Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 3389643, 12 pages
https://doi.org/10.1155/2021/3389643

*Research Article*

# Analysis of Japanese Expressions and Semantics Based on Link Sequence Classification

**Yanyan Shi[1] and Yuting Liang[2]**

[1] *School of Foreign Languages, Harbin University of Commerce, Harbin Heilongjiang 150028, China*
[2] *Department of Foreign Languages, East University of Heilongjiang, Harbin Heilongjiang 150066, China*

Correspondence should be addressed to Yanyan Shi; shiraohekaw@163.com

Received 17 July 2021; Revised 8 August 2021; Accepted 17 August 2021; Published 6 September 2021

Academic Editor: Yuanpeng Zhang

Based on locness corpus, this paper uses Wordsmith 6.0, SPSS 24, and other software to explore the use of temporal connectives in Japanese writing by Chinese Japanese learners. This paper proposes a method of tense classification based on the Japanese dependency structure. This method analyzes the results of the syntactic analysis of Japanese dependence and combines the tense characteristics of the target language to extract tense-related information and construct a maximum entropy tense classification model. The model can effectively identify the tense, and its classification accuracy shows the effectiveness of the classification method. This paper proposes a temporal feature extraction algorithm oriented to the hierarchical phrase expression model. The end-to-end speech recognition system has become the development trend of large-scale continuous speech recognition because of its simplicity and efficiency. In this paper, the end-to-end technology based on link timing classification is applied to Japanese speech recognition. Taking into account the characteristics of Japanese hiragana, katakana, and Japanese kanji writing forms, through experiments on the Japanese data set, different suggestions are explored. The final effect is better than mainstream speech recognition systems based on hidden Markov models and two-way long and short-term memory networks. This algorithm can extract the temporal characteristics of rules that meet certain conditions while extracting expression rules. These tense characteristics can guide the selection of rules in the expression process, make the expression results more in line with linguistic knowledge, and ensure the choice of relevant vocabulary and the structural ordering of the language. Through the analysis of time series and static information, we combine the time and space dimensions of the network structure. Using connectionist temporal classification (CTC) technology, an end-to-end speech recognition method for pronunciation error detection and diagnosis tasks is established. This method does not require phonemic information nor does it require forced alignment. The extended initials and finals are the error primitives, and 64 types of errors are designed. The experimental results show that the method can effectively detect the wrong pronunciation, the detection accuracy rate is 87.07%, the false rejection rate is 7.83%, and the error rate is 87.07%. The acceptance rate is 25.97%. This method uses network information more comprehensively than traditional methods, and the model is more effective. After detailed experiments, this article evaluates the prediction effect of this method and previous methods on the data set. This method improves the prediction accuracy by about 15% and achieves the expected goal of the work in this paper.

## 1. Introduction

Statistical language expression is one of the challenging frontier topics in the field of natural language processing, which has a wide range of application value and important commercial application prospects [1]. In recent years, statistical language expression technology has developed rapidly, and a series of impressive results have been achieved. However, in practical applications, how to effectively use linguistic knowledge in statistical language expression models to improve the quality of expression is still a research hotspot [2]. At present, in the statistical machine expression, the research on tense is mainly limited to the aspect of tense recognition, and there are few studies on the expression of tense

[3]. Temporal information is important linguistic information, so the tense problem studied in this paper is transformed into a problem of incorporating tense and other linguistic knowledge into statistical expressions [4]. Driven by many applications, the research on link prediction has achieved fruitful results. At present, a method based on node similarity is widely used, and the possibility of link generation is predicted by the size of the similarity score [5]. In the static method, the network changes over time are ignored. If only the network diagram under the most recent time snapshot is used, when the network changes frequently, the prediction effect will drop sharply [6]. With the development of the Internet, there are more and more extensive scenarios where links occur repeatedly, and the evolution of networks is becoming more and more common. Static link prediction methods are far from being able to adapt to the needs of the new situation. Therefore, in recent years, information has gradually gained attention [7].

In recent years, with the rapid development of the Internet at home and abroad and the integration of the world economic market, the amount of network data information has increased sharply, international exchanges have become more frequent, and the language expression market is broad [8]. Language expression, as an important way to overcome language communication barriers, has a profound impact on the promotion of political, economic, cultural, and military exchanges between countries [9]. Language expression is a powerful guarantee for active and healthy exchanges and dialogues between the two countries. However, traditional manual expression is inefficient and costly. For the huge number and increasing number of expression tasks, it can no longer meet the needs of society and the market [10]. Language expression is the automatic expression of text, which is an important content in text processing, and most applications in text processing need to reason and filter according to the temporal relationship of text, such as time extraction and automatic summarization, and tense can provide them important clues, so tense plays an indispensable role in these applications [11]. Similarly, the correct expression of the tense in the text can convey the information of the source language as accurately as possible, but different languages have greater differences in the expression of tense, which is a great challenge to the expression of the tense in terms of language [12]. At the technical level, the existing language expression methods are still mainly limited to the use of rules to solve tense problems. These methods are inefficient and costly. Even statistical language expression methods cannot solve language well [13].

This article takes Japanese as the research object and studies the expression of tense from the perspectives of Japanese-Chinese. Japanese belongs to the cohesive language family, and its tense is determined by the deformation of the predicate ending, and the changes of the predicate ending are various. There are similar endings in different simultaneous expressions, which leads to the low accuracy of the tense expression of statistical expressions. In response to the above problems, this article proposes a statistical expression method that incorporates tense characteristics. The difference between a recursive network and a feedforward network lies in this feedback loop that constantly uses its own output at the last moment as input. The purpose of adding memory to the neural network is the sequence itself contains information, and the recursive network can use this information to complete tasks that the feedforward network cannot complete. This sequence information is stored in the hidden state of the recursive network and is continuously passed to the previous layer, spanning many time steps, affecting the processing of each new sample. Human memory will continue to cycle invisible in the body, affecting our behavior without showing a complete appearance, and information will also circulate in the hidden state of the recursive network. This method uses a deep control gating function to connect multilayer LSTM units and introduces linear correlation between the upper and lower layers in the recurrent neural network, which can build a deeper voice model. At the same time, we use the training criteria of linking time series classification for model training. We build an end-to-end speech recognition system to solve the hidden Markov model's need to force the alignment of labels and sequences and use CTC training criteria to realize an end-to-end speech system, by combining with the traditional LSTM-CTC model. The model is compared to verify the effectiveness of the deep LSTM neural network [14, 15] in speech recognition.

