Attention-Based Bimodal Neural Network Speech Recognition System on FPGA

Aiwu Chen
College of Intelligent Manufacturing (CIM), Hunan University of Science and Engineering (HUSE), Yongzhou 425199, China
E-mail: caiwu9050@126.com

Keywords: speech recognition, field programmable gate array, neural network, attention mechanism

Received: September 13, 2024

To further improve the accuracy of speech recognition technology, a neural network speech recognition system based on a field programmable gate array is designed. Firstly, a neural network audiovisual bimodal speech recognition algorithm based on an attention mechanism is designed. Then, a speech recognition platform based on on-site programmable gate arrays is built. The results showed that the word error rate and the character error rate of this research algorithm were 3.17% and 1.56%, respectively, which were significantly lower than the traditional Lip-Reading Network algorithm's 26.24% and 12.56%. The algorithm converged quickly when the training rounds were less than 10 and tended to stabilize when it was 20. The proposed speech recognition platform used many DSP units in its design, with a utilization rate of 83.2%, the lowest power consumption of 2.21W, the highest energy efficiency ratio of 26.15, and the shortest processing time and faster running speed. In summary, the research algorithm can reasonably allocate learning weights, improve training speed, and has certain feasibility and effectiveness because of introducing attention mechanism. It has good application effects in speech recognition, which helps to improve the accuracy of language recognition algorithms and promote communication between humans and machines.

Povzetek: Članek predstavlja nov hibridni algoritem za optimizacijo parametrov permanentnega magneta sinhronega motorja (PMSM), ki združuje samooptimizacijsko simulirano kaljenje (SA) in optimizacijo rojev delcev (PSO).

1 Introduction

The progress of Artificial Intelligence (AI) technology has also been further developed and gradually applied to various fields. Speech recognition is an important field in AI. With the application and development of deep learning in Speech Recognition Technology (SRT), the error rate of speech recognition systems has been significantly reduced [1]. Lip Reading Network (LipNet) was jointly proposed by the Artificial Intelligence Laboratory at the University of Oxford, the Google DeepMind team, and the Canadian Institute of Advanced Studies. It is a lip-reading program that combines deep learning technology. It utilizes machine learning to achieve sentence-level automatic lip-reading technology, which has high human lip-reading accuracy [2-3]. Field Programmable Gate Array (FPGA) is a new type of high-density programmable logic device, which not only has high speed and reliability but also has a large scale of user-defined logic functions. It is suitable for logic circuit applications such as timing and combination and can replace dozens or even hundreds of general medium-scale WI chips [4-5]. AI technology is closely linked to speech recognition and promotes each other. In this context, this study designs a Neural Network Audiovisual Bimodal Speech Recognition (NNA-BSR) algorithm based on Attention Mechanism (AM) and an FPGA-based speech recognition platform. This study aims to improve the accuracy of SRT and lay the foundation for more efficient and convenient human-computer interaction. There are two main innovations in this study. The first point is to construct a bimodal Speech Recognition Algorithm (SRA) that combines sound and visual features. The second point is to facilitate the introduction of AMs, thereby enabling the network to exercise the capacity to select information. The main structure of the study consists of four parts. Part 1 analyzes the current research status. Part 2 is to design the AM-based NNA-BSR algorithm and build a speech recognition platform based on FPGA. The third part is to analyze the application effect of the proposed model. The final part is a summary of the entire study.

2 Related works

FPGA is a new type of programmable logic device with excellent performance, which can significantly reduce the development cost of digital systems. Kumar *et al.* stated that protecting voice communication is a challenging task. A method for encrypting speech signals in peer-to-peer communication using an improved lightweight advanced encryption standard algorithm was proposed to address the issue of asynchronous communication between chaotic signals and oscillators, which limited their ability to protect speech communication. The tests were carried out on two different FPGAs [6]. Rao *et al.* stated that in terms of data security and anonymity, some publicly facing programs were insecure and had environmental issues. The method of online public opinion was based on FPGA and machine education. In this popular emotional network, the public's perception of disaster networks was closely related to society. This brought unprecedented pressure on the government's ability to respond to crises and their consequences [7]. Koyuncu *et al.* used a multi-layer feedforward Artificial Neural Network (ANN) to model the economic analysis of hydrogen production and liquefaction systems and implemented it on FPGA. Firstly, a dataset was created using the engineering equation solver program for ANN-based system modeling, and then the model was trained and tested using the Matlab program [8]. Chu *et al.* focused on the hardware structure of a Raman scattering distributed fiber optic transducer platform based on FPGA implementation and analyzed the principle of Raman scattering. The output analog electrical signal was converted into a digital signal at a 16-bit sampling rate through an analog-to-digital converter. The test results on the fiber optic transducer platform indicated that the program successfully returned all operations and had a certain degree of effectiveness [9]. RodríGuez-Borbón *et al.* first applied FPGA as a new, customizable hardware architecture for fast and efficient quantum dynamics simulations of large-scale chemical and material systems. The real-time electronic dynamics calculation performance of this method was good, and FPGA could play a promising role in the upcoming applications of quantum chemistry and materials science [10]. Gholami *et al.* believed that the hardware implementation of spike neural networks (PNNs), known as neural morphological structures, provided a clear understanding of brain performance. A PNN using Izhikevich neurons and a gradient descent learning algorithm was proposed to approximate S-shaped and other nonlinear functions. The results of hardware synthesis and the physical implementation of PNN on FPGA were also reported [11].

Table 1: Summary table of related work

Related Work	Performance Index	Limitation
Koenecke *et al.*	The average word error rate is 0.35	The ASR system shows significant racial differences
Haeb-Umbach *et al.*	Can effectively solve the problem of acoustic distortion	Low accuracy of speech recognition
Song *et al.*	The accuracy of English speech recognition is over 80%	Not combined with video information
Yang *et al.*	Unweighted average recall rate reaches 63.98%	Without combining video information, the efficiency is low
Lin *et al.*	The tag error rate is 3.95%	Not combined with video information
Muraru *et al.*	The highest accuracy is over 95%	Not exploring the applicability in different speech recognition environments
Ma *et al.*	The average recognition accuracy is as high as 87%	Not combined with video information
Gao *et al.*	The average online speech recognition latency is only 1 μs	Not combined with video information

