Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 21 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i21.58851

PAPER

Design and Evaluation of an Augmented Reality and Generative AI-Driven English Speaking Training System

Jingfang Wu[1] (✉), Yueying Shen[2]

[1]Hunan City University, Yiyang, China

[2]Guangzhou Institute of Science and Technology, Guangzhou, China

wujingfang@hncu.edu.cn

ABSTRACT

This paper addresses the demand for English speaking training that integrates augmented reality (AR) with generative artificial intelligence (AI). We design and implement a system that supports highly real-time and robust speech interaction. At its core, the system employs mobile-oriented streaming speech recognition, enhanced with a Conformer encoder using causal convolution, a hybrid Connectionist Temporal Classification (CTC)/Attention decoding mechanism, dynamic block-wise streaming training, dynamic signal-to-noise ratio (SNR) data augmentation, and a multi-speaker speech filtering model. These techniques effectively address challenges in AR environments, such as complex noise, heterogeneous devices, and multi-speaker interference, thereby significantly improving recognition accuracy and response speed. Building on this foundation, we develop an English speaking training application that integrates AR scenarios with a generative AI dialogue engine, offering an immersive and adaptive practice environment. Experimental results demonstrate that the system substantially reduces word error rates in noisy and multi-speaker conditions, achieves latency suitable for real-time interaction, and provides a positive user experience, validating both its technical effectiveness and application feasibility.

KEYWORDS

augmented reality (AR), generative artificial intelligence (AI), streaming speech recognition, English speaking training

1 INTRODUCTION

As augmented reality (AR) technologies [1–3] and generative artificial intelligence (AI) [4] continue to advance rapidly, the creation of immersive and interactive environments for language learning [5, 6] is increasingly becoming a reality. Against this background, constructing an English speaking training system that achieves low latency, high robustness, and strong interactivity has important practical significance for improving learners' oral expression ability and language application literacy. However, existing English learning applications [7, 8] mostly rely on preset

Wu, J., Shen, Y. (2025). Design and Evaluation of an Augmented Reality and Generative AI-Driven English Speaking Training System. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(21), pp. 34–48. https://doi.org/10.3991/ijim.v19i21.58851

Article submitted 2025-06-24. Revision uploaded 2025-09-15. Final acceptance 2025-09-28.

dialogues or fixed voice commands, lacking truly open, dynamic, and multimodal interaction capabilities, and are especially difficult to support the complex and variable acoustic conditions and real-time human-computer dialogue requirements in mobile AR environments [9]. This bottleneck mainly comes from two aspects of challenges: first, mobile speech recognition technology [10, 11] still has insufficient recognition accuracy and real-time response capability in real scenarios such as noise and multi-speaker conditions; second, there is a lack of learning application frameworks that deeply integrate AR context awareness [12] and generative AI dialogue capability [13], which makes it difficult to provide highly realistic and personalized speaking training experiences [14].

Aiming at the above problems, this paper conducts research on the "English speaking training system based on AR and generative AI," focusing on two main parts: first, a mobile streaming speech recognition technology for AR generative AI human-computer interaction is proposed. By introducing a Conformer encoder improved with causal convolution, a two-stage decoding architecture combining Connectionist Temporal Classification (CTC) and Attention, a dynamic block-wise streaming training strategy, dynamic signal-to-noise ratio noise augmentation, and a lightweight multi-speaker speech filtering model, the system solves recognition difficulties such as diverse noise, device heterogeneity, and multi-speaker overlap in mobile AR environments, significantly improving recognition robustness while ensuring low latency. Second, based on the above speech technologies, an English speaking training application is designed and implemented that integrates AR visual scenarios with a generative AI dialogue engine, providing learners with an immersive, adaptive, and feedback-rich speaking practice environment through the combination of virtual and real learning contexts and intelligent dialogue agents. This study not only provides a systematic technical solution for streaming speech recognition in complex mobile environments, but also offers important theoretical references and practical paradigms for constructing the next generation of intelligent language learning systems, with strong technical innovation value and application prospects.

2 STREAMING SPEECH RECOGNITION FOR AR AND GENERATIVE AI INTERACTION

2.1 Speech recognition model

To meet the requirements of AR and generative AI human-computer interaction for low-latency and high-accuracy speech recognition, this paper selects the Conformer model [15] as the basic architecture of the audio encoder. This model combines the advantages of Transformer in capturing long-term dependencies and convolutional neural network (CNN) in extracting local features, and is particularly suitable for processing continuous and variable speech input in AR environments. However, the original Conformer, due to its deep separable convolution structure depending on right-side audio context, leads to an increase of receptive field with the number of layers, which not only introduces unacceptable response latency but also destroys the control of streaming training by the attention mask mechanism, making it difficult to be directly applied in real-time interaction scenarios.

For this reason, this paper introduces causal convolution [16] to transform the Conformer into streaming. Causal convolution ensures that the output at each time step depends only on the current and historical input by truncating the right-side context, thus strictly satisfying the causality constraint of streaming processing.

In the specific implementation, the original depth-wise separable convolution in Conformer is replaced by causal convolution, and only the left-side padding strategy is used to adjust the input. In this way, without changing the model structure, the inter-layer right-side information dependence is eliminated. Taking a convolution layer with kernel size 3 and stride 1 as an example, traditional convolution requires padding 1 time step on each side of the input to achieve "Same Padding," and at this time the receptive field of moment t_3 covers [t_2, t_3, t_4]; while causal convolution only pads 2 time steps on the left side, so that the output at moment t_3 depends only on [t_1, t_2, t_3], completely meeting the requirements of real-time processing. Assuming that the kernel size is represented by *KS*, the amount of padding required for standard convolution and causal convolution to achieve "Same Padding" can be calculated by the following formula:

$$PAD_{CNN} = \frac{KS-1}{2}, PAD_{CA-CNN} = KS-1 \tag{1}$$

Through the above improvement, the cumulative latency caused by stacking layers is effectively eliminated, and the attention module is also guaranteed to rely on the mask to achieve precise streaming control, providing a stable and efficient encoding foundation for real-time speech interaction and feedback of generative AI in AR environments.

