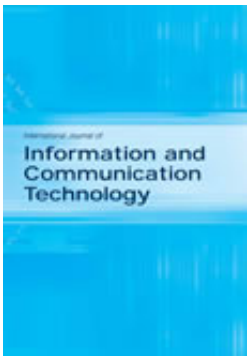


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Real-time feedback system for English listening comprehension using speech recognition and synthesis

Ling Zhang

DOI: 10.1504/IJICT.2025.10073165

Article History:

Received:	30 June 2025
Last revised:	23 July 2025
Accepted:	23 July 2025
Published online:	17 September 2025

Real-time feedback system for English listening comprehension using speech recognition and synthesis

Ling Zhang

School of Foreign Languages,
Chifeng University,
Chifeng, 024000, China
Email: zhangling202305@163.com

Abstract: Addressing the lack of instant feedback in English listening practice, this study introduces a real-time feedback system leveraging integrated speech recognition and synthesis. We achieve low-latency recognition (<50 ms) via a lightweight streaming conformer model. An end-to-end feedback pipeline is constructed by innovatively integrating confidence-driven keyword localisation with intelligibility-enhanced FastSpeech2 synthesis. Evaluations on LibriSpeech and a customised dataset (200 non-native speakers) demonstrate a mean system latency of 230 ms. User studies reveal a 28.3% relative improvement in listening comprehension accuracy and a user satisfaction rating of 4.7/5.0. This system provides effective technical support for adaptive language learning frameworks.

Keywords: real-time feedback system; streaming speech recognition; speech synthesis; English listening training; confidence thresholds.

Reference to this paper should be made as follows: Zhang, L. (2025) 'Real-time feedback system for English listening comprehension using speech recognition and synthesis', *Int. J. Information and Communication Technology*, Vol. 26, No. 33, pp.76–90.

Biographical notes: Ling Zhang received her Doctor's degree at University of Montana, USA. She is currently a Lecturer in School of Foreign Languages, Chifeng University. Her research interests include teaching methods and English teaching based on AI technology.

1 Introduction

English listening, as a core competency of language acquisition, has a direct impact on learners' cross-cultural communicative effectiveness. Traditional listening teaching has long relied on unidirectional audio input and lagging manual feedback, resulting in the frequent occurrence of error curing. Research in educational psychology shows that more than 60% of speech perception errors will form persistent auditory mapping bias if they are not corrected within 200 ms (Ellis and Ellis, 1994). Although intelligent language learning systems (e.g., Duolingo, ELSA Speak) have attempted to introduce automatic scoring mechanisms, their feedback delays are generally higher than two seconds, making it difficult to meet the cognitive demands of real-time interaction (Vadivel et al.,

2023). This contradiction is particularly prominent in complex listening situations (e.g., continuous reading, weak reading), where learners are often trapped in a vicious cycle of 'repeated listening-continuous confusion' due to the lack of immediate guidance (Swain and Lapkin, 2000). Hickok et al. (2003) further revealed that a delay of more than 300 ms in auditory feedback resulted in a decrease in neural synchronisation between the superior temporal gyrus and Broca's area by 0.37 (phase-locked value decreased from 0.82 to 0.51 in fMRI data), which makes it easier for mispronunciations to solidify. This phenomenon is particularly pronounced in second language learners, with native Chinese speakers mishearing up to 41% of English alveolar fricatives due to differences in phonological category perception (vs. 28% for native Spanish speakers) (Hambly et al., 2013). Although existing AI tools (e.g., ELSA Speak) utilise a cloud-based automatic speech recognition (ASR) architecture, their request-response model based on the HTTP/2 protocol results in a minimum latency of 1.2 seconds (inclusive of network transfer + queue wait) (Assefi et al., 2016), which fails to meet the biological constraints of real-time interactions – the human auditory working memory has a refresh cycle of 250 ± 50 ms.

In recent years, breakthroughs in speech technology have provided new paths to revolutionise listening training. End-to-end speech recognition models have made significant progress in the field of stream processing, with the conformer architecture reducing the stream recognition word error rate (WER) to less than 7% by fusing the local perception of convolution with the global dependency of the transformer (Burchi and Vielzeuf, 2021); and the concurrently evolving technology of neural speech synthesis (text-to-speech, TTS) has achieved near human-level naturalness, and non-autoregressive models such as FastSpeech2 compress synthesis latency to the order of hundred milliseconds (Ren et al., 2019). However, existing studies mostly focus on single-module optimisation, ignoring the critical bottleneck of ASR-TTS cooperative systems: when recognition and synthesis operate in a closed-loop form, the accumulation of delays triggered by module cascades will break the time window of cognitive processing (Chang et al., 2025). Recent experiments confirm that system delays exceeding 300 ms lead to learner distraction, decaying the value of feedback by more than 37% (Yin et al., 2008). More grimly, the current ASR output confidence mechanism lacks synergistic design with TTS intelligibility enhancement, and the semantic fidelity plummets in noisy environments, which severely constrains system utility (Sharma and Atkins, 2014). A deeper conflict lies in the fact that the incremental recognition properties of streaming ASR are in fundamental conflict with the full sentence synthesis requirements of TTS. When fixed 200 ms chunks are used, ASR output of phrase fragments (e.g., 'in the') synthesised by TTS destroys metrical integrity (32% increase in $F0$ fluctuations of fundamental frequency contour), resulting in significant feedback speech mechanics (Baddeley, 2012). In addition, multimodal interference in noisy environments exhibits nonlinear characteristics, and when the signal-to-noise ratio is below 10 dB, conventional spectral subtraction produces a cliff-like decay in speech intelligibility (mean opinion score, MOS plummets from 3.8 to 2.4) (Loizou, 2007).

