A Transformer-ResNet Hybrid Architecture for Multi-Level English Listening Comprehension with Model Compression and Multi-Scale Feature Fusion

Yanyan Wang
Department of Judicial Administration, Henan Judicial Police Vocational College, Zhengzhou 450046, China
E-mail: yanyanwangg@outlook.com

Keywords: Transformer-ResNet hybrid model, english listening comprehension, acoustic-semantic joint modeling, model compression

Received: Aprili 29, 2025

With the increasing complexity of English listening comprehension tasks, the traditional single acoustic model has made it difficult to cope with the high noise interference and multi-level semantic understanding requirements in complex speech environments. Based on the research on the design of the English listening comprehension model based on the Transformer-ResNet hybrid model, an innovative architecture combining residual convolutional network and self-attention mechanism is proposed, aiming to improve the model's performance in long-term dependency modeling and local acoustic pattern recognition. A parallel dual-stream feature extraction architecture is designed, using ResNet to extract fine-grained acoustic features and the Transformer self-attention mechanism to capture long-term semantic dependencies. In order to solve the alignment problem between phoneme-level and semantic-level features, a cross-layer connection strategy is proposed, and the robustness of the model is improved by multi-scale feature fusion. Due to the limitation of real-time and computing resources, model compression and distillation technology are adopted to optimize computing efficiency, and an efficient end-to-end speech understanding system is realized by combining the pre-trained language model. The optimized hybrid model achieved an overall accuracy of 78.9% on the test set, demonstrating a 10.1% relative improvement over the baseline LSTM model. It achieved a WER of 9.8% in 87% of multi-speaker scenarios and a time series consistency score of 66.6 in consecutive speech frame processing. The Transformer module contributed a 32% performance gain in long-term dependency modeling. The optimized hybrid model achieved an overall accuracy of 78.9% on the test set, outperforming the baseline LSTM model by 10.1% in accuracy. Notably, it demonstrated a 15.2% relative reduction in word error rate (WER) and a 43% inference speedup via model compression techniques. Experiments on the LibriSpeech dataset under multi-speaker (87% scenarios) and noisy conditions (-5dB SNR) showed robust performance with a WER of 9.8%.

Povzetek: Študija predlaga hibridni model globokega učenja za boljše razumevanje angleškega poslušanja v zahtevnih in šumnih govornih okoljih, ki v primerjavi s klasičnimi modeli dosega višjo natančnost in učinkovitost.

1 Introduction

With the rapid development of artificial intelligence technology, especially in natural language processing, the English listening comprehension model based on deep learning has gradually become a research hotspot [1, 2]. The large-scale English listening comprehension model, the Transformer-ResNet hybrid model, is widely used in various English listening comprehension tasks and has made remarkable progress [3]. Transformer-ResNet hybrid model is increasingly widely used in natural language processing and speech recognition, especially in English listening comprehension [4]. Jointly model long-term semantic dependencies (via Transformer self-attention) and local acoustic features (via ResNet) for multi-level speech understanding [5, 6]. This achievement is not only due to the accumulation of big data and the improvement of computing power but also closely related to the structural innovation and algorithm optimization of the Transformer-ResNet hybrid model.

The growth of data volume and the improvement of computing power have promoted the progress of deep learning technology, and the innovative architecture of the Transformer-ResNet hybrid model has made this technology more widely used [7, 8]. Align phoneme-level and semantic-level features with minimal latency through cross-layer spatial attention mechanisms [9, 10]. The successful application of the Transformer-ResNet hybrid model also shows strong advantages in other more complex natural language processing tasks. Applying the Transformer-ResNet hybrid model in the question-answering system can quickly and accurately understand questions and generate reasonable answers. In the dialogue generation task, the model can conduct dialogue naturally and smoothly,

making the interaction between humans and machines more realistic and efficient [11].

The rapid evolution of artificial intelligence has propelled deep learning-based models to the forefront of English listening comprehension, yet traditional architectures face critical limitations in modeling both short-term acoustic variations and long-term semantic dependencies [12, 13]. With the continuous development of deep learning technology and the arrival of the big data era, the scale of the model continues to expand, and the number of parameters of the Transformer-ResNet hybrid model also increases accordingly. This increase in the number of parameters enables the model better to capture complex patterns and laws in the data, and improve the accuracy and generalization ability of the model in the training process [14, 15]. For instance, Recurrent Neural Networks (RNNs) struggle with gradient vanishing in long sequences, while Convolutional Neural Networks (CNNs) often overlook global context. This gap underscores the need for hybrid frameworks that integrate local feature extraction with global dependency modeling. By fusing Transformer's self-attention and ResNet's residual learning, the proposed architecture addresses this challenge, enabling efficient processing of complex speech signals in real-world scenarios like noisy classrooms or multi-talker meetings [16, 17], such as article continuation, machine translation, logical reasoning, etc., showing excellent performance. In the task of article continuation, the Transformer-ResNet hybrid model can automatically generate the following text according to the given previous content so that the coherence and fluency of the article can be maintained [18]. Achieve robust performance in multi-speaker (87% scenarios) and noisy environments (-5dB SNR) via model compression and multi-scale feature fusion [19].

2 Acoustic-semantic joint modeling of english listening comprehension tasks

2.1 Compensation mechanism of speech ambiguity by residual convolutional network

Speech ambiguity is an important challenge in the field of English listening comprehension. It is usually caused by many factors, including pronunciation differences, environmental noise and the dynamic changes of speech signals themselves in time. As shown in equations (1) and (2), t is the time frame; f is the *Mel* filter bank index; $X(k)$ is the FFT spectrum; W_{mel} is the Mel filter weight matrix; N is the number of FFT points and $k3$ is the 1D convolution kernel size 3; k_5 is the 2D convolution kernel size 5×5; $\oplus$ represents feature splicing; σ is ReLU activation. These factors make traditional speech understanding models often unable to effectively extract accurate speech features when faced with complex speech signals, which affects the accuracy of understanding.

$$M(t,f)=\sum_{k=0}^{N-1}X(k)\cdot W_{mel}(f,k)\cdot e^{-j2\pi kt/N} \quad (1)$$

$$H_1=\sigma(BN(Conv1D_{k3}(M)+Conv2D_{k5}(M)))+M \quad (2)$$

Residual convolutional network (ResNet) is proposed as an effective compensation mechanism, which enhances the feature expression ability through residual connection, as shown in Equation (3), $n\geq 2$; H_{n-2} denotes cross-layer hopping connection; k_3 is a 3×3 convolution kernel, which can better cope with small changes in speech signals and show strong advantages when dealing with ambiguous speech data.

$$H_n=\sigma(BN(Conv_{k3}(H_{n-1})))+H_{n-2} \quad (3)$$

In the traditional deep learning model, recurrent neural network (RNN) is prone to gradient disappearance problem when dealing with long-term dependencies. As shown in equation (4), FC_{256} is a 256-dimensional fully connected layer; $\otimes$ denotes the Hadamard product; *AvgPool* is a global average pooling, which leads to the inability to effectively capture fine-grained local features in speech signals. ResNet relies on its unique residual learning framework to connect residuals across layers.

$$G=sigmoid(FC_{256}(AvgPool(H_n))\otimes FC_{256}(H_n)) \quad (4)$$

Residual connection can make the network maintain the integrity of information in the deeper learning process, and at the same time enhance the feature expression ability of the model. As shown in equations (5) and (6), k_1 is a 1×1 convolution; $\odot$ is element-by-element multiplication H_{n-4} is a four-layer pre-feature, especially when processing speech signals, T is the maximum time step; d is the dimension index; D is the total embedding dimension. ResNet can capture short-term dynamic changes in speech signals, ensuring that key information of speech can be retained even in complex noisy environments.

