Application of Speech Recognition Technology Based on Multimodal Information in Human-Computer Interaction

Yuan Zhang
Xuchang Vocational Technical College
Henan Province, Data Intelligence and Security Application Engineering Technology Research Center, Xuchang 461000, China

***Abstract*—Multimodal human-computer interaction is an important trend in the development of human-computer interaction field. In order to accelerate the technological change of human-computer interaction system, the study firstly fuses Connectionist Temporal Classification algorithm and attention mechanism to design a speech recognition architecture, and then further optimizes the end-to-end architecture of speech recognition by using the improved artificial swarming algorithm, to obtain a speech recognition model suitable for multimodal human-computer interaction system. One of them, Connectionist Temporal Classification, is a machine learning algorithm that deals with sequence-to-sequence problems; and the Attention Mechanism allows the model to process the input data in such a way that it can focus its attention on the relevant parts. The experimental results show that, the hypervolume of the improved swarm algorithm converges to 0.861, which is 0.099 and 0.059 compared to the ant colony and differential evolution algorithms, while the traditional swarm algorithm takes the value of 0.676; the inverse generation distance of the improved swarm algorithm converges to 0.194, while that of the traditional swarm, ant colony, and differential evolution algorithms converge to 0.263, 0.342, and 0.246, respectively. Hypervolume and Inverse Generation Distance Measures the diversity and convergence of the solution set. The speech recognition model takes higher values than the other speech recognition models in the evaluation metrics of accuracy, precision, and recall, and the lowest values of the error rate at the character, word, and sentence levels are respectively 0.037, 0.036 and 0.035, ensuring higher recognition accuracy while weighing the real-time rate. In the multimodal interactive system, the experimental group's average opinion scores, objective ratings of speech quality, and short-term goal comprehensibility scores, and the overall user experience showed a significant advantage over the control group of the other methods, and the application scores were at a high level. The speech processing technology designed in this study is of great significance for improving the interaction efficiency and user experience, and provides certain references and lessons for the research in the field of human-computer interaction and speech recognition.**

Keywords*—Multimodal information; speech recognition; intelligent optimization algorithm; multimodal human-computer interaction; CTC; attention mechanisms; artificial bee colony algorithms

I. Introduction

Human-computer interaction (HCI) aims to realize information exchange and interaction between human and computer through computer technology. Single-modal human-computer interaction has been unable to meet the growing user needs, and multimodal human-computer interaction system is gradually becoming a research hotspot [1, 2]. Multimodal human-computer interaction (MMHCI) is a system that uses multiple input and output modes to carry out human-computer interaction, integrating various technologies such as speech recognition, gesture recognition, speech synthesis, expression parsing, etc., which is able to better adapt to the user's personalized needs, and has the advantages of improving interaction experience and interaction efficiency. It has the advantage of improving interaction experience and interaction efficiency [3].

At present, MMHCI is still a relatively new type of human-computer interaction and most of its key technologies are still in the exploration stage. Speech recognition is a key technology in the cognitive decision-making aspect of MMHCI, which affects the overall performance of the system. Currently, Hidden Markov, Recurrent Neural Network, Transformer structure and Connectionist Temporal Classification (CTC) are mostly used in the construction of speech recognition models, but the key technology enhancement of the models focuses on the accuracy rate, feature extraction, etc., and does not pay much attention to the specific speech recognition tasks and application environments. The recognition accuracy, real-time efficiency, semantic and emotional understanding, and single-signal processing of existing speech recognition technology still cannot meet the application requirements of MMHCI [4, 5].

In order to enhance the accuracy, robustness and improve the interactivity and flexibility of speech recognition in the interaction process of MMHCI speech recognition technology, the study utilizes inputs in the form of audio and video formed by multimodal information such as speech, semantics, expression, text and video, and designs a framework for speech recognition model with the help of improved CTC algorithm and a mixture of attention mechanisms. The adaptive multiple search strategy is also implemented on the basis of Artificial Bee Colony Algorithm (ABC) and this is used to complete the improved optimization of speech recognition model.

The study takes MMHCI as the research object, and carries out research on the key technologies of its speech recognition module, which helps to improve the accuracy and robustness of speech recognition, and obtains technological innovations, and provides theoretical support for the speech recognition algorithm of MMHCI. The introduction of ABC algorithm

provides a new solution to the optimisation problem of speech recognition. The research is expected to improve the reliability and efficiency of speech recognition technology in practical applications, enriching the user interaction experience; at the same time, it expands the application scenarios of speech recognition, promotes the development of MMHCI, and helps to realize a more intelligent and humanized interaction experience. The optimization research results of the key technologies of speech recognition in MMHCI provide new ideas and technologies for the development of the field of human-computer interaction, and promote the integration of the industrial and computer fields.

The study consists of six main sections, Section II is to carry out a summary of existing related research work in the field of speech recognition and human-computer interaction; then Section III elaborates on the construction and improvement of the speech recognition framework of the CTC joint attention mechanism, and designs the optimization algorithm of the improved ABC speech recognition framework; in Section IV, completes the performance test of the speech recognition model and the application analysis; Discussion is given in Section V and finally, Section VI concludes the paper.

II. Related Work

With the development of computer technology and artificial intelligence, multimodal human-computer interaction and speech recognition technology related to human-computer interaction have become a hot topic of concern in the community. Various neural networks and deep learning are widely used in speech recognition. Among the speech recognition researches in different countries, there have been relatively more researches on the recognition of English, Japanese or Chinese. Mukhamadiyev et al. designed an end-to-end deep neural network-hidden Markov model speech recognition model and a hybrid CTC-attention network with the small language Uzbek as the research object. The method can effectively utilize the connection time classification objective function and achieves improved recognition efficiency and accuracy, with a speech recognition error rate of 14.3% on the Uzbek dataset [6]. Deep learning models have been used effectively in speech recognition tasks, Dua et al. extended the application of convolutional neural networks to the recognition of speech signals and developed a speech-to-text recognition system. The method achieves recognition accuracy of 89.15 per cent and word error rate of 10.56 per cent on continuous and extensive lexical sentences of speech signals of different pitches [7]. Świetlicka et al. have used principal component analysis in combination with multilayer perceptual networks for fluent and interfering speech signals to analyse their application in describing the dimensionality reduction of speech signal variables. The experimental results show that the method achieves 76% total classification accuracy compared to Kohonen network [8]. Based on the network training techniques, Reza et al. designed a stacked five-layer custom residual convolutional neural network and seven-layer bi-directional gated recurrent units, where the network units all contain learnable layer normalization techniques based on element affine parameters. The character error rate of the model is 4.7 and 3.61% based on the public datasets librisspeech, LJ Speech validation [9].