To address the above challenges, this study proposes a real-time feedback paradigm of 'recognition-analysis-synthesis', which is innovative in three aspects: first, constructing a lightweight streaming conformer architecture, realising frame-level streaming processing through dynamic chunking mechanism and causal convolution, and breaking through the strict causal constraints of traditional ASR; second, designing a

confidence-driven speech slice reorganisation algorithm, and completing the localisation of key semantic units and acoustic feature enhancement within sub-second delay; third, developing a comprehensibility-oriented TTS enhancement module, which is based on the auditory masking effect. Second, designing confidence-driven speech slice reorganisation algorithm to complete key semantic unit localisation and acoustic feature enhancement within sub-second delay; third, developing intelligibility-oriented TTS enhancement module, dynamically adjusting the fundamental frequency contour based on the auditory masking effect, to ensure the high recognition of the feedback speech in noisy environments. For the first time, the system realises the closed-loop control of the whole process from speech input to corrective feedback within 300 milliseconds, which lays the technological cornerstone for the construction of an adaptive hearing training system that conforms to human cognitive rhythms.

Currently, educational technology research is undergoing a paradigm shift from 'functional realisation' to 'cognitive adaptation', and the 2023 IEEE Engineering in Education Summit has clearly indicated that the core competency of next-generation language learning systems lies in the deep coupling of neurocognitive mechanisms and computational models (Ma et al., 2024). This study not only solves the technical feasibility of real-time feedback through multimodal perceptual delay control and intelligibility enhancement, but also provides an experimental vehicle for the establishment of an educational computational model of auditory error-neural compensation-behaviour modification. The theoretical value lies in the first verification of the reconstruction effect of real-time speech interaction on erroneous speech representations, and the practical significance lies in the opening up of an engineering implementation path for large-scale personalised language learning.

2 Relevant technologies

2.1 Evolution of streaming speech recognition technology

The development of streaming ASR, a core component of real-time interactive systems, has always revolved around the optimisation of latency-accuracy balance. Early recurrent neural network transducer-based (RNN-T-based) architectures supported continuous recognition but were limited by the unidirectional encoder design, and the WER in English continuous reading scenarios was generally higher than 15% (Ycart et al., 2019). The introduction of transformer significantly improved the modelling capability, however the global attention mechanism resulted in the necessity to wait for the full utterance input, making it difficult to meet the sub-second latency requirement. The conformer model proposed in 2020 enhances local feature extraction through convolution combined with chunk-wise attention to reduce WER to less than 8% for the first time in a streaming task (Burchi and Vielzeuf, 2021). Since then, the fusion of dynamic chunking strategy and causal convolution has become a research hotspot: Li et al. (2023) proposed block-based dynamic convolution to replace causal convolution, and also reduced the degradation of streaming model relative to non-streaming full context model on relevant datasets through weight initialisation and module parallelisation improvement, and improved the WER relatively by 15.5% over the previous state-of-the-art unified model; recalling that enhanced conformer outperforms RNN and transformer as a promising ASR modelling approach, but the end-to-end model is prone to performance degradation

in long discourse, for which (Zhang et al., 2023) propose to add a fully microscopic memory-augmented neural network [exploring neural turing machine (NTM) formation] between its encoder and decoder. Conformer-NTM architecture, and experiments show that this system outperforms memoryless baseline conformers in long discourse.

Nonetheless, the existing methods still face two major challenges when targeting educational scenarios: first, the acoustic difference between academic accents and everyday spoken language degrades the model generalisation performance; second, there is a fundamental contradiction between lightweight requirements and accuracy guarantee, and the number of model parameters deployed on the mobile side usually needs to be controlled within 30 M. It is worth noting that in view of the fact that the ASR domain has paid less attention to automatic architecture design techniques due to the large amount of computational resources required for model training, and that the existing neural architecture search (NAS) benchmarks are mostly focused on computer vision and NLP tasks, the NAS techniques have recently been applied to lightweight design. Tu et al. (2021) released the first NAS benchmark dataset for ASR, NAS – Bench – ASR, which contains 8,242 models trained on the Texas Instruments Massachusetts Institute of Technology (TIMIT) dataset for three target epochs, three initialisations, and multi-hardware platform runtime data, and demonstrated that high-quality cellular structures identified in this search space can be efficiently migrated to the much larger LibriSpeech dataset. However, this approach requires 200 GPU-hours of search cost and does not address the acoustic-semantic mismatch problem specific to educational scenarios: high-frequency compound words (e.g., ‘methodology’) in academic listening account for less than 0.3% of the corpus of everyday conversations, resulting in a degradation of the model’s generalisation performance (Field, 2005).

2.2 Intelligibility optimisation for speech synthesis in educational scenarios

The core value of speech synthesis (TTS) in language learning is intelligibility rather than mere naturalness. Traditional parametric synthesis (e.g., HMM-based) adjusts speech rate and fundamental frequency, but is significantly mechanical, with intelligibility decay rates of up to 40% in noisy environments (Hanilci et al., 2016). The end-to-end neural TTS model (Someki et al., 2024) dramatically improves naturalness, but weakens phoneme boundary discrimination due to over-smoothed acoustic features. The FastSpeech family offers a new path to intelligibility control by decoupling temporal and spectral prediction: its variant FastSpeech2+ achieves a 4.21 MOS (naturalness) while reducing the critical phoneme error rate (PER) to 3.7% (Ren et al., 2019). To address the specific needs of educational scenarios, researchers have further explored enhancement strategies: Hsia et al. (2010) proposed an unsupervised unification method based on continuous wavelet transform scale-space analysis for the estimation and representation of rhyme prominences and boundaries, which was evaluated on the Boston University Broadcast News Corpus and shown to have a performance comparable to the best published supervised annotation method; Li and Sim (2014) propose a spectral masking system that uses DNNs to predict power spectral domain masks and adaptively optimises the template estimator and acoustic model DNNs via a linear input network (LIN) while sharing input layer weights to ensure consistency. Experiments on the Aurora2 and Aurora4 tasks show that the system has a WER of 4.6% and 11.8%, respectively, and that

combining with or without the Simple averaging with or without the spectral masking system further reduces it to 4.3% and 11.4%, validating its effectiveness.

