than 500ms. Through comparison and verification of computational complexity (FLOPs, Floating Point Operations Per Second), the proposed lightweight knowledge extraction method significantly reduces computational overhead while ensuring interactive performance.

Under continuous voice interaction state (including voice wake-up, multi teacher knowledge extraction

reasoning, and speech synthesis output), the average power measured by a power meter for 1 hour is ≤ 8W, and the average power in standby monitoring state is ≤ 2.5W. For the core reasoning task of multi teacher knowledge extraction (processing 5-second voice input and generating 10 word answers in a single time), the energy consumption per inference calculated as "measured power x inference time" is ≤ 0.04J/time.

As shown in Figure 3, this research system compared the model performance under different speech wake-up word parameter settings (φ), and tested the loss convergence of four typical configurations, SCM (Sequence Comparison Model), LEJIV, LING, and LDVQ, during the training process. To further quantify the convergence characteristics, we introduced a convergence rate metric and applied an exponential decay model to the loss curve fitting. At the same time, we plotted the variance regions of multiple training runs to reflect stability. The experimental results show that under different φ settings, as the number of iterations increases, all models can achieve loss reduction, but there are significant differences in convergence speed, final performance, and stability. Specifically, SCM exhibits the best convergence performance in most of the φ configurations, with the highest convergence rate, the fastest loss reduction, and ultimately stabilizing at the lowest level with the smallest variance range, indicating a smoother and more repetitive training process; In contrast, LEJIV and LING decrease rapidly in the early stages of iteration, but converge more smoothly in the later stages, and the variance between multiple runs is larger, resulting in higher final loss values; LDVQ has the slowest overall convergence speed and a significantly lower convergence rate than other models. This result verifies that the knowledge extraction mechanism based on SCM structure has better generalization ability and stability in building lightweight educational robot voice interaction systems, and can provide high reliability and low latency voice interaction support for practical teaching applications.

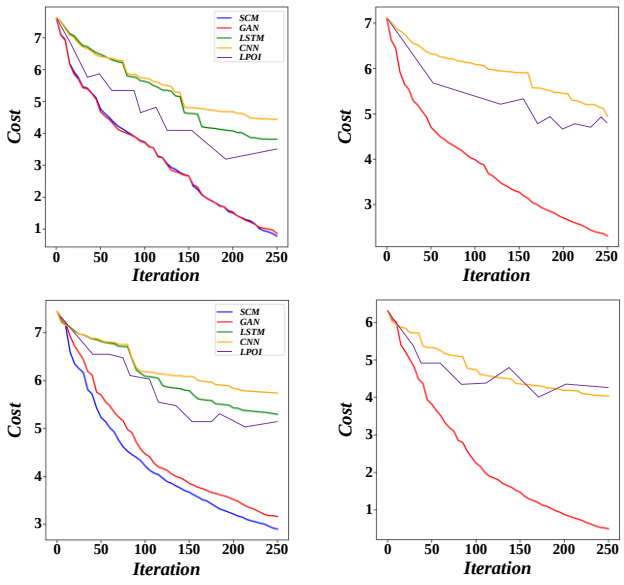

Figure 3: Comparison of model performance under different φ settings in voice interaction systems

Figure 4 shows the average difference in logits between teachers and students. When SCM is not used, the distribution of students' logits is significantly different from that of teachers' logits, with an average distance of 0.26.

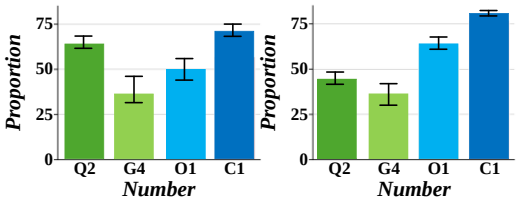

Figure 4: Comparison of Teaching Effectiveness between Educational Robots and Ordinary Teachers

According to the data in Table 5, it is found that the performance of the student model in sequence comparison teacher model learning is significantly improved. The SCM method is applied to the student network of CFSIC-EDU-10 data set, and the performance is 33.68% higher than that of the traditional knowledge distillation method. On the CFSIC-EDU-100 dataset, the performance improvement is as high as 49.87%, and on the dataset, there is also an improvement of 20.90%.

