SIN-CHN SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(3), 1136.
https://doi.org/10.62617/mcb1136

Article

Development of multimodal English translation system based on biomechanics

Yaodan Liang

School of Foreign Languages, Yulin Normal University, Yulin 537000, China; echolianglyd@163.com

CITATION

Liang Y. Development of multimodal English translation system based on biomechanics. Molecular & Cellular Biomechanics. 2025; 22(3): 1136. https://doi.org/10.62617/mcb1136

ARTICLE INFO

Received: 17 December 2024
Accepted: 15 January 2025
Available online: 24 February 2025

COPYRIGHT

Copyright © 2025 by author(s). *Molecular & Cellular Biomechanics* is published by Sin-Chn Scientific Press Pte. Ltd. This work is licensed under the Creative Commons Attribution (CC BY) license. https://creativecommons.org/licenses/by/4.0/

Abstract: With the increasing demand for translation quality, improving the translation quality of English translation systems has become a severe challenge. Traditional English translation systems mostly use single-modal information of text for translation, ignore multimodal information such as speech, and fail to combine the biomechanical movement data of the mouth, throat, and tongue, resulting in poor English translation quality. This paper develops an English translation system based on text, speech and biomechanical motion data, using the translation modeling of Transformer and the temporal feature processing advantages of BiLSTM (Bidirectional Long Short-Term Memory). The study first used MFCC to extract speech features and the spectral features of the audio, and used oral pressure sensors, optical sensors and other sensors to synchronously collect the students' oral opening angles and lip curvature, vocal cord vibration frequency, glottal opening and closing degree, tongue tip position, tongue back curvature and other biomechanical motion feature data, and then input the features into the BiLSTM model for time series modeling and the time series relationship between speech and biomechanical features. Then, a multimodal Transformers framework was constructed, and text, speech, and biomechanical data were integrated using a weighted fusion method, and the multi-head attention mechanism was combined for translation. Finally, the BiLSTM model and the multimodal Transformers framework were integrated in series, and the module was designed and integrated into the English translation system. The experiment was based on the LibriSpeech ASR corpus. The results showed that the multimodal combination of text + audio + biomechanical information performed best, reaching 0.82, 0.76 and 0.79 in BLEU, METEOR, and ROUGE-L indicators respectively, which were 0.06 higher than the multimodal combination without the introduction of biomechanical information, and the loss of emotion transfer was only 0.01. The experimental results show that combining biomechanical information can significantly improve the translation quality of the multimodal English translation system and maintain the original emotion of the language.

Keywords: biomechanical information; multimodal information; English translation system; transformer model; oral opening angle

1. Introduction

With the acceleration of globalization and the increase in cross-border communication, English translation systems play an increasingly important role in all walks of life [1,2]. Current English translation systems mostly rely on a single text input, ignoring the integration of multimodal information such as voice and images. The single information input mode often results in translation quality failing to meet diverse practical needs, especially when faced with complex contexts and non-standard language expressions. Traditional systems fail to take into account biomechanical motion data, the movement of the mouth, tongue and larynx, which are actually closely related to speech generation and recognition. How to effectively

integrate different types of data based on a biomechanical perspective to improve the accuracy and naturalness of the translation system has become an important challenge in current research.

In the fields of linguistics and neuroscience, the relationship between biomechanical motion data and language expression has been widely studied. Studies have shown that the movement patterns of the mouth and throat can affect the understanding and expression of language. In addition, neuroscience research has also shown that the brain integrates signals from the auditory and sensorimotor systems when processing language to achieve a comprehensive understanding of language. Therefore, incorporating biomechanical motion data into the translation system can better simulate the natural production process of human language and improve the accuracy and naturalness of the translation system.

This paper aims to improve the quality of English translation systems, mainly in terms of the accuracy of speech context and emotional expression. In view of the limitations of traditional unimodal translation systems, an innovative solution based on multimodal features is adopted. The experiment combines text, speech and biomechanical motion data, and designs a multimodal Transformer architecture and a BiLSTM (Bidirectional Long Short-Term Memory) model to achieve fusion modeling of multimodal features and capture of temporal features. In terms of speech and biomechanical features, the study used MFCC (Mel Frequency Cepstral Coefficents) to extract speech features, synchronously collected biomechanical motion data of the mouth, throat and tongue through multiple sensors, and performed data integration and translation modeling based on weighted fusion strategy and multi-head attention mechanism. Experimental results show that the system significantly improves translation performance, outperforming existing methods in multiple indicators such as BLEU, while maintaining the accuracy of language sentiment transfer, providing new ideas and practical support for multimodal translation research.

Paper Contribution:

(1) For the first time, this paper combines biomechanical motion data such as mouth opening angle, lip curvature, and vocal cord vibration frequency with speech and text data, and adopts a translation model based on multimodal feature fusion. It breaks through the limitation of traditional translation systems that only rely on text and speech data, and can fully consider the impact of biomechanical characteristics on language expression, thereby improving the accuracy of the translation system in context understanding and emotional transmission.

(2) The study innovatively uses the BiLSTM model for temporal modeling of multimodal features to capture the temporal relationship between speech and biomechanical features. At the same time, the multimodal Transformer architecture is used to translate the fused data, which significantly improves the performance of the translation system, especially in terms of emotion retention and contextual adaptability.

(3) This paper uses experiments to verify the effectiveness of biomechanical information in a multimodal English translation system, proving that the introduction of biomechanical features significantly improves the translation quality and maintains the emotional consistency of the language.

2. Related work

In the field of English translation, many scholars have begun to gradually shift from traditional single-modal English translation to multi-modal English translation, and have achieved a lot of research results. Zhang [3] and Sitender et al. [4] combined text to translate English and designed an intelligent teaching English translation system, which improved efficiency compared with manual translation. Ma et al. [5] used the GAN (Generative Adversarial Network) model to translate English based on text, and the translation effect was improved by 8%. Speech modality has a large number of applications in English translation, which enables English translation to contain rich semantic information, reduce word error rate, and improve translation scores, but it is only based on a single speech modality [6,7]. Meetei et al. [8] combined visual and text information to translate English and found that the multimodal translation system was better than the baseline single modality. The combination of visual and auditory modalities has shown good application in the translation of traditional Chinese cultural terms and low-resource languages, improving the accuracy of translation [9,10]. The above scholars have improved the quality of translation to a certain extent by combining visual and auditory modalities, visual and text information, etc., from text-based single modality to English translation. However, the above scholars have not considered all modalities, and biomechanical motion data has been ignored by most scholars.

The biomechanical motion data of the mouth, throat, and tongue are of great significance in the multimodal English translation system, which directly affects the production of speech and the clarity of pronunciation. Studies have shown that speech generation depends not only on the shape of the vocal tract, but also on the movement of the tongue and larynx, which is crucial to the accuracy of the translation system [11,12]. Serrurier [13] and Mielke et al. [14] further emphasized that biomechanical data can help the system better understand and generate natural pronunciation and improve the quality of speech translation. The above-mentioned scholars' research shows that the integration of biomechanical motion data helps to improve the overall performance and accuracy of multimodal translation systems.

In recent years, with the rapid development of deep learning technology, many scholars have begun to use models such as Transformer for English translation to improve the quality of translation. As a representative model for processing long texts and semantic understanding, the Transformer model is widely used in the translation of various languages such as English, and has shown good translation performance [15,16]. Gamage et al. [17] applied the CNN (Convolutional Neural Networks) model to English speech input and converted it into gestures and text of other speech, which improved the interpretability of the translation and facilitated communication for people with disabilities. Liu et al. [18] used a method that combined cascaded RCNN (Region-Convolutional Neural Networks) and word embedding fusion to perform Chinese-English multimodal translation, improving the ability to focus on the semantic information of words and improving the quality of translation. Kumhar [19] and Zhili et al. [20] used the LSTM (Long Short-Term Memory) model to translate English into languages such as Urdu, capturing the similarity between the semantics of key words and thus improving the quality of translation.