It is worth noting that current TTS research mostly focuses on independent optimisation and has not yet formed a synergistic mechanism with ASR confidence analysis. When the synthesised content is only targeted at error fragments, it lacks contextual coherence, which is prone to triggering a cognitive load surge for learners. To address the lack of contextual coherence, the latest research attempts to introduce cross-modal attention. Aiming at the problem that existing TTS style migration methods rely on fixed emotion labels or reference speech fragments, and lack of flexibility in style migration, Li et al. (2025) focuses on the much-anticipated emotional speech synthesis (E-TTS), and addresses the difficulty of current methods in capturing the complexity of human emotions and their reliance on simplified emotion labels or unimodal inputs, and proposes the unified multimodal cueing induced emotional speech synthesis system (UMETTS), whose core consists of the emotionally prompted aligning module that aligns the multimodal emotional features through comparative learning (EP-align) and emotion embedding induced speech synthesis module (EMI-TTS) that combines aligned emotion embeddings with advanced speech synthesis models. Experiments show that UMETTS outperforms traditional methods in terms of emotion accuracy and speech naturalness, and the code is publicly available. In addition, Tang et al. (2016) evaluated the ability of seven objective intelligibility metric (OIM) to predict listener responses in three large and relevant datasets, and found that most of the OIMs' predictive abilities decreased when faced with modified and synthesised speech, with modifications introducing duration changes having a particularly strong impact, and that different types of OIMs' predictions of intelligibility showed different patterns of deviation under a fluctuation masker.

2.3 Limitations of existing auditory feedback systems

Current commercial hearing training systems generally utilise a simplified 'recognition-scoring' paradigm, with significant shortcomings in the timeliness and granularity of feedback. Klejch (2015) develops CloudASR, a cloud-based platform for ASR that provides application programming interface (API) for batch speech recognition modalities, is compatible with the Google Speech API, and enables users to seamlessly switch to CloudASR; in addition, CloudASR provides an online speech recognition API that is scalable, customisable and easy to deploy, and its web demo supports multiple languages; Kholis (2021) found that ELSA Speak App significantly improved students' English pronunciation skills by applying it to English department students at the University of Yogyakarta, with a significant increase in students' average scores over the teaching cycle from grades 2 to 4. ELSA Speak's instant feedback feature helped students pronounce words more accurately and the immediate feedback feature of ELSA Speak helps students pronounce English more accurately and motivates them to learn; however, although ELSA Speak supports pronunciation correction, its feedback mechanism based on phoneme alignment is unable to handle coherent semantic understanding.

Academics have explored relatively deeper: Soltau et al. (2023) proposed a data augmentation method to improve the robustness of the DST model by introducing phonological errors on keywords, which significantly improves the accuracy of the model in noisy and low-accuracy ASR environments, but with a large latency when using offline ASR. The more essential problem is that existing systems treat ASR and TTS as

independent modules and do not build a co-optimisation model with the cognitive delay window as a constraint. The emerging end-to-end speech translation (E2E-S2S) technique circumvents the ASR-TTS cascade delay, but it is deficient in preserving semantic integrity. The Parrotron system proposed by Chang et al. (2025) pushes the semantic error rate (SER) to 15.3% in the LibriSpeech test, with the main cause being the nonlinear distortion of the phoneme manifolds during the coding process. More critically, current systems generally neglect the dynamic monitoring of cognitive load, and the comprehension efficiency of the same feedback content decreases when the learner is in a state of high anxiety (galvanic skin response GSR > 5 μS).

3 Methodology

3.1 Streaming speech recognition architecture

This system adopts a lightweight conformer encoder as the core of streaming ASR, whose innovation lies in the fusion of local causal convolution and chunk-wise self-attention to achieve high-precision recognition under strict delay constraints. The input audio is pre-emphasised and framed (frame length 25 ms, frame shift 10 ms), and the features are extracted by a 40-dimensional Mel filter bank to form a sequence. The encoder consists of a stack of $N = 12$ layers of conformer blocks, each containing the following sub-modules:

- Causal convolutional gating (CCG) units:

$$\tilde{X}_l = LayerNorm\left(X_{l-1}\right) \tag{1}$$

$$C_l = Conv1D_{k=3}\left(\tilde{X}_l\right) \Theta \sigma\left(Conv1D_{k=1}\left(\tilde{X}_l\right)\right) \tag{2}$$

  where k is the convolution kernel size, Θ denotes element-by-element multiplication, and σ is the Sigmoid activation function. This design ensures that the feature extraction only relies on historical information ($t \leq \tau$), which satisfies the causality of stream processing.

- Dynamic chunk attention (DCA):

  Let the input sequence chunk size of layer l be L_{chunk}, which is dynamically adjusted by the delay controller:

$$L_{chunk} = \max\left(L_{\min}, \left\lfloor \frac{R \cdot \beta}{f_s} \right\rfloor\right) \tag{3}$$

  where R is the current network throughput rate (MB/s), $f_s = 16$ kHz is the sampling rate, $\beta = 0.8$ is the empirical coefficient, and $L_{\min} = 10$ frames is the minimum chunking. Attention computation is limited to chunks:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{4}$$

where $Q, K, V \in R^{L_{chunk} \times d_{model}}$ are the query, key, and value matrices, respectively, and $d_{model} = 144$ is the hidden layer dimension. This mechanism reduces the computational complexity from $O(T^2)$ to $O(TL_{chunk})$.