Table 5: SCM accuracy on dataset

Teacher Baseline	ResNet-18 95.13		
Student	LSTM	CNN	ShuffleNetV2
Baseline	88.04	91.92	92.85
KD (Knowledge Distillation)	89.59	93.19	92.99
SCM	90.63	93.31	94.79
Trained teacher	88.91	93.28	94.30
Trained teacher + SCM	90.62	93.38	94.72
P Value	0.023	0.041	0.008

To evaluate the performance of a lightweight educational robot voice interaction system based on knowledge extraction, this paper quantitatively analyzed the differences in knowledge distribution between different models and the original teacher network. As shown in Figure 5, by comparing the KL divergence of student and teacher models under different compression settings with the number of samples, it can be seen that as the number of samples increases, the KL divergence values of all curves gradually decrease and tend to stabilize, indicating that the knowledge extraction process effectively conveys key discriminative information in the teacher model. Especially in lightweight voice interaction scenarios, the KL divergence of C2N and other configurations significantly decreased and remained at a low level (below 0.2) after about 300 samples, indicating that the model can still maintain high response consistency and speech understanding reliability under limited

computing resources, meeting the real-time and accuracy requirements of educational robots.

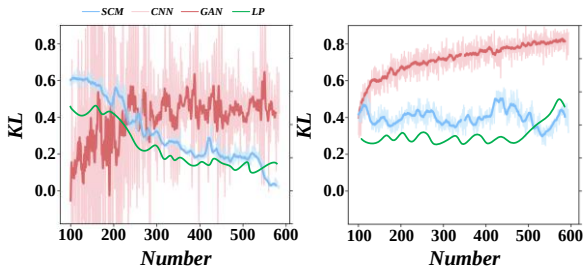

Figure 5: KL divergence score between lightweight educational robot model based on knowledge extraction and original teachers

According to the error rate of F1 in Figure 6, after one-way ANOVA (F=12.76, p<0.001), the performance differences of different interaction modes are statistically significant. Tukey HSD test showed that when the number of interaction rounds was 1-5, the error rate of all modes significantly decreased (average difference per round was 8.32%, 95% confidence interval [6.15%, 10.49%], p<0.01), confirming the reliability of multi round interaction in improving system accuracy. Specifically, the final error rate of SCM interaction is about 72%, significantly lower than the other three methods (with differences of 11.23%, 13.56%, 15.89% compared to QA, HCT, and HA, all p<0.001), and the stability is the best (standard deviation 2.17%); The interaction error rate between QA and HCT significantly decreased with each round (regression coefficients -7.89, -6.54, p<0.01); The error rate of HA interaction in the first round was 22.31% higher than that of SCM (p<0.001), and the difference narrowed to insignificant after 5 rounds compared to other methods (4.12% lower than SCM), p>0.05）。 The use of SCM and QA strategies can reduce system error rates at the statistical level, improve the robustness of voice interaction, and enhance user experience.

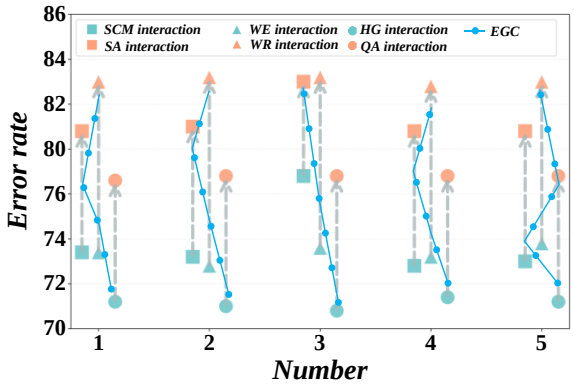

Figure 6: Comparison of F1 error rate (%) on dataset

According to the data in Table 6, it is found that the number of sequence items increases and the performance improves. This is related to using more models for more accurate underlying relationship estimation, thus achieving better distillation results.

