CIT. Journal of Computing and Information Technology, Vol. 34, No. 3, September 2026, 203–218
doi: 10.20532/cit.2026.1006584

Optimization and Development of Edge-Embedded Speech Recognition Module of Improved DFCNN-CTC

Fang Nie[1] and Lijuan Zheng[2]

[1] Chengdu Jincheng University; Chengdu, Sichuan Province, China
[2] School of Artificial Intelligence and Electronic Engineering, Sichuan Technology and Business University, Chengdu, Sichuan Province, China

In view of the challenges faced by edge-device speech recognition of smart toys, such as limited computing resources, environmental noise interference, and specificity of children's pronunciation, this study proposes a new lightweight speech recognition model. The study uses original spectrograms as input and introduces transformers to improve the deep full-sequence convolutional neural network-connected temporal classification model to achieve high-precision end-to-end recognition. This model uses dual-student difference structure mutual learning, sequence-level distillation, and KL imitation loss to narrow the performance gap between the student model and the teacher model. Experiments demonstrated that the improved model's word error rate on the children's speech data set reached 8.2%, which was 3.9%~6.1% lower than Conformer, Branchformer, and other models. After the lightweight improvement, the model size was reduced from 125MB to 24MB, with an increase in WER of only 0.7%. After further compression, the model size was reduced to 8.1 MB, the inference delay was reduced to 155ms, the average power consumption was 45mW, and the energy efficiency ratio was improved by 4.4 times. In the home environment test, the wake-up success rate was 96.5%, the command recognition accuracy rate was 93.2%, and the false wake-up rate was 0.8 times/hour. The proposed method can improve the speech recognition efficiency of smart toys in edge environments and reduce their energy consumption.

ACM CCS (2012) Classification: Computing methodologies → Artificial intelligence → Natural language processing → Speech recognition

Keywords: smart toys, edge-embedding, speech recognition, DFCNN-CTC, mutual learning of knowledge distillation

1. Introduction

As a typical product integrating artificial intelligence and consumer electronics, smart toys have seen rapid growth in market size in recent years [1]. According to industry data, the global AI toy market is expected to exceed US$60 billion in 2033 [2]. However, existing products commonly have problems such as low speech recognition accuracy, poor adaptability to environmental noise, high power consumption, and strong dependence on the cloud [3]. 68% of consumers report that smart toys can only recognize fixed instructions, and dialogue is like reciting lines [4]. Children's speech has the characteristics of high pitch, large acoustic variation, and non-standard pronunciation. The misjudgment rate of general speech recognition models in toy scenes exceeds 30%, and the effective recognition rate in complex family environments is only about 65% [5]. At the same time, toy products have rigid constraints on the low power consumption, offline availability, and privacy security of speech recognition modules, requiring speech recognition algorithms to achieve efficient reasoning under limited end-side computing power [6].

To improve the efficient fusion of convolution and self-attention in automatic speech recognition, Kim et al. proposed a new E-Branchformer model, which enhanced Branchformer by introducing a more effective fusion method and stacking point-by-point modules. On LibriSpeech's test-clean and test-other test sets, this

method achieved word error rates of 1.81% and 3.65% respectively, setting a new state-of-the-art level without using any external training data [7]. Voice communication is susceptible to interference in noisy environments. Therefore, Liu et al. proposed a new skin-attached acoustic sensor. The sensor used a high-sensitivity, wide-bandwidth piezoelectric micromachined ultrasound transducer combined with a flexible package to capture the vibrations of the sound-producing organs and skin movements. Experiments demonstrated that the system's accuracy in classifying laryngeal speech features exceeded 96%, and its accuracy in recognizing daily sentences reached 99.8%. This solution had the advantages of simple process, stable performance, easy integration, and low cost [8]. Yao et al. proposed a faster, memory-saving, and better-performing model to address the deficiencies in efficiency and performance of traditional encoders in automatic speech recognition. The model adopted a U-Net coding structure, reorganized modules, and reused attention weights, and introduced a bias normalization layer. Experiments on multiple data sets demonstrated that this model significantly outperformed existing mainstream speech models [9]. Aiming at the expensive and time-consuming problem of annotating data in audio and video speech recognition, Ma et al. proposed a method of using pre-trained ASR models to automatically transcribe unlabeled data sets to expand the training set. This method added automatic transcription data to the basic data set, and the model reached a word error rate of 0.9% in the new data set, which was a relative improvement of 30% compared to the current optimal method, and the performance was better than the model trained using 26 times non-public data [10]. Fathullah et al. proposed a small new audio encoder to be applied in the field of speech recognition. This encoder prepended a sequence of audio embeddings to a sequence of text embeddings to perform automatic speech recognition. Experiments on the Multilingual LibriSpeech data set indicated that the improved model improves performance by 18% compared to the monolingual baseline model and could achieve multilingual speech recognition [11]. To optimize the training and inference efficiency of the Conformer model, Rekesh et al. proposed a new Fast Conformer model. The model used limited contextual attention to replace global attention and added global token fine-tuning, redesigning the architecture through a down sampling scheme. Experiments demonstrated that this model was 2.8 times faster than the original Conformer, supported expansion to one billion parameters, and achieved the most advanced accuracy on automatic speech recognition benchmarks [12].

To sum up, existing studies have achieved substantial progress in speech recognition. However, artificial feature extraction like acoustic model Mel frequency cepstral coefficients and Mel filter bank features are prone to losing high-frequency information in traditional speech recognition methods, and N-gram in language models is difficult to capture long-distance semantic dependencies. Therefore, the research proposes an improved model of Deep Fully Convolutional Neural Network-Connectionist Temporal Classification (DFCNN-CTC), which aims to improve the adaptability of the speech recognition environment and recognition robustness, achieve real-time response, and enhance the interactive experience. The innovation of the research lies in replacing artificial features with original spectrograms to retain complete frequency domain information, using DFCNN to accumulate long-term context, combining CTC to achieve end-to-end alignment, and introducing a Transformer language model to replace N-grams to effectively capture long-distance semantic dependencies. The study designs a Mutual Learning of Knowledge Distillation (MLKD) framework using a wide-deep dual student model structure (wide network captures local features, deep network models long-term dependencies), and achieves complementary learning between student models through sequence-level knowledge distillation and KL divergence imitation loss.

2. Research Methodology

2.1. Improved DFCNN-CTC smart toy speech recognition based on transformer

Traditional speech recognition systems mostly use an architecture of acoustic model plus language model. The acoustic model faces in-

formation loss caused by artificial feature extraction, and the N-gram in the language model is difficult to capture long-distance semantic dependencies [13]. Therefore, the study proposes a speech recognition model based on Transformer improved DFCNN-CTC, as illustrated in Figure 1.

In Figure 1, the model includes three parts: preprocessing, acoustic model, and language model. The original spectrogram is used as the input of the DFCNN module, which can avoid the loss of high-frequency information. The study uses deep convolutional pooling structures to accumulate long-term context, and combines CTC decoding to achieve end-to-end aligned output. Finally, Transformer is introduced to replace the N-gram module to fully model sequence semantics through the self-attention mechanism. Speech data preprocessing is a key step in converting original digital signals into computer-recognizable feature vectors, including frame windowing and acoustic feature extraction. The speech signal has short-term stationarity, and its characteristics are relatively stable within the range of 10 to 30 ms. Therefore, the signal needs to be divided into short-term frames for analysis. The study adopts the overlapping segmentation method, setting the frame length to 25ms and the frame shift to 10ms to make overlapping transitions between frames and to avoid the loss of speech information [14]. However, segmenting the speech signal directly will produce a truncation effect and lead to spectrum information leakage, so window processing is required. The research uses a Hamming window that is smoother and has better expressive ability of speech spectrum characteristics, and the calculation is demonstrated in Equation (1) [15].

$$W(t)=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi t}{N-1}\right), 0\le t\le N-1\\ 0, other\end{cases}. \quad (1)$$

