RT-AVTC: A Real-Time Audio-Visual Tone Correction Network Using Multimodal Deep Learning and Causal Convolution

Min Wang[1, 2, *], Yehui Duan[1]
[1]Nanjing Normal University, Nanjing 210024, China
[2]Henan Vocational Institute of Arts, Zhengzhou 451464, China
E-mail: fishbook2025@163.com
[*]Corresponding author

Keywords: Multimodal deep learning, Chinese tone correction, dynamic time regulation, causal convolution, self-attention mechanism, real-time audio-visual feedback

Received: August 5, 2025

Real-time feedback and robustness against environmental noise are critical challenges in computer-aided Chinese tone learning. This paper introduces RT-AVTC, a novel real-time audio-visual tone correction network designed to address those limitations through multimodal deep learning. The proposed architecture integrates a Multi-task Cascaded Convolutional Network (MTCNN) for audio-visual feature alignment, a causal convolution module and a Bidirectional Long Short-Term Memory (BiLSTM) network for robust temporal sequence classification, and a feedback module incorporating Dynamic Time Warping (DTW) for quantitative error analysis. The model was rigorously evaluated on the public LRS3 dataset. Experimental results demonstrated that the RT-AVTC model achieved a state-of-the-art accuracy of 94.26%, significantly outperforming strong baselines including Conformer and Whisper. Notably, in a challenging -5 dB signal-to-noise ratio environment, the multimodal approach yielded a 17.41% performance improvement over its audio-only counterpart. Furthermore, the error correction module provided targeted feedback, improving the average fundamental frequency (F0) curve deviation by over 55% for learners. With its high accuracy, real-time processing capability, and low computational overhead, the RT-AVTC framework presents a valuable and practical solution for effective computer-aided language learning.

Povzetek: Študija predstavi RT-AVTC, večmodalni avdio-vizualni sistem za sprotno korekcijo kitajskih tonov, ki z MTCNN poravnavo, vzročno konvolucijo + BiLSTM ter DTW-povratno zanko robustno razpoznava in usmerjeno popravlja tonalne napake tudi v šumnem okolju.

1 Introduction

As a tonal language, the pitch fluctuations of Chinese syllables (i.e., tones) are the core elements for distinguishing word meanings and achieving effective communication [1]. For learners of Chinese as a second language, accurately mastering tones is not only the key to standardizing pronunciation but also the necessary path to fluent expression [2]. However, due to differences in native language backgrounds, tone learning often becomes a key and difficult point in the teaching process. With the advancement of Artificial Intelligence (AI) and Human-Computer Interaction (HCI) technology, computer-aided language learning systems have shown great potential in personalized teaching and instant feedback, providing a new technological path for solving the challenges of tone learning [3]. Traditional computer tone recognition and error correction research is mostly based on single modal audio signal processing. Early research mainly relied on extracting acoustic features such as fundamental frequency (F0), energy, and duration from speech, and then combining them with classical machine learning algorithms, including hidden Markov models and SVM for classification. In the era of deep learning, researchers have begun to use models including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to achieve significant improvements in the accuracy of tone recognition [4, 5].

In recent years, deep learning has become increasingly widespread in Natural Language Processing (NLP), providing new support for language learning. Yan et al. noted that deep neural networks can automatically extract high-level features from raw language data, improving accuracy and generalization. Combining RNN and CNN significantly optimizes language recognition, translation, and text generation, providing a theoretical basis for building efficient language learning platforms [6]. Xu demonstrated that deep learning has transformed language processing tasks, particularly in named entity recognition, intent recognition, sentiment analysis, and machine translation. Deep learning models effectively capture semantic and contextual relationships through large-scale corpus training, making language learning more intelligent. New architectures such as Transformer have further advanced personalized learning and semantic understanding [7]. Yang reviewed applications and optimization strategies of deep learning in language understanding and generation, introducing mechanisms such as word embedding, language modeling, and dialogue systems, while focusing on model optimization (LSTM, GRU, Transformer) and computational acceleration. The author proposed applying federated learning, model compression, and self-supervised learning to address data security and resource limitations, supporting intelligent development of language learning systems [8]. Arkhangelskaya and Nikolenko systematically reviewed deep learning evolution in language processing, examining distributed word representation, character-level models, and sequence modeling, and analyzing how pre-trained models like BERT and GPT achieved breakthroughs in dialogue systems, grammatical analysis, and sentiment classification [9]. Li discussed deep learning's role in English speech recognition and teaching feedback, finding that advanced models effectively assist learners in correcting pronunciation and improving accuracy, enhancing the learning experience. The intelligent evaluation function provided teachers with real-time feedback and personalized suggestions, significantly improving teaching quality in non-native language environments [10].

The transfer learning and spectral enhancement techniques adopted by Khurana et al. to solve the data shortage problem in emotion recognition also provided a reference for studying the representation ability of enhancing audio features in limited data [11]. The accuracy of tones was directly related to the clarity of semantics. The multilingual speech semantic alignment framework proposed by Khurana et al. provided a new research path for exploring the deep correlation between tone acoustic features and Chinese semantic information [12]. In neuroscience, Choi et al.'s study on infant speech perception revealed that multimodal speech networks already exist before vocalization. This provided theoretical support for studying the effectiveness of tone correction using audio-visual fusion [13]. Finally, to achieve real-time feedback of the system, Zhong et al. developed a fully simulated reservoir computing system. The efficient and low-power spatiotemporal signal processing capability demonstrated by this system provided a hardware level implementation approach for building error correction systems that could run smoothly on user devices and meet low latency interaction requirements [14].

A review of existing research highlights advancements across various aspects of speech processing while revealing a lack of a cohesive framework that achieves robust pitch recognition in real-world noisy environments and provides actionable feedback for learners. Therefore, the architecture selection in the proposed model is not arbitrary, but rather a synthesis of insights from seemingly different fields to construct a system that is both methodological and practically valuable. For instance, the well-documented vulnerability of audio-only systems to noise motivates the adoption of a multimodal audio-visual approach. This design choice is substantiated by findings in neuroscience that demonstrate the human brain's reliance on visual cues, such as lip movements, to enhance speech perception in challenging acoustic conditions. Furthermore, the challenge of accurately modeling the dynamic and subtle contours of Chinese tones has important similarities with the research on pathological speech analysis. Among them, detecting small time deviations in speech patterns is crucial. This parallel strongly justifies the use of a Bidirectional Long Short-Term Memory (BiLSTM) network, a model renowned for its efficacy in capturing long-range temporal dependencies within sequential data. Finally, to bridge the gap between mere recognition and effective pedagogy, the development of the error-correction module is guided by principles from HCI and computer-aided language learning. The integration of Dynamic Time Warping (DTW) is a direct answer to the need for specific, quantitative feedback, moving beyond a simple binary of "correct" or "incorrect" to offer targeted guidance. Consequently, the presented solution is not merely a collection of high-performing components, but a principled integration of multimodal fusion, sequential deep learning, and user-centric feedback, designed to form a cohesive and robust system for real-time Chinese tone correction.

Many experts have researched feature extraction methods for single-modal tone recognition and the application of multimodal fusion in speech emotion and anomaly detection. However, current research still has shortcomings, such as low accuracy in tone recognition in noisy environments and insufficient domain adaptability across speakers and dialects. This study proposes a Real-time Audio-Visual Tone Correction (RT-AVTC) network to address challenges such as the efficient fusion of multimodal features and the generation of real-time corrective feedback. DTW is a well-established algorithm for sequence alignment. The novelty of this framework lies not in the use of DTW alone, but in its principled integration within a cohesive audio-visual learning system. The core innovation is the design of a feedback module that synergistically combines DTW-based quantitative analysis with a multimodal feedback loop. This approach scientifically aligns the learner's fundamental frequency (F0) curve with a standard template, enabling a precise, quantitative comparison of tonal contours. The system then leverages this alignment to provide targeted, multi-faceted feedback: the visual representation of the F0 curve highlights specific deviation points and supplements quantitative data on key parameters such as starting spacing and range, resolving skill-related errors. This unique combination of robust multimodal fusion for recognition and integrated DTW for targeted, visual, and quantitative feedback is expected to significantly advance the intelligent assistance and optimization of computer-aided Chinese tone acquisition. A summary of related work is shown in Table 1.

