Research Article

Danhua Huang* and Shuaiqiu Xiang

Speech recognition and intelligent translation under multimodal human–computer interaction system

https://doi.org/10.1515/jisys-2023-0192
received October 11, 2023; accepted April 15, 2024

Abstract: The traditional translation robot is limited to the translation of single-mode text images and text videos, which has the problem of low translation accuracy. Therefore, speech recognition and intelligent translation in multimodal human–computer interaction (HCI) system are proposed. First, the network structure of speech recognition model in multi-channel HCI system is established, and the multi-head self-attention mechanism is constructed. Then, the artificial intelligence voice wake-up function is designed, and a multimodal machine translation model is constructed. On this basis, selective attention is added to obtain visual recognition of perceived text, and the decoder is used for multimodal gating fusion to realize the output of encoder translation results. Experimental results show that this method has high BLUE value and high translation accuracy.

Keywords: multimodal human–computer interaction, speech recognition, intelligent translation, attention mechanism

1 Introduction

Multimodal human–computer interaction (HCI) technology from the character stage, graphical interface stage to the touch screen stage, the touch screen brings convenience while there are also many inconveniences. With the normalization of the Internet and the maturity of artificial intelligence (AI) technology, people's needs are changing, and they start to explore an interaction mode that can free their hands while also having a more realistic interaction experience [1]. Multimodal HCI is the development trend of natural HCI, and multimodal information interaction is used as input and output to make full use of the characteristics of various modalities and improve the efficiency and realism of HCI. In recent years, the rapid development of machine learning has prompted more and more researchers to consider how to use the increasingly mature computer vision, speech recognition, speech synthesis, and other technologies in the field of natural language processing to realize the application of multimodal HCI [2]. The current continuous improvement in machine learning methods and information processing techniques, and the accompanying increase in computer computing power, has produced many mature algorithmic models in the field of deep learning, such as speech recognition, speech synthesis, computer graphics, and natural language processing. How to use these mature AI technologies to realize large-scale applications has gradually become a problem for researchers in these fields to explore and think about; based on this, "multimodality" is also increasingly talked about by researchers, and multimodal

*** Corresponding author: Danhua Huang,** School of English Studies, Zhejiang Yuexiu University, Shaoxing, 312000, China,
e-mail: 20051009@zyufl.edu.cn
Shuaiqiu Xiang: School of Software, Shenzhen Institute of Information Technology, Shenzhen, 518172, China,
e-mail: xiangsq@sziit.edu.cn

HCI mode will be a very important research direction in the field of AI in the future. The multimodal HCI model will be a very important research direction in the future AI field. Early speech recognition methods were simple template-matching methods, and the research direction was mainly for the recognition of isolated words and small vocabularies. After the 1980s, the research object of speech recognition shifted from isolated words to continuous words, and the research method gradually shifted to statistical template methods [3]. There were two very important modeling approaches at that time: acoustic models modeled using Gaussian mixture distribution–hidden Markov models (GMM–HMM) and language models based on N-element grammar modeling, and they dominated speech recognition for a long time. Although deep neural networks (DNN)–HMM-based speech recognition systems had achieved good recognition rates at that time, since neural networks actually work only on phoneme recognition, separate training is required for acoustic models, language models, and pronunciation models, and the aforementioned models have to be doubled if discriminative training is used, in addition to the fact that expertise is required to define pronunciation dictionaries and phoneme sets for specific languages, DNN The network structure and training process of traditional HMM-based speech recognition systems are very cumbersome [4]. Therefore, end-to-end structure-based modeling algorithms are a hot research topic for speech recognition tasks in recent years. The end-to-end structure-based speech recognition algorithm essentially solves the sequence-to-sequence mapping problem by mapping acoustic feature sequences to text sequences through a neural network structure, which does not require additional separate modeling of language models and pronunciation models and is much easier to train than traditional recognition frameworks. In this regard, this article proposes a speech recognition and intelligent translation method based on multimodal HCI system, aiming to improve the speech recognition effect and provide reliable reference for machine translation [5]. Speech recognition technology is an important means of achieving HCI. By converting speech signals into understandable text or commands, users can more conveniently interact with computers, and it can be applied in various fields. The research and development of speech recognition technology can improve the efficiency and comfort of people using computers and other intelligent devices. Intelligent translation refers to the process of using computer technology to convert text or spoken language from one language into another. The development of intelligent translation technology can effectively reduce communication barriers between different languages and promote cooperation and communication among humans in various fields such as economy, technology, and culture. A comparison between the proposed work and recent work on the subject of investigation is shown in Table 1.

2 Speech recognition technology under multimodal HCI system

2.1 Speech recognition network structure construction

The multimodal HCI system consists of three parts: input audio, intermediate link for cognitive and decision control, output audio and video, where the intermediate link for cognitive and decision control is the core algorithm part of the system, which combines technologies such as speech recognition, text dialogue, speech synthesis, and generating face expressions, and contains multimodal information such as speech, text, expressions, and faces. The input of the system is audio, and the output is video and audio, for which the speech recognition technology needs to be designed, and the specific implementation process is as follows [6].

The task of speech recognition converts the input audio into the corresponding text, which is the first link of the multimodal HCI system. In the task of speech recognition, a variety of speech data are usually used to evaluate the performance of the system. The number of voices used depends on the size of the data set. Generally speaking, in order to obtain reliable evaluation results, a certain amount of voice data is needed to test the system. Hundreds or thousands of speech samples can be used for testing to evaluate the accuracy of speech recognition and other indicators. Therefore, the accuracy of the model recognition will directly determine whether the output audio of the multimodal HCI system designed in this article can reasonably answer the input audio, which is an important key factor affecting the system performance [7]. The network structure of the model will be described below.

