Research on the Construction of English Translation Model for Speech Recognition Based on Multiple Information Sources and Lexical Assistance

Limei Ma
School of Foreign Languages, Ningxia Medical University, Yinchuan, Ningxia 750004, China
E-mail: mlmeima@outlook.com

Keywords: machine translation, speech recognition, english text, lexical aids, multiple information sources

Received: December 26, 2023

Speech recognition and machine translation technologies have been continuously evolving and improving to meet the growing demand for multilingual communication. Meanwhile, diverse information source inputs make traditional speech recognition and translation systems even more challenging. In view of this, the study takes multiple information sources as an entry point to improve the existing English speech recognition and translation systems. Aiming at the process of speech comprehension and text generation in speech recognition, the study utilizes the lexical labels in the lexical auxiliary function to make functional adjustments to the input of the speech source and the output of the translated text. Secondly, after incorporating multiple information source conditions, the study proposes a novel English speech recognition translation model. The experimental results show that compared with other speech recognition models, the new model proposed by the study recognizes up to 87% average accuracy. In addition, the research proposed model can further improve the BLEU scores of the translations and strengthen the textual information with a maximum score of 0.81. In summary, lexical assistance under multiple information sources can significantly improve the performance of English translation models for speech recognition, providing new ideas and methods for the development of speech recognition and translation technology.

Povzetek: Predstavljen je model za prevajanje angleškega govora z leksikalno pomočjo iz več virov, ki izboljšuje natančnost prepoznavanja in BLEU ocene prevodov.

1 Introduction

English translation model for speech recognition is an important research area in the field of natural language processing [1]. In recent years, with the rapid development of speech technology and machine translation technology, researchers have been exploring how to construct more accurate and effective English translation models for speech recognition [2]. Some foreign researchers have widely adopted deep learning techniques, especially recurrent neural networks, convolutional neural networks and attention mechanisms, to construct speech recognition and translation models [3]. For example, Vlter et al. proposed a neural perception method incorporating visual presentation in order to improve the speech recognition perception of cochlear implants. This method has a greater improvement in speech perception rate than the traditional method, and the speed of speech word acquisition and retrieval is significantly improved [4]. In addition, Bibin et al. proposed a low-complexity extraction system of Mel frequency cepstrum coefficients for speech recognition translation in order to improve the performance of existing speech recognition translation models. The experimental results show that the recognition and translation speed of this new recognition and translation system is greatly improved and the normalized energy consumption is reduced by 5.4% [5]. Meanwhile, domestic researchers have also focused on the research and development of multi-language support technology, so that the speech recognition English translation model can be applied to different languages, not only limited to English [6]. For example, Lin et al. in order to improve the accuracy of multilingual recognition and translation of the existing traffic control system, after fusing recurrent neural networks and hybrid speech feature embedding blocks. The research team proposed a new multilingual automatic recognition translation model. The experimental results show that the Chinese and English character recognition error rates of this new model are reduced by 4.8% and 7.2% respectively [7]. These deep learning methods have made significant progress in improving model performance, and this multilingual support is expected to facilitate the development of cross-lingual communication [8]. However, despite the impressive results of many studies, there is still the problem of degradation of recognition and translation performance due to lexical complexity and multiple sources of linguistic input. With the development of technology, lexical aids have gradually come into the public's view [9]. Lexical assistance is a technique commonly used in natural language processing to enhance the performance of named entity recognition and other text processing tasks [10]. The main goal of lexical assistance is to improve the accuracy of recognition by considering the lexical properties of each word, and many existing scholars have also studied lexical assistance to varying degrees. For example, in order to improve the effectiveness of natural language processing techniques on text and speech data, Shao et al. proposed a sequence labeling framework that incorporates lexical assistance features. The experimental results show that the method's semantically labeled sequence prediction is better than other methods of the same type [11]. In addition, Samuel et al. proposed a semi-supervised lexicality-assisted model combined with Word2Vec word embedding in order to improve the novel acquisition of perceptually labeled corpora. Experimental results show that this new model has a high word embedding feature extraction rate and obtains superior competitiveness compared to supervised algorithms [12]. Combining the above literature, the study drew a summary table of relevant writings as shown in Table 1.

Table 1: Summary table of related works.

Author	**Literatures**	**Method**	**Index**	**Key findings**
Vlter C	Non-Auditory Functions in Low-Performing Adult Cochlear Implant Users	Integration of Neuroperception in Visual Presentation	Speech perception rate/retrieval speed	The method's speech perception rate and speech word retrieval speed are significantly improved.
Bibin S P S	A low latency modular-level deeply integrated MFCC feature extraction architecture for speech recognition - ScienceDirect	Mel-frequency cepstrum coefficient extraction	Translation speed, normalized energy consumption	The system's speed of recognition and translation was dramatically increased and the normalized energy consumption was reduced by 5.4%.
Lin Y	Towards multilingual end-to-end speech recognition for air traffic control	Recursive Neural Network Optimized Speech Feature Embedding Module	Bilingual character recognition error rate	The model's Chinese and English character recognition error rates were reduced by 4.8% and 7.2%, respectively
Shao Y	Self-attention-based conditional random fields latent variables model for sequence labeling - ScienceDirect	Lexical aids	Semantic Tag Sequence Prediction Rate	This method has better semantic tag sequence prediction than other methods of the same type
Samuel S	Deep analysis of word sense disambiguation via semi-supervised learning and neural word representations	Word2Vec Word Embedding	Word embedding feature extraction rate	Higher word embedding feature extraction rate for the new model

To summarize, there are few applications related to speech recognition translation and lexical assistance, while experts and scholars at home and abroad have conducted in-depth research on both speech recognition translation and lexical assistance respectively in recent years, and have also achieved certain research results. However, due to the diversity of language nature and semantics, relying on a single source of information as input often results in translation errors. In view of this, the study innovatively integrates lexical assistance into the construction of the recognition translation model, and at the same time, based on multiple sources of information, constructs a speech recognition English translation model suitable for complex environments, aiming to improve the comprehensive performance of the speech recognition translation model under multiple sources of information. This research is divided into four parts, the first part analyzes and summarizes the research of others, the second part introduces how the lexical assistance model and the new speech recognition translation model are constructed, the third part tests the performance of the new speech recognition translation model, and the last part summarizes the article.