## 2. Related Work

Rule-based expression technology is highly dependent on humans. It mainly relies on linguists to summarize rules and manually compile the rules. The workload is large, and problems such as rule conflicts are prone to occur, and it is difficult to cope with large-scale expression tasks. The corpus-based method pays more attention to automatically obtaining rules from a large-scale corpus, which has the advantages of language independence and automatic knowledge acquisition, which greatly improves the efficiency of expression. With the increase of the amount of network information, the rule-based method has become more and more limited, and the corpus-based method has more and more obvious advantages, has developed rapidly, and has achieved a series of remarkable results [16]. Allen et al. [17] proposed an expression model based on word alignment, which entered the development stage of language expression and marked the birth of modern statistical expression methods. McCune [18] proposed the phrase expression model, which greatly improved the expression effect. This phrase expression model used a logarithmic linear model and its weight tuning method, which has made a great contribution to the expression performance improvement of the phrase expression model. At the same time, due to its strong ease of use and other advantages, the phrase expression model has begun to attract wide attention from all walks of life. In practical applications, the online translation systems of large domestic and foreign Internet companies such as Baidu and Google all use this model as backend support, which has certain commercial value. However, this model uses phrases as the basic processing unit, and its ability to adjust order is limited, and the quality of long sentences is not good, which needs further improvement. Izard et al. [19] introduced the concept of generalized variables. A hierarchical phrase expression model was further proposed, which used hierarchical phrase rules to achieve expression, which made the entire expression process more hierarchical, to a certain extent, alleviates the problem of insufficient global ordering ability in the phrase expression model, and solved other problems of ordering. The method of the problem has a certain guiding effect. As an expression model based on formal grammar, this model introduces more effective information for the expression process. Similarity can be described by many methods. According to the type of information used, link prediction mainly includes similarity methods based on network topology and similarity methods based on node attributes. The network topology can be divided into local information and global information. Asahara and Matsumoto [20] also considered the number of neighbors in common neighbors. Later, these neighbor-based methods were also extended, not only considering whether there is a link between neighbors but also the number of links, that is, the addition based on the basic neighbor method. In addition to neighbor-based methods, global information was also often used for link prediction, such as path information between nodes. Another way to consider global information is the random walk model. This model can be considered a general expression based on the neighbor method and the path-based method, which can better reflect the network topology information. The similarity method based on node attributes is rarely used solely for link prediction. Because on the one hand, node attributes are often used in specific types of networks, and the processing is relatively complicated; on the other hand, node attributes may be subjective and sometimes not as reliable as the network topology. However, node attributes and network topology, respectively, reflect information from different aspects of the network, and combining the two can often achieve a better effect. In addition to node attributes and network topology, community information has recently been shown to be helpful for link prediction. At present, various types of information in networks such as local information, global information, and community information are often targeted at specific fields, and the relatively comprehensive and comprehensive similarity score model is not very satisfactory.

Since tenses have different and complex expressions in different languages, it is very difficult to maintain the same tense information from the source language to the target language in terms of expression. In the machine expression system based on intermediate language conversion [21], the tense information of the source language was first converted into language-independent abstract expressions. For example, in the study of Chinese-Japanese tense expression, Cheng et al. [22] proposed Lexical Conceptual Structures (LSC) based on two levels of knowledge expression are used to assist language expression, combining tense and lexical semantics, and through some rule transformations in the expression process, it can generate Japanese sentences with correct tenses for Chinese sentences expression. However, the above research is based on a language expression system based on intermediate language conversion. It is a rule-based language expression. It is not only highly subjective but also highly dependent on language types, and it is difficult to expand. It is also significantly different from the existing SMT system. On this basis, others jointly build a temporal model with the context of the target end. First, they syntactically analyzed the bilingual parallel corpus and used the syntactic analysis result to automatically extract the temporal information contained in the sentence, and then according to the temporal continuity and text classification, a temporal model based on the classifier was proposed. In the SMT system, someone proposed a corpus-based method to study Japanese and Chinese tense expression, but mainly demonstrated that bilingual corpus contains rich tense information, which can help solve the problem of Japanese and Chinese tense expression. On this basis, we can consider integrating bilingual tense information to solve the problem of tense expression [23]. In the statistical language expression system, on the one hand, there is less research on the expression of tense. On the other hand, the expression research on the integration of linguistic knowledge is in its infancy. Various methods need to be further improved. The full use of linguistic knowledge in the research of machine translation has important value and also has important guiding significance for the research of tense in this article.

## 3. Japanese Language Expression and Semantic Model Construction Based on Link Timing Classification

*3.1. Link Timing Classification.* Time series has achieved good results in describing time information. There are two main ways to represent the network information of each time period in history as discrete time series diagrams and link predictions. Link timing classification (CTC) is mainly used to deal with timing classification tasks, especially when the alignment result between the input signal and the target label is unknown. Link timing classification technology can perform label prediction at any point in the entire input sequence, which solves the problem of forced alignment in traditional speech recognition. The criterion for neural network training by linking time series classification technology is called the CTC criterion. Figure 1 shows the hierarchical distribution of link timing classification.

The existing information of the static method is a graph. According to the link between nodes $x$ and $y$ and other nodes, the similarity scores of $x$ and $y$ are calculated to predict the current unlinked node pair $(x, y)$ will be generated in the next time period possibility of linking.

$$X[t] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cdots & 0 \\ 0 & 0 & t \end{bmatrix}, \tag{1}$$

$$Y(x) = x(1) + 2 \times x(2) + \cdots + t \times x(t). \tag{2}$$

One is the time series of the number of links between nodes, which only predicts the future link situation based on the past links between nodes, and achieves similar results

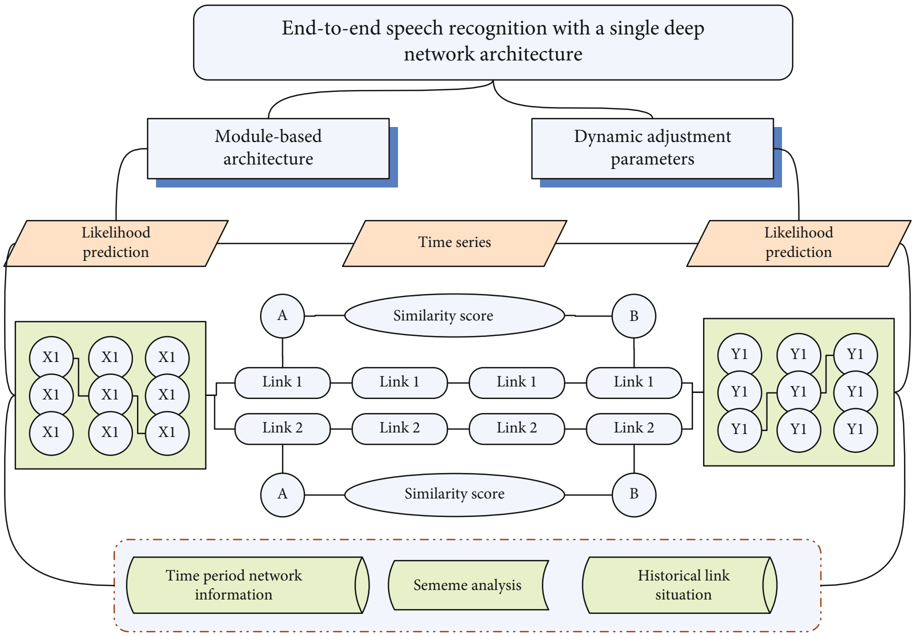
End-to-end speech recognition with a single deep network architecture
Module-based architecture
Dynamic adjustment parameters
Likelihood prediction
Time series
Likelihood prediction
A
Similarity score
B
X1
Y1
Link 1
Link 2
A
Similarity score
B
Time period network information
Sememe analysis
Historical link situation