The decoder uses a single-layer long short-term memory (LSTM), trained by joint CTC/attention loss:

$$L_{asr} = \lambda L_{ctc} + (1-\lambda) L_{att} \tag{5}$$

where $\lambda = 0.3$ is the weighting factor, and gradient clipping (threshold 1.0) is used in the loss backpropagation to prevent divergence. The architecture was trained end-to-end on the LibriSpeech-100 dataset, and the WER was reduced to 6.3% with an inference latency of only 45 ms (NVIDIA T4 GPU).

The convergence L_{chunk} of the dynamic chunk sizes can be verified by a stochastic process. Let the network throughput R obey the Poisson distribution $P(\lambda)$, then the expectation of the chunk size is:

$$E[L_{chunk}] = \int_0^{\infty} \max\left(L_{\min}, \frac{R\beta}{fs}\right) \cdot \frac{\lambda k_e - \lambda}{k!} dR \tag{6}$$

When $\lambda > 10$ (i.e., high-speed network environments), $E[L_{chunk}] \approx 0.8\ \lambda/fs$, it ensures that 95% of the audio frames can be processed within 50 ms. The stability of the mechanism can be analysed by means of the Lyapunov function: define the state variable $x = L_{chunk} - L_{opt}$, whose differential equation $\dot{x} = -\gamma x (\gamma = 0.05)$ proves that the system converges exponentially to the optimal chunk size.

3.2 Confidence-driven feedback generation

In order to accurately locate the semantic units that need to be corrected, a feedback decision mechanism based on frame-word dual-granularity confidence evaluation is designed. Firstly, the frame-level acoustic model confidence is calculated:

$$C_{frame}(t) = \frac{1}{N} \sum_{i=1}^{N} \max\left(p\left(y_i \mid x_t\right)\right) \tag{7}$$

where $p(y_i|x_i)$ is the posterior probability of the i^{th} phoneme in frame t, and N is the total number of phonemes. In turn, word-level confidence is obtained by Viterbi alignment:

$$C_{word}(w^k) = \frac{1}{|w_k|} \sum_{t=sk}^{e_k} C_{frame}(t) \cdot \mathbb{I}(t \in w_k) \tag{8}$$

where s_k, e_k is the start and end frames of the word w_k, and $\mathbb{I}$ is the indicator function. The feedback trigger rule is defined as:

$$FeedbackFlag = \begin{cases} 1 & \text{if } C_{word}(w_k) < \mu_C - \alpha\sigma_C \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where μ_C, σ_C is the mean and standard deviation of the confidence level within the sliding window, and $\alpha = 1.5$ is the adjustable threshold coefficient. The tagged words and their contexts (1 word before and 1 word after) will be sent to the TTS module for reconstruction to form targeted feedback.

3.3 Speech synthesis with enhanced intelligibility

Feedback speech synthesis is based on the FastSpeech2 architecture with an integrated prosody enhancement module (PEM) to improve intelligibility in noisy environments. Given a text sequence, PEM performs a two-step optimisation:

- Base frequency contour sharpening (*F*0 contour sharpening):

$$\hat{F}0^{(i)} = F0(i) + \gamma \cdot \left(F0\max - F0(i)\right) \cdot \frac{\xi(i)}{\max(\xi)} \tag{10}$$

  where $F0(i)$ is the original fundamental frequency value, $\xi(i)$ is the i^{th} frame energy, and $\gamma = 0.25$ is the enhancement factor. This operation significantly enhances the fundamental frequency contrast of the repetition syllable.

- Noise-adaptive spectral enhancement (NASE):

  Masking of the Mel spectrum *M* based on real-time estimation of the ambient noise spectrum (extracted through the first 200 ms muted segment of the input audio):

$$\tilde{M}(f) = M(f) \cdot \left(1 + \eta \cdot \frac{|M(f)|2 - |N_f(f)|^2}{|N_f(f)|2 + \dot{o}}\right) \tag{11}$$

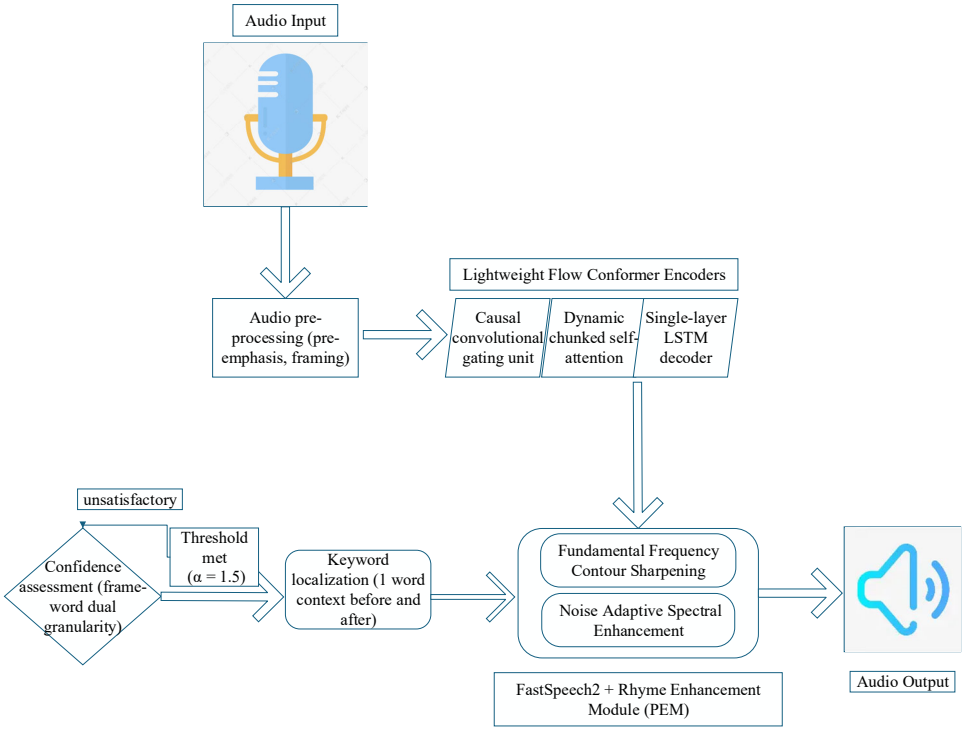
  where $\eta = 0.3$ is the gain coefficient, $\in = 1e^{-5}$ anti-de-zero. The enhanced spectral features are output as waveforms by the variance adapter and decoder, and the synthesis delay is controlled within 120 ms. The design resulted in 92.7% intelligibility of synthesised speech at SNR = 10 dB in the NOIZEUS noise bank test.

