# A Hybrid Framework Integrating Speech Recognition, Lexical Frequency Analysis, and BERT-BiLSTM for Computational Modeling of Age-Based Emotional Expression Preferences in Nanjing Wu Dialect

Zuming Wang*

Nanjing University of Science and Technology School of Marxism, Nanjing, China

*wangzuming_joseph@outlook.com

## Abstract

**Computational modeling of dialectal emotional expression holds significant value for dialect digitalization and intelligent speech service development, yet existing studies generally overlook the moderating effect of age variables on emotional expression. This study constructs a hybrid computational framework integrating speech recognition, lexical frequency analysis, and BERT-BiLSTM to achieve quantitative modeling of age-based emotional expression preferences in Nanjing Wu dialect. The framework realizes dialect speech transcription through Wav2Vec 2.0 transfer learning, extracts age-group vocabulary preference vectors using the TF-IDF method, and captures deep semantic features through the BERT-BiLSTM cascaded architecture. Experiments based on Nanjing Wu dialect corpus demonstrate that the hybrid framework outperforms all baseline models in sentiment classification tasks, while age-stratified analysis reveals a significant pattern wherein positive sentiment proportion increases progressively with age while negative sentiment proportion decreases. The research outcomes can provide a reusable technical solution for dialect emotion computing and offer empirical evidence for the design of age-friendly intelligent speech services.**

## Keywords

**Nanjing Wu dialect, emotional expression preferences, age stratification, BERT-BiLSTM, hybrid framework.**

## 1. Introduction

Dialect, as a lively language resource carrying regional cultures and emotional cognitive patterns, has important values for dialect digitalization and the intelligent voice service system based on computational models simulating its characteristics of emotional expression. Current voice recognition systems for Chinese dialects have a basic technical foundation, but they are still confronted with severe performance limitations for the recognition of complex dialects because the dialect phonetic features are very heterogeneous and the corpora are insufficiently annotated[1]. Nanjing Wu dialect, a crucial branch of the Wu dialect group, also has its special tone and prosodic patterns, and the computational model study based on its emotional expression is untouched.

Age, as a prominent social factor that affects language use, plays a significant role in emotion expression preferences, with this age-grading phenomenon arising from intergenerational variations in culture as well as changes in cognitive functions across a lifetime trajectory[2]. Deep leaning techniques have proved to offer better performance for emotion recognition tasks, which combine multimodal information[3]. Technical support exists for a computational characterization of dialect emotion expression preferences for different ages. In current studies,

research into speech recognition for Southern dialects has made initial breakthroughs[4] and algorithms for emotion recognition with multimodal fusion strategies have also made great progress[5]. Meanwhile, studies on Wu dialect recognition have started to explore transfer learning to overcome difficulties related to a small amount of data[6]. However, to date, existing studies all ignore aspects involving moderating age factors for dialect emotion expression, without a systematic approach which can combine speech recognition, lexical statistics, and deep semantic models.

On the basis of the above shortcomings, this paper establishes a hybrid computing model that combines speech recognition, lexical frequency analysis, and BERT-BiLSTM to quantitatively model age-emotion expression preference in Nanjing Wu dialect. The technical breakthroughs lie in the following three parts: the speech recognition part uses the Wav2Vec 2.0 model that is pre-trained and then fine-tuned for the characteristics of Wu dialect, the lexical frequency analysis part identifies the high-frequency words of each age group for emotional expression by TF-IDF and establishes the preference vector, and the deep learning part uses BERT to represent the semantic features and BiLSTM to model the sequential expression behaviors. The technical results of this study have the potential to offer a reusable technical method for dialect emotion computing and provide an empirical basis for the intelligent speech services that are friendly to all age groups.

# 2. Methodology

## 2.1. Dataset and Preprocessing

The experimental data used in this study were incorporated from various publicly available dialect speech corpora, mainly including the THCHS-30 Chinese speech dataset and its dialect speech corpus subsets, as well as open-source Chinese sentiment corpora for emotion label transformation. The corpus selection involves a dual criterion of geographical division and linguistic features, where speech samples were extracted from Nanjing as well as the Wu dialects, and were stratified according to age, namely young (18-35 years), middle-aged (36-55 years), and senior (56+ years) age groups, with the stratified sampling of the original dataset to ensure a well-balanced filtered and marked dataset, as described in Table 1 below:

**Table 1.** Experimental Dataset Statistics

| Age Group | Samples | Positive | Negative | Neutral | Avg. Duration (s) |
|---|---|---|---|---|---|
| Young (18-35) | 387 | 147 | 134 | 106 | 4.6 |
| Middle-aged (36-55) | 342 | 134 | 109 | 99 | 4.9 |
| Elderly (56+) | 289 | 122 | 79 | 88 | 5.3 |
| Total | 1,018 | 403 | 322 | 293 | 4.9 |

Table 1 shows the structure of the Nanjing Wu dialect emotional corpus. The number of subjects in the three age groups is distributed with more subjects in the young and middle-aged groups, and fewer subjects in the elderly group, which objectively reflects the fact that it is hard to get Voice data from elderly users of the Nanjing Wu dialect. Positive emotions took the lead across all age groups (39.6%), and the difference between positive and negative emotion distributions provided an analytical data basis for further age-based modeling regarding emotional expression preferences.