Table 1: Summary of related work

Model	Primary Datasets Used	Key Architectural Features	Reported F1-Score / Accuracy (%)	Real-Time Factor (RTF)	Known Limitations
Conformer	LibriSpeech, AISHELL-1	Transformer with convolution layers for local and global feature capture.	92.41 / 92.53 (on LRS3)	0.45	Performance degrades significantly in high-noise environments without specialized augmentation.
AV-HuBERT	LRS3, VoxCeleb2	Self-supervised multimodal learning; masked prediction on audio and visual streams.	92.88 / 92.94 (on LRS3)	0.53	High computational cost; real-time performance is a challenge for interactive applications.
Whisper	680k hours of diverse, weakly supervised web data.	Large-scale, weakly supervised Transformer model trained for general speech recognition.	91.53 / 91.68 (on LRS3)	1.24	Not optimized for specific tasks like tone correction; high latency makes it unsuitable for real-time feedback.
RT-AVTC (This Work)	LRS3, MMEG	Causal convolution with BiLSTM and audio-visual fusion for real-time processing.	94.18 / 94.26 (on LRS3)	0.31	Robust to noise but performance is dependent on clear, frontal-view visual input.

2 Methods and materials

2.1 Design of input data for tone correction model

To provide a clear and structured framework, this research is guided by a set of explicit objectives and testable hypotheses. The investigation seeks to answer the following primary research questions: (1) To what extent can an audio-visual multimodal approach improve the accuracy and robustness of Chinese tone recognition compared to a conventional audio-only model, particularly in noisy environments? (2) Is it feasible to design a system that not only recognizes tones accurately but also operates in real-time with low latency to provide immediate, actionable feedback for language learners? (3) How effective is a quantitative feedback mechanism, based on DTW, in diagnosing and correcting users' tonal errors? Correspondingly, this study posits three central hypotheses. H1: The integration of visual features will significantly enhance tone recognition accuracy and robustness, especially under low Signal-to-Noise Ratio (SNR) conditions. H2: A computationally efficient architecture can achieve real-time processing suitable for interactive learning applications. H3: Providing learners with quantitative feedback on their pitch contour deviation will lead to a measurable improvement in their tone production accuracy. The validation of these hypotheses is based on several key measurable outcomes: classification accuracy on the LRS3 dataset across various noise levels, the RTF and system latency metrics, and the percentage reduction in the average F0 curve deviation observed in a user study.

.

This study focuses on the research of a real-time Chinese tone correction algorithm based on Multimodal Deep Learning (MDL), with the core of designing and implementing an MDL-based algorithm model [15]. The overall architecture follows a complete closed-loop process from data acquisition, signal processing, multimodal feature analysis, to intelligent decision-making and feedback, as shown in Fig. 1.

In Fig. 1, the system starts with a multi-modal optimized data synchronization acquisition end, using microphones, cameras, etc., to collect voice, facial video, and muscle activity data. The data undergo audio processing, video processing, and electromyographic signal fusion processing, and are further processed by the audio and video sub-network. The correct tone results and error diagnosis analysis are output, and the deep classification diagnostic model is input. Finally, through the real-time feedback generation module, tone correction feedback is provided to Chinese learners, achieving a multimodal data-driven tone learning loop.

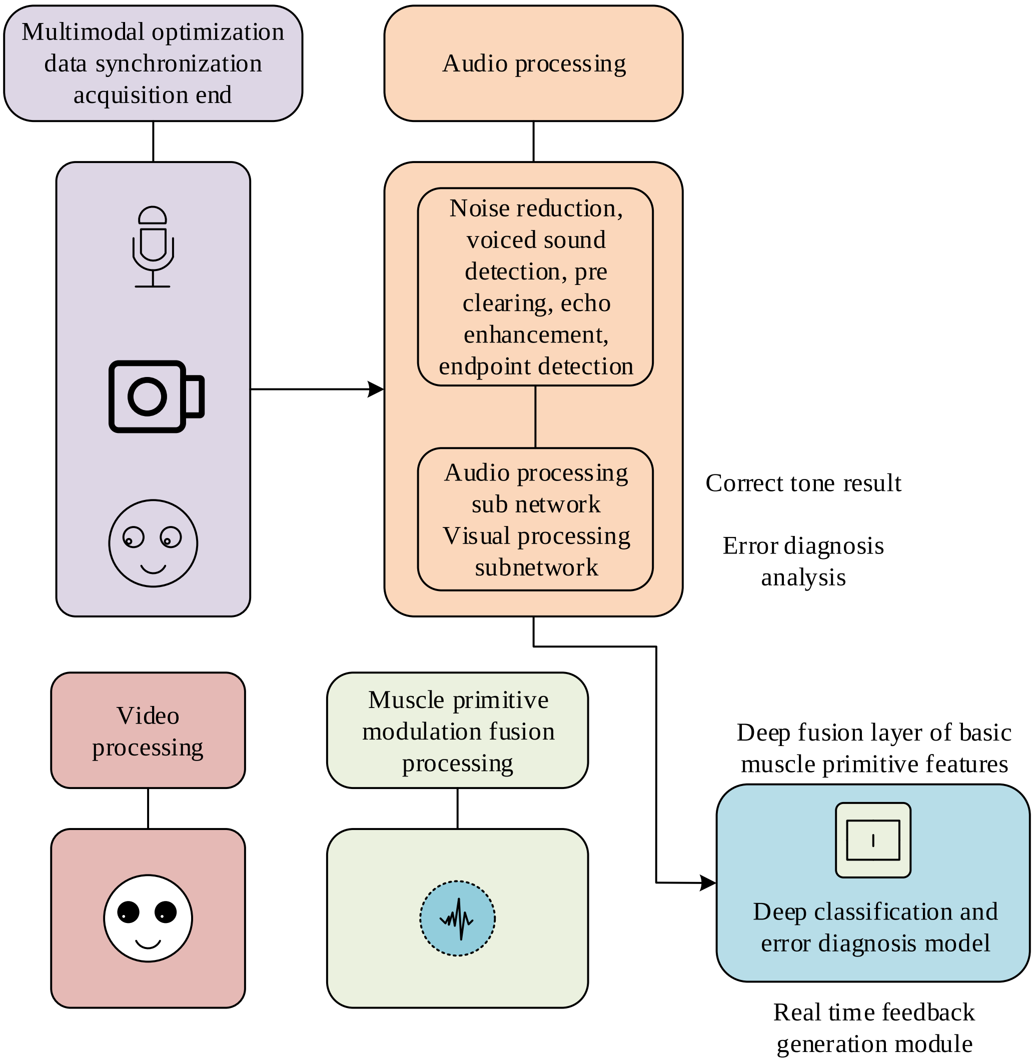

Multimodal optimization data synchronization acquisition end
Audio processing
Noise reduction, voiced sound detection, pre clearing, echo enhancement, endpoint detection
Audio processing sub network Visual processing subnetwork
Correct tone result
Error diagnosis analysis
Video processing
Muscle primitive modulation fusion processing
Deep fusion layer of basic muscle primitive features
Deep classification and error diagnosis model
Real time feedback generation module

Figure 1: Overall framework of tone correction system for Chinese learners

However, the collected raw signals cannot be directly used for model training. They are often mixed with various noises and irrelevant information, and must go through a fine preprocessing process [16]. To compensate for the loss of high-frequency parts during the pronunciation process and improve the SNR of the signal, it is necessary to perform pre-emphasis processing on the signal. This is equivalent to passing the signal through a high pass filter, whose transfer function is shown in equation (1).

