Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2025.CH23276

Research on the Application of Speech Recognition Technology Based on Transformer Model

Hang Xi

Computer Science and Communication Engineering College, Jiangsu University, Zhejiang, China
dufflejf@163.com

Abstract: Speech recognition technology has developed from the 1950s to the present, evolving from template matching methods to Hidden Markov Model (HMM) statistical methods, then to machine learning techniques, and finally to the current use of Transformer technology for speech recognition tasks. However, the Transformer model has not yet been widely adopted in the field of speech recognition. This paper explores the characteristics of Transformer model, combines it with the characteristics of speech recognition tasks, analyzes the challenges associated with using Transformer model for these tasks, and provides suggestions for directions of future research, so as to facilitate the application of Transformer models in speech recognition. The paper finds that the reasons for the limited application of Transformer models in speech recognition tasks mainly include their numerous parameters, complex structure, and high computational costs, which have prevented their extensive use in this field. In the future, efforts should focus on enhancing model compression and lightweight design, and improving the attention mechanism to boost the applicability of Transformer models in speech recognition.

Keywords: Transformer, Automatic Speech Recognition, Deep Learning, Computer Science

1. Introduction

Language is one of the most vital forms of communication for humans. As a branch of natural language processing, speech recognition has attracted extensive interest and study from academics in recent years due to the development of artificial intelligence technology. At the same time, numerous speech recognition products have been developed and are widely used in various aspects of people's lives, from mobile applications to in-car computers. Since its inception in the 1950s, speech recognition has evolved from template matching technology, to Hidden Markov Models (HMM) statistics, then to machine learning, and finally to the use of Transformer models today. The accuracy and efficiency of speech recognition have also been continuously optimized and improved. However, the use of Transformer models in speech recognition remains limited.

To enhance the application of Transformer models in speech recognition, many scholars have proposed improvement directions and suggestions based on the characteristics of the model itself. Scholar Wang Yanhong proposed designing a lightweight Transformer model by replacing the linear operations of Query, Key, and Value with lightweight convolutional operations, optimizing the multi-head attention mechanism to improve attention distribution, and introducing block low-rank decomposition in the feedforward neural network to maximize model compression [1]. Scholar Li Junhua proposed a feature enhancement algorithm (CLformer) based on Transformer model, which

DOI: 10.54254/2755-2721/2025.CH23276

enhances local feature interaction and low-level feature utilization, thereby improving the performance of automatic speech recognition (ASR) tasks [2].

This article employs a literature review method for research. It focuses on the inherent characteristics of the Transformer model, combines them with the characteristics of speech recognition tasks, and analyzes the difficulties, pain points, and related technical deficiencies of applying the Transformer model to speech recognition tasks. This article provides constructive suggestions and a theoretical basis for the future improvement and research directions of the Transformer model.

2. Literature review

2.1. Basic model and types of speech recognition technology

Speech recognition, commonly referred to as ASR, is an important branch in artificial intelligence. Its basic principle is to convert human language into text information, thereby enabling natural language interaction between humans and machines. This is regarded as the most direct way of human-machine interaction.

Scholar M.A. Anusuya mentioned that a speech recognition process consists of four units: front-end unit, model unit, language model unit and search unit, as shown in Figure 1 [3]. The simple equation composed of these four modules is the mathematical representation of the speech recognition system.

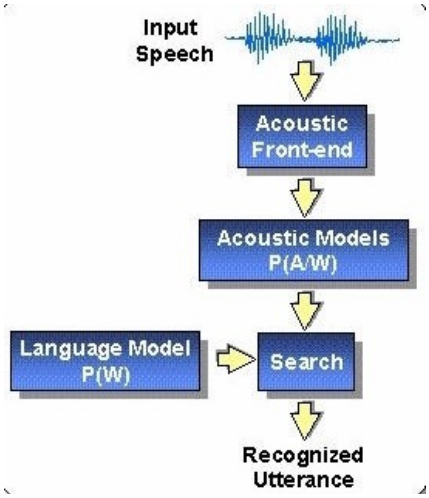

Figure 1: Basic model of speech recognition [3]

Scholar Santosh K. Gaikwad mentioned that speech recognition can be classified into four different categories based on the types of ullerances [4].