Table 1: Comparison between proposed work and recent work on the subject of investigation

Contrast index	Proposed work	HMM	Language model based on n-gram grammar modeling
Solve problems	Solve the problem that only single-mode text images and text videos can be translated, and the translation accuracy is not high.	Because neural network only works in phoneme recognition, it is necessary to train acoustic model, language model, and pronunciation model separately.	Solve the problem of using specialized knowledge to determine the pronunciation dictionary and phoneme combination of a specific language.
Methods	First, the network structure of speech recognition model in multi-channel HCI system is established, and the multi-head self-attention mechanism is constructed. Then, the voice wake-up function of AI is designed, and a multimodal machine translation model is established. Selective attention is added to obtain visual recognition of perceived text, and multi-modal gating fusion is carried out by decoder to realize the output of encoder translation results.	Acoustic model based on Gaussian mixture distribution modeling	Acoustic model based on Gaussian mixture distribution modeling
Innovation/ Deficiency	Modeling algorithm based on the end-to-end structure is a research hotspot in the field of speech recognition in recent years. The speech recognition algorithm based on the end-to-end structure maps the acoustic feature sequence to the text sequence through the neural network structure, which essentially solves the mapping problem from sequence to sequence, and it is easier to train than the traditional recognition framework without additional language model and pronunciation model.	Translation robots can only translate single-mode text images and text videos.	There is a problem of low translation accuracy.

The network structure of the speech recognition model in this article is borrowed from the transformer model, which has superior performance in processing sequential data, and several modifications are made on this basis: first, the frame rate is reduced and the sequence length of acoustic features is reduced to improve the computational efficiency; second, the decoder does not completely use the teacher forcing technique during training, but sets a threshold between the predicted characters and the real characters, which determines whether the decoder uses predicted characters or real characters for the next moment of input; third, the minimum granularity of the units modeling the text in the decoder is subwords, and the network structure of speech recognition in this article is shown in Figure 1 [8].

The model structure is a typical encoder-a-decoder-based network structure. In the multi-channel HCI system, the network structure of speech recognition model plays a vital role. This network structure needs strong feature extraction ability to effectively capture important information in speech signals; at the same time, it also needs to have the ability of context modeling to consider the influence of context information in speech signals on recognition results; in addition, the ability of multi-channel fusion is also important, so that the speech features from different channels can be integrated to improve the overall recognition performance and robustness; finally, in order to meet the requirements of real time, the network structure should have a

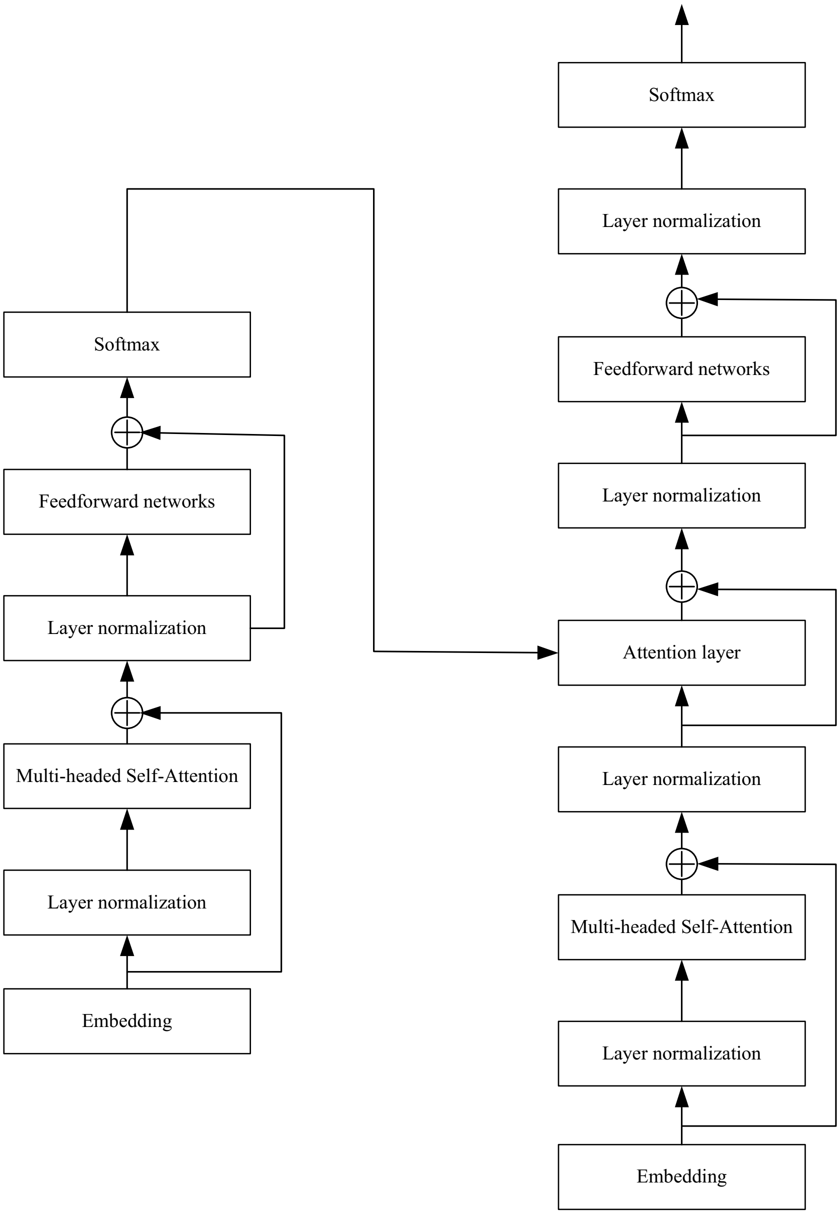

Figure 1: Network structure of speech recognition model under multimodal HCI system.