In Equation (1), $W(t)$ signifies the value of the window function at time t. N signifies the length of the window function, that is, the number of sampling points. The acoustic feature extraction efficiency of the speech signal processed by frame and window processing is higher. The study uses spectrogram features to reduce the computational complexity of feature extraction. Fast Fourier transform of the signal is required

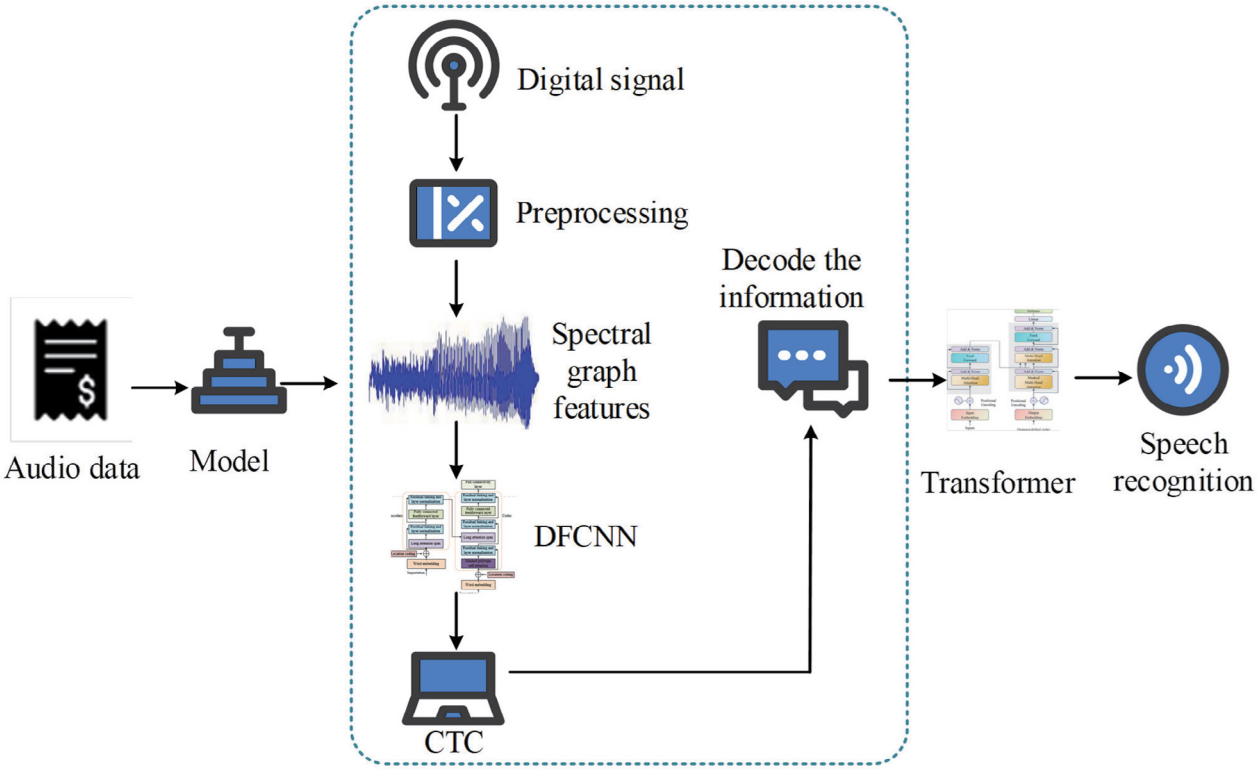


Figure 1. Specific structure of the T-DFCNN-CTC speech recognition model.

during extraction, and the calculation is demonstrated in Equation (2) [16].

$$P(i)=\sum_{n=0}^{N-1}S(i)^{-j\frac{2\pi}{N}kn}. \quad (2)$$

In Equation (2), $P(i)$ represents the frequency domain representation of the speech signal after fast Fourier transform, $S(i)$ represents the time domain speech signal of frame i, j represents the imaginary unit $\sqrt{-1}$, and k represents the frequency domain index. According to the Nyquist theorem, the frequency range of a speech signal with a sampling rate of 16 kHz after FFT is 0 to 8 kHz. After completing the FFT, the phase spectrum containing a small amount of information is discarded, the amplitude spectrum is retained, and the spectral line energy is obtained [17]. The DFCNN-CTC module in the T-DFCNN-CTC speech recognition model is responsible for mapping the processed spectrogram features into pinyin sequences. Figure 2 presents the specific structure of the DFCNN-CTC.

In Figure 2, the input layer is responsible for receiving spectrogram features, normalizing the data, and using a zero-padding layer to standardize the length of all sentences. The convolutional pooling layer includes five groups of convolutional neural networks and maximum pooling structure. The specification of each convolutional neural network is 3×3, and the maximum pooling layer reduces the number of parameters. After multi-layer convolution pooling, the model can accumulate enough historical and future information to effectively express the long-term correlation of speech signals. The model employs a fully connected layer to connect all neurons of the previous pooling layer to integrate the classification features extracted by the convolutional pooling layer, using ReLU as the activation function. The CTC output layer outputs the predicted pinyin sequence. The decoding method is beam search and introduces blank placeholders to achieve alignment of the input and output sequences [18]. In speech recognition, the language model can convert the pinyin sequence output by the acoustic model into the final text. The text conversion is demonstrated in Equation (3) [19].

$$W^{*}\approx\arg\max_{W}P(X|W)P(W). \quad (3)$$

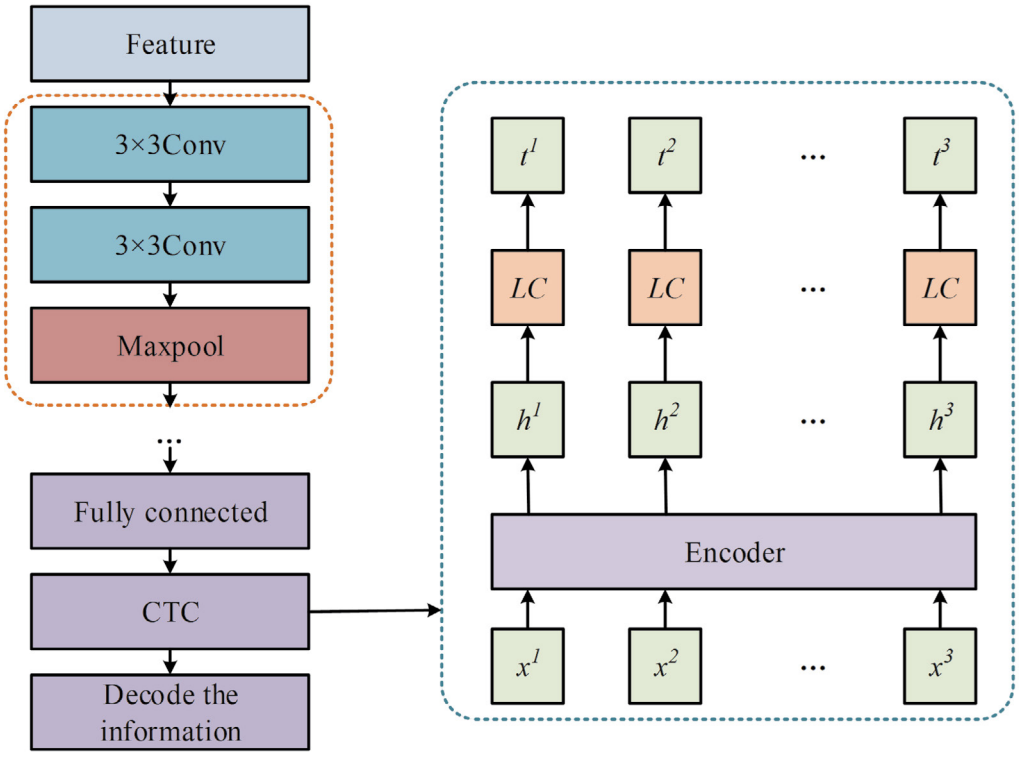


Figure 2. Specific structure of the DFCNN-CTC module.