With the rapid development of deep learning, artificial intelligence and other computer technologies, gesture recognition, speech recognition and other technologies related to human-computer interaction have also made great progress. The application of speech recognition technology in the field of human-computer interaction is gradually increasing, and researchers have conducted performance improvement studies around specific application tasks of speech recognition technology. Lv et al. summarized the research on human-computer interaction and speech recognition based on Web of Science, an academic literature database. The study found that intelligent human-robot interaction and deep learning have made great progress in gesture recognition, speech recognition and emotion recognition, and deep learning can effectively improve the recognition accuracy [10]. In order to improve the service quality and control effect of robots, Pan designed a command understanding method based on command intent understanding and key information extraction, as well as a human-robot voice interaction system with good application effect based on microphone array, voice wake-up and speech recognition [11]. Mavropoulos et al. designed an MMHC system based on knowledge representation, speech recognition and synthesis, sensor data analysis and Computer vision designed a MMHCI. The system can collect and monitor patient related information for healthcare [12]. Liu et al. conducted a study on speech interaction based on emotional Internet of Things, designing a multi-stage deep transfer learning scheme for the problem of limited large-scale emotion labeled datasets, and the experimental results show that the model is effective and superior in terms of naturalness and emotional expressiveness [13]. The use of speech recognition technology in the teaching process to assist teachers in correcting the pronunciation of spoken English has significant application effects. Ran et al. improved the speech recognition algorithm based on artificial intelligence speech recognition technology and designed a speech cutting model based on phonemic level speech error correction, including speech front-end processing and feature parameter extraction. The experimental results verify the effectiveness of the model [14].

In summary, there have been many applications and researches on speech recognition technology and human-computer interaction, but the technical improvement of the model mainly focuses on improvement of recognition accuracy and the reduction of error rate, and does not pay much attention to the specific speech recognition task and application environment. Moreover, there are relatively few studies on speech recognition for human-computer interaction with multimodal information, and the application of the existing speech recognition technology in multimodal human-computer interaction system is still immature and poorly adapted. In this study, speech recognition technology is improved by utilizing multimodal information.

III. Optimized Speech Recognition Model Design Based on Multimodal Information

In order to meet the technical demands of multimodal human-computer interaction, the study firstly designs a speech recognition framework for multimodal human-computer interaction based on the CTC and Attention mechanism; then improves and optimizes the CTC-Attention decoding scheme;

and finally introduces the ABC algorithm to improve the global searching ability and adaptive ability of the hybrid speech recognition model.

A. Speech Recognition Model Design Based on CTC and Attention Mechanism

Traditional human-computer interaction is mostly limited to keyboard, mouse or touch screen inputs, with a single mode of interaction, mostly relying on independent speech or gesture recognition. MMHCI technology is an important innovation in the field of artificial intelligence and interaction, integrating multiple input modes, including speech, gesture, touch, facial expression and eye movement, etc., and fusing the input data of different modes. Combining data from various input modes enhances the accuracy and intelligence of responses in multimodal human-computer interaction. At the same time, MMHCI can adaptively learn according to the user's interactive behavior habits and provide more convenient and personalized interactive services. MMHCI integrates multiple technologies such as speech recognition, image recognition, motion sensing, etc., and utilizes multi-modal information to complete the interaction, which involves multiple modules such as information input, multi-modal interaction information fusion and processing, multi-modal interaction information feedback, etc. Speech recognition technology is one of the key components of MMHCI, and it is also the most important component of MMHCI, as it can be used for the interaction between different modalities. Speech recognition technology is an important part of MMHCI, which is the key link to enhance user experience and interaction richness [15]. The MMHCI framework designed in the study consists of three parts: data input, cognitive decision control and output. The cognitive decision control includes the recognition of multimodal information such as speech, image and video, text dialogue, and facial expression. The study is aimed at the improvement of speech recognition accuracy, cross-modal information fusion and recognition efficiency, and the basic techniques such as the CTC algorithm and attention mechanism are utilized for speech recognition.

The speech recognition model designed in the study is an end-to-end encoder-decoder network structure. CTC is an algorithm for processing sequence data, which mainly solves the label alignment problem due to the change in the length of the sequence data [16, 17]. CTC contains a CTC loss function, which can be applied to the sequence data with variable-length labels during the training of the neural network without the need to align the speech. The calculation process of CTC network update is shown in Eq. (1), in which X denotes the speech feature sequence, $X = x_1, x_2, ..., x_T$; Y denotes the text sequence, $Y = y_1, y_2, ..., y_U$; $A(Y)$ denotes the set of all the aligned sequences corresponding to Y , a_i ; a_i denotes the path mapped from X to Y ; P denotes the likelihood probability of the mapping; and $blank(\in)$ is the special character introduced by the CTC to solve the problem of the model's input/output alignment.

$$\begin{cases} CTC_{Loss} = \sum_{(X,Y)\in D} -\log P(Y|X) \\ P(Y|X) = \sum_{A\in A(Y)} P(A|X) = \sum_{A\in A(Y)} \prod_{i}^{T} P(a_i|X) \end{cases} \quad (1)$$

Attention Mechanism (AM) is a model that simulates the mechanism of human attention allocation for solving the problem of information filtering and weighting when processing sequence data. The network architecture of the speech recognition model designed for the study draws on the Transformer model. The Transformer model is a model that solves the sequence-to-sequence problem based on the Attention mechanism, and the structure of the Transformer feature extractor is shown in Fig. 1.

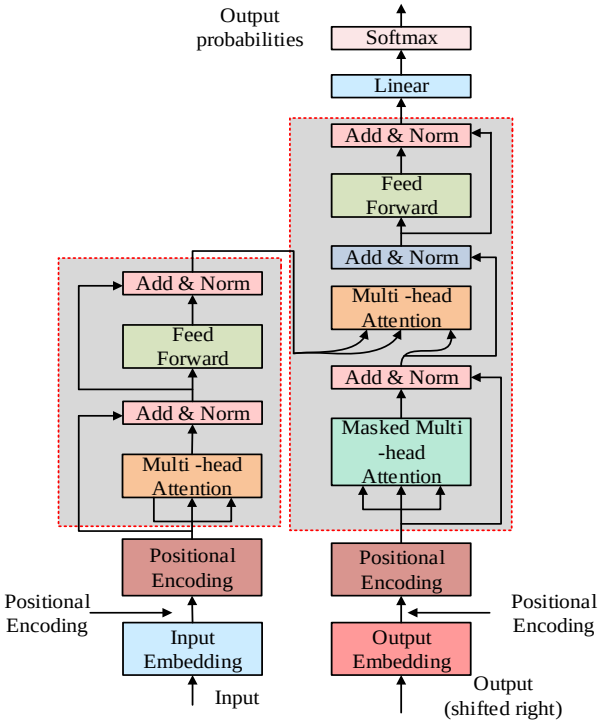

Fig. 1. Structure of transformer feature extractor.