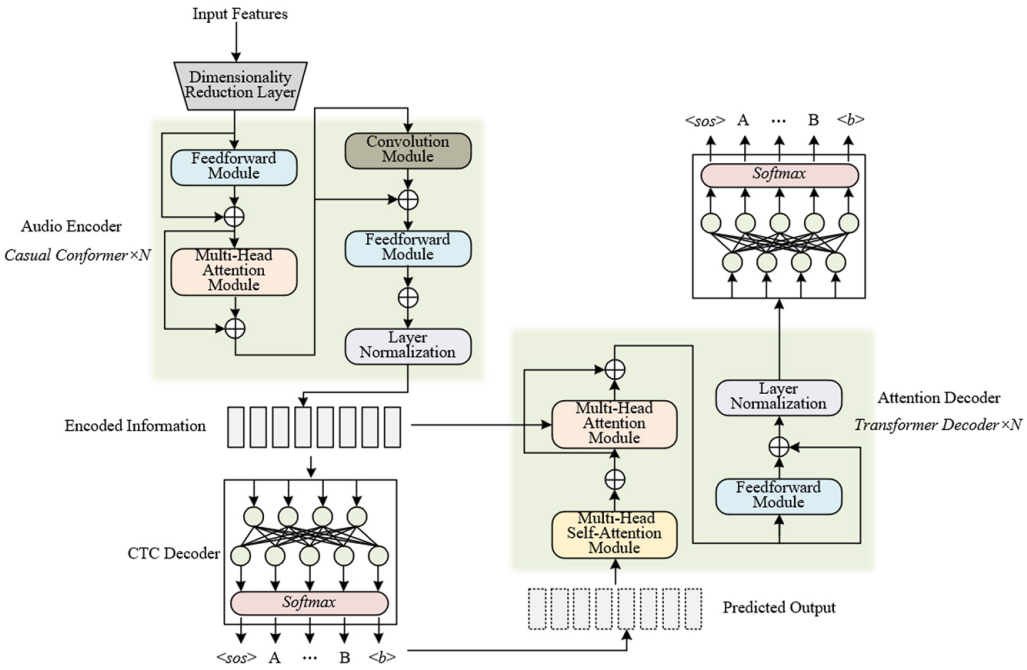

Fig. 1. Mobile streaming speech recognition model for AR and generative AI human-computer interaction

Facing the dual requirements of high real-time performance and high accuracy in AR and generative AI human-computer interaction, this paper proposes a two-stage streaming speech recognition architecture based on CTC and Transformer, aiming to achieve low-latency response on mobile terminals and recognition performance close to non-streaming systems. In the first stage, pure streaming CTC decoding is adopted to ensure real-time performance, and in the second stage, a Transformer rescoring mechanism based on complete audio context is introduced to significantly improve

recognition accuracy, thereby effectively solving the contradiction between insufficient recognition rate of traditional streaming models and the inability of attention models to perform streaming processing. The model architecture is shown in Figure 1.

The first-stage model consists of a dimensionality reduction layer, a causal Conformer encoder, and a CTC prediction layer. The dimensionality reduction layer compresses the input audio features from 320 dimensions to 256 dimensions through a fully connected network, reducing the subsequent computational load, and embeds relative position encoding to enhance temporal modeling capability and state reusability, supporting efficient use of historical information in streaming inference. The audio encoder adopts 12 layers of low-dimensional, deep-structured causal Conformer modules, which integrate local and global acoustic representations under strict causal constraints, balancing model representation ability and mobile terminal computing limitations. The CTC prediction layer maps the encoded features to the vocabulary space and outputs the character probability distribution. The entire first-stage structure meets streaming processing requirements and can generate preliminary recognition results in real time under partial speech input. Specifically, for the input feature sequence $a = (a_1, a_2, \ldots, a_s)$, the processing of the first stage can be expressed by the following formulas:

$$\begin{aligned} a_{\dim\_reduction} &= PositionEncoding(Linear(a)) \\ a_{encoder} &= Conformer(a_{\dim\_reduction}) \\ b_{first-pass} &= Soft\max(Linear(a_{encoder})) \end{aligned} \tag{2}$$

The second stage introduces a Transformer-based decoder as a rescoring module, which re-evaluates multiple candidate paths of the first stage in parallel using complete audio encoding. This module takes the entire audio features as key-value pairs and combines the preliminary recognition results to perform cross-attention calculation, optimizing the output distribution with the help of the complete semantic context required by generative AI, thereby significantly improving recognition accuracy with almost no additional latency, especially suitable for robust understanding of complex commands and multi-turn dialogues in AR environments.