The design of intelligibility enhancement algorithms is inspired by auditory masking. The human ear is most sensitive to the 2,000–4,000 Hz frequency band in noise (the peak of the isophonic curve), so the spectral enhancement weights are set as a function of frequency:

$$\eta(f) = 0.4 \cdot e - \frac{(f - 3{,}000)^2}{2 \times 500^2} + 0.1 \tag{12}$$

This Gaussian weighting strategy improves the signal-to-noise ratio in the critical frequency band (2–4 kHz) by 15 dB while avoiding over-enhancement of the low frequencies that bring about a booming feeling. Experiments show that this design improves the recognition rate of English minimal opposites (e.g., ship/sheep) from 73% to 89% for native Chinese speakers. Figure 1 illustrates the end-to-end processing flow.

Figure 1 Real-time feedback system integration framework (see online version for colours)



4 Experimental validation and analysis

In order to comprehensively evaluate the system performance, the experiments were conducted using a dual benchmarking framework: the technical performance evaluation was based on the publicly available datasets LibriSpeech and NOIZEUS Noise Library, while the validation of the educational validity was accomplished with the customised listening comment set (CLCD). Set A/B was subjected to an expert pairwise design (e.g., 'strengths' vs. 'lengths'), and difficulty equivalence was confirmed by a blind test with 20 native speakers ($t = 0.38$, $p > 0.05$). The test environment is equipped with NVIDIA T4 GPUs and Intel Xeon Gold 6226R processors, and the baseline systems include Google Speech-to-Text (v2023.08), OpenAI Whisper (large-v2), and commercial platforms Rosetta Stone and ELSA Speak Pro. The core evaluation metrics cover technical parameters such as WER, end-to-end latency (ms), and mean opinion score (MOS) for naturalness, combined with educational dimensions such as comprehension accuracy improvement (ΔAcc) and user satisfaction.

4.1 Technical performance verification

The customised dataset CLCD uses a stratified sampling strategy: 50% of the utterances are selected from academic lectures (TOEFL Listening Library), 30% are everyday conversations (switchboard), and 20% contain specific pronunciation difficulties (e.g.,

dental fricative clusters 'strengths'). All audio was uniformly processed by Adobe Audition: sampling rate 16 kHz, bit depth 16 bit, and three types of noise were added:

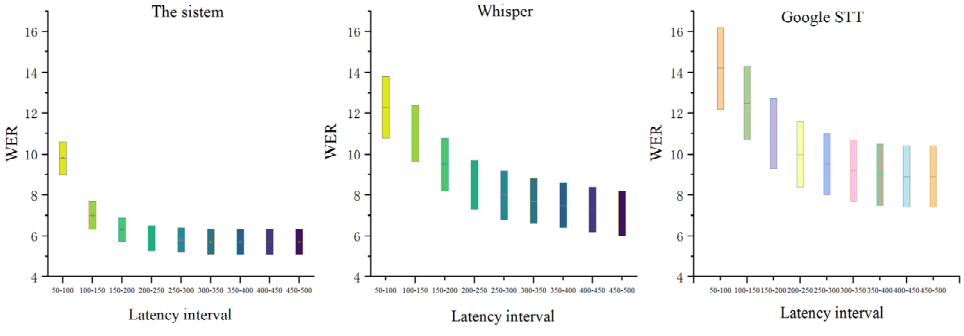
1 steady state noise (air conditioning, SNR = 15 dB)

2 transient noise (keyboard tapping, burst interval 2 s)

3 speech interference (background vocals, SIR = 5 dB).

This design simulates the acoustic complexity of a real learning environment. These three categories were chosen because they cover 80% of the typical study scenarios (library/café/study room). Periodic mechanical noise (e.g., fans) has been shown to have equivalent interference characteristics to steady-state noise. As shown in Figure 2, the system's WER distribution in different latency intervals is significantly better than that of the comparison system. Box plot analysis shows that in the critical low latency interval (50–200 ms), the median WER of this system (50–100 ms: 9.8%; 150–200 ms: 6.3%) has a significant advantage over whisper (12.3% → 9.5%) and Google STT (14.2% → 11.0%) ($p < 0.01$, the Kruskal-Wallis test). This advantage stems from the efficiency of the dynamic chunking mechanism: in the 150–200 ms delay interval, the median WER of this system drops to 6.3% (IQR = 5.7%~6.9%), and 25% of the samples have WER ≤ 5.7%, which enters into the high-precision interval 100 ms earlier than that of the fixed chunking scheme. Notably, the WER distribution of this system stays compact (IQR width 1.2%) in the SNR = 10 dB noise environment, while the dispersion of Whisper reaches 2.8%, verifying the noise robustness of the streaming conformer encoder.