low inference delay and can quickly process and recognize the input speech signal. By optimizing and improving the network structure, the accuracy and robustness of speech recognition can be improved to meet the needs of multi-channel HCI system. The encoder structure is a stack of N encoder layers; each layer contains two sub-layers: a multi-headed self-attentive mechanism and a simple fully connected feedforward network; the two sub-layers are connected by a residual network and then normalized by a layer [9]. The decoder structure is a stack of M decoder layers; each layer contains three sub-layers: a multi-headed self-attentive mechanism, an encoder–decoder multi-headed self-attentive mechanism, and a simple fully connected feedforward network; and each sub-layer is also connected using residuals and then layer normalization operations, and several important modules of the network structure are described below [10].

(1) Location code

There is no convolution and loop operation in the network structure used in this article, and it is necessary to add some tagged position information to the input sequence, which can be the absolute or relative position information of the sequence. Therefore, the position encoding is added to the lowermost embedding layer of the codec structure, which has the same dimension as the input embedding d_{model}, and the two can be added together [11]. The position encoding can be obtained by learning or calculated by a fixed calculation using sine and cosine functions with different frequencies, as shown below:

$$P_{\text{E(pos},2i)} = \sin\left(\frac{p_{\text{os}}}{d_{\text{model}}}\right), \tag{1}$$

$$P_{\text{E(pos},2i+1)} = \cos\left(\frac{p_{\text{os}}}{d_{\text{model}}}\right), \tag{2}$$

where p_{os} denotes the position in the sequence, P_{E} denotes the vector corresponding to the position p_{os}, and d_{model} denotes the dimension of the vector. The trigonometric function calculates the position encoding with a generative law that can be expected to have some extrapolation, and according to the trigonometric principle, the elements in $P_{\text{E}}(p_{\text{os}} + n)$ can be represented by the elements in $P_{\text{E}}(p_{\text{os}})$, so that the position encoding of each input feature is associated with [12].

(2) Layer normalization

Layer normalization is to normalize the output of the activation function to accelerate the convergence of the model so that all hidden units in the same layer of the network share the mean and variance, independent of the batch size.

2.2 Multi-headed self-attentive mechanism construction

Attention mechanism is a kind of data processing, and some scholars have previously achieved better results by applying attention mechanism to image classification tasks, and then, attention mechanism has gained wide attention in the field of deep learning [13,14]. Suppose that the attention mechanism is used to learn the context vector for an input sequence *X* (sequence length *L* and dimension *d*), a query vector *Q* is needed (the generation of this vector is determined by the specific task), and the attention score is obtained by calculating the correlation between the query vector and each input in the input sequence *X* through the scoring function (in this article, the attention score is calculated using dot product scaling), and then, the attention score is mapped into a softmax function. The softmax function maps the attention score into a probability distribution between (0, 1), and finally, the input information is weighted and summed to obtain the context vector. The attention calculation method mentioned earlier is called soft attention, and the attention mechanisms in this article use the soft attention method, as shown in the following expressions [15]:

$$A(Q, K, V) = \text{softmax}Q\text{KT}dV, \tag{3}$$

where $A(Q, K, V)$ represents the attention mechanism that generates the query vector *Q* from the input sequence *X*, which is called the self-attentive mechanism. The self-attentive mechanism generally adopts the query one-key-one-value model, which maps the input sequence into *Q*, *K*, and *V*, respectively, by different weight vectors, where the attention score calculated by the scoring function is the correlation between *Q* and *K*. The attention mechanism that does multiple mappings of the query one-key-value vector is called the multi-headed attention mechanism, and the specific structure is shown in Figure 2 [16].

Multi-attention mechanism is a common mechanism in neural networks, which is widely used in natural language processing and computer vision. Each attention header gets a set of attention weights by inner product of query and key. These attention weights reflect the correlation between the query and each key, i.e., the degree to which the query pays attention to the key. The final output of multi-head attention mechanism can be obtained by weighted summation of attention weights and corresponding values. Each attention head can learn different correlation patterns, thus capturing the diversity within the data. Specifically, in the network structure, input features are mapped into query, key, and value spaces through

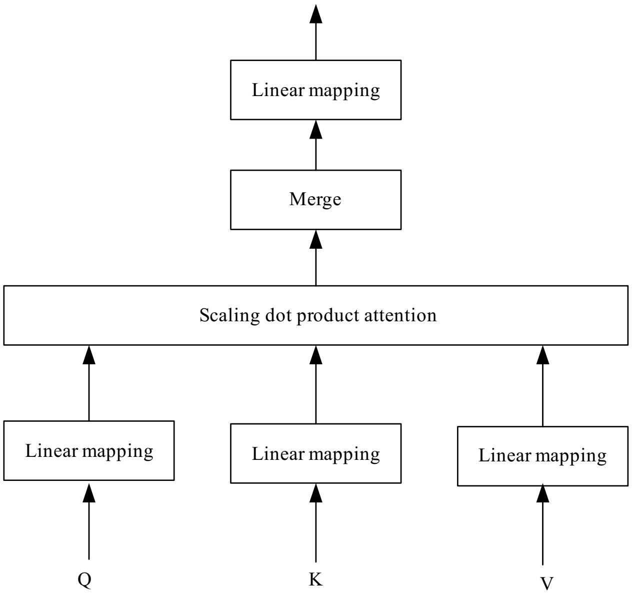

Figure 2: Network structure diagram of the multi-headed attention mechanism.

linear transformation, and these mapping matrices are learnable parameters of the network. The final output can be transmitted to the next layer of network for further processing or application.

