EAI Endorsed Transactions
on Internet of Things

Research Article EAI.EU

Conformer network-guided speech recognition for smart home Internet of Things system

Shengjun Huang[1,2], and Dong-hyun Kim[1,*]

[1]Department of Computer and Information Engineering, Graduate School Youngsan University, Busan, 612-022 South Korea
[2]Nanjing Technical Vocational College, No. 58, Huangshan Road, Jianye District, Nanjing City 210019, China

Abstract

In response to the fact that the World is gradually entering an aging society, and the problem that traditional Internet of Things (IoT) systems are operationally complex and lack humanization, a conformer network-guided speech recognition for smart home Internet of Things system is proposed in this paper. Firstly, by introducing a voice recognition module with an embedded processor, not only traditional voice recognition has been achieved, but also cloud transmission of voice has been realized, breaking through the bottleneck of low computing and storage capabilities of the main control chip. Then, by using Internet of Things technology, the complex algorithms are transferred to the cloud for execution. There is a significant improvement in voice recognition. By leveraging the distributed storage feature of the cloud, a user-specific voice database can be established categorically. This enables the provision of a vast amount of data basis when users are learning. In response to the shortcomings of the existing Conformer speech recognition model, such as insufficient extraction ability of time-frequency features, redundant model structure and large number of parameters, this paper proposes a speech recognition model based on asymmetric convolution and gated feed-forward neural network. Different-sized asymmetric convolutions are used to perform multi-scale fusion and down-sampling on the time-frequency features of the speech sequence. This not only enhances the model's ability to extract time-frequency features but also effectively reduces the information loss during down-sampling. At the same time, the gated feed-forward module is introduced to replace the double half-step feed-forward network in Conformer, reducing the number of network parameters while simplifying the model structure. Finally, based on a large amount of data, the system gradually builds a personalized speech recognition library for the user through learning. Through experiments, the effectiveness of the proposed intelligent fusion-based Internet of Things system in terms of speech recognition accuracy, computing power, and the intelligence level of voice interaction has been verified.

Keywords: Internet of Things, conformer network, speech recognition, asymmetric convolution, gated feed-forward neural network

Received on 30 July 2025, accepted on 09 November 2025, published on 20 November 2025

doi: 10.4108/eetiot.9818

1. Introduction

In recent years, with the development of science and technology, speech recognition and Internet of Things (IoT) technologies have been widely applied in intelligent retrieval, telemedicine, intelligent transportation, and smart home, among others [1-3]. Among them, smart home is an important application field of speech recognition and Internet of Things technology, and it has great development potential [4-6]. However, in the aspect of intelligent interaction in smart homes, there are problems such as complicated operations, insufficiently user-friendly design for the elderly, low voice recognition rate, rigid communication, and slow data transmission speed. To address these issues, scholars have conducted extensive researches. For example, Zhang et al. [7] proposed a new idea of combining Internet of Things technology with Kinect depth sensors, using three-dimensional body-sensing camera sensors in the home stereoscopic space to recognize human body postures and perform some voice

[*]Corresponding author. Email: lihchesu@foxmail.com

recognition. However, Chinese voice control was not currently supported and had certain limitations. Zhang et al. [8] proposed an Internet of Things (IoT) smart home system based on a microcontroller of a single-chip microcomputer, comprehensively elaborating on the development level of IoT technology and using IoT technology and mobile APPs to control household appliances. However, it was relatively cumbersome in terms of human-computer interaction, and the functions of the mobile APP and other features were not particularly perfect. The consideration for elderly users was not comprehensive enough. In terms of the combination of speech recognition and the Internet of Things, Juluru et al. [9] proposed a speech recognition-based smart home assistant, integrating IoT technology and speech recognition technology to simplify the human-computer interaction method. However, it merely utilized the LD3320 speech module as a command switch for use, without uploading the voice data to the cloud server. The communication would be rather rigid, and the data processing capability was not strong enough. Zhou et al. [10] utilized LD3320 and Cortex-M3 cores for voice interaction control. However, due to the limited computing power and storage space of a single chip, it would restrict the computing speed and affect the user's experience.

In order to enhance people's experience with smart homes, applying voice recognition and self-learning algorithms to smart homes is an effective solution. Ahmed et al. [11] proposed a speech endpoint detection based on deep learning, and described the research on the hardware and software algorithms of voiceprint recognition. Here, the algorithms such as endpoint recognition, Multi-Resolution Aural Cepstrum Coefficient (MRACC), and Deep Neural Network (DNN) have provided good references for enhancing the user experience of smart homes. Wu et al. [12] also made many attempts in the application of artificial intelligence voiceprint recognition technology in the Internet of Things, and comprehensively combined Internet of Things technology and voiceprint recognition technology [13], but they only discussed this technology and did not provide practical solutions.