The above scholars used Transformer models, CNN, LSTM and other models for translation, which significantly improved the translation quality. However, they failed to integrate the advantages of Transformer models and BiLSTM models for translation, and did not integrate biomechanical motion data, leaving a large blank research space. Recent studies have begun to focus on the application of multimodal fusion in English translation. Most of these studies focus on the combination of text and visual information, and the use of biomechanical information is still insufficient. In addition, although some studies have attempted to combine different models to improve performance, in-depth exploration of the complementary advantages between models is still limited. Therefore, this study further explores the application of multimodal fusion in English translation by combining Transformer and BiLSTM models, in order to overcome the shortcomings of existing research.

3. Development of a multimodal English translation system

3.1. Multimodal transformer architecture

In the development of a multimodal English translation system, multimodal data includes text, audio, and biomechanical data. For audio and biomechanical data, this paper first uses the BiLSTM model to model the time series and extract features to capture the dependencies between the previous and next parts. Then, the text, audio, and biomechanical data are input into the multimodal Transformer architecture for multimodal fusion and other processing to achieve multimodal English translation.

In the encoder and decoder, the multi-head attention layer, feedforward neural network layer, and normalization layer work closely together to achieve effective processing and feature extraction of multimodal data. The multi-head attention layer performs self-attention calculations on the input multimodal data to generate a comprehensive feature representation. The feedforward neural network layer further extracts and integrates feature information, and the normalization layer eliminates the dimensional differences between different modal data. Through this collaborative work between layers, the encoder and decoder can extract feature information that is useful for translation tasks.

3.1.1. Multi-head attention mechanism

In the multimodal Transformer architecture, the multi-head attention mechanism is a core component, and its core idea is to pay attention to different positions of the input sequence in parallel [21–23]. The multi-head attention mechanism calculates multiple attention functions in parallel, projects the input information into multiple subspaces, and captures the complex relationship between different modal features.

In the multimodal Transformer architecture, the multi-head attention mechanism achieves multimodal information fusion by focusing on different positions of the input sequence in parallel. When processing text, speech, and biomechanical data, the query vector Q, key vector K, and value vector V of each modality are calculated separately, and the attention weight is obtained through dot product calculation to fuse information from different modalities. For the input feature sequence C, define the query vector Q, key vector K, and value vector V. The calculation formulas for the query vector Q, key vector K, and value vector V are shown in Equations (1)–(3)

respectively.

$$Q = CW_Q \tag{1}$$

$$K = CW_K \tag{2}$$

$$V = CW_V \tag{3}$$

W_Q, W_K, and W_V represent learned weight matrices.

In the multi-head attention mechanism, multiple different attention heads are calculated in parallel. For each attention head, a different weight matrix is learned and the relationship between query, key, and value is calculated in different subspaces. The calculation formula for the attention weight of each head is shown in Equation (4).

$$\mathrm{A}\,(Q,K,V) = \mathrm{softmax}\,(\frac{QK^{\top}}{\sqrt{\mathrm{d}_k}})V \tag{4}$$

The calculation formula for parallel computing of the multi-head attention mechanism is shown in Equation (5).

$$Z(Q,K,V) = Con(head_1,\ldots,head_s)W_\alpha \tag{5}$$

$$head_s = A(QW_Q^s, KW_K^s, VW_V^s) \tag{6}$$

In order to make full use of the information in multimodal data, this paper extends the multi-head attention mechanism to speech and biomechanical data in addition to processing text information. In the case of multimodality, the input data includes text, speech, and biomechanical features, which are processed by their respective attention mechanisms. For each modality input, the query, key and value matrices are calculated respectively, and the outputs are Z^{test}, Z^{audio}, and Z^{bio}. The experiment uses the weighted fusion method [24,25] to fuse the feature information of different modalities, and the fused expression is shown in Equation (7).

$$Z = \beta_1 Z^{test} + \beta_2 Z^{audio} + \beta_3 Z^{bio} \tag{7}$$

β_1, β_2, and β_3 all represent weight coefficients. Z^{test}, Z^{audio}, and Z^{bio} represent the attention output results of text, speech, and biomechanical data, respectively.

3.1.2. Encoder design

In the encoder design, it includes a multi-head attention mechanism, a feedforward neural network, and a normalization layer [26,27]. For the text, speech and biomechanical data, they are projected onto the same dimension. After projection, the three modal data are concatenated into a whole and processed using a multi-head attention layer. The output of the multi-head attention layer is then transformed nonlinearly through a feedforward neural network to further extract features.

The calculation formula for the nonlinear transformation is shown in Equation (8).

$$FFN(z) = ReLU(zW_1 + \zeta_1)W_2 + \zeta_2 \tag{8}$$

W_1 and W_2 represent the weight matrix of the feedforward neural network, and ζ_1 and ζ_2 represent the bias terms.

In order to improve the training stability of the model, the encoder adds a

normalization layer and residual connection after each sublayer. The normalization calculation formula is shown in Equation (9).

$$\mathrm{LN}（\mathrm{z}） = \frac{\mathrm{z}-\mu}{\eta+\epsilon} \cdot \vartheta_1 + \vartheta_2 \tag{9}$$

μ and η represent the mean and standard deviation respectively, ϑ_1 and ϑ_2 represent learnable parameters, and ϵ represents a smoothing term.

3.1.3. Decoder design

In this experiment, the decoder is used to convert the feature representation generated by the encoder into the target output. The decoder design is similar to the encoder, but a cross-attention layer is added to combine the output features of the encoder [28,29].

For the decoder, the first step is to input the target modality data, the second step is to use the masked multi-head attention layer to process the dependencies in the sequence, the third step is to use the cross attention layer to combine the output features of the encoder, and the fourth step is to use the feedforward neural network for nonlinear transformation and stabilization. The calculation formula of the attention in the masked multi-head attention layer is shown in Equation (10).

$$A（Q,K,V） = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}} + H\right)V \tag{10}$$

H represents the masking matrix.

After multiple layers of stacking, the decoder outputs the target features and maps them to the probability distribution of the target vocabulary, as shown in Equation (11).

$$P(y_t|y < t, Z_{en}) = softmax(Y_{ou}W_p + \zeta_p) \tag{11}$$

ζ_p represents the projection parameters and Z_{en} represents the encoder output sequence.

The multimodal Transformer architecture is shown in **Figure 1**.

Figure 1 shows that the multimodal Transformer architecture integrates text, speech, and biomechanical data for translation. The translation of “I am a super English translation assistant” greatly highlights the emotional expression of the original sentence.