$$H(z)=1-az^{-1} \qquad (1)$$

In equation (1), z^{-1} is the unit delay operator. The coefficient a is a pre-emphasis factor, typically ranging from 0.9 to 1.0. After pre-emphasis processing, the overall flatness of the speech signal spectrum is improved, and the details of the high-frequency part are effectively enhanced. To maintain a smooth transition between frames, a certain amount of overlap is usually set between adjacent frames. Short-term energy reflects changes in signal amplitude, with voiced segments typically having much greater energy than voiceless and muted segments. The calculation is given by equation (2).

$$E_n = \sum_{k=-\infty}^{\infty} [x(k)w(n-k)]^2 \qquad (2)$$

In equation (2), E_n is the short-term energy of frame n, $x(k)$ is the original speech signal, and $w(n-k)$ is the window function. The short-term Zero Crossing Rate (ZCR) refers to the number of times a frame signal passes through zero values. Due to its high-frequency characteristics, the ZCR of voiceless segments is significantly higher than that of voiced segments. This definition is shown in equation (3).

$$Z_n = \sum_{k=-\infty}^{\infty} | sgn(x(n)) - sgn(x(n-1)) | \omega(n-k) \qquad (3)$$

In equation (3), Z_n is the short-term ZCR of frame n, and $sgn(\cdot)$ is the sign function. By setting two energy thresholds, high and low, and corresponding ZCR thresholds, the attributes of each frame are comprehensively judged to accurately determine the starting and ending points of the speech signal, and separate the voiced segments for analysis. For synchronously collected video signals, the goal of preprocessing is to extract and normalize the lip region containing key pronunciation information. The process mainly includes: firstly, applying Multi-task Cascaded Convolutional Networks (MTCNN) to each frame of the image to locate the position of the face. The MTCNN architecture is shown in Fig. 2.

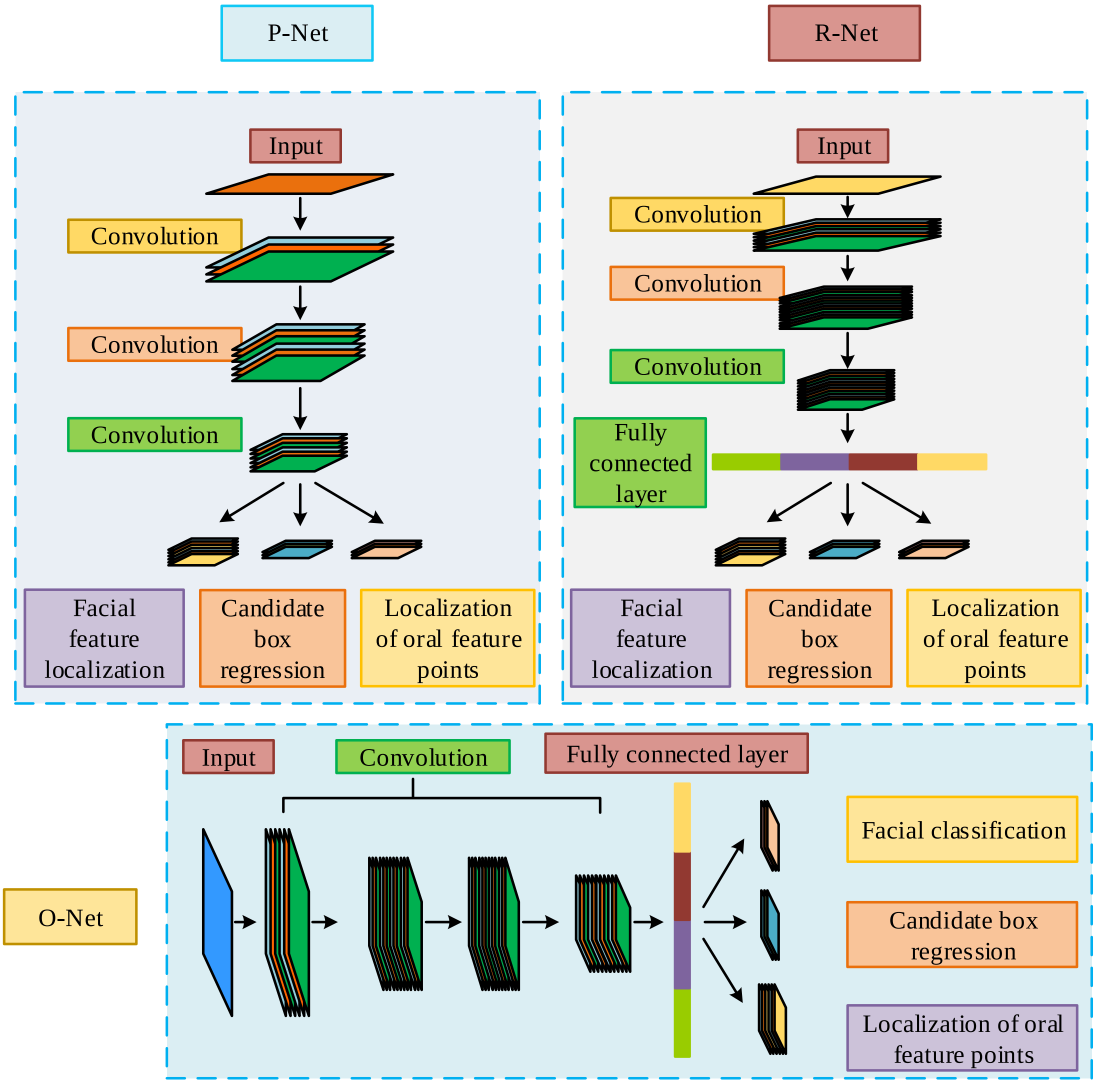


Figure 2: Architecture of MTCNN

In Fig. 2, the function of the MTCNN in this workflow is to accurately detect and isolate the lip region from each video frame, which serves as the visual input for the model. The process adheres to the standard, sequential application of MTCNN's three-stage cascaded architecture. Initially, the Proposal Network (P-Net) scans the image to generate a set of candidate facial bounding boxes. These candidates are then passed to the Refine Network (R-Net), which filters out false positives and performs initial calibration of the bounding boxes. Finally, the Output Network (O-Net) conducts a final refinement to output two key pieces of information: a high-precision facial bounding box and the coordinates of five primary facial landmarks, including the corners of the mouth. Based on the coordinates of these mouth landmarks, a precise Region of Interest (ROI) that tightly crops the lip area is calculated and extracted from the frame. This extracted lip ROI is then normalized to a fixed size and converted to grayscale, serving as the standardized visual feature sequence for the subsequent multimodal fusion module.

2.2 Multimodal feature fusion and tone classification module

After the front-end data collection and preprocessing process, the system obtained audio and visual feature sequences that are accurately synchronized in time. The multimodal feature fusion and tone classification module is the core and brain of the entire real-time error correction system, carrying two key tasks. The first task is to efficiently and intelligently fuse features from different modes that contain part of the information about pronunciation to form a unified feature representation that is more robust and more complete than any single mode. Task two is to build a powerful classification model based on the fused features, which not only needs to accurately recognize the actual tone spoken by the learner, but also needs to be able to compare it with the target tone [17]. To achieve the two key tasks mentioned above, this study designs a real-time multimodal fusion model with causal convolution and Self-Attention Mechanism (SAM) as the core. The overall architecture is shown in Fig. 3.