Automatic Speech Recognition (ASR) is a technology that uses algorithms to convert human speech signals containing lexical content into text information. With the rapid development of artificial intelligence, speech recognition technology has become one of the hottest topics in the current field of science and technology, playing an important role in both military and civilian voice automation control systems. Among them, the end-to-end speech recognition model is a new type of speech recognition technology. This model can map the acoustic feature sequence to the output labels using a single neural network, without the need for complex system design and pre-aligned training. The end-to-end speech recognition model has become a research hotspot in the current field of speech recognition due to its simplicity, efficiency and flexibility.

End-to-end speech recognition methods can be classified into two categories: the connectionist temporal classification (CTC) algorithm [14] and the attention-based encoder-decoder (AED) algorithm [15]. The CTC method directly maps the input sequence to the output sequence without the need for label alignment. However, CTC is highly dependent on the pronunciation dictionary and language model, and requires independent assumptions. For complex speech sequences, it may result in translation errors [16]. In contrast, AED can adaptively focus on the key parts of the input speech sequence through the attention mechanism, thereby enhancing the flexibility and robustness of the model and reducing its dependence on the pronunciation dictionary and language model. During the actual training process, AED can adopt the multi-task learning method and combine the CTC loss to optimize the shared encoder. This training strategy effectively improves the convergence performance of the model and mitigates the impact of the alignment problem.

For the widely used AED method, researchers can adopt various different model structures. Kamal et al. [17] introduced the Transformer model into the field of speech recognition. This method utilized the parallel processing of the self-attention mechanism inside the Transformer to handle long sequence information and capture global context information. The Conformer model [18] used a hybrid network structure based on the Transformer, combining the convolutional neural network (CNN) that was good at extracting local features with the self-attention mechanism. This structure enabled the model to have the ability to capture both global context and local correlations. In the field of speech recognition, Conformer has achieved excellent performance and has become one of the models that attract much attention in the current research field of speech recognition.

The Conformer speech recognition model has achieved remarkable results, but there are two aspects that are worthy of discussion and research: (1) Speech information has time-frequency characteristics. The Conformer model down-samples the time-frequency features of the speech information through two sub-sampling layers. This structure will lose a considerable amount of frequency features, which directly affects the recognition accuracy of the model [19]. (2) The Conformer architecture draws on the design concept of Macaron-Net [20] and employs a Macaron-style dual half-step feed-forward module. This design can effectively enhance the model's recognition performance. However, its dual feed-forward modules contain more weight parameters, thereby increasing the computational cost of the entire network, which limits the practical deployment scenarios and recognition speed of the Conformer model.

To enhance the ability of the acoustic model to extract time-frequency features in speech recognition, Rajab et al. [21] used the VGG network architecture as the initial layer of the model and performed down-sampling of the speech's time-frequency features through two max pooling layers. Aich et al. [22] employed a MobileNetV2-like architecture to fuse two speech time-frequency feature maps of different resolutions, aiming to achieve multi-stream and multi-scale feature extraction. Berghi et al. [23] replaced the sub-sampling layers in Conformer with the stack-based ConvNeXt architecture to leverage the information contained in the time-frequency speech features. The above methods can effectively extract the time-frequency features of speech. However, introducing additional model architectures will result in a more complex network structure. Moreover, to improve the recognition efficiency of the model, Yu et al. [24] used local generative attention instead of self-attention, reducing the computational complexity of the model and enhancing its performance. Burchi et al. [25] introduced progressive downsampling and group attention mechanisms in the Conformer model, improving the method and accelerating the training and inference speed of the model. Although these methods reduce the computational cost of the model, they do not involve a reduction in model parameters.

In response to the above issues, this paper proposes a conformer network-guided speech recognition method for smart home Internet of Things system. This system consists of a main control board, a speech acquisition module, a cloud server, a mobile APP, and an off-site control board. This system has two working modes: local area network and wide area network. It uses a built-in processor-based speech recognition module. While conducting traditional recognition, it can also selectively transmit the collected speech information to the cloud server through the built-in processor, and execute the complex speech processing algorithms on the cloud, thereby improving the accuracy of voiceprints. At the same time, a user's voice database is classified and established in the cloud, and a personal voiceprint database belonging to the user is created based on the data storage. The intelligent integrated IoT system in this paper abandons traditional interaction technologies such as keyboards and touch screens, adopts voice interaction while retaining the functions of the mobile phone APP, takes into account the needs of elderly users while also catering to the usage habits of young people. Compared with the traditional LD3320 speech recognition module, it breaks through the limitations of the computing and storage capabilities of the main control chip, and can provide certain technical support for the arrival of the artificial intelligence era.