Figure 1: Hierarchical distribution of link timing classification.

to the static method. Combining it with the static method can further improve the prediction effect.

$$C(t, s) = \{t \in A : s(t) = t * (t-1), t \longrightarrow N\}. \tag{3}$$

For new links, because the time series of link times is lost, the hybrid model is downgraded to a static similarity method; in addition, the hybrid model multiplies the final static method prediction value with the time series prediction value, which makes it difficult to describe the network in each time period.

$$z = \begin{cases} i \times (x(t) + x(t-1)), t > i, \\ i \times (x(t) - x(t-1)), t < i, \end{cases} \tag{4}$$

$$H[x] = [x(1) \cdots x(t)] \times \begin{bmatrix} x(t) \\ \cdots \\ x(1) \end{bmatrix}. \tag{5}$$

The existing information of this method is from the first time period in history to the current time period $t$, a total of $t$ pictures. According to $(x, y)$ historical link situation (including the bold dashed line at time $t$), we establish a unary time series model to predict the link situation in the next time period. The probability number of links is greater than 0 is the link probability.

$$p(t \mid x) - \prod_{t=1}^{T} H(x) \times (x(t) - \bar{x}) = 0, \tag{6}$$

$$\frac{\partial \ln p(z \mid x)}{\partial x} \times p(z \mid x) \times H(x) = 1. \tag{7}$$

However, because the model is too simple and fails to describe the relationship between the similarity score and the number of links, the changing laws of the two are different, and the results obtained by the mixed model are not as good as just using the similarity score time series.

*3.2. Temporal Feature Extraction.* Time series is a sequence of ordered data recorded in chronological order. We observe the time series in order to find the law of its development and change, and then, we select a suitable model to fit the observations to complete the prediction of the future value based on the model. The application of time series analysis and forecasting methods is very rich, such as forecasting the load of the power system and the price of a stock in the stock market.

Figure 2 shows the flow of sentence temporal feature extraction. The first part is the input layer, which accepts the acoustic characteristics of the input. As a branch of mathematical statistics, time series has its own set of analysis and prediction methods. In the unary time series model, in order to predict the future value, the historical value of the series itself is the research focus. Different from the unitary time series, in addition to its own influence, the changing laws of some series are also related to other series. The next is the batch normalization layer and the zero-filling layer. The reason for adding the zero-filling layer is to ensure batch processing with the same length; the second part is convolution, which contains 10 CNN layers, 5 maxpool layers, then the batch normalization layer, and the dropout layer. The function of the dropout layer is to prevent overfitting and improve generalization ability during training; the third part is the fully connected layer (dense layer), each neuron in the fully connected layer is connected with all neurons in the previous pooling layer, and the fully connected layer integrates the classification features of the convolution and

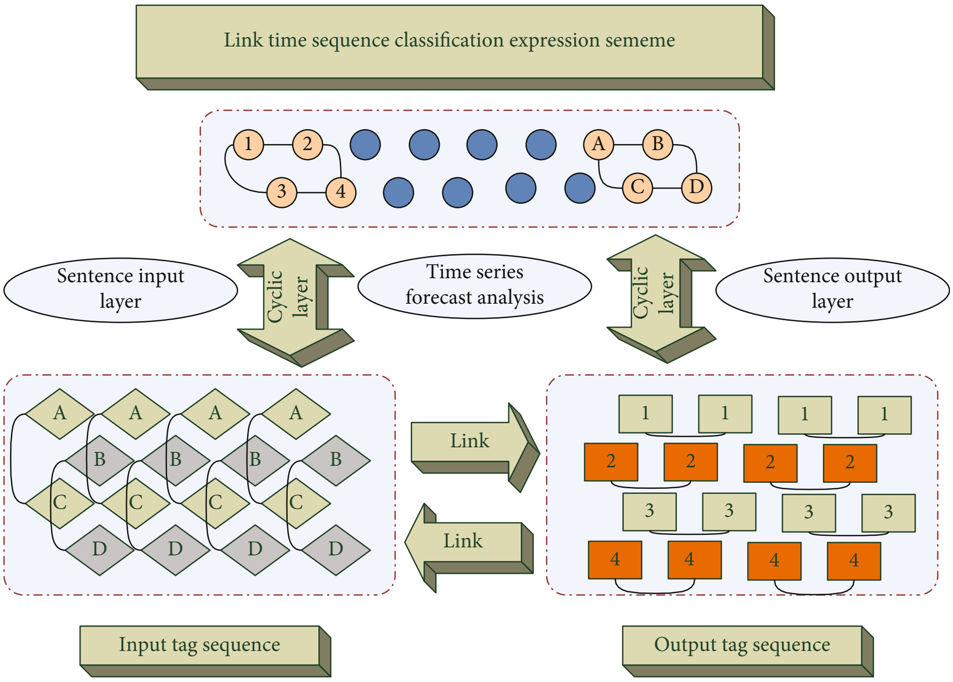

Figure 2: Sentence temporal feature extraction process.

pooling layers and distinguishes them. The activation function of each neuron uses the linear rectification function. The output value of the last layer is passed to the softmax logistic regression for classification. Finally, the CTC output layer is used to generate the predicted phoneme sequence. This method also tries to compare the similarity score calculated by the entire network between nodes and the actual occurrence between nodes. Combining the number of links, the hybrid model normalizes the similarity score of each time period and adds the number of links, which is used as the input of the time series.

*3.3. Phrases and Semantic Fusion.* The language model of the encoder-decoder structure [24] integration is only reflected in the prediction of the current moment by using the prediction of the previous moment during decoding. There are certain limitations and cannot make full use of linguistic information. Therefore, in order to explicitly introduce it in the decoding process in the language model, we further improve the performance of speech recognition. Based on the transformer structure, this article removes the decoder part and combines the encoder with the link timing classification as the end-to-end model used in this article. The main body of the entire model is composed of several identical coding layers stacked, and each coding layer is divided into two parts: multihead attention [25, 26] and feed-forward network. The main method of current speech recognition is to train the acoustic model by combining the recurrent neural network (RNN) and its variants with the hidden Markov model. The cyclic neural network uses the past information to input the output of the hidden layer at the last moment into the hidden layer at the current moment, retaining the previous information. As a time series, the speech signal has a strong context dependence, so the recurrent neural network is quickly applied to speech recognition. Theoretically, RNN can handle arbitrarily long sequences, but due to the disappearance of the gradient, the RNN cannot use the information at a longer time. A residual connection is added after each part, and then, layer normalization is performed. Figure 3 shows the three-dimensional histogram of the residuals in the coding layer of the language model. In addition, the model also includes a downsampling and upsampling module, a position encoding module, and an output layer. The output adopts the link timing classification criterion and predicts a label for each frame of speech to minimize the CTC loss function.