For the dual requirements of recognition accuracy and training stability in AR and generative AI human-computer interaction scenarios, this paper designs a multi-task loss function that integrates CTC [17] and attention mechanism [18] to achieve efficient collaborative training of the streaming speech recognition model. This hybrid loss is composed of CTC loss and KL divergence attention loss with label smoothing weighting, which fully utilizes the strong guidance of CTC sequence alignment and the powerful contextual modeling ability of the attention mechanism, significantly improving the model's generalization performance in complex acoustic environments of mobile terminals and multi-turn interactions. Specifically, the CTC loss explicitly models the monotonic alignment relationship between input and output sequences through the forward-backward algorithm, providing temporal constraints for streaming decoding and ensuring the low-latency characteristics of generated responses; while the KL divergence loss with label smoothing enhances the calibration ability and generalization of the model on the output distribution, effectively alleviating overfitting and improving the modeling ability of diverse output distributions required by generative AI. The two are jointly optimized through a multi-task learning framework, which not only compensates for the shortcomings of pure attention training such as fuzzy alignment, strong data dependence, and slow convergence, but also avoids the problem of overly strong independence assumption in CTC training, enabling the model to have both real-time processing capability

and semantic understanding accuracy in AR interaction scenarios. This loss design significantly accelerates model convergence, improves adaptability to complex AR acoustic environments such as noise and multi-speaker conversations, and provides a reliable optimization basis for real-time generative AI speaking interaction and feedback. Assuming that the real label and predicted label embedding vectors with dimension v are represented by b and $\bar{b}$, the calculation formula of the attention loss function is given as follows:

$$LOSS_{ATT}(\bar{b},b) = KL(\bar{b},b) = \sum_{u=1}^{v} b_u \cdot \log \frac{b_u}{\bar{b}_u} \tag{3}$$

Assuming that the speech features and their corresponding label sequence are represented by a and b, the CTC loss and attention loss are represented by $LOSS_{ZSZ}$ and $LOSS_{ATT}$, and the joint loss is represented by $LOSS_{Total}$, with the weight factor controlling the CTC loss and attention loss represented by η, the expression of the joint loss function is:

$$LOSS_{Total}(a,b) = \eta LOSS_{CTC}(a,b) + (1-\eta) LOSS_{ATT}(a,b) \tag{4}$$

In the mobile speech recognition system for AR and generative AI human-computer interaction, the decoding mechanism needs to balance response speed, accuracy, and resource efficiency. In the decoding process, the proposed model first encodes the speech stream block by block through the feed-forward dimensionality reduction layer and the causal Conformer encoder, and uses the CTC decoder to realize the first-stage streaming decoding, generating candidate results in real time; after the input ends, the attention decoder performs second-stage decoding based on complete encoding information to further improve recognition accuracy. The system supports the following four decoding algorithms to adapt to the trade-off between real-time performance and accuracy in different interaction scenarios:

1. CTC greedy search: selects the output with the highest probability at each time step with extremely low latency, and generates the final result after compression of repetitions and blank labels, suitable for AR real-time command recognition with very high real-time requirements;
2. CTC prefix beam search: retains multiple candidate paths through beam width control, and merges probabilities of the same prefix at each step, which can significantly improve the candidate quality in the streaming decoding stage and provide more reliable preliminary hypotheses for the second stage;
3. Attention beam search: generates outputs in an autoregressive manner in the second stage, retaining the top-k optimal sequences at each step, relying on complete speech context to achieve more accurate semantic modeling, suitable for scenarios such as generative AI dialogue with higher accuracy requirements;
4. Attention rescoring: performs parallel rescoring on the Top-N candidates output by the CTC stage in a teacher-forcing manner, avoiding cumulative errors and latency caused by autoregression, significantly improving recognition rate with almost no additional computational overhead, particularly suitable for real-time interaction requirements in mobile AR environments with diverse noise and limited resources.

The system achieves low-latency and high-accuracy speech recognition in AR interaction scenarios through flexible combination of the above decoding strategies, providing a reliable decoding basis for generative AI speaking training and multimodal interaction.

2.2 Model training and data augmentation

For the heterogeneity of mobile devices and the diversified recognition requirements in AR and generative AI human-computer interaction scenarios, this paper proposes a new dynamic training algorithm, aiming to enable a single model to flexibly adapt to deployment environments with different latency-accuracy trade-offs. The algorithm randomly samples the speech block size within a certain range during the training process, forcing the model to adapt to streaming inputs with different block sizes, significantly improving its generalization ability to variable block decoding. Different from traditional dynamic training, this method abandons non-streaming training samples, and all training is conducted on block-wise speech, so as to strengthen the consistency of model encoding and inference in pure streaming scenarios. In the inference stage, the following strategies are adopted to further improve system efficiency and adaptability:

1. Cache reuse mechanism: The historical state of each layer's encoder output is cached, and after concatenation with new speech blocks, it is used as key-value input, avoiding redundant computation, supporting efficient cross-block state transmission, and significantly reducing the computational load in AR real-time interaction;
2. Block parallel computation: Based on the attention mask of blocks, multiple frames within the same block can be computed in parallel, improving frame computation efficiency without requiring additional future context, overcoming the problem of high latency and slow response in traditional frame-by-frame computation, and meeting the strict requirements of low-latency interaction for generative AI dialogue in augmented reality.

By adopting the above training and inference mechanisms, the model can dynamically adjust block size according to actual device performance and scenario requirements, realizing "one training, diverse deployment" under heterogeneous mobile hardware environments, and providing a highly adaptive, low-latency speech recognition foundation for real-time speech interaction and generative AI feedback in AR environments.