SRT is a high-tech technology that enables machines to transform speech signals into corresponding text or commands through recognition and comprehension processes. It involves multiple fields such as pattern recognition, AI, and signal processing. The Automatic Speech Recognition (ASR) system, which employs sophisticated machine learning algorithms to transform spoken language into text, is a driving force behind the capabilities of popular virtual assistants. However, it is important to recognize that this technology may not be equally applicable to all populations. Koenecke *et al.* investigated the transcriptional ability of five state-of-the-art automated speech systems in structured interviews with 42 white and 73 black individuals to address this issue [12]. Haeb Umbach *et al.* stated that far-field ASR leads to a significant improvement in recognition accuracy, resulting in a completely different processing pipeline compared to ASR for close-range speech. Therefore, a signal enhancement front-end with de-reverberation, source separation, and acoustic beamforming was adopted to clean up speech, and the back-end ASR engine was enhanced through multi-condition training and adaptation [13]. Song believed that degree learning technology achieved unparalleled performance in many tasks due to its hierarchical feature learning ability and data modeling ability compared to shallow learning technology. Therefore, taking English pronunciation as the object, a deep learning SRA combining features and attributes of speech was proposed. The proposed algorithm could significantly improve the performance of English speech recognition systems [14]. Yang *et al.* put forth a compact speech recognition network with spatiotemporal characteristics for edge computing, aiming to address the gap in research concerning the design of an efficient speech recognition network for the advancement of edge computing. A simplified AM was also proposed. This indicated that the accuracy of the proposed mechanism has been improved in both speech emotion recognition and keyword recognition [15]. Lin *et al.* designed a new processing paradigm that integrates multilingual speech recognition into a framework using three cascading modules. Moreover, to address the diversity between radio transmission noise and speakers, a multi-scale convolutional neural network architecture was proposed to adapt to different data distributions and improve performance. The proposed architecture had certain effectiveness and scalability on open corpora [16]. Muraru S *et al.* proposed a speech recognition method based on a k-nearest neighbor algorithm and analyzed its application effect using the AudioMNIST dataset. The results showed that when the k parameter was 5, the highest accuracy of the proposed method was over 95% [17]. Ma *et al.* proposed an English speech recognition model that combines multiple information sources to address the issues of speech understanding and text generation in speech recognition. The results indicated that the average recognition accuracy of the proposed model was as high as 87% [18]. Gao *et al.* proposed a speech recognition method based on long short-term memory and spatiotemporal sparsity. The results indicated that the proposed method could support real-time online speech recognition when implemented on small and large FPGAs, with an average delay of only 1 µs [19]. The summary table of the above-related work is shown in Table 1.

In summary, many previous scholars have conducted extensive research on speech recognition. However, most SRAs have complex network structures and high hardware requirements, which hinder the development of SRAs. Therefore, the research on NNA-BSR system based on FPGA has important practical application value and prospects.

3 Design of a neural network speech recognition system based on FPGA

In recent years, multi-modal information processing has received increasing attention. To further improve the accuracy of SRT by combining visual and auditory features, this study will design an AM-based NNA-BSR algorithm and build an FPGA-based speech recognition platform to test speech recognition.

3.1 Design of NNA-BSR algorithm based on attention mechanism

Due to the complexity of audiovisual bimodal speech recognition tasks, traditional machine learning methods find it difficult to effectively fuse different dimensional features between modalities. Therefore, it is necessary to use deep learning with stronger feature extraction capabilities to build audio-visual bimodal SRAs [20]. LipNet is composed of Spatiotemporal Convolutional Neural Networks (STCNN), Recurrent Neural Network (RNN), and Connectivist Temporal Classification (CTC) loss. It is the first end-to-end sentence level lip reading model that considers both spatiotemporal visual features and sequence information. The specific framework of LipNet is shown in Figure. 1.

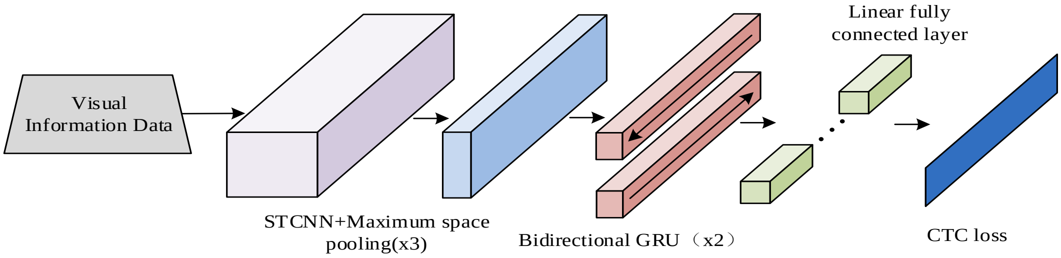


Figure 1: Structure diagram of LipNet

The input data of LipNet are an image sequence, introducing a time dimension. Traditional two-dimensional convolutions are difficult to apply to their feature extraction tasks. Therefore, the STCNN is selected to process image sequence data through convolution in both temporal and spatial dimensions, as shown in formula (1).

$$[stconv(x,w)]_{d'tij}=\sum_{d=1}^{D}\sum_{t'=1}^{k_t}\sum_{i'=1}^{k_w}\sum_{j'=1}^{k_h} w_{d'dt'i'j'}x_{d,t+t',i+i',j+j'} \tag{1}$$

In formula (1), d is the quantity of layers of STCVV. k_t represents the length of the time step. k_w and k_h are the width and height of the convolutional kernel. w represents the weight parameter. x represents feature mapping. LipNet uses ReLU as the activation function, as shown in formula (2).

$$f(x)=\max(0,x) \tag{2}$$

Nevertheless, ReLU has the potential to render certain neurons incapable of activation. Consequently, LipNet opts for the Adam optimizer, which exhibits superior adaptability. In addition, LipNet connects a maximum space pooling layer after each layer of STCNN to perform feature dimensionality reduction processing, which not only extracts the core features of visual information as much as possible but also filters redundant information. The space pooling operation is shown in formula (3).

$$MaxPool(x_v^{(t)})=\max(x_v^{(t)}) \tag{3}$$

In formula (3), $x_v^{(t)}$ represents the feature mapping output of STCNN layer B. Traditional RNNs cannot consider the subsequent state information, so bidirectional RNNs have emerged. In a bidirectional RNN, the input data at each moment are transmitted to both RNNs simultaneously. Subsequently, the two RNNs transmit the extracted features in opposite directions, allowing the entire network to combine contextual features [21]. The final predicted value is determined by the output results of both RNNs. The specific structure of bidirectional RNN is shown in Figure. 2.

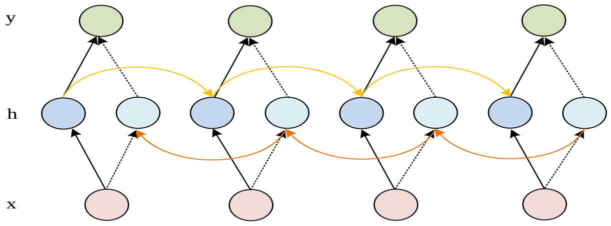


Figure 2: Structure diagram of bidirectional RNN

After extracting the spatiotemporal features of visual information sequences through the STCNN layer, LipNet uses two Bidirectional GRU (B-GRU) layers to further extract the features. This enables deep features to have contextual information and can extract subtle changes between adjacent image frames to improve accuracy [22]. The extraction process of B-GRU is shown in formula (4).

$$\begin{cases} H_t^{(1)} = f(U^{(1)}H_{t-1}^{(1)} + W^{(1)}X_t + b^{(1)}) \\ H_t^{(2)} = f(U^{(2)}H_{t+1}^{(2)} + W^{(2)}X_t + b^{(2)}) \\ H_t = H_t^{(1)} \oplus H_t^{(2)} \end{cases} \quad (4)$$