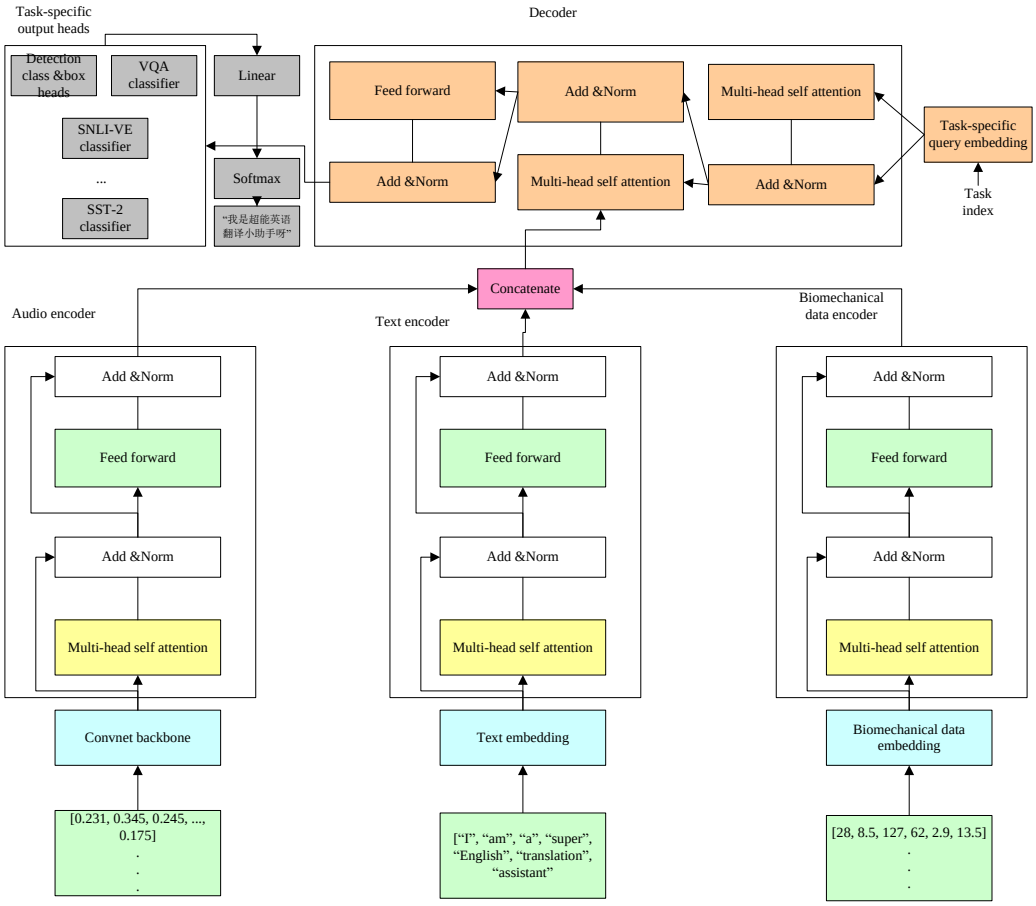

Figure 1. Multimodal transformer architecture.

3.2. Construction of BiLSTM model and processing of temporal feature data

3.2.1. Construction of BiLSTM model

In order to capture the temporal dependency between multimodal speech data and biomechanical motion signals, this paper adopts the BiLSTM model [30–32] to provide high-quality feature representation for subsequent multimodal fusion and translation modeling.

In BiLSTM, its forward and backward units calculate the corresponding hidden states in chronological order and reverse order respectively. The calculation formulas for the forward and backward hidden states at each time step are shown in Equation (12).

$$\begin{aligned} \overrightarrow{m}_t &= LSTM_1(a_t, \overrightarrow{m}_{t-1}) \\ \overleftarrow{m}_t &= LSTM_2(a_t, \overleftarrow{m}_{t-1}) \end{aligned} \tag{12}$$

$\overrightarrow{m}_t$ represents the hidden state of the forward unit, and $\overleftarrow{m}_t$ represents the hidden state of the backward unit.

The hidden state of BiLSTM is connected and output using the forward and backward hidden states, and the expression is shown in Equation (13).

$$m_t = [\overrightarrow{m}_t, \overleftarrow{m}_t] \tag{13}$$

The BiLSTM model consists of two LSTM units, one is the forward LSTM,

which runs from the beginning to the end of the sequence; the other is the reverse LSTM, which runs from the end to the beginning of the sequence. Each LSTM unit includes three gates: input gate, forget gate, and output gate. The input gate determines how much information of the current time step is written into the memory cell, the forget gate determines the degree of forgetting of information in the memory cell, and the output gate determines how much information of the memory cell is output to the hidden state. This structure enables BiLSTM to effectively capture long-term and short-term dependencies in sequence data, providing high-quality feature representation for subsequent multimodal fusion and translation modeling.

3.2.2. Speech data processing

In speech data processing, MFCC is used for feature extraction [33,34]. The speech signal is subjected to short-time Fourier transform (STFT) to convert the time domain signal into a frequency domain signal and obtain the spectrogram of the speech signal. STFT captures the local frequency information of the signal by sliding a window on the signal and calculating the Fourier transform of the signal in the window. Then, the spectrum in the STFT result is nonlinearly scaled using the Mel filter group to match the human ear's perception of frequency. The Mel frequency scale is closer to the way the human ear perceives sound and can better reflect the characteristics of the speech signal [35]. Finally, the spectrogram processed by the Mel filter is subjected to discrete cosine transform to obtain MFCC features. Discrete cosine transform can remove redundant information in the spectrum and extract the main features of the speech signal. Through these steps, the speech signal is converted into a feature vector suitable for model processing, providing effective feature support for subsequent speech recognition and translation tasks.

3.2.3. Biomechanical data processing

The biomechanical data include the movement data of the oral cavity, larynx, and tongue.

(1) Biomechanical characteristics of the oral cavity

In this paper, the biomechanical characteristics of the oral cavity include the oral opening angle and the curvature of the lips. The experiment uses oral pressure sensors and image processing technology to obtain dynamic data inside and outside the oral cavity.

For the opening and closing degree, this paper uses an optical sensor to obtain the angle of the oral opening. The expression of the oral opening and closing degree and the curvature of the lips is shown in Equation (14).

$$\begin{aligned} O_i &= arctan(\frac{y_2 - y_1}{x_2 - x_1}) \\ C_i &= \frac{\sum_{j=1}^{n} ||p_j - p_{j+1}||}{n} \end{aligned} \tag{14}$$

O_i represents the degree of oral opening, and C_i represents the curvature of the lips. $(x_1,\ y_1)$ and $(x_2,\ y_2)$ represent the position coordinates of the two ends of the lips. p_j represents the coordinates of the lips at different time points.

(2) Biomechanical characteristics of the larynx

The biomechanical characteristics of the larynx include the vibration frequency of the vocal cords and the degree of glottis opening and closing.

The experiment uses a pressure sensor to obtain the vibration frequency of the vocal cords and a glottis sensor to measure the degree of glottis opening and closing. The expression of the vocal cord vibration frequency and the degree of glottis opening and closing is shown in Equation (15).

$$\begin{matrix} f_s=\frac{1}{T}\sum_{k=1}^{n}|\frac{Q_k}{T}| \\ E_i=\frac{1}{T}\int_0^T e(t)dt \end{matrix} \tag{15}$$

T represents the period, Q_k represents the vibration intensity, and $e(t)$ represents the opening and closing degree.

(3) Biomechanical characteristics of the tongue

The biomechanical characteristics of the tongue include the position of the tongue tip and the curvature of the tongue back. This paper uses electromagnetic sensors to obtain the position of the tongue in space. The expression of the tongue position and curvature is shown in Equation (16).

$$\begin{matrix} L_i = \sqrt{(x_t - x_0)^2 + (y_t - y_0)^2 + (z_t - z_0)^2} \\ B_i = \int_{t_0}^{t_1} \frac{dy}{dx} \end{matrix} \tag{16}$$

Among them, $(x_t,\ y_t)$ represents the coordinates of the tongue at a certain moment, and $(x_0,\ y_0)$ represents the initial coordinates. x and y represent the coordinates of the tongue on the plane.

In order to improve the translation quality of the translation system, this paper introduces the biomechanical characteristic data of the oral cavity, larynx, and tongue, and fuses the three using the weighted average method. The study standardized the features of each part to fuse them at a unified scale, and weighted the sum of the biomechanical features of the oral cavity, tongue, and larynx to merge them into a comprehensive feature vector. The merged biomechanical feature vector is shown in Equation (17).

$$F_{bio} = \delta_1 O_i + \delta_2 C_i + \delta_3 f_s + \delta_4 E_i + \delta_5 L_i + \delta_6 B_i \tag{17}$$

Among them, $\delta_1,\ \delta_2,\ \delta_3,\ \delta_4,\ \delta_5,\ \delta_6$ represent the weighting coefficients.