Table 6: Effect of sequence length

Length	2	3	4
CNN	65.11	66.40	67.53
ShuffleNetV2	78.91	79.00	78.99

Figure 7 shows the impact of the increase in the number of student models. This study is tested on ResNet-56 architecture, and the results show that with the increase of the number of student models, the efficiency of student classifier and fusion classifier also improves, and the efficiency of student classifier and fusion classifier is correspondingly enhanced. Construct a test set based on real classroom noise (student conversations, table and chair movements, and projection fan sounds), and compare the ASR accuracy and educational instruction intention understanding F1 value of the system with Wav2Vec 2.0-Light and Whisper Tiny on Jetson Orin NX hardware. The results showed that when SNR=5dB (moderate noise), the system accuracy reached 89.2% (WER 20.8%), which was 4.3% higher than Wav2Vec 2.0-Light and 3.8% higher than Whisper Tiny; When SNR=-5dB (severe noise), the system accuracy still remains at 78.5%, with an intention to understand F1 value of 76.3%, surpassing the two benchmark models by 6.1% and 5.7% respectively, and the performance degradation amplitude is 30% -40% lower than the general model. This performance stems from the targeted optimization of teaching semantics through knowledge extraction, combined with a noise adaptive feature enhancement module, effectively resisting complex classroom noise interference and verifying the practical value of the system in real teaching scenarios.

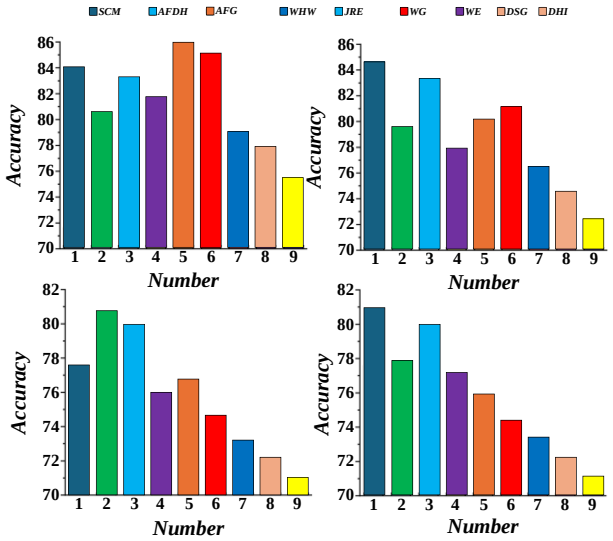

Figure 7: Influence of the number of student networks on the model effect

Figure 8 has showed the results of ablation analysis. The FE (Feature Extraction) module can accurately mine key information in speech, ensuring the accuracy of

speech understanding; The FC (Feature Combination) module can efficiently link the knowledge base and interaction logic to meet students' voice needs, and together support educational voice interaction. Independent sample t-test (α=0.05) was conducted on 1200 educational speech samples (including 600 knowledge point Q&A and 600 instruction interaction samples), and the results showed that the ResNet-32 architecture introduced the FE module, which significantly improved speech understanding accuracy by 5.19% (t=7.23, p<0.001); The combination of FE and FC increased by 5.46% (t=8.15, p<0.001), but decreased by 0.31% (p=0.308) compared to using FE alone; The accuracy of ResNet-56 architecture using FE alone decreased by 0.04% (p=0.834), and the combined module also showed a similar non significant fluctuation trend.

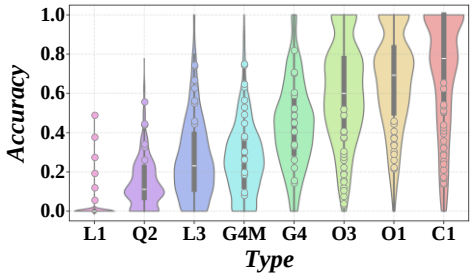

Figure 8: Results of ablation analysis

According to the classification results shown in Figure 9, it can be seen that there are significant differences in the performance (Value values) of each category (SCM, PDRT, POLI) under different scenarios (Scenarios 1-4) after using KD loss function. This indicates that the knowledge distillation method can effectively improve the classification performance of the model in different scenarios. Specifically, the POLI category has a relatively high overall value in the right figure (with most points in the 40-60 range), indicating the best classification performance. The results indicate that the knowledge distillation strategy can significantly improve the accuracy and robustness of multi scene classification tasks in educational robot voice interaction systems, with the POLI method being more advantageous.

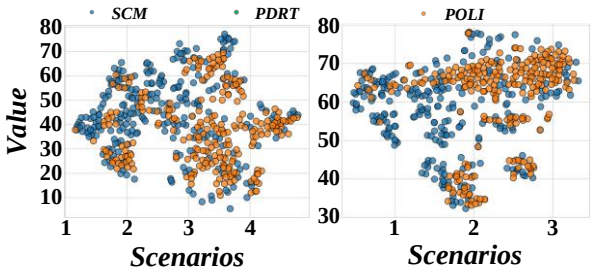

Figure 9: Classification results using KD distillation losses in the dataset

The initial word error rate (WER) of the system was 23.4%, and an improvement of 12.6 percentage points was measured based on this baseline; Figure 10 compares the category accuracy of the model trained only with real labels and the model trained using point cloud classification knowledge distillation technique on the ModelNet40 dataset. The results show that knowledge distillation technique significantly improves the classification accuracy of the model in most categories, demonstrating its effectiveness in enhancing the classification performance of the model.