2. Intelligent integrated IoT system architecture

Figure 1 presents the key architecture diagram of the system, including the wide area network closed-loop system structure and the local area network closed-loop system structure. Users can select different working modes according to the usage scenarios.

In Figure 1, a closed-loop control system is constructed through three communication methods: WiFi, ZigBee and Bluetooth, to complete a complete set of operations for receiving, processing and executing commands of voice information. Since it does not involve the use of peripheral resources [26], when the device is in a long-term working state, it can effectively reduce channel pressure and power consumption, and increase the battery life. In order to enhance the user experience and increase the intelligence level of the system, a closed-loop network is established among the central controller, cloud server, remote devices, and users through communication methods such as WiFi, 4G and 5G.

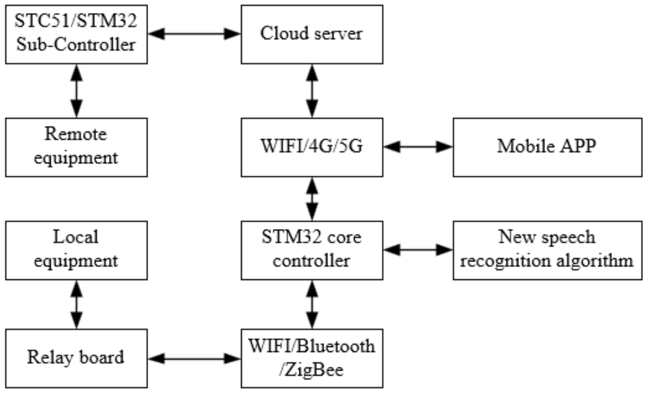


Figure 1. Intelligent integrated IoT system architecture

The introduction of cloud servers has overcome the computational capacity bottleneck of the STM32 central controller. During each heartbeat cycle, the controller communicates with the cloud server once, reporting the command data, device status, and user usage during the heartbeat cycle. The cloud server uses each device terminal as a data node of the neural network and then utilizes the computing power of the central server to form a central system with learning capabilities [27,28].

3. Conformer-based speech recognition

3.1. Conformer model

The Conformer model combines deep separable convolution with the self-attention mechanism to integrate local and global feature relationships and achieves excellent results in many speech recognition tasks. The structure of the Conformer encoder model is shown in Figure 2.

The convolutional down-sampling layer consists of a 3×3 convolution with a stride of 2, which is responsible for down-sampling the time-frequency features of the input audio sequence to reduce the computational complexity in the subsequent stages. The fully connected layer converts the time-frequency features into time-series features and inputs them into the Conformer module to complete feature extraction.

The Conformer module consists of four sub-modules: two feed-forward modules (FFM), a multi-head self-attention module (MHSA), and a convolution module

(Conv). Residual connection methods and layer normalization strategies are employed at the beginning and end of each module. Among them, two feedforward modules that use half-step residual weights sandwich the multi-head self-attention module and the convolution module in between. This structure is called the Macaron style.

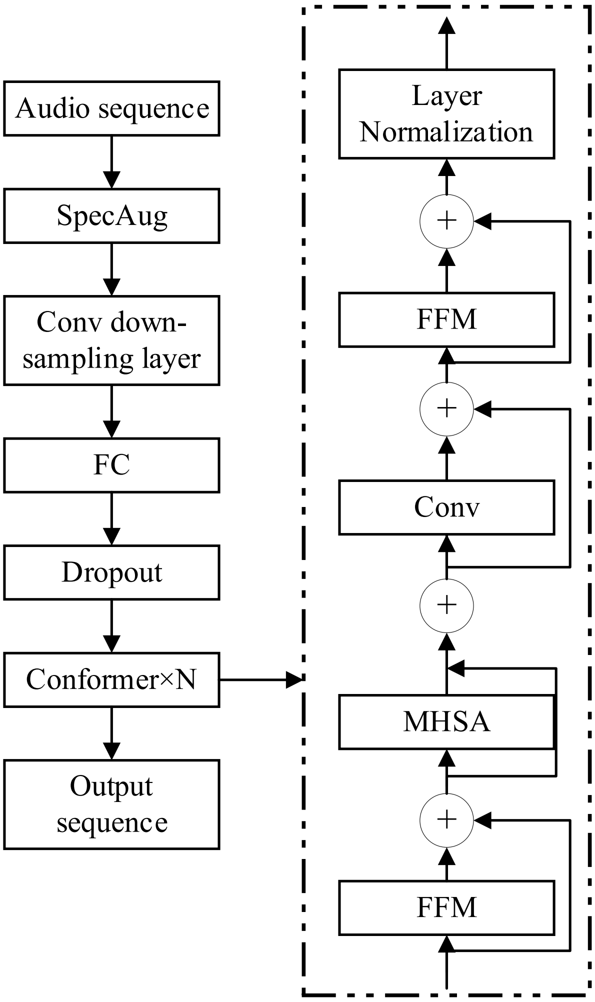


Figure 2. Conformer encoder model structure

For the input x_i of the i-th layer in the Conformer module and the output y_i, the calculation is as shown in equations (1)~(4).