After fusion, the experiment inputs all biomechanical characteristics of the mouth, tongue and larynx into the BiLSTM model for time series data processing, enhancing the integrity of multimodal data and enabling the translation system to more accurately capture the key biomechanical information in the pronunciation process.

3.3. Fusion of transformer and BiLSTM

The BiLSTM model extracts key time series information by capturing the temporal relationship between speech and biomechanical features. This information includes the acoustic features of speech and the motion features of the vocal organs, providing comprehensive contextual support for translation. The Transformer model uses its multi-head attention mechanism to integrate text, speech, and biomechanical features to capture the complex relationship between different modalities. This integration improves the ability to understand the language context and enhances the ability to capture emotional and semantic information.

For Transformer and BiLSTM, this paper uses a series connection method to fuse

the Transformer model architecture and BiLSTM. The experiment applies BiLSTM to the time series modeling of audio and biomechanical data. While capturing the time series features, the output time series representation can be input into the Transformer architecture together with the text information for multimodal fusion. Among them, BiLSTM is responsible for extracting temporal features from audio and biomechanical data, while Transformer integrates cross-modal information through its powerful multi-head attention mechanism to achieve the purpose of multi-modal English translation.

The Transformer architecture uses the multi-head attention mechanism to simultaneously process the interactive information of text, audio, and biomechanical data to achieve effective cross-modal information fusion. The fusion method of Transformer and BiLSTM in this paper can ensure that the complementarity of multimodal data can be fully utilized and improve the translation performance of the system. In particular, when capturing the relationship between complex language and biomechanical movements, it fully utilizes the temporal and contextual characteristics of each data source.

The fusion structure diagram of Transformer and BiLSTM is shown in **Figure 2**.

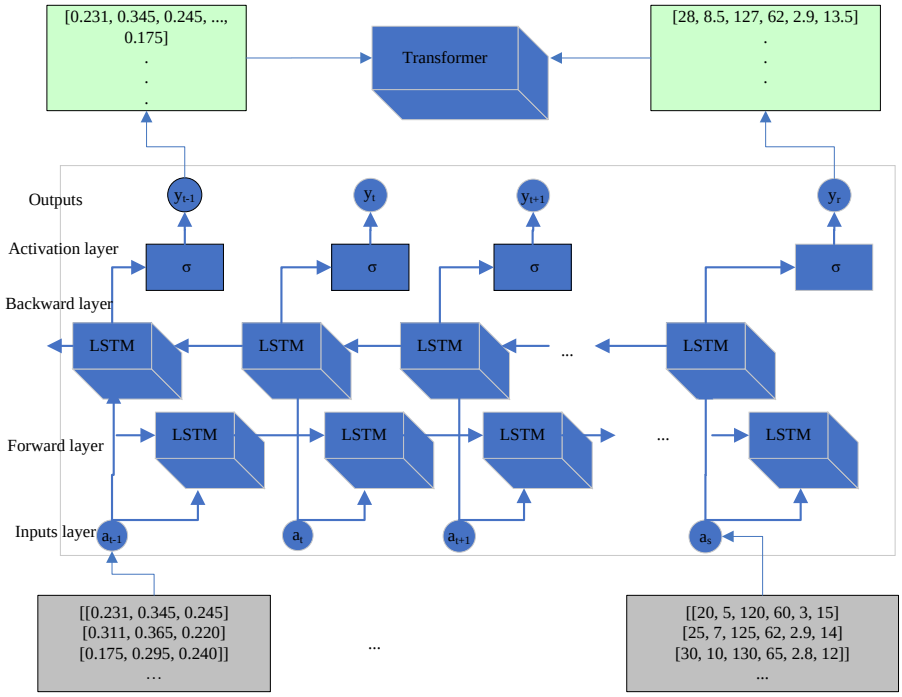

Figure 2. Fusion structure diagram of transformer and BiLSTM.

In **Figure 2**, it can be seen that Transformer and BiLSTM are fused in a direct series manner. The speech data and biomechanical data are directly sent to Transformer for multimodal fusion and translation processing after being modeled in time series by the BiLSTM model.

3.4. Training and optimization of fusion model

(1) Loss function design

This paper comprehensively considers the translation quality, temporal feature learning and the effect of model fusion, and designs a joint optimization function. The

joint optimization function includes the cross entropy loss of the translation output and the L2 regularization loss of the features extracted by the BiLSTM model during the temporal modeling process. The calculation formula of the cross entropy loss is shown in Equation (18).

$$Loss_1 = -\sum_{i=1}^{n} y_{true}^{(i)} log(y_{pred}^{(i)}) \tag{18}$$

$y_{true}^{(i)}$ is the true translation label and $y_{pred}^{(i)}$ is the translation output predicted by the model.

The calculation formula of the regularization loss is shown in Equation (19).

$$Loss_2 = \iota \sum_{i=1}^{n} ||\kappa_i||^2 \tag{19}$$

ι represents the regularization hyperparameter.

(2) Optimizer design and learning rate adjustment

In order to improve the stability and convergence speed of training, the experiment uses the Adam optimizer and adopts a dynamic learning rate adjustment strategy. The update of the Adam optimizer is shown in Equation (20).

$$\theta_{t+1} = \theta_t - \xi \frac{\varpi_t}{\sqrt{\varrho_t} + \varsigma} \tag{20}$$

ξ represents the learning rate, ϖ_t and ϱ_t represent the first-order moment estimate and the second-order moment estimate. ς represents a very small constant.

This paper adds a Dropout layer to both the BiLSTM and Transformer architectures to prevent overfitting. In the BiLSTM model, the Dropout layer is added to the input layer and the hidden layer output layer of each layer, while the Dropout layer in the Transformer model is applied to the input and output layers of the self-attention layer and the feedforward network to enhance the generalization ability of the model.

During the training process, the hyperparameter settings are shown in **Table 1**.

Table 1. Hyperparameters.

Parameters	Value	Parameters	Value
Learning rate	0.001	Dropout probability	0.3
Learning rate decay factor	0.9	Number of BiLSTM hidden layer units	256
Batch size	32	Number of transformer hidden layer units	512
Number of training rounds	50	Number of encoder layers	6
Regularization parameter	0.01	Number of decoder layers	6
Number of BiLSTM time steps	100	Number of multi-head attention heads	8
Gradient clipping threshold	1.0	Optimizer	Adam

In **Table 1**, the learning rate is 0.001, the learning rate decay factor is 0.9, the batch size is 32, the number of training rounds is 50, the regularization parameter is

0.01, the gradient clipping threshold is 1.0, and the Dropout probability is 0.3. The number of BiLSTM time steps is 100, and the number of hidden layer units is 256. The number of hidden layer units of Transformer is 512, the number of encoder layers and decoder layers are both 6, and the number of multi-head attention heads is 8.

4. Module design of multimodal English translation system

In this paper, the designed multimodal English translation system includes text processing module, speech processing module, biomechanical data processing module, multimodal information fusion module and translation output module. In the text processing module, its task is to preprocess the input English text. For English text, this paper performs clear steps such as removing irrelevant symbols, expanding abbreviations, and correcting spelling errors, and uses word segmentation technology to segment the text into words or subwords. Then, the BERT (Bidirectional Encoder Representations from Transformers) tokenizer can be used to encode the vocabulary and output the corresponding word vector representation.

In the speech processing module, the speech signal is first input into the preprocessing module for denoising, silence removal and signal enhancement, and the STFT technology is used to convert the speech signal into a spectrogram. Then the Mel filter bank is used to nonlinearly scale the spectrum in the result after STFT processing to match the Mel scale, and the discrete cosine transform is applied to the spectrum after Mel filter processing to output MFCC features. After obtaining the MFCC features, the experiment uses BiLSTM to further perform time series modeling to capture the long-term and short-term dependencies in the audio signal and enhance the time series characteristics of the speech signal.