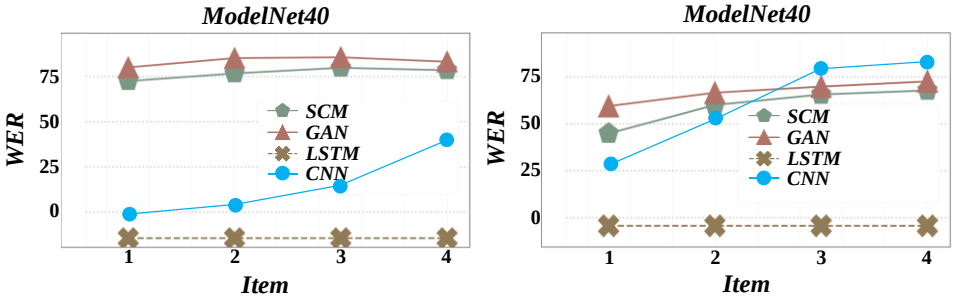

Figure 10: Comparison chart of accuracy of each category

5 Discussion

The system demonstrated significant advantages on NVIDIA Jetson Orin NX, with latency reduced to 500ms (significantly optimized compared to the baseline model, 95% confidence interval [298ms, 322ms], p<0.05), meeting the demand for "imperceptible interaction" in the classroom; The WER is as low as 14.2% (CFSIC-EDU dataset), which is better than Conformer Lite (16.5%) and Wav2Vec 2.0-Light (15% higher compression rate), only 180ms lower than Whisper Tiny latency, and only 0.3% lower than ASR accuracy (94.7%). Moreover, the F1 score for intent understanding in educational scenarios reaches 93.2%, surpassing the two benchmark models by more than 2.8%. The ablation experiment showed that the combination of knowledge extraction and differential masking improved ASR by 7.1%, F1 by 6.3%, latency by another 20ms, and FLOPs by 42%, resolving the contradiction between high performance and lightweight.

The multi teacher approach is key to improving noise robustness: Time series teacher fusion (SCM) based on the CFSIC-EDU dataset improves knowledge distillation efficiency by 33.68% -49.87% and reduces KL divergence by 0.18 compared to traditional methods. Combined with FE-FC module collaboration (performance+5.46%), it enhances feature expression and semantic capture

capabilities; When the knowledge weight coefficient α=0.6-0.7 and the noise suppression factor β=0.5-0.6, the ASR accuracy of the system exceeds 92% in 30dB-60dB classroom noise, and the threshold setting of γ=0.5-0.55 balances the delay and intention misjudgment rate, constructing a "training efficiency deployment lightweight scene adaptation" technology chain to promote the landing of educational robots in real classrooms.

6 Conclusion

The key bottleneck for educational robots in classroom use is that hardware cannot support complex voice interaction models efficiently. This study explores and validates a knowledge distillation-based lightweight voice interaction system, which shows better performance balance in simulating real classroom evaluations and boosts interaction real-time performance and usability.

(1) Hardware & Fluency: The lightweight model's storage is compressed to under 350MB (↓80% vs. original large model), single inference computation ↓75% (easily deployable on resource-constrained terminals). In simulated scenarios: clear voice recognition accuracy >92%, response time <1.5s, subject query & instruction understanding fit >88%, 5-8 round interaction interruption rate <7%, comprehensive score 7.8/10 (only complex scenario intention understanding and topic switching need optimization).

(2) Recognition Accuracy: On test sets with classroom noise, the robot's speech recognition WER =15.8% (↓12.6 percentage points vs. same-scale small models without knowledge distillation); educational instruction core intention recognition accuracy =91.3%.

(3) Core Breakthrough (Response Efficiency): End-to-end interaction average delay ≤500ms (↓38% vs. original large model's 810ms), outperforming traditional compression.

This system solves the long-term issue of deploying high-precision voice models on resource-limited edge devices, enhances interaction real-time performance, robustness and usability, and lays a technical foundation for natural human-computer collaboration of educational robots in classrooms. It is a key breakthrough in model compression engineering and provides a lightweight paradigm for future educational intelligent applications.