$$\bar{x}_i = x_i + \frac{1}{2}FFM(x_i) \quad (1)$$

$$x'_i = \bar{x}_i + MHSA(\bar{x}_i) \quad (2)$$

$$x''_i = x'_i + Conv(x'_i) \quad (3)$$

$$y_i = x''_i + LayerNorm(x''_i + \frac{1}{2}FFM(x'')) \quad (4)$$

3.2. Overview of the proposed speech recognition model

This paper proposes a speech recognition model based on asymmetric convolution and gated feed-forward neural network (ACGFNN). While keeping the decoder unchanged, the convolutional down-sampling layer and the encoder of the Conformer model are optimized. The overall structure of this proposed model is shown in figure 3, consisting of a pre-processing module, an asymmetric convolution front-end, and an encoder. The pre-processing module performs Log-Mel feature extraction on the original audio sequence. The asymmetric convolution front-end is responsible for down-sampling the audio features and adding positional encoding to the down-sampled feature sequence. The encoder is composed of N stacked ACGFNN modules and is responsible for mapping the down-sampled feature sequence to the hidden layer. Finally, the decoder maps the hidden layer to the sequence of natural speech and obtains the final output result.

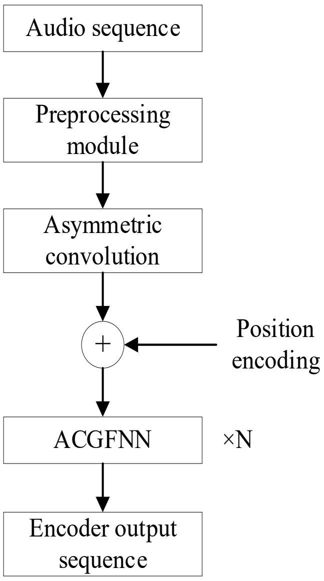


Figure 3. ACGFNN model

3.3. Asymmetric convolution front-end

The asymmetric convolution front-end consists of an asymmetric convolution block (ACB), a fully connected layer (FC), and an activation function GELU. As shown in Figure 4(a), for the audio features with an input frame rate of 10ms, two asymmetric convolution layers with a step size of 2 are used to down-sample the features along the time axis and frequency axis to 40ms respectively, in order to reduce the computational complexity of subsequent steps. Meanwhile, the GELU activation function is used to perform nonlinear transformation on the features. Finally, the down-sampled time-frequency features are converted into time series features through a fully connected layer for output.

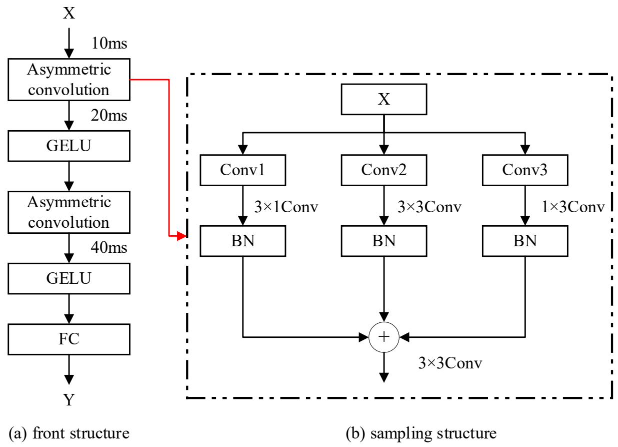
The calculation of the output Y for the input audio feature X is shown in equations (5) to (7).

$$\bar{X} = GELU(ACB(X)) \quad (5)$$

$$\tilde{X} = GELU(ACB(\bar{X})) \quad (6)$$

$$Y = FC(\tilde{X}) \quad (7)$$

Unlike text sequences that only have temporal features, speech sequences possess both temporal and frequency features. Like other AED-based methods, the Conformer model devotes most of its model capacity to temporal feature modeling. In the model, only two sub-sampling layers are used to down-sample the time-frequency features of the speech signal to reduce the computational complexity. Then, these features are directly converted into time-series features for feature extraction by the encoder. This down-sampling strategy is insufficient in extracting the frequency features of the speech sequence, resulting in significant information loss and affecting the recognition accuracy of the model. To address the aforementioned issues, this paper designs a lightweight multi-scale down-sampling structure based on asymmetric convolutional layer, replacing the ordinary down-sampling layer in the original Conformer structure.