For the high-noise and variability characteristics of mobile speech input in AR and generative AI human-computer interaction, this paper further proposes a dynamic noise data augmentation algorithm to significantly improve the robustness and generalization ability of the streaming speech recognition model in real complex environments. The algorithm dynamically controls the signal-to-noise ratio during the mixing of speech and noise, which not only increases the diversity of training data but also avoids strong noise from destroying key features of speech, thereby better fitting the actual speech interaction in AR environments where background noise is complex and acoustic conditions are variable. Specifically, assuming that clean speech data is represented by $X_{CL}(s)$, and noise audio data is represented by $X_{NO}(s)$, the speech mixing process of traditional static noise data augmentation is represented by the following formula:

$$MIX_{ST}(s) = X_{CL}(s) + X_{NO}(s) \tag{5}$$

Different from traditional methods, the proposed method dynamically samples signal-to-noise ratio values from a preset SNR range for each construction of training samples, and adaptively adjusts the noise coefficient accordingly, ensuring that

the synthesized speech not only contains rich noise types but also maintains speech recognizability. Assuming the noise coefficient is ε, then:

$$MIX_{DY}(s) = X_{CL}(s) + \varepsilon \cdot X_{NO}(s) \tag{6}$$

The formula for calculating the signal-to-noise ratio of the speech signal with noise coefficient ε used in this method is:

$$SNR\left(X_{CL}(s), \varepsilon \cdot X_{NO}(s)\right) = 10 \cdot \log_{10} \frac{\sum_s {X_{CL}}^2(s)}{\sum_s \varepsilon^2 \cdot {X_{NO}}^2(s)} \tag{7}$$

$$\varepsilon = \sqrt{\frac{\sum_s X_{CL}^2(s)}{10^{\frac{tve}{10}} \sum_s X_{CL}^2(s)}} \tag{8}$$

2.3 Speech filtering model

To address the problem of simultaneous speech and speech overlap interference commonly present in AR environments, this paper proposes a lightweight streaming multi-speaker speech filtering model, serving as a front-end module for speech recognition, effectively improving the robustness of speech recognition in complex acoustic scenarios. The model is guided by speaker voiceprint identity information and combines a streaming processing mechanism to achieve real-time extraction and enhancement of the target speaker's voice, significantly reducing the impact of irrelevant voices and environmental noise on recognition results.

The model uses a d-vector voiceprint encoder trained based on GE2E loss to extract a 256-dimensional speaker embedding vector as identity representation, which has good speaker generalization capability. The speech filtering module concatenates the temporally expanded voiceprint features with the mixed speech FBank features to form a 576-dimensional mixed representation, which is input into a streaming LSTM network for encoding. The network outputs a soft mask with the same dimension as the original speech, and the target speaker's voice is enhanced through element-wise multiplication. Assuming the pre-recorded voiceprint speech of the target speaker is represented by a_{SPK}, the target speaker's voiceprint features are represented by a_{EMB}, and a_{MIX} contains the mixed speech features of multiple speakers, the filtered target speaker's speech features are represented by $a_{enhance}$. The computation process of the model is as follows:

$$\begin{aligned} a_{EMB} &= Embeeder(a_{SPK}) \\ a_{IN} &= concat(a_{EMB}, a_{MIX}) \\ a_{enhance} &= a_{MIX} \cdot Soft\max(Linear(LSTM(a_{IN}))) \end{aligned} \tag{9}$$

The model uses mean square error (MSE) as the loss function, and during training, the parameters of the voiceprint encoder are fixed while only the filtering network is optimized, ensuring system stability and modular training efficiency. The specific expression is:

$$LOSS = MSE(a_{CL}, a_{EN}) = \frac{\sum_{d=0}^{D-1} \sum_{s=0}^{S-1} a_{CL}^{(s,d)} - a_{EN}^{(s,d)}}{S \times F} \tag{10}$$

The model has the following prominent advantages: first, it does not rely on a known number of speakers in advance and can dynamically adapt to changes in speakers in AR scenarios; second, it uses FBank features as input, abandoning high-sample-rate raw waveform processing, greatly reducing computational complexity.

3 ENGLISH SPEAKING TRAINING VIA AR AND GENERATIVE AI

Figure 2 shows the overall architecture of the English speaking training system proposed in this paper. The system deeply integrates AR and generative AI technologies and possesses multimodal perception, real-time speech interaction, and intelligent feedback capabilities. The architecture is divided top-down into the AR interaction layer, multimodal input processing layer, and virtual-real fused information synchronization process: the AR scene management module is responsible for constructing and rendering the learning scenarios; the speech enhancement front-end integrates noise suppression and speaker separation functions, effectively improving speech quality in complex acoustic environments; the streaming speech recognition core combined with the generative dialogue engine realizes low-latency speech recognition and context-aware response generation; the pronunciation and semantic evaluation module relies on recognition results and AI feedback to provide multi-dimensional assessment of users' oral performance. The entire system realizes a closed loop from real-time speech processing to immersive AR interaction through module collaboration, providing learners with a highly realistic and strongly adaptive speaking training environment.