The Transformer model treats the input and output sequences as a series of encoder and decoder stacked layers, with the different layers being composed of Multi-Headed Self-Attention Mechanism (MHSA) and Feedforward neural network (FFN). FFN), AM can weigh different positions in sequence processing. However, Transformer is weak in dealing with fine-grained local feature extraction problems, the study adopts the Conformer model, an improved structure of Transformer, to fully learn the local feature information, and the model structure is shown in Fig. 2. The Conformer model contains a Conformer module, which contains the Self-AM and CNN which is placed between the FFN layers. The Conformer introduces a position-sensitive sinusoidal function when calculating the AM scores, which helps the model to handle the weighting of acoustic features at different time steps. In addition, the Conformer model employs forward weighting and layer normalization to further improve the model performance [18, 19]. The study hybridizes CTC with Conformer model, firstly, Fbank features and 3-dimensional fundamental

frequency features are selected for training the speech recognition model, and SpecAugment method is utilized for data enhancement of speech. Then after convolutional sampling, linear mapping and regularization the input encoder Conformer block and finally decoding is done by CTC and AM decoder.

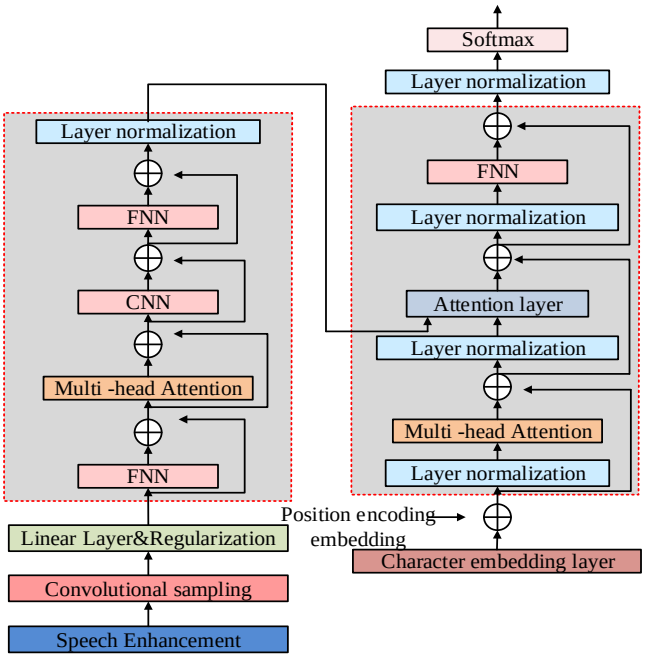

Fig. 2. Structural diagram of the transformer model.

The completion process of the Conformer model is shown in Eq. (2) and x denotes the input.

$$\begin{cases} x_{FFN1} = x + \frac{1}{2} FFN(x) \\ x_{MHSA} = x_{FFN1} + MHSA\left(x + \frac{1}{2} FFN(x)\right) \\ x_{Conv} = x_{MHSA} + Conv\left(x_{MHSA}\right) \\ x_{FFN2} = Layernorm\left(x_{Conv} + \frac{1}{2} FFN\left(x_{Conv}\right)\right) \end{cases} \tag{2}$$

The loss function calculation of the CTC-Conformer hybrid model is shown in Eq. (3), where x, y denotes the acoustic features, the real text, respectively; and λ denotes the hyperparameters balancing CTC and AM.

$$Loss_{combined}(x,y) = \lambda * Loss_{ctc}(x,y) + Loss_{attention}(x,y) \tag{3}$$

The decoding score for the CTC-Conformer hybrid model is calculated in Eq. (4).

$$Score_{final} = \lambda * Score_{ctc} + Score_{attention} \tag{4}$$

B. Improvement of CTC-AM Speech Recognition Framework

CTC-Conformer hybrid model has achieved a large advantage over the traditional Transformer model in speech recognition, but it still has shortcomings such as model convergence difficulty and complex calculation. In this study, improvements are made to AM and CTC algorithms. The decoding process of CTC-Conformer is shown in Fig. 3, as seen in Fig. 3, it is difficult for the model to converge in the face of a longer feature sequence input; and the computational complexity of Self-AM is $O(L^2)$, and the computational complexity of decoding of the hybrid model is higher [20].

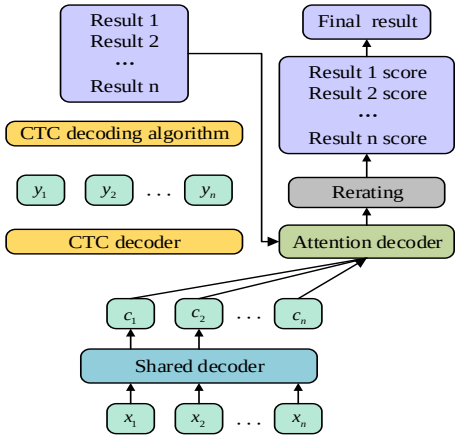

Fig. 3. Schematic diagram of CTC transformer decoding process.

For the long sequence speech recognition task, the study introduces the concept of Probability Sparse (Prob-Sparse). The AM used in the study is soft attention, as shown in Eq. (5). In Eq. (5), Q, K, V corresponds to query, key, and value, respectively; d denotes the sequence dimension;

$$A(Q,K,V) = Soft\max\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{5}$$

The query matrix Q has some sparsity, which will lead to more redundant computation if the attention of all queries of the query vector is computed. The study has calculated the attention score and distribution difference using Kullback-Leibler (KL) scatter, which measures the difference between the generated sample distribution and the target distribution. The calculation procedure is shown in Eq. (6). In Eq. (6), L denotes the length of the sequence; q_i, k_i, v_i denotes the i th line of Q, K, V; and p, U denotes the distribution of the attention scores and the uniform distribution, respectively.

$$KL(p\|U) = \ln \sum_{j=1}^{L} e^{\frac{q_i k_j^T}{\sqrt{d}}} - \frac{1}{L}\sum_{j=1}^{L}\frac{q_i k_j^T}{\sqrt{d}} - \ln L \tag{6}$$

The sparsity metric value $M_{Sparse}(q_i, K)$ of the query matrix Q is calculated from Eq. (6), see Eq. (7). The calculation process can be accelerated by keeping the larger $M_{Sparse}(q_i, K)$ queries. In Eq. (7), K denotes the random sampling of K; L denotes the number of samples.

$$M_{Sparse}\left(q_i,K\right)=\max_j\left\{\frac{q_i k_j^T}{\sqrt{d}}\right\}-\frac{1}{L}\sum_{j=1}^{L}\frac{q_i k_j^T}{\sqrt{d}} \tag{7}$$

L the calculation process is shown in Eq. (8), and r_{sample} denotes the sample sampling factor.