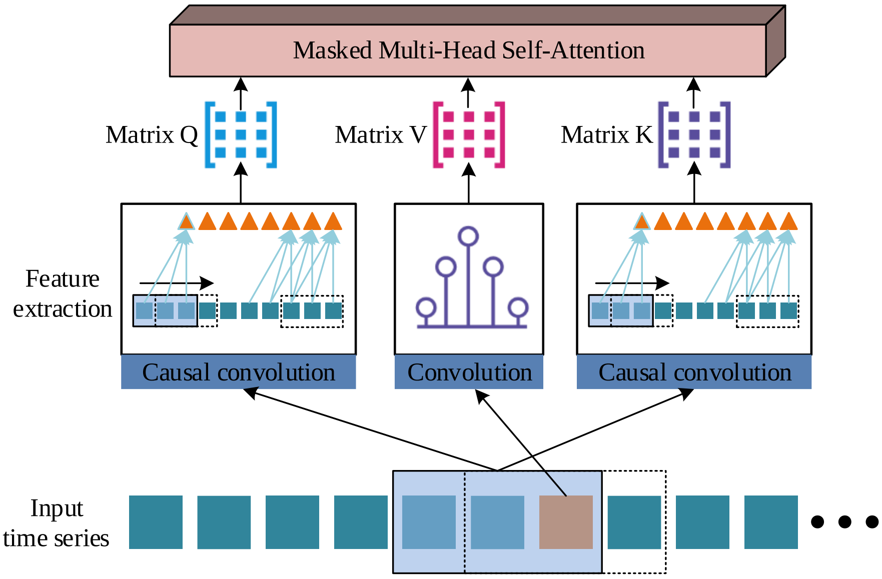


Figure 3: Multimodal fusion module structure based on causal convolution and SAM

In Fig. 3, the model adopts a parallel multi-branch structure to encode the input multimodal features. Among them, for the two modalities with strong temporal dependence, audio and visual, the model adopts a causal convolution module. Unlike standard convolution, causal convolution only covers the past and current time points when calculating the output of the current time point, and never touches on future information. At each time step t, the system first calculates an attention score e_t through a small feedforward neural network, as shown in equation (4).

$$e_t = \boldsymbol{v}^T \tanh(\boldsymbol{W}_a h_t^{audio} + \boldsymbol{W}_v h_t^{video} + \boldsymbol{b}) \quad (4)$$

In equation (4), h_t^{audio} and h_t^{video} are the audio and visual feature vectors for that time step. $\boldsymbol{W}_a$ and $\boldsymbol{W}_v$ are the weight matrices of the corresponding modes, $\boldsymbol{b}$ is the bias vector, and $\boldsymbol{v}$ is a parameter vector used to map the result to a scalar. $\tanh$ is the hyperbolic tangent activation function. Subsequently, the scores e_t of all time steps are normalized using the Softmax function to obtain the attention weights α_t of each modality at each time step. Finally, the original modality features are weighted and summed using these weights to generate the final fused feature vector c_t, as shown in equation (5).

$$c_t = \alpha_t^{audio} h_t^{audio} + \alpha_t^{video} h_t^{video} \quad (5)$$

In this way, the model can dynamically focus on the most reliable or informative modality at present, generating a more robust and reflective fusion feature sequence $\{c_1, c_2, ..., c_T\}$ that better reflects the essence of pronunciation. Among them, T is the total frame number of syllables. To fully utilize the contextual information in the fused feature sequence, this study selected BiLSTM as the core classifier, as shown in Fig. 4.

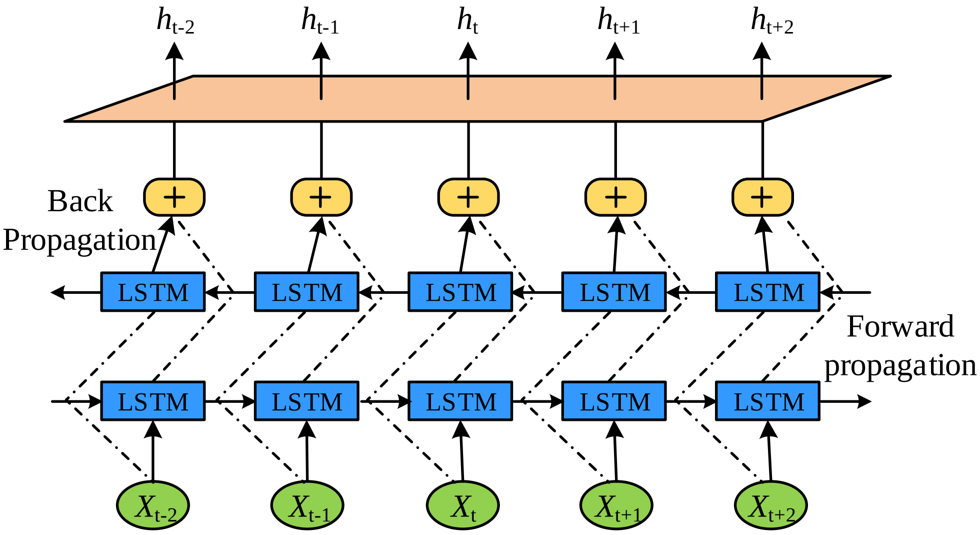


Figure 4: BiLSTM classification module structure

In Fig. 4, the LSTM layer located below processes the input fused feature sequence in chronological order from left to right. The LSTM layer located above processes the same sequence in reverse order. The processing process of BiLSTM at each time step t is shown in equation (6).

$$\begin{cases}\overrightarrow{h_t} = LSTM_{\text{forward}}(c_t, \overrightarrow{h_{t-1}}) \\ \overleftarrow{h_t} = LSTM_{\text{backward}}(c_t, \overleftarrow{h_{t+1}})\end{cases} \qquad (6)$$

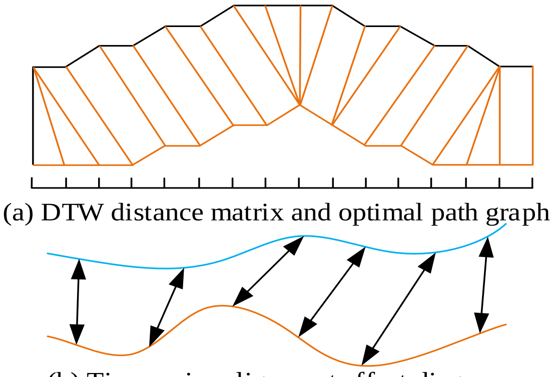
In equation (6), $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are the hidden states of the forward and backward LSTM at t, encoding information from the beginning and end of the sequence to the current time. Finally, this vector representing the tone information of the entire syllable is fed into a fully connected layer and, through a Softmax, outputs a probability distribution vector $\boldsymbol{p}$, as shown in equation (7) [18].

$$\boldsymbol{p} = \text{softmax}(\boldsymbol{W}_y[\overrightarrow{h_T};\overleftarrow{h_T}]+\boldsymbol{b}_y) \qquad (7)$$

In equation (7), $\boldsymbol{W}_y$ and $\boldsymbol{b}_y$ are the weight matrix and bias vector of the output layer. The dimension of vector $\boldsymbol{p}$ is the number of tone categories (for example, four tones in Mandarin are 4). The model ultimately identifies the category with the highest probability value as the tone recognition result.

2.3 Tone error diagnosis and feedback mechanism based on DTW

Error diagnosis is an extended function of the classification module. The system will compare the recognition results output by the model with the target tone of the current exercise. If the two do not match, the system will determine a pronunciation error and enter the diagnostic process. To achieve accurate diagnosis and generate quantitative feedback, the second step of the module is to conduct quantitative comparative analysis. Due to the varying speed of speech among different individuals, even the duration of each pronunciation by the same person may differ. Therefore, the length of the F0 sequence generated by learners is usually inconsistent with that of the F0 sequence in the standard model [19]. To make meaningful comparisons between these two time series with different lengths, this study uses the DTW algorithm, as shown in Fig. 5.