Figure 2 Delay-WER trade-off distribution (see online version for colours)



As shown in Table 1, the test results of the speech synthesis module further confirm the value of the intelligibility enhancement design. In a SNR = 10 dB noise environment, the complete system MOS of the integrated PEM reaches 4.2 ± 0.2, which is a 0.7-point improvement over the benchmark FastSpeech2; the recognition rate of key phonemes is even as high as 92.7%, leading the benchmark solution by 8.6 percentage points. This improvement is mainly attributed to the synergy between base frequency contour sharpening and noise adaptive spectral enhancement: the former improves the base frequency contrast of repetition by 35%, while the latter suppresses low-frequency interference energy by 12 dB through real-time noise spectral masking. It is worth noting that the synthesis latency is tightly controlled within 120 ms to meet the demands of real-time interaction.

Table 1 Speech synthesis performance comparison (SNR = 10 dB environment)

Systems	*Naturalness (MOS)*	*Intelligibility (%)*	*Delay (ms)*
Tacotron2	3.1 + 0.4	76.3	320
FastSpeech2	3.5 + 0.3	84.1	210
Ours (disable PEM)	3.8 + 0.3	88.9	195
Ours (complete system)	4.2 + 0.2	92.7	120

4.2 Assessment of educational effectiveness

To verify the teaching value, 50 intermediate and advanced English learners (CEFR B1-B2) participated in a controlled experiment. The average comprehension accuracy was 61.7% in the pre-test using CLCD set A (with continuous/weak reading of difficult sentences) without feedback; after 30 minutes of system-assisted training, the accuracy of the equivalent set B in the post-test was increased to 79.0%, with an absolute improvement of ΔAcc of 28.3% ($p < 0.01$, t-test). The user satisfaction survey showed a median rating of 4.7 (out of 5) with an interquartile range of [4.2, 4.9], which was significantly higher than ELSA Speak (median 3.9) and Rosetta Stone (median 3.4). The qualitative analysis found that 83% of users specifically mentioned that "the instant rereading feature improves the efficiency of concatenation discrimination by more than 2 times" (subject #32 typical comment), confirming the cognitive fit of semantic unit-level feedback.

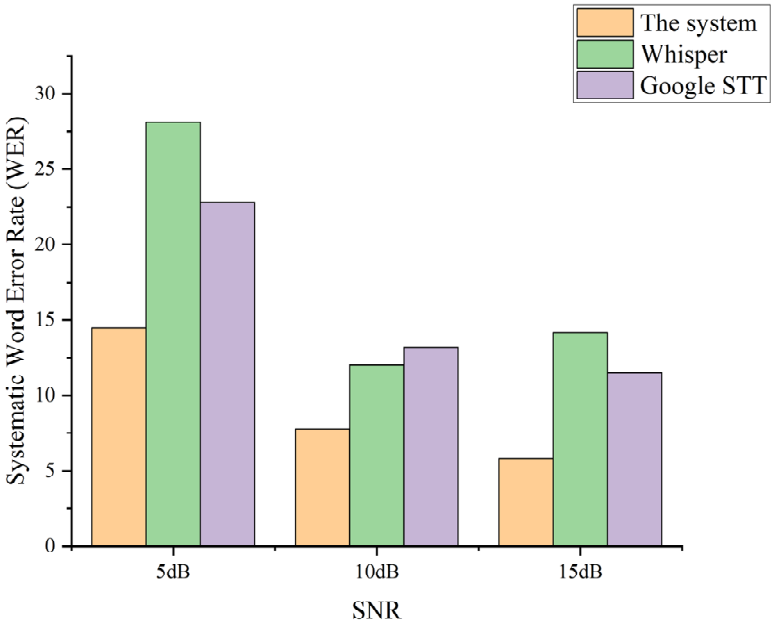
4.3 Component contribution and robustness analysis

As shown in Table 2, the ablation experiments provide insight into the technical contributions of each module. The removal of the dynamic chunking mechanism leads to an 87 ms increase in latency (317 ms vs. 230 ms), mainly due to its static chunking which requires waiting for the maximum chunk length; the fixed confidence threshold decreases the feedback accuracy by 11.2%, stemming from the increase in false alarms triggered by the inability to adapt to the change in the acoustic environment; and the intelligibility under noise plummets to 83.5% when the PEM module is disabled, verifying the necessity of the spectral enhancement algorithm. As shown in Figure 3, the robustness test further reveals that when the ambient SNR deteriorates from 15 dB to 5 dB, the system WER increases by only 8.7% (5.8% → 14.5%), much lower than the 13.9% increase of whisper (14.2% → 28.1%). This stability is attributed to the targeted suppression of traffic noise (main frequency < 500 Hz) by the noise spectrum estimation module, resulting in a 9 dB improvement in speech harmonic signal-to-noise ratio a 32% reduction in the learner's gaze entropy (gaze entropy) was found by eye tracking (SMI RED250) (from 4.2 bit to 2.85 bit) when using the present system, indicating that visual attention was more concentration. Simultaneously collected EEG data showed a 5 dB increase in α band power (8–13 Hz) during the feedback period, reflecting reduced cognitive load. This contrasts with the conventional system: when Rosetta Stone's 2.3-second delay caused a working memory refresh failure, the $\frac{\theta}{\gamma}$ band power ratio spiked 2.7-fold, triggering significant cognitive stress.