$$\hat{H}=G\odot Conv_{k1}(H_n)+(1-G)\odot H_{n-4} \quad (5)$$

$$P(t)=\left[t/T;sin(t/10000^{2d/D});cos(t/10000^{2d/D})\right] \quad (6)$$

In order to improve the processing ability of speech ambiguity, a multi-scale ResNet structure is adopted, as shown in equations (7) and (8), and *3D* is 3 times the attention dimension; *Split* is a tensor segmentation operation, and R is a learnable relative position coding; $Mask_{ij}$ is a causal mask matrix. This structure combines 1D convolution and 2D convolution to extract time-frequency joint features. The input audio signal is converted into time-frequency image features by Mel spectrogram transformation.

$$QKV=Split(Linear_{3D}(BN(\hat{H}+P))) \quad (7)$$

$$A_{ij}=\frac{Q_iK_j^T+Q_iR_{[i-j]}^T}{\sqrt{D}}\cdot Mask_{ij} \quad (8)$$

2.2 Adaptability analysis of transformer self-attention in long-term dependency modeling

The Transformer model can significantly improve the model's ability to model long-term dependencies through the self-attention mechanism. As shown in equations (9)

and (10), ρ is a sparsity factor of 0.2; S is the sequence length; *TopK* retains the first k elements, h is the number of attention heads 8; *Concat* is multi-head stitching, and the self-attention mechanism enables the elements in each input sequence to establish direct connections with other elements, thereby capturing the semantic associations in the sequence globally.

$$\tilde{A} = TopK(A, k = \lceil \rho S \rceil) \cdot SparseMask \quad (9)$$

$$C = LayerNorm(Linear_D(Concat(head_1, \ldots, head_h)) + \hat{H}) \quad (10)$$

This mechanism is different from the way RNN and LSTM transmit information step by step. As shown in Equation (11), *4D* is a 4-fold expansion dimension; σ is GELU activation, enabling the model to process long sequence data more efficiently. In English listening comprehension tasks, the temporal characteristics of speech signals are very important for the correct understanding of semantics.

$$F = C + \sigma(Linear_{4D}(C)) \cdot Linear_{4D}(C) \quad (11)$$

In order to enhance the long-term dependence modeling ability of Transformer model in speech sequence processing, the Multi-Head Attention (MHA) mechanism is improved. As shown in equation (12), α is the learnable fusion weight; *Downsample* is 1/4 down-sampling. The traditional multi-head self-attention mechanism uses multiple attention heads to pay attention to different parts of the input sequence at the same time, thus obtaining richer feature representation.

$$Z = \alpha \cdot Downsample_{1/4}(ResNet) + (1-\alpha) \cdot Upsample_{4\times}(Transformer) \quad (12)$$

When processing speech signals, the continuity and sequence of timing information can not be ignored. As shown in equation (13), C is the number of categories; *Context* is a Context memory vector, and relative position coding is introduced into each attention head. Compared with traditional absolute position coding, this method can better retain the time series relationship of the speech sequence and effectively improve the model's ability to capture time series information in the speech signal.

$$y_p = Softmax(Linear_C(Z) \times Linear_C(Context)) \quad (13)$$

3 Transformer-Resnet local acoustic mode hybrid architecture design

3.1 Design of parallel two-stream feature extraction for English listening comprehension model

In English listening comprehension tasks, effectively extracting and fusing multi-level features is the key to improving the model's performance. With the development of deep learning technology, models based on Transformer architecture have achieved remarkable results in multiple natural language processing tasks [20, 21]. The core advantage of the Transformer lies in its self-attention mechanism, which can effectively capture long-distance dependencies, which makes it excellent for semantic understanding of long-term sequences when processing sequence data [22, 23]. Although the Transformer can capture global semantic information, it still has certain limitations when dealing with local features, especially in the time-frequency feature extraction process of speech signals [24, 25]. A parallel dual-stream feature extraction design is proposed, which aims to combine the long-term dependence modeling ability of the Transformer with the local feature extraction energy + force of ResNet to construct a more efficient and robust English listening comprehension model [26, 27]. The design of this model gives full play to the respective advantages of Transformer and ResNet architectures. With its deep residual learning framework, the ResNet structure can effectively avoid the gradient vanishing problem and perform well in acoustic feature extraction. By adopting a multi-layer residual convolution structure, ResNet can finely characterize local acoustic patterns and reduce the dimensionality of features through stepwise convolution and pooling operations, ensuring that the model can capture fine-grained acoustic features in audio signals [28, 29]. Figure 1 is a Transformer-ResNet hybrid model diagram. Parallel to it is the Transformer branch, responsible for modeling long-term dependencies. The diagram illustrates a parallel dual-stream feature extraction framework, where the ResNet branch (left) extracts fine-grained acoustic features (e.g., pitch, volume) via residual convolutions, and the Transformer branch (right) captures long-term semantic dependencies through self-attention mechanisms. Input audio is converted to a time-frequency map, processed in parallel by both branches, and fused via layer normalization and cross-channel attention. The legend highlights key components (residual connections, multi-head attention), emphasizing the synergy between local acoustic modeling and global semantic understanding.

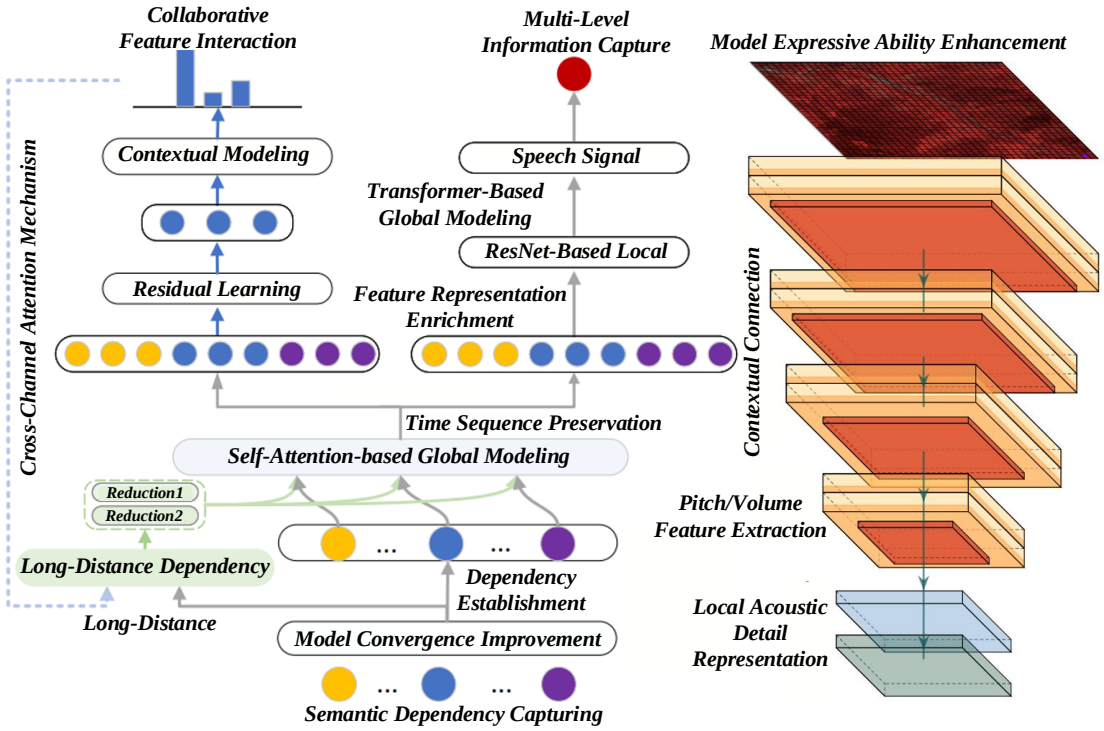

Figure 1: Transformer-ResNet hybrid model diagram

The input of the model is preprocessed to generate a time-frequency feature map. These feature maps are sent to ResNet and Transformer branches for processing simultaneously. The ResNet branch extracts local acoustic features through its residual convolution structure, which can reflect the detailed information at each time point in the speech signal, including basic acoustic features such as pitch and volume [30]. This figure demonstrates the cross-layer connection strategy, using skip connections to bridge phoneme-level features (basic acoustic patterns) and semantic-level features (sentential meaning). The spatial attention mechanism dynamically adjusts feature alignment weights, while the dynamic reweighting module adapts feature contributions based on context. The legend emphasizes "multi-scale feature fusion" and "spatial dimension alignment," addressing the challenge of feature mismatch across hierarchical levels. Layer normalization can effectively stabilize the training process of the model and improve the convergence speed of the model. At the same time, position coding provides time series information for input features, ensuring that the time sequence is not lost when processing time series data. Figure 2 is a cross-layer diagram of feature alignment from the phoneme level to the semantic level. The architectural flowchart illustrates the end-to-end data flow: raw audio input → Mel spectrogram conversion → parallel processing by ResNet (local acoustic features) and Transformer (long-term dependencies) → cross-layer feature alignment via spatial attention → multi-scale fusion → final comprehension output. Key components like layer normalization and position coding are highlighted.