Speech recognition technologies can be categorized into four main types: Isolated Word, Connected Words, Continuous Speech, and Spontaneous Speech. Isolated Word recognition processes single words or utterances at a time, ensuring high accuracy but requiring precise segmentation. Connected Words allows multiple utterances to be run together with minimal pauses in between. Continuous Speech recognition, akin to computer dictation, identifies entire sentences after natural speech, with the challenge lying in accurately detecting pauses and sentence boundaries. Spontaneous Speech recognition, the most advanced, handles natural language features such as assimilation, stuttering, and pauses, closely mimicking human conversational patterns.

DOI: 10.54254/2755-2721/2025.CH23276

2.2. The three major methods of speech recognition

From the perspective of the development history, the methods of speech recognition can be classified into the following three most common ones.

The acoustic phonetic approach is one of the earliest methods of speech recognition, relying on analysis of the waveform features of speech to assign appropriate labels. It begins by analyzing and extracting the spectrum of the speech, followed by segmenting the spectrum into stable acoustic regions. Each region is then labeled, and utterances are determined based on the sequence of these labels. Despite its foundational role, this method has not seen widespread commercial application due to its limitations in handling variability and complexity in speech.

The pattern recognition approach, which has dominated the field for over 70 years, involves two key steps: pattern training and pattern comparison. This method uses a well-defined mathematical framework to establish consistent speech patterns. During the comparison stage, the input speech is matched against these pre-trained patterns to determine the best match. This approach is divided into template-based and stochastic methods, with the latter, particularly the Hidden Markov Model (HMM), being more widely used. The stochastic approach excels at handling uncertainties such as variations in speaking speed and volume, making it highly suitable for speech recognition tasks.

The artificial intelligence approach combines elements of both pattern recognition and the acoustic phonetic methods. It leverages deep learning technologies, like deep neural networks (DNN), convolutional neural networks (CNN), and recurrent neural networks (RNN), to automatically learn deep-level features from speech signals. Variants like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) are particularly effective in capturing temporal dependencies within speech sequences, improving recognition accuracy. However, traditional deep learning models face challenges with computational efficiency and adaptability in noisy environments. Recent advancements, such as the Transformer model, have addressed these issues by using self-attention mechanisms to capture both global and local dependencies efficiently, enabling faster training and better performance in speech recognition tasks.

2.3. Overview of Transformer model

The Transformer model is a brand-new architecture proposed by Ashish Vaswani et al in their paper Attention Is All You Need [5]. It eliminates the recurrent and convolutional operations in Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN), and completely relies on the attention mechanism to capture the global dependency between the input and output.

This model follows the overall architecture of encoder-decoder, in which both the encoder and the decoder consist of multiple identical layers stacked together. Each encoder layer in the Transformer model is made up of two sub-layers: the multi-head self-attention layer and the feed-forward neural network layer, as shown in Figure 2.

DOI: 10.54254/2755-2721/2025.CH23276

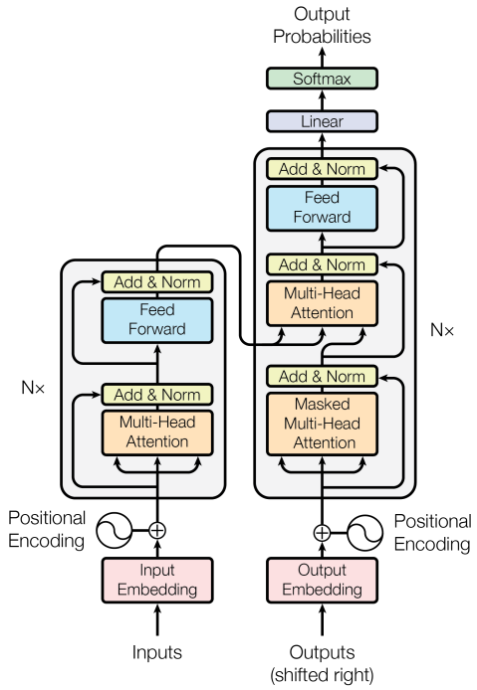

Figure 2: The Transformer - model architecture [5]

Residual connections and layer normalizations are adopted between these two sub-layers. Although the decoder layers are similar to the encoder levels, the encoder layers' output is handled by an extra masked multi-head attention layer. Every position of the decoder can concentrate on every position in the input sequence thanks to this extra attention layer.