The "multihead attention" module is composed of several identical layers stacked, and each layer is a self-attention mechanism that uses scaled dot-product attention (scaled dot-product attention). Self-attention is a mechanism that uses the connection between different positions of the input sequence to calculate the input representation. Specifically, it has three inputs, queries, keys, and values, which can be understood as the speech feature after encoding. The output of query is obtained by the weighted summation of value, and the weight of each value is calculated by the design function of query and its related key. The "multihead" mechanism is used to combine multiple different layers. The self-attention representation of that is calculated, and $h$ represents the number of "heads". Figure 4 shows the fusion of linguistic expressions and sememe characters. Combining the characters in the word with spaces as separators to form character pairs, we combine word frequency and count all character pairs and their appearance frequency. We select the pair of characters with the highest frequency, remove the spaces in the middle, and merge them. The resulting character sequence is used as a new symbol to replace the original character pair.

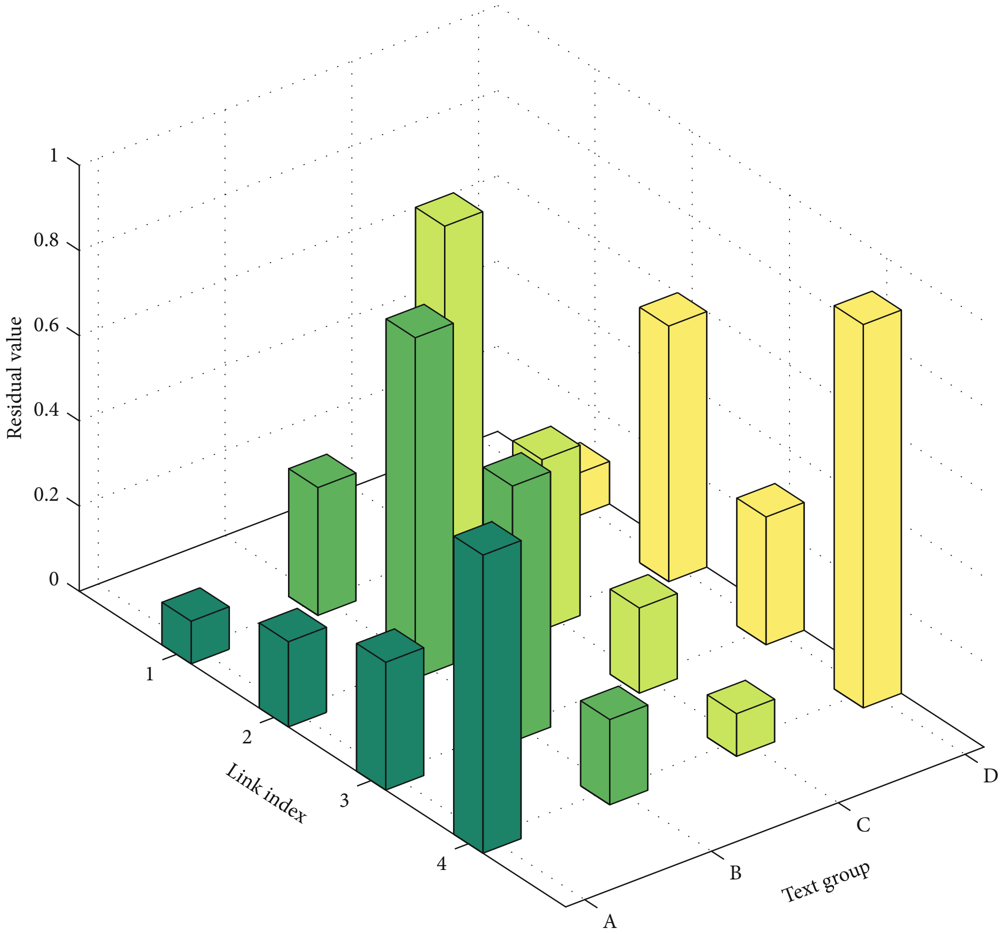

Figure 3: Three-dimensional histogram distribution of residuals in the coding layer of the language model.

Then, we repeat the two steps of counting the frequency of character pairs and merging the most frequent character pairs until the number of merging reaches the specified number. We output the merged character pairs according to the statistical frequency from high to low, then use the character pair list generated in the previous section. The corpus that needs to be processed is segmented. When segmenting is performed, the same word is used as the unit, and the characters are first segmented. Then, the character pairs are merged in the order of the frequency of appearance in the character pair list from high to low. Finally, it is retained that does not appear in units in the list of character pairs and individual characters not participating in the merging.

## 4. Application and Analysis of Japanese Language Expression and Semantic Model Based on Link Timing Classification

*4.1. Link Timing Score Prediction.* In the LSTM-based end-to-end speech recognition system, we use a bidirectional long and short-term memory network to model the timing of speech features, and the output uses the link timing classification criterion to directly predict the label sequence. In this experiment, we use 3 hidden layers. They are combined with a LSTM network with 1024 hidden nodes in each layer and 108-dimensional filterbank features for acoustic model training. The modeling unit uses ub-word units. The 3-gram language model is combined when decoding. Figure 5 shows the sentence level accuracy deviation box type figure. In recognition, in order to use the knowledge of Japanese linguistics to further improve the recognition performance, we combined the traditional language model when decoding and unified the dictionary, language model, and acoustic model to decode, and the system recognition performance was significantly improved. On the Japanese data set, using the algorithm recommended in this article, the recognition performance is significantly better than the current mainstream hybrid system based on the hidden Markov model and the end-to-end model based on the bidirectional long and short-term memory network. In order to reduce the memory occupied during model training and speed up the training, we first downsample the original speech feature frame and, then, encode it through the linear layer. At the same time, because the output adopts the CTC criterion, it is necessary to predict a label for each frame of speech, so the speech feature is performed before the output layer One-step upsampling operation restores the length of the speech frame to the original length. In this way, while speeding up the calculation and reducing memory requirements, it