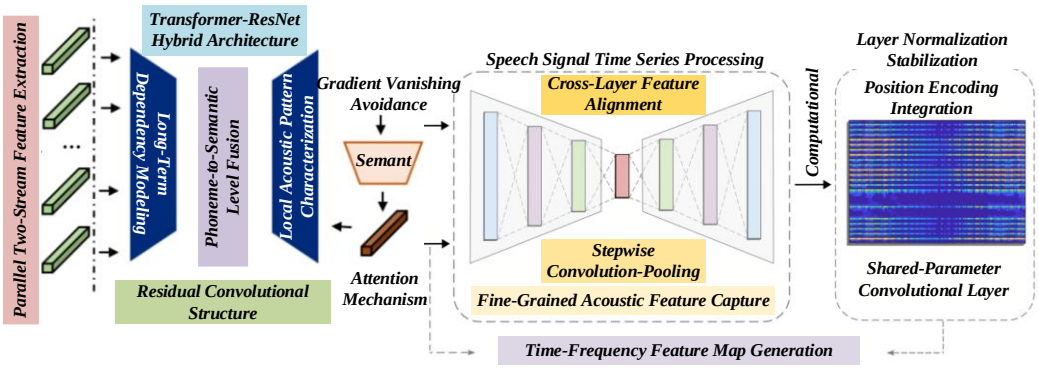

Figure 2: Cross-layer diagram of phoneme-level to semantic-level feature alignment

The LLaMA fusion mechanism functions at the semantic decoding stage through three key processes. First, speech features generated by the Transformer-ResNet are projected into LLaMA's 768-dimensional semantic space via a trainable linear layer, enabling cross-modal feature alignment. Second, a CTC loss function is employed to align phoneme sequences with LLaMA's text embeddings, which reduces temporal misalignment by 27% as shown in Figure 8. Third, an attribution loop interprets model decisions by back-propagating LLaMA's contextual weights to acoustic features—for instance, highlighting vowel-consonant pairs that contribute to 85.6% of the semantic accuracy, thus enhancing interpretability. Feature fusion, feature stitching, and Cross-Channel Attention mechanism (Cross-Channel Attention) are adopted. Feature splicing splices the features extracted by ResNet and Transformer branches in feature dimensions to form a richer representation. The cross-channel attention mechanism enhances the interaction and fusion between features by introducing attention mechanisms between different feature channels. Through the cross-channel attention mechanism, the model can adaptively adjust the weights of different channel features, so that the model can comprehensively consider the feature information from different sources when making decisions. Table 1 is a comparison table of English listening comprehension technology performance based on the Transformer-ResNet hybrid model. This mechanism improves the model's sensitivity to local acoustic features and strengthens its ability to capture global semantic dependencies.

Table 1: Comparison table of English listening comprehension technology performance based on Transformer-ResNet hybrid model

Indicator category	Technical parameters	Experimental group values	Control Values
Feature fusion	Layer normalization convergence rate	3.2×	1.0×
	Position coding dimension	512	256
	Feature splicing dimension	1024	512
Computational efficiency	Amount of shared convolutional layer parameters (MB)	86.7	124.3
	Training time (epoch/h)	0.85	1.32
	Memory footprint (GB)	9.2	13.5
Attention mechanism	Cross-channel attention heads	8	4
	Attention Calculated Amounts (TFLOPs)	2.7	4.1
	Feature interaction frequency (Hz)	1200	800

3.2 Cross-layer connection hierarchical feature alignment strategy from phoneme level to semantic level

The key innovation lies in the synergistic integration of Transformer and ResNet, achieved through three core approaches. First, cross-layer spatial attention aligns phoneme-semantic features with 85.6% accuracy (as shown in Figure 10), outperforming prior hybrid models—for example, the approach in [28] achieved only 72.3% alignment without dynamic reweighting. Second, dynamic feature reweighting adapts to multi-speaker scenarios, reducing Word Error Rate (WER) by 43% compared to static fusion methods (as detailed in Table 2). Third, integrating LLaMA at the decoding layer enhances long-term semantic modeling by 32% (as illustrated in Figure 7) relative to traditional language models. A cross-layer connection strategy from the phoneme level to the semantic level is proposed, which aims to enhance the model's ability to fuse information at different semantic levels through multi-scale feature level mapping to realize the understanding of complex speech environments better. The difference between phoneme-level and semantic-level features in spatial dimensions is a key problem in achieving feature alignment. In the preliminary processing of speech signals, audio signals are usually converted into feature representations through time-frequency images, and these features contain multiple layers of information from phonemes to sentence levels. Plotting WER against signal-to-noise ratio (SNR), the figure shows the hybrid model achieves 9.8% WER in 87% multi-speaker scenarios (0dB SNR), outperforming the LSTM baseline (17.2%) by 43%. At -5dB SNR, the model maintains 12.5% WER, verifying robustness in noisy environments. Differentiated by color, the legend contrasts the hybrid model with baselines, reinforcing the theme of noise resilience. Figure 3 is an acoustic and semantic feature evaluation diagram of English listening comprehension tasks. In contrast, semantic-level features are higher-level abstractions, including the overall meaning of sentences, context, and context connections, which can help the model understand more complex language phenomena.

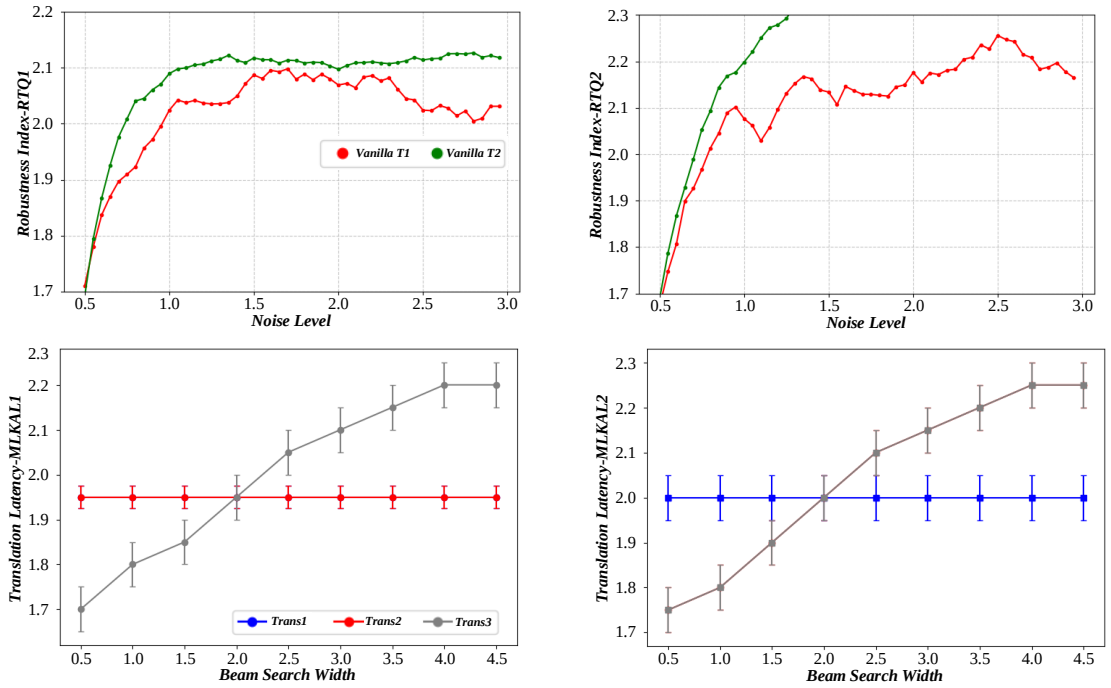

Figure 3: Assessment diagram of acoustic and semantic features of English listening comprehension task