(a) front structure
(b) sampling structure

Figure 4. Asymmetric convolutional layer structure

The asymmetric convolution layer employs two one-dimensional convolutions to enhance the convolution of the square frame in both horizontal and vertical directions. Compared with the traditional symmetric convolution layer, this method can enhance the influence of local significant features and has achieved success in many computer vision tasks. The lightweight multi-scale down-sampling structure based on asymmetric convolution layer designed in this paper is shown in Figure 4(b). The asymmetric convolution layer consists of four sub-layers: Conv1, Conv2, Conv3, and the batch normalization layer (BN). Conv1 and Conv3 are one-dimensional asymmetric convolutions with kernel sizes of 3×1 and 1×3 respectively, Conv2 is a square convolution with a kernel size of 3×3, and the batch normalization layer is used to stabilize the feature sequence.

For the input audio features, Conv1 and Conv3 respectively perform down-sampling along their frequency axis and time axis. The sampling results undergo batch normalization to ensure the stability of the feature gradients. Then, the normalized features are fused with the time-frequency features obtained by down-sampling Conv2 simultaneously along the frequency axis and time axis. This processing method enriches the feature space of the audio and enhances the generalization ability and robustness of the model. The calculation of the down-sampling process is shown in Equation (8).

$$I = Conv1(X) + Conv2(X) + Conv3(X) \quad (8)$$

Here, X represents the input audio features. I represents the output result after down-sampling.

Compared with encoder-decoder models based on attention mechanisms (such as Transformer, Conformer, etc.), the proposed method achieves multi-scale down-sampling of input audio features by using stacked asymmetric convolutional layers, thereby extracting and integrating features with different receptive field sizes on the time axis and frequency axis. This new method enhances the model's ability to handle features of different time domains and frequency domain scales. Compared with the convolutional down-sampling layer, this method can extract the time-frequency features of speech sequences more effectively without significantly increasing the number of model parameters. It also reduces the information loss during the down-sampling process and improves the model's expressive power.

3.4. Encoder structure

The encoder is composed of N identical ACGFNN modules. As shown in Figure 5, each encoder layer consists of three sub-modules: the multi-head self-attention module (MHSA), the convolution module (Conv), and the gated feed-forward module (GFFM). Each module uses residual connections and layer normalization (LN) strategies before and after to enhance the performance of the model.

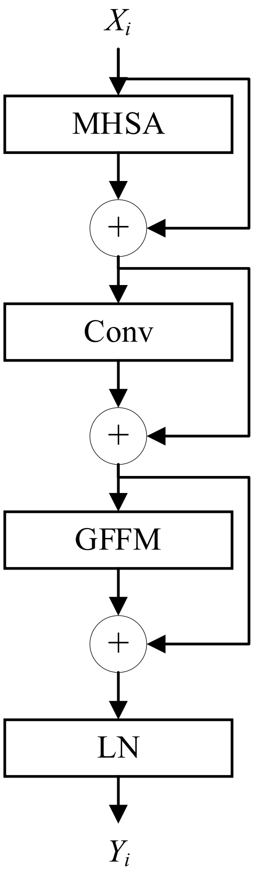


Figure 5. ACGFNN encoder structure

For the input x_i of the $i-th$ layer and the output y_i, the calculation is as shown in equations (9) to (12).

$$\tilde{x}_i = x_i + MHSA(x_i) \tag{9}$$

$$x'_i = \tilde{x}_i + Conv(\tilde{x}_i) \tag{10}$$

$$x''_i = x'_i + GFFM(x'_i) \tag{11}$$

$$\mathrm{y}_i = x''_i + LN(x''_i) \tag{12}$$

The feed-forward module (FFM) in the Conformer architecture has a calculation as shown in equation (13).

$$FFM(x) = Dropout(Swish(xW))W_2 \tag{13}$$

Where, $Dropout(\cdot)$ represents the Dropout operation. $Swish(\cdot)$ represents the Swish activation function. $x \in R^{T\times d}$ indicates a feature sequence with an input length of T and a dimension of d. $W_1 \in R^{d_i\times d_h}$ represents a feed-forward layer with an input dimension of d_i and a hidden layer dimension of d_h. $W_2 \in R^{d_h\times d_i}$ represents a feed-forward layer with an input dimension of d_h and a hidden layer dimension of d_i.

The Conformer model introduces a Macaron-style dual half-step feed-forward module based on the Transformer architecture. Specifically, two half-step feed-forward layers are inserted before the multi-head self-attention module and after the convolution module. Although this method can significantly improve the recognition accuracy, it also adds more network parameters to the already complex Conformer model. In the current context where there is an increasing focus on model light-weighting, excessive network parameters make it difficult for the model to run and deploy in resource-constrained scenarios such as embedded devices or mobile devices. Moreover, larger network models are not conducive to real-time inference, which is crucial for applications that require rapid response and real-time speech recognition systems. To address this issue, this paper replaces the dual feed-forward layer architecture in Conformer with a single gated feed-forward layer.