In the biomechanical data processing module, the biomechanical movement data of the mouth, larynx, tongue, etc., are mainly processed. For biomechanical data, the processing module first standardizes the biomechanical data to eliminate dimensional differences. The weighted average method can be used to fuse multiple biomechanical features into a comprehensive feature and pass it as input to the BiLSTM model to capture the key biomechanical timing information during speech pronunciation.

For the multimodal information fusion module, it uses a weighted average mechanism to fuse text features, speech features, and biomechanical features into a unified vector representation. The final translation output module generates translations based on the Transformer decoder, where the decoder uses an autoregressive mechanism to gradually generate translation vocabulary in the target language based on the input context information.

5. Evaluation of the English translation system

5.1. Experimental environment

Hardware environment: The server is configured with Intel Xeon Platinum 8280, NVIDIA A100 Tensor Core GPU (Graphics Processing Unit), 512 GB DDR4 (Double-Data-Rate Fourth), 4 TB NVMe SSD (Solid State Disk). The experimental workstation is configured with Intel Core i9-13900K, NVIDIA RTX 4090, 64 GB DDR5 (Double Data Rate 5), and 2 TB NVMe SSD.

Software environment: Server environment Ubuntu 20.04 LTS, workstation environment Windows 11 Professional, Python 3.10, PyTorch 2.0, Transformers 4.32, torchaudio 2.1, NumPy 1.24 and SciPy 1.11.

5.2. Experimental data and preprocessing

The experimental data in this paper include text data, speech data and biomechanical data. The speech data comes from the LibriSpeech ASR corpus, which covers about 1000 h of 16kHz read English speech corpus. For the text data, this paper organized experts to compile it into a manuscript. In terms of biomechanical data, the experiment randomly selected 100 h of speech and called 122 students from a school to simulate conversation scenarios by combining speech and text. Oral pressure sensors, optical sensors, laryngeal pressure sensors, electromagnetic sensors, etc., can be used to synchronously collect students' oral opening angles and lip curvature, vocal cord vibration frequency, glottal opening and closing, tongue tip position, tongue back curvature and other biomechanical movement data. The data was divided using the ten-fold cross validation method, and the mean was taken as the final experimental result. Some of the collected biomechanical data are shown in **Table 2**.

Table 2. Partial biomechanical data.

Student number	Voice segment number	Oral opening angle (°)	Lip curvature (°)	Vocal cord vibration frequency (Hz)	Glottal opening (%)	Tongue tip position (cm)	Tongue dorsum curvature (°)
1	1	20	12	230	70	2.3	15
2	2	18	10	215	65	2.1	13
3	3	22	14	240	75	2.4	16
4	4	19	11	220	68	2.2	14
5	5	21	13	235	72	2.5	17
6	6	23	15	250	78	2.6	18
7	7	20	12	225	70	2.2	14
8	8	19	10	210	60	2.1	12
9	9	21	14	230	73	2.4	16
10	10	22	13	240	75	2.3	15

Table 2 visualizes the data of 10 students' oral opening angles and lip curvature, vocal cord vibration frequency, glottal opening and closing degree, tongue tip position, tongue back curvature, etc.

Data preprocessing

(1) Text data

For text data, this paper first standardizes the original text data, removes extra spaces and special characters, and uses BERT tokenizer to segment the text into subword units. After segmentation, each word is screened to remove common stop words, and a pre-trained BERT model is used to map each word or subword into a high-dimensional word vector.

(2) Speech data and biomechanical processing

In the speech data, the experiment uses spectral subtraction to remove background noise and silent segments, and uses Wiener filtering to optimize and

standardize speech clarity, and combines STFT technology to convert speech signals into spectrograms. In the preprocessing of biomechanical data, the linear interpolation method is first used to synchronize the data collected by different sensors to ensure that all data are aligned on the same time axis. Then the Z-score method is used to detect and remove outliers in the sensor data. In addition, the data is standardized and each feature is normalized to between [0,1]. Finally, a low-pass filter is used to smooth the data to reduce the impact of noise on the data.

5.3. Evaluation indicators

In this paper, the evaluation indicators include BLEU, METEOR, ROUGE-L, TER (Translation Edit Rate), WER (Word Error Rate), CER (Character Error Rate).

BLEU (Bilingual Evaluation Understudy):

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} \tau_n \cdot logP_n) \tag{21}$$

BP represents the penalty factor, P_n represents the n-gram accuracy, and τ_n represents the n-gram weight.

METEOR (Metric for Evaluation of Translation with Explicit ORdering):

$$METEOR = \frac{P \cdot R}{\varphi \cdot P + （1 - \varphi） \cdot R + \phi} \tag{22}$$

P represents precision, R represents recall, φ represents balance factor, and ϕ represents penalty factor.

ROUGE-L (Longest Common Subsequence):

$$ROUGE - L = \frac{LCS(\chi, \psi)}{|\psi|} \tag{23}$$

$|\psi|$ represents the length of the reference sentence, and $LCS(\chi, \psi)$ represents the length of the longest common subsequence between the candidate sentence and the reference sentence.

5.4. Experimental design

Experimental Design Steps:

(1) The experiment is based on the data collected from the public text dataset, and the actual students' speech data and biomechanical data in the experiment are statistically analyzed, and preprocessed by denoising and outliers.

(2) MFCC is used to extract speech features, extract the spectral features of the audio, and input the features into the BiLSTM model for time series modeling to capture contextual semantic dependencies. For biomechanical data, oral pressure sensors, optical sensors, laryngeal pressure sensors, electromagnetic sensors, etc., are used to synchronously collect biomechanical movement data such as the student's oral opening angle and lip curvature, vocal cord vibration frequency, degree of glottis opening and closing, tongue tip position, tongue back curvature, etc., and send them to the BiLSTM model for processing.

(3) A multimodal Transformers framework can be built, and text, speech, and

biomechanical data can be integrated using a weighted fusion method, and then translated using a multi-head attention mechanism.

(4) This paper uses a serial method to integrate the BiLSTM model and the multimodal Transformers framework, and designs a module that is integrated into the English translation system.

5.5. Experimental results

5.5.1. Translation quality statistics

In order to evaluate the translation performance of the multimodal English translation system, this paper uses BLEU, METEOR, ROUGE-L, TER, WER, and CER indicators to quantify the quality of translation. The translation quality statistics are shown in **Figure 3**. In **Figure 3**, the comparison models include Transformers, LSTM, RCNN, and Seq2Seq (Sequence-to-Sequence).

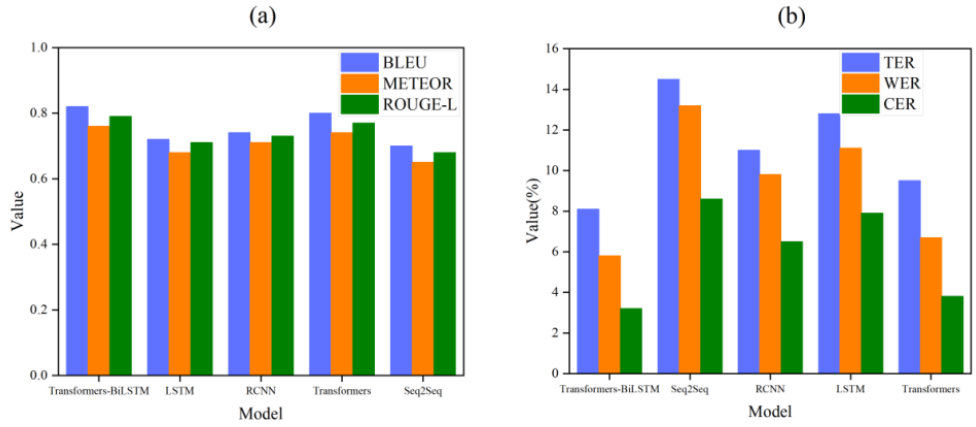

Figure 3. Translation quality statistics: **(a)** BLEU, METEOR, ROUGE-L statistical results; **(b)** TER, WER, CER statistical results.