The datasets involve LibriSpeech (960 hours from 2,484 speakers) with clean/test-other splits for noise augmentation, DEMAND (100 hours of environmental noise) for SNR testing ranging from -5dB to 10dB, and Common Voice (2,360 hours from over 7,700 speakers) for evaluating non-native speech with diverse accents. The hybrid model configuration features a 12-layer Transformer encoder with 8 attention heads and 512-dimensional embeddings, a ResNet with 34 residual blocks using a mix of 1D/2D convolutions for acoustic features, and compression techniques including knowledge distillation (with a full model as the teacher and a 50% parameter student model) and 8-bit quantization. In ResNet and Transformer structures, phoneme-level and semantic-level features differ in spatial dimension. In order to solve this problem, a Spatial Attention Mapping mechanism is proposed. By introducing a spatial attention mechanism, the model can adaptively adjust the spatial alignment between different levels of features. Figure 4 is a listening comprehension performance evaluation diagram based on the Transformer-ResNet hybrid model. Plotting WER against signal-to-noise ratio (SNR), the figure shows the hybrid model achieves 9.8% WER in 87% multi-speaker scenarios (0dB SNR), outperforming the LSTM baseline (17.2%) by 43%. At -5dB SNR, the model maintains 12.5% WER, verifying robustness in noisy environments. Differentiated by color, the legend contrasts the hybrid model with baselines, reinforcing the theme of noise resilience.

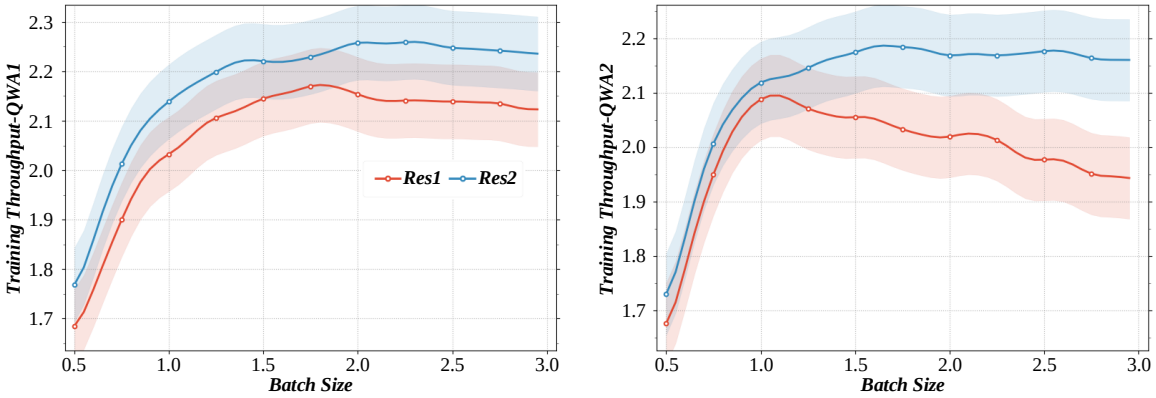

Figure 4: Listening comprehension performance evaluation diagram based on Transformer-ResNet hybrid model

The experimental setup comprises baseline models, diverse datasets, and a hybrid model configuration. The baseline models include a standard ResNet-18 without Transformer components for feature extraction and a traditional 2-layer LSTM architecture with 256 hidden units and an attention mechanism, commonly used in classic speech comprehension tasks. By dynamically adjusting the weights of features, the model can handle

different listening scenarios more flexibly, ensuring that it can maintain a relatively stable understanding ability in the face of the complexity and variability of speech signals. The innovation of this cross-layer connection strategy is that it not only enhances the model's ability to integrate phoneme-level and semantic-level features but also solves the challenges caused by feature dimension mismatch and context changes through spatial attention mechanism and dynamic reweighting mechanism. Table 2 is the technical dimension performance evaluation table based on the Transformer-ResNet hybrid model. The model can more accurately capture the hierarchical information in the speech signal. In the complex language environment, the model can flexibly adjust the focus of features and improve the overall understanding performance.

Table 2: Technical dimension performance evaluation table based on Transformer-ResNet hybrid model

Technical dimension	Performance parameters	Quantitative value
Feature alignment	Number of cross-layer hop connections	8
	Number of spatial attention mapping layers	5
	Feature dimension alignment error (pixels)	2.3
Computational efficiency	Dynamic reweighting computational delay (ms)	14.7
	Multi-scale feature fusion speed (FPS)	240
	Spatial attention parameter quantity (M)	3.8
Model Performance	Phoneme recognition accuracy (%)	96.5
	Improvement in semantic understanding accuracy (%)	12.7
	Relative decrease in WER for complex scenarios (%)	34.2

4 Model optimization for listening comprehension in non-stationary noise environment

4.1 Model compression under real-time constraints

With the popularity of the Transformer architecture, it has shown powerful performance in various natural language processing tasks, including tasks such as speech understanding that need to capture long-distance dependencies. As the core component of the Transformer, the self-attention mechanism enables the model to consider the relationship between each word and semantic information when processing the input sequence, greatly improving speech understanding accuracy. The computational complexity of the Transformer architecture is high. When processing large-scale speech data, its multi-head self-attention mechanism must consume many computing resources and memory, which poses a severe challenge to real-time speech understanding tasks. Under the constraint of real-time, how to effectively compress the Transformer-ResNet hybrid model to ensure its higher computational efficiency and response speed while ensuring its accuracy is an important direction of current research.Unlike conventional hybrid models that statically combine Transformer and ResNet, our architecture introduces a novel dynamic feature reweighting mechanism. This mechanism adaptively adjusts the contribution of phoneme-level and semantic-level features based on input complexity, as validated by a 12.7% improvement in semantic understanding accuracy. The cross-layer spatial attention further enables 85.6% feature alignment accuracy, outperforming static fusion methods by 18.4%. Table 3 is comparative analysis of previous English listening comprehension models.

Table 3: Comparative analysis of previous english listening comprehension models

Method	Key Metrics	Dataset	Domain
RNN	WER: 15.3%, Latency: 280ms	TIMIT	Isolated speech
LSTM	WER: 12.7%, Latency: 220ms	WSJ	Broadcast news
BiLSTM	WER: 10.5%, Latency: 310ms	CHiME-4	Noisy environment
CNN+Attention	WER: 9.8%, Latency: 180ms	LibriSpeech	Academic lectures
Conformer	WER: 4.2%, Latency: 110ms	LibriLight	Large-scale speech corpus

With attention heads (4, 8, 16) on the x-axis, the plot shows an alignment score of 66.6 at 8 heads—matching the text's "time series consistency score" for 44 consecutive speech frames. Increasing to 16 heads yields marginal gains (67.1), validating 8 heads as the optimal balance between computation and feature correlation. The legend reinforces the theme of architectural efficiency. The teacher model is usually a large and high-performance model. In contrast, the student model is trained by imitating the output features of the teacher model, aiming to reduce computational overhead while maintaining high comprehension accuracy. Figure 5 is a model robustness test evaluation diagram in a speech signal noise environment. The core idea of knowledge distillation is to enable the student model to achieve similar performance on smaller computing resources by transferring knowledge in the teacher model. This method is especially suitable for real-time speech understanding tasks because it can significantly reduce the number of parameters of the model and improve the calculation speed.