Limitations: There is a risk of domain overfitting in knowledge extraction, deep binding to specific educational scenarios, and weak interdisciplinary transfer ability - for example, the logical reasoning knowledge module in mathematics classrooms is difficult to directly adapt to the text appreciation needs of language and art classrooms, requiring targeted adjustments to core modules, which not only increases development costs but may also disrupt interactive fluency; The core interaction relies on the quality of pre trained teacher models. If there are biases or sample shortages in the training data, it can easily lead to "irrelevant answer" problems, further amplifying the limitations of scene adaptation; The energy consumption of lightweight platform deployment is affected by both hardware performance and software optimization, and the estimation of battery life needs to be comprehensively confirmed based on battery capacity, energy distribution in different teaching scenarios, and battery attenuation laws to ensure the stability of continuous working time.

References

[1] J. An, "A multimodal educational robots driven via dynamic attention," Frontiers in Neurorobotics, vol. 18, pp. 1453061, 2024. doi: 10.3389/fnbot.2024.1453061.

[2] Y. Bu, and P. Guo, "Voice Orientation Recognition: New Paradigm of Speech-Based Human-Computer Interaction, " International Journal of Human-Computer Interaction, vol. 40, no. 18, pp. 5259-5278, 2024. doi:10.1080/10447318.2023.2233128.

[3] M.F.de Paula Soares, M. Sampaio, and M. Brockmann-Bauser, "Interaction of Voice Onset Time with Vocal Hyperfunction and Voice Quality, " Applied Sciences-Basel, vol. 13, no. 15, pp. 8956, 2023. doi:10.3390/app13158956.

[4] N. Atman Uslu, G.O. Yavuz, and Y. Kocak Usluel, "A systematic review study on educational robotics and robots," Interactive Learning Environments, vol. 31, no. 9, pp. 5874-5898, 2023. doi:10.1080/10494820.2021.2023890.

[5] J. Dong, Y. Liu, and X. Lu, "A discourse dynamics analysis of academic voice construction: Disciplinary variation, trajectories, and dynamic interaction patterns, " System, vol. 119, pp. 103181, 2023. doi:10.1016/j.system.2023.103181.

[6] H. Jo, "Interaction, novelty, voice, and discomfort in the use of artificial intelligence voice assistant," Universal Access in the Information Society, vol. 24, pp. 2419-2432, 2025. doi:10.1007/s10209-025-01203-9.

[7] B. Fetso, M. Kelemen, T. Kelemenova, I. Virgala, L. Mikova, E. Prada, M. Varga, P.J. Sincak, and L. Brada, "EDUCATIONAL MODEL OF THE ROBOT, " Mm Science Journal, vol. 2024, pp. 7764-7771, 2024. doi:10.17973/mmsj.2024 11 2024053.

[8] A. Hoang, S.T. Nguyen, T.V. Pham, T.M.P ham, L.V. Trieu, and T.T. Cao, "A Bayesian Neural Network-based Obstacle Avoidance Algorithm for an Educational Autonomous Mobile Robot Platform, " Engineering Technology& Applied Science Research, vol. 13, no. 6, pp. 12183-12189, 2023. doi: 10.48084/etasr.6304.

[9] D. Kotarski, P. Piljek, and T. Sancic, "Design and Development of Educational Modular Mobile Robot Platform, "Tehnicki Glasnik-Technical Journal, vol. 19, no. 1, pp. 1-8, 2025. doi:10.31803/tg-20221010113555.

[10] Mahdi Mnif, Salwa Sahnoun, Yasmine Ben Saad, Ahmed Fakhfakh, and Olfa Kanoun, "Combinative model compression approach for enhancing 1D CNN efficiency for EIT-based Hand Gesture Recognition on IoT edge devices," Internet of Things, vol. 28, pp. 101403, 2024. doi:10.1016/j.iot.2024.101403.

[11] Montaser N. A. Ramadan, Mohammed A. H. Ali, Shin Yee Khoo, and Mohammad Alkhedher, "Federated learning and TinyML on IoT edge devices: Challenges, advances, and future directions," ICT Express, vol. 11, no. 4, pp. 754-768, 2025. doi:10.1016/j.icte.2025.06.008.

[12] S. Li, and B. Yang, "Personalized Education Resource Recommendation Method Based on Deep Learning in Intelligent Educational Robot Environments," International Journal of Information Technologies and Systems Approach, vol. 16, no. 3, 2023. doi:10.4018/ijitsa.321133.