Self-attention mechanism is the core of the Transformer model. It enables the model to simultaneously pay attention to other positions in the sequence when processing the information at a certain position. This mechanism achieves its purpose by calculating the similarity between the query and the key, and then weighting and summing the values based on the similarity as shown in figure (a). To capture various dependencies between different positions in the sequence, the Transformer model adopts the multi-head attention mechanism. It projects the queries, keys, and values through multiple different linear transformations into multiple subspaces, and independently performs self-attention operations in each subspace as shown in figure (b). Finally, the outputs of various subspaces are concatenated and subjected to a linear transformation to obtain the final output.

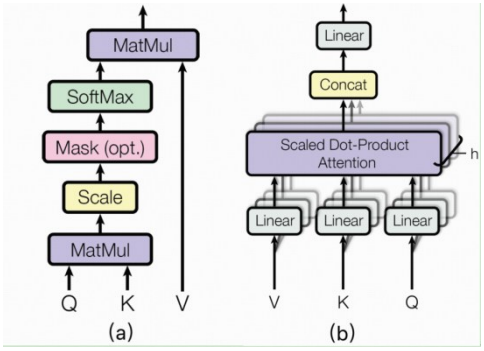

Figure 3: (a) Scaled Dot-Product Attention. (b) Multi-Head Attention consists of several attention layers running in parallel [5]

DOI: 10.54254/2755-2721/2025.CH23276

Since the Transformer model does not employ recurrent or convolutional operations to capture the positional information within the sequence, positional encoding is implemented to address this deficiency. Positional encoding generates a unique representation for each position through a combination of sine and cosine functions. To provide positional information, these representations are added to the input embeddings.

The Transformer model has achieved extensive applications and remarkable performance improvements in natural language processing. Its advantages mainly lie in its higher parallelization capability, the ability capability of identifying long-term dependencies, and its stronger expressive power. Since the Transformer model is entirely dependent on the attention mechanism, it can achieve parallelized computing more easily, thereby accelerating the training process. Meanwhile, through the self-attention mechanism, the Transformer model can directly capture the dependency relationship between any two positions in the sequence without having to pass information step by step like RNN does. This makes the Transformer model more advantageous in handling long sequences. Furthermore, the multi-head attention mechanism enables the Transformer model to capture various different dependencies within the sequence, thereby enhancing the model's expressive power.

The Transformer model is a model architecture that is innovative, efficient and has strong expressive power. It has achieved remarkable performance improvements and broad application prospects in natural language processing.

3. Improvement of Transformer model

The speech recognition technology based on Transformer model has achieved remarkable research progress in recent years. Early studies mainly focused on how to effectively apply the Transformer model to speech recognition tasks and explore its performance in various scenarios.

3.1. Speech recognition technology employing the Transformer model

The Transformer model performs well in machine translation jobs because it can handle sequence data more effectively and save training time. Therefore, Dong et al. were the first to apply the Transformer model to the field of speech recognition. They innovatively developed a model named Speech-Transformer, which is a non-recurrent sequence-to-sequence (seq2seq) model specifically designed for speech recognition tasks [6].

They first conducted an analysis and research on the progress made by the sequence-to-sequence model in speech recognition. They discovered that it has the drawback of slow training speed. Due to the internal loop, the training is not parallelized, which makes it particularly time-consuming when dealing with long speech sequences. To solve this problem, they proposed the Speech-Transformer model.

The Speech-Transformer model is entirely dependent on the attention mechanism to learn position dependence, thereby enabling faster and more efficient training. This model adopts an encoder-decoder architecture, as seen in Figure 4, in which the encoder transforms the voice feature sequence into hidden representations, and the decoder uses these hidden representations to produce the matching character sequence.

DOI: 10.54254/2755-2721/2025.CH23276

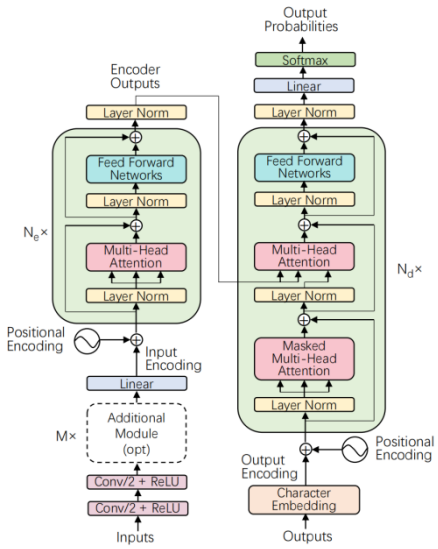

Figure 4: Model architecture of the Speech-Transformer [6]

They also proposed a 2D-Attention mechanism as shown in Figure 5, which can jointly focus on the time axis and frequency axis of two-dimensional speech, thereby providing richer expression for Speech-Transformer. This mechanism is inspired by the Long Short-Term Memory Network (LSTM) for time and frequency, replacing the time-frequency recurrence with the temporal-frequency dependency captured by attention.