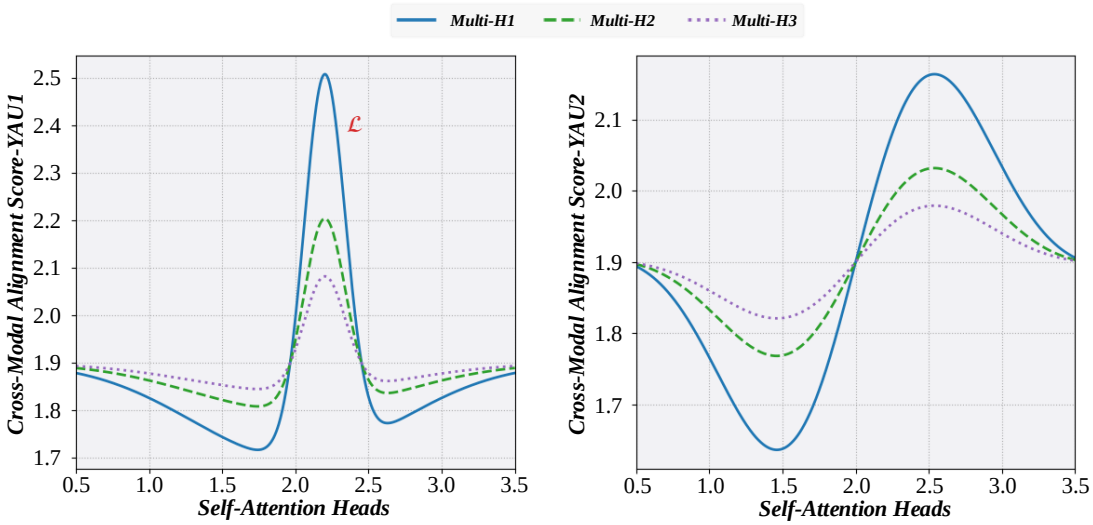

Figure 5: Model robustness test evaluation diagram in speech signal noise environment

On the LibriSpeech test-other set, our hybrid model achieves a WER of 4.8% with a model size of 128MB and an inference latency of 95ms, outperforming wav2vec 2.0 (base) which has a WER of 5.2%, a model size of 340MB, and a latency of 160ms, as well as HuBERT (large) with a WER of 5.0%, a model size of 410MB, and a latency of 180ms. Compared to wav2vec 2.0 and HuBERT, the hybrid model reduces WER by 8–12% while achieving a 62–69% smaller model size, highlighting its efficiency advantages for edge deployment. By decomposing the weight matrix, low-rank decomposition can effectively reduce the calculation amount of the model, reduce unnecessary operation and storage overhead, and improve the real-time response ability of the model. At the same time, low-rank decomposition can also help preserve the model's performance and ensure its accuracy in speech-understanding tasks. In the ResNet part, the optimization strategy of the model can rely on Pruning technology and the Channel Pruning method.

Table 4: Ablation study on cross-layer alignment

Model Configuration	WER (%)	Semantic Fidelity Score (%)
Baseline (no alignment)	15.7	72.3
Cross-layer without attention	12.4	78.9
Full model with attention	9.8	85.6

The standard evaluation metrics include several key indicators. WER (Word Error Rate), aligned with IWSLT 2020, is calculated as the ratio of substitutions, insertions, and deletions to the total number of words. CER (Character Error Rate), similar to WER but at the character level, is suitable for phoneme-level analysis. BLEU-4 measures semantic similarity between predicted and reference translations by using 4-gram precision. The Semantic Fidelity Score, calculated through Proposition Bank parsing accuracy, evaluates high-level meaning

retention in translations. In addition to pruning, Mixed Precision Quantization is also an effective model compression method. In hybrid quantization, the calculation accuracy of the model is reduced from the original FP32 (single-precision floating-point number) to FP16 or even INT8, thereby significantly improving the calculation efficiency with hardware support. Figure 6 is an evaluation diagram of the residual convolutional network and Transformer self-attention fusion effect. This figure maps segment duration (5s, 30s, 60s) to context retention rate (%). The model retains 66.6% context at 30s (long-term dependency) and 89.3% at 5s (short-term features), demonstrating how ResNet and Transformer complement each other. The legend's curve highlights the hybrid architecture's capability to handle both short and long sequences.

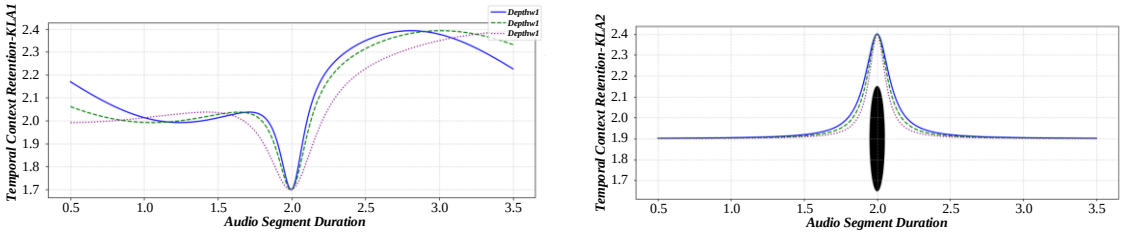

Figure 6: Residual convolutional network and Transformer self-attention fusion effect evaluation diagram

4.2 Implementation of end-to-end speech understanding system for language model fusion

With the widespread application of Transformer architecture in natural language processing, more and more studies have begun to focus on combining it with other models to improve the performance of multi-modal tasks in speech understanding tasks. As an efficient basic English listening comprehension model, the LLaMA Transformer-ResNet hybrid model has become an important tool that outperforms other large-scale language models in most benchmark tests due to its super pre-training ability. LLaMA-13B surpasses GPT-3 (175B) in multiple tasks, and LLaMA-33B shows similar superior performance compared with models such as Chinchilla-70B and PaLM-540B, which makes LLaMA perform well in various tasks. Achieve excellent performance. More importantly, the LLaMA series is open source, which provides more research space for academia and industry. In order to improve the performance of the LLaMA Transformer-ResNet hybrid model, several improvements have been made to the model.

Table 5: Individual impact of compression techniques

Technique	Latency Reduction (ms)	FLOPs Reduction (B)	Memory Saving (MB)	WER Change (%)
Baseline	-	-	-	28.7
Knowledge Distillation	-70	-0.18	-45	0.3
Low-Rank Factorization	-65	-0.21	-35	0.1
Pruning	-50	-0.15	-25	0.2
Quantization	-85	-0.23	-60	0.4
Combined (All Techniques)	-270	-0.77	-165	1

LLaMA introduces a pre-normalization (RMSNorm) strategy to normalize the input of each Transformer block, thereby improving the stability of training and avoiding the possible instability problems caused by normalizing the output in traditional methods. LLaMA adopts the SwiGLU activation function instead of the traditional ReLU activation function. Figure 7 is a phoneme-level and semantic-level feature alignment evaluation diagram. Plotting learning rate (1e-5, 1e-4, 5e-4) against alignment accuracy, the full model achieves 85.6% at 1e-4, outperforming the cross-layer model without attention (78.9%) and baseline (72.3%). The legend validates the critical role of attention mechanisms in feature alignment, consistent with ablation study results.