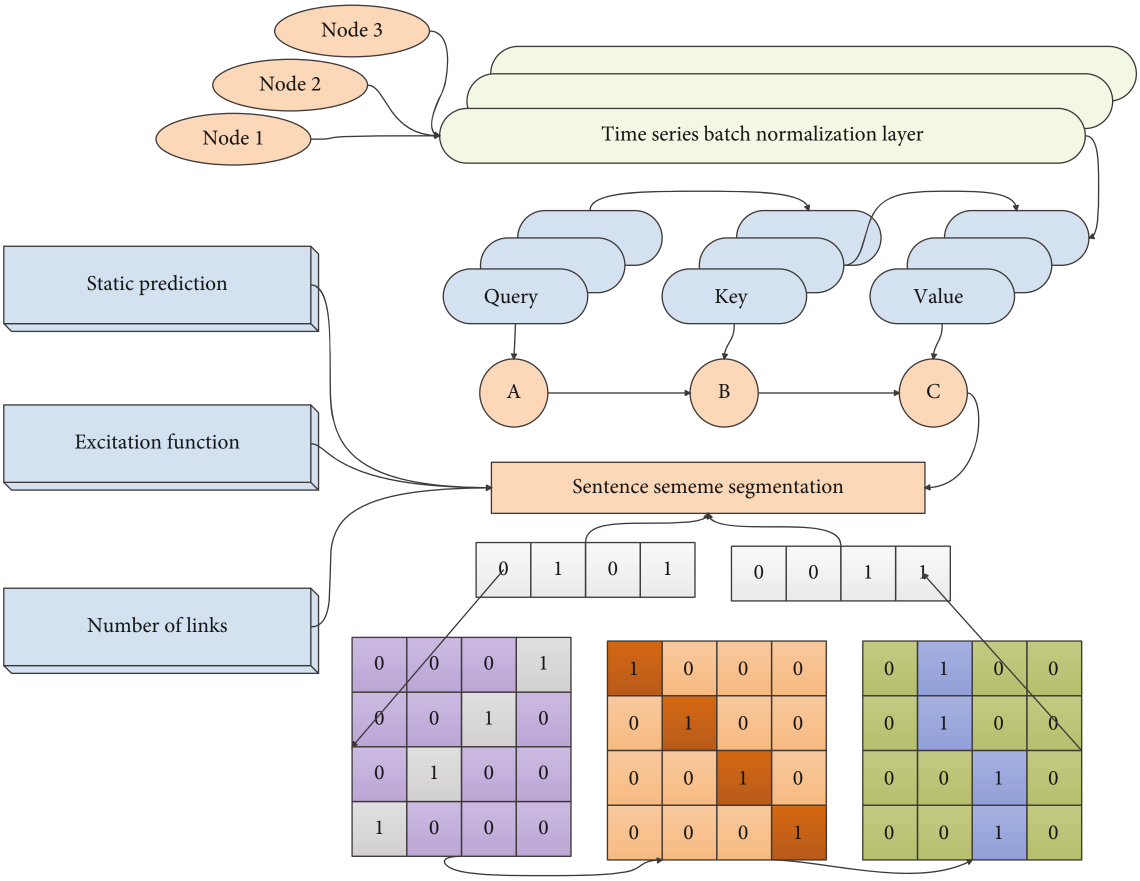

Figure 4: The fusion of expressions and semantic characters.

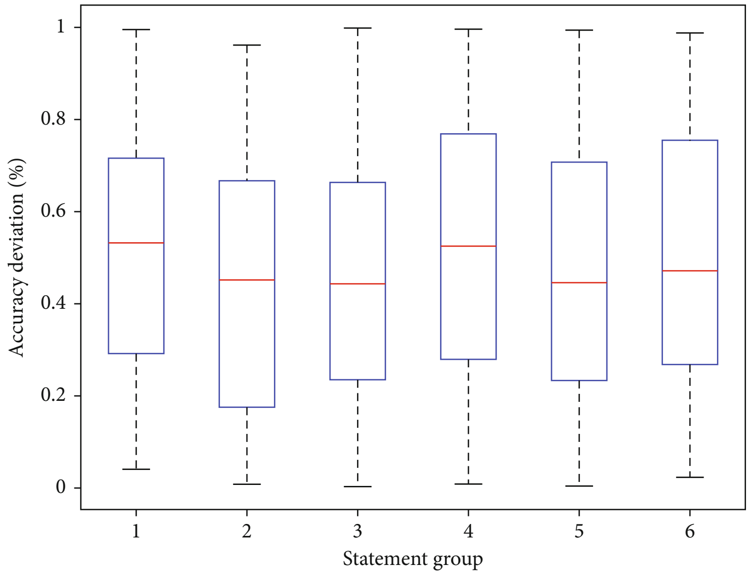

Figure 5: Box plot of sentence level accuracy deviation.

also ensures the accuracy of the model and improves the recognition performance.

For the CTC-based end-to-end model described in this article, in the experiment, we use 6 layers of coding, the "multihead" attention part uses 8 "heads," and the original speech features are 108-dimensional filterbank features. For downsampling, the dimension after encoding is 512 dimensions, and then, the position-coding information is added as the input of the network. The feedforward network dimension uses 1024 dimensions. Algorithm utilization for the idea

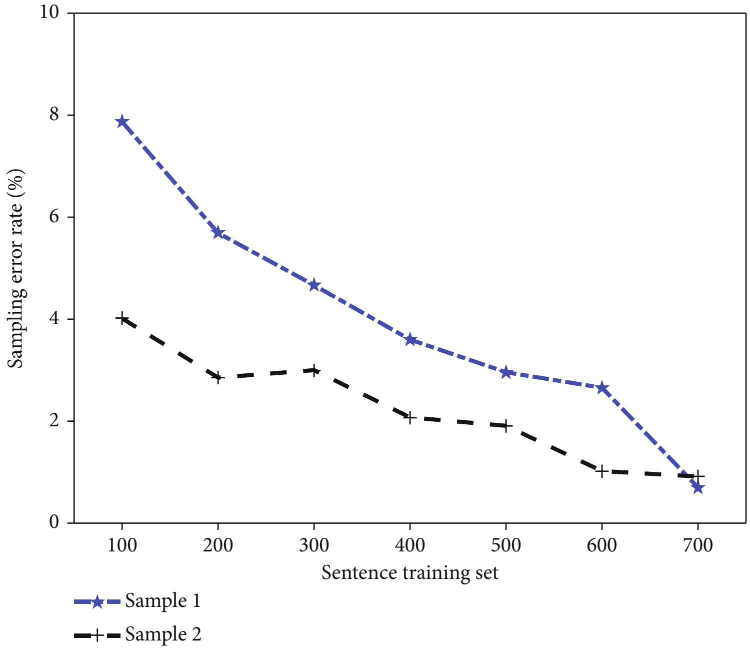

FIGURE 6: Line chart of comparison of term data recognition error rate.

of the greedy algorithm is an efficient data compression algorithm. In the process of data compression, the BPE algorithm searches for the byte pair that appears most frequently in the program code on the target memory (RAM) page; the single-byte token that appears in the specific code replaces the byte pair. In order to compare with the traditional HMM-based Mandarin speech recognition method, the HMM-based Mandarin recognition is performed under the same data set, in which the acoustic features use 39-dimensional MFCC features (dimensional cepstral coefficient features, dimensional energy feature, and its first and second order differences), the word error rate based on monophones in the experimental results is 50.9%, which is worse than the results in the text, indicating that the method in the text is superior to the monophone based on HMM in Mandarin speech recognition method. The method in the text does not require the use of pronunciation dictionaries and language models, which simplifies the implementation process of the speech recognition system. We repeat these two steps until all byte pairs are replaced or no byte pairs appear with a frequency greater than 1, and finally, the compressed data is output. With the dictionary containing all the replaced byte pairs, the dictionary is used to restore the data.