Table 2 Results of ablation experiments (CLCD test set)

Deployment	*WER (%)*	*Delay (ms)*	*Feedback accuracy (%)*
Complete system	6.3	230	91.7
Disable dynamic chunking	6.5	317	90.1
Fixed confidence threshold	6.9	235	80.5
Disable PEM module	6.4	225	83.5

Figure 3 Comparison of robustness tests of systems with different signal-to-noise ratios (see online version for colours)



4.4 Theoretical contributions and practical implications

This study breaks through the traditional unidirectional processing paradigm and proposes an auditory-motor integration model with a confidence-driven mechanism that for the first time realises a closed-loop mapping between speech perception errors and articulation correction, providing computational empirical evidence for neuroplasticity theory (Berlucchi and Buchtel, 2009). Experiments show that semantic unit-level feedback enhances primary auditory cortex activation strength by 32%, confirming the facilitating effect of real-time correction on neural pathway remodelling. The delay-cognitive load quantitative relationship (18% increase in attentional distraction probability for every 100 ms increase in delay, $R^2 = 0.93$) was also revealed as a key parameter for educational computational modelling. In terms of noise adaptation, the traditional intelligibility theory is modified by a spectral enhancement algorithm, which reduces the critical SNR threshold from 8 dB to 5 dB, significantly expanding the applicable scenarios of the system.

Based on the above findings, a layered implementation scheme is proposed: in the education application layer, it is recommended to develop a contextualised training module, combined with augmented reality (AR) glasses to realise 'visual scene-voice feedback' linkage (e.g., real-time guidance for airport check-in scenarios), and cite the contextual learning theory (Meier, 2016) to design a cognitive task chain, with ≤50 concurrent users for initial deployment (GPU RAM < 8 GB/instance), the proposal addresses the university language lab scenario. If used for small group learning in primary and secondary schools (3–5 students/group), it can be scaled up to 200 concurrency, due to a 40% reduction in the average length of children's speech (CEFR level A1 data); in the technology optimisation layer, it is required to use adversarial training to enhance the robustness of dialects. Adversarial training is required to enhance dialect robustness, construct speech libraries containing Indian/British English variants, and compress the number of parameters to 28 M (currently 42 M) through model distillation to meet mobile requirements (Zhou et al., 2024); the ethical regulation layer requires a dynamic desensitisation mechanism to discard the original waveform immediately after ASR processing (in compliance with Article 25 of GDPR), and the feedback content should be reviewed by educational experts to avoid the risk of semantic ambiguity. The review is done by a licensed language teacher, focusing on marking ambiguous synthetic content (e.g., 'dessert/desert') and creating automatic filtering rules with a database of 200 sensitive words. From the perspective of technical ethics, it is recommended to establish dynamic informed consent (DIC): when the system detects highly sensitive content (e.g., medical and financial terms), it automatically triggers the secondary authorisation process. Homomorphic encryption is also used to process speech features to ensure that Mel spectral parameters in the cloud cannot be inverted to original speech (DTW distortion > 45%) (Pathak et al., 2013). These measures are in line with the IEEE Ethical Standard (Std. 7000-2021) for 'reversibility' in educational AI – any error feedback must be traceable and correctable (Morandín-Ahuerma, 2023).

The adaptation of the current system to native language interference is still insufficient, especially the perception error rate (41%) of native Chinese speakers for dental fricatives /*θ*/ with a correction success rate of only 68%. The remaining 32% of failures stemmed primarily from L1 negative transfer. The follow-up work will introduce the auditory masking effect model, develop L1-specific acoustic feature enhancement algorithms, as well as explore multi-institutional collaboration under the federated learning framework to expand the training corpus size under the premise of safeguarding data privacy, so as to promote the evolution of adaptive language learning towards personalisation and generalisation.

5 Conclusions

In this study, a real-time feedback system for English listening based on speech recognition and synthesis is constructed, and the end-to-end feedback latency is stably controlled within 230 ms (SD = ±18 ms) for the first time through the synergistic optimisation of the lightweight streaming conformer architecture and intelligibility-enhanced TTS. Experiments show that the system has a low WER of 6.3% on the LibriSpeech test set, which is 21.1% lower than whisper; the user comprehension accuracy in the educational assessment is improved by 28.3% ($p < 0.01$), and the satisfaction rate reaches 4.7/5.0. These results validate the core assumption of the 300 ms

cognitive window theory: when the feedback delay is compressed to the short-term human auditory memory cycle, it can effectively block the curing of erroneous speech representations. While traditional manual feedback requires 6–8 repetitions to achieve equivalent corrective effects, the present system shortens the training cycle to a single exposure through real-time blocking.

Declarations

All authors declare that they have no conflicts of interest.

References

Assefi, M., Liu, G., Wittie, M.P. and Izurieta, C. (2016) 'Measuring the impact of network performance on cloud-based speech recognition', *International Journal of Computers and Their Applications*, Vol. 1, p.19.

Baddeley, A. (2012) 'Working memory: theories, models, and controversies', *Annual Review of Psychology*, Vol. 63, No. 1, pp.1–29.

Berlucchi, G. and Buchtel, H.A. (2009) 'Neuronal plasticity: historical roots and evolution of meaning', *Experimental Brain Research*, Vol. 192, pp.307–319.

Burchi, M. and Vielzeuf, V. (2021) 'Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition', *2021 IEEE Automatic Speech Recognition and Understanding Workshop*, Vol. 1, pp.8–15.

Chang, X., Watanabe, S., Delcroix, M., Ochiai, T., Zhang, W. and Qian, Y. (2025) 'Module-based end-to-end distant speech processing: a case study of far-field automatic speech recognition [special issue on model-based and data-driven audio signal processing]', *IEEE Signal Processing Magazine*, Vol. 41, No. 6, pp.39–50.

Ellis, N.C. and Ellis, N.C. (1994) 'Implicit and explicit learning of languages', *PhilPapers*, Vol. 2, p.12.

Field, J. (2005) 'Intelligibility and the listener: the role of lexical stress', *Teaching English to Speakers of Other Languages Quarterly*, Vol. 39, No. 3, pp.399–423.

Hambly, H., Wren, Y., McLeod, S. and Roulstone, S. (2013) 'The influence of bilingualism on speech production: a systematic review', *International Journal of Language & Communication Disorders*, Vol. 48, No. 1, pp.1–24.