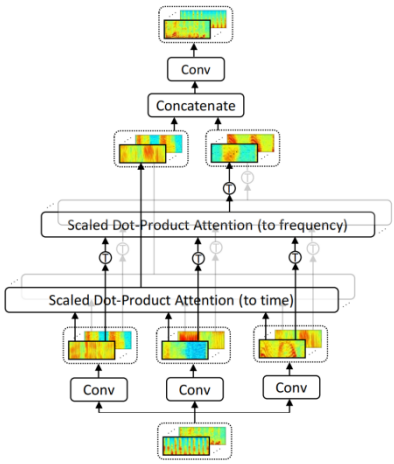

Figure 5: Illustration of the Proposed 2D-Attention mechanism [6]

In the experimental part, Dong et al. evaluated the Speech-Transformer using the Wall Street Journal (WSJ) speech recognition dataset. The experimental results indicated that the best model of this model achieved a word error rate (WER) of 10.9%, and the entire training process only took 1.2 days on a single GPU, which was significantly faster than the training speed of the published recurrent sequence-to-sequence models [6]. Furthermore, they also compared Speech-Transformer with other speech recognition models, such as those based on convolutional neural networks and recurrent neural networks. The results demonstrated that Speech-Transformer outperforms the other models in terms of performance and training speed [6].

Overall, the Speech-Transformer model successfully addresses the problem of slow training speed of recurrent sequence-to-sequence models in speech recognition by introducing attention

DOI: 10.54254/2755-2721/2025.CH23276

mechanisms and 2D-Attention mechanisms. This model provides new ideas and methods for the field of speech recognition and has broad application prospects.

3.2. Conformer model

Gulati et al. utilized the benefits of using attention mechanisms and convolution, respectively, to extract global and local feature information. They then merged the Transformer model with CNN to develop the Conformer model, which further improves the model's accuracy. He pointed out that in the field of ASR, the interaction between global and local information is of vital importance for model performance. Through previous studies, they found that while Transformer and similar models successfully capture global dependencies through self-attention processes, they fall short in terms of extracting enough local features. Convolutional neural networks, on the other hand, are good at capturing local features but relatively weak in modeling global information. Therefore, Gulati et al. hypothesized that by integrating global and local interactions, an efficient and superior ASR model with parameters can be achieved [7].

The Conformer model they designed adopts an encoder structure similar to Transformer, but adds a convolution module in each encoding layer. The Figure 6 shows that the Conformer block consists of two feed-forward modules and a convolution module, forming a "sandwich" structure. This structure is based on Macaron-Net, replacing the feed-forward layer in the original Transformer block with two half-step feed-forward layers, one before the attention layer and the other after the attention layer. Moreover, Conformer uses relative position embeddings to enhance the modeling ability of the multi-head self-attention module for position information.

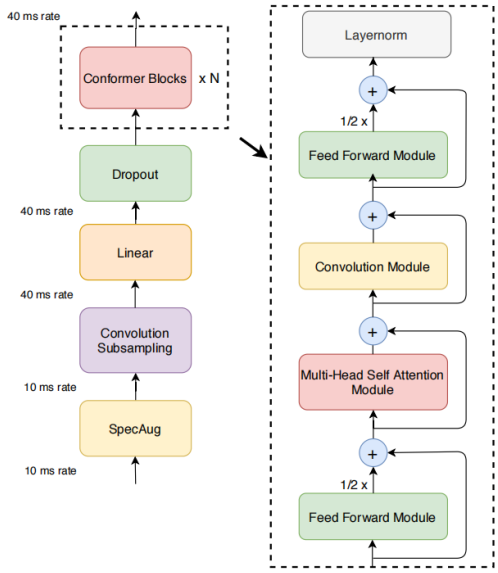

Figure 6: Conformer encoder model architecture [7]

In the experimental part, Gulati et al. conducted extensive experiments on the LibriSpeech dataset to verify the effectiveness of the Conformer model. The results demonstrated that the Conformer model achieved significantly better performance than the baseline models (such as Transformer and Jasper, etc.) in multiple evaluation metrics. Moreover, they conducted ablation experiments to analyze in detail the effect of different components (such as the number of attention heads, the size of convolution kernels, activation functions, etc.) on the model performance, further proving the rationality of the Conformer model design.