In formula (4), $U^{(1)}$, $U^{(2)}$, $W^{(1)}$, $W^{(2)}$, $b^{(1)}$, and $b^{(2)}$ are the parameters of the two GRU layer. $\oplus$ represents the concatenation or addition operation of feature states. H_t represents the output of B-GRU. The SoftMax activation function is mainly found in multi-classification task $y \in \{1, 2, ..., N\}$. After being processed by the SoftMax activation function, the conditional probability that the predicted input belonging to category n is shown in formula (5).

$$\hat{y} = p(y = n|x) = soft\max(x_n^{(f)}) = \frac{\exp(x_n^{(f)})}{\sum_{n'}^{N} \exp(x_{n'}^{(f)})} \quad (5)$$

In formula (5), $\hat{y}$ represents the prediction probability of the category label. $x^{(f)}$ is the input. $x_n^{(f)}$ is the n-th dimensional feature value in the input. The value processed by SoftMax is the predicted result of the entire model. The predicted results are compared with the labels, and the CTC loss function and backpropagation algorithm are used to update the training parameters. After a large amount of repeated training, LipNet with loss convergence is finally obtained, and the process is shown in formula (6).

$$\begin{aligned} \underset{\theta\_lipnet}{\arg\max}\, L(D) &= -\ln\left(\Pi_{(x,l)\in D}\, p(l|x)\right) \\ &= -\sum\nolimits_{(x,l)\in D} \ln\left(\sum_{\pi \in B^{-1}(t)} p(\pi|x)\right) \end{aligned} \quad (6)$$

In formula (6), θ_lipnet represents the learning parameter of the LipNet model. D represents training data. x represents the input image data. l represents the real label. To enable the network to extract both visual and auditory information, this study builds an AV LipNet based on LipNet. AV LipNet is divided into Feature Extraction Networks (FEN) and feature fusion networks. The FEN is further divided into auditory flow and visual flow. The visual flow structure consists of three STCNN layers, three maximum space pooling layers, and one B-GRU layer. The auditory flow structure consists of three CNN layers, three maximum pooling layers, and one B-GRU layer. CNN uses the Mel scale Frequency Cepstral Coefficients (MFCC) features obtained from auditory information data through Short Time Fourier Transform (STFT) as input to extract detailed features of auditory information, as shown in formula (7).

$$x_{c',t,i}^{(l)} = \varphi\left(\sum_{c=1}^{C}\sum_{t'=1}^{T}\sum_{i'=1}^{F} w_{c'ct'i'}^{(l)} x_{c,t+t',i+i'}^{(l-1)} + b_c^{(l)}\right) \quad (7)$$

In formula (7), $x_{c',t,i}^{(l)}$ represents the output result of the l-th layer convolution. $\varphi(\cdot)$ represents the activation function. C represents the number of channels. T represents the time step. F represents the spatial size of the output result of the $l-1$-th layer convolution. $w_{c'ct'i'}^{(l)}$ is the weight parameter of the l-th layer convolution. $b_c^{(l)}$ represents the offset of the l-th layer convolution. After extracting detailed features, each CNN layer will be sent to the maximum pooling layer for dimensionality reduction processing, further extracting deep features, as shown in formula (8).

$$MaxPool(x_a^{(t)}) = \max(x_a^{(t)}) \quad (8)$$

In formula (8), $x_a^{(t)}$ represents the auditory detail features extracted through the l-layer CNN layer. After completing the deep feature extraction of visual and auditory features, the two features are concatenated using a feature fusion network, as shown in formula (9).

$$X = X_v \oplus X_a \quad (9)$$

In formula (9), X represents the fusion feature matrix. X_v represents the visual feature matrix. $\oplus$ represents the linking operation of the feature matrix. X_a represents the auditory feature matrix. After obtaining the fusion feature matrix, two B-GRU layers are used for feature extraction of the fusion features, mining temporal information, and sending it into a linear fully connected network for mapping. Finally, the conditional probability of each output is calculated using the SoftMax activation function, and the optimal result is selected as the output. Simultaneously comparing the labels and output results, the loss is calculated using the CTC loss function, and the network learning parameters are updated based on backpropagation. Through many repeated trainings, the loss-converged AV LipNet algorithm is finally obtained, as shown in formula (10).

$$\underset{\theta\_AV-lipnet}{\arg\max}\ L(D) = -\ln\left(\Pi_{(x,l)\in D}\, p(l|X)\right) = -\sum_{(x,l)\in D} \ln\left(\sum_{\pi\in B^{-1}(t)} p(\pi|X)\right) \tag{10}$$

In formula (10), $\theta\_AV-lipnet$ represents the AV LipNet parameter. X represents the fusion feature matrix. To enable the network to focus on learning effective features, this study introduces an AM based on AV LipNet. The specific structure of the Convolutional Block Attention Module (CBAM) is shown in Figure. 3.