The gated feed-forward module is a feed-forward neural network that incorporates gated linear units [4] (GLU). This module captures long-term dependencies in the sequence through a gating mechanism, enhancing the processing ability of noise and variations in the speech sequence. The module calculation is shown in equation (14).

$$GFFM(x) = Dropout(GELU(xW_1) \otimes (xV))W_2 \tag{14}$$

Where $W_1 \in R^{d_i\times d_h}$, $V \in R^{d_i\times d_h}$, $W_2 \in R^{d_i\times d_h}$, $GELU$ represents GELU activation function. $\otimes$ indicates element-wise multiplication.

Compared with the original Conformer model that has dual half-step residual connection feed-forward layers, this new method replaces the dual feed-forward module structure in the original model with a single feed-forward module integrated with a gated linear unit. The model is more concise and efficient. It has a smaller number of parameters while achieving better performance.

4. Experiments and Result Analysis

We compare the performance of the Transformer and Conformer model on the AISHELL-1 and aidatatang200zh datasets [29]. The experimental results are shown in Table 1. In terms of the parameter number, compared with the Conformer model, the ACGFNN model reduces the parameter quantity by 12.8%. In terms of experimental performance, the word error rate decreases by 0.44% and 0.4% on the validation set and test set of the AISHELL-1 dataset, and by 0.43% and 0.39% on the aidatatang200zh dataset. The results show that the ACGFNN model significantly reduces the word error rate while maintaining a relatively low parameter quantity. Its excellent performance in different datasets verifies the model's strong generalization ability

Table 1. Comparison of experimental results of each model on different datasets

Model	Parameter size/M	AISHELL-1	aidatatang200zh

		validation	Test	validation	Test
Transformer	30.5	6.22	6.88	6.56	6.80
Conformer	46.3	4.53	4.89	3.97	4.68
ACGFNN	40.4	4.09	4.49	3.54	4.29

To further verify the effectiveness of the proposed method in this paper, experiments are conducted on the mainstream AISHELL-1 dataset in the field of speech recognition, and the experimental results are compared with those of advanced methods. The comparison methods are all encoder-decoder models based on the attention mechanism. The results are shown in Table 2. The experimental results indicate that the ACGFNN model is with only 40.4M parameters, which achieves better word error rates on the test set than other comparison models, it demonstrates the most advanced performance.

Table 2. Comparison of experimental results of each model on the AISHELL-1 dataset

Model	Parameter size/M	Validation	Test
Speech-Transformer [30]	52.2	6.58	7.38
LDSA [31]	49.8	5.80	6.50
ESPnet [32]	46.2	4.84	5.18
Lite-Transformer [33]	51.1	4.71	5.07
WeNet [34]	46.3	4.46	4.62
RoPE [35]	49.5	4.35	4.70
U2 [36]	48.4	4.15	4.64
ACGFNN	40.4	4.09	4.49

We also compare the model training duration and decoding speed of our method with those of the Transformer and Conformer speech recognition models on the AISHELL-1 and aidatatang200zh datasets based on the ESPnet speech recognition framework. For the evaluation of model training duration, the average training time per epoch of each model is used as the measurement standard. In the evaluation of decoding speed, the Real-Time Factor (RTF) of speech recognition is adopted as the assessment criterion, where a smaller RTF value indicates better performance in terms of time consumption of the model. The experimental results are shown in Table 3.

Table 3. Comparison of training time (TT) and RTF

Model	AISHELL-1		aidatatang200zh	
	TT	RTF	TT	RTF
Transformer	10.53	0.29	9.89	0.43
Conformer	13.93	0.40	12.97	0.55
ACGFNN	15.02	0.37	15.24	0.48

According to the experimental results in Table 3, the ACGFNN model outperforms the other two models in terms of training time, but it outperforms the Conformer model in terms of real-time rate for speech recognition. This is mainly because the Conformer model adds convolution and double half-step feed-forward modules in the Transformer framework, thus requiring more training parameters. The proposed ACGFNN model is based on the Conformer model, results in an increase in model complexity. However, by replacing the dual half-step feed-forward module with a single gated feed-forward neural network, it successfully improves the real-time rate of speech recognition. Compared to Conformer, the real-time rate (RTF) index of the ACGFNN method in this study on the AISHELL-1 dataset decreases by 0.03, and on the aidatatang200zh dataset, it decreases by 0.07. These results further prove the superiority of the ACGFNN model in terms of speech recognition speed.