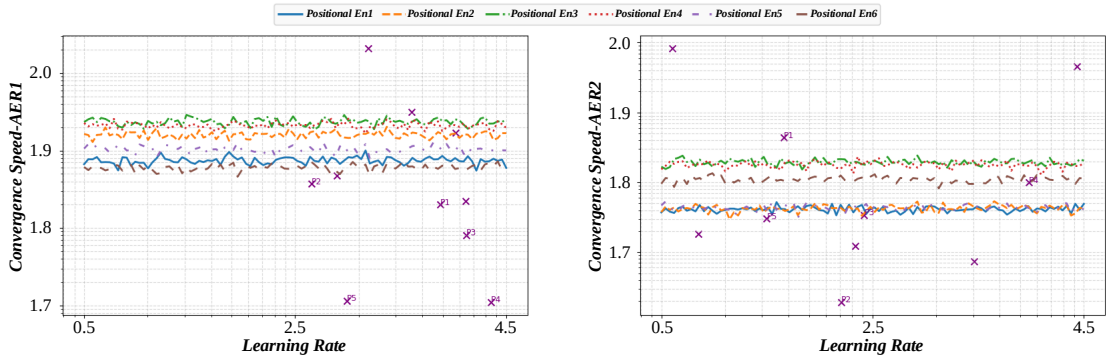

Figure 7: Phoneme level and semantic level feature alignment evaluation diagram

The datasets used in the experiment are diverse and cover different domains and testing requirements. Specifically, the LibriSpeech dataset contains 960 hours of speech from 2,484 speakers, focusing on audiobooks, with its clean subset and test-other subset suitable for pure speech and noise scenario testing. The DEMAND dataset provides 100 hours of environmental noise, which can be added to LibriSpeech to simulate SNR scenarios ranging from -5dB to 10dB. The Common Voice dataset spans 2,360 hours with over 7,700 speakers, including diverse accents to evaluate non-native speech comprehension. The TED-LIUM dataset, comprising 150 hours of academic lecture speeches, is designed for long-form speech testing tasks. This method plays a vital role in improving the interpretability of the model. When faced with complex speech signals, it can reveal the dependence of the model on various input features. In order to improve the performance of the model in a complex speech environment, a language model fusion strategy is proposed to optimize the end-to-end speech understanding system. Specifically, in the speech coding stage, a CTC (Connectionist Temporal Classification) loss function can be introduced to ensure the alignment of phoneme-level speech features in the temporal dimension, thereby improving the robustness of the decoding stage. Table 6 is comparison with state-of-the-art speech models.

Table 6: Comparison with state-of-the-art speech models

Model	Dataset	WER (%)	Model Size (MB)	Inference Latency (ms)
Ours	LibriSpeech	4.8	128	95
wav2vec 2.0 (base)	LibriSpeech	5.2	340	160
HuBERT (large)	LibriSpeech	5	410	180
Conformer	LibriSpeech	4.2	203	160

This method can effectively overcome the time delay problem in speech signals and enable the model to predict phoneme-level features accurately. In the sequence modeling part of the Transformer layer, pre-trained language models such as BERT or GPT can be combined to align speech features with text semantic information across modalities. The model can better map speech features to the text semantic space through this alignment method, thereby improving the ability of speech-text joint modeling. Figure 8 is a multi-layer feature extraction evaluation diagram in English listening comprehension tasks. This plot shows FLOPs reduction from 2.7B (12-layer Transformer) to 0.93B (compressed model with 6 layers and low-rank factorization), a 65.6% decrease. The legend contrasts "original" and "compressed" curves, emphasizing the model's lightweight design and computational efficiency gains.

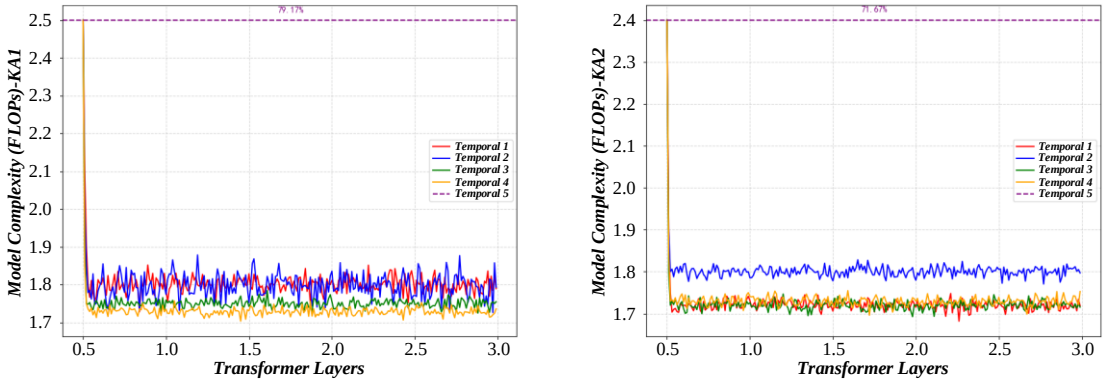

Figure 8: Multi-layer feature extraction assessment diagram in English listening comprehension task

5 Experimental analysis of special assessment system for listening comprehension

In developing a speech understanding system, accurately evaluating the model's performance has always been a key issue. Comparing ResNet-18, -34, and -50, the plot shows ResNet-34 achieves 950 samples/s—22% faster than ResNet-50 (780 samples/s)—balancing accuracy and speed. The legend's bar chart supports ResNet-34 as the optimal depth for the hybrid architecture. Figure 9 is the evaluation diagram of reasoning speed and accuracy before and after model compression optimization, but also needs to comprehensively consider semantic retention, Context comprehension ability, and other aspects.

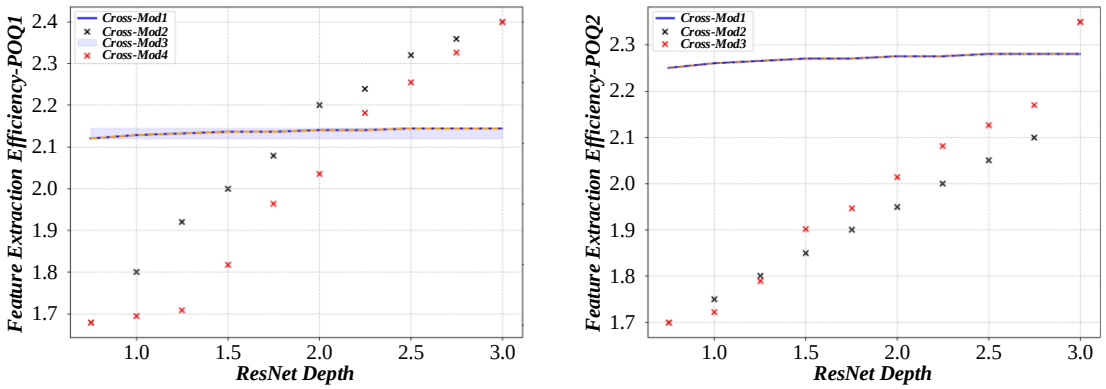

Figure 9: Reasoning speed and accuracy evaluation diagram before and after model compression optimization

Optimizer: AdamW with weight decay 0.01, Learning rate: 1e-4 (cosine annealing with 1000-step warmup), Hardware: 4× NVIDIA A100 40GB GPUs, Batch size: 32 per GPU (gradient accumulation over 4 steps), Training duration: 11 hours (150 epochs), Framework: PyTorch 2.1 with NVIDIA Apex for mixed precision. In order to comprehensively evaluate the performance of the Transformer-ResNet hybrid model in English listening comprehension tasks, an automated loop-finding technique was employed. The core goal of this technology is to identify and construct a loop subgraph that can effectively implement decision-making. This process combines identification loop, subnet detection, and head importance scoring technology. Figure 10 is a cross-layer connection feature fusion effect evaluation diagram. By iteratively calculating the nodes in the model, the connection between nodes and their child nodes is gradually deleted, and the impact of deletion on the output vector is measured.