2 Construction of english translation model for speech recognition based on multiple information sources and lexical assistance

In order to construct an English translation model for speech recognition under multiple information sources, the study firstly introduces the existing translation technology and introduces two kinds of auxiliary models based on this technology respectively. The first section introduces the models of lexically-assisted comprehension and lexically-assisted generation as constructed, respectively, and the second section combines these two models and incorporates the consideration of multiple information sources to propose a model of English translation for speech recognition under multiple information sources.

2.1 Lexically assisted comprehension and lexically assisted generative modeling

Speech recognition translation system is a language representation technology that transforms speech information recognition into textual modality, which is more complex than machine translation that deals with textual information modality alone [13]. Speech recognition translation modality firstly utilizes automatic speech recognition technology to convert speech signals into the same transcribed text and secondly utilizes machine translation to translate the transcribed text into the text of the target language. However, the encoder's insufficient ability to characterize the input utterance often results in translation errors [14]. With the development of deep learning technology, lexically-assisted understanding has gradually come into the public's view, and the method has gained good performance in the Transformer architecture with an attention mechanism. Schematic diagram of traditional Transformer architecture. As can be seen from Figure 1, the framework mainly consists of an encoder and a decoder. Among them, the encoder consists of two sublayers stacked together, namely the multi-head self-attention sublayer and the fully connected network sublayer. The decoder causes the elements to be unassociated after masking operation through the multi-head self-attention sublayer, and then outputs them after connecting the encoder through similar sublayers. The computational procedure of the attention mechanism is shown in Eq. (1).

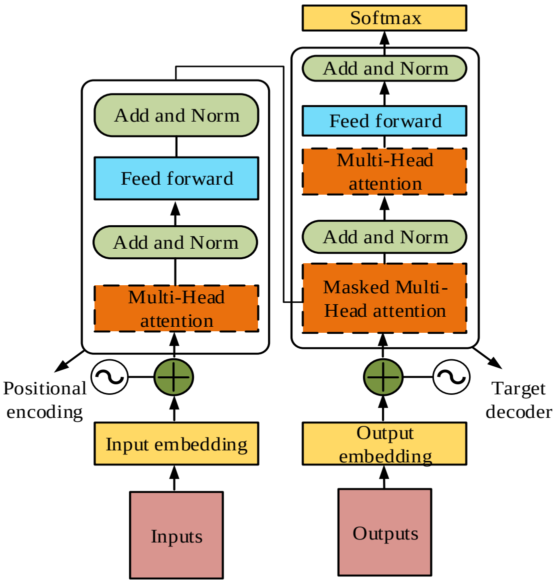


Figure 1: Schematic diagram of traditional transformer architecture.

$$s(q,k_i)=k_i^T q \quad (1)$$

In Eq. (1), k denotes the key variable, s denotes the scoring function, q denotes any target element, and i denotes the source text composed of the i vector. For the results of different scoring functions, the numerical normalization conversion is performed by softmax function, and the calculation process is shown in Eq. (2).

$$\alpha = soft\max(s(q,k_i)) \quad (2)$$

In Eq. (2), α denotes the similarity calculation coefficient. After this coefficient is calculated, the final output vector of the attention mechanism is obtained by weighted summation with the value variables. The formula for this process is shown in equation (3).

$$A = Attention(s(q,k_i)) = \sum_{i=1}^{N} \alpha_i v_i \quad (3)$$

In Eq. (3), N denotes the text sequence matrix composed of N vectors, α_i denotes the value variable weight coefficients, and v_i denotes the source text vector coefficients. After a long time of use, it is found that the performance of traditional Transformer architecture decreases when dealing with long sequences of speech text information, and at the same time, it requires a large amount of labeled data for training, which rather increases the burden of the system [15]. Therefore, the research continues to explore deeply with the Transformer architecture that incorporates lexical aids, and the schematic diagram of the Transformer model that incorporates lexical aids for comprehension is shown in Figure 2.

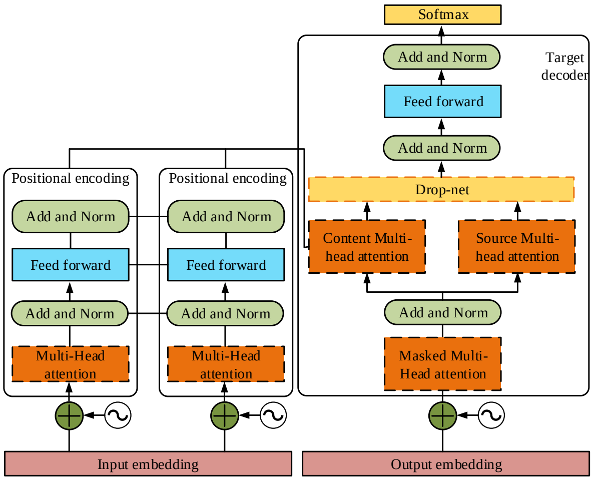


Figure 2: Transformer model for part of speech assisted understanding.