The advantage of multi-head attention mechanism is that it can capture the information in the input data from different angles at the same time, and accurately weight and integrate the important information. By learning the weights between multiple attention heads, the network can flexibly adjust different attention modes and carefully interpret and deal with specific tasks.

The attention mechanism is characterized by focusing limited attention on important information, acquiring critical information and it is data-driven, with fewer training parameters than convolutional neural networks (CNNs) and recurrent neural networks while having the ability of CNN parallel processing. The self-attentive mechanism focuses too much on its own location while encoding the features at the current location, thus ignoring other information [17]. The multi-headed self-attentive mechanism is similar to the role of using multiple filters simultaneously in CNN, which can solve the problem that the self-attentive mechanism focuses too much on its own location features and can capture richer features, and the output of the attention layer collects the encoded features of all subspaces, making the model more expressive.

There are three forms of keys for the multi-headed self-attentive mechanism in the network structure of the model in this article. The query and key value in the self-attentive module of the encoder layer are derived from the output of the preceding layer. Similarly, in the decoder layer, the query and key value for the first self-attentive module are also sourced from the output of the preceding layer. However, due to the parallel nature of the attention mechanism, a mask is applied to the input to prevent the decoder from accessing irrelevant information during decoding. In the second self-attentive module of the decoder layer, the query is still obtained from the output of the previous layer, but the key values are now derived from the context vector features encoded by the encoder layer specifically for acoustic features [18].

Speech recognition is essentially a multi-classification problem, and the model uses a cross-entropy loss function, which is a loss function commonly used in multi-classification tasks. And the label smoothing strategy is used to reduce the confidence level of correctly classified samples to improve the adaptive ability of the model, with the following formulas [19]:

$$H(q', p) = \sum_{K=1}^{K} \log P_{\mathrm{E}}(p_{\mathrm{os}}), \tag{4}$$

$$q'(k) = (1 - \varepsilon)K, \tag{5}$$

where H represents the cross-entropy, ε represents the smoothing parameter, K represents the number of categories of token, (q', p) represents the model adaptive value, and $q'(k)$ represents the loss function. The aforementioned steps are organized to obtain the speech recognition process, as shown in Figure 3.

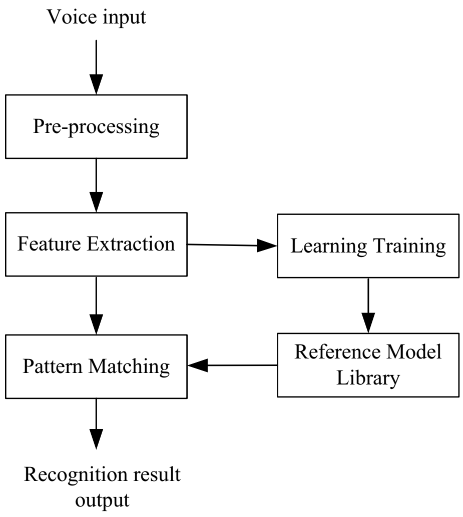

Figure 3: Speech recognition process.

As can be seen from Figure 3, based on the following four working principles, the speech recognition system is able to complete the recognition of speech. First, the speech recognition library and detection terminal are targeted and selected, and combined with anti-aliasing band-pass filtering technology, which can effectively eliminate the noise of individual speech differences, sampling equipment, and sampling environment; then, the speech acoustic parameters such as average energy and vibration peak and average over-zero rate are extracted, which can quickly and accurately reflect the phonetic characteristics of speech; then, the speech pattern database is established, and the link of language repetition training mainly is to let the speaker repeat the pronunciation, directly delete the redundant speech information one by one from the original speech samples, only keep part of the key speech data, and scientifically classify the key speech data according to the relevant scheme; finally, accurately output the relevant semantics of speech according to the speech similarity [20]. Through the aforementioned steps, the speech recognition network structure and the multi-headed attention mechanism network structure can be constructed and the loss function can be obtained to complete the design of speech recognition technology under the multimodal HCI system.

3 Intelligent translation technology under multimodal HCI system

3.1 AI voice wake-up function design

In order to implement the voice wake-up function, a process needs to be set up in the background to regularly monitor the environment around the device in real time and to detect the presence of the keyword required by the device in the signal, as shown in Table 2.

As can be seen from Table 3, after completing the detection of the keyword signal, the detection callback or static detection is performed and the action is recorded. In the static detection of the recorded action, the process returns the recorded pulse code modulation (PCM) audio data to the main process and then continues to the next step [21]. Based on this, the data slicing process is performed on the captured pulse code modulated

Table 2: AI voice wake-up process

Steps	Process and instructions
01	Detection of process signals, hibernation for a specified length of time
02	Get PCM data and detect keywords
03	Determine whether to callback or not
04	Static detection of recording
05	Determine if the audio strength is less than the threshold value
06	Perform recording callback and return PCM audio data

Table 3: Machine translation process for HCI function

Steps	Process and instructions
01	Read the machine translation model and select the pre-stored corresponding domain according to the user of the translation domain
02	Splitting the text into sentences
03	Search for matching translations on the search network

signal using the prime multistage dynamic labeling model. The machine translation steps of the HCI function are shown in Table 3.