*4.2. Expression and Semantic Simulation.* This article conducts experiments on the King-ASR-450 data set. The database collects 79,149 voice data in a quiet environment, which is 121.3 hours long. The dictionary formed by the transcribed text contains a total of 66027 Japanese words. All voice data are 8 KHz sampling rate, 16bit, single-channel format. In the experiment, 76.6 k voice data (117.45 h) were selected as the training set, 0.5 k voice data (0.77 h) were used as the development set, and 2.0 k speech data (3.07 h) are as the test set. This paper uses them as the experimental platform to compare the experimental effects under different models and explore the improved the impact of the unit on the recognition performance. In the experiment, we built three systems for comparison. The first is the mainstream LSTM-HMM-based baseline system, and the second system is the LSTM-based CTC end-to-end recognition system, and the third is the CTC end-to-end recognition system based on LSTM. The system is the SA-CTC end-to-end recognition system proposed in this paper. Figure 6 shows the comparison of the word data recognition error rate.

In the experiment, 39 Mel-Frequency Cepstral Coefficients (MFCC characteristics) are used as the input signal of the GMM-HMM hybrid system. In the GMM-HMM system, the final number of bound states is 3334 through Gaussian splitting and decision tree clustering. The model forcibly aligns the training data to obtain frame-level labels, which are used as training data for the subsequent neural network. In the LSTM-HMM training, 108-dimensional filterbank features are used for training, and 40 frames of speech data are used before and after the current frame to obtain the before and after information. The network has 3 hidden layers, the hidden layer nodes are 1024, and the modeling unit is the bound 3334 states. Considering the impact of resource sparseness on the experimental results, we use a pronunciation dictionary and use phonemes as modeling units to conduct experiments. There are 237 phonemes in the data set. After blank is added, the output node of the network is 238, and the frequency of each phoneme is counted. The relative balance is better, and the trained model should be more robust. The 3-gram language model is combined when decoding.

In order to evaluate the performance of the pronunciation error detection system, the article refers to the hierarchical evaluation structure developed in the literature, and

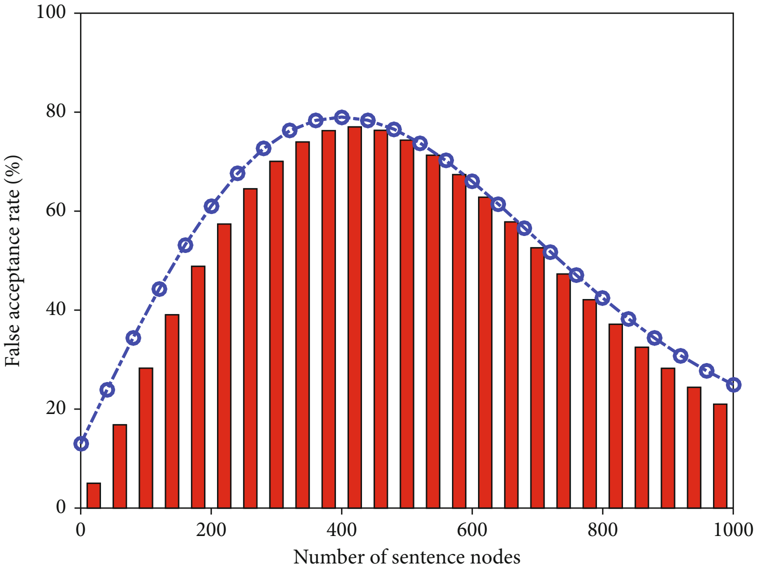

Figure 7: The histogram distribution of sentence node detection error acceptance rate.

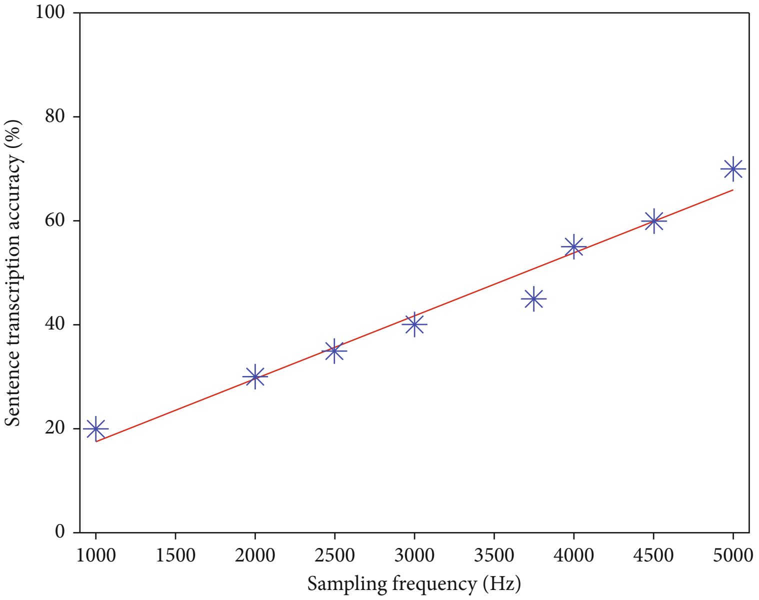

Figure 8: Linear fitting of sentence transcription rate under different average sampling frequencies.

design evaluation indicators. There are 4 kinds of test results in the experiment: correct acceptance (AT), correct rejection (RT), false rejection (RF), and false acceptance (AF). Figure 7 shows the histogram of the false acceptance rate of sentence node detection. According to these 4 detection results, the performance of the system is measured by the false acceptance rate (RFA), false rejection rate (RFR), and correct diagnosis (AD). RFA indicates that the learner's incorrect pronunciation is detected by the system as the correct pronunciation percentage; RFR represents the percentage of learners' correct pronunciation detected by the system as incorrect pronunciation; AD is the correct rate of system diagnosis, that is, the system's detection results are consistent with the labeled results. In order to analyze their Mandarin pronunciation errors in more detail, these 64 kinds of errors are classified and counted.