As can be seen from Figure 2, based on the original Transformer framework, lexically assisted understanding restructures the entire encoder and decoder. The encoder is decomposed into source word encoding and real word encoding, which are composed of multiple identical layers stacked on top of each other, and each layer is dominated by a Multi-head Self-Attention Mechanism (MATT). The decoder adds an Attention Module and a Residual Connection Module to better explore the auxiliary role of real words on the source sentence representation. The MATT formula for the encoder is shown in equation (4).

$$MATT(Q_x,K_x,V_x)=Concat(head_1,\cdots,head_H)$$
(4)

In Eq. (4), Q_x , K_x , V_x denote the three input matrices of the source word encoder, x denotes the source sequence labeling, $head$ denotes the attention head, and H denotes the number of M-attention heads, respectively. Where the expression of the attention mechanism of the input matrices is shown in Eq. (5).

$$S(Q_x,K_x,V_x)=soft\max(\frac{Q_xK_x^T}{\sqrt{d_{model}}}) \quad (5)$$

In Eq. (5), d_{model} denotes the dimension of the model. The formula for processing operations for each attention module by residual linkage and layer normalization is shown in Eq. (6).

$$H_x^l=LN(MATT_e^l(Q_x^{l-1},K_x^{l-1},V_x^{l-1})) \quad (6)$$

In Eq. (6), H_x^l denotes the source sentence representation of the l th layer. The feature matrix $(Q_x^{l-1},K_x^{l-1},V_x^{l-1})$ is linearized as shown in Eq. (7).

$$H_a^l=LN(FNN_e^l(H_x^l)+H_x^l)) \quad (7)$$

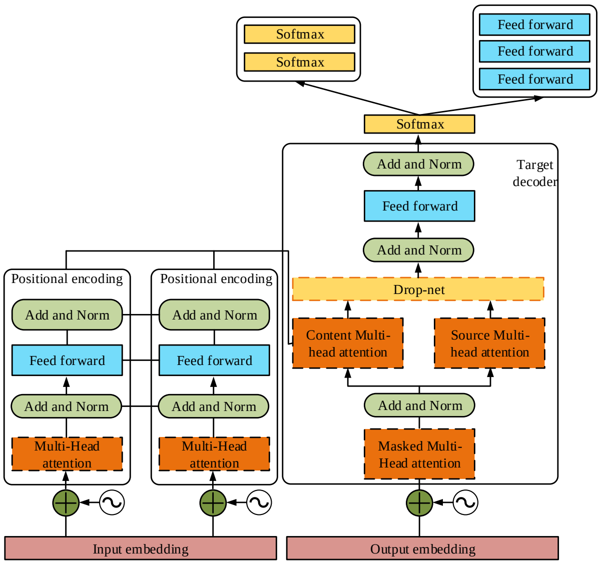


Figure 3: A transformer model for part of speech assisted generation.

In Eq. (7), FNN denotes the fully connected feed-forward network, and H_a^l denotes the source sentence sequence at the l layer. After real words of enhanced source sentence representation are achieved, lexical tagging can be realized to extract real words, but to realize the real word features incorporated into the translation system, lexical assisted generation is needed to optimize text generation to enhance the capability of the decoder. The Transformer model incorporating lexically assisted generation is shown in Figure 3.

As can be seen in Figure 3, the model structure is similar to the Transformer model for lexically assisted understanding, but a lexical classifier is connected to the decoder at the last layer, and at this point the source sentence attention module and the auxiliary sentence module in the decoder no longer share parameters. Overall, a multitasking approach, aided by a lexical recognition task, is used to integrate with a general speech recognition system to achieve the model's ability to generate correct multi-subsentence words by subword combining and predicting the corresponding lexical labels for the decoder's generated words. The lexical classifier is responsible for constraining the lexical information for decoding and mapping all the subwords in the same word into a pseudo-word by a multitasking approach. This process requires a length factor to regulate the degree of influence of words of different lengths, at which point the translation task and lexical classification assistance are realized as shown in equation (8).

$$P`=\arg\max\{\lambda * P(T|X)\} \quad (8)$$

In Eq. (8), λ denotes the length factor, T denotes the lexical label sequence, and X denotes the source sequence. When extracting lexical information, it is inevitable to produce wrong lexical labels, so in order to reduce the probability of error and be able to better incorporate the lexical auxiliary information, compared with the traditional Transformer framework, at this time, the decoder incorporates the drop-net module and the shared decoder module. The calculation formula is shown in equation (9).

$$s_i^l = MATT(s_i^{l-1}, s_{<i+1}^{l-1}, s_{>i+1}^{l-1}) \quad (9)$$

In Eq. (9), s_i^l denotes the l layer decoder output result of the random variable at time i . Overall, the drop-net module enables the complete utilization of both sentence encoder and source sentence encoder outputs.

2.2 English translation model construction for speech recognition under multiple information sources

The general use of single source information as input often results in erroneous recognition translation results, so the study addresses the multimodal information of multiple information sources, such as text, language, image, video, etc., that exist in the actual speech recognition process, and selectively adds them to the subsequent speech recognition translation model. An example of English-Chinese recognition translation with multiple information sources is shown in Figure 4.

As can be seen in Figure 4, unlike the conventional speech recognition task, the combination of video and image-based speech recognition results in more accurate semantic understanding. For example, the original sentence in a single source language, "He is easily the best student in the class", can be easily interpreted as "He is easily the best student in the class" after conventional translation, whereas the correct translation is "He is absolutely undoubtedly the best student in the class". The message is "He is definitely the best student in the class". The reason for this is that "easily" and the fixed collocation "be the best" in the original English sentence are not correctly recognized grammatically and semantically, and it is also attributed to the single-source input translation model. Assuming that multiple sources of video information are added for semantic feature assistance, the English-Chinese translation is shown in Figure 4 above. Introducing effective visual audio information as a single source for the translation model can improve the process of recognizing the translation and corroborate the accuracy of the translation results. Through the above analysis, the research will be based on the Transformer model, through the cross-modal audio-video fusion technology and the contextual sentence opening technology to fuse with the language and video information respectively, and finally input into the encoder and decoder. The structure of the cross-modal audio and video fusion network is schematically shown in Figure 5.

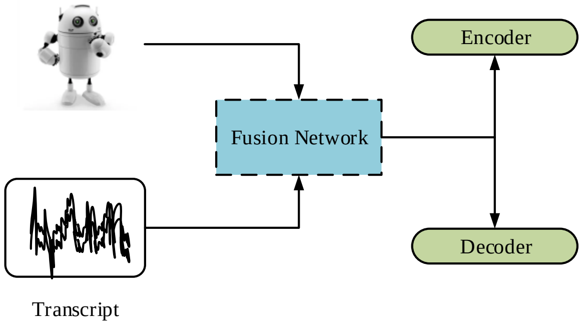


Figure 4: An example of english chinese recognition translation with multiple information sources.