Since the number of states in the W_n set increases exponentially with n, it would take a lot of time if the set is not pruned, so pruning is required. The pruning process is shown in the following:

For the identified source language sentence $R_1, R_2, \dots, R_n$, there exists a phrase model (f, g, s, d), and the four elements in this model represent the phrase base, grammar model, distortion restrictions, and distortion parameters.

Assume that W_n translates n words into m sets, and if an element of an m statement in W_2 consists of two words, it means that only two words are translated into the statement. For each state of W_i, there is a translation state, and all possible states are added to the corresponding set, and the highest rated of them is finally returned [22].

Let β be the search parameter and p be the transfer parameter, and the resulting grammar model expression is shown in the following:

$$g = n_{\text{ext}}(p, \beta). \tag{6}$$

After determining the search parameters, the pruning purpose is achieved after removing all the parameters in the set that do not satisfy equation (2):

$$a(g) \geq a(p) - \beta. \tag{7}$$

When translating a source language sentence, the translation options of the source language sentence are first read, and then, the translation hypotheses are expanded from the small container to the large container. At each transition stage, if the difference between a score and the highest score in the container is greater than a threshold, the state decreases; if the state remains the same, all available translation options are expanded; if the old and new hypotheses are the same, the score is increased and the best translation result is to find the highest scoring translated statement in the largest container [23].

Voice wake-up function is one of the basic functions of multi-channel HCI system, which activates the system by recognizing specific voice commands and starts the subsequent interaction process. Perceptual visual representation of text is to transform text information into a representation form that computers can understand and process in a multimodal environment. The design and implementation of these two functions are helpful to improve the user experience and interaction effect of multi-channel HCI system.

3.2 Visual representation acquisition of perceptual text

There is a close relationship between designing AI voice wake-up function and obtaining visual representation of perceived text. As the entrance of the system, the voice wake-up function starts the system by recognizing specific voice commands; the visual representation of perceptual text provides a treatable representation form for text input in multimodal environment. The integration and collaborative design of these two functions is helpful to improve the efficiency and user experience of multi-channel HCI system and promote the development and application of HCI.

The main method to perceive the visual representation of the text is to collate the visual and textual information, align them, and find the similarity score between them to obtain the attention matrix, after which the visual representation weights are reassigned according to this matrix [24]. The specific steps are as follows:

(1) Transform the input text into word embeddings according to the word embedding table;
(2) Spatial visual features of images are obtained using the image pre-training model VCC19;
(3) Determining the visual feature dimension and the feature dimension of the text representation;
(4) If the dimensions of the two features are different, then a linear fully connected network is used for dimensional transformation so that the two dimensions are the same;
(5) Dot product processing and normalization of textual and visual features to obtain the attention matrix;
(6) Based on the attention matrix and spatial visual features from Steps (3) and (2), the visual representation of the perceived text is obtained;
(7) To encode and process textual and visual representations of perceptual text, two independent Transformer encoders are utilized, each responsible for one type of representation;
(8) Two text states are obtained, which are the hidden state and the visual hidden state;
(9) Reassign and fuse the hidden state weights.

3.3 Intelligent translation model construction and optimization

In the intelligent translation model in multimodal HCI systems, visual information can only play an auxiliary role, and if too much visual information is added to the model, it may make the redundant information in the model increase, thus making the model translation accuracy decrease. To solve this problem, this article proposes an intelligent translation model in multimodal HCI systems based on increasing the selection attention to replace the traditional feature summation by the important part of image features, which effectively reduces the interference of redundant information.

Assuming that $E_{\text{text}}[e_1, e_2, \ldots, e_n]$ and $F_{\text{img}}[f_1, f_2, \ldots, f_n]$ represent the textual representation and the spatial visual representation, respectively, the attention matrix can be obtained by finding the similarity between them, and the specific matrix expression is shown in the following:

$$M_{\text{atrix att}} = E_{\text{text}} \times F_{\text{img}}^{T}, \tag{8}$$

where $M_{\text{atrix att}}$ represents the attention matrix, where each element represents the similarity of a word to a visual feature. The dot product of the textual representation e_i and the visual representation f_i is calculated to obtain $M_{\text{atrix att}}$. Each element is selected by Gumbel−sigmoid to obtain the region of the visual feature that is more relevant to the current one. Once the attention matrix is obtained, it can be exponentially normalized to a probability distribution where the similarity score does not sum to 1 and then multiplied with each element of the sampling result matrix $M_{\text{atrix sel}}$ to obtain the multimodal attention matrix $M_{\text{atrix mm}}$, which is calculated as shown in the following:

$$M_{\text{atrix mm}} = M_{\text{atrix sel}} \times \text{soft max}(M_{\text{atrix att}}). \tag{9}$$

To further enhance the modeling capability of selective attention, a multi-headed attention mechanism can also be added to this attention, as shown in the following expression:

$$M_{\mathrm{HGA}}(Q,K,V)=C_{\mathrm{oncat}}(G_{A1},G_{A2},\ldots,G_{Ah})W^{0},\tag{10}$$

where M_{HGA} and h denote the multi-head selective attention and the number of this attention head, respectively; $G_{A1}, G_{A2}, \ldots, G_{Ah}$ both denote the sub-attention module, which is processed by splicing and linear transformation to obtain the final output [25]. Based on the improvement of the aforementioned attention matrix, the encoder of the multimodal translation model based on selective attention is obtained, as shown in Figure 4.