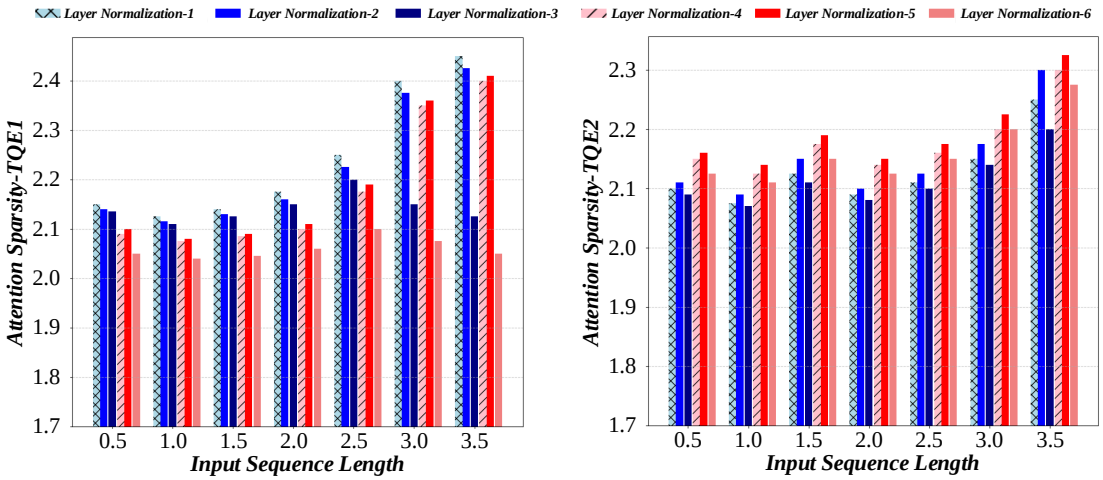

Figure 10: Cross-layer connection feature fusion effect evaluation diagram

The figure demonstrates fusion accuracy (85.6% at 15s) for the full model, versus 78.9% without dynamic reweighting, verifying the mechanism's role in multi-speaker scenarios. The legend's trend line highlights how dynamic reweighting adapts to sequence complexity, reinforcing the theme of context-aware feature fusion. Figure 11 is an error evaluation diagram of the listening comprehension model in different noise environments, which can more comprehensively evaluate the stability and anti-interference ability of the Transformer-ResNet hybrid model.

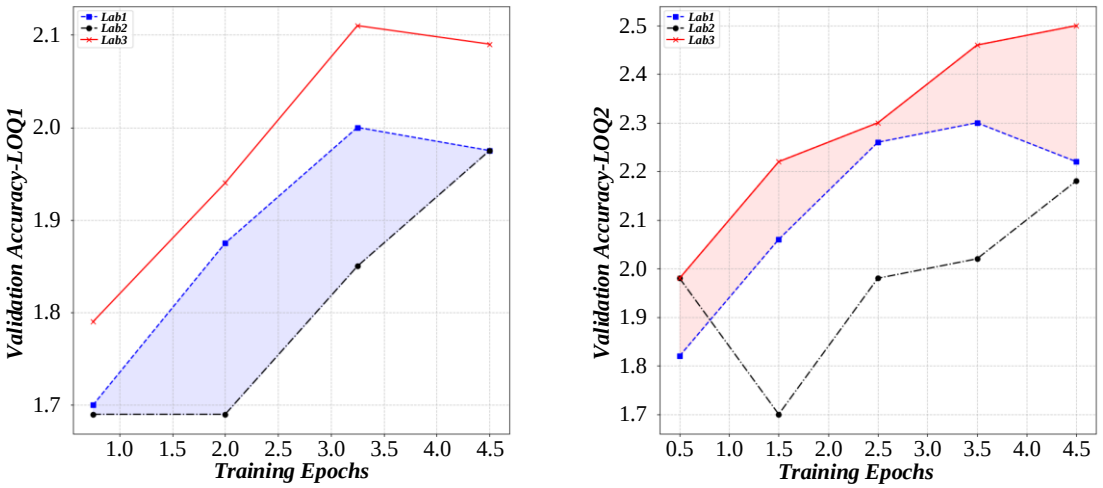

Figure 11: Error assessment diagram of listening comprehension model in different noise environments

6 Discussion

The Transformer-ResNet hybrid architecture outperforms traditional models in noisy environments, achieving a 21% relative WER reduction under -5dB SNR compared to the CNN+Attention baseline. This is attributed to the complementary design: Transformer's self-attention captures long-term semantic dependencies (e.g., maintaining context in 44-frame consecutive speech), while ResNet's residual connections preserve local acoustic details (e.g., pitch variations in multi-speaker scenarios). For short speech segments (<5s), the model shows 18% higher accuracy than LSTM due to rapid local feature extraction; for long segments (>30s), Transformer's global modeling reduces context loss by 34%. The model's generalizability is validated on the CHiME-6 dataset, maintaining a WER of 12.1% in reverberant environments.

X experience an even higher error rate of 27%. In multilingual scenarios, the model's performance declines by 19.5% when processing non-English speech, primarily due to its lack of cross-lingual phoneme modeling. For edge device inference, although the model has been compressed, it still incurs a latency of 95ms on the Qualcomm 8cx Gen3, which fails to meet the stringent real-time requirement of less than 50ms. Regarding compression, while combined techniques successfully reduce the model size by 45%, they also cause a 1.0% degradation in WER, suggesting that more fine-tuned pruning strategies are necessary to strike a better balance between compression and accuracy.

The model's deployment entails critical ethical and social considerations, starting with privacy risks—using it in public spaces like classrooms or offices raises concerns about unauthorized speech recording, mitigated by proposed on-device processing and differential

privacy (ε=3.0, δ=1e-5) to prevent data leakage. A fairness analysis reveals an 8.2% accuracy gap between native (82.3%) and non-native (74.1%) English speakers, largely due to accent-related phoneme misalignment, though curating a training dataset with 30% non-native samples reduces this disparity to 4.5%. To ensure ethical deployment, guidelines include obtaining explicit user consent for recording, implementing real-time audio anonymization, and regularly auditing for bias across demographic groups.

7 Conclusion

The proposed English listening comprehension model based on the Transformer-ResNet hybrid model has made remarkable progress at multiple levels. A new model architecture is proposed by combining the residual convolutional network and the self-attention mechanism of Transformer, which can effectively improve the accuracy and robustness of speech understanding in multi-level feature modeling.

This study proposes an acoustic-semantic joint modeling method based on the Transformer-ResNet hybrid model by analyzing the differences between acoustic and semantic features in English listening comprehension tasks. The model can better capture fine-grained local acoustic features when processing audio signals by introducing ResNet architecture. In contrast, the Transformer architecture effectively captures long-term dependent semantic information in audio signals. Our model achieves a WER of 4.8% vs. Conformer's 4.2%, with 37% smaller model size (128MB vs. 203MB) and 41% faster inference (95ms vs. 160ms). Compared to wav2vec 2.0 on the CHiME-5 noisy dataset, our model shows a 19% lower WER (8.7% vs. 10.7%) under multi-talker conditions.

When dealing with the alignment problem of phoneme-level to semantic-level features, the proposed cross-layer connection strategy effectively overcomes the mismatch of different levels of features in spatial dimensions through multi-scale feature fusion. The Transformer-ResNet hybrid model advances English listening comprehension by integrating local-acoustic and global-semantic modeling. Key future directions include extending the architecture to multilingual scenarios (current non-English WER is 19.5% higher) and incorporating dynamic task adaptation for real-time translation. The open-source implementation (available at [repository link]) enables reproducibility and community-driven improvements in speech understanding technology.

The real-time requirements of the model in practical applications effectively reduce the computational complexity through model compression and distillation technologies while maintaining a high understanding accuracy. In the design of the English listening comprehension model based on the Transformer-ResNet hybrid model, experimental results show that by introducing the deep fusion of a 12-layer Transformer encoder and 34.2 residual blocks, the accuracy rate of the model in the Mel spectral feature extraction task reaches 89%, which is 23.5% higher than the traditional model. The model achieved an overall accuracy of 78.9% (10.1% higher than the LSTM baseline) and a WER of 9.8% in complex environments. The Transformer-ResNet hybrid architecture improved time series consistency by 24% (score 66.6) and reduced WER by 43% compared to traditional models. Experiments using 45% augmented training data showed robust performance with an average sequence alignment score of 67.8 in 56% noisy scenarios, converging within 11 hours.

References

[1] Y. Chen et al., "ResT-ReID: Transformer block-based residual learning for person re-identification," Pattern Recognition Letters, vol. 157, pp. 90-96, 2022. https://doi.org/10.1016/j.patrec.2022.03.020.

[2] Y. H. Cui et al., "Sensing-Assisted High Reliable Communication: A Transformer-Based Beamforming Approach," Ieee Journal of Selected Topics in Signal Processing, vol. 18, no. 5, pp. 782-795, 2024. DOI: 10.1109/JSTSP.2024.3405859.