DOI: 10.54254/2755-2721/2025.CH23276

3.3. Lightweight design of Transformer

Although the Transformer model has a powerful ability for sequence modeling, its huge parameter scale and high computational complexity limit its deployment on resource-constrained edge devices. Wang Yanhong et al. improved the Transformer model and achieved a lightweight speech recognition model that maintains a high recognition rate while significantly reducing computational complexity and memory usage [1].

They have developed a lightweight speech recognition model CLSC-Transformer based on the improvement of the Transformer model. This model adopts a two-dimensional attention mechanism, imitating the ability of humans to predict sounds by observing frequency changes, enabling the model to dynamically model in both time and frequency dimensions and extract richer acoustic features more accurately. Regarding the large number of parameters and high computational intensity of the feedforward network and attention mechanism in the Transformer model, they have made the following optimizations.

The first is the Lightweight feedforward network module. According to the research results of scholar WANG J, a lightweight block-based feedforward network method is presented to divide the input sequence into several small blocks, and then apply the feedforward network to each small block separately, thus reducing the computation amount [8].

The second is the Lightweight Multi-head Attention Layer. According to the research results of scholar ZENG K, Conv2D operation is used to replace the operation of Q, K, V matrix generated by linear transformation in conventional multi-head attention, and low-rank decomposition technology is introduced to further reduce parameter redundancy and computational complexity [9]. At the same time, a layer of linear transformation is retained to better extract input features and enhance the linear representation of the model.

In order to verify the recognition effect of the CLSC-Transformer model, the experiments they conducted adopted the Pytorch learning library under the Windows 10 system and used the AISHELL-1 dataset for comparative tests. The experimental results indicated that the CLSC-Transformer model maintained excellent recognition performance while significantly reducing the number of parameters. Compared with other models, the CLSC-Transformer model reduced the proportion of parameters by a relatively large margin and also decreased the word error rate.

Furthermore, the paper also further verified the generalization ability of the CLSC-Transformer model. The model was trained and tested using the LRS2 dataset which involved diverse background noises, variations in speaking speed, and more complex environments. Different configurations were tried to explore the optimal parameter settings. The experimental results demonstrated that the CLSC-Transformer model also exhibited excellent performance in complex speech scenarios.

3.4. Hotword enhancement algorithm

The Transformer model has demonstrated its remarkable computational advantages in the field of speech recognition. However, in certain specific scenarios, such as in the medical domain, there are numerous proper nouns or hot words. In such specific scenarios, the accuracy of the conventional Transformer model will decline. Therefore, a hot word enhancement algorithm is needed to optimize the model for these professional terms in order to improve the recognition accuracy.

Yiyang Qian et al. proposed a speech recognition hot word enhancement algorithm based on soft beam pruning and prefix word module [10]. This algorithm is based on the Transformer model and improves the beam search algorithm. Through the soft beam pruning algorithm, it dynamically trims the decoding paths with lower confidence levels, reducing the computational load and improving the decoding speed. Meanwhile, the algorithm also introduces a prefix word module, which utilizes the

DOI: 10.54254/2755-2721/2025.CH23276

connection between professional terms and the results of the previous time step to improve the recognition accuracy of professional terms.

By conducting experiments on the Chinese medical dialogue test set, they found that the proposed algorithm for enhancing hot words can significantly promote the recognition accuracy of professional terms and also enhance the decoding speed.

Furthermore, they also explored the issue of hot word enhancement in streaming speech recognition and proposed a re-ranking model based on Connectionist Temporal Classification

(CTC) decoding and beam search, which further expanded the performance of the model. This model optimizes the recognition results through secondary ranking and correction, thereby improving the accuracy of streaming speech recognition.

Finally, Yiyang Qian et al. designed and implemented a voice recognition hot word enhancement system based on PyQt, applying the proposed algorithm to practical scenarios [10]. This system supports both medical hot word enhancement and custom hot word enhancement, and can adjust the intensity of hot word enhancement. It realizes the functions of real-time voice recognition and hot word enhancement.