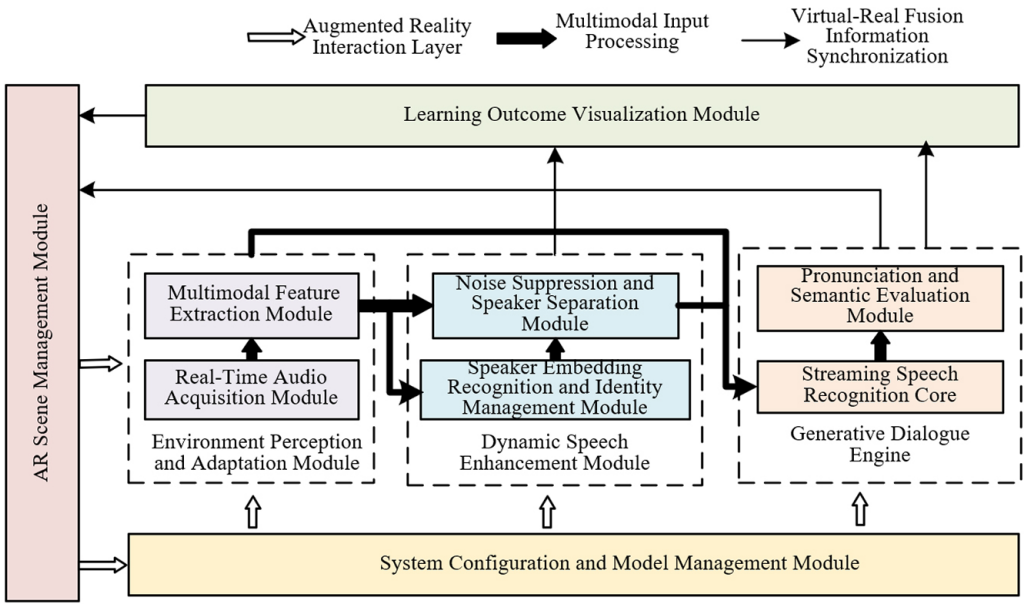

Fig. 2. Architecture of the English speaking training system integrating AR and generative AI

Figure 3 depicts the parallel multi-thread processing pipeline in this English speaking training system, showing the complete data flow and control logic from raw audio input to the final generation of intelligent feedback. The process begins with the multimodal audio acquisition thread, which is responsible for real-time capturing of user and environmental sounds, and transmits the raw audio frames to the multimodal feature extraction thread through a thread-safe data queue. This thread not only extracts acoustic features but also prepares data for subsequent AR

fusion processing. The processed multimodal feature vectors are sent to the next queue, where the real-time speech enhancement and separation module purifies them to improve the signal-to-noise ratio, creating clean input conditions for the core recognition task.

The purified features finally reach the core streaming recognition and AI response thread. This thread first performs low-latency first-stage decoding through the streaming speech recognition core, generating real-time intermediate recognition results and ensuring immediacy of interaction. Subsequently, the system does not immediately output results but enters the semantic optimization and response generation stage. In this stage, complete contextual information is used to rescore and refine intermediate results, driving the generative dialogue engine to produce contextually appropriate and semantically rich dialogue content integrated with AR scene context. Finally, the output is no longer pure text but an evaluation result and AI feedback that integrates accuracy assessment, pronunciation suggestions, and AI dialogue content, forming a complete technical closed loop, efficiently supporting an immersive and interactive speaking training experience.

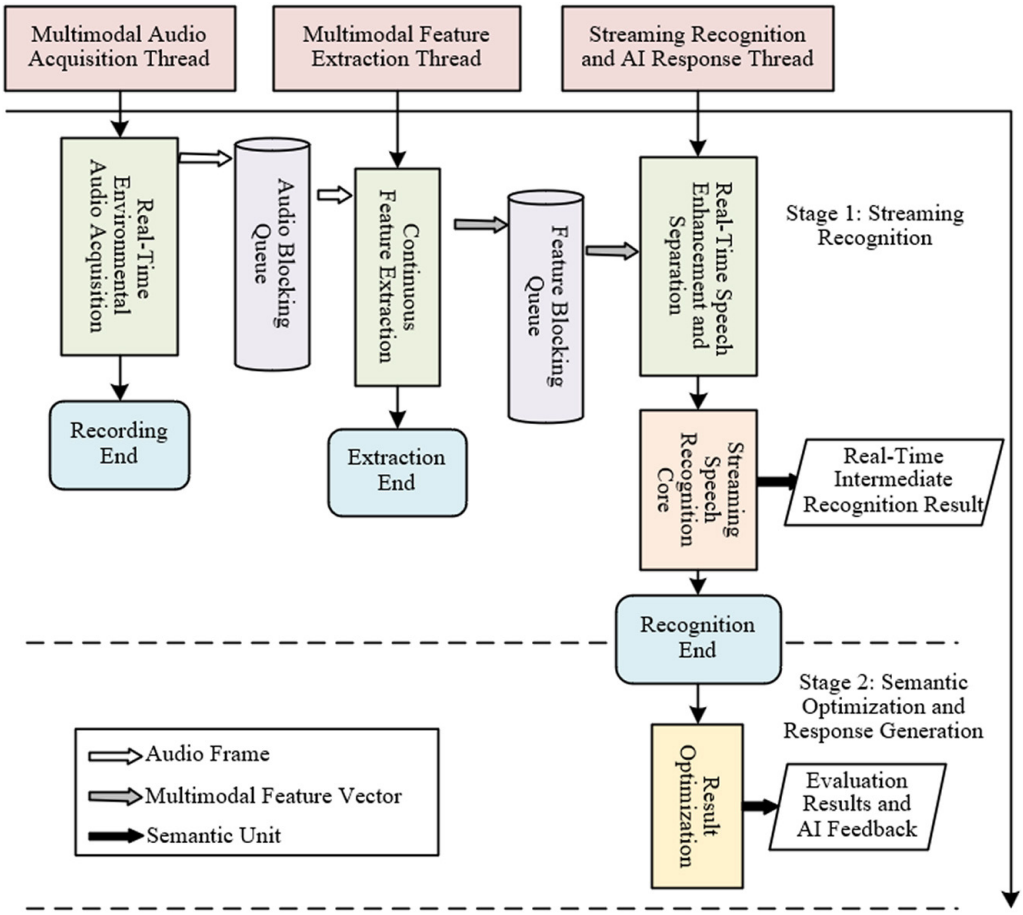

Fig. 3. Parallel multi-thread processing pipeline in the English speaking training system