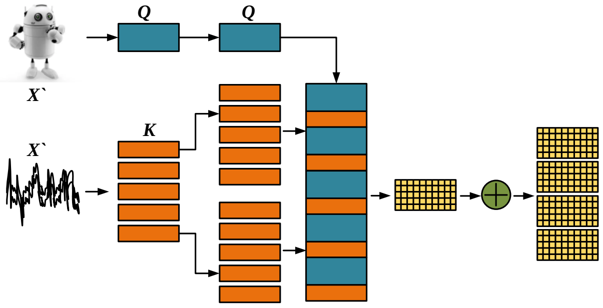


Figure 5: Schematic diagram of cross modal audio video fusion network structure.

As can be seen in Figure 5, the query vector Q is extracted from the video features $X^{`}$ while the key matrix K and the value matrix are extracted from the speech features. The output of the audio-video network is a weighted set of each value matrix, where each element in the value matrix is computed from the vector Q and the corresponding fusion function. The computation of the attention score for the cross-modal audio-video fusion network is shown in equation (10).

$$\alpha = MATT(Q, K, V) \quad (10)$$

In Eq. (10), α denotes the score of the multi-attention mechanism and also expressed as the relative importance of each audio frame. In addition, in order to obtain the speech features guided by the global video features, the obtained attention values are continued to be transmitted to the linear layer and the context is generated using the cross-attention mechanism. This process is calculated as shown in equation (11).

$$x_c = a^T V W^c + b^c \quad (11)$$

In Eq. (11), x_c denotes the context obtained after cross-computation, W^c and b^c both denote the parameters of the linear layer, and $a^T V$ denotes the visual adaptive training result of a lexical labeling. The fusion features obtained by introducing the speech vectors guided by the video information into the speech feature sequence of each frame are shown in Eq. (12).

$$x_i^f = x_i^\alpha + x^c, i = 0,1,\cdots,n \quad (12)$$

In Eq. (12), x_i^f denotes the fusion feature, x^c denotes the speech representation vector, and x_i^α denotes the speech feature sequence of each frame. The fusion features of each frame are combined through the visual adaptive theory to get the final audio-video fusion feature, which is calculated as shown in Eq. (13).

$$x_i^{vat} = x_i^\alpha + x^v, i = 0,1,\cdots,n \quad (13)$$

In Eq. (13), x_i^{vat} denotes the final audio and video fusion features. In summary, the fused audio and video multi-information sources are introduced into the Transformer model and the speech recognition translation model under multi-information sources is constructed, and the results of this model are shown in Figure 6.

As can be seen from Figure 6, the input information is compared with the traditional Transformer model, and the new model adds the combined video and audio signals. At the same time, the audio and video signals are more complex than ordinary speech signals, so two additional convolutional layers are added to the model to collect features from the input audio and video, which are used to reduce the length of the sequence in order to improve the translation speed, and finally, one all-connecting layer is added to complete the purpose of converting speech features. In the cross-modal audio-video fusion network model, the audio modal information dominates while the video modal information plays a secondary role to emphasize the speech information. Meanwhile, the study does not consider the relevance of other contents in the video for the time being to avoid the interference of other video factors.

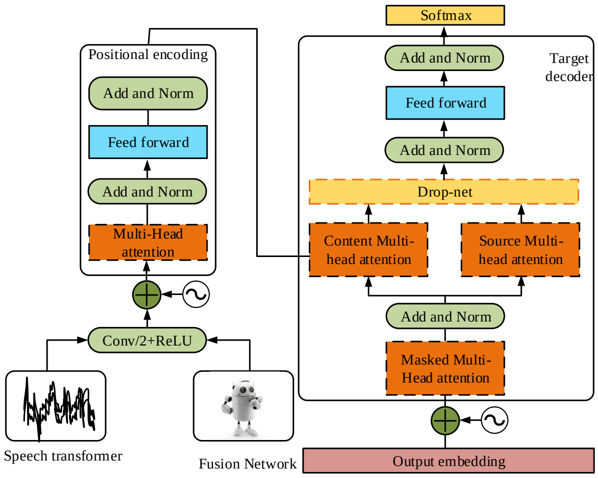


Figure 6: A transformer model for integrating audio and video with multiple information sources.

3 Performance test of english translation model for speech recognition combined with lexical assistance under multiple information sources

In order to test the comprehensive performance of the study's proposed multi-information source speech recognition translation model incorporating lexical aids, the first section of the study tests and compares one of the lexically aided comprehension models, and lexically aided generation models, respectively. The second section then tests the simulation performance of the final proposed model.

3.1 Performance test of lexically assisted speech recognition translation model

In order to validate the performance status of the multi-information source speech recognition translation model proposed in the study, the study builds a suitable experimental environment. The CPU is Intel®Corei7-9700CPU@3.00GHz×8 and the GPU is NVIDIA GeForce RTX 2060 SUPER), the batch size is set to 2000, and the update frequency is set to 10. The recognition-translation model for lexically-assisted comprehension is labeled as POS-Ud, the recognition-translation model for lexically-assisted generation is labeled as PSO-Gen, and the final model that combines lexically-assisted comprehension and generation is labeled as POS-UdGen. There are two important hyper-parameters in the POS-UdGen model, which are the sampling rate during lexical categorization r and the weights of the loss function that regulates the lexical-assisted task λ. In order to verify the degree of influence of the parameters on the POS-UdGen model, and also to find the optimal parameter settings in preparation for subsequent tests. The study was judged by the BLEU score, which ranges from 0-1. The size of the score reflects the similarity between the model translation results and the reference translation results, and the higher the score, the more similar the translations are and the higher the translation quality. The test results are shown in Table 2.