In Equation (3), W^* represents the optimal text sequence, W represents the candidate text sequence, X represents the input acoustic feature vector, $P(X|W)$ signifies the acoustic probability model, and $P(W)$ signifies the language probability model. To solve the problem that N-gram cannot capture the semantic similarity between words, and the parameter space expands sharply as the order increases, the study utilizes the Transformer based on the self-attention mechanism to replace N-gram. Figure 3 illustrates the specific structure.

In Figure 3, the input terminal is the data sequence output by DFCNN-CTC. The text embedding layer is used for high-dimensional vector conversion, and the position information is added through the position encoding layer, which are mixed and sent to the encoder. The output of self-attention utilizes all input sequence information and can be calculated in parallel. However, the fitting ability of a single self-attention head is limited. The research uses multiple parallel self-attention heads to extract features from different representation subspaces and finally splices and linearly transforms to obtain the output result. The calculation is demonstrated in Equation (4) [20].

$$\begin{cases} head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \\ MultiHead\left(Q,K,V\right) = Concat\left(head_1,...,head_n\right)W^O \end{cases} \tag{4}$$

In Equation (4), $head_i$ signifies the output of the ith attention head, Q, K and V signify the query, key, and value matrices. W_i^Q, W_i^K and W_i^V signify the learnable weight matrices of the ith attention head of the three matrices. *MultiHead* represents the final output of multi-head attention, *Concat* represents feature splicing, and W^O represents the linear transformation matrix of the output. There are six sets of identical modules in the encoder, and each module contains multi-head attention, residual connection, normalization, and feed-forward neural network layers. Residual connections prevent gradient disappearance, and normalization prevents data deviation from being too large. Based on the encoder, the feedforward network decoder adds a multi-head self-attention structure to ensure that only existing output data is queried when predicting the current position to

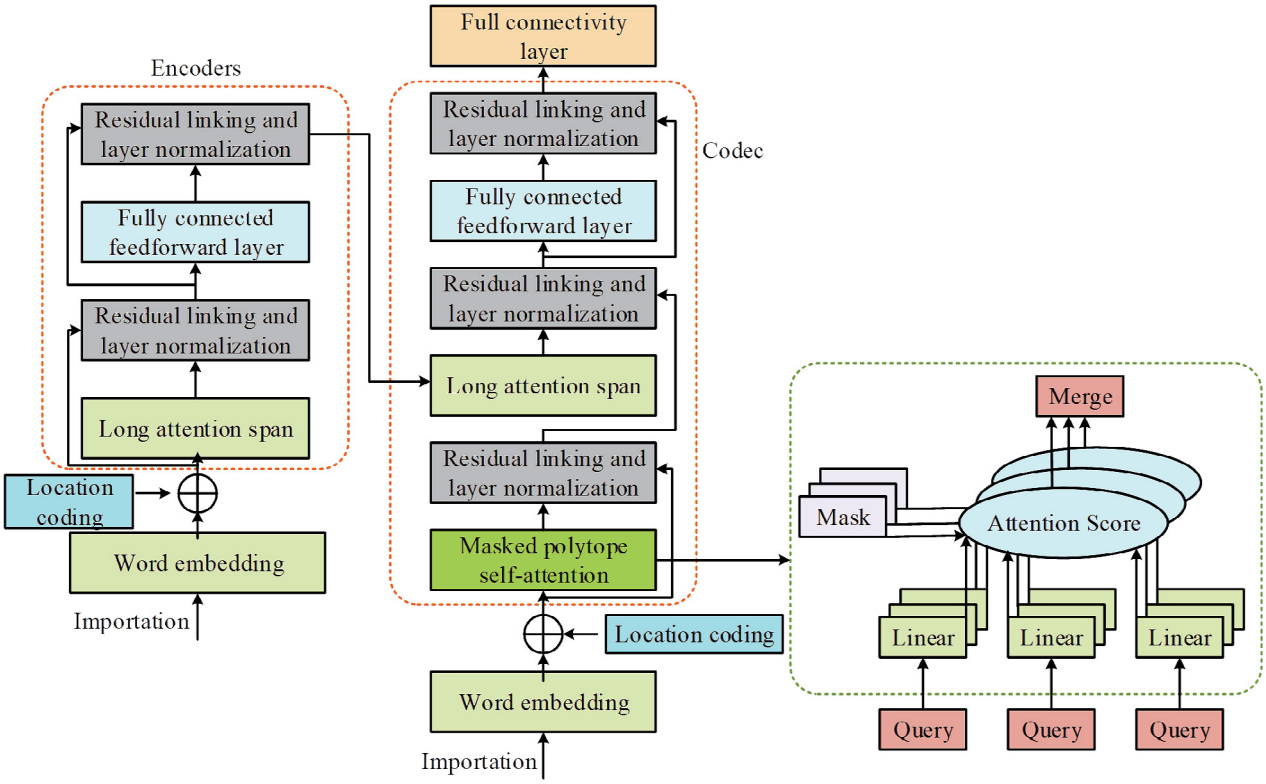


Figure 3. Specific structure of the self-attention-based Transformer.

prevent future information leakage. The decoder uses an autoregressive method to gradually generate the output sequence and finally outputs the probability distribution of the word through the linear layer and softmax layer.

2.2. Edge-embedding optimization for speech recognition based on knowledge distillation

The T-DFCNN-CTC model can achieve good performance in general speech recognition tasks. However, in resource-constrained edge devices (such as smart toys, *etc.*), the model needs to compress further its volume while maintaining high recognition accuracy [21]. Therefore, the study introduces the MLKD architecture to improve the performance of speech recognition tasks through knowledge sharing and mutual learning between models. The specific structure is demonstrated in Figure 4.

In Figure 4, two student models with different structures not only learn from the teacher model during the training stage, but can also communicate with each other and share knowledge, making up for their respective deficiencies through complementary learning mechanisms. The network structure of student 1 is wider, and the network structure of student 2 is deeper. Compared with learning the knowledge of a fixed teacher model, this mutual learning method can further narrow the performance gap between student and teacher models. Moreover, the study introduces sequence-level knowledge distillation. By comparing the prediction results with the ground-truth label sequence and calculating the corresponding error, the student model is guided to adjust its output, so that the student model can gradually approach the performance of the teacher model. Each student model is optimized and trained using three loss functions. The CTC loss is used to ensure the end-to-end optimization capability of the model in speech recognition tasks. The sequence-level knowledge distillation loss function allows the student model to learn richer feature representations. The imitation loss function aligns the class probability distributions between different student models. The total loss function is calculated as demonstrated in Equation (5).