In **Figure 3a**, the Transformers model achieved 0.8, 0.74, and 0.77 in BLEU, METEOR, and ROUGE-L indicators, respectively. Transformers-BiLSTM (Transformers-Bidirectional Long Short-Term Memory) achieved the best performance in BLEU, METEOR, and ROUGE-L indicators, which were 0.82, 0.76, and 0.79, respectively. In comparison, the Transformers model improved by 0.02, significantly outperforming other comparison models. It can be seen that the introduction of BiLSTM for time series modeling is effective, and the architecture that integrates multimodal data can better capture translation context and semantic information, and enhance the fluency and semantic integrity of translation. The traditional Seq2Seq model performed the worst in terms of BLEU, METEOR and ROUGE-L indicators, which were only 0.70, 0.65 and 0.68 respectively. LSTM and RCNN perform similarly, but both have worse translation quality than the translation system designed by the Transformers-BiLSTM model.

In **Figure 3b**, for the TER, WER, and CER indicators, a low TER value indicates that the model has fewer insertion, deletion, and substitution errors in translation, while lower WER and CER reflect the high accuracy of the output text at the word and character levels. Transformers-BiLSTM achieved the best performance, with TER, WER, and CER values of only 8.1%, 5.8%, and 3.2%, respectively. Seq2Seq's TER,

WER, and CER reached 14.5%, 13.2%, and 8.6%, respectively, with the most translation errors. Overall, Transformers-BiLSTM performed well in all indicators, while Seq2Seq's unimodal characteristics and simple encoding mechanism caused the translation quality to lag behind other models in all indicators.

5.5.2. Multimodal fusion and single-modal effects

In this paper, three modalities are included, namely text, audio, and biomechanical information. The experiment uses the Transformers-BiLSTM model to explore the translation quality under single modality and different multimodalities. The results are shown in **Figure 4**. In **Figure 4**, single modality includes text and audio, and multimodality includes text + audio, text + biomechanical information, audio + biomechanical information, and text + audio + biomechanical information.

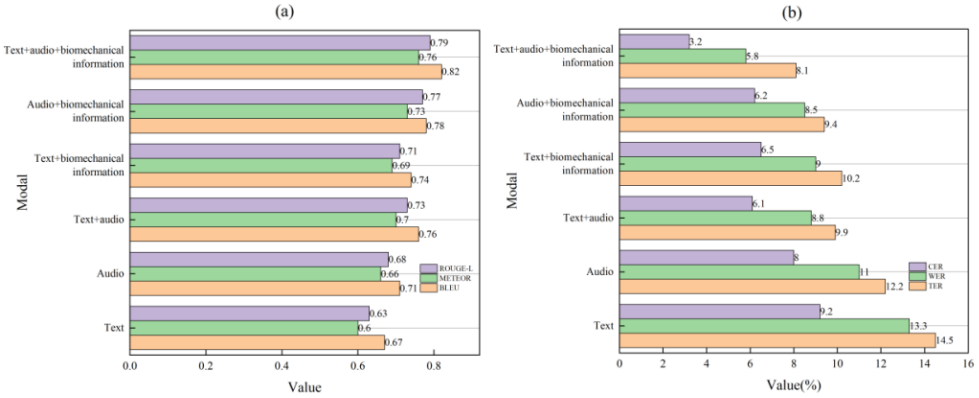

Figure 4. Analysis of multimodal fusion and single modal effects: **(a)** BLEU, METEOR, ROUGE-L statistical results under multimodal and single modal; **(b)** TER, WER, CER statistical results under multimodal and single modal.

In **Figure 4a**, from the perspective of BLEU, METEOR, and ROUGE-L indicators, the multimodal combination of text + audio + biomechanical information performs best, at 0.82, 0.76, and 0.79, respectively. This shows that this multimodal combination can most effectively capture context and semantic information, and improve the semantic accuracy and fluency of translation. In a single modality, audio performs better than text, with a BLEU of 0.71, an improvement of 0.04 over text. This is because audio data contains intonation and contextual information, which helps translate emotional and semantic expressions. In bimodal fusion, the BLEU of audio + biomechanical information reached 0.78, significantly better than other combinations. The BLEU of text + audio was only 0.76, and the BLEU of text + biomechanical information was only 0.74. It can be seen that biomechanical information and audio are highly complementary, which helps to improve translation quality.

In **Figure 4b**, the combination of text + audio + biomechanical information has the lowest TER, WER and CER, which are 8.1%, 5.8% and 3.2% respectively. It can be seen that multimodal fusion significantly reduces vocabulary and character-level errors in translation. In single modality, the TER of audio reaches 12.2%, while that of text reaches 14.5%. It can be seen that the error rate of audio is lower than that of text, indicating that audio features can more effectively capture the details of pronunciation and intonation. In the dual-modal combination, audio + biomechanical

information has the lowest error rate, with a TER of only 9.4%, while text + biomechanical information has a higher error rate, with a TER of 10.2%. Audio is more intuitive than text in terms of semantic transmission, and biomechanical information can more accurately describe the dynamic characteristics of the vocal organs.

In summary, the multimodal information fusion of text + audio + biomechanical information achieves the best English translation quality and meets actual needs.

5.5.3. Model translation speed and translation time, training convergence speed and parameter amount

The translation speed and translation time, training convergence speed and parameter amount of the model are shown in **Figure 5**.

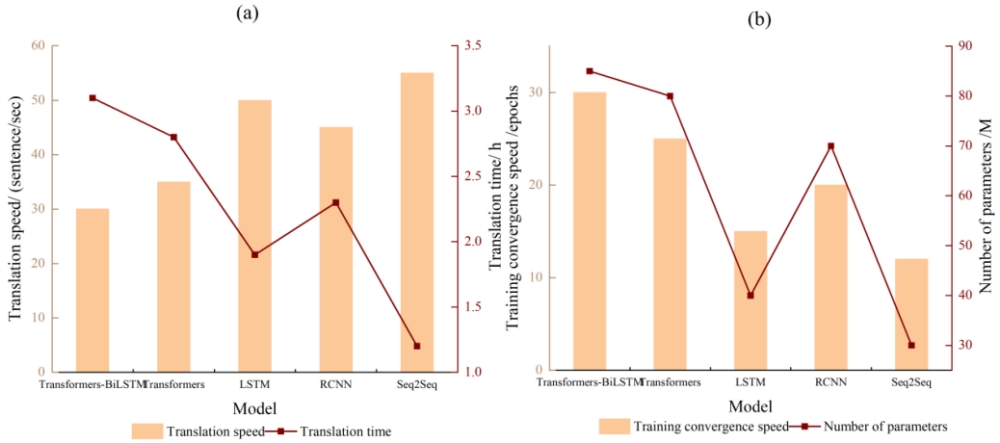

Figure 5. Model translation speed and translation time, training convergence speed and parameter amount: **(a)** Model translation speed and translation time; **(b)** Model training convergence speed and parameter amount.

In **Figure 5**, different models have obvious differences in translation speed and translation time. Seq2Seq has the fastest translation speed, reaching 55 sentences/sec, and the shortest translation time, only 1.2 h. Transformers-BiLSTM has a translation speed of only 30 sentences/sec and a translation time of 3.1 h. Transformers has a translation speed of 35 sentences/sec and a translation time of 2.8 h.

In terms of training convergence speed, Seq2Seq performs best, requiring only 12 epochs, while Transformers-BiLSTM requires 30 epochs to converge, with a parameter count of 85M. LSTM and RCNN have achieved good results in terms of parameter quantity and training convergence speed. From the above, the Transformers-BiLSTM model is complex and has a large number of parameters, which leads to slower translation speed, but it has stronger feature learning capabilities. In the future, the Transformers-BiLSTM model can be optimized by combining edge computing and model optimization strategies.