(a) DTW distance matrix and optimal path graph

(b) Time series alignment effect diagram

Figure 5: DTW algorithm processing time series

Fig. 5 (a) shows the DTW distance matrix and optimal path, while Fig. 5 (b) shows the time series alignment effect. The DTW algorithm uses dynamic programming to find the optimal path in the cumulative distance matrix, maps two unequal time series point-to-point, and minimizes distortion by scaling the time axis. Assuming that the F0 sequence produced by the learner is $U = (u_1, u_2, ..., u_n)$, and the F0 sequence of the standard pronunciation model is $S = (s_1, s_2, ..., s_m)$, where n and m are the lengths of the two sequences. The DTW algorithm constructs an $n \times m$ cumulative distance matrix γ. Each element $\gamma(i, j)$ in the matrix represents the minimum cumulative distance between the top i points of sequence U and the top j points of sequence S. The calculation method is shown in equation (8).

$$\gamma(i,j)=d(u_i,s_j)+\min\{\gamma(i-1,j),\gamma(i-1,j-1),\gamma(i,j-1)\} \quad (8)$$

In equation (8), $d(u_i,s_j)$ is the local distance between sequence points u_i and s_j, usually using Euclidean distance, i.e. $|u_i,s_j|$. The core of this recursive relationship is that the best path to point (i,j) must pass through one of its adjacent 3 points $(i-1,j)$, $(i-1,j-1)$, or $(i,j-1)$. Ultimately, the value of $\gamma(n,m)$ is the DTW distance between two sequences, which can serve as an indicator of the overall similarity between learners' pronunciation and standard pronunciation [20]. The workflow of the error correction and feedback module is shown in Fig. 6.

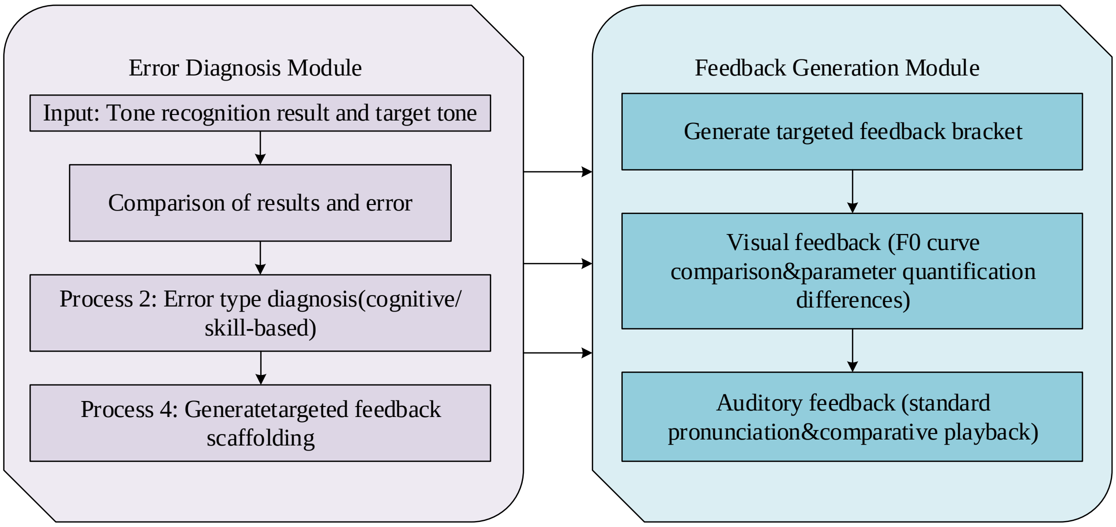

Error Diagnosis Module
Input: Tone recognition result and target tone
Comparison of results and error
Process 2: Error type diagnosis(cognitive/ skill-based)
Process 4: Generatetargeted feedback scaffolding
Feedback Generation Module
Generate targeted feedback bracket
Visual feedback (F0 curve comparison&parameter quantification differences)
Auditory feedback (standard pronunciation&comparative playback)

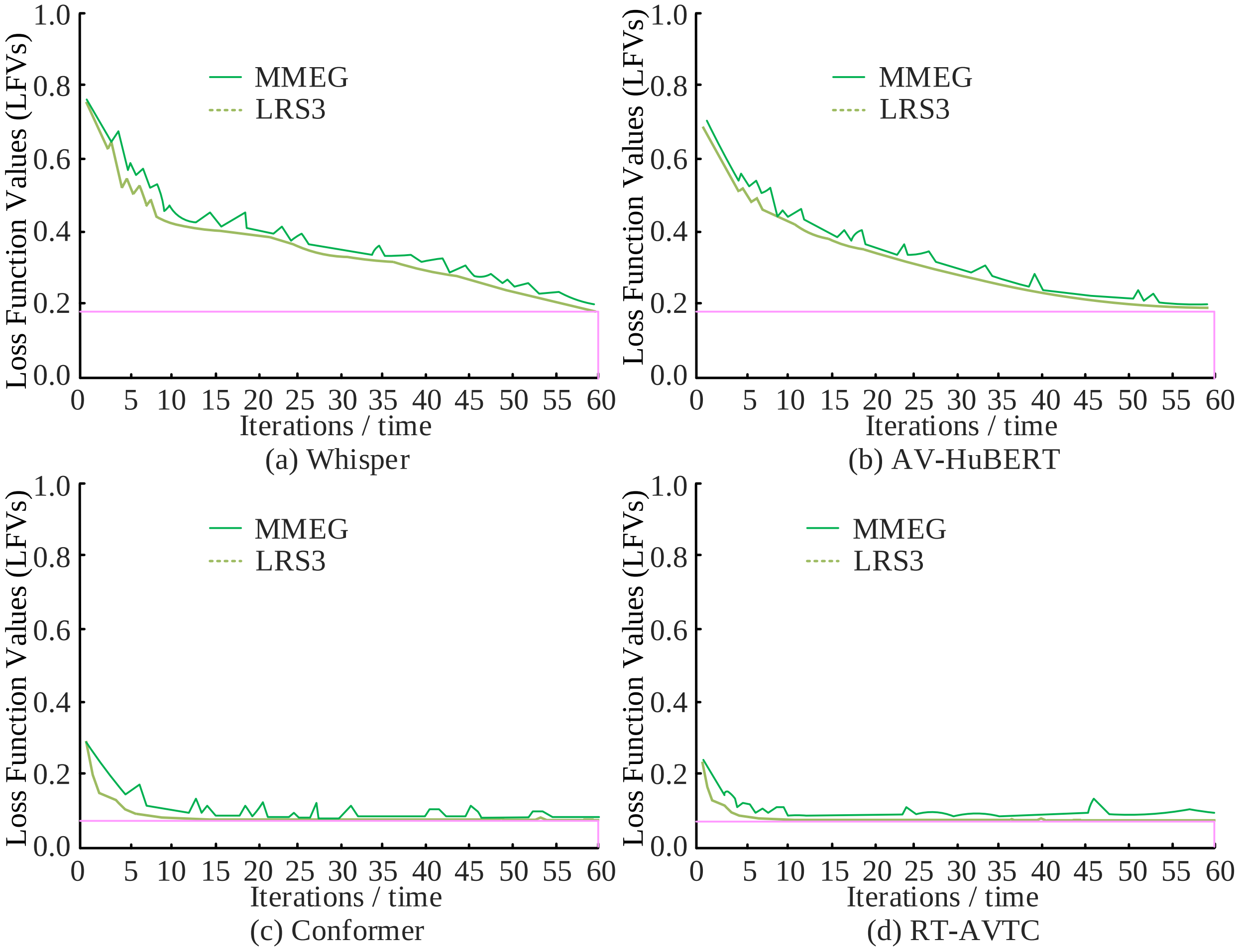
Figure 6: Workflow of error correction and feedback module

In Fig. 6, the system first receives the tone recognition results and compares them with the target tone to determine if there are any errors by judging the nodes. If there is no bias, the correct prompt will be directly output. If there is a bias, an error signal will be triggered, and a type diagnosis will be entered, distinguishing it as cognitive or skill bias. Subsequently, the system generates targeted feedback scaffolds and outputs two forms of feedback. One is visual feedback, which visualizes the location of errors through F0 curve comparison and parameter differences; The second is auditory feedback, which helps learners correct their tone by playing standard pronunciation and contrasting audio. This process realizes the closed-loop design of real-time diagnosis and multimodal feedback. To contextualize the practical application of this real-time system, the user-facing component is conceptualized as a desktop or web-based Graphical User Interface (GUI). In a typical use case, the GUI presents the learner with a target syllable or word. Upon recording, the system captures the user's audio and video input. Immediately after the utterance is complete, typically within the sub-200-ms real-time threshold, the interface updates to display the feedback. The feature of this feedback screen is a central panel that displays a visual comparison of the user-extracted F0 pitch profile overlaid with the standard template profile, highlighting areas of significant deviation. Accompanying this visual graph is the quantitative DTW distance score and specific corrective advice (e.g., "start pitch too high"). This immediate, multi-faceted presentation of results is what constitutes the real-time learning loop for the end-user.