$$L_{stu} = (1-\alpha)L_C + \alpha(L_S + \beta L_K) \qquad (5)$$

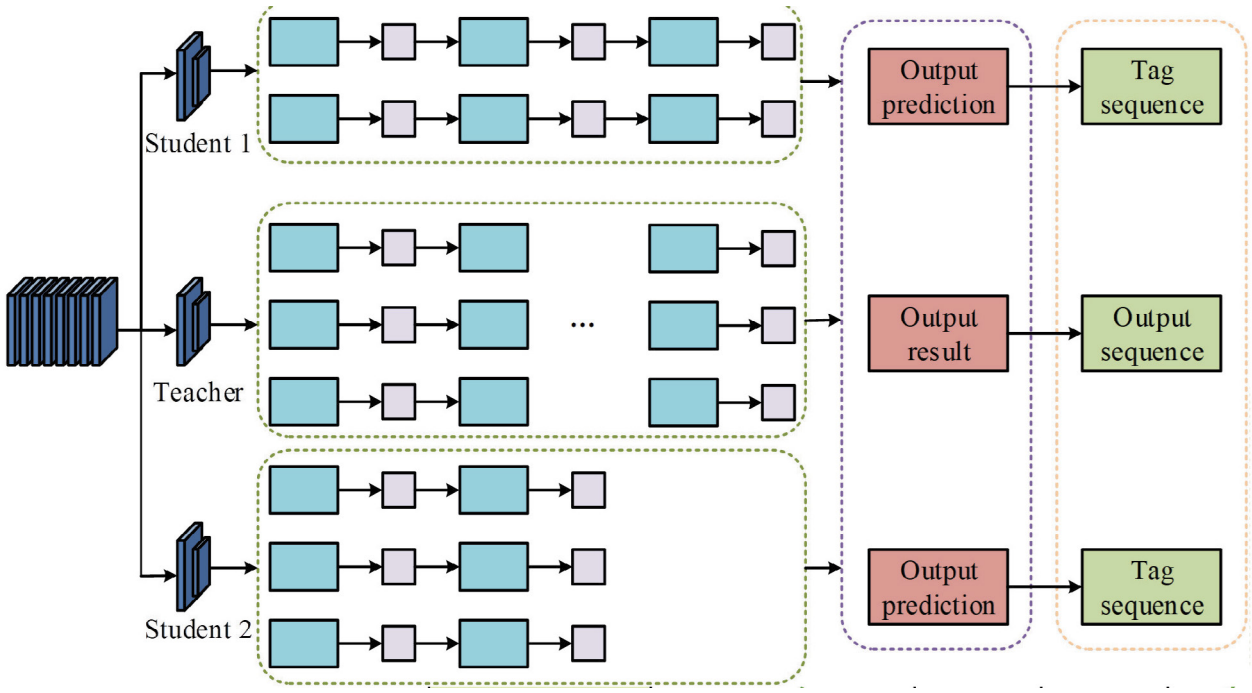


Figure 4. Specific structure of MLKD.

In Equation (5), L_{stu} represents the total loss function, L_C represents the CTC loss, L_S represents the sequence-level knowledge distillation loss, L_K represents the imitation loss, α represents the balance weight coefficient between the CTC loss and the sequence-level knowledge distillation loss, and β represents the imitation loss weight coefficient. The three loss functions are calculated as shown in Equation (6) [22].

$$\begin{cases} L_C = -\ln \sum\limits_{\pi \in B^{-1}(Y)} p(\pi|X) \\ L_S = -\dfrac{1}{T}\sum\limits_{t=1}^{T}\sum\limits_{c=1}^{C} p_t^{tea}(c)\cdot \log p_t^{stu}(c) \\ L_K = \dfrac{1}{T}\sum\limits_{t=1}^{T} KL\left(p_t^{stu,1} \middle\| p_t^{stu,2}\right) \end{cases} \tag{6}$$

In Equation (6), $B^{-1}(Y)$ represents the set of all paths that can obtain the real label Y sequence by removing repeated labels and blanck space characters. $p(\pi|X)$ represents the probability of the model predicting path π. T signifies the number of time steps of the speech feature sequenc′ C represents the total number of output categories. $p_t^{tea}(c)$ signifies the probability that the teacher model predicts category c at t. $p_t^{stu}(c)$ signifies the probability that the student model predicts category c at t. $KL(\cdot)$ signifies KL divergence. $p_t^{stu,1}$ signifies the output probability distribution of student model 1 at t. $p_t^{stu,2}$ represents the output probability distribution of student model 2 at t. The loss function L_{stu}^m of the improved m learning model is calculated as demonstrated in Equation (7) [23].

$$\begin{cases} L_{stu}^m = (1-\alpha)L_c + \alpha\left(\sum\limits_{i\neq m}^{T_i}\lambda_{i,m}L_S + \beta\sum\limits_{i\neq m}^{T_i} L_K\right) \\ \lambda_{i,m} = \dfrac{\exp(w_{i,m})}{\sum\limits_{m\neq i}^{T_m}\exp(w_{i,m})} \end{cases} \tag{7}$$

In Equation (7), T_i signifies the total teacher models. $\lambda_{i,m}$ signifies the learnable distillation weight of the ith teacher to the mth student model, and $w_{i,m}$ represents the parameters automatically optimized during the training process.

2.3. Device-side deployment and engineering optimization for smart toys

After the model compression was completed, further engineering optimization was required for the edge environment of the smart toy. To adapt the MLKD compressed speech recognition model to the edge environment of smart toys, adjustments needed to be made in aspects such as hardware quantization, instruction customization, power consumption optimization, and mass production adaptation [24]. Smart toys used ARM Cortex-M series edge chips with a computing power of 0.5 ~ 2TOPS and power consumption requirements of <1W. The study used quantization-aware training on the distilled student model, mapping the weights and activations from FP32 to INT8, so that the model volume was compressed from 120MB of the original teacher model to 8MB [25]. At the same time, channel pruning was performed on the convolutional layer to remove redundant parameters and further reduced the amount of inference calculations.

Smart toys are powered by rechargeable nickel-metal hydride batteries and are sensitive to power consumption. Optimization is performed at both the system and algorithm levels. The chip operating frequency is dynamically adjusted based on the voice activity status. When the voice detection module detects no valid voice input for a long time (>5 seconds), the system switches to deep sleep mode, shuts down the analog-to-digital converter, clock, and most devices, leaving only the low-power listening circuit [26]. After detecting the wake-up word, it quickly returns to the high-performance operating frequency within 10ms to achieve instantaneous "sleep-wake" switching. The research uses operator fusion and memory reuse (intermediate tensors share the same memory pool) technology to reduce data handling overhead. For continuous instruction scenarios, research is conducted on merging adjacent short speech frames into micro-batches for reasoning to reduce the number of model calls. When the signal-to-noise ratio is lower than the threshold, high-noise frames are selectively discarded to further reduce the amount of calculation while ensuring the recognition rate.

3. Results and Discussion

3.1. Experimental analysis of T-DFCNN-CTC

The training platform was Intel Xeon Gold 6248 CPU (40 cores), 128 GB RAM, NVIDIA Tesla V100 (32 GB) GPU, used to train the T-DFCNN-CTC model. The T-DFCNN-CTC model used the PyTorch framework and Adam optimizer. The initial learning rate was set to 0.0001 and the model was trained for 200 rounds. The data sets used in the experiment included ASR-SCEChilSC (https://magichub.com/cn/datasets/chinese-english-scripted-speech-corpus-children/#content) and Chinese Kids Speech DB (https://catalog.elra.info/en-us/repository/browse/ELRA-S0496/). ASR-SCEChilSC is an open-source Chinese-English bilingual children's speech data set, which lasts 1.44 hours and has a total of 2266 sentences. Chinese Kids Speech DB is a large-scale Chinese children's reading data set, with a duration of 87 hours and a total of 43,667 items. The audio parameters of both data sets were 16kHz, 16bit. The study randomly divided the two data sets into training, test, and validation sets in a 7:2:1 ratio. The comparative methods used in the study included Convolution-augmented Transformer for Speech (Conformer), Parallel MLP-Attention Architectures to Capture Local and Global Context for Speech Recognition and Understanding (Branchformer), Branchformer with Enhanced Merging for Speech Recognition (E-Branchformer). A comparison of the speech recognition performance of the different models is shown in Figure 5.

In Figure 5(a), the word error rate of the T-DFCNN-CTC model decreased the fastest and tended to converge at 40 iterations. The minimum word error rate was 8.2%, which was 6.1%, 4.3%, and 3.9% lower than Conformer, Branchformer, and E-Branchformer, respectively. In Figure 5(b), the word error rates of all models facing the test set increased, with the T-DFCNN-CTC model increasing by 0.3%, and the second-best E-Branchformer increasing by 3.0%. The comparison of the recall rates of the T-DFCNN-CTC model in different types of data sets is demonstrated in Figure 6.

In Figure 6(a), in Chinese Kids Speech DB, the T-DFCNN-CTC model had excellent recall rate, with a maximum recall rate of 88.5%, which was 8.3% higher than the second-best E-Branchformer (80.2%). For the mixed Chinese-English dataset shown in Figure 6(b) (ASR-SCEChilSC), the T-DFCNN-CTC model recall rate dropped by 12.1%. This decrease may be caused by interference between adjacent Chinese and English words, leading to recognition errors.