*4.3. Example Application and Analysis.* In this paper, the pronunciation characteristics of Japanese and Chinese are compared, and a corpus of students' Mandarin pronunciation errors is designed. Figure 8 shows the linear fitting of sentence transcription rates under different average sampling frequencies. The corpus is recorded under silent office conditions with a microphone and smartphone. Its average sampling rate is 3,000 Hz and the sampling size is 16 bits. The sentences in the corpus (text prompts in the recording) are everyday words and cover all syllables. People participating in the recording are young students who have serious

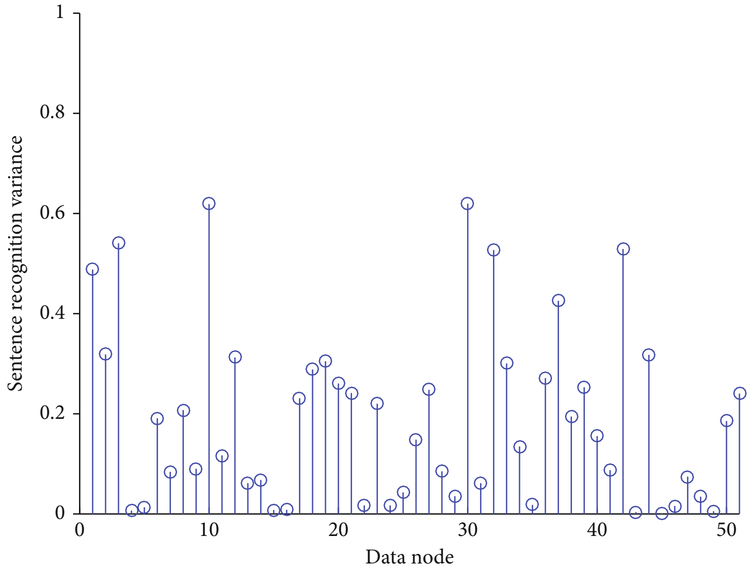

Figure 9: Match stick graph of training corpus identification variance of different nodes.

accents in their pronunciation. The recorded corpus is cross-labeled by 10 graduate students majoring in phonetics. When there are inconsistencies, phonetics experts will be asked to judge them. This corpus is only used for testing.

In the DNN-based acoustic model modeling process, 75-dimensional fiherbank features are used, and the network trained with 10000 h Japanese corpus is used as the initial network, which effectively avoids the local optimal solution in network training. The DNN model [27–29] includes 6 hidden layers, each layer includes 2,048 nodes, and the output layer includes 6,004 nodes. In order to improve the distinguishing degree of the DNN model, the input layer adopts the framing operation, and the input nodes are 825 (11 × 75) nodes. In addition, the HMMs acoustic model is forced to align the training data to obtain frame-level annotations for DNN training [30, 31].

In order to further verify that when the training corpus is sufficient and the modeling unit does not have sparseness, the CTC criterion is better than the CE criterion. This article conducts experiments on the Japanese language full database provided by the NIST2015 Keyword Search (OpenKWS) competition. The training set is about 40 hours, and the test set is 1 h. Figure 9 shows the training corpus identification variance matchstick graph of different nodes. The initial network of DNN training is also for the 10000 h Japanese training network. The last layer is cut off, and it is initialized to 3054 nodes randomly. The sequential connectives used by learners and native speakers are mainly concentrated on sequential connectives: connectives that show order and sequence. There are 50 sequential connectives in the corpus, accounting for 65.86% of the sequential words used by learners; there are 30 sequential connectives, accounting for 57.97% of the sequential words used by native speakers, especially in the top four frequently used words. The LSTM training network is exactly the same as the aforementioned Japanese initial network except that the last layer becomes 3054 nodes. It can be seen that the decoding efficiency required by the end-to-end acoustic model is higher than that of the traditional hidden Markov-based acoustic model, which is increased by about 50%. This is mainly because in the HMM-based acoustic model decoding process, in addition to the dictionary and language model, the HMM model is also packaged into a WFST network, which greatly increases the search space for decoding. In the CTC acoustic model, the HMM model is no longer needed. In addition, it can be seen that when the modeling unit is a Japanese word, although the recognition result is poor, the decoding speed is the fastest. This is because the word is used as the modeling unit and even a dictionary is not needed, so the WFST network only contains language models. Compared with the acoustic model whose modeling unit is tri-phone, the decoding search space based on the end-to-end acoustic model is smaller, and the decoding speed is faster.

In Japanese end-to-end tri-phone acoustic modeling, the output layer of the LSTM network contains 3596 nodes (3595 triphone and 1 blank). In addition to the output layer, the initial network is similar to that. The CTC initial network of the Japanese experimental part is exactly the same. Figure 10 shows the pie chart of the recognition percentages of the endpoints at different levels of the model. It can be seen that the recognition performance of the recognition system based on the end-to-end model is significantly better than the hybrid system based on the hidden Markov model, and the recognition accuracy is above 90%. At the same time, the recognition based on the end-to-end model of the SA-CTC effect is better than the end-to-end model based on LSTM-CTC, with an accuracy rate of 91.35%.

## 5. Conclusion

This paper proposes a method of integrating tense characteristics in statistical expressions. This method realizes the selection and filtering of rules of different tenses without increasing the complexity of the decoder. And there is no

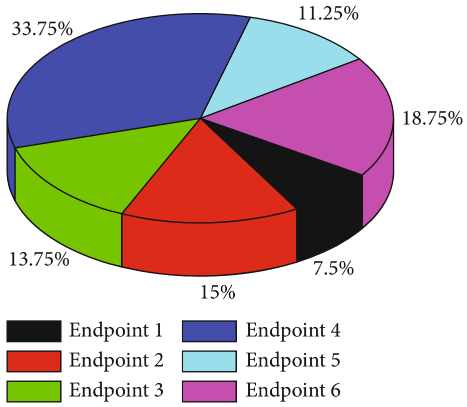

Figure 10: Fan chart of the recognition proportions of endpoints at different levels of the model.

dependence on language, only need to choose to integrate monolingual tense features or bilingual tense features according to the difference of language grammar. In the end, the recognition performance of the SA-CTC-based end-to-end model surpasses the HMM-based hybrid model and the BiLSTM-based end-to-end model, and the recognition accuracy reaches 91.35%. This paper studies the end-to-end technology based on the self-attention mechanism and link timing classification and builds a complete speech recognition system on the Japanese data set. At the same time, according to the characteristics of Japanese large vocabulary, the algorithm is introduced, and the subword unit is used as Japanese recognition modeling unit. The experimental results of Japanese-Chinese and Japanese-Japanese expressions show that the method proposed in this paper can not only effectively improve the tense expression accuracy of the hierarchical phrase model but also achieve the purpose of word sense disambiguation and improvement of sentence structure adjustment. In response to the above shortcomings, this paper proposes a new link prediction method SOTS (Similarities and Occurrences Time Series) based on the combination of node similarity and link times. First, calculate the similarity score between the nodes in each time period through a trending random walk and, then, use the time series model to combine it with the actual number of links between the nodes in each time period to predict the occurrence of each node pair in the next time period possibility of linking. Through two combined time series models, this paper studies the relationship between the similarity scores between nodes and the actual number of links. This method can be used to predict new links and recurring links in the evolving network in the future.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

All the authors do not have any possible conflicts of interest.