3 Results

3.1 Model performance evaluation and analysis

The experiment is run on dual Intel Xeon Gold 6248R CPUs, 256 GB DDR4 memory, and 4×NVIDIA A100 80 GB GPU servers. The operating system is Ubuntu 20.04, the deep learning framework is PyTorch 1.12, equipped with Python 3.9, CUDA 11.4, and cuDNN 8.2. The model is trained for 100 epochs with a batch size of 64, using cross-entropy loss. The AdamW optimizer is selected over the standard Adam optimizer for its improved implementation of weight decay, which helps prevent overfitting and enhances model generalization. This is paired with a cosine annealing learning rate scheduler, starting at an initial rate of 1e-4. This scheduler is chosen because it allows the learning rate to decrease smoothly, enabling the model to converge into a more stable and robust minimum, which is particularly effective for complex sequence modeling tasks. The one with the lowest validation set loss is used for testing. The LRS3 dataset, although primarily designed for lip reading, is motivated by its large-scale, diverse, and precisely synchronized audiovisual data, which is crucial for training robust multimodal fusion models. Due to the lack of explicit tone annotations in the dataset, tone labels required for generating Mandarin speech segments are generated through pseudo-labeling. A state-of-the-art, pre-trained acoustic model is used to perform forced alignment on the audio tracks, automatically extracting phoneme-level information, from which the corresponding tone for each syllable is derived. This standard procedure provides the necessary ground-truth labels for training and evaluating the tone correction task. This study compares the Conformer: Convolution-augmented Transformer for Speech Recognition (Conformer) model, the Audio-Visual HuBERT: Self-Supervised Learning of Audio-Visual Speech Representations via Masked Multimodal Cluster Prediction (AV-HuBERT) model, and the Robust Speech Recognition via Large-Scale Weak Supervision (Whisper) model with the proposed RT-AVTC. The changes in the Loss Function Values (LFVs) of each model when training is shown in Fig. 7.



(a) Whisper

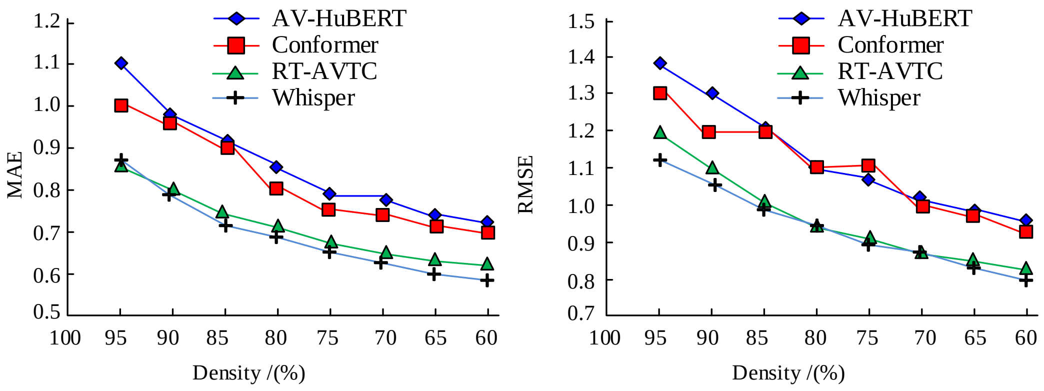
(b) AV-HuBERT

(c) Conformer

(d) RT-AVTC

Figure 7: Comparison of training losses of different models on different datasets

Figs. 7 (a) and (b) show the loss function changes of Whisper and AV HuBERT models. Both models show a relatively gentle downward trend, gradually converging after about 60 iterations, and the final loss value remains around 0.2. The Conformer model in Fig. 7 (c) exhibits faster learning efficiency, with its loss function rapidly decreasing to a lower level within the first 10 iterations. In Fig. 7 (d), the RT-AVTC model exhibits the best performance, with its loss function achieving the fastest descent after training begins and converging to a stable state close to zero in almost 5 iterations. The final convergence values on both datasets are the lowest among the four models. The performance comparison of different models under different data densities is shown in Fig. 8.



(a) MAE

(b) RMSE

Figure 8: Comparison of different models under various data densities

Fig. 8 shows a comparison of Mean Absolute Error (MAE) indicators. Under all data density conditions (from 95% to 60%), the curve corresponding to the RT-AVTC model is always at the bottom, indicating that its MAE value is the lowest among all comparison models. This indicates that the RT-AVTC model has the highest prediction accuracy. Fig. 8 (b) shows the results of the Root Mean Square Error (RMSE) index, which has been further validated. The RMSE value of RT-AVTC remains the lowest among all models. To ensure the practical applicability of the system for real-time feedback, it is crucial to define the constraints for system latency. For an interactive computer-aided language learning tool, the feedback loop must be perceived as instantaneous by the user to be effective. Drawing from established principles in HCI, a total system latency of under 200 ms is generally considered the threshold for real-time perception. The RTF is directly related to this latency, where Latency = RTF × Input Audio Duration. Therefore, to meet this requirement, the system is designed with a target RTF that ensures the processing time remains well below this perceptual threshold for typical utterance lengths. For instance, for a short utterance of one second, the model's measured RTF of 0.18 corresponds to an approximate processing latency of 180 ms. This value falls within the acceptable range, confirming the system's suitability for providing immediate, real-time corrective feedback in a live learning scenario. Table 2 shows the comprehensive performance comparison of various models on the LRS3 and MMEG datasets.

Table 2: Comprehensive results of different models on LRS3 and MMEG datasets

Model	Params (M)	RTF	LRS3's F1-Score (%)	LRS3's Accuracy (%)	MMEG's F1-Score (%)	Data Scarcity (Density @ 60%) (%)
RT-AVTC	85.27	0.31	94.18	94.26	92.65	/
Conformer	118.51	0.45	92.41	92.53	90.17	/
AV-HuBERT	145.88	0.53	92.88	92.94	90.56	/
Whisper	244.16	1.24	91.53	91.68	89.24	/

In Table 2, the RT-AVTC achieves an F1-Score of 94.18% and an accuracy of 94.26% on the LRS3 dataset with a parameter size of 85.27 M and an RTF of 0.31, and an F1-Score of 92.65% on the MMEG. The parameters of the Conformer model are 118.51 M, RTF is 0.45, and other corresponding indicators are 92.41%, 92.53%, and 90.17%. AV-HuBERT with 145.88 M parameters RTF 0.53, obtained corresponding values of 92.88%, 92.94%, and 90.56%. Whisper has 244.16 M and RTF 1.24, with F1-Score of 91.53% and 91.68% in LRS3 and 89.24% in MMEG.