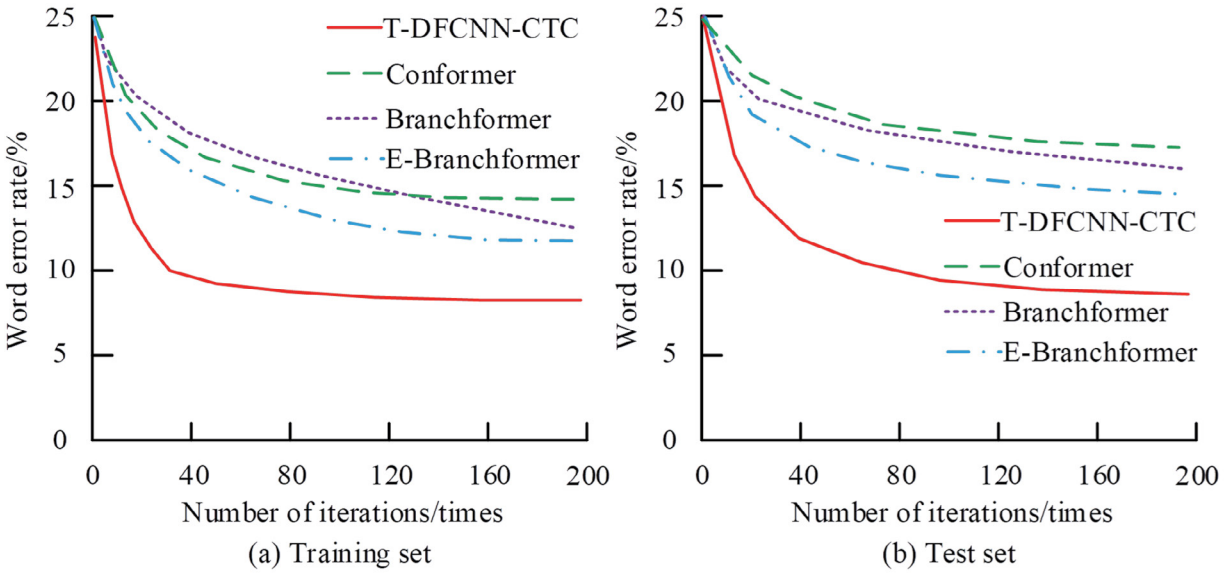


Figure 5. Comparison of speech recognition performance of different models.

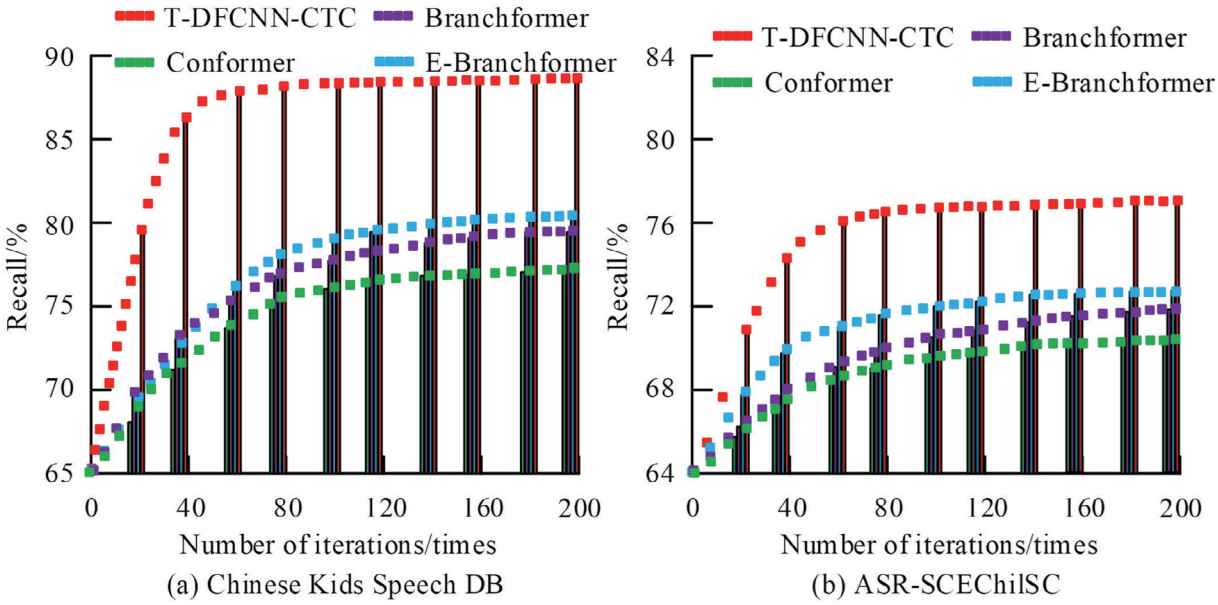


Figure 6. Comparison of recall rates of the T-DFCNN-CTC model on different types of datasets.

3.2. Experimental analysis of edge embedding speech recognition model

The study introduced two lightweight speech recognition models, Unified Two-pass Bidirectional End-to-end Model for Speech Recognition (Unified Two-pass), and Moonshine: Speech Recognition for Live Transcription and Voice Commands (Moonshine), for experimental testing. The robustness comparison of the model under different SNR is demonstrated in Figure 7.

In Figure 7(a), under white noise, as the signal-to-noise ratio increased, the word error rates of different models gradually decreased, and the T-DFCNN-CTC model decreased the fastest. At 10dB, the word error rate of the T-DFCNN-CTC model was 18.1%, which was 3.7% and 2.1% lower than the Unified Two-pass and Moonshine models, respectively. In Figure 7(b), the TV background sound had the greatest impact on the speech recognition performance. At 0dB, the word error rates of the three models were all above 30%. In Figure 7(c), under the influence of toy music, the T-DFCNN-CTC model still achieved the lowest word error rate, reaching 11.3% at 25dB. The comparison of knowledge distillation compression effects is demonstrated in Figure 8.

In Figure 8(a), T-DFCNN-CTC-1 represented single-student distillation, and T-DFCNN-CTC-2 represented dual-student mutual learning. The word error rate of the T-DFCNN-CTC model was higher than that of the two distilled student models and gradually decreased as the speech duration increased. At 4s, the word error rate was slightly higher than that of 2s. This may be due to a larger number of adjacent Chinese and English words, which reduced speech recognition accuracy. In Figure 8(b), T-DFCNN-CTC-2 could effectively reduce the model performance degradation caused by knowledge distillation. Its word error rate was much lower than T-DFCNN-CTC-1 and slightly higher than the teacher model T-DFCNN-CTC. The results of the edge embedding speech recognition model ablation experiment are demonstrated in Table 1.

In Table 1, the FLOPs of configuration F were reduced from 1840M to 460M, a reduction of 75%, and the inference delay was shortened from 668ms to 155ms, which met the real-time interaction requirements of toys (<200ms). The average power consumption dropped from 85mW to 45mW, a reduction of 47%, and the energy efficiency ratio increased from 1.21 times/Wh to 5.33 times/Wh, an increase of 4.4 times. This meant that with the same battery capacity, the proposed solution could support more