In order to verify the effectiveness of each module in the ACGFNN model and the impact of different encoder layers on the model performance, this paper conducts multiple ablation experiments on the AISHELL-1 dataset. The experimental results are shown in Table 4. Among them, the first row in Table 4 presents the results of the ACGFNN model; the second row uses an ordinary convolutional front-end instead of the asymmetric convolutional front-end; the third row replaces the gated feed-forward with the Conformer dual feed-forward module; the fourth row adds a double half-step gated feed-forward with a Macaron pattern on the basis of the ACGFNN model; the fifth row shows the results of the ACGFN1 model with 15-layer encoder, which has a similar number of parameters to the Conformer model.

Table 4. The ablation experiment results on the AISHELL-1 dataset

Method	Encoder layer	Parameter size/M	Validation	Test
ACGFNN	12	40.4	4.09	4.49
-ACB	12	40.0	4.37	4.75
-GFNN	12	46.7	4.19	7.61
+Mac	12	59.3	5.22	4.58
ACGFN1	15	46.7	4.15	4.55

According to the experimental results in Table 4, the following conclusions can be drawn. Firstly, the the multiple improved methods proposed in this paper can enhance the overall performance of the model. For example, removing the asymmetric convolutional layer can result in 0.26% decrease in word error rate on the test set, and removing the gated feed-forward module can cause 0.12% decrease in word error rate on the test set. This verifies the effectiveness of the proposed module in this paper; secondly, removing the non-symmetric convolutional layer has a greater impact on the model's performance. The non-symmetric convolutional layer enriches the feature space of the audio by multi-scale fusion of the input's time-frequency characteristics, thereby reducing the interference of noise and significantly improving the convergence speed of the model. Furthermore, whether by adding a dual gated feed-forward structure to the ACGFNN model or by using ACGFN1 with an increased number of encoder layers, the model performance will deteriorate.

5. Conclusion

This paper presents a new system that integrates intelligent voice and the Internet of Things (IoT) system. The intelligent voice recognition method is based on asymmetric convolution and gated feed-forward neural networks, effectively enhancing the intelligence level of the IoT system. The proposed intelligent fusion-based new IoT system has significant advantages over traditional IoT systems in terms of the accuracy of voice recognition, the computing capacity of data processing, and the intelligence level of voice interaction. This further validates the accuracy and convenience of interaction of the system.

Acknowledgements.
None.

References

[1] Mu X, Antwi-Afari M F. The applications of Internet of Things (IoT) in industrial management: a science mapping review[J]. International Journal of Production Research, 2024, 62(5): 1928-1952.

[2] Gajić T, Petrović M D, Pešić A M, et al. Innovative approaches in hotel management: integrating artificial intelligence (AI) and the Internet of Things (IoT) to enhance operational efficiency and sustainability[J]. Sustainability, 2024, 16(17): 7279.

[3] Yin S, Li H, Laghari A A, et al. An anomaly detection model based on deep auto-encoder and capsule graph convolution via sparrow search algorithm in 6G Internet of Everything[J]. IEEE Internet of Things Journal, 2024, 11(18): 29402-29411.

[4] Popoola O, Rodrigues M, Marchang J, et al. A critical literature review of security and privacy in smart home healthcare schemes adopting IoT & blockchain: problems, challenges and solutions[J]. Blockchain: Research and Applications, 2024, 5(2): 100178.

[5] Amru M, Kannan R J, Ganesh E N, et al. Network intrusion detection system by applying ensemble model for smart home[J]. International Journal of Electrical and Computer Engineering, 2024, 14(3): 3485-3494.

[6] Yin S, Li H, Laghari A A, et al. FLSN-MVO: edge computing and privacy protection based on federated learning Siamese network with multi-verse optimization algorithm for industry 5.0[J]. IEEE Open Journal of the Communications Society, 6:3443-3458, 2024.

[7] Zhang X, Wang X, Jia Y. The visual internet of things system based on depth camera[C]//Proceedings of 2013 Chinese Intelligent Automation Conference: Intelligent Automation & Intelligent Technology and Systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013: 447-455.

[8] Zhang L, Wu H. Application of single chip technology in internet of things electronic products[J]. Journal of Intelligent & Fuzzy Systems, 2021, 40(2): 3223-3233.

[9] Juluru T K, Golamari J M, Chitti S, et al. Human-Computer Interaction in Audio Systems: An IoT-Based Gesture Control Approach[C]//2025 7th International Conference on Inventive Material Science and Applications (ICIMA). IEEE, 2025: 795-802.

[10] Zhou F Y, Li J H, Tian G H, et al. Research and Implementation of Embedded Voice Interaction System Based on ARM in Intelligent Space[J]. Advanced Materials Research, 2012, 433: 5620-5627.