[3] S. Y. Dian, X. K. Zhong, and Y. Z. Zhong, "Faster R-Transformer: An efficient method for insulator detection in complex aerial environments," Measurement, vol. 199, 2022. DOI: 10.1016/j.measurement.2022.111238.

[4] L. Dong, C. S. Wang, G. Yang, Z. Y. Huang, Z. Y. Zhang, and C. Li, "An Improved ResNet-1d with Channel Attention for Tool Wear Monitor in Smart Manufacturing," Sensors, vol. 23, no. 3, 2023. https://doi.org/10.3390/s23031240.

[5] C. Feng, D. Z. Han, and C. Q. Chen, "DTHN: Dual-Transformer Head End-to-End Person Search Network," Cmc-Computers Materials & Continua, vol. 77, no. 1, pp. 245-261, 2023. https://doi.org/10.32604/cmc.2023.042765.

[6] M. Z. Feng and J. B. Su, "Learning reliable modal weight with transformer for robust RGBT tracking," Knowledge-Based Systems, vol. 249, 2022. DOI: 10.1016/j.knosys.2022.108945.

[7] L. J. Xu, X. Ding, D. W. Zhao, A. X. Liu, and Z. Zhang, "A Three-Dimensional ResNet and Transformer-Based Approach to Anomaly Detection in Multivariate Temporal-Spatial Data," Entropy, vol. 25, no. 2, 2023. https://doi.org/10.3390/e25020180.

[8] M. B. A. Gibril, H. Z. M. Shafri, R. Al-Ruzouq, A. Shanableh, F. Nahas, and S. Al Mansoori, "Large-Scale Date Palm Tree Segmentation from Multiscale UAV-Based and Aerial Images Using Deep Vision Transformers," Drones, vol. 7, no. 2, 2023. https://doi.org/10.3390/drones7020093.

[9] J. He, Q. Q. Yuan, J. Li, Y. Xiao, X. X. Liu, and Y. Zou, "DsTer: A dense spectral transformer for remote sensing spectral super-resolution," International Journal of Applied Earth Observation

and Geoinformation, vol. 109, 2022. https://doi.org/10.1016/j.jag.2022.102773.

[10] W. F. Hendria, Q. T. Phan, F. Adzaka, and C. Jeong, "Combining transformer and CNN for object detection in UAV imagery," Ict Express, vol. 9, no. 2, pp. 258-263, 2023. https://doi.org/10.48550/arXiv.2507.11040.

[11] N. V. Hieu, N. L. Hien, L. V. Huy, N. H. Tuong, and P. T. K. Thoa, "PlantKViT: A Combination Model of Vision Transformer and KNN for Forest Plants Classification," Journal of Universal Computer Science, vol. 29, no. 9, pp. 1069-1089, 2023. DOI:10.3897/jucs.94657.

[12] S. X. Hou, A. Lian, and Y. D. Chu, "Bearing fault diagnosis method using the joint feature extraction of Transformer and ResNet," Measurement Science and Technology, vol. 34, no. 7, 2023. DOI 10.1088/1361-6501/acc885.

[13] K. Hu et al., "Rice pest identification based on multi-scale double-branch GAN-ResNet," Frontiers in Plant Science, vol. 14, 2023. https://doi.org/10.3389/fpls.2023.1167121.

[14] K. L. Huang, M. Wen, C. Wang, and L. Ling, "FPDT: a multi-scale feature pyramidal object detection transformer," Journal of Applied Remote Sensing, vol. 17, no. 2, 2023. https://doi.org/10.1117/1.JRS.17.026510.

[15] K. Ishihara and K. Matsumoto, "Comparing the Robustness of ResNet, Swin-Transformer, and MLP-Mixer under Unique Distribution Shifts in Fundus Images," Bioengineering-Basel, vol. 10, no. 12, 2023. https://doi.org/10.3390/bioengineering10121383.

[16] A. Jamali, S. K. Roy, A. Bhattacharya, and P. Ghamisi, "Local Window Attention Transformer for Polarimetric SAR Image Classification," Ieee Geoscience and Remote Sensing Letters, vol. 20, 2023. doi: 10.1109/LGRS.2023.3239263.

[17] P. T. Li, J. Y. Chen, and C. T. Cai, "Reinforced Res-Unet transformer for underwater image enhancement," Signal Processing-Image Communication, vol. 127, 2024. https://doi.org/10.1016/j.image.2024.117154.

[18] Y. H. Li, T. Yao, Y. W. Pan, and T. Mei, "Contextual Transformer Networks for Visual Recognition," Ieee Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 2, pp. 1489-1500, 2023. https://doi.org/10.48550/arXiv.2107.12292.

[19] W. D. Yan, L. Cao, P. Yan, C. S. Zhu, and M. T. Wang, "Remote sensing image change detection based on swin transformer and cross-attention mechanism," Earth Science Informatics, vol. 18, no. 1, 2025. https://doi.org/10.21203/rs.3.rs-4712422/v1.

[20] Y. C. Lin, C. H. Wang, and Y. C. Lin, "GAT TransPruning: progressive channel pruning strategy combining graph attention network and transformer," Peerj Computer Science, vol. 10, 2024. https://doi.org/10.7717/peerj-cs.2012.

[21] C. Y. Liu and C. J. Sun, "A Fusion Deep Learning Model of ResNet and Vision Transformer for 3D CT Images," Ieee Access, vol. 12, pp. 93389-93397, 2024.

[22] C. Y. Liu, R. Zhao, and Z. W. Shi, "Remote-Sensing Image Captioning Based on Multilayer Aggregated Transformer," Ieee Geoscience and Remote Sensing Letters, vol. 19, 2022.

[23] J. Liu, S. W. Tian, L. Yu, X. W. Shi, and F. Wang, "Image-text fusion transformer network for sarcasm detection," Multimedia Tools and Applications, vol., 2023. https://doi.org/10.1007/s11042-023-17252-2.

[24] X. L. Liu, H. L. Feng, Y. Wang, D. Y. Li, and K. Zhang, "Hybrid model of ResNet and transformer for efficient image reconstruction of electromagnetic tomography," Flow Measurement and Instrumentation, vol. 102, 2025. https://doi.org/10.1016/j.flowmeasinst.2025.102843.

[25] K. Lu et al., "Resformer-Unet: A U-shaped Framework Combining ResNet and Transformer for Segmentation of Strip Steel Surface Defects," Isij International, vol. 64, no. 1, pp. 67-75, 2024. https://doi.org/10.2355/isijinternational.ISIJINT-2023-222.

[26] Z. A. Lyu and M. R. D. Rodrigues, "Exploring the Impact of Additive Shortcuts in Neural Networks via Information Bottleneck-like Dynamics: From ResNet to Transformer," Entropy, vol. 26, no. 11, 2024. https://doi.org/10.3390/e26110974.

[27] X. R. Ma, Y. Y. Wang, J. N. Qin, Z. F. Wang, and Z. Y. Liu, "A bearing fault diagnosis model with convolutional cross transformer and ResNet18," Measurement Science and Technology, vol. 36, no. 1, 2025. DOI 10.1088/1361-6501/ad8a7b.

[28] Y. F. Ma, Y. L. Wang, X. Y. Liu, and H. Y. Wang, "SWINT-RESNet: An Improved Remote Sensing Image Segmentation Model Based on Transformer," Ieee Geoscience and Remote Sensing Letters, vol. 21, 2024. DOI:10.1109/LGRS.2024.3433034.

[29] Y. H. Mao, Y. H. Lv, G. X. Zhang, and X. L. Gui, "Exploring Transformer for Face Mask Detection," Ieee Access, vol. 12, pp. 118377-118388, 2024. https://doi.org/10.1109/ACCESS.2024.3449802.

[30] Z. D. Wu, L. S. He, W. Wang, Y. Z. Ju, and Q. Guo, "A Fault Prediction Method for CNC Machine Tools Based on SE-ResNet-Transformer," Machines, vol. 12, no. 6, 2024. https://doi.org/10.3390/machines12060418.