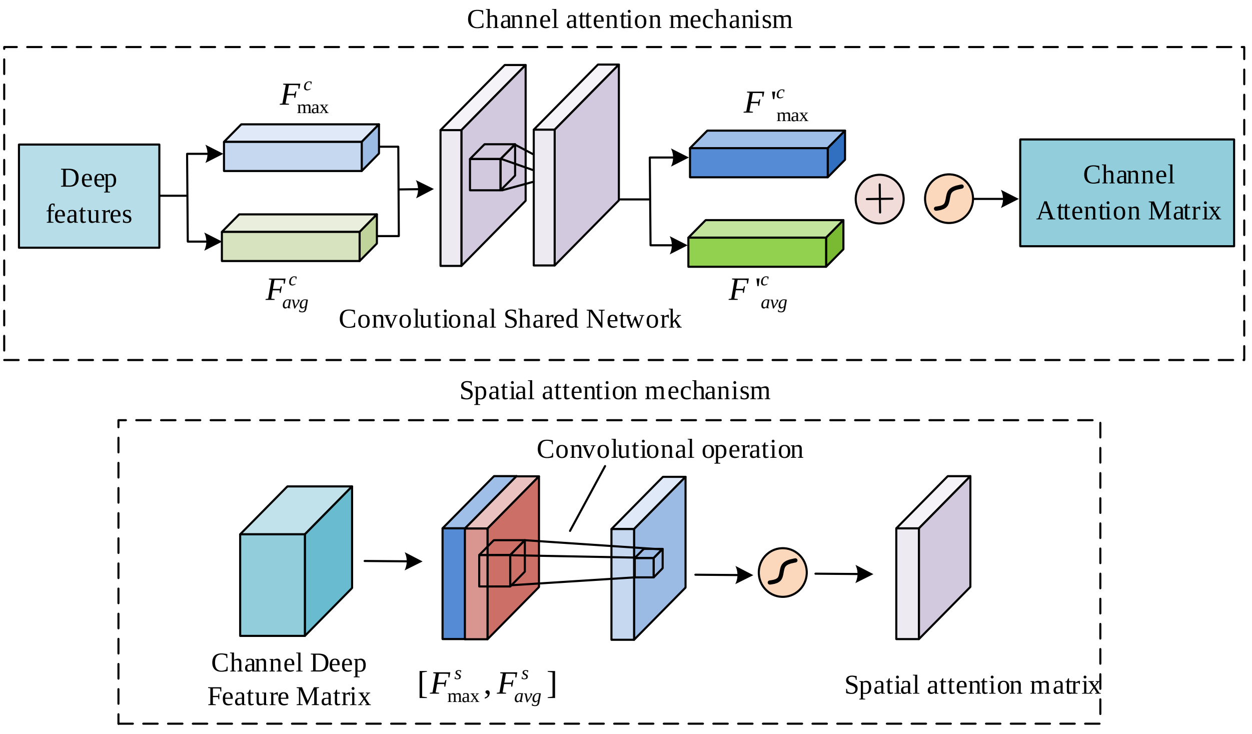

Channel attention mechanism
Deep features
$F^c_{\max}$
F^c_{avg}
Convolutional Shared Network
$F'^c_{\max}$
F'^c_{avg}
Channel Attention Matrix
Spatial attention mechanism
Convolutional operation
Channel Deep Feature Matrix
$[F^s_{\max}, F^s_{avg}]$
Spatial attention matrix

Figure 3: Structure diagram of CBAM

The process of CBAM is shown in formula (11).

$$M_c(F) = \sigma\left(MLP(AvgPool(F)) + MLP(MaxPool(F))\right) = \sigma\left(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{\max}))\right) \tag{11}$$

In formula (11), $M_c(F)$ represents the channel attention feature weight. F represents the input feature mapping. MLP represents a convolutional shared network. F^c_{avg} and $F^c_{\max}$ represent deep feature maps focused on different background features. $\sigma(\cdot)$ represents the activation function.

3.2 Building a speech recognition platform based on FPGA

FPGA is a semi-custom device developed based on various programmable logic devices. The internal structure includes registers, logical unit arrays, lookup tables, and hard-core resources. Among them, the logical unit array consists of configurable logical blocks and input output blocks. Hard core resources can improve hardware capabilities. The traditional FPGA development and design method uses Hardware Description Language (HDL) and corresponding design software to code the functions to be implemented, as shown in Figure. 4.

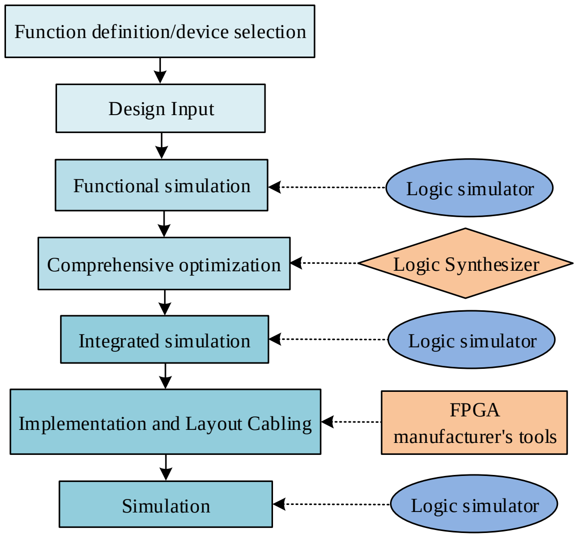


Figure 4: Traditional FPGA development process

The input for FPGA development can be schematic, IP core, or HDL. The mainstream language used for HDL input is Verilog HDL or VHDL. Functional behavior simulation represents the verification of the logical functionality of the designed circuit before compilation. The comprehensive stage transforms higher-level abstract levels into lower-level descriptions and optimizes logical connections in the design based on corresponding indicators, thereby simplifying the hierarchical structure in the design. The final layout and wiring will be implemented by FPGA-related software, and after synthesis, simulation can help designers verify the correctness and performance of the design [23]. The use of FPGA for neural network construction can also be developed using High Level Synthesis (HLS). HLS has a relatively brief development cycle and can be ported between platforms, as well as reused between modules, which effectively reduces the time required for program debugging. HLS also allows developers to directly encapsulate hardware kernel IPs using functions such as C or C++. Function modules can be regarded as equivalent RTL descriptions.

The execution efficiency of convolutional layers has a significant impact on the overall computational efficiency of CNN, so designing a convolutional module is the core content of the entire hardware acceleration design. Convolutional layers require many weight parameters and input/output feature map data. Due to the limited amount of RAM stored on FPGA chips, it is difficult to achieve large-scale data exchange solely relying on RAM. Therefore, input feature maps and weight parameters can be pre-stored in dynamic memory (Double Data Rate, DDR), and then data exchange between embedded block RAM and DDR can be carried out to effectively improve data throughput and transmit more data. The Advanced extensible Interface (AXI) bus has the characteristics of high bandwidth and low latency. As a bus standard that can achieve high data throughput, it can simultaneously perform data read and write tasks, separating execution operations from data read and write requests. The AXI4 interfaces supported in Vitis HLS include memory interface AXI4, register interface AXI4 lite, and serial interface AXI-stream. Since both AXI4 and AXI4 lite are memory-mapped, although they are relatively easy to implement, they can result in the inefficient use of resources and may lead to issues with the occupation of read and write response channels for data and addresses. Therefore, this study chooses AXI-stream and uses the AXI-HP interface of the ZYNQ platform to access external DDR data. The memory access process relies on the Direct Memory Access (DMA) engine, as shown in Figure. 5 for its basic structure.

This study employs a block caching mechanism for data caching, whereby the input feature map and convolutional weight parameters are partitioned into distinct blocks of varying degrees, and the output feature map is calculated in a cyclic sequence. The selected hardware verification platform is the MZ7030FA development board. The development board consists of a bottom board and a core board. The core board covers all hardware resources of FPGA. The base plate has added peripheral interfaces, which helps promote project development. In this study, when using FPGA for speech recognition testing, the preprocessing, feature extraction, and decoding processes of speech signals are completed in software. On the one hand, the feature extraction process of speech signals usually requires high-precision data operations. On the other hand, FPGA hardware computing and storage resources are limited, and the decoding process requires a large search space. To conduct FPGA language recognition testing, it is necessary to input real-time speech into the CPU to initiate the preprocessing and feature extraction process of the speech. To ensure the normal operation of speech recognition tasks, it is necessary to write corresponding code for driver control design, as shown in Figure. 6.