$$L=r_{sample}\ln L \tag{8}$$

In summary, the computation of Prob-Sparse Attention for the research design is shown in Eq. (9), where I_{Sparse} denotes the index of the L_{Sparse} query and $L_{Sparse}=r_{sample}L$.

$$\left(q_i,K,V\right)=\begin{cases}\sum_j^L p\left(k_j\middle|q_i\right)v_j & if\ \ i\in I_{Sparse}\\ v_i & else\end{cases} \tag{9}$$

The CTC algorithm employs a dynamic programming algorithm to learn the mapping relationship between sequences, and uses maximum likelihood estimation to learn the probability of mapping paths, but the increase in the length of the input sequences is not conducive to the CTC algorithm to find feasible paths, and the blank labels introduced by the CTC will lead to the model easily falling into the local optimal situation. The conditional probability $p\left(l\middle|X_{1:T}\right)$ for a given target sequence $X_{1:T}$ is computed in Eq. (10), where ϖ denotes the path of temporally concatenated observation labels; l denotes the true output; and B denotes the many-to-one mapping from π to l.

$$p\left(l\middle|X_{1:T}\right)=\sum_{\pi\in B^{-1}(l)}p\left(\varpi\middle|X_{1:T}\right) \tag{10}$$

The training process of CTC is constantly optimizing the CTC loss function $Loss_{ctc}$ to complete, but CTC exists a large number of ϖ, the optimization of the loss function will cause the model to produce poorly aligned output; and the error signal is positively correlated with π, the positive feedback of the error signal makes the probability prone to fall into a certain single path, which leads to the model overfitting phenomenon. In this regard, the study introduces the maximum conditional entropy to improve the CTC, and the network structure of the improved CTC-Attention decoding scheme is shown in Fig. 4.

The CTC damage function based on maximum conditional entropy $Loss'_{ctc}$ is calculated in Eq. (11), α denotes the coefficient of maximum conditional entropy regularization; $H\left(p\left(\pi\middle|l,X\right)\right)$ denotes the entropy of feasible paths of input and target sequences. The maximum conditional entropy reduces the influence of the positive feedback of error information on the model training search, and the loss value calculated by Eq. (11) and the loss function value of the attention mechanism can be calculated according to Eq. (3) to obtain the loss value of the hybrid Improve CTC-Prob-Sparse Attention model.

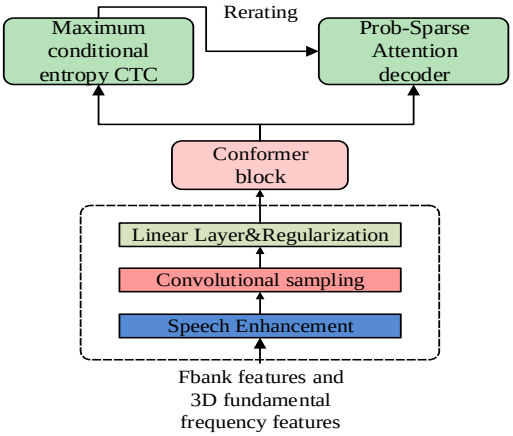

Fig. 4. Network structure diagram of improved CTC attention decoding scheme.

$$Loss'_{ctc}=Loss_{ctc}-\alpha H\left(p\left(\varpi\middle|l,X\right)\right) \tag{11}$$

C. CTC-AM Speech Recognition Framework Optimized by Fusion and Improved ABC Algorithm

After the optimization of CTC loss function and AM, in order to realize the adaptive recognition of speech recognition model, the study introduces ABC algorithm to optimize the Improve CTC-Prob-Sparse Attention hybrid speech recognition model. ABC is a swarm intelligence global optimization algorithm that draws on the honey harvesting behavior of honeybee colonies. ABC algorithm divides honeybees into hiring bees, following bees and scout bees, and considers feasible solutions as food sources for bees. The hired bees search for new honey sources based on the old honey source information and determine whether to update the solution based on the evaluated objective function value. The following bee joins the honey source search process based on the information shared by the hiring bee; the scouting bee's task is to randomly select a new honey source in the whole solution space to try when the hiring bee does not make further progress, and the algorithm flow is shown in Fig. 5. The ABC algorithm is chosen for the study to further improve the global search ability and adaptive capability of the hybrid speech recognition model.

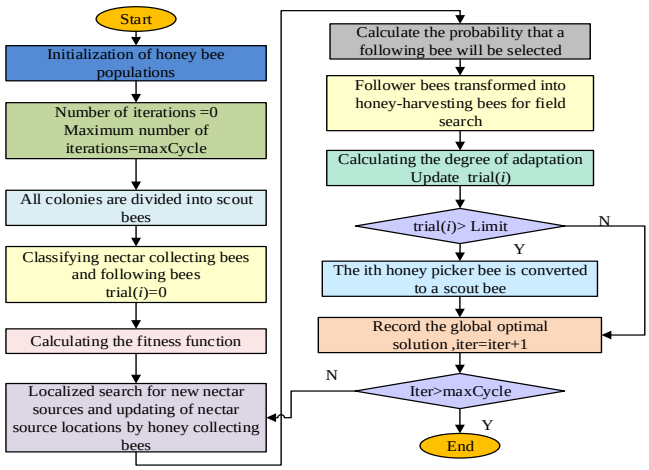

Fig. 5. Schematic diagram of ABC algorithm process.

The total number of bees is N_s, the population size of hired bees is N_e, the population of following bees is N_u, and the search space is defined as S. The number of honey sources is equal to the number of hired bees, the search feasible solution for the i th hired bee X_i^j is computed in Eq. (12), $X_{\max}^j, X_{\min}^j$ denotes the maximum and minimum values of the j dimensional components of the honey sources, $j \in \{1,2,..,D\}$, D denote the individual vector dimensions, respectively.

$$X_i^j = X_{\min}^j + rand(0,1)\left(X_{\max}^j - X_{\min}^j\right) \tag{12}$$

The benefit degree value of the honey source $fitness_i$ is calculated in Eq. (13), where f_i represents the objective function of the optimization problem.

$$fitness_i = \begin{cases} \dfrac{1}{1+f_i} \\ 1+abs(f_i) \end{cases} \tag{13}$$

The nectar source $fitness_i$ determines the probability that the following bee will be selected P_i, which is calculated as shown in Eq. (14).

$$P_i = \frac{f_i}{\sum_{n=1}^{N_e} f_n} \tag{14}$$

The search position generation calculation for hired bees is shown in Eq. (15), and the nectar source update occurs when the fitness value of the new search position is larger. However, when the position of the nectar source has not been updated, the hired bees of this nectar source are converted to scout bees, and the position updating process is the same as Eq. (12).

$$\begin{aligned} &new\_X_i^j = X_i^j + rand[-1,1]\left(X_i^j - X_k^j\right) \\ &j \in \{1,2,...,D\} \quad k \in \{1,2,...,N_e\} \end{aligned} \tag{15}$$

However, the ABC algorithm still has some application defects, ABC has a good global search ability in the hiring bee, following bee, and scouting bee stages, but the local search ability in different stages is weakened. In this regard, the study introduces the Adaptive Double Search (ADS) strategy, which can improve the convergence of the ABC algorithm, and the improved ADSABC workflow is shown in Fig. 6.