5.5.4. Contribution of biomechanical features in English translation system

Biomechanical features are an important part of the multimodal translation system. This paper now counts the contribution of sub-category features in biomechanical features to the translation quality. The results are shown in **Figure 6**.

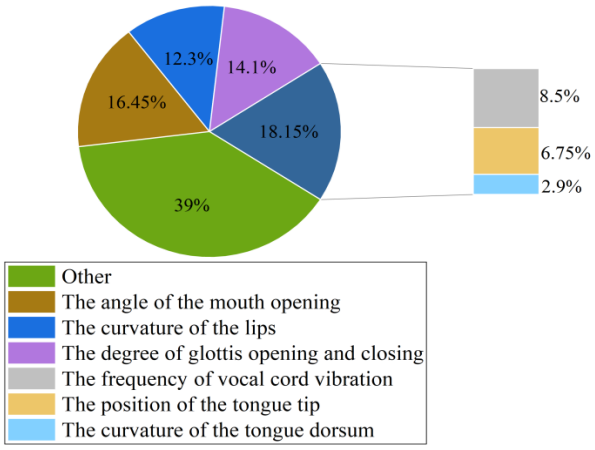

Figure 6. Contribution of biomechanical features in the English translation system.

In **Figure 6**, it can be seen that biomechanical features contribute 61% to the multimodal translation system, accounting for the majority, while the others only account for 39%. Among the biomechanical features, the angle of oral opening contributes the most, reaching 16.45%, the degree of glottal opening and closing reaches 14.1%, and the curvature of the lips reaches 12.3%. This shows that biomechanical features contribute greatly to the modal English translation system and promote the quality of English translation.

5.5.5. Contribution and role of audio and biomechanical features to emotional transmission

Different modalities are crucial to the expression of emotional transmission in translation systems. This paper takes a passage of English text as an example to statistically analyze the contribution and role of audio and biomechanical features to emotional transmission. The results are shown in **Figure 7**. In **Figure 7**, the sentiment intensity is obtained by using the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool, and the sentiment polarity score is calculated by using VADER based on sentiment vocabulary and context. For sentiment transfer loss, the experiment uses the sentiment polarity scores of the source text and the translated text to calculate statistics.

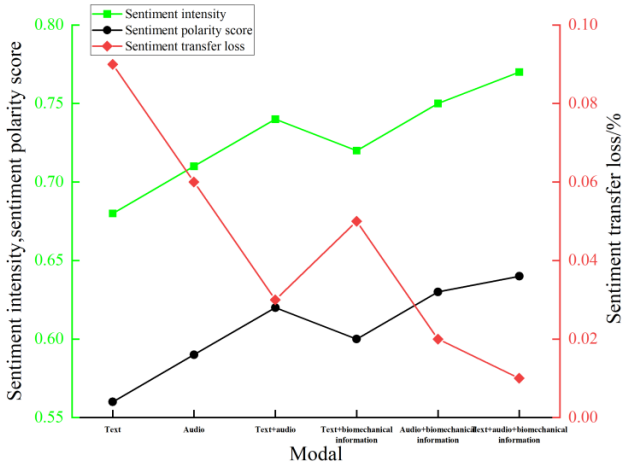

Figure 7. Analysis results of the contribution and role of audio and biomechanical features to emotional transmission.

The closer the emotional intensity and emotional polarity scores are to the original translation, the more similar the translated text is to the original translation. The emotional intensity of the original translation text is 0.78, and the emotional polarity score is 0.65. In **Figure 7**, text + audio + biomechanical information, reached 0.77 and 0.64, respectively, and the original translation of 0.78 and 0.65 are 0.01 different. In the single modality, the emotional intensity and emotional polarity scores of audio are 0.71 and 0.59, while the text reaches 0.68 and 0.56, respectively. It can be seen that audio signals have more advantages than text in expressing emotions. When two modalities are fused, text + audio reaches 0.74 and 0.62 respectively, and audio + biomechanical information reaches 0.75 and 0.63 respectively, both significantly better than single modality. This shows that the complementarity between modalities helps to capture emotional features more accurately, especially the combination of audio and biomechanical signals, which can enhance the delicacy of emotional expression.

In terms of emotional transfer loss, the richer the modality combination, the lower the loss. The emotion transfer loss of text + audio + biomechanical information is only 0.01, which is closest to the original translation. In the single modality, the emotion transfer loss of audio is 0.06, and that of text is 0.09. In the dual-modality combination, the loss of audio + biomechanical information is 0.02, text + biomechanical information is 0.05, and text + audio is 0.03. In summary, biomechanical information has a significant contribution to supplementing the emotion transfer of audio and text modalities, and it performs well in capturing subtle emotional changes.

6. Experimental discussion

The multimodal English translation system developed in this paper based on the Transformers-BiLSTM model is significantly better than single-modal and other comparison models in translation quality evaluation indicators such as BLEU, METEOR, and ROUGE-L. In multimodal fusion, the combination of text, audio, and biomechanical information performs best in all indicators. Multimodal fusion performs well in sentiment transmission analysis, can more accurately capture sentiment polarity and intensity, and has the highest sentiment match with the original translation. In this paper's multimodal English translation system, multimodal fusion can effectively capture translation context, semantic information, and pronunciation dynamics, improving translation fluency and semantic integrity. At the same time, the contribution of biomechanical features to the translation system is fully utilized, among which the angle of oral opening and the degree of glottal opening and closing significantly enhance the expressiveness and accuracy of speech.

The multimodal English translation system in this study demonstrates the potential of introducing biomechanical information and audio features to improve translation quality and emotional transmission. This significantly reduces the error rate of words and characters in the translated text, and provides new ideas for the study of the emotional accuracy of machine translation systems. The multimodal translation method provides a more natural and smooth solution for cross-language communication in complex scenarios, and is of great significance to the application of speech generation, intelligent learning platforms and multimodal interaction

technologies. The introduction of biomechanical information provides a quantitative basis for studying the impact of the dynamics of the vocal organs on language expression

7. Conclusions

This paper combines text, speech and biomechanical motion data, and adopts an English translation system based on multimodal features to significantly improve the performance of traditional single-modal translation systems in context understanding and emotional expression. The study integrates text, speech and biomechanical motion data into the BiLSTM model for time series modeling, and combines it with the multimodal Transformer architecture for translation modeling. The experimental results show that the system achieves the best performance in BLEU, METEOR and ROUGE-L indicators, with low emotional transfer loss, proving the effectiveness of biomechanical features in improving translation quality and maintaining emotional consistency. This study found that multimodal features have great potential in improving the performance of translation systems, but there are shortcomings in the cost of collecting biomechanical data and the real-time nature of data processing. In addition, the applicability and limitations in actual application scenarios still need to be further optimized. Factors such as different accents, speaking speeds, and language habits may affect the accuracy of biomechanical data acquisition and translation. In addition, real-time translation scenarios place higher demands on the system's data processing speed and translation speed, and the current system cannot fully meet these requirements. Future research can optimize data acquisition and processing algorithms, improve the system's real-time and adaptability, and better cope with challenges in actual application scenarios. Future work can focus on optimizing sensor equipment, improving data processing efficiency, and exploring the development of multimodal translation systems in more languages and cultural backgrounds, further promoting the intelligentization and popularization of translation technology.

Funding: This work was supported by The Research Capability Improvement Project of Young Teachers in Universities in Guangxi "Evaluation and Improvement Path of Language Service Competitiveness in Guangxi under the Framework of RCEP" (No. 2023KY0581).

Ethical approval: Not applicable.

Conflict of interest: The author declares no conflict of interest.