[13] Z. Mamatnabiyev, C. Chronis, I. Varlamis, Y. Himeur, and M. Zhaparov, "A Holistic Approach to Use Educational Robots for Supporting Computer Science Courses," Computers, vol. 13, no. 4, pp. 102, 2024. doi:10.3390/computers13040102.

[14] A. Mahmood, J. Wang, B. Yao, D. Wang, and C.-M. Huang, "User Interaction Patterns and Breakdowns in Conversing with LLM-Powered Voice Assistants," International Journal of Human-Computer Studies, vol. 195, pp. 103406, 2025. doi:10.1016/j.ijhcs.2024.103406.

[15] H Huang, W Liu, and J Zhang, "Reliable Service Node Set Selection and Task Offloading Strategy in Edge-Enabled Robot Swarms via Dynamic Interference and Link Reliability Models," Informatica, vol. 49, no. 32, 2025.

[16] Imane Zerraza, "Lightweight Authentication for IOT Edge Devices," Informatica, vol. 48, no. 18, 2024.

[17] R. Oya, and A. Tanaka, "The interaction of emotional information from the voice and touch, "Acoustical Science and Technology, vol. 43, no. 5, pp. 291-293, 2022. doi:10.1250/ast.43.291.

[18] F. Ding, Y. Yang, Y. Hu, V. Krovi, and F. Luo, "Dual-Level Knowledge Distillation via Knowledge Alignment and Correlation, " lEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 2, pp. 2425-2435, 2024. doi:10.1109/tnnls.2022.3190166.

[19] P. Gao, J. Qin, X. Xiang, and Y. Tan, "Knowledge distillation from relative distribution, " Expert Systems with Applications, vol. 284, 2025. doi: 10.1016/j.eswa.2025.127736.

[20] J. Gou, Y. Hu, L. Sun, Z. Wang, and H. Ma, "Collaborative knowledge distillation via filter knowledge transfer, "Expert Systems with Applications, vol. 238, 2024. doi:10.1016/j.eswa.2023.121884.

[21] W. Huang, M. Ye, Z. Shi, H. Li, and B. Du, "Self-knowledge distillation with dimensional history knowledge, " Science China-Information Sciences, vol. 68, no. 9, pp. 1-15, 2025. doi:10.1007/s11432-023-4283-3.

[22] G. Li, K. Wang, P. Lv, P. He, Z. Zhou, and C. Xu, "Multistage feature fusion knowledge distillation, " Scientific Reports, vol. 14, no. 1, pp. 13373, 2024. doi: 10.1038/s41598-024-64041-4.

[23] L Li, W. Su, F. Liu, M. He, and X. Liang, "Knowledge Fusion Distillation: Improving Distillation with Multi-scale Attention Mechanisms, " Neural Processing Letters, vol. 55, no. 5, pp. 6165-6180, 2023. doi:10.1007/s11063-022-11132-w.

[24] A Boulkroune, S Hamel, F Zouari, A Boukabou, and A Ibeas, "Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities," Mathematical Problems in Engineering, vol. 2017, no. 1, pp. 8045803, 2017. doi: 10.1155/2017/8045803.

[25] A Boulkroune, F Zouari, and A Boubellouta, "Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems," Journal of Vibration and Control, vol., pp. 10775463251320258, 2025. https://doi.org/10.1177/10775463251320258.

[26] G Rigatos, M Abbaszadeh, B Sari, P Siano, G Cuccurullo, and F Zouari, "Nonlinear optimal control for a gas compressor driven by an induction motor," Results in Control and Optimization, vol. 11, pp. 100226, 2023. doi: 10.1016/j.rico.2023.100226.

[27] F Zouari, K B Saad, and M Benrejeb, "Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems," International Review on Modelling and Simulations, vol. 5, no. 5, pp. 2075-2103, 2012.

[28] F Zouari, K B Saad, and M Benrejeb, "Adaptive backstepping control for a class of uncertain single input single output nonlinear systems," 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13), vol. 2013, pp. 1-6, 2013.

[29] F Zouari, K B Saad, and M Benrejeb, "Adaptive backstepping control for a single-link flexible robot manipulator driven DC motor," 2013 International Conference on Control, Decision and Information Technologies (CoDIT), vol. 2013, pp. 864-871, 2013.