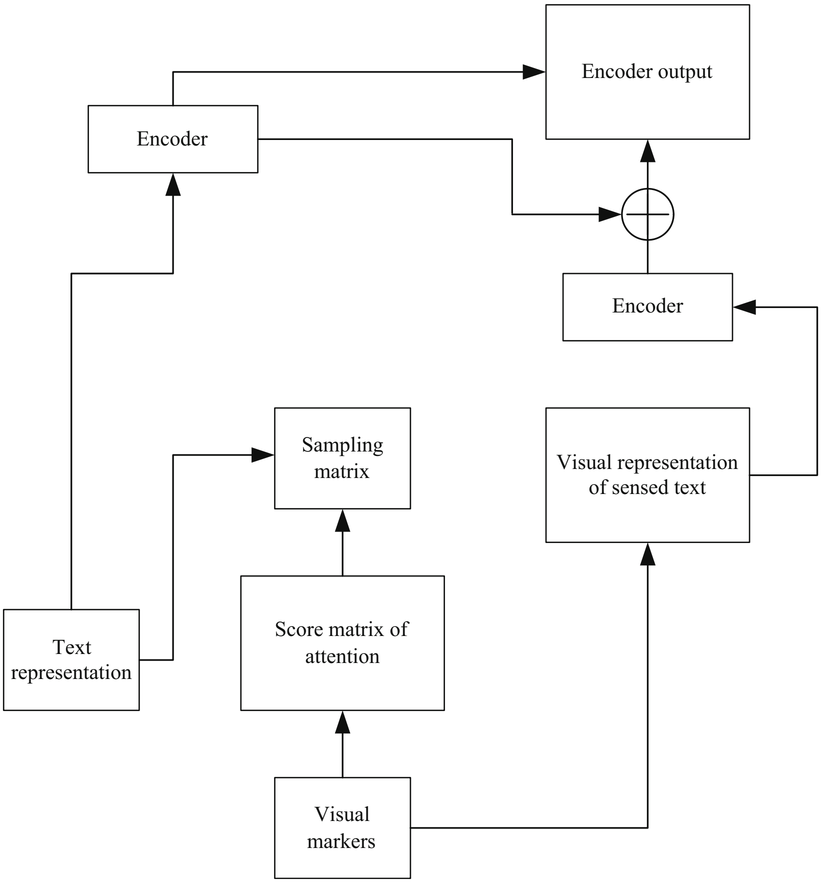

Figure 4: Encoder structure of multimodal translation model based on selective attention.

Through the aforementioned steps, the optimization and design of multimodal machine translation model can be completed, and the encoder can be obtained to realize the machine translation function. By combining the machine translation-related contents in this section with the speech recognition technology mentioned earlier, the design of speech recognition and intelligent translation technology under multimodal HCI system is completed.

4 Testing and analysis

4.1 Test preparation

In order to prove that the speech recognition and intelligent translation techniques under the multimodal HCI system proposed in this article are better than the conventional speech recognition and intelligent translation techniques in terms of translation effect, after the theoretical part of the design is completed, an experimental session is constructed to test the evaluation effect of the method in this article. In order to improve the reliability of the experimental results, in addition to the method of this article, two conventional methods are selected as the experimental control group, which are the speech recognition and intelligent translation technology based on fuzzy algorithm and the speech recognition and intelligent translation technology based on support vector machine (SVM).

To obtain better experimental results, the experimental data were chosen to be obtained from the Multi30K dataset. This dataset contains 30,000 text-image data pairs with a total of 30 000 entries. Among them, the training set, test set, and validation set are 25,000, 4,200, and 800 items, respectively. The text-a-video data was derived from the VATEX dataset with a total of 42,000 entries. The data percentages of training set, test set, and validation set are 60, 20, and 20%, respectively, and the classification results of the test set are shown in Table 4.

Table 4: Test set classification

Types of sentences	Number of words	Number of sentences
Simple sentences	$N < 8$	458
Normal sentences	$8 \le N < 18$	368
Complex sentences	$N > 18$	178

To ensure the experimental effect, different parameters will be set according to the difference of datasets for the text-image and text-video machine translation tasks. The numbers of layers of encoder and decoder are set to 4 and 5 for text-images and text-video machine translation tasks, respectively; the word vector dimension is set to 128 and 256, respectively; the self-attention and multimodal attention between encoder and decoder are set to 4, and the attention is set to 4 and 8, respectively; the initial learning rate is set to 0.001 Optimizer is selected as Adam, Batch Size.

4.2 Analysis of test results

The comparison criterion chosen for this experiment is the translation performance of the method, and the specific measure is the BLUE index, which is mainly used to evaluate the accuracy of translation, and the basic principle is to find out the degree of N-gram contribution between the generated translation and the reference translation, and the specific calculation formula is shown in the following formula:

$$B_{\mathrm{LUE}} = \frac{\sum_i \sum_k \min h_k(c_i)}{B_{\mathrm{P}}}, \tag{11}$$

where B_{P} represents the penalty factor, i and k represent the kth N-gram fragment and the ith sentence, respectively, h_k represents the current utterance, and c_i represents the standard reference translation.