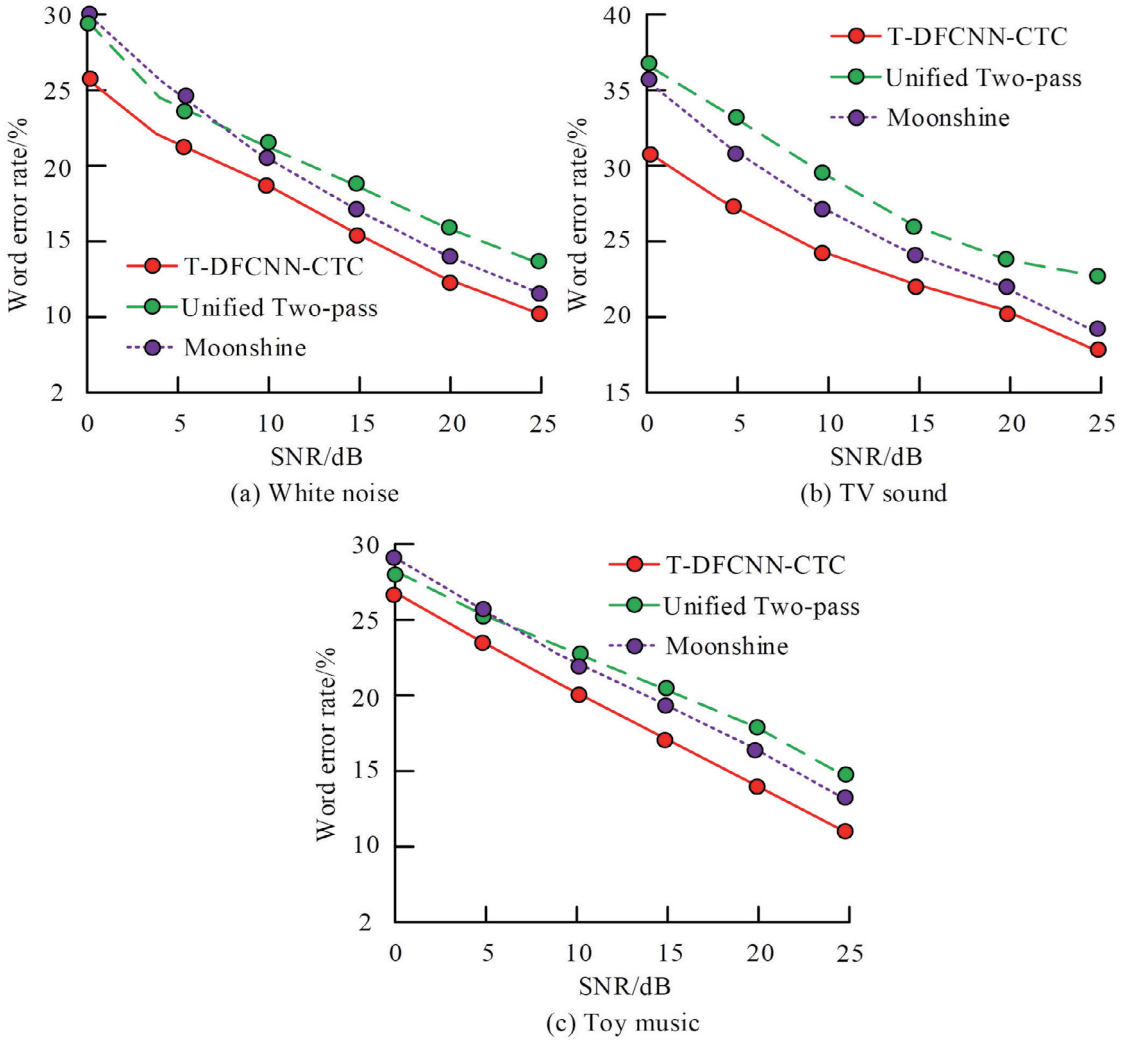


Figure 7. Comparison of robustness of models under different signal-to-noise ratios.

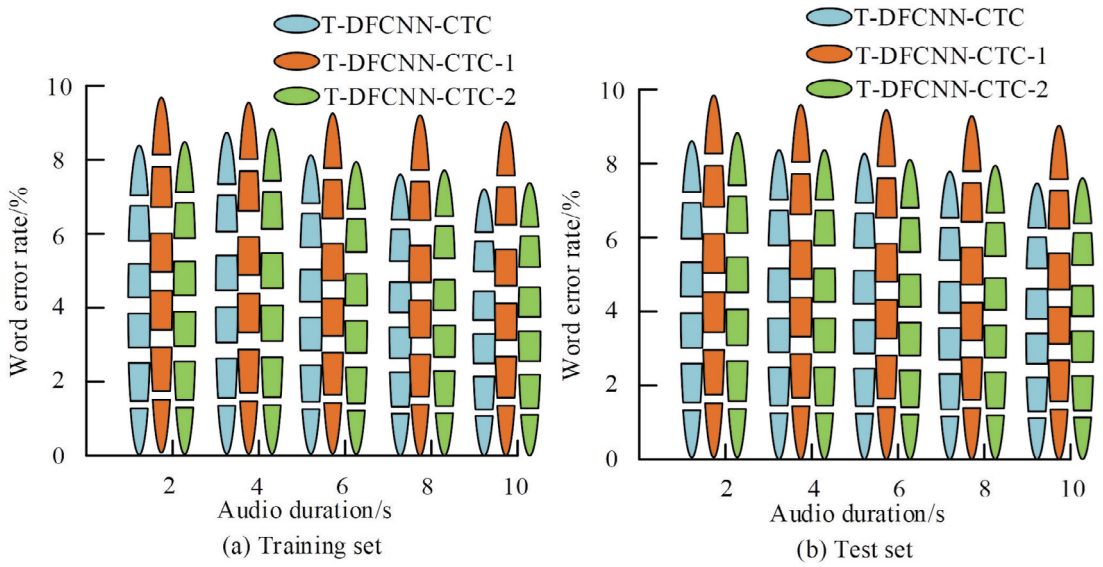


Figure 8. Comparison of knowledge distillation compression effects.

Table 1. Ablation experiment results of edge-embedded speech recognition models.

Configuration	WER (%)	Volume (MB)	Parameter quantity (M)	Memory usage (KB)	Inference delay (ms)	FLOPs (M)	Average power consumption (mW)
A (T-DFCNN-CTC)	9.5	125	125.0	3229	320	12500	280
B (Single Student)	13.8	24	24.2	620	680	1850	89
C (Two Students, without Sequence-level Knowledge Distillation)	15.2	26	26.1	670	695	1920	92
D (Two Students, without KL)	11.5	25	24.8	635	672	1880	87
E (Complete MLKD)	10.2	24	24.0	615	668	1840	85
F (Complete MLKD + INT8)	11.0	8.1	36.0	192	155	460	45

than 4 times the number of interactions as the single-student solution. Mutual learning, sequence-level knowledge distillation, and KL divergence jointly ensured the accuracy of the compressed model.

3.3. Smart toy performance test

The experiment recruited 100 children aged 2-10 years (50 boys and 50 girls). Young children aged 2-4 years need to be able to express simple vocabulary. All children conducted free interaction tests in their respective home living room environments, and a total of 1,200 effective interaction sessions were collected. The accuracy of recognition of toy real scene instructions is demonstrated in Table 2.

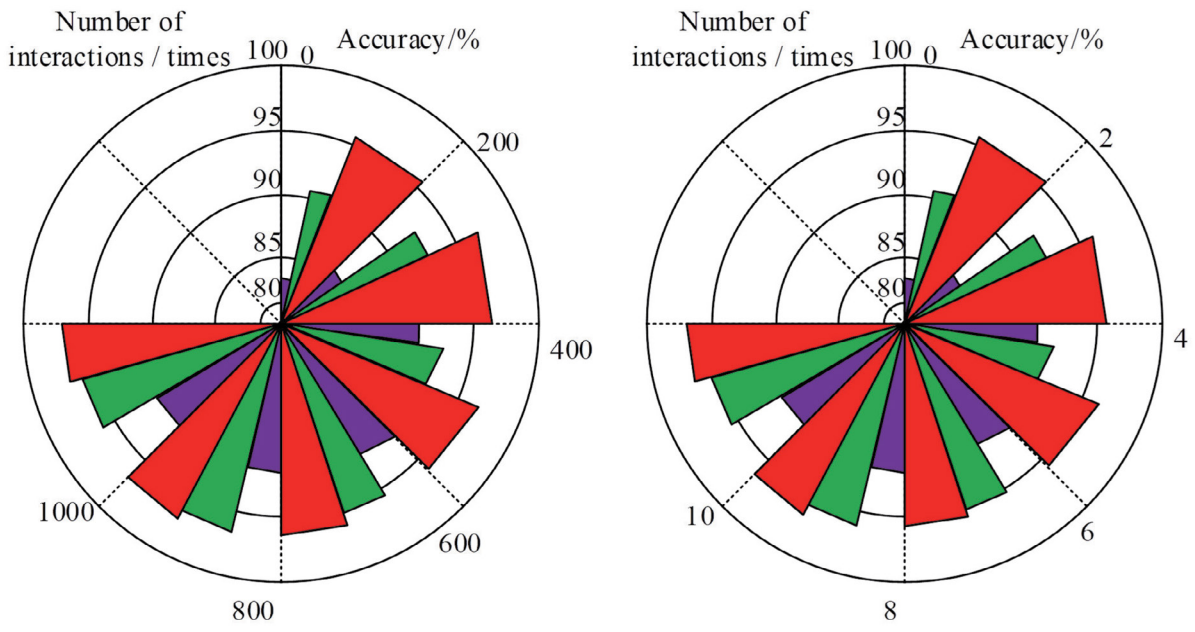
In Table 2, the wake-up success rate was the proportion of responses within 1 second after calling the wake-up word for the first time. Due to low pronunciation accuracy, the wake-up success rates of children aged 2-4 were only 94.2% and 95.1%, while the wake-up success rates of children aged 8-10 could reach 97.5% and 98.0%.