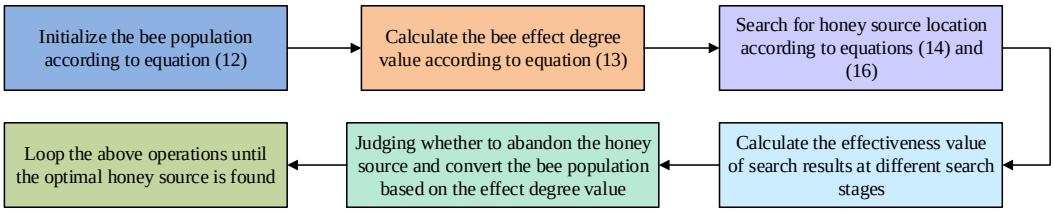

Fig. 6. Improved ADSABC workflow.

The search process of ADS is shown in Eq. (16), in which $x_{i,j}$, $x_{r,j}$ and $v_{i,j}$ denote the positions before and after the search of the hired bees following the bees, respectively; $x_{i,Global}$ denotes the global optimal nectar position under the current number of iterations; $x_{best,j}$ denotes the j elements of the global optimal position; α and μ denote the neighborhood search coefficients; and γ denotes the Kersey's variability factor, which helps the following bees to jump out of the local optimum. $\gamma = \tan\left[(\xi - 0.5)\pi\right]$ The following are used as the random numbers: ξ is the random number, $\xi \in [0,1]$; $\psi(t)$ is the adaptive adjustment factor, t is the number of cycles, and n is the number of iterations.

$$\begin{cases} v_{i,j} = x_{i,j} + \alpha^{n+1} rand\left(x_{best,j} - x_{i,j}\right) + \left(1-\psi(t)\right)\beta^{n+1}\left(x_{i,Global} - x_{i,j}\right) \\ v_{i,j} = x_{r,j} + \mu^{n+1} rand\left(x_{best,j} - x_{i,j} + \gamma\right) + \left(1-\psi(t)\right)\eta^{n+1}\left(x_{r,Global} - x_{r,j}\right) \\ \psi(t) = 1 - rand^{\left(1-t/MaxCycle\right)^2} \end{cases} \tag{16}$$

IV. Performance and Application Analysis of Speech Recognition Technology Based on Multimodal Information

In order to test the effectiveness of the research-designed speech recognition algorithm in multimodal human-computer interaction systems, the research designed two parts of performance test experiments, namely, the performance test of the improved ABC algorithm and the performance and application effect analysis experiments of the Improve CTC-Prob-Sparse Attention speech recognition framework.

A. Performance Test of Improved Swarm Algorithm

The performance of the ADSABC algorithm designed for the study is first analyzed by selecting the traditional ABC algorithm, Differential Evolution Algorithm (DE) and Ant Colony Optimization (ACO) algorithms for comparison, which are all based on Java language implementation. The performance and convergence of the optimization algorithms are analyzed by choosing the single-peak functions: The Sphere function, the Schwefel function, and the multi-peak functions, the Rastrigin function, the Griewank function, and the Ackley function.

Forty independent optimization experiments were set up to evaluate the convergence of different optimization algorithms from four angles, and the statistical results of the experiments are shown in Table I. Table I shows that the ADSABC algorithm has the best convergence performance for solving different test functions. The optimization value of the ADSABC algorithm for solving single-peak and multi-peak functions is smaller than that of the other algorithms, and the optimal solution can be found in all 40 independent experiments with a convergence rate of 100%. The ADSABC algorithm has the smallest number of solution iterations and the fastest convergence speed on average.

Hypervolume Indicator (HV) was selected to be associated with the Inverted Generational Distance Inverted Generational Distance (IGD) are chosen as the evaluation indexes of the algorithm, and the experimental results are shown in Fig. 7. HV is used to measure the size of the solution set occupied by the algorithm in the target space, and the larger the HV, the better the diversity and uniformity of the solution set is, and the IGD mainly focuses on the distance between the algorithm-generated solution set and the real optimal solution set, and the smaller the IGD, the better the performance of the algorithm. The smaller the IGD index, the better the performance of the algorithm. IGD and HV can jointly evaluate the iterative process and search effect of the algorithm, and the experimental results are shown in Fig. 7. As it can be seen in Fig. 7(a), the HV index of ADSABC algorithm takes the largest value, converges to 0.861 with the increase of iteration number, and the diversity and uniformity of the solution set are good. As seen in Fig. 7(b), the IGD curve of the ADSABC algorithm converges around the minimum value of 0.194, which is significantly different from other algorithms. It can be seen that the solution performance of the improved ABC algorithm of the study is good.

TABLE I. COMPARISON OF CONVERGENCE PERFORMANCE OF DIFFERENT ALGORITHMS

Test function	Algorithm	Average convergence value	Frequency of convergence	Minimum number of iterations	Average time (s)
Sphere	ABC	1.53E-6	29	461	2.91
	ADSABC	2.826E-16	40	215	1.94
	DE	1.947E-9	16	367	3.64
	ACO	3.672E-11	27	406	4.09
Schwefel	ABC	6.462E-10	21	403	3.08
	ADSABC	1.723E-13	40	216	2.16
	DE	9.012E-10	29	310	4.61
	ACO	9.306E-8	31	403	5.06
Rastrigin	ABC	5.77E-6	31	343	3.26
	ADSABC	9.283E-11	40	271	2.54
	DE	8.735E-9	29	302	4.16
	ACO	1.374E-8	36	425	6.17
Griewank	ABC	1.565E-7	35	461	11.66
	ADSABC	2.853E-12	40	329	9.54
	DE	5.563E-8	29	464	11.65
	ACO	2.618E-9	31	506	10.36
Ackley	ABC	7.042E-6	26	349	9.54
	ADSABC	3.983E-12	40	321	8.32
	DE	3.786E-10	30	406	11.58
	ACO	2.853E-9	25	496	11.22