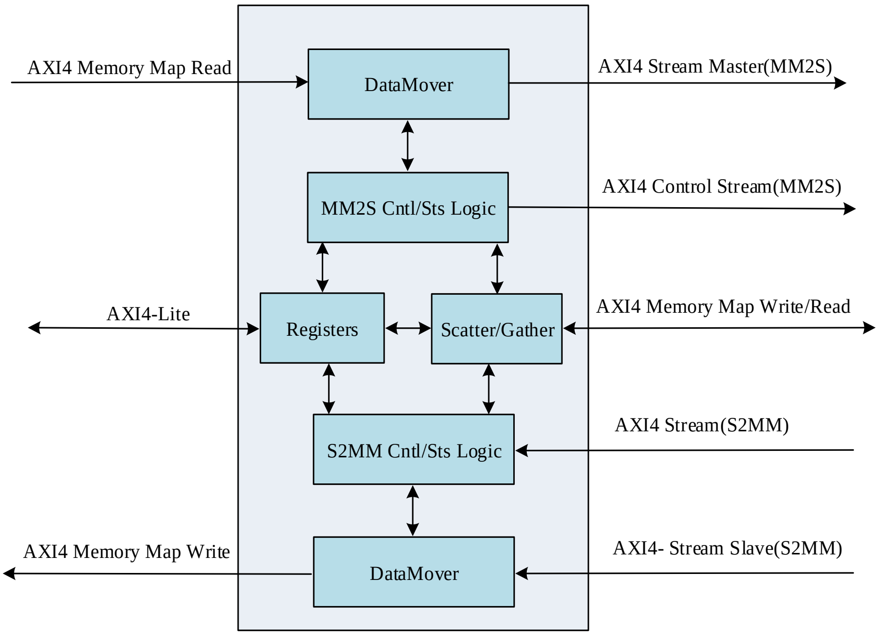


Figure 5: The Basic Structure of DMA

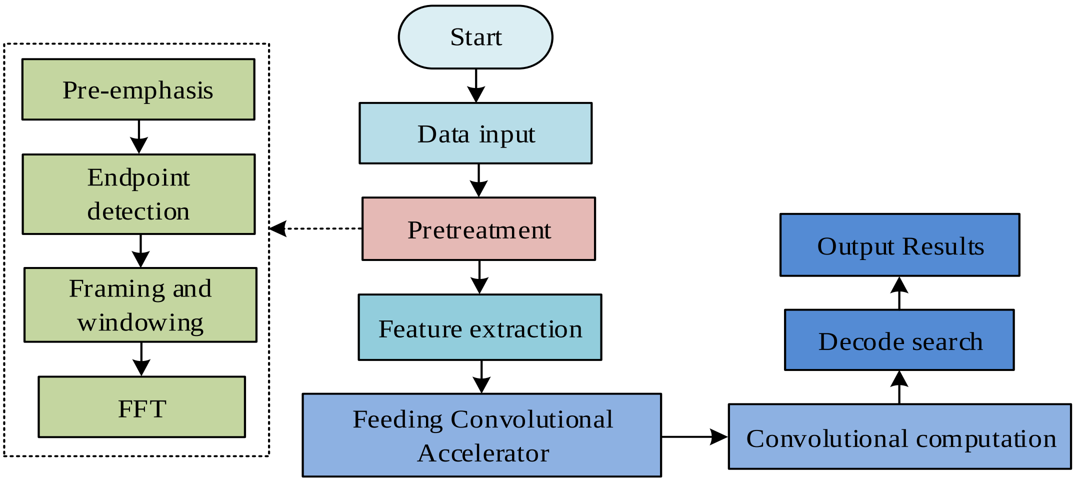


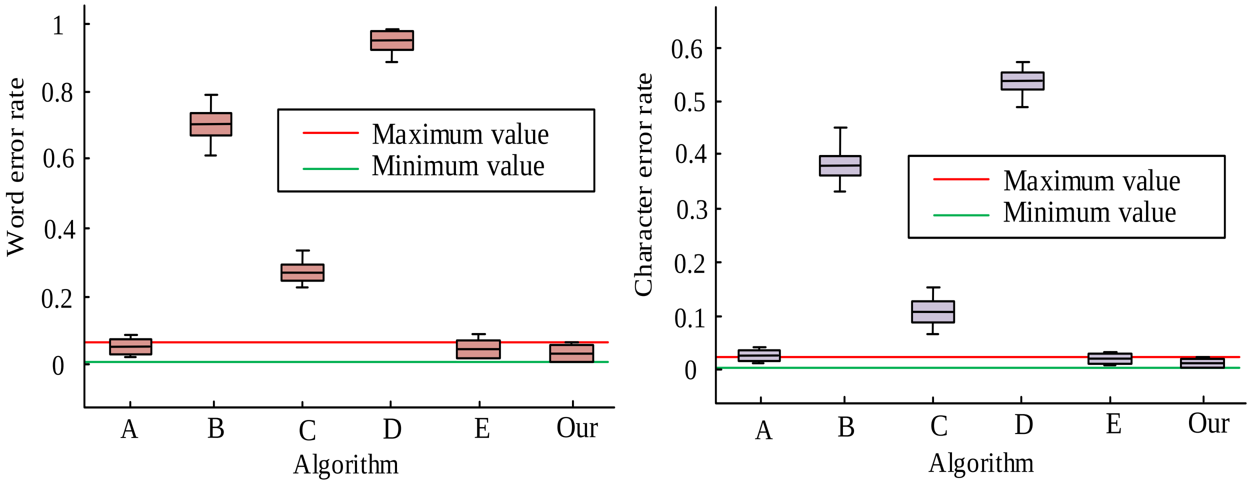
Figure 6: FPGA language recognition testing flowchart

4 Performance analysis of neural network speech recognition system based on FPGA

This study designs the AM-based NNA-BSR algorithm and a speech recognition platform based on FPGA, which helps to promote the development of SRT. However, its effectiveness needs further verification. This study mainly analyzes from two aspects. Firstly, the feasibility and effectiveness of the NNA-BSR based on AM are analyzed, and then the performance of the FPGA-based speech recognition platform is analyzed.

4.1 Effect analysis of NNA-BSR algorithm based on attention mechanism

The GRID dataset is a sentence level audio-visual bimodal speech recognition dataset that includes various parts of speech in English sentences. The video information has a resolution of 360×288 and a frame rate of 25fps. When used for model training, it can enable the model to complete more complex speech recognition tasks. To verify the superiority of the AM-based NNA-BSR algorithm, a GRID dataset is used and compared with the Word Error Rate (WER) and Character Error Rate (CER) of five algorithms (denoted as A, B, C, D, E): A-LipNet, Asymmetric BLSTM, LipNet, TM seq2seq, and AV LipNet. The red line in the figure represents the maximum value of the research method's indicators, while the green line represents the minimum value of the indicators. As shown in Figure. 7, among the six algorithms, the WER and CER of this research algorithm are the lowest, with 1.59% and 0.68%, respectively. The AV LipNet algorithm exhibits a significantly lower error rate than the traditional LipNet algorithm, with respective values of 4.01% and 1.69%. The TM seq2seq algorithm performs the worst, with a WER of over 95%. The results show that the AM-based NNA-BSR algorithm has high accuracy in language recognition, good speech recognition performance, and certain feasibility and superiority.