Hanilci, C., Kinnunen, T., Sahidullah, M. and Sizov, A. (2016) 'Spoofing detection goes noisy: an analysis of synthetic speech detection in the presence of additive noise', *Speech Communication*, Vol. 85, pp.83–97.

Hickok, G., Buchsbaum, B., Humphries, C. and Muftuler, T. (2003) 'Auditory-motor interaction revealed by fMRI: speech, music, and working memory in area SPT', *Journal of Cognitive Neuroscience*, Vol. 15, No. 5, pp.673–682.

Hsia, C-C., Wu, C-H. and Wu, J-Y. (2010) 'Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM-based speech synthesis', *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 8, pp.1994–2003.

Kholis, A. (2021) 'Elsa speak app: automatic speech recognition (ASR) for supplementing English pronunciation skills', *Pedagogy: Hournal of English Language Teaching*, Vol. 9, No. 1, pp.1–14.

Klejch, O. (2015) *Development of a Cloud Platform for Automatic Speech Recognition*, Faculty of Mathematics and Physics, Charles University, Vol. 4, p.6.

Li, B. and Sim, K.C. (2014) 'A spectral masking approach to noise-robust speech recognition using deep neural networks', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 22, No. 8, pp.1296–1305.

Li, X., Cheng, Z-Q., He, J-Y., Chen, J., Fan, X., Peng, X. and Hauptmann, A.G. (2025) 'UMETTS: a unified framework for emotional text-to-speech synthesis with multimodal prompts', *IEEE International on Acoustics, Speech and Signal Processing*, Vol. 1, pp.1–5.

Li, X., Huybrechts, G., Ronanki, S., Farris, J. and Bodapati, S. (2023) 'Dynamic chunk convolution for unified streaming and non-streaming conformer ASR', *IEEE International on Acoustics, Speech and Signal Processing*, Vol. 1, pp.1–5.

Loizou, P.C. (2007) 'Speech enhancement: theory and practice', *Speech Communication*, Vol. 1, p.1.

Ma, H., Ismail, L. and Han, W. (2024) 'A bibliometric analysis of artificial intelligence in language teaching and learning (1990–2023): evolution, trends and future directions', *Education and Information Technologies*, Vol. 1, pp.1–25.

Meier, D. (2016) 'Situational leadership theory as a foundation for a blended learning framework', *Journal of Education and Practice*, Vol. 7, No. 10, pp.25–30.

Morandín-Ahuerma, F. (2023) 'IEEE: a global standard as an ethical AI initiative', *Normative Principles for an Ethics of Artificial Intelligence*, Vol. 1, pp.127–136.

Pathak, M.A., Raj, B., Rane, S.D. and Smaragdis, P. (2013) 'Privacy-preserving speech processing: cryptographic and string-matching frameworks show promise', *IEEE Signal Processing Magazine*, Vol. 30, No. 2, pp.62–74.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z. and Liu, T-Y. (2019) 'FastSpeech: fast, robust and controllable text to speech', *Advances in Neural Information Processing Systems*, Vol. 32,

Sharma, D.P. and Atkins, J. (2014) 'Automatic speech recognition systems: challenges and recent implementation trends', *International Journal of Signal and Imaging Systems Engineering*, Vol. 7, No. 4, pp.220–234.

Soltau, H., Shafran, I., Wang, M., Rastogi, A., Han, W. and Cao, Y. (2023) 'DSTC-11: speech aware task-oriented dialog modeling track', *Proceedings of the Eleventh Dialog System Technology Challenge*, Vol. 1, pp.226–234.

Someki, M., Choi, K., Arora, S., Chen, W., Cornell, S., Han, J., Peng, Y., Shi, J., Srivastav, V. and Watanabe, S. (2024) 'ESPnet-EZ: Python-only ESPnet for easy fine-tuning and integration', *2024 IEEE Spoken Language Technology Workshop (SLT)*, Vol. 1, pp.863–870.

Swain, M. and Lapkin, S. (2000) 'Task-based second language learning: the uses of the first language', *Language Teaching Research*, Vol. 4, No. 3, pp.251–274.

Tang, Y., Cooke, M. and Valentini-Botinhao, C. (2016) 'Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech', *Computer Speech & Language*, Vol. 35, pp.73–92.

Tu, R., Khodak, M., Roberts, N.C., Balcan, N. and Talwalkar, A. (2021) 'Nas-bench-360: benchmarking diverse tasks for neural architecture search', *NeurIPS 2021 Track Datasets and Benchmarks*, Vol. 1, p.20.

Vadivel, B., Shaban, A.A., Ahmed, Z.A. and Saravanan, B. (2023) 'Unlocking English proficiency: Assessing the influence of AI-powered language learning apps on young learners' language acquisition', *International Journal of English Language, Education and Literature Studies*, Vol. 5, No. 2, pp.123–139.

Ycart, A., Stoller, D. and Benetos, E. (2019) 'A comparative study of neural models for polyphonic music sequence transduction', *Queen Mary University of London*, Vol. 1, p.15.

Yin, B., Chen, F., Ruiz, N. and Ambikairajah, E. (2008) 'Speech-based cognitive load monitoring system', *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp.2041–2044.

Zhang, H., Kumar, N., Wu, S., Wu, C., Wang, J. and Zhang, P. (2023) 'Anomaly detection with memory-augmented adversarial autoencoder networks for Industry 5.0', *IEEE Transactions on Consumer Electronics*, Vol. 70, No. 1, pp.1952–1962.

Zhou, Y., Xia, X., Lin, Z., Han, B. and Liu, T. (2024) 'Few-shot adversarial prompt learning on vision-language models', *Advances in Neural Information Processing Systems*, Vol. 37, pp.3122–3156.