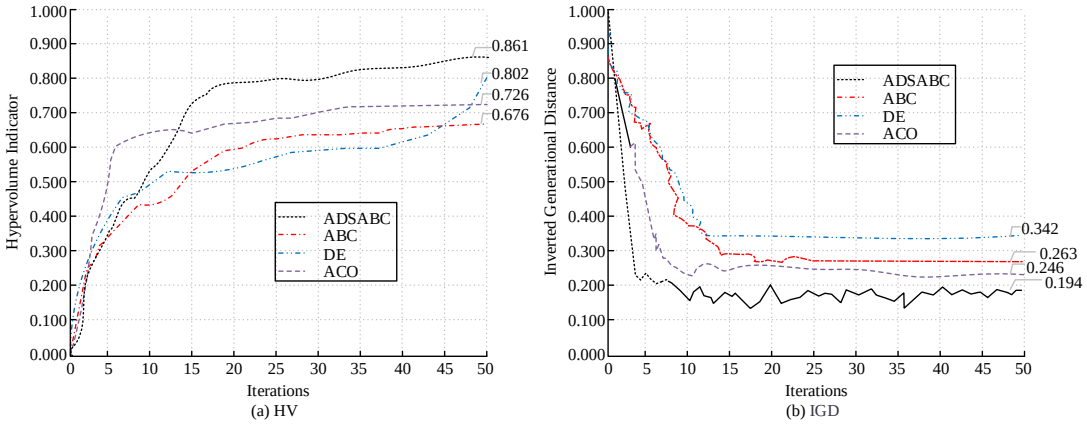

Fig. 7. Comparison of HV and IGD for different optimization algorithms.

B. Hybrid Improved Speech Recognition Model Performance Testing and Application Analysis

Based on Windows 7 operating system, hardware environment Intel Core i7, Intel Q270 series chipset, memory is 32 GB, hard disk space is 8 TB capacity, based on Python implementation programming. The acoustic features input to the model are 80-dimensional Fbank features combined with 3-dimensional fundamental frequency features. The hybrid improved model (HI-CTC-Conformer) designed in the study is compared and analyzed with the CTC-Conformer model before improvement, Wavelet Neural Network (WNN), and Transformer model.

LibriSpeech, Fisher, Mozilla, VoxPopuli, and AN4 datasets are selected as the experimental dataset, and the experimental analyzed data is selected to be divided into training and test sets in the ratio of 9:1. The accuracy and precision-recall curve (PR) results of different speech recognition models are shown in Fig. 8. As seen in Fig. 8(a), the HI-CTC-Conformer model designed for the study has the highest recognition accuracy curve, with a maximum accuracy of 94.54%; compared to the other models, the HI-CTC-Conformer model is more accurate in its recognition results. As can be seen in Fig. 8(b), the PR curve of HI-CTC-Conformer model is in the upper rightmost part of the coordinate axis, and the Average Precision (AP), which represents the area of the PR curve, takes the largest value, and the recall of HI-CTC-Conformer model can reach 0.84 when the precision rate is 90%. In the same experimental environment, the accuracy and PR curve meticulously and comprehensively validate the overall excellent performance of the HI-CTC-Conformer model.

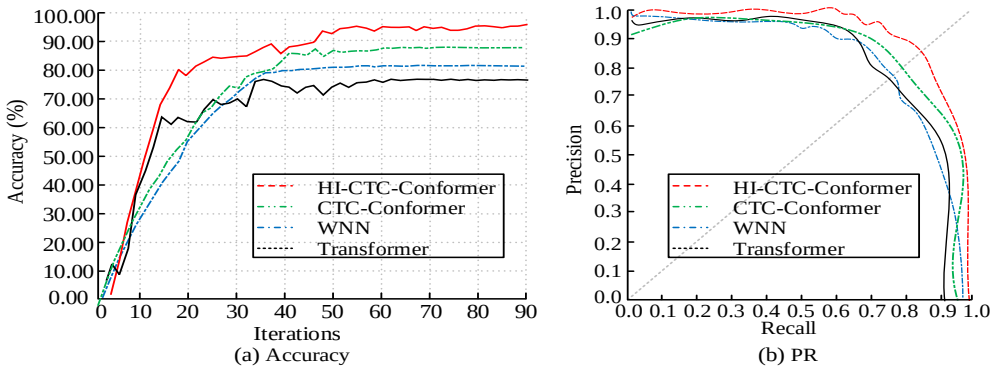

Fig. 8. Comparison of accuracy and PR curve of different recognition models.

TABLE II. RECOGNITION ERROR RATES OF DIFFERENT RECOGNITION MODELS

Model	Index	Index	LibriSpeech	Fisher	Mozilla	VoxPopuli	AN4
HI-CTC-Conformer	Test	WER	0.041	0.036	0.046	0.044	0.036
		CER	0.049	0.047	0.037	0.054	0.047
	Training	WER	0.043	0.043	0.044	0.043	0.039
		CER	0.041	0.046	0.044	0.054	0.054
CTC-Conformer	Test	WER	0.066	0.063	0.064	0.061	0.068
		CER	0.065	0.060	0.067	0.064	0.065
	Training	WER	0.063	0.066	0.074	0.071	0.063
		CER	0.054	0.067	0.065	0.075	0.075
WNN	Test	WER	0.073	0.069	0.068	0.073	0.064
		CER	0.060	0.073	0.060	0.074	0.078
	Training	WER	0.073	0.084	0.085	0.081	0.072
		CER	0.082	0.074	0.084	0.078	0.073
Transformer	Test	WER	0.086	0.083	0.084	0.087	0.088
		CER	0.091	0.097	0.086	0.088	0.091
	Training	WER	0.131	0.103	0.094	0.105	0.108
		CER	0.144	0.123	0.114	0.095	0.091

Word Error Rate (WER) and Character Error Rate (CER) of speech recognition are selected as the evaluation indexes of different models, and the experimental results are shown in Table II. As seen in Table II, the WER and CER of the HI-CTC-Conformer model are significantly lower than those of the other three models on different datasets, with the lowest WER of 0.036 and the lowest CER of 0.037. In contrast, the WER of the Transformer model reaches the highest of 0.131 and the CER reaches the highest of 0.144. The WER and CER are calculated as the edit distance between the recognition results output by the system and the standard reference text, which can indicate the degree of inaccuracy of the speech recognition model. The difference between the two is that CER is more sample than WER, and the editing operations at character level are more fine-grained than those at word level. It can be seen that the model designed in the study has a more significant improvement in recognition accuracy compared to various types of baseline models.

The statistics of the Sentence Error Rate (SER) and Real-time Factor (RTF) metrics of the model are shown in Fig. 9. As can be seen in Fig. 9(a), on different datasets, the SER of the HI-CTC-Conformer model designed for the study takes the lowest level, and the median level of the SER is under 0.05, while the SER of the other three models takes the value above 0.06, and the highest value reaches 0.14. The SER denotes the editing distance between the sentence outputted by the system and the reference text, and it can be seen that the overall accuracy of the HI-CTC-Conformer model is at a high level. As seen in Fig. 9(b), the RTF of the HI-CTC-Conformer model is also at the lowest level, with values taken from different datasets under 4.00%. The RTF measures the decoding speed of the speech recognition model as the ratio of the recognition time to the speech duration, and is used to evaluate the real-time performance of the system. Comprehensively, it can be seen that the HI-CTC-Conformer model has achieved a good balance between RTF and SER metrics, and the system can have both better real-time performance and recognition accuracy.