Analyzing the data in Table 5, we can see that compared with the other two conventional models, the BLUE value of the machine translation model under multimodal HCI system proposed in this article is higher, which proves that the model proposed in this article has higher translation accuracy and better model performance.

Table 5: Comparison results of BLUE values for different translation models

Number of tests	Fuzzy algorithms	SVMs	Multimodal HCI
01	37.5	39.2	42.6
02	36.8	40.1	43.7
03	38.2	38.5	44.6
04	36.4	38.4	43.2
05	35.3	39.7	46.1

Choose accuracy as the evaluation index, the higher the accuracy value, the better the representation effect, and its calculation formula is as follows:

$$F_{\mathrm{PN}} = \frac{F_{\mathrm{P}}}{F_{\mathrm{P}} + F_{\mathrm{N}}} \times 100\%, \tag{12}$$

where F_{PN} represents the accuracy of speech recognition and intelligent translation, F_{P} represents the number of correctly predicted samples, and F_{N} represents the total number of samples. Using formula (12), the sentences in the experimental dataset are translated by fuzzy algorithm, SVM, and multi-channel HCI, respectively, and the accuracy of translation results is shown in Figure 5.

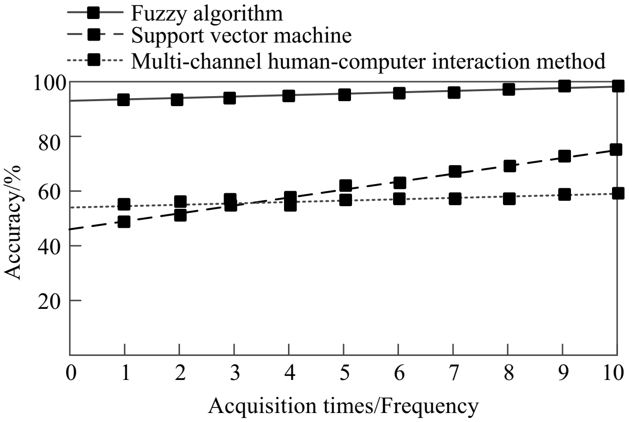

Figure 5: Comparison results of translation accuracy of three methods.

As can be seen from Figure 5, compared with fuzzy algorithm and SVM method, the accuracy of multi-channel HCI method is higher, with the highest value of 99%, which indicates that the research method has better speech recognition and intelligent translation effect.

Use confusion matrix to verify the effect of speech recognition and intelligent translation in multimodal HCI system. Confusion matrix, also known as error matrix, is a standard format for expressing accuracy evaluation, usually expressed in the form of matrix with m rows and n columns, specifically as follows:

$$P_{\mathrm{E}} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nm} \end{bmatrix}. \tag{13}$$

Specifically, each column of the confusion matrix represents the prediction category, and the total number of each column represents the number of data predicted as this category; each row represents the real category of data, and the total number of data in each row represents the number of data instances in this category. Through the confusion matrix, we can intuitively understand the performance of speech recognition and intelligent translation. Under the application of different methods, three test samples are selected, and the samples are identified by mistake and missed speech, and their numbers are counted, and then, the errors are calculated. Finally, the results are compared, as shown in Table 6.

As can be seen from Table 6, the error of multimodal HCI method on test sample 1 is 1.2%, and the average error is 1.5%. The error of fuzzy algorithm on test sample 1 is 6.9%, and the average error is 7.3%. The error of SVM method on test sample 1 is 7.5%, and the average error is 7.7%. In the test sample 1, the error of multimodal HCI method is the lowest, while the error of SVM method is the highest. In terms of average value, multimodal HCI method is still the most accurate, and the average error of fuzzy algorithm and SVM method is high; through the aforementioned analysis, we can know that the multimodal HCI method performs best in these test samples, with higher accuracy and stability; The performance of fuzzy algorithm on test samples is

Table 6: Error comparison results of different methods

Different methods	Test sample	Error (%)	Average value (%)
Multi-modal HCI method	Test sample 1	1.2	1.5
	Test sample 2	2.1	
	Test sample 3	1.2	
Fuzzy algorithm	Test sample 1	6.9	7.3
	Test sample 2	7.3	
	Test sample 3	7.7	
SVM method	Test sample 1	7.5	7.7
	Test sample 2	6.7	
	Test sample 3	8.9	

inferior to that of multimodal HCI method, but better than that of SVM method. SVM performs worst among these test samples and has a high error rate. This shows that the research method is more accurate and comprehensive, and the recognition effect is better.

To sum up, the speech recognition and intelligent translation methods in this multi-modal HCI system have achieved better performance. By evaluating the BLUE value, the similarity between the translation results and the manual reference translation can be quantified. It has higher BLUE value and accuracy, which shows that it has better translation accuracy and model performance, and provides a more reliable language processing tool for cross-cultural communication and cooperation.

5 Conclusion

In the multi-modal HCI system, speech recognition and intelligent translation are two key technologies, and their combination can provide strong support for cross-cultural communication and cooperation. This article discusses the research of speech recognition and intelligent translation in multimodal HCI system and draws the following conclusions: the model proposed in this article has higher translation accuracy and better model performance. It shows that the model in this article has a higher BLUE value than the traditional translation model, which verifies its advantages in translation accuracy and model performance.

However, there are still some challenges and room for improvement. First of all, speech recognition technology is still affected by background noise and speaker changes, and further research is needed to improve robustness and reliability. Second, in the field of intelligent translation, although the model has achieved good results in a multimodal environment, it still needs to solve some specific problems between languages, such as structural differences and cultural differences.