## References

[1] H. Brock, I. Farag, and K. Nakadai, "Recognition of non-manual content in continuous Japanese sign language," *Sensors*, vol. 20, no. 19, p. 5621, 2020.

[2] M. Asahara, S. Kato, H. Konishi, M. Imada, and K. Maekawa, "BCCWJ-TimeBank: temporal and event information annotation on Japanese text," *Journal of Computational Linguistics & Chinese Language*, vol. 4, pp. 20–24, 2019.

[3] N. Laokulrat, M. Miwa, Y. Tsuruoka, and T. Chikayama, "Uttime: temporal relation classification using deep syntactic features," *Lexical and Computational Semantics*, vol. 3, pp. 88–92, 2019.

[4] R. Bansal, M. Rani, H. Kumar, and S. Kaushal, "Temporal information retrieval and its application: a survey," in *Emerging Research in Computing, Information, Communication and Applications*, pp. 251–262, Springer, Singapore, 2019.

[5] J. Pustejovsky, R. Knippen, J. Littman, and R. Saurí, "Temporal and event information in natural language text," *Language Resources and Evaluation*, vol. 39, no. 2, pp. 123–164, 2020.

[6] M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, J. Moszkowicz, and J. Pustejovsky, "The TempEval challenge: identifying temporal relations in text," *Language Resources and Evaluation*, vol. 43, no. 2, pp. 161–179, 2019.

[7] N. Osaka, M. Osaka, M. Morishita, H. Kondo, and H. Fukuyama, "A word expressing affective pain activates the anterior cingulate cortex in the human brain: an fMRI study," *Behavioural Brain Research*, vol. 153, no. 1, pp. 123–127, 2020.

[8] K. E. Moore, "Ego-perspective and field-based frames of reference: temporal meanings of FRONT in Japanese, Wolof, and Aymara," *Journal of Pragmatics*, vol. 43, no. 3, pp. 759–776, 2019.

[9] T. Ogihara, "The ambiguity of the-te iru form in Japanese," *Journal of East Asian Linguistics*, vol. 7, no. 2, pp. 87–120, 2018.

[10] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding*, vol. 91, no. 1-2, pp. 160–187, 2019.

[11] F. Cheng and Y. Miyao, "Classifying temporal relations by bidirectional lstm over dependency paths," *Computational Linguistics*, vol. 7, pp. 1–6, 2019.

[12] S. Kita, A. Özyürek, S. Allen, A. Brown, R. Furman, and T. Ishizuka, "Relations between syntactic encoding and co-speech gestures: implications for a model of speech and gesture production," *Language and cognitive processes*, vol. 22, no. 8, pp. 1212–1236, 2017.

[13] L. Chen-Hafteck, "Music and language development in early childhood: integrating past research in the two domains," *Early Child Development and Care*, vol. 130, no. 1, pp. 85–97, 2019.

[14] Y. Peng, N. Kondo, T. Fujiura et al., "Dam behavior patterns in Japanese black beef cattle prior to calving: automated detection using LSTM-RNN," *Computers and Electronics in Agriculture*, vol. 169, p. 105178, 2020.

[15] K. Nishikawa, R. Hirakawa, H. Kawano, K. Nakashi, and Y. Nakatoh, "Detecting system Alzheimer's dementia by 1d CNN-LSTM in Japanese speech," in *2021 IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, USA, 2021, January.

[16] A. Bender, A. Rothe-Wulf, L. Hüther et al., “Moving forward in space and time: how strong is the conceptual link between spatial and temporal frames of reference?,” *Frontiers in Psychology*, vol. 3, p. 486, 2012.

[17] S. Allen, A. Özyürek, S. Kita et al., “Language-specific and universal influences in children’s syntactic packaging of manner and path: a comparison of Japanese, Japanese, and Turkish,” *Cognition*, vol. 102, no. 1, pp. 16–48, 2018.

[18] L. McCune, “A normative study of representational play in the transition to language,” *Developmental Psychology*, vol. 31, no. 2, p. 198, 2019.

[19] C. E. Izard, “Innate and universal facial expressions: evidence from developmental and cross-cultural research,” *Language Sciences*, vol. 9, pp. 4–9, 2019.

[20] M. Asahara and Y. Matsumoto, “Constructing a temporal relation tagged corpus of Chinese based on dependency structure,” *Japanese Society for Artificial Intelligence*, vol. 7, pp. 311–315, 2020.

[21] M. Sotirova-Kohli, D. H. Rosen, S. M. Smith, P. Henderson, and S. Taki-Reece, “Empirical study of Kanji as archetypal images: understanding the collective unconscious as part of the Japanese language,” *Journal of Analytical Psychology*, vol. 56, no. 1, pp. 109–132, 2019.

[22] F. Cheng, M. Asahara, I. Kobayashi, and S. Kurohashi, “Dynamically updating event representations for temporal relation classification with multi-category learning,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1352–1357, 2020.

[23] P. Pettenati, K. Sekine, E. Congestrì, and V. Volterra, “A comparative study on representational gestures in Italian and Japanese children,” *Journal of Nonverbal Behavior*, vol. 36, no. 2, pp. 149–164, 2019.

[24] N. T. Ly, C. T. Nguyen, and M. Nakagawa, “An attention-based row-column encoder-decoder model for text recognition in Japanese historical documents,” *Pattern Recognition Letters*, vol. 136, pp. 134–141, 2020.

[25] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, “Speech enhancement using self-adaptation and multi-head self-attention,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 181–185, Barcelona, Spain, 2020, May.

[26] R. Imaizumi, R. Masumura, S. Shiota, and H. Kiya, “Dialect-aware modeling for end-to-end Japanese dialect speech recognition,” in *In 2020 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC)*, pp. 297–301, Honolulu, Hawaii, 2020, December.

[27] Z. Huang, P. Zhang, R. Liu, and D. Li, “Immature apple detection method based on improved Yolov3,” *ASP Transactions on Internet of Things*, vol. 1, no. 1, pp. 9–13, 2021.

[28] J. Zhang, J. Sun, J. Wang, and X. G. Yue, “Visual object tracking based on residual network and cascaded correlation filters,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 8427–8440, 2021.

[29] W. Chu, P. S. Ho, and W. Li, “An adaptive machine learning method based on finite element analysis for ultra low-k chip package design,” *IEEE Transactions on Components, Packaging and Manufacturing Technology*, pp. 1–1, 2021.

[30] E. Yamada, “Fostering criticality in a beginners’ Japanese language course: a case study in a UK higher education modern languages degree programme,” *Language Learning in Higher Education*, vol. 6, no. 2, pp. 453–471, 2020.

[31] L. Mealier, G. Pointeau, and S. Mirliaz, “Narrative constructions for the organization of self experience: proof of concept via embodied robotics,” *Frontiers in Psychology*, vol. 8, p. 1331, 2017.