## 2.2. Hybrid Computational Framework

For the problem of preference modeling on the expression of emotions on the basis of age in the Nanjing Wu dialect, a three-stage cascaded framework has been developed based on the

requirements mentioned above by integrating the fields of speech recognition, frequency analysis, and deep semantics, as depicted in the architectural design below in Figure 1.

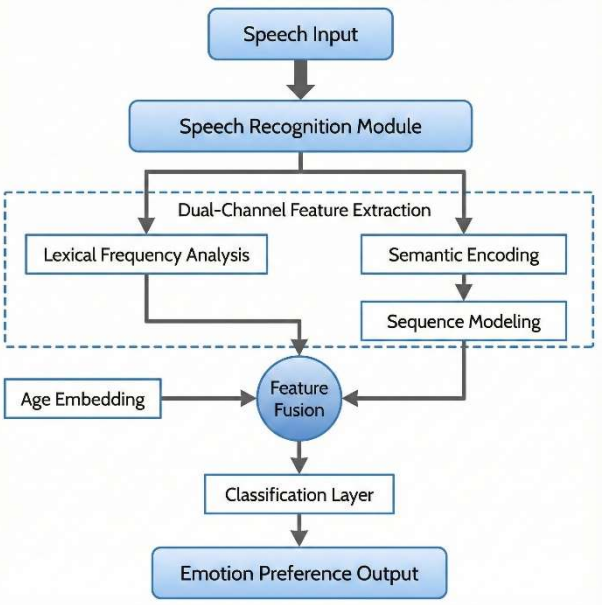


**Figure 1.** Overall Architecture of the Hybrid Computational Framework

The hybrid paradigm in Figure 1 takes a "speech recognition - dual-channel feature extraction - multi-source feature fusion" architecture. The text sequence output from the speech recognition component is utilized as a common input for dual-channel feature extraction. In this, the left part represents feature extraction based on lexical statistics, and the right part represents feature extraction based on deep semantic information using semantic encoding and modeling of sequences. The output from the word frequency analysis component, sequence modeling component, and embedding vector of age are combined in three different manners at a fusion node, followed by output of the sentiment preference prediction using a classification layer.

The lexical frequency analysis component applies the TF-IDF algorithm to obtain the distinctive vocabulary for the different age groups. This has shown special superiority in the explicit emotional lexical characterization[7], with the calculation formula for the weighting given by:

$$w_{t,d} = (1 + \log tf_{t,d}) \times \log \frac{N}{df_t} \tag{1}$$

where $tf_{t,d}$ represents term frequency, $N$ denotes the total number of documents, and $df_t$ indicates the number of documents containing term $t$. On the basis of these weights, vectors are then formed that express the emotional vocabulary selection tendency of each age group.

The deep semantic modeling component uses a BERT-BiLSTM cascading architecture, where BERT derives dynamic semantic features of the vocabulary using its bidirectional Transformer Encoder[8], whose outputs are passed to the bidirectional LSTM layer:

$$\overrightarrow{h_t} = \mathrm{LSTM}(x_t, \overrightarrow{h_{t-1}}), \quad \overleftarrow{h_t} = \mathrm{LSTM}(x_t, \overleftarrow{h_{t+1}}) \tag{2}$$

$$h_t^* = [\overrightarrow{h_t} \oplus \overleftarrow{h_t}] \tag{3}$$

In order to fully achieve the integration of the lexical frequency statistical features and the deep-level semantic features, the strategy of feature concatenation is adopted in this study. The strategy involves the concatenation of the TF-IDF vocabulary preference vectors, the BiLSTM output sequence vectors, and the age group embedding vectors. The hybrid feature vectors contain the contextual semantic meaning, the lexical choice preferences, and the age group information, all of which are used for the prediction of the emotion category through the fully connected layer. The training of the model is conducted through the use of the age group-weighted cross-entropy function:

$$\mathcal{L} = -\sum_{i=1}^{N} \omega_{k_i} \sum_{c=1}^{C} y_{ic} \log(\hat{y}_{ic}) \tag{4}$$

where $\omega_{k_i}$ represents the age group weight, and $y_{ic}$ and $\hat{y}_{ic}$ denote the ground truth label and predicted probability, respectively. Training utilizes the Adam optimizer with a learning rate of $2\times10^{-5}$ and batch size of 32, with the complete procedure detailed in Algorithm 1.