Finally, several subjects were recruited to analyze the application effect of the research-designed speech recognition model based on the multimodal human-computer interaction system with the subjective evaluation index Mean Opinion Score (MOS), the objective measurement index Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI), and the experimental results are shown in Fig. 10. Quality (PESQ) and Short-Time Objective Intelligibility (STOI) to analyze the application effect, and the experimental results are shown in Fig. 10. As can be seen in Fig. 10, the difference between the experimental group and the control group scores of the recovered scoring results is obvious, and the MOS scores of the design results of the study are higher than 3, indicating that the overall quality of the method is available and high. The PESQ and STOI scores are in the range of 2-4.5 and 0.5-1.0, respectively, and the scores are at a high level, and the quality of the speech recognition is high.

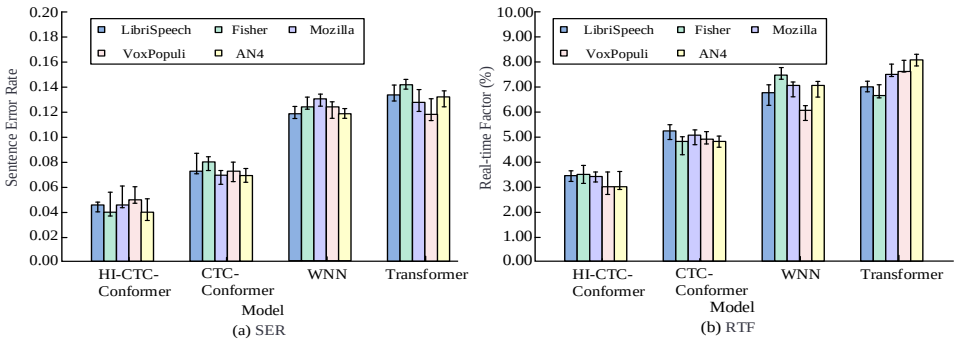

Fig. 9. Comparison of SER and RTF metrics for different speech recognition models.

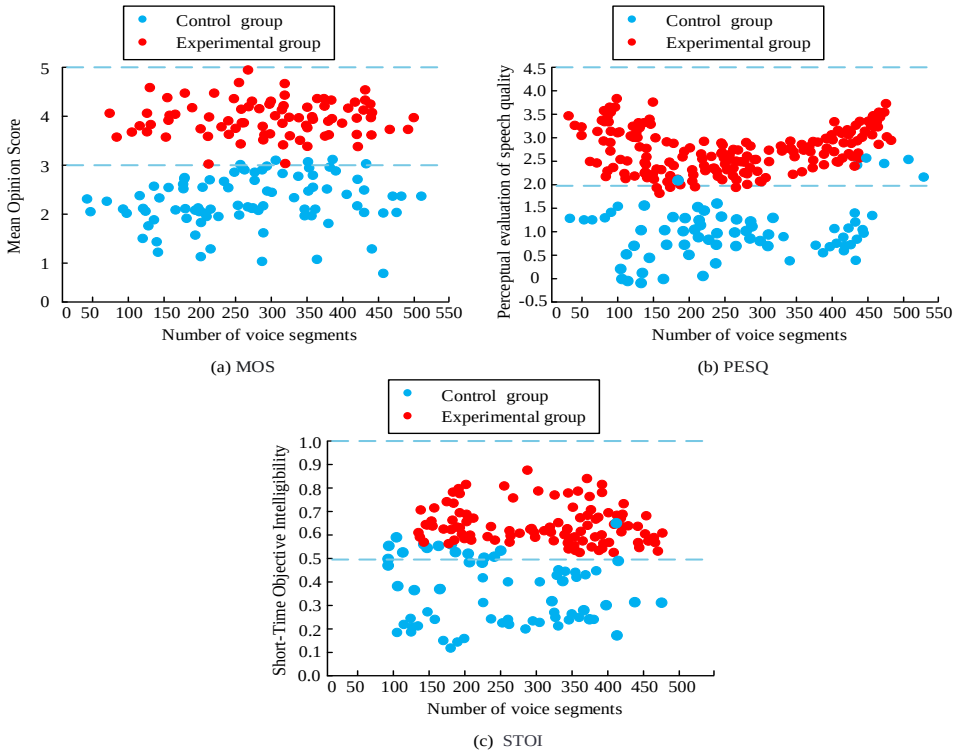

Fig. 10. Application effect analysis of speech recognition model.

V. DISCUSSION

MMHCI is a technology that has gradually emerged with the advancement of AI technology, using multiple perceptual modalities such as vision, sound, and touch to enhance and optimize the human-computer interaction experience. Compared with single-modal interaction, MMHCI is more natural, efficient, and intelligent, and has been widely used in the fields of smart home, intelligent healthcare, and intelligent transportation. Through the fusion of multimodal information, the HCI system can obtain more precise semantic information, which improves the accuracy and robustness of the system. Speech recognition technology is an important part of MMHCI, which can convert speech signals into text or commands to realize efficient communication between humans and computers. Meanwhile, speech recognition can be combined with other modalities to form a more comprehensive and intelligent interaction. In the process of summarizing the existing research work, it was found that researchers in study [6], study [7], study [8], and study [9] mainly used neural network and deep learning techniques to construct speech recognition technology, and their optimal performance achieved a vocabulary sentence recognition accuracy of 89.15%, a word error rate of 10.56%, and a character error rate of 3.61%. However, this type of research lacks targeted improvement analysis of speech recognition techniques in specific tasks. In addition, Mavropoulos et al. also applied speech recognition technology to MMHCI [12], but did not carry out performance improvement and adaptation optimization studies for speech recognition technology. The application of existing speech recognition techniques to MMHCI still faces performance and application challenges.

In this regard, the study designed a basic speech recognition framework based on the improved CTC algorithm and attention mechanism, and introduced the ABC algorithm in the intelligent optimization algorithm to improve and optimize the framework in depth. The research results show that the method outperforms other speech recognition models in terms of accuracy, precision, and recall evaluation metrics, with a maximum accuracy of 94.54%. The minimum values of error rates at the character, word, and sentence levels are 0.037, 0.036, and 0.035, respectively, ensuring higher recognition accuracy while weighing the real-time rate. The performance metrics take a significant advantage over existing work. The design of the study improves the accuracy and robustness of MMHCI speech recognition and achieves innovative and adaptive optimization of the algorithm. By improving the CTC algorithm, AM, it provides a new framework for speech recognition model, which provides a valuable reference and reference for subsequent research. Meanwhile, the introduction of intelligent optimization algorithms into the field of speech recognition improves the performance of the speech recognition model and assists the model in adapting to the complex and changing application scenarios in real interaction.