a)Word error rate

b)Character error rate

Figure 7: Comparison of WER and CER of Six Algorithms

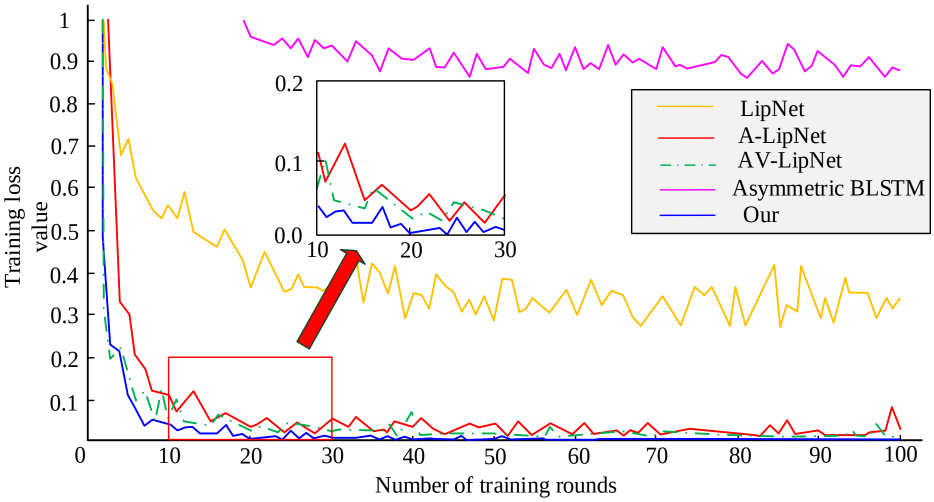
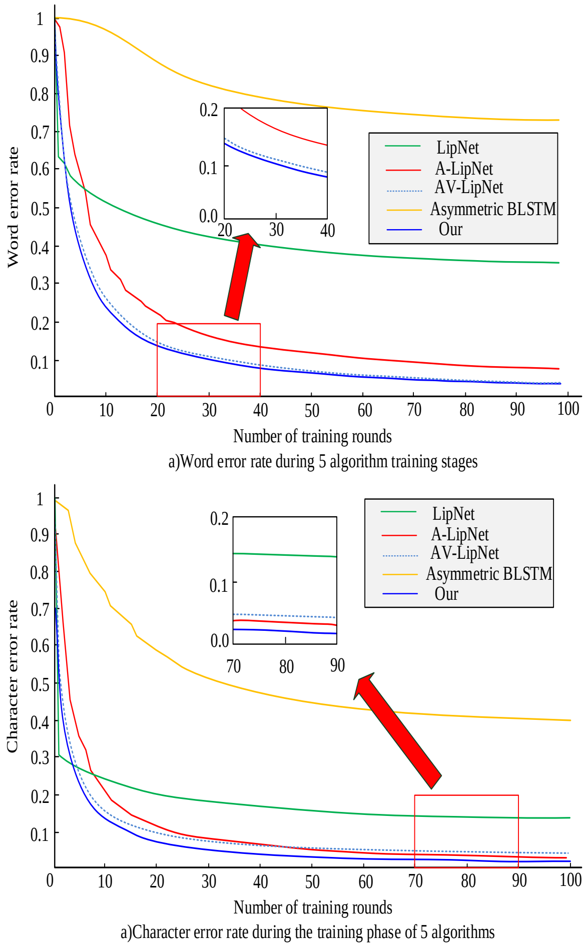


Figure 8: Training loss curves for five algorithms



a)Word error rate during 5 algorithm training stages

a)Character error rate during the training phase of 5 algorithms

Figure 9: CER curve of five algorithm training stages

To verify the performance of the AM-based NNA-BSR algorithm, this study compares its convergence speed with four algorithms: A-LipNet, Asymmetric BLSTM, LipNet, and AV LipNet. Figure. 8 is the results. Among the five algorithms, the research algorithm has the fastest convergence speed and the final loss value is lower. It converges quickly when the quantity of training rounds is less than 10 and tends to stabilize when the amount of training rounds is 20. Next is the AV LipNet algorithm. The performance of Asymmetric BLSTM is the worst, with unsatisfactory changes in losses. The results show that the AM-based NNA-BSR algorithm has a fast convergence speed and good performance.

The changes in WER and CER of the above 5 algorithms with increasing training times are displayed in Figure. 9. As the training increases, the error rates of all five algorithms decrease. However, the algorithm in this study performs the best, with the WER and CER also decreasing the fastest. The CER has stabilized in about 30 rounds and remains at a very low error rate. The results indicate that the AM-based NNA-BSR algorithm can allocate learning weights reasonably and improve training speed due to the introduction of AM, which has certain feasibility and effectiveness.

The speech recognition accuracy results of the above five algorithms under different signal-to-noise ratios are shown in Figure. 10. As the signal-to-noise ratio increases, the speech recognition accuracy of all five algorithms gradually increases. Among them, the speech recognition accuracy of the proposed AM-based neural network audio-visual dual-mode SRA has always been the highest. When the signal-to-noise ratio is 5dB, the speech recognition accuracy is 97.38%.

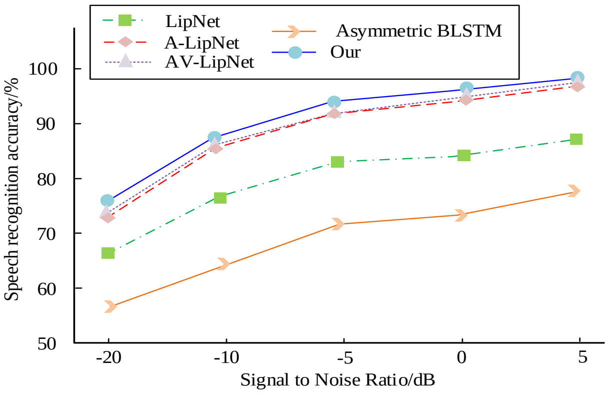


Figure 10: Comparison of speech recognition accuracy under different signal-to-noise ratios

4.2 Performance analysis of FPGA-based speech recognition platform

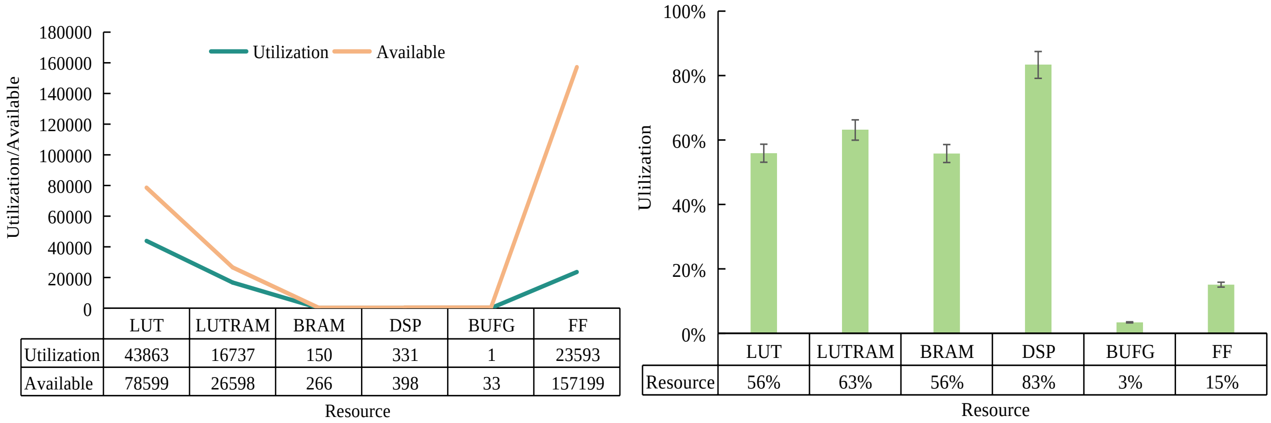
Using an FPGA based speech recognition platform for thinking and speech testing, speech signals are collected through an external microphone in a quiet environment. During this process, it is necessary to maintain the voice signal as mono, with a sampling frequency of 16kHz, a size of 16bit, and a duration of no more than 16s. After voice endpoint detection, effective voice data are fed into the feature extraction step. The input format of speech signals has a direct impact on the processing efficiency of FPGA platforms, and different formats require different decoding and processing algorithms. The sampling rate has a significant impact on the processing performance and storage requirements of FPGA platforms. A higher sampling rate increases the computational burden and storage requirements of FPGA, while a lower sampling rate may lead to a decrease in speech signal quality, thereby affecting recognition accuracy. Larger and longer speech signals require more computing resources and storage space, which may lead to a decrease in processing speed. The study adopts the Valid Activity Detection (VAD) algorithm for continuous recognition of speech and divides it into two categories: effective and ineffective. Subsequently, the short-term energy and zero crossing rate of the speech segment are analyzed. When the short-term energy is high and the zero-crossing rate is low, the region is classified as a speech segment, and conversely, as a speech blank area. The MZ7030FA development board is used as the hardware verification platform for the study, and the interval between the output channels of the convolutional layer is set to 16. The hardware resource usage of FPGA is shown in Figure. 11. The FPGA-based speech recognition platform uses many DSP units in its design, with a utilization rate of 83.2%. This is due to the large number of parameters calculated in parallel by the convolutional module.