3.2 Noise robustness and pronunciation deviation correction results

The dataset consists of 8,000 samples of Mandarin four-tone monosyllabic speech recorded by Chinese learners and native speakers, with a speech sampling rate of 16 kHz and 16-bit PCM. Although surface Electromyographic (sEMG) data were collected as part of a broader experimental design to explore articulatory muscle activity, the analysis of this third modality is beyond the scope of the current paper. This study focuses specifically on the performance and robustness of the audio-visual fusion model. The integration and comparative analysis of EMG signals are therefore reserved for future work. The data are divided into training set, validation set, and testing set in proportions of 70%, 15%, and 15%. The sample ratio of learners to native speakers is about 3:1, and different noise conditions are constructed by artificially adding street noise and white noise (SNRs of 30, 20, 10, 0, -5 dB) to verify the robustness. During the experiment, each participant completes the tone reading task in quiet and different noise environments. The system synchronously records the audio, video, and EMG signal three-mode data, and outputs the recognition results and error correction feedback in real-time. Meanwhile, the changes in DTW distance and the number of visual errors before and after error correction are tracked and recorded to comprehensively evaluate the robustness and error correction effect. The accuracy of tone recognition under different noise conditions using the multimodal fusion model is listed in Table 3.

Table 3: Tone recognition accuracy of multimodal fusion model under different noise conditions

Noise Level (dB)	Audio-Only Model (%)	Visual-Only Model (%)	Multimodal Fusion Model (%)	Improvement over Audio-Only (%)
30	91.23 (±1.1)	72.18 (±2.5)	95.67 (±0.9)	+4.44
20	85.79 (±1.4)	69.54 (±2.8)	93.12 (±1.0)	+7.33
10	78.46 (±1.9)	66.82 (±3.1)	89.97 (±1.3)	+11.51
0	69.32 (±2.6)	64.19 (±3.5)	83.46 (±1.7)	+14.14
–5	61.87 (±3.2)	62.03 (±3.8)	79.28 (±2.1)	+17.41

The noise robustness of the proposed multimodal fusion model is rigorously evaluated against its single-modality counterparts, with the results detailed in Table 3. To ensure the reliability of these findings, all accuracy metrics are calculated as the mean and standard deviation over 10 independent trials. The data clearly show that the multimodal fusion model significantly outperforms both the audio-only and visual-only models across all tested SNRs. Notably, as the noise level increases, the performance gap widens, culminating in a 17.41% absolute improvement over the audio-only model in the highly challenging -5 dB environment. Furthermore, the standard deviation values provide crucial support for the model's robustness claim. The multimodal model consistently exhibits a lower standard deviation compared to the single-modality models, especially under severe noise conditions. This low variance indicates that its superior performance is not only high on average but also stable and consistent across different trials, confirming its reliability for practical applications in real-world, non-ideal acoustic settings. Group-level statistics on the efficacy of the error correction and feedback module (n=6) is shown in Table 4.

Table 4: Group-level statistics on the efficacy of the error correction and feedback module (n=6)

Metric	Mean	Standard Deviation (SD)
Initial F0 DTW Distance (Hz)	30.66	2.54
Corrected F0 DTW Distance (Hz)	13.45	1.96
Deviation Improvement (%)	56.32	3.28
Visual Feedback Error Reduction (count)	4.17	1.17
Auditory Feedback Repetition (times)	1.83	0.75

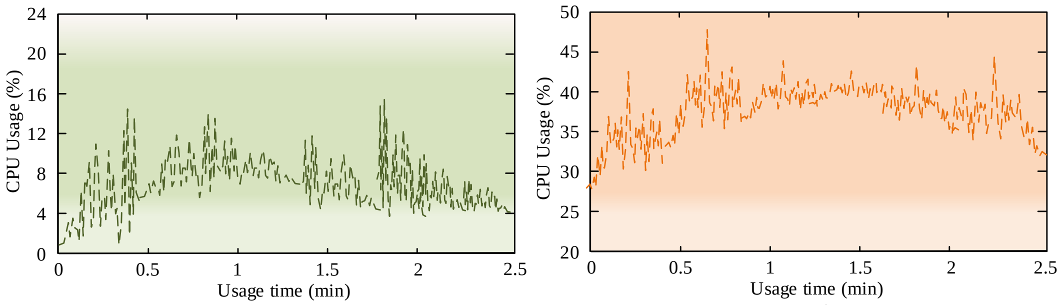
The efficacy of the error correction and feedback module is evaluated by quantifying the change in learners' tone pronunciation deviation. The group-level results, summarized in Table 4, demonstrate a substantial and consistent improvement following the intervention. On average, the initial fundamental frequency (F0) DTW distance from the standard template is 30.66 Hz (SD = 2.54). After receiving corrective feedback from the system, this distance is significantly reduced to a mean of 13.45 Hz (SD = 1.96). This corresponds to an average deviation improvement of 56.32% (SD = 3.28), indicating the module's strong positive effect on tone accuracy. Furthermore, the feedback mechanism is efficient, requiring an average of only 1.83 (SD = 0.75) auditory repetitions to facilitate correction.

To formally verify that the observed performance gains and error reductions are not products of random chance, a series of statistical tests is conducted on the key results. The primary comparisons, namely, the accuracy improvement of the RT-AVTC model over the next-best baseline and the reduction in DTW distance for learners before and after using the system, are subjected to significance testing. Table 5 summarizes the statistical methods employed and their outcomes. The results of these tests provide robust, quantitative evidence for the efficacy and superiority of the proposed model and its corrective feedback mechanism.

Table 5: Summary of statistical significance tests for key results

Comparison	Statistical Test Used	Result (*p*-value)
Accuracy: RT-AVTC vs. Whisper Model	Independent Samples t-test	$p < 0.01$
Error Correction: Pre- vs. Post-Intervention DTW Distance	Paired Samples t-test	$p < 0.001$

The comparison of CPU usage between the RT-AVTC model and the Conformer baseline model is shown in Fig. 9.



(a) RT-AVTC

(b) Conformer

Figure 9: Comparison of CPU usage between RT-AVTC model and Conformer baseline model

Fig. 9 (a) depicts the changes in CPU usage during the operation of the RT-AVTC model. Its occupancy rate curve fluctuates steadily, maintaining an overall low range of 4% to 15%, demonstrating excellent lightweight characteristics. The CPU usage of Conformer in Fig. 9 (b) fluctuates within the high range of 30% to 48%, with an average resource utilization much higher than the model proposed in the study. Through direct comparison of the two graphs, RT-AVTC achieves high-precision recognition and error correction while having extremely low computational overhead, with an average CPU usage rate of less than one-third of the Conformer model. This significant superiority in operational efficiency makes the RT-AVTC model a strong candidate for deployment in resource-constrained environments. Although the current evaluation is conducted on high-performance hardware, the model's lightweight nature presents a promising pathway toward mobile implementation. To substantiate this potential, future work should focus on model compression and optimization strategies. Investigating techniques such as post-training quantization, pruning, and conversion to mobile-friendly formats like TFLite or ONNX will be critical next steps to validate and enable the model's deployment on personal computers and mobile devices. To experimentally validate the architectural choice of the causal convolution module and demonstrate its specific contribution to the model's performance, an ablation study is conducted. Two variants of the RT-AVTC model are created: replacing causal convolution with standard 1D convolutional layers and additional LSTM layers. These variants are then evaluated against the complete RT-AVTC model on the core metrics of accuracy, noise robustness, and real-time performance. The comparative results are presented in Table 6.

Table 6: Ablation study results comparing causal convolution with alternative modules

Model Variant	Accuracy (%)	Accuracy at -5 dB SNR (%)	RTF
RT-AVTC (with Causal Convolution)	94.26	78.50	0.18
Variant A (with Standard 1-D Conv)	93.85	74.23	0.17
Variant B (with LSTM Layer)	94.10	78.15	0.45

The results from the ablation study confirm the superiority of using a causal convolution module in the proposed architecture. The variant with standard 1D convolution shows a significant decrease in accuracy while maintaining similar RTF, especially under -5 dB SNR conditions. This suggests that preserving the temporal causality of the input sequence is crucial for robust feature extraction in noisy environments. Conversely, the variant employing an LSTM layer achieves comparable accuracy to the final model but incurs a significant computational cost, with its RTF increasing by 150%, rendering it unsuitable for the real-time feedback task. Therefore, the causal convolution module provides the optimal trade-off between high accuracy, strong noise robustness, and low-latency processing, justifying its selection for the RT-AVTC network.