4 EXPERIMENTAL RESULTS AND ANALYSIS

The experiments first compared the character error rate (CER) of different speech recognition models under clean speech, high-noise environment, and multi-speaker

conversation scenarios. Analysis of Table 1 shows that first, the CausalConformer architecture using causal convolution outperforms traditional Transformer models in all scenarios due to its streaming processing capability; second, the proposed dynamic streaming training strategy effectively improves the model's generalization ability to variable input lengths, further reducing CER of the baseline model; critically, after introducing dynamic SNR noise data augmentation, the model's robustness in high-noise environments is significantly enhanced, with CER reduced by approximately 3% points; finally, the full system optimization scheme integrated with a speech filtering front-end achieves the best performance (CER = 13.26%) in the most challenging multi-speaker conversation scenario, reducing the error rate by nearly 17% points compared to the baseline model without any optimization. These results demonstrate that the streaming architecture, dynamic training strategy, targeted data augmentation, and front-end filtering techniques adopted in this paper jointly form an efficient and reliable solution, significantly improving speech recognition accuracy in mobile AR interaction environments and providing a solid technical foundation for generative AI oral training systems.

Table 1. Performance comparison of streaming speech recognition models for AR interaction environments (CER/%)

Model Architecture	Training and Augmentation Strategy	Decoding Algorithm	Clean Speech	High-Noise Environment	Multi-Speaker Conversation Scenario
Transformer-CTC	Static Training	CTC Greedy Search	8.06	25.15	30.22
Transformer-AED	Static Training	Attention Beam Search	7.10	22.43	28.91
CausalConformer-CTC	Static Training	CTC Prefix Beam Search	7.02	20.58	26.74
Proposed model	Static Training	Attention Rescoring	6.48	18.67	24.05
Proposed model	Dynamic Training	Attention Rescoring	5.95	15.32	19.88
Proposed model	Dynamic Training + Data Augmentation	Attention Rescoring	5.90	12.07	16.54
Proposed model	Full System Optimization (Dynamic Training + Data Augmentation + Speech Filtering)	Attention Rescoring	5.85	9.83	13.26

To evaluate the practical performance of the proposed streaming speech recognition model in AR interaction environments, this experiment compared the character error rate and latency of different training and decoding strategies in various typical AR scenarios. The results in Table 2 indicate that first, the dynamic training strategy significantly reduces CER in all scenarios compared with static training, especially improving the multi-speaker conversation scenario most noticeably, verifying the generalization capability of dynamic training in complex acoustic environments; second, the proposed "dynamic training + data augmentation" algorithm further reduces CER while maintaining similar latency, particularly reducing CER by about 2.3% in high-noise environments compared with dynamic training alone, proving the effectiveness of enhanced streaming modeling; finally, the scheme integrating data augmentation achieves the best performance, with CER as low as 14.92% in the most challenging multi-speaker scenario, and overall latency meets real-time interaction requirements (<150 ms). These results demonstrate that the dynamic training and targeted augmentation strategies adopted in this paper can effectively balance recognition accuracy and response real-time performance, providing a reliable and efficient speech recognition solution for generative AI oral training systems in AR environments.

Table 2. Performance comparison of streaming recognition under different training and decoding strategies in AR interaction scenarios (CER/%)

Training Strategy	Model Configuration	Attention Rescoring (Decoding Algorithm)	Clean Environment	High-Noise Environment	Multi-Speaker Conversation Scenario	Average Latency (ms)
Static Training	*CausalConformer-CTC*	No	13.31	22.15	28.44	120
	+ *Attention Decoder*	Yes	11.02	18.92	24.30	180
Dynamic Training	Proposed model	Yes	7.07	15.54	20.61	150
Dynamic Training + Data Augmentation	Proposed model	Yes	6.95	13.27	17.85	140
Dynamic Training + Data Augmentation	Proposed model	Yes	6.82	11.03	14.92	145

To verify the effectiveness of the proposed data augmentation algorithm in improving speech recognition robustness under complex AR acoustic environments, this experiment compared the performance of different training strategies in clean and simulated AR noisy environments. Results in Table 3 indicate that the baseline model without data augmentation exhibits a sharp CER increase to ~25% in noisy environments, significantly reducing usability. With traditional static noise augmentation, the model's noise resistance is significantly improved, reducing CER by over 12% points in noisy environments, demonstrating the necessity of data augmentation. The proposed data augmentation algorithm further optimizes performance, achieving the lowest CER in AR noisy environments (11.78% when Chunk = 30), showing stable performance across different streaming chunk sizes, while causing no performance degradation in clean environments. These results demonstrate that the proposed algorithm effectively improves model generalization to typical AR noise patterns, providing key technical support for constructing highly robust real-time speech interaction systems.

Furthermore, experiments were designed to verify the effect of the speech filtering front-end on improving speech recognition accuracy in complex AR acoustic environments. Analysis of Table 4 indicates that in clean environments, integrating the speech filtering front-end maintains the original system performance (CER ~7.5%), showing that this module does not introduce additional distortion; in AR noisy environments, CER is significantly reduced by approximately 2.3–2.5% points, demonstrating its effectiveness in suppressing environmental noise and reverberation. Most notably, in the most challenging multi-speaker conversation scenario, speech filtering leads to a revolutionary performance improvement, reducing CER from approximately 60% to around 18%, a decrease of over 40 percentage points, completely resolving recognition failures caused by overlapping speech. These results demonstrate that the proposed speech filtering model, as a front-end module for speech recognition, greatly enhances system robustness and practicality in real AR environments, providing critical technical support for generative AI oral interaction in multi-speaker scenarios.