Therefore, in practical applications, the speech recognition framework can provide users with a more natural and smooth interaction experience, which significantly improves user satisfaction and convenience. At the same time, the algorithmic innovation allows MMHCI's speech recognition technology to be applied to more complex scenarios, which promotes cross-fertilization of multimodal information processing, deep learning, optimization algorithms, and other fields.

In future research work, researchers can further explore how to more effectively fuse information from different modalities, and utilize different modal information to achieve the complementarity and association of multiple information. Moreover, the fusion of more techniques can be further attempted to improve the efficiency and performance of the model. In this way, MMHCI can realize wider application and popularization.

VI. CONCLUSION

With the rapid development of intelligent technology, speech recognition technology has been increasingly used in human-computer interaction. In order to enhance the application effect of multimodal interaction system, the study designed a speech recognition framework based on improved CTC algorithm and attention mechanism.

The experimental results show that, the improved ABC algorithm has better convergence performance on different test functions, and the convergence value, convergence number and convergence rate are better than other algorithms. Meanwhile, the diversity and homogeneity of the solution set are better, with the maximum HV of 0.861 and the minimum IGD of 0.194. The maximum accuracy of the HI-CTC-Conformer model is 94.54%, the area of the precision vs. recall curve is the largest, and the model's recall is up to 0.84 when the precision rate is 90%. Compared with the baseline model, the recognition accuracy and efficiency of this model are significantly improved because of the low recognition error rate of "word", "character" and "sentence". In the multimodal interaction system, the MOS scores, PESQ and STOI scores of the experimental group are in the range of 3-5, 2-4.5, and 0.5-1.0, respectively, and the application effect is superior.

The study improved the accuracy and comprehension of speech recognition and provided a more natural and convenient interaction experience. However, the study did not involve speech synthesis and expression analysis for multimodal interaction systems, which can be a future research direction for multimodal information analysis in the field of human-computer interaction.

ACKNOWLEDGMENT

This work was supported in part by the Key Research and Development Program of Zhejiang Province under Grant 2024C01108.

REFERENCES

[1] J. Zhang, S. Wang, W. He, J. Li, Z. Cao, and B. Wei, "Projected augmented reality assembly assistance system supporting multi-modal interaction," Int J Adv Manuf Tech, vol. 123, no. 3, pp. 1353-1367, November 2022.

[2] E. Y. Oh, and D. Song, "Developmental research on an interactive application for language speaking practice using speech recognition technology," Etr&D-Educ Tech Res, vol. 69, no. 2, pp. 861-884, April 2021.

[3] A. Moin, F. Aadil, Z. Ali, and D. Kang, "motion recognition framework using multiple modalities for an effective human‐computer interaction," J Supercomput, vol. 79, no. 8, pp. 9320-9349, May 2023.

[4] A. S. Dhanjal, and W. Singh, "A comprehensive survey on automatic speech recognition using neural networks," Multimed Tools Appl, vol. 83, no. 8, pp. 23367-23412, March 2024.

[5] S. Ambrogio, P. Narayanan, A. Okazaki, A. Fasoli, C. Mackin, K. Hosokawa, and G. W. Burr, "An analog-AI chip for energy-efficient speech recognition and transcription," Nature, vol. 620, no. 7975, pp. 768-775, August 2023.

[6] A. Mukhamadiyev, I. Khujayarov, O. Djuraev, and J. Cho, "Automatic speech recognition method based on deep learning approaches for Uzbek language," Sensors-Basel, vol. 22, no. 10, pp. 3683-3705, May 2022.

[7] S. Dua, S. S. Kumar, Y. Albagory, R. Ramalingam, A. Dumka, R. Singh, and A. S. AlGhamdi, "Developing a speech recognition system for recognizing tonal speech signals using a convolutional neural network," Appl Sci-Basel, vol. 12, no. 12, pp. 6223-6235, June 2022.

[8] I. Świetlicka, W. Kuniszyk-Jóźkowiak,and M. Świetlicki, "Developing a speech recognition system for recognizing tonal speech signals using a convolutional neural network," Sensors-Basel, vol. 22, no. 1, pp. 321-336, January 2022.

[9] S. Reza, M. C. Ferreira, J. J. Machado, and J. M. R. Tavares, "A customized residual neural network and bi-directional gated recurrent unit-based automatic speech recognition model," Expert Syst Appl, vol. 215, no. 4, pp. 119293-119304, April 2023.

[10] Z. Lv, F. Poiesi, Q. Dong, J. Lloret, and H. Song, "Deep learning for intelligent human‐computer interaction," Appl Sci-Basel, vol. 12, no. 22, pp. 11457-11484, November 2022.

[11] S. Pan, "Design of intelligent robot control system based on human‐computer interaction," Int J Syst Assur Eng, vol. 14, no. 2, pp. 558-567, April 2023.

[12] T. Mavropoulos, S. Symeonidis, A. Tsanousa, P. Giannakeris, M. Rousi, E. Kamateri, and I. Kompatsiaris, "Smart integration of sensors, computer vision and knowledge representation for intelligent monitoring and verbal human-computer interaction," J Intell Inf Syst, vol. 57, no. 2, pp. 321-345, June 2023.

[13] R. Liu, Q. Liu, H. Zhu, H. and M. Cao, "Multistage deep transfer learning for EmIoT-Enabled Human‐Computer interaction," IEEE Internet Things, vol. 9, no. 16, pp. 15128-15137, August 2022.

[14] D. Ran, W. Yingli, and Q. Haoxin, "Artificial intelligence speech recognition model for correcting spoken English teaching," J Intell Fuzzy Syst, vol. 40, no. 2, pp. 3513-3524, February 2021.

[15] Y. Wu, and J. Li, "Multi-modal emotion identification fusing facial expression and EEG," Multimed Tools Appl, vol. 82, no. 7, pp. 10901-10919, September 2023.

[16] U. Maniscalco, P. Storniolo, and A. Messina, "Bidirectional multi-modal signs of checking human-robot engagement and interaction," Int J Soc Robot, vol. 14, no. 5, pp. 1295-1309, April 2022.

[17] L. Jia, X. Zhou, and C. Xue, "Non-trajectory-based gesture recognition in human-computer interaction based on hand skeleton data," Multimed Tools Appl, vol. 81, no. 15, pp. 20509-20539, March 2022.

[18] G. Doras, Y. Teytaut, and A. Roebel, "A linear memory CTC-based algorithm for text-to-voice alignment of very long audio recordings," Appl Sci-Basel, vol. 13, no. 3, pp. 1854-1879, January 2023.

[19] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," Nat Mach Intell, vol. 4, no. 11, pp. 930-939, October 2022.

[20] P. Preethi, and H. R. Mamatha, "Region-based convolutional neural network for segmenting text in epigraphical images," AIA, vol. 1, no. 2, pp. 119-127, January 2023.