Table 2. Instruction recognition accuracy in real toy scenarios.

Grouping	Age / Years	Sample Size (times)	Wake-up Success Rate (%)	Instruction Recognition Accuracy (%)	False wake-up Rate (times/hour)	Average Response Delay (ms)
Boy	2-4	180	94.2 ± 1.5	87.5 ± 2.3	1.2 ± 0.4	162 ± 14
	5-7	270	96.8 ± 0.9	93.6 ± 1.8	0.8 ± 0.3	138 ± 10
	8-10	150	97.5 ± 0.6	96.2 ± 1.2	0.5 ± 0.2	128 ± 8
Female	2-4	180	95.1 ± 1.2	89.2 ± 2.0	1.0 ± 0.3	154 ± 12
	5-7	270	97.3 ± 0.8	94.8 ± 1.5	0.6 ± 0.2	132 ± 9
	8-10	150	98.0 ± 0.5	97.1 ± 1.0	0.4 ± 0.1	124 ± 6

The instruction recognition accuracy of children aged 5 and above could reach more than 90%, and the response delays were less than 200ms.

The relationship between the speech recognition accuracy rate and the number of interactions and time is demonstrated in Figure 9.



(a) The impact of interaction frequency (b) Interaction time impact

Figure 9. Relationship between speech recognition accuracy and number of interactions and time.

In Figure 9(a), the recognition accuracy of children aged 2-4 increased the fastest in the first 600 interactions. After 1,000 interactions, the recognition accuracy was 94.8%, which was 2.9% and 3.4% lower than that of children aged 5-7 and 8-10, respectively. The language recognition accuracy rate of children aged 8-10 years old changed slightly with the number of interactions. In Figure 9(b), the recognition accuracy of children aged 2-4 and 5-7 increased rapidly in the first 6 days, and the final recognition accuracy increased by 6.5% and 5.3% respectively compared with the initial situation. Table 3 demonstrates the comparison of speech recognition performance under different command complexity and noise levels.

In Table 3, short instructions had the highest accuracy under each noise level (91.8% in noisy environments) and the shortest response delay (124~162ms). The accuracy of long instructions dropped to 86.5% in a noisy environment, but it was still higher than the 83.4% of mixed Chinese and English instructions. Mixed instructions between Chinese and English were affected by language switching and were the most difficult to recognize. The accuracy was 92.3% in a quiet environment and only 83.4% in a noisy environment. The latency under all conditions was less than 200ms, meeting real-time interaction requirements.

Table 3. Recognition performance under different instruction complexity and noise levels.

Instruction type	**Noise level**	**Sample size (times)**	**Instruction recognition accuracy (%)**	**Average response delay (ms)**
Short instruction (≤ 3 characters)	Quiet (< 45 dB)	150	97.2 ± 1.0	124 ± 8
Short instruction	Moderate (45 - 60 dB)	150	95.5 ± 1.5	138 ± 10
Short instruction	Noisy (> 60 dB)	150	91.8 ± 2.2	162 ± 14
Long instruction (≥ 5 characters)	Quiet	120	94.5 ± 1.8	142 ± 12
Long instruction	Moderate	120	91.2 ± 2.3	158 ± 15
Long instruction	Noisy	120	86.5 ± 3.0	182 ± 18
Mixed Chinese-English instruction	Quiet	100	92.3 ± 2.1	148 ± 13
Mixed Chinese-English instruction	Moderate	100	88.6 ± 2.8	166 ± 16
Mixed Chinese-English instruction	Noisy	100	83.4 ± 3.5	192 ± 20

4. Conclusion

To solve the low accuracy of speech recognition of smart toys, the study proposed an edge-embedded speech recognition model based on the improved DFCNN-CTC. The word error rate of the T-DFCNN-CTC model decreased the fastest, with a minimum value of 8.2%, which was 6.1%, 4.3% and 3.9% lower than Conformer, Branchformer and E-Branchformer, respectively. The maximum recall rate of T-DFCNN-CTC could reach 88.5%, which was 8.3% higher than the second-best E-Branchformer (80.2%). T-DFCNN-CTC-2 could effectively reduce the degradation of model performance caused by knowledge distillation. Its word error rate was much lower than T-DFCNN-CTC-1 and slightly higher than the teacher model T-DFCNN-CTC. The teacher model size was compressed from 125MB to 24MB, and the word error rate only increased from 9.5% to 10.2%. Through INT8 quantization and channel pruning, the model size was reduced to 8.1MB, and the inference delay was reduced from 668ms to 155ms, meeting real-time interaction requirements. The average power consumption of the improved model was 45mW, and the energy efficiency was 4.4 times higher than the single-student solution. In the actual test, the overall wake-up success rate of the smart toy was 96.5%, the command recognition accuracy rate was 93.2%, and the false wake-up rate was 0.8 times/hour. The proposed solution is superior to existing offline voice solutions in terms of recognition accuracy, real-time performance, power consumption, cost, and mass production feasibility. There are still some problems in this study. For example, the recognition accuracy of mixed Chinese and English instructions dropped to 83.4% in a noisy environment, and the confusion problem caused by language switching has not been completely resolved. In the future, end-to-end modeling in mixed Chinese and English scenarios can be explored, and language recognition front end or bilingual shared encoders can be introduced to reduce language switching errors.

Declaration of Competing Interests

The authors declare no conflict of interest.

Acknowledgements

This work was supported by the Design and Development of AI Toys Based on Multimodal Embodied Interaction (Grant No. JCKY22.01.0184) and the Design and Development of Embedded Intelligent Children's Playmate (Grant No. 2026JCKY0006).

Funding

This research was funded by [1] Design and Development of AI Toys Based on Multimodal Embodied Interaction (Project No: JCKY22.01.0184) and [2] Design and Development of Embedded Intelligent Children's Playmate (Project No: 2026JCKY0006).

Data Availability

Data used in this article are available upon request from the authors.