Funding information: Shaoxing Higher Education Teaching Reform Research Program (SXSJG202402).

Author contributions: All authors participated in some part of the work for this article. In the investigation, DH proposed the idea and conceived the design, carried out the simulation, and wrote the original draft; SX analyzed and discussed the results and reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

Data availability statement: Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

References

[1] Badrinath S, Balakrishnan H. Automatic speech recognition for air traffic control communications. Transp Res Rec. 2022;2676(1):798–810.

[2] Zeng T, Yang X, Wan Y, Mao Y, Liu Z. Effectiveness assessment of improvement measures in physical protection system monitoring center. Kerntechnik. 2021;86(1):33–8.

[3] Song T, Zhao H, Liu Z, Liu H, Hu Y, Sun D. Intelligent human hand gesture recognition by local-global fusing quality-aware features. Future Gener Comput Syst. 2021;115(7043):298–303.

[4] Roda-Sanchez L, Olivares T, Garrido-Hidalgo C, de la Vara JL, Fernández-Caballero A. Human-robot interaction in industry 4.0 based on an internet of things real-time gesture control system. Integr Comput Eng. 2021;28(2):1–17.

[5] Porcheron M, Fischer JE, Reeves S. Pulling back the curtain on the wizards of oz. Proc ACM Human-Comput Interact. 2021;4(CSCW3):1–22.

[6] Zhang Y. Interactive intelligent teaching and automatic composition scoring system based on linear regression machine learning algorithm. J Intell Fuzzy Syst. 2021;40(2):2069–81.

[7] Zhang D. Intelligent recognition of dance training movements based on machine learning and embedded system. J Intell Fuzzy Syst. 2021;1:1–13.

[8] Jasim M, Khaloo P, Wadhwa S, Zhang AX, Sarvghad A, Mahyar N. Community click: capturing and reporting community feedback from town halls to improve inclusivity. Proc ACM Human-Comput Interact. 2021;4(CSCW3):1–32.

[9] Yang B, Xia X, Wang S, Ye L. Development of flight simulation system based on leap motion controller. Proc Comput Sci. 2021;183(2):794–800.

[10] Jing W, Tao H, Rahman MA, Kabir MN, Yafeng L, Zhang R, et al. RERS-CC: Robotic facial recognition system for improving the accuracy of human face identification using HRI. Work. 2021;68(7):1–12.

[11] Yu J, Ji H, Song Q, Zhou L. Design and implementation of business access control in new generation power grid dispatching and control system. Proc Comput Sci. 2021;183(22):761–7.

[12] Carlos Alberto PJ, Sonia Karina PJ, Francisca Irene SA, Adrielly Nahomee RÁ. Waste reduction in printing process by implementing a video inspection system as a human machine interface. Proc Comput Sci. 2021;180:79–85.

[13] Sha Y, Feng T, Xiong X, Yang T. Designing online psychological consultation expert system using human-computer interaction. Mob Inf Syst. 2021;2021(1):1–12.

[14] Su KW, Chiu PC, Lin TH. Establishing a blockchain online travel agency with a human-computer interaction perspective. J Hosp Tour Technol. 2022;13(3):559–72.

[15] Mitchell EG, Maimone R, Cassells A, Tobin JN, Davidson P, Smaldone AM, et al. Automated vs. human health coaching: exploring participant and practitioner experiences. Proc ACM Human-Comput Interact. 2021;5(CSCW1):1–37.

[16] Zhang H. Voice keyword retrieval method using attention mechanism and multimodal information fusion. Sci Program. 2021;2021(8):1–11.

[17] Yuan Q, Wang R, Pan Z, Xu S, Gao J, Luo T. A survey on human-computer interaction in spatial augmented reality. J Comput Des Comput Graph. 2021;33(3):321–32.

[18] Sreekanth NS, Narayanan NK. Multimodal human computer interaction with context dependent input modality suggestion and dynamic input ambiguity resolution. Int J Eng Trends Technol. 2021;69(5):152–65.

[19] Evers K, Chen S. Effects of automatic speech recognition software on pronunciation for adults with different learning styles. J Educ Comput Res. 2021;59(4):669–85.

[20] Alhumsi MH, Belhassen S. The challenges of developing a living Arabic phonetic dictionary for speech recognition system: A literature review. Adv J Soc Sci. 2021;8(1):164–70.

[21] Kempfle JS, Panda A, Hottin M, Vinik K, Kozin ED, Ito CJ, et al. Effect of powered air-purifying respirators on speech recognition among health care workers. Otolaryngol-Head Neck Surg. 2021;164(1):87–90.

[22] Ji YJ, Bahng J, Lee JH. Efficacy of a closed-set auditory training protocol on speech recognition of adult hearing aid users. Korean J Otorhinolaryngol – Head Neck Surg. 2021;64(2):70–6.

[23] Folkeard P, Eeckhoutte MV, Levy S, Dundas D, Abbasalipour P, Glista D, et al. Detection, speech recognition, loudness, and preference outcomes with a direct drive hearing aid: Effects of band width. Trends Hearing. 2021;25(4):8–13.

[24] Sun Z, Tang P. Automatic communication error detection using speech recognition and linguistic analysis for proactive control of loss of separation. Transp Res Rec. 2021;2675(5):1–12.

[25] Kumar LA, Renuka DK, Rose SL, Shunmuga priya MC, Wartana IM. Deep learning based assistive technology on audio visual speech recognition for hearing impaired. Int J Cognit Comput Eng. 2022;3:24–30.