4. The challenges

The speech recognition technology based on Transformer model has received extensive attention and research in recent years. However, in practical applications, this technology still encounters a series of challenges and problems, which have a significant impact on the performance of Transformer model in speech recognition.

Firstly, data sparsity is a widespread issue. In the task of speech recognition, especially when dealing with multiple languages and dialects, the available training data is often very limited. Although the Transformer model has strong modeling capabilities, its performance will be greatly compromised in the case of scarce data. To tackle this problem, researchers have proposed some methods, such as data augmentation and transfer learning, to leverage the limited data resources to enhance the generalization ability of the model. However, these methods still face certain limitations in practical applications. How to more effectively utilize sparse data to improve the performance of the Transformer model remains an unsolved problem.

Secondly, noise interference is also an issue that cannot be ignored in speech recognition. In practical environments, speech signals are often disturbed by various noises, such as environmental noise and equipment noise. These noises seriously affect the clarity and recognizability of speech signals, thereby reducing the recognition accuracy of Transformer models. To address noise interference, researchers have proposed a series of denoising techniques and robust training methods.Even though these techniques have somewhat increased the model's resistance to interference, it is still difficult to determine how to make Transformer models even more resilient in complicated and changing real-world settings.

Apart from the aforementioned two issues, the Transformer model also encounters other challenges in speech recognition. For instance, there are problems related to the model's complexity and the recognition of multiple languages and dialects. To address these issues, researchers need to continuously explore new methods and technologies to further advance the field of speech recognition. Future studies can concentrate on ways to reduce model complexity, boost the model's anti-interference capability, increase model performance with sparse data, and accomplish universal recognition for a variety of languages and dialects.

DOI: 10.54254/2755-2721/2025.CH23276

5. Conclusion

This article analyzes the inherent structure of the Transformer model and interprets the characteristics of speech recognition tasks, combining the features of the Transformer model with those of speech recognition tasks for analysis and research. It also studies and analyzes current researches and experiments conducted by scholars in this field, identifying the difficulties, pain points, and related technical deficiencies of applying the Transformer model to speech recognition tasks. Based on the research findings, suggestions for future research directions are proposed to better apply the Transformer model to speech recognition tasks. However, this study has some limitations, such as not conducting detailed test analysis on current speech recognition products, not conducting a thorough survey on user experience, and thus lacking understanding of the practical application experience of current speech recognition products.

Speech recognition, as the most direct means of communication between humans and computers, will undoubtedly attract the attention of many researchers in the future. As one of the most powerful large models today, the Transformer model will also play a pivotal role in this field. In the future, speech recognition will inevitably be integrated into all aspects of people's lives and widely used in various corners of daily existence.

References

[1] *Yanhong Wang, Liang Zhao, Guanjun Wang. (2025). Improved voice recognition lightweight design for the Transformer model, Computer Engineering and Applications.*

[2] *Junhua Li, Zhikui Duan, Xinmei Yu. (2025). A feature enhancement algorithm based on Transformer model and its application. Journal of Foshan University (Natural Sciences Edition).*

[3] *M. A. Anusuya, S. K. Katti. (2009). Speech Recognition by Machine: A Review. International Journal of Computer Science and Information Security, 6(3).*

[4] *Santosh K. Gaikwad, Bharti W. GawaIi, Pravin Yannawar. (2010). A Review on speech Recognition Technique. International Journal of computer Applications, 10(3).*

[5] *VASWANI A., SHAZEER N., PARMAR N., et al. (2017). Attention is all you need [J]. Advances in Neural Information Processing Systems.*

[6] *DONG L., XU S., XU B. (2018). Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 5884-5888.*

[7] *GULATI A., QIN J., CHIU C. C., et al. (2020). Conformer: Convolution-augmented Transformer for speech recognition. In Proceedings of Interspeech, 5036-5040.*

[8] *WANG J., LIANG Z., ZHANG X., et al. (2024). EfficientASR: Speech Recognition Network Compression via Attention Redundancy and Chunk-Level FFN Optimization [J]. arXiv preprint arXiv: 2404.19214.*

[9] *ZENG K., PAIK I. A. (2020). Lightweight transformer with convolutional attention. In 2020 11th International Conference on Awareness Science and Technology (iCAST). IEEE, 1-6.*

[10] *Yiyang Qian. (2023). Automatic speech recognition and hotword enhancement Algorithm based on Transformer.*