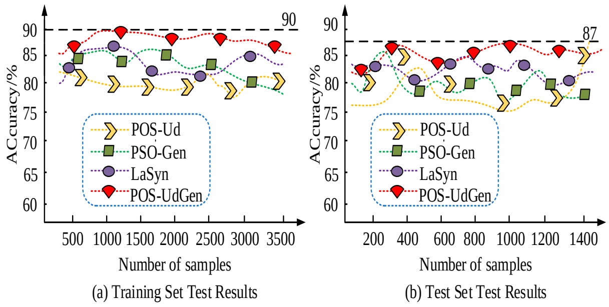
Table 2: The performance impact of different parameters on the pos udgen model.

Parameter	**BCN**		**Parameter**	**BCN**	
Sampling rate r	Training speed (1000-word symbols/second)	BLEU	Loss function weight λ	Training speed (1000-word symbols/second)	BLEU
0.1	20	0.77	1	38	0.59
0.2	25	0.74	2	37	0.63
0.3	22	0.63	3	29	0.62

0.4	23	0.69	4	32	0.49
0.5	24	0.82	5	41	0.73
0.6	24	0.80	6	33	0.64
0.7	27	0.79	7	36	0.74
0.8	31	0.67	8	34	0.79
0.9	28	0.62	9	38	0.72
1.0	28	0.42	10	32	0.61

As can be seen from Table 2, the POS-UdGen model BLEU scores highest when the sampling rate r = 0.5 and the loss function weight λ = 0.8. It indicates that the model recognition and translation results under these two sets of parameter settings are closest to the reference translation results. Therefore, the subsequent recognition and translation effect test is set at r = 0.5 and λ = 0.8. The recognition translation task is carried out using the BCN corpus, which is a large-scale corpus jointly established by Oxford Press, Longman Publishing Company British Library, and Oxford University Computer Center. The collection includes newspapers, periodicals, novels, books and other types of corpus. The study uses nearly 20,000 bilingual pairs from it and divides them into training set and test set according to the ratio of 8:2 to train and test the POS-Ud model, PSO-Gen model, POS-UdGen model and LaSyn model of the same type, and finally the recognition accuracy is used as the reference index, and the test results are shown in Figure 7.

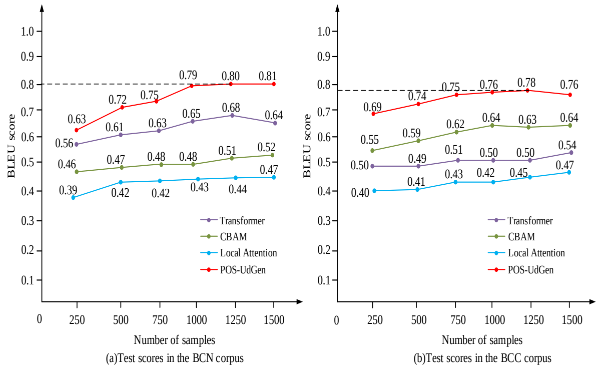
Figure 7(a) shows the recognition accuracy results of the four recognition translation models on the training set, and Figure 7(b) shows the recognition accuracy results of the four recognition translation models on the test set. From Figure 7, it can be seen that the recognition results of the four models on the training set are significantly better than those on the test set, and the best performing model is POS-UdGen. the highest recognition accuracy is up to 87%. Compared to the lowest performing POS-Ud model, the performance improvement is about 12.7%. It shows that the proposed POS-UdGen can significantly improve the recognition accuracy of the translation system. After comparing the recognition rates, the study compares the POS-UdGen model, the Transformer model, the Recognized Translation Model with Local Attention and the Recognized Translation Model with Convolutional Block Attention Module (CBAM) with the BLEU score. (Convolutional Block Attention Module, CBAM) are compared. Meanwhile, in order to ensure the validity of the data, the study was tested in BCN corpus and BCC modern Chinese corpus respectively. The specific test results are shown in Table 8.



(a) Training Set Test Results

(b) Test Set Test Results

Figure 7: Recognition and translation accuracy results of different models.



(a)Test scores in the BCN corpus

(b)Test scores in the BCC corpus

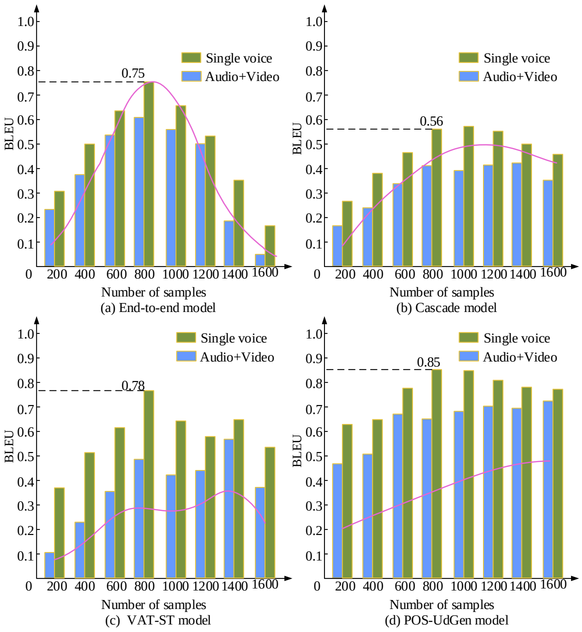
Figure 8: BLEU score results of different models in different corpora.