To verify the performance of an FPGA-based speech recognition platform, this study compares it with the FPGA-based speech recognition platform proposed by Li *et al.* and Yu *et al.* [24-25]. Table 2 shows the specific results. The FPGA-based speech recognition platform designed in this study has the lowest power consumption of 2.21W and the highest energy efficiency ratio of 26.15.

To verify the running speed of the FPGA-based speech recognition platform, the AISHELL-ASR0009-OS1 speech dataset is used for testing and compared with the Feiteng CPU and IntelCPU. AISHELL-ASR0009-OS1 is an open-source Chinese speech dataset with a recording time of 178 hours, containing 400 speakers from different accent regions in China. However, the AISHELL-ASR0009-OS1 dataset primarily encompasses 11 specific domains, including smart homes and autonomous driving. Consequently, it may not comprehensively encompass all daily communication scenarios, which, to some extent, constrains its scalability. From Figure. 12, as the test voice strips increase, the speech processing time of all three platforms increases. However, the platform designed in this study has always had the shortest speech processing time, faster running speed, and good speech processing performance.

Table 2: Performance comparison of different speech recognition platforms

-	Yu Y	Li J	Our
Quantitative approach	8bit	16bit	8bit
LUT	94763	121472	43863
BRAM	165	467	150
DSP	516	664	331
FF	150848	159872	23593
Computational performance/GOPS	398	232.4	57.8
Consumption/W	16.8	9.63	2.21
Energy efficiency ratio	23.69	24.13	26.15



a)Chip resource distribution and usage

b)Utilization

Figure 11: Hardware resource usage of FPGA

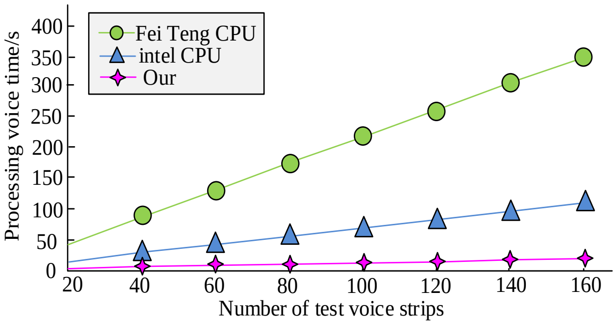


Figure 12: Comparison of speech processing time on three platforms

5 Discussion

To further improve the accuracy of SRAs, this study designed an AM-based NNA-BSR algorithm and an FPGA-based speech recognition platform. Unlike the language recognition methods proposed by Song Z, Yang S, Lin Y, Ma L, and Gao C, this research algorithm combined visual information and could effectively improve the robustness of the algorithm for speech recognition in noisy environments by combining visual and auditory information. The data showed that compared with the five algorithms, A-LipNet, Asymmetric BLSTM, LipNet, TM-seq2seq, and AV LipNet, the proposed AM-based NNA-BSR algorithm had the lowest WER and CER, with 1.59% and 0.68%, respectively. It had high accuracy and good speech recognition performance. This proved that combining visual information could effectively improve the speech recognition performance of algorithms. The proposed algorithm had the fastest convergence speed and the final loss value was lower. It converged quickly when the number of training rounds was less than 10 and tended to stabilize when the number of training rounds was 20. As the number of trainings increased, the WER and CER of the research algorithm also decreased the fastest. The CER has stabilized around 30 rounds and remained at a very low error rate. When the signal-to-noise ratio was 5dB, the speech recognition accuracy of the proposed SRA was 97.38%. This was because the proposed feature fusion structure could more effectively utilize visual and auditory information to establish potential correlations, thereby enhancing the learning ability of the algorithm. The FPGA-based speech recognition platform used many DSP units in its design, with a utilization rate of 83.2%. In addition, compared with the FPGA-based speech recognition platform proposed by Li J *et al.* and Yu Y *et al.*, the FPGA-based speech recognition platform designed in this study had the lowest power consumption of 2.21W, the highest energy efficiency ratio of 26.15, the shortest processing time, and faster running speed. This might be because this study used fewer hardware resources in the design of the speech recognition platform, thus having certain advantages in power consumption and energy efficiency ratio.

6 Conclusion

With the advent of the intelligent era, SRAs have become an important channel for communication between people and machines. Currently, the accuracy of speech recognition is still not ideal, which to some extent hinders the further development and application of AI. In response to this issue, this study designed an AM-based neural network audio-visual dual-mode SRA and an FPGA-based speech recognition platform. The feasibility and effectiveness of the designed algorithm and platform were demonstrated through experimental results. Applying the proposed method to practical speech recognition can help improve the accuracy of speech recognition and enhance the real-time and flexibility of human-computer interaction. However, when designing a speech recognition platform based on FPGA, the research adopted the HLS approach, which not only shortened the development cycle and efficiency but also to some extent affected the circuit quality. Therefore, in future research, Verilog HDL should be further adopted to optimize hardware design.

Funding

The research is supported by the Research Foundation of the Natural Science Foundation of Hunan Province, (Grant No. 2024JJ7189); The Scientific Research Project of the Hunan Provincial Department of Education (Grant No. 22C0539); the Social Science Project of Hunan Provincial Achievement Review Association, (Grant No. XSP24YBC319).

Conflict of interest

The authors declare that they have no conflicts of interest.