3.3 Error analysis

To provide a deeper insight into the model's performance and identify specific failure cases, a detailed error analysis is conducted. The analysis focuses on inter-tone confusion patterns and performance variations across different user types (native speakers vs. language learners). A confusion matrix is generated to visualize the misclassification patterns among the four Mandarin tones, as presented in Table 7. The matrix diagonal represents correct classifications, while off-diagonal values indicate errors where the predicted tone (column) differs from the actual tone (row).

Table 7: Confusion matrix for mandarin tone classification

Actual Tone	Predicted Tone 1	Predicted Tone 2	Predicted Tone 3	Predicted Tone 4
Tone 1	98.1 %	0.5 %	0.8 %	0.6 %
Tone 2	0.7 %	92.5 %	5.8 %	1.0 %
Tone 3	1.1 %	6.2 %	91.3 %	1.4 %
Tone 4	0.9 %	0.4 %	0.6 %	98.1 %

Furthermore, the model's accuracy is broken down by both tone and user type to examine its performance on the distinct pronunciation patterns of native speakers versus learners. The results are summarized in Table 8.

Table 8: Model accuracy breakdown by tone and user type (%)

User Type	Tone 1	Tone 2	Tone 3	Tone 4	Overall
Learner	97.6 %	91.5 %	90.1 %	97.8 %	94.25 %
Native Speaker	99.1 %	94.8 %	93.9 %	98.9 %	96.68 %

The error analysis reveals two key findings. Firstly, the confusion matrix in Table 8 indicates that the majority of classification errors occur between Tone 2 (rising) and Tone 3 (dipping-rising). This is a well-documented area of difficulty for Chinese language learners, as the phonetic contours of these tones can be similar, particularly in continuous speech. The model's confusion mirrors this linguistic challenge. Secondly, Table 8 shows that the model achieves slightly higher accuracy for native speakers compared to learners across all tone types. This performance gap is most pronounced in Tone 2 and Tone 3, indicating that greater variability and non-standard pronunciation in learners' speech pose greater challenges to the model, and the integrated feedback system aims to address this issue.

4 Discussion and comparative analysis

Although the experimental results quantitatively demonstrate the superiority of the proposed RT-AVTC model, it is necessary to conduct a more in-depth qualitative analysis to understand the factors contributing to its success and acknowledge its inherent limitations. The model's significant advantage in noisy environments can be attributed to its core architectural designs. The primary source of robustness is the audio-visual fusion mechanism, which leverages the visual stream of lip movements as a noise-invariant source of phonetic information. When the audio signal is corrupted, the fusion architecture's attention mechanism can dynamically learn to assign greater weight to the stable visual modality, thereby guiding the tone prediction process and ensuring reliability. This is complemented by the temporal feature extraction modules. Causal convolution excels at capturing local speech patterns that may persist even in degraded signals, providing stable input for BiLSTM networks. This network models the long-term dependency relationship of the complete pitch contour in reverse. Despite these strengths, the model has several dependencies and weaknesses. Its performance is contingent upon high-quality visual input. Suboptimal conditions such as poor lighting, non-frontal camera angles, or partial facial occlusions will likely diminish its performance advantage. Furthermore, the model's speaker-independent generalization capabilities warrant further investigation, as its performance may vary when encountering speakers with accents, speaking styles, or facial physiognomies that differ significantly from the LRS3 training distribution. Therefore, future research should focus on addressing these limitations, perhaps by improving visual robustness through advanced data augmentation techniques or enhancing generalization through training on more diverse multi-speaker datasets.

References

[1] Nawroly S S, Popescu D, Thekekara Antony M C. Category-based and target-based data augmentation for dysarthric speech recognition using transfer learning. Studies in Informatics and Control, 2024, 33(4): 83-93. DOI:10.24846/v33i4y202408.

[2] Xu Z Y. Research on deep learning in natural language processing. Advances in Computer and Communication, 2023. DOI: 10.26855/acc.2023.06.018

[3] Yang Z J. Deep Learning Applications in Natural Language Processing and Optimization Strategies. Journal of Modern Education and Culture, 2024. DOI: 10.70767/jmec.v1i2.257

[4] Li L. Application of deep learning technology in speech recognition and language teaching. Lecture Notes in Education Psychology and Public Media, 2024. DOI: 10.54254/2753-7048/59/20241721

[5] Wang Y., Perrin S. Deep Chinese teaching and learning model based on deep learning. International Journal of Languages, Literature and Linguistics, 2024. DOI: 10.18178/ijlll.2024.10.1.479

[6] Yan F, Wang J, Li W. Research on the application of deep learning in natural language processing. Frontiers in Computing and Intelligent Systems, 2024. DOI: 10.54097/m9sxpv44

[7] Xu Z. Research on deep learning in natural language processing. Advances in Computer and Communication, 2023. DOI: 10.26855/acc.2023.06.018

[8] Yang Z. Deep learning applications in natural language processing and optimization strategies. Journal of Modern Education and Culture, 2024. DOI: 10.70767/jmec.v1i2.257

[9] Arkhangelskaya E O, Nikolenko S I. Deep learning for natural language processing: a survey. Journal of Mathematical Sciences, 2023, 273: 533-582. DOI: 10.1007/s10958-023-06519-6

[10] Li L. Application of deep learning technology in speech recognition and language teaching. Lecture Notes in Education Psychology and Public Media, 2024. DOI:10.54254/2753-7048/59/20241721

[11] Khurana Y, Gupta S, Sathyaraj R, Raja S P. RobinNet: A multimodal speech emotion recognition system with speaker recognition for social interactions. IEEE Transactions on Computational Social Systems, 2022, 11(1): 478-487. DOI:10.1109/TCSS.2022.3228649

[12] Khurana S, Laurent A, Glass J. Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation. IEEE Journal of

Selected Topics in Signal Processing, 2022, 16(6): 1493-1504. DOI:10.48550/arXiv.2205.08180

[13] Choi D, Yeung H H, Werker J F. Sensorimotor foundations of speech perception in infancy. Trends in Cognitive Sciences, 2023, 27(8): 773-784. DOI:10.1016/j.tics.2023.05.007

[14] Zhong Y, Tang J, Li X, Liang X, Liu Z, Li Y, et al. A memristor-based analogue reservoir computing system for real-time and power-efficient signal processing. Nature Electronics, 2022, 5(10): 672-681. DOI:10.1038/s41928-022-00838-3

[15] Mnasri Z, Rovetta S, Masulli F. Anomalous sound event detection: a survey of machine learning based methods and applications. Multimedia Tools and Applications, 2022, 81(4): 5537-5586. DOI:10.1007/s11042-021-11817-9

[16] Geng L, Liang Y, Shan H, Xiao Z, Wang W, Wei M. Pathological voice detection and classification based on multimodal transmission network. Journal of Voice, 2025, 39(3): 591-601. DOI:10.1016/j.jvoice.2022.11.018

[17] Yan J, Cheng Y, Wang Q, Liu L, Zhang W, Jin B. Transformer and graph convolution-based unsupervised detection of machine anomalous sound under domain shifts. IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, 8(4): 2827-2842. DOI:10.1109/TETCI.2024.3377728

[18] Akram A, Sabir A. Fine-Tuning BERT for aspect extraction in multi-domain ABSA. Informatica, 2023, 47(9): 123-132. DOI:10.31449/inf.v47i9.5217

[19] Prasad S, Gupta H, Ghosh A. Leveraging the potential of large language models. Informatica, 2024, 48(8): 1-16. DOI:10.31449/inf.v48i8.5635

[20] Arkhangelskaya E. O., Nikolenko S. I. Deep learning for natural language processing: a survey. Journal of Mathematical Sciences, 2023, 273: 533-582. DOI:10.1007/s10958-023-06519-6