Figure 8(a) shows the BLEU score graphs of different models in BCN corpus, and Figure 8(b) shows the BLEU score graphs of different models in BCC corpus. As can be seen from Figure 8, the POS-UdGen model proposed by the study achieves excellent results in both different corpora, with a maximum BLEU score of 0.81. In addition, the score of the Transformer model is much higher than that of the other recognizing translation models, which indicates that it is meaningful for the study to make improvement enhancements in the Transformer model, and also proves that lexical aids enhance the characterization of source utterances and effectively improves the translation quality. To verify the robustness of the POS UdGen model, the study continued to introduce the Common Voice multilingual dataset and the TIMIT acoustic feature dataset for testing. The Common Voice multilingual dataset is a multilingual dataset that includes nearly 80000 people from over 60 countries around the world. This dataset is suitable for model testing in different languages. The TIMIT acoustic feature dataset consists of broadband recordings from 630 speakers and American English from eight major dialects. Use Common Voice and TIMIT as multiple input sources, and BCN and BCC as single input sources. The above four models were tested using processing time and resource utilization as indicators, and the test results are shown in Table 3.

Table 3: Model processing time and resource utilization results under different datasets.

Data source type	**Data set**	**Modle**	**Processing time/s**	**Resource utilization rate/%**
Single information source	BCN	Transformer	4.25	52.11
		Local Attention	6.21	69.44
		CBAM	4.38	48.77
		POS-UdGen	2.11	24.16
	BCC	Transformer	3.71	44.59
		Local Attention	5.22	49.63
		CBAM	3.14	36.58
		POS-UdGen	1.72	21.47
Multiple information sources	Common Voice	Transformer	6.28	61.27
		Local Attention	8.28	67.69
		CBAM	4.83	43.68
		POS-UdGen	2.69	27.16
	TIMIT	Transformer	7.41	54.17
		Local Attention	9.68	53.66
		CBAM	6.87	29.39
		POS-UdGen	3.17	20.17

As can be seen from Table 3, the processing time of the Local Attention model is generally preferred to be long in both single and multiple sources, followed by the Transformer and CBAM models, and the optimal performance is still the POS-UdGen model proposed in the study. The POS-UdGen has the shortest processing time of 1.72 seconds in single source, and the lowest resource utilization rate of 21.47%. The shortest processing time under multiple novel sources is 2.69 seconds and the lowest resource utilization is 20.17%. It can be seen that the time consumption of the proposed method in the study for multiple source recognition and translation is more but there is a lack of reduction in the overall resource utilization. It shows that the model still has high performance in different linguistic and acoustic environments, and the multi-information source recognition method is more capable of stimulating the real effect of the model.



(a) End-to-end model

(b) Cascade model

(c) VAT-ST model

(d) POS-UdGen model

Figure 9: Performance of different models under single or multiple source information.

3.2 Practical application test of speech recognition english translation model under multiple information sources

In order to verify the testing effect of the POS-UdGen model proposed in the study in practical applications, the study adopts the how2 multi-information source dataset as the data background, which contains a total of 200,000 pairs of audio and video information. In addition, the speech recognition translation model encoder is set to be 10 layers, the decoder is 6 layers, contains 6 attention heads, and the dropout ratio is 0.1. Using BLEU as the reference index, the end-to-end model, the cascade model, the VAT-ST model, and the POS-UdGen model are tested in a single-source information session and a multiple-information-source environment, respectively, and the test results are shown in Figure 9.

Figure 9(a) shows the performance results of the end-to-end model under single or multiple information sources, Figure 9(b) shows the performance results of the cascade model under single or multiple information sources, Figure 9(c) shows the performance results of the VAT-ST model under single or multiple information sources, and Figure 9(d) shows the performance results of the POS-UdGen model under single or multiple information sources. The purple curve in the figure represents the recognition and translation performance status of each model, and it can be intuitively found that the POS-UdGen model has the best overall performance from this line. In addition, the recognition and translation performance of each model in the audio+video information source environment is significantly larger than that in the single audio environment, indicating that the signal input from multiple information sources will play a positive role in the recognition and translation of the model. And the proposed POS-UdGen model will be easier for speech recognition and translation under multiple information sources because it incorporates cross-modal audio and video fusion techniques.

In order to reveal more intuitively and effectively the alignment of source speech and target translation under multiple information sources, the study randomly selects an audio and video clip in the dataset, and analyzes the performance of the POS-UdGen model for recognizing translations in terms of a confusion matrix, and the confusion matrix results are shown in Figure 10.

As can be seen in Figure 10, for the source sentence "He is the best student in the class", the POS-UdGen model recognizes and translates it with good accuracy, especially for the nouns and pronouns in it. However, the recognition of the adjective "best" is misaligned with the similar "the", and the second "the" is misaligned with "in". The second "the" is misaligned with "in". Therefore, the POS-UdGen model needs to be further tested for recognizing translations of real words. In order to investigate the effect of real word lexicality in speech recognition on the performance of the proposed POS-UdGen model that combines lexically assisted comprehension and lexically assisted generation, ablation tests were conducted on nouns, verbs, adjectives, adverbs, pronouns, etc., which are contained in real words in English utterances. The test results are shown in Table 4.

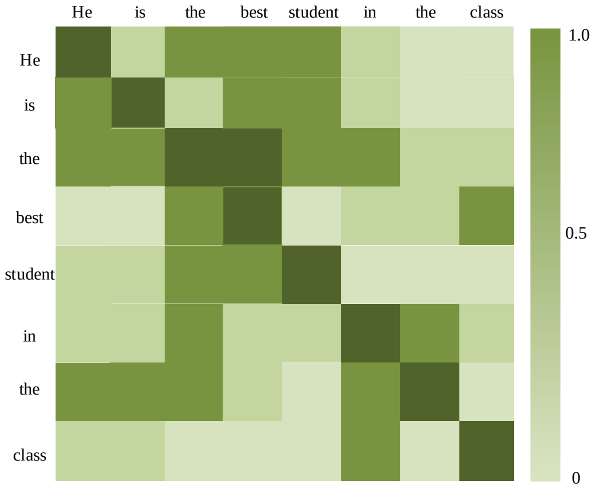
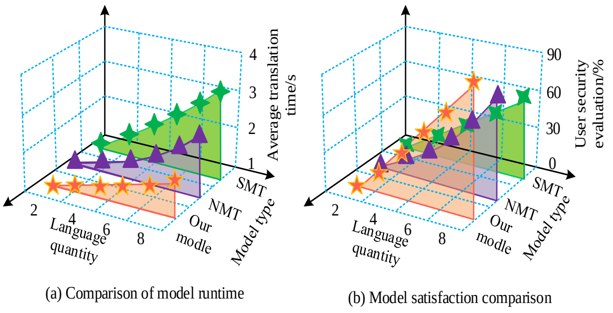


Figure 10: Confusion matrix graph of pos udgen model.