Finally, this paper evaluated the optimization effect of dynamic quantization on model deployment to verify its ability to maintain model performance under mobile resource constraints (refer to Table 5). Results indicate that INT8 quantization reduces model size by approximately 66% and inference latency by about 30%, significantly improving deployment efficiency. In terms of recognition performance, the quantized model shows only slight increases in CER across all test environments

(Clean Environment +0.2–0.34%, Noisy Environment +0.33–0.42%, Multi-Speaker Scenario +0.71–1.28%), with performance loss controlled within an acceptable range. Moreover, the absolute CER in noisy environments remains far below the baseline system without the speech filtering model. These results demonstrate that dynamic quantization can effectively achieve model compression and acceleration, providing key technical support for real-time deployment of speech recognition systems on mobile AR devices, successfully balancing performance and efficiency.

Table 3. Robustness verification of different data augmentation algorithms in AR noise environments (CER/%)

Training Strategy	Test Environment	*Chunk* = 30	*Chunk* = 16	*Chunk* = 8
No Augmentation	Clean Environment	7.85	8.14	8.50
	AR Noisy Environment (High Noise + Reverberation)	24.40	25.49	26.54
Traditional Static Noise Augmentation	Clean Environment	7.36	7.57	7.85
	AR Noisy Environment (High Noise + Reverberation)	12.37	12.82	13.43
Proposed Augmentation Algorithm	Clean Environment	7.37	7.54	7.85
	AR Noisy Environment (High Noise + Reverberation)	11.78	12.19	12.85

Table 4. Impact of speech filtering model on CER across different datasets (%)

System Configuration	Test Environment	*Chunk* = 30	*Chunk* = 16	*Chunk* = 8
Baseline	Clean Environment	7.37	7.54	7.85
	AR Noisy Environment (High Noise + Reverberation)	11.78	12.19	12.85
	Multi-Speaker Conversation Scenario (Speech Overlap)	59.61	60.52	61.38
Integrated Speech Filtering Front-End	Clean Environment	7.35	7.52	7.83
	AR Noisy Environment (High Noise + Reverberation)	9.42	9.87	10.35
	Multi-Speaker Conversation Scenario (Speech Overlap)	17.90	18.22	18.79

Table 5. Performance-efficiency trade-off analysis of dynamic quantization for mobile model deployment (CER/%)

System Configuration	Model Size (Compression Rate)	Clean Environment (*Chunk* = 16)	AR Noisy Environment (*Chunk* = 16)	Multi-Speaker Conversation Scenario (*Chunk* = 16)	Inference Latency (*ms*)
*FP*32	200.9 *MB* (/)	7.54	12.19	60.52	142
+ *Quantization (INT*8)	69.1 *MB* (65.6%)	7.74 (+ 0.20)	12.52 (+ 0.33)	61.80 (+ 1.28)	98
Integrated Speech Filtering Front-End (*FP*32)	233.2 *MB* (/)	7.52	9.87	18.22	155

5 CONCLUSION

This paper focused on the "AR- and Generative AI-based English Oral Training System," with key breakthroughs in efficient and robust streaming speech recognition under complex mobile acoustic environments. In terms of model structure, the Conformer encoder was improved with causal convolution and combined with a CTC/Attention hybrid mechanism, constructing a low-latency, high-accuracy two-stage recognition architecture. In terms of training methodology, a dynamic training strategy was proposed to enhance the model's generalization ability to different streaming chunk sizes; a dynamic augmentation algorithm was designed to significantly improve robustness in noisy environments. For multi-speaker scenarios, a lightweight speaker embedding-based speech filtering front-end was developed to effectively suppress non-target speaker interference. Finally, dynamic quantization technology was applied to achieve significant model compression, supporting efficient deployment on mobile devices. Experimental results show that the proposed methods outperform baseline models in clean, noisy, and multi-speaker scenarios, with CER reduced by more than 40% points at maximum, while inference latency is controlled within 150 ms, meeting the strict requirements of AR real-time interaction.

The main value of this study lies in systematically addressing multiple challenges faced by speech recognition in mobile AR environments, including latency, noise, multi-speaker speech, and resource constraints, providing a complete technical solution and solid theoretical support for constructing practical immersive generative AI oral training systems. However, certain limitations remain, such as discrepancies between simulated AR acoustic environments and real scenarios, performance of the speaker embedding module under highly overlapping speech, and insufficient verification of system generalization in multilingual and multi-dialect scenarios. Future work will focus on the following directions: first, constructing dynamic multimodal datasets closer to real AR environments; second, exploring visual-speech joint modeling methods to further improve speech separation and recognition performance using AR camera visual information.