References

1. Su Y. Intercultural Communication and Language Conversion in Translation Studies. International Journal of Education and Humanities. 2023; 10(1): 186-189. doi: 10.54097/ijeh.v10i1.11116
2. Wang J. The Role of Sentiment Analysis in Machine Translation- A Review on the Example of ChatGPT. Science and Technology of Engineering, Chemistry and Environmental Protection. 2024; 1(9). doi: 10.61173/cqqm2k31
3. Zhang C, Yu T, Gao Y, et al. Design of a Smart Teaching English Translation System Based on Big Data Machine Learning. International Journal of Web-Based Learning and Teaching Technologies. 2023; 18(2): 1-14. doi: 10.4018/ijwltt.330144
4. Sitender, Bawa S, Kumar M, et al. A comprehensive survey on machine translation for English, Hindi and Sanskrit languages. Journal of Ambient Intelligence and Humanized Computing. 2021; 14(4): 3441-3474. doi: 10.1007/s12652-021-

03479-0
5. Ma W, Yan B, Sun L. Generative Adversarial Network-Based Short Sequence Machine Translation from Chinese to English. Ding B, ed. Scientific Programming. 2022; 2022: 1-10. doi: 10.1155/2022/7700467
6. Zhang H, Si N, Chen Y, et al. Improving Speech Translation by Cross-Modal Multi-Grained Contrastive Learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2023; 31: 1075-1086. doi: 10.1109/taslp.2023.3244521
7. Wu Y, Qin Y. Machine translation of English speech: Comparison of multiple algorithms. Journal of Intelligent Systems. 2022; 31(1): 159-167. doi: 10.1515/jisys-2022-0005
8. Meetei LS, Singh SM, Singh A, et al. Hindi to English Multimodal Machine Translation on News Dataset in Low Resource Setting. Procedia Computer Science. 2023; 218: 2102-2109. doi: 10.1016/j.procs.2023.01.186
9. Yang Y, Pan X, Wang R, Zheng Y. English Translation Strategies of Traditional Chinese Cultural Terms in Multimodal Context. China Terminology. 2023.
10. Gambier Y. Audiovisual translation and multimodality: what future?. Media and intercultural communication: a multidisciplinary journal. 2023.
11. Shi X, Yu Z. Adding Visual Information to Improve Multimodal Machine Translation for Low-Resource Language. Yousaf MH, ed. Mathematical Problems in Engineering. 2022; 2022: 1-9. doi: 10.1155/2022/5483535
12. Kent RD. The Feel of Speech: Multisystem and Polymodal Somatosensation in Speech Production. Journal of Speech, Language, and Hearing Research. 2024; 67(5): 1424-1460. doi:10.1044/2024_JSLHR-23-00575
13. Serrurier A, Neuschaefer-Rube C. Morphological and acoustic modeling of the vocal tract. The Journal of the Acoustical Society of America. 2023; 153(3): 1867-1886. doi: 10.1121/10.0017356
14. Mielke J, Hussain Q, Moisik SR. Development of a new vowel feature from coarticulation: Biomechanical modeling of rhotic vowels in Kalasha. Laboratory Phonology. 2023; 14(1). doi: 10.16995/labphon.9019
15. Zhao Y, Zhang J, Zong C. Transformer: A General Framework from Machine Translation to Others. Machine Intelligence Research. 2023; 20(4): 514-538. doi: 10.1007/s11633-022-1393-5
16. Rahali A, Akhloufi MA. End-to-End Transformer-Based Models in Textual-Based NLP. AI. 2023; 4(1): 54-110. doi: 10.3390/ai4010004
17. Gamage S, Lakshan K, Wickramarathna V, et al. A Multimodal Approach for Real-Time Sinhala Sign Language Translation. International Research Journal of Innovations in Engineering and Technology. 2023.
18. Liu H. A new double attention decoding model based on cascade RCNN and word embedding fusion for Chinese-English multimodal translation. International Journal of Reasoning-based Intelligent Systems. 2024; 16(1): 26-36. doi: 10.1504/ijris.2024.137429
19. Kumhar SH, Ansarullah SI, Gardezi AA, et al. Translation of English Language into Urdu Language Using LSTM Model. Computers, Materials & Continua. 2023; 74(2): 3899-3912. doi: 10.32604/cmc.2023.032290
20. Zhili W, Qian Z. A Deep Learning-based Method for Determining Semantic Similarity of English Translation Keywords. International Journal of Advanced Computer Science and Applications. 2024; 15(5). doi: 10.14569/ijacsa.2024.0150531
21. Gamal D, Alfonse M, Jiménez-Zafra SM, et al. Case Study of Improving English-Arabic Translation Using the Transformer Model. International Journal of Intelligent Computing and Information Sciences. 2023; 23(2): 105-115. doi: 10.21608/ijicis.2023.210435.1270
22. Li Y, Shan Y, Liu Z, et al. Transformer fast gradient method with relative positional embedding: a mutual translation model between English and Chinese. Soft Computing. 2022; 27(18): 13435-13443. doi: 10.1007/s00500-022-07678-5
23. Badawi S. Transformer-Based Neural Network Machine Translation Model for the Kurdish Sorani Dialect. UHD Journal of Science and Technology. 2023; 7(1): 15-21. doi: 10.21928/uhdjst.v7n1y2023.pp15-21
24. Liu S, Gao P, Li Y, et al. Multi-modal fusion network with complementarity and importance for emotion recognition. Information Sciences. 2023; 619: 679-694. doi: 10.1016/j.ins.2022.11.076
25. Li L, Tayir T, Han Y, et al. Multimodality information fusion for automated machine translation. Information Fusion. 2023; 91: 352-363. doi: 10.1016/j.inffus.2022.10.018
26. Zheng W, Gong G, Tian J, et al. Design of a Modified Transformer Architecture Based on Relative Position Coding. International Journal of Computational Intelligence Systems. 2023; 16(1). doi: 10.1007/s44196-023-00345-z
27. Guo Z, Hou Y, Hou C, et al. Locality-Aware Transformer for Video-Based Sign Language Translation. IEEE Signal Processing Letters. 2023; 30: 364-368. doi: 10.1109/lsp.2023.3263808

28. Tian T, Song C, Ting J, et al. A French-to-English Machine Translation Model Using Transformer Network. Procedia Computer Science. 2022; 199: 1438-1443. doi: 10.1016/j.procs.2022.01.182
29. Sameer M, Talib A, Hussein A, Husni H. Arabic Speech Recognition Based on Encoder-Decoder Architecture of Transformer. Journal of Techniques, 2023.
30. Xiang Y, Chen Y, Fan W, et al. Enhancing computer-aided translation system with BiLSTM and convolutional neural network using a knowledge graph approach. The Journal of Supercomputing. 2023; 80(5): 5847-5869. doi: 10.1007/s11227-023-05686-2
31. Safder I, Ali M, Aljohani NR, et al. Neural machine translation for in-text citation classification. Journal of the Association for Information Science and Technology. 2023; 74(10): 1229-1240. doi: 10.1002/asi.24817
32. Rao YSN, Chong YT, Khan RU, et al. Dynamic Sign Language Recognition and Translation Through Deep Learning: A Systematic Literature Review. Journal of Theoretical and Applied Information Technology. 2024.
33. Gourisaria MK, Agrawal R, Sahni M, et al. Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. Discover Internet of Things. 2024; 4(1). doi: 10.1007/s43926-023-00049-y
34. Biswas M, Rahaman S, Ahmadian A, et al. Automatic spoken language identification using MFCC based time series features. Multimedia Tools and Applications. 2022; 82(7): 9565-9595. doi: 10.1007/s11042-021-11439-1
35. Nayak SS, Darji AD, Shah PK. Machine learning approach for detecting Covid-19 from speech signal using Mel frequency magnitude coefficient. Signal, Image and Video Processing. 2023; 17(6): 3155-3162. doi: 10.1007/s11760-023-02537-8