Table 4: The ablation test results of POS UdGen model for real words.

Content word	BCN			BCC		
/	Number of word symbols	Accuracy	BLEU	Number of word symbols	Accuracy	BLEU
Noun	2000	85.40%	0.82	2400	82.70%	0.86
Verb	1800	72.60%	0.67	3000	80.10%	0.72
Adjective	3400	42.70%	0.43	2800	57.30%	0.37
Adverb	1000	45.90%	0.47	3300	48.60%	0.34
Pronoun	3600	69.80%	0.67	2100	77.80%	0.69



(a) Comparison of model runtime

(b) Model satisfaction comparison

Figure 11: Traffic signal translation results of different models in multilingual scenarios traffic signal translation results of different models in multilingual scenarios

As can be seen from Table 4, the POS-UdGen model has the highest recognition rates for nouns, verbs and pronouns in both corpora, with the highest values of 85.4%, 80.1% and 77.8%, respectively, and the corresponding BLEU scores are also the highest. While adjectives and adverbs have the lowest recognition accuracy and BLEU scores, this result is not much different from the confusion matrix results above. In summary, the POS-UdGen model proposed in the study performs better in the simulation test, and the recognition and translation results are close to the reference source text. In addition, the study introduced the more popular Neural Machine Translation (NMT) and Statistical Machine Translation (SMT) models using actual traffic control scenarios in different countries as the test environment. And these models are deployed at the traffic control indicator to try to analyze the translation effect of different models. Firstly, in order to achieve effective deployment, the study optimized the training of various models to reduce the complexity. Second, the translation time and user safety score were used as indicators for testing, and the test results are shown in Figure 11. In addition, the study introduced the more popular Neural Machine Translation (NMT) and Statistical Machine Translation (SMT) models using actual traffic control scenarios in different countries as the test environment. And these models are deployed at the traffic control indicator to try to analyze the translation effect of different models. Firstly, in order to achieve effective deployment, the study optimized the training of various models to reduce the complexity. Secondly, the translation time and user security score were used as indicators for testing and the test results are shown in Figure 11.

Figure 11 shows the comparison of the average translation time consumption of the three models in the multilingual scenario, and Figure 12 shows the comparison of the security satisfaction evaluation of the three models in the multilingual scenario. As can be seen from Fig. 11, the translation time consumption of the model increases gradually with the increase of language categories. When it reaches 8 types of languages, the average translation time of the POS-UdGen model at this point is 1.8 seconds, which is a reduction of 1.2 seconds compared with the 3 seconds of the SMT model. In addition, the safety rating of the POS-UdGen model for traffic light control is close to 85%, while the rating of the SMT model is only 60%. In summary, it can be shown that the proposed model of the study is able to adapt to the translation work under multi-source and multi-language information input scenarios, and can efficiently guide the translation of traffic control language for the development of different countries and the needs of tourists.

4 Conclusion

With the increasing cultural exchanges, speech recognition and translation in transnational languages is of particular importance. In order to successfully construct a more powerful and accurate English translation model for speech recognition, the study improves the existing Transformer framework and proposes a final recognition translation model, i.e., the POS-UdGen model, with the help of multiple information sources and lexical aids. The experimental results show that the performance of the proposed POS-UdGen model in the study works best when the sample rate r = 0.5 and the weight of the loss function λ = 0.8, at which time the best recognition and translation results are achieved. Comparison with other recognition and translation models of the same type reveals that the POS-UdGen model has more accurate and stable performance results, with a maximum recognition accuracy of 87% and a maximum BLEU score of 0.81. The model has the shortest processing time of 1.72 seconds in single information source and the lowest resource utilization of 21.47%. The shortest processing time in multiple novel sources is 2.69 seconds and the lowest resource utilization is 20.17%.

In the simulation test results, the BLEU scores of the POS-UdGen model are much larger than those of a single audio environment in a multi-information source environment and the the POS-UdGen model has the highest recognition rate for real words, verbs and pronouns in the source utterance, with the highest values of 85.4%, 80.1% and 77.8%, respectively. In addition, the average translation time of this model in real traffic light control translation is 1.8 seconds, and the safety evaluation is close to 85% In addition, the average translation time of this model in real traffic light control translation is 1.8 seconds, and the safety evaluation is close to 85%. In summary, the POS-UdGen model proposed in the study shows significant improvements in speech recognition and translation tasks, improving recognition through the lexical labeling function in lexical assistance and enhancing the performance of speech translation through the multi-information source approach of audio and video. However, the study only discusses the speech recognition process and translation process, and has not yet explored how to effectively acquire speech information. In order to improve the completeness of the recognition-translation model, subsequent studies can continue to explore this area in depth.