References

[1] H. Ahlawat, N. Aggarwal and D. Gupta, "Automatic Speech Recognition: A survey of deep learning techniques and approaches", *Int. J. of Cognitive Computing in Engineering*, vol. 6, no. 4, pp. 201–237, 2025. http://dx.doi.org/10.1016/j.ijcce.2024.12.007

[2] H. Kheddar, M. Hemis and Y. Himeur, "Automatic speech recognition using advanced deep learning approaches: A survey", *Information Fusion*, vol. 109, no. 2, pp. 102422–102431, 2024. http://dx.doi.org/10.1016/j.inffus.2024.102422

[3] A. S. Dhanjal and W. Singh, "A comprehensive survey on automatic speech recognition using neural networks", *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 23367–23412, 2024. http://dx.doi.org/10.1007/s11042-023-16438-y

[4] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter and S. Watanabe, "End-to-end speech recognition: A survey", *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, no. 7, pp. 325–351, 2023. http://dx.doi.org/10.1109/taslp.2023.3328283

[5] Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu, and G. Y. Li, "Deep learning enabled semantic communications with speech recognition and synthesis", *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6227–6240, 2023. http://dx.doi.org/10.1109/twc.2023.3240969

[6] D. l-Fraihat, Y. Sharrab, F. Alzyoud, A. Qahmash, M. Tarawneh and A. Maaita, "Speech recognition utilizing deep learning: A systematic review of the latest developments", *Human-centric Comput. Inf. Sci.*, vol. 14, no. 15, pp. 1–33, 2024. http://dx.doi.org/10.1109/mcna59361.2023.10185880

[7] K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe, "E-branchformer: Branchformer with enhanced merging for speech recognition", in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, 2023, vol. 5, no. 9, pp. 84–91. http://dx.doi.org/10.1109/SLT54892.2023.10022656

[8] T. Liu, M. Zhang, Z. Li, H. Dou, W. Zhang, J. Yang et al., "Machine learning-assisted wearable sensing systems for speech recognition and interaction", *Nature Communications*, vol. 16, no. 1, pp. 2363–2367, 2025. http://dx.doi.org/10.1038/s41467-025-57629-5

[9] Z. Yao, L. Guo, X. Yang, W. Kang, F. Kuang and Y. Yang et al., "Zipformer: A faster and better encoder for automatic speech recognition", arXiv preprint arXiv:2310.11230, vol. 4, no. 6, pp. 157–169, 2023. http://dx.doi.org/10.1109/icme59968.2025.11209883

[10] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels", in *Proceedings of the IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2023, vol. 4, no. 8, pp. 1–5. http://dx.doi.org/10.1109/icassp49357.2023.10096889

[11] Y. Fathullah, C. Wu, E. Lakomkin, J. Jia, Y. Shangguan and K. Li et al., "Prompting large language models with speech recognition abilities", in *Proceedings of the IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2024, vol. 7, no. 5, pp. 13351–13355. http://dx.doi.org/10.48550/arXiv.2307.11795

[12] D. Rekesh, N. R. Koluguri, S. Kriman, S. Majumdar, V. Noroozi and H. Huang et al., "Fast conformer with linearly scalable attention for efficient speech recognition",in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, vol. 6, no. 2, pp. 1–8. http://dx.doi.org/10.48550/arXiv.2305.05084

[13] M. Motiejauskas and G. Dzemyda, "EfficientNet Convolutional Neural Network with GramMatrices Modules for Predicting Sadness Emotion", *Int. J. of Computers Communications & Control*, vol. 19, no. 5, pp. 6697–6708, 2024. http://dx.doi.org/10.15837/ijccc.2024.5.6697

[14] J. Park, T. Yoo, S. Lee and T. Kim, "Urban Noise Analysis and Emergency Detection System using Lightweight End-to-End Convolutional Neural Network", *Int. J. of Computers Communications & Control*, vol. 18, no. 5, pp. 5814–5822, 2023. http://dx.doi.org/10.15837/ijccc.2023.5.5814

[15] J. L. Pang, "Adaptive fault prediction and maintenance in production lines using deep learning", *Int. J. of Simulation Modelling*, vol. 22, no. 5, pp. 734–745, 2023. http://dx.doi.org/10.2507/IJSIMM22-4-CO20

[16] P. Kannapiran and M. M. R. Sindha, "Voice-Based Gender Recognition Model Using FRT and Light GBM", *Tehnicki vjesnik - Technical Gazette*, vol. 30, no. 1, pp. 282–291, 2023. http://dx.doi.org/10.17559/TV-20220302182704

[17] S. S. Nawroly, D. Popescu and M. C. Thekekara Antony, "Category-based and Target-based Data Augmentation for Dysarthric Speech Recognition Using Transfer Learning", *Studies in Informatics and Control*, vol. 33, no. 4, pp. 83–93, 2024. http://dx.doi.org/10.24846/v33i4y202408

[18] Z. Chen, H. Huang, A. Andrusenko, O. Hrinchuk, K. C. Puvvada and J. Li et al., "Salm: Speech-augmented language model with in-context learning for speech recognition and translation", in the *Proceedings of the IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2024, vol. 2, no. 4, pp. 13521–13525. http://dx.doi.org/10.1109/icassp48485.2024.10447553

[19] C. Hema and F. P. G. Marquez, "Emotional speech recognition using CNN and deep learning techniques", *Applied Acoustics*, vol. 211, no. 3, pp. 109492–109503, 2023. http://dx.doi.org/10.1016/j.apacoust.2023.109492

[20] N. K. Dennis, "Using AI-Powered Speech Recognition Technology to Improve English Pronunciation and Speaking Skills", *IAFOR Journal of Education*, vol. 12, no. 2, pp. 107–126, 2024. http://dx.doi.org/10.22492/ije.12.2.05

[21] H. Kheddar, Y. Himeur, S. Al-Maadeed, A. Amira and F. Bensaali, "Deep transfer learning for automatic speech recognition: Towards better generalization", *Knowledge-Based Systems*, vol. 277, no. 6, pp. 110851–110862, 2023. http://dx.doi.org/10.1016/j.knosys.2023.110851

[22] Q. Yang, W. Jin, Q. Zhang, Y. Wei, Z. Guo, and X. Li et al., "Mixed-modality speech recognition and interaction using a wearable artificial throat", *Nature Machine Intelligence*, vol. 5, no. 2, pp. 169–180, 2023. http://dx.doi.org/10.1038/s42256-023-00616-6

[23] S. Ambrogio, P. Narayanan, A. Okazaki, A. Fasoli, C. Mackin and K. Hosokawa et al., "An analog-AI chip for energy-efficient speech recognition and transcription", *Nature*, vol. 6, no. 7, pp. 768–775, 2023. http://dx.doi.org/10.1038/s41586-023-06337-5

[24] C. Chen, Y. Hu, C. H. H. Yang, S. M. Siniscalchi, P. Y. Chen and E. S. Chng, "Hyporadise: An open baseline for generative speech recognition with large language models", *Advances in Neural Information Processing Systems*, vol. 36, no. 9, pp. 31665–31688, 2023. http://dx.doi.org/10.52202/075280-1375

[25] J. Jeon, S. Lee, and S. Choi, "A systematic review of research on speech-recognition chatbots for language learning: Implications for future directions in the era of large language models", *Interactive Learning Environments*, vol. 32, no. 8, pp. 4613–4631, 2024. http://dx.doi.org/10.1080/10494820.2023.2204343

[26] Y. O. Sharrab, H. Attar, M. A. H. Eljinini, Y. Al-Omary and W. A. E. Al-Momani, "Advancements in speech recognition: A systematic review of deep learning transformer models, trends, innovations, and future directions", *IEEE Access*, vol. 13, no. 4, pp. 46925–46940, 2025. http://dx.doi.org/10.1109/access.2025.3550855

Received: May 2026
Revised: May 2026
Accepted: May 2026

Contact addresses:

Fang Nie
Chengdu Jincheng University
Chengdu, Sichuan Province
China
e-mail: 15982283985@163.com

Lijuan Zheng*
School of Artificial Intelligence and Electronic Engineering
Sichuan Technology and Business University
Chengdu, Sichuan Province
China
e-mail: zhenglijuan_117@163.com
*Corresponding author

Fang Nie received the B.S. degree in Communication Engineering from Sichuan Normal University, Sichuan, China, in 2005, and the M.S. degree in Computer Application Technology from the University of Electronic Science and Technology of China, Sichuan, China, in 2016. She is currently a Lecturer with the School of Electronic Information, Chengdu Jincheng University, Sichuan, China. Her research interests include artificial intelligence and embedded systems.

Lijuan Zheng received the M.S. degree in Computer Technology from Xihua University, Chengdu, China, in 2012. She is currently a Professor and a dual-qualified teacher with Sichuan Technology and Business University, China. Her research interests include computer applications and software development.