[11] Ahmed G, Laway e A A. CNN-based speech segments endpoints detection framework using short-time signal energy features[J]. International Journal of Information Technology, 2023, 15(8): 4179-4191.

[12] Wu Q, Liu Y. A speech endpoint detection method based on cascaded speech enhancement[C]//2021 International Conference on Electronic Information Technology and Smart Agriculture (ICEITSA). IEEE, 2021: 1-6.

[13] Jiang Y, Yin S. Heterogenous-view occluded expression data recognition based on cycle-consistent adversarial network and K-SVD dictionary learning under intelligent cooperative robot environment[J]. Computer Science and Information Systems, 2023, 20(4): 1869-1883.

[14] Chao L, Chen J, Chu W. Variational connectionist temporal classification[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 460-476.

[15] Du S, Li T, Yang Y, et al. Multivariate time series forecasting via attention-based encoder–decoder framework[J]. Neurocomputing, 2020, 388: 269-279.

[16] Ji Z, Xiong K, Pang Y, et al. Video summarization with attention-based encoder–decoder networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(6): 1709-1717.

[17] Kamal M B, Khan A A, Khan F A, et al. An Innovative Approach Utilizing Binary-View Transformer for Speech Recognition Task[J]. Computers, Materials & Continua, 2022, 72(3).

[18] Lo W C, Wang W J, Chen H Y, et al. Feasibility study regarding the use of a conformer model for rainfall-runoff modeling[J]. Water, 2024, 16(21): 3125.

[19] Chen S, Wu Y, Chen Z, et al. Continuous speech separation with conformer[C]//ICASSP 2021-2021 IEEE International

Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 5749-5753.

[20] Chan T K, Chin C S. Multi-branch convolutional macaron net for sound event detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 2972-2985.

[21] Rajab M A, Abdullatif F A, Sutikno T. Classification of grapevine leaves images using VGG-16 and VGG-19 deep learning nets[J]. TELKOMNIKA (Telecommunication Computing Electronics and Control), 2024, 22(2): 445-453.

[22] Aich U, Saha A, Woźniak M, et al. Schizophrenia detection from electroencephalogram signals using image encoding and wrapper-based deep feature selection approach[J]. Scientific Reports, 2025, 15(1): 21390.

[23] Berghi D, Wu P, Zhao J, et al. Fusion of audio and visual embeddings for sound event localization and detection[C]//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 8816-8820.

[24] Yu W, Zhu M, Wang N, et al. An efficient transformer based on global and local self-attention for face photo-sketch synthesis[J]. IEEE Transactions on Image Processing, 2022, 32: 483-495.

[25] Burchi M, Vielzeuf V. Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition[C]//2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2021: 8-15.

[26] El Mattar S, Baghdad A. Beyond Traditional RFID: Unveiling the Potential of Wi-Fi, 5G, Bluetooth, and Zigbee for Backscatter Systems[J]. Transactions on Emerging Telecommunications Technologies, 2025, 36(2): e70062.

[27] Varriale V, Cammarano A, Michelino F, et al. Critical analysis of the impact of artificial intelligence integration with cutting-edge technologies for production systems[J]. Journal of Intelligent Manufacturing, 2025, 36(1): 61-93.

[28] Abatal A, Mzili M, Mzili T, et al. Intelligent Interconnected Healthcare System: Integrating IoT and Big Data for Personalized Patient Care[J]. International Journal of Online & Biomedical Engineering, 2024, 20(11).

[29] Bu H, Du J, Na X, et al. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline[C]//2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA). IEEE, 2017: 1-5.

[30] Dong L, Xu S, Xu B. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition[C]//2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018: 5884-5888.

[31] Xu M, Li S, Zhang X L. Transformer-based end-to-end speech recognition with local dense synthesizer attention[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 5899-5903.

[32] Li C, Shi J, Zhang W, et al. ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration[C]//2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021: 785-792.

[33] Wenxuan Z, Yaqin Z, Zhaoxiang Z, et al. Lite transformer network with long–short range attention for real-time fire detection[J]. Fire Technology, 2023, 59(6): 3231-3253.

[34] Ma J, Reda S. WeNet: Configurable Neural Network with Dynamic Weight-Enabling for Efficient Inference[C]//2023 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED). IEEE, 2023: 1-6.

[35] Liu B, Han Z, Chen X, et al. Rope-net: deep convolutional neural network via robust principal component analysis[J]. Machine Learning, 2025, 114(7): 150.

[36] Tsunoo E, Futami H, Kashiwagi Y, et al. Decoder-only architecture for streaming end-to-end speech recognition[J]. arXiv preprint arXiv:2406.16107, 2024.