5 Discussion

The study introduces multiple information sources and lexical aids for technical modification from the nature of language and semantic points. Among the multiple information sources such as video, image and text to enrich and enhance the contextual understanding of speech recognition. This integrated approach of utilizing multiple information sources demonstrates higher accuracy and adaptability when dealing with complex context and speech information. For example, in the model performance test, Local Attention and CBAM models enhance speech recognition and translation only from the perspective of attention mechanism. However, the general interval of test time under datasets with different acoustic features and language types is 6-8 seconds, which is nearly double compared to the average time of 3 seconds for studying the proposed models. The results are similar to those of Vlter C et al. who proposed a fused visual neuropresentation approach for recognizing translations, both of which do not involve data processing with multiple information sources. In addition, the BLEU test results of the end-to-end model have more ups and downs in the comparison test between single and multiple information sources, and the change curve of the test results of the VAT-ST model is similar to that of the proposed method of the study, but with significant numerical differences. The respective BLEU maxima are 0.75, 0.78 and 0.85. The above data illustrate that the signal input from multiple information sources will optimize the recognition and translation of the model. And the POS-UdGen model proposed in the study will be easier for speech recognition and translation under multiple information sources due to the integration of cross-modal audio-video fusion techniques. In addition, the research proposed model considers lexical labels in the translation process through lexical aids, and the model is able to recognize and understand the structure of sentences more accurately, especially when dealing with sentences that are highly polysemous and complex in structure. As shown by the confusion matrix test results and the English real word detection results, the POS-UdGen model's label recognition rates for nouns, verbs and pronouns are 85.4%, 80.1% and 77.8%, respectively, which show excellent detection and recognition efficiency compared to other models. The results perform in line with those of Shao Y et al. However, the study integrates the subsequent translation process with lexical assistance, opening up a new layer of applications for speech recognition and lexical assistance. It can be said that the novelty and effectiveness of this approach is clearly demonstrated in the experimental results.

The above findings can illustrate that the English translation model combining multiple information sources has high speech recognition accuracy in complex environments, and the method provides a new perspective for processing multimodal data and improves the translation embarrassment of relying on traditional single speech recognition. In addition, the combination of lexical assistance is of great practical significance for improving the processing of semantic understanding and contextual relevance in speech recognition. It not only improves the performance of the model in processing complex utterances and polysemous words, but also provides a new approach to understand the linguistic structure of speech. Future research can expand the usage domain to other languages, such as German, French, etc., further highlight the performance advantages of the technique in different language contexts through the study of social media technology and cultural contexts, and apply the model to cross-cultural international conferences, tourism services and other interactive domains in order to evaluate its performance effects in practical applications.

References

[1] Ntalampiras S (2021). Speech emotion recognition via learning analogies. *Pattern Recognition Letters* 144(5), pp. 21-26. https://doi.org/10.1016/j.patrec.2021.01.018

[2] Atack J M, Guo C, Yang L, Zhou Y, Jennings M P (2020). DNA sequence repeats identify numerous Type I restriction-modification systems that are potential epigenetic regulators controlling phase-variable regulons; phasevarions. *The FASEB Journal*, 34(1), pp. 1038-1051. https://doi.org/10.1096/fj.201901536RR

[3] Li S, Xing X, Fan W, Cai B, Fordon P, Xu X (2021). Spatiotemporal and Frequential Cascaded Attention Networks for Speech Emotion Recognition. *Neurocomputing*, 448(11), pp. 238-248. https://doi.org/10.1016/j.neucom.2021.02.094

[4] Vlter C, Oberlnder K, Carroll R, Dazert S, Lentz B, Martin R, Thomas J (2021). Non-Auditory Functions in Low-Performing Adult Cochlear Implant Users. *Otology & Neurotology*, 42(5), pp. 543-551. https://doi.org/10.1097/MAO.0000000000003033

[5] Bibin S P S, Glittas A X, Gopalakrishnan L (2021). A low latency modular-level deeply integrated MFCC feature extraction architecture for speech recognition - ScienceDirect. *Integration*, 76(1), pp. 69-75. https://doi.org/10.1016/j.vlsi.2020.09.002

[6] Alsayadi H A, Abdelhamid A A, Hegazy I, Fayed Z T (2021). Arabic speech recognition using end-to-end deep learning. *IET Signal Processing*, 15(8), pp. 521- 534. https://doi.org/10.1049/sil2.12057

[7] Lin Y, Yang B, Guo D, Fan P (2021). Towards multilingual end-to-end speech recognition for air traffic control. *IET Intelligent Transport Systems*, 15(9), pp. 1203-1214. https://doi.org/10.1049/itr2.12094

[8] Dong Y, Yang X (2021). Affect-salient event sequence modelling for continuous speech emotion recognition. *Neurocomputing*, 458, pp. 246-258. https://doi.org/10.1016/j.neucom.2021.06.036

[9] Zhao L, Zhang A, Liu Y, Fei H (2020). Encoding Multi-Granularity Structure Information for Joint Chinese Word Segmentation and POS Tagging. *Pattern Recognition Letters*, 138(10), pp. 163-169. https://doi.org/10.1016/j.patrec.2020.07.017

[10] Chen X, Hai Z, Wang S, Li D, Wang C, Luan H (2020). Metaphor Identification: a Contextual Inconsistency based Neural Sequence Labeling Approach. *Neurocomputing*, 428(7), pp. 268-279. https://doi.org/10.1016/j.neucom.2020.12.010

[11] Shao Y, Lin J, Srivastava G, Jolfaei A, Guo D, Hu Y (2021). Self-attention-based conditional random fields latent variables model for sequence labeling - ScienceDirect. Pattern Recognition Letters, 2021, 145(5):157-164. https://doi.org/10.1016/j.patrec.2021.02.008

[12] Samuel S, José Marcio D, Evangelos M, Lilian B (2021). Deep analysis of word sense disambiguation via semi-supervised learning and neural word representations. *Information Sciences*, 570(1), pp. 278-297. https://doi.org/10.1016/j.ins.2021.04.006

[13] Guo Y, Mustafaoglu Z, & Koundal D (2022). Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms. *Journal of Computational and Cognitive Engineering*, 2(1), pp. 5-9. https://doi.org/10.47852/bonviewJCCE2202192

[14] Lei Y (2022). Research on microvideo character perception and recognition based on target detection technology. *Journal of Computational and Cognitive Engineering*, 1(2), pp. 83-87. https://doi.org/10.47852/bonviewJCCE19522514

[15] Lin Y, Li Q, Yang B, Yan Z, Tan H, Chen Z (2021). Improving speech recognition models with small samples for air traffic control systems. *Neurocomputing*, 445(20), pp. 287-297. https://doi.org/10.1016/j.neucom.2020.08.092