References

[1] Manoharan S, Ponraj N. Analysis of complex non-linear environment exploration in speech recognition by hybrid learning technique. Journal of Innovative Image Processing (JIIP), 2(4): 202-209, 2020. http://dx.doi.org/10.36548/jiip.2020.4.005

[2] Xue F, Yang T, Liu K, Hong Z, Cao M, Guo D, Hong R. LCSNet: End-to-end Lipreading with Channel-aware Feature Selection. ACM Transactions on Multimedia Computing, Communications and Applications, 19(1): 1-21, 2023. http://dx.doi.org/10.1145/3524620

[3] Zhou S, Pereida K, Zhao W, Schoellig A P. Bridging the model-reality gap with lipschitz network adaptation. IEEE Robotics and Automation Letters, 7(1): 642-649, 2021. http://dx.doi.org/10.1109/lra.2021.3131698

[4] Beasley A E, Clarke C T, Watson R J. An OpenGL compliant hardware implementation of a graphic processing unit using field programmable gate array–system on chip technology. ACM Transactions on Reconfigurable Technology and Systems (TRETS), 14(1): 1-24, 2020. http://dx.doi.org/10.1145/3410357

[5] Radner H, Stange J, Büttner L, Czarske J. Field-programmable system-on-chip-based control system for real-time distortion correction in optical imaging. IEEE Transactions on Industrial Electronics, 68(4): 3370-3379, 2020. http://dx.doi.org/10.1109/tie.2020.2979557

[6] Kumar K, Ramkumar K R, Kaur A. A lightweight AES algorithm implementation for encrypting voice messages using field programmable gate arrays. Journal of King Saud University-Computer and Information Sciences, 34(6): 3878-3885, 2022. http://dx.doi.org/10.1016/j.jksuci.2020.08.005

[7] Rao A S, Krishna B V, Saravanan D, David D B, Devi O R, Asokan A, David D S. Supervision calamity of public opinion actions based on field programmable gate array and machine learning. International Journal of Nonlinear Analysis and Applications, 12(2): 1187-1198, 2021. http://dx.doi.org/10.22075/IJNAA.2021.5195

[8] Koyuncu I, Yilmaz C, Alcin M, Tuna M. Design, and implementation of hydrogen economy using artificial neural network on field programmable gate array. International Journal of Hydrogen Energy, 45(41): 20709-20720, 2020. http://dx.doi.org/10.1016/j.ijhydene.2020.05.181.

[9] Chu K C, Chang K C, Wang H C, Wang H C, Lin Y C, Hsu T L. Field-programmable gate array-based hardware design of optical fiber transducer integrated platform. Journal of Nanoelectronics and Optoelectronics, 15(5): 663-671, 2020. http://dx.doi.org/10.1166/jno.2020.2835

[10] Rodríguez-Borbón J M, Kalantar A, Yamijala S S, Oviedo B, Najjar W, Wong B M. Field programmable gate arrays for enhancing the speed and energy efficiency of quantum dynamics simulations. Journal of chemical theory and computation, 16(4): 2085-2098, 2020. http://dx.doi.org/10.1021/acs.jctc.9b01284

[11] Gholami M, Zaman Farsa E, Karimi G. Reconfigurable field-programmable gate array-based on-chip learning neuromorphic digital implementation for nonlinear function approximation. International Journal of Circuit Theory and Applications, 49(8): 2425-2435, 2021. http://dx.doi.org/10.1002/cta.3075

[12] Koenecke A, Nam A, Lake E, Nudell J, Quartey M, Mengesha Z, Toups C, Rickford J R, Jurafsky D, Goel S. Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14): 7684-7689, 2020. http://dx.doi.org/10.1073/pnas.1915768117

[13] Haeb-Umbach R, Heymann J, Drude L, Watanabe S, Delcroix M, Nakatani T. Far-field automatic speech recognition. Proceedings of the IEEE, 109(2): 124-148, 2020. http://dx.doi.org/10.1109/JPROC.2020.3018668

[14] Song Z. English speech recognition based on deep learning with multiple features. Computing, 102(3): 663-682, 2020. http://dx.doi.org/10.1007/s00607-019-00753-0

[15] Yang S, Gong Z, Ye K, Wei Y, Huang Z. EdgeRNN: a compact speech recognition network with spatio-temporal features for edge computing, IEEE Access. 8: 81468-81478, 2020. http://dx.doi.org/10.1109/ACCESS.2020.2990974

[16] Lin Y, Guo D, Zhang J, Chen Z, Yang B. A unified framework for multilingual speech recognition in air traffic control systems. IEEE Transactions on Neural Networks and Learning Systems, 32(8): 3608-3620, 2020. http://dx.doi.org/10.1109/TNNLS.2020.3015830.

[17] Muraru S, Cocianu C L. Spoken Digit Recognition using the k-Nearest-Neighbor method and Mel Frequency Cepstral Coefficients. Informatica Economica, 28(2), 2024. http://dx.doi.org/10.24818/issn14531305/28.2.2024.01

[18] Ma L. Research on the Construction of English Translation Model for Speech Recognition Based on Multiple Information Sources and Lexical Assistance. Informatica, 48(6), 2024. http://dx.doi.org/http://dx.doi.org/10.31449/inf.v48i6.5588

[19] Gao C, Delbruck T, Liu S C. Spartus: A 9.4 TOp/s FPGA-based LSTM accelerator exploiting spatio-temporal sparsity. IEEE Transactions on Neural Networks and Learning Systems, 35(1): 1098-1112, 2022. http://dx.doi.org/10.48550/arXiv.2108.02297

[20] Wang X, Cheng M, Eaton J, *et al.* Fake node attacks on graph convolutional networks. Journal of Computational and Cognitive Engineering, 1(4):

165-173, 2022. http://dx.doi.org/10.47852/bonviewjcce2202321

[21] Dhyani M, Kumar R. An intelligent Chatbot using deep learning with Bidirectional RNN and attention model. Materials today: proceedings, 34(5): 817-824, 2021. http://dx.doi.org/10.1016/j.matpr.2020.05.450

[22] Liu Y, Song Z, Xu X, Rafique W, Zhang X, Shen J, Khosravi M R, Qi L. Bidirectional GRU networks-based next POI category prediction for healthcare. International Journal of Intelligent Systems, 37(7): 4020-4040, 2022. http://dx.doi.org/10.1002/int.22710

[23] Jiang S, Pan P, Ou Y, Batten C. PyMTL3: A Python framework for open-source hardware modeling, generation, simulation, and verification. IEEE Micro, 40(4): 58-66, 2020. http://dx.doi.org/10.1109/MM.2020.2997638

[24] Li J, Un K F, Yu W H, Mak P, Martins R P. An FPGA-based energy-efficient reconfigurable convolutional neural network accelerator for object recognition applications. IEEE Transactions on Circuits and Systems II: Express Briefs, 68(9): 3143-3147, 2021. http://dx.doi.org/10.1109/TCSII.2021.3095283

[25] Yu Y, Wu C, Zhao T, Wang K, He L. OPU: An FPGA-based overlay processor for convolutional neural networks. IEEE Transactions on Very Large-Scale Integration (VLSI) Systems, 28(1): 35-47, 2019. http://dx.doi.org/10.1109/TVLSI.2019.2939726