6 REFERENCES

[1] N. Terentieva, V. Karpenko, N. Yarova, N. Shkvyria, and M. Pasko, "Technological innovation in digital brand management: Leveraging artificial intelligence and immersive experiences," *Journal of Research, Innovation and Technologies*, vol. 4, no. 2, pp. 201–223, 2025. https://doi.org/10.57017/jorit.v4.2(8).06

[2] R. Efendi, Ambiyar, Estuhono, and R. A. Wulandari, "Bridging the Industry 4.0 skills gap: An immersive augmented reality mobile learning approach for vocational education," *International Journal of Interactive Mobile Technologies*, vol. 19, no. 6, pp. 60–74, 2025. https://doi.org/10.3991/ijim.v19i06.53825

[3] X. Lyu, S. S. Ramasamy, and F. Ying, "Digital virtual anchors impact in entertainment industry: An exploration of user acceptance and market insights," *Journal of Research, Innovation and Technologies*, vol. 4, no. 2, pp. 125–141, 2025. https://doi.org/10.57017/jorit.v4.2(8).01

[4] N. Immorlica, B. Lucier, and A. Slivkins, "Generative AI as economic agents," *ACM SIGecom Exchanges*, vol. 22, no. 1, pp. 93–109, 2024. https://doi.org/10.1145/3699824.3699832

[5] J. Cardenas-Valdivia, J. Flores-Alvines, O. Iparraguirre-Villanueva, and M. Cabanillas-Carbonell, "Augmented reality for Quechua language teaching-learning: A systematic review," *International Journal of Interactive Mobile Technologies*, vol. 17, no. 6, pp. 116–138, 2023. https://doi.org/10.3991/ijim.v17i06.37793

[6] M. Makoe and T. Shandu, "Developing a mobile app for learning English vocabulary in an open distance learning context," *International Review of Research in Open and Distributed Learning*, vol. 19, no. 4, pp. 208–221, 2018. https://doi.org/10.19173/irrodl.v19i4.3746

[7] F. W. Chen, "Who is teaching English in English language learning applications? Investigating native-speakerism in the mobile learning age," *Changing English*, vol. 32, no. 3, pp. 260–274, 2025. https://doi.org/10.1080/1358684X.2024.2431012

[8] J. Lee, S. Hwang, J. Lee, and S. Kang, "Comparative performance characterization of mobile AR frameworks in the context of AR-based grocery shopping applications," *Applied Sciences*, vol. 10, no. 4, p. 1547, 2020. https://doi.org/10.3390/app10041547

[9] S. Deligne, S. Dharanipragada, R. Gopinath, B. Maison, P. Olsen, and H. Printz, "A robust high accuracy speech recognition system for mobile applications," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 551–561, 2003. https://doi.org/10.1109/TSA.2002.804541

[10] M. J. Kim, S. Y. Suk, H. Y. Jung, and H. Y. Chung, "A speech and character combined recognition engine for mobile devices," in *Embedded and Ubiquitous Computing: International Conference (EUC 2006)*, in Lecture Notes in Computer Science, Seoul, Korea, vol. 4096, 2006, pp. 549–559. https://doi.org/10.1007/11802167_56

[11] M. S. A. Chowdhury, M. F. Khan, M. S. Ali, M. Z. Islam, M. A. Mannan, and M. A. Ullah, "Bangla speech processing: An analytical study of feature extraction and recognition methods," *Mathematical Modelling of Engineering Problems*, vol. 12, no. 7, pp. 2387–2404, 2025. https://doi.org/10.18280/mmep.120718

[12] Z. Zhang, Z. Pan, W. Li, and Z. Su, "X-board: An egocentric adaptive AR assistant for perception in indoor environments," *Virtual Reality*, vol. 27, no. 2, pp. 1327–1343, 2023. https://doi.org/10.1007/s10055-022-00742-3

[13] M. Firdaus, N. Thangavelu, A. Ekbal, and P. Bhattacharyya, "I enjoy writing and playing, do you?: A personalized and emotion grounded dialogue agent using generative adversarial network," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2127–2138, 2022. https://doi.org/10.1109/TAFFC.2022.3155105

[14] Z. Arias *et al.*, "Personalized preclinical training in dental ergonomics and endodontics in undergraduate dentistry students (Pilot Study)," *Acta Med Okayama*, vol. 77, no. 2, pp. 147–159, 2023. https://doi.org/10.18926/amo/65144

[15] B. C. Patel, N. A. Sapp, and R. Collin, "Standardized range of conformers and symblepharon rings," *Ophthalmic Plastic & Reconstructive Surgery*, vol. 14, no. 2, pp. 144–145, 1998. https://doi.org/10.1097/00002341-199803000-00013

[16] R. A. Hamad, M. Kimura, L. Yang, W. L. Woo, and B. Wei, "Dilated causal convolution with multi-head self attention for sensor human activity recognition," *Neural Computing and Applications*, vol. 33, no. 20, pp. 13705–13722, 2021. https://doi.org/10.1007/s00521-021-06007-5

[17] P. V. Rouast and M. T. Adam, "Single-stage intake gesture detection using CTC loss and extended prefix beam search," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2733–2743, 2020. https://doi.org/10.1109/JBHI.2020.3046613

[18] W. Messaoud, R. Trabelsi, A. Cabani, and F. Abdelkefi, "Maritime object detection using attention mechanism," *Signal, Image and Video Processing*, vol. 18, no. 2, pp. 1833–1845, 2024. https://doi.org/10.1007/s11760-023-02897-1

7 AUTHORS

Jingfang Wu is an Associate Professor at the School of Humanities, Hunan City University and holds a Ph.D. in Linguistics. Her research focuses on cognitive linguistics and language education. She has published over 30 articles in both Chinese and international journals and has led four provincial-level research projects in linguistics and language teaching (E-mail: wujingfang@hncu.edu.cn).

Yueying Shen is an Associate Professor at the Foreign Language School, Guangzhou Institute of Science and Technology and holds a Ph.D. in Rhetoric and Linguistics. Her research focuses on applied linguistics and language education. She has published over 10 articles in both Chinese and international journals and has participated in four provincial-level research projects in linguistics and language teaching (E-mail: syy_winnie@gzist.edu.cn).