---

Algorithm 1: Age-Aware Hybrid Feature Emotional Preference Classification

---

Input: Speech samples $S = \{s_1, s_2, ..., s_n\}$, Age labels $A = \{a_1, a_2, ..., a_n\}$

Output: Emotion preference predictions $\hat{Y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_n\}$

Initialize: learning rate $\alpha = 2 \times 10^{-5}$, batch size $B = 32$, convergence threshold $\varepsilon$

Initialize: age group weights $\omega = \omega_1, \omega_2, \omega_3$

Speech Recognition:

for each speech sample $s_i$ do

$T_i \leftarrow \text{Wav2Vec2.0\_Transcription}(s_i)$ with Wu dialect fine-tuning

end for

Dual-Channel Feature Extraction:

for each text sequence $T_i$ do

Compute TF-IDF weights: $w_{t,d} = (1 + \log tf_{t,d}) \times \log \frac{N}{df_t}$

$v_{tfidf} \leftarrow \text{BuildPreferenceVector}(T_i, a_i)$

$H_{bert} \leftarrow \text{BERT\_Encoder}(T_i)$

for each time step $t$ do

$\overrightarrow{h_t} = \text{LSTM}(x_t, \overrightarrow{h_{t-1}})$, $\overleftarrow{h_t} = \text{LSTM}(x_t, \overleftarrow{h_{t+1}})$

$h_t^* = [\overrightarrow{h_t} \oplus \overleftarrow{h_t}]$

end for

$h_{lstm} \leftarrow \text{BiLSTM\_Output}(H_{bert})$

$e_{age} \leftarrow \text{Embedding\_Lookup}(a_i)$

end for

Feature Fusion and Classification:

Repeat

for each batch $b = 1$ to $\lceil N / B \rceil$ do

$h_{fused} \leftarrow \text{Concat}(v_{tfidf}, h_{lstm}, e_{age})$

$\hat{y}_i \leftarrow \text{Softmax}(\text{FC}(h_{fused}))$

Compute weighted cross-entropy loss: $\mathcal{L} = -\sum_{i=1}^{N} \omega_{k_i} \sum_{c=1}^{C} y_{ic} \log(\hat{y}_{ic})$

$\theta^{t+1} \leftarrow \text{Adam}(\theta^t, \nabla\mathcal{L}, \alpha)$

end for

until $\|\Delta\| < \varepsilon$

return Final emotion preference predictions $\hat{Y}$

Algorithm 1 demonstrates the process for model training, including speech recognition, TF-IDF feature extraction, encoding based on the use of the BERT and BiLSTM model, embedding for the age variable, and the fusion of features. The proposed architecture combines lexical features, semantic features, and stratified features based on the age variable to capture variations in preference for emotional expression.

# 3. Results

## 3.1. Framework Performance Validation

To ensure the efficacy of the proposed combined scheme, two aspects were used for evaluation in the proposed task: the recognition of speeches, as well as the recognition of sentiments. With respect to the recognition of speeches, the model that involved the fine-tuning of Wav2Vec 2.0 through the concept of transfer learning reached an average Character Error Rate (CER) measure of 18.6% for the test dataset for the Nanjing Wu dialect, where the CER for the youth group recorded 15.2%, the middle-aged group recorded 18.4%, and the seniors recorded 23.1%, due to the nature of the seniors having lower clarity and variations in speech rate.

**Table 2.** Model Performance Comparison and Ablation Experiment Results

| Model | Accuracy (%) | Macro-F1 (%) |
|---|---|---|
| SVM (TF-IDF) | 63.2 | 59.8 |
| BiLSTM | 68.7 | 64.3 |
| BERT | 72.4 | 69.1 |
| BERT-BiLSTM (w/o TF-IDF) | 73.8 | 70.5 |
| Hybrid Framework | 76.9 | 73.6 |
| — w/o TF-IDF Module | 74.1 | 70.8 |
| — w/o Age Embedding | 75.3 | 71.7 |

For evaluating the performances of the models in the task of sentiment classification, metrics of accuracy and macro F1-value (Macro-F1) were used, and there were four comparative experiments involving one model with BERT, one model with BiLSTM, a cascaded model involving both BERT and BiLSTM but not involving term frequency features, and a traditional Support Vector Machine method using TF-IDF features. For comparing the effects of individual

components on the overall performances, ablation experiments involving the removal of the module containing the analysis of the term frequencies and the module containing the embedding of the age were performed, whose results are given in Table 2.

The effectiveness of the proposed hybrid model is evident from the results shown in Table 2, as it performs better compared to all the rival models with an improvement of 3.1% in the F1 score when no term frequency is considered in the BERT-BiLSTM model. Furthermore, ablation tests show that the removal of the analysis module for the term frequency analysis task contributes to the 2.8% decrease in the F1 score, and the removal of the age embedding module contributes to the 1.9% decrease in the F1 score.

### 3.2. Age-Stratified Emotional Expression Preference Analysis

With the hybrid model developed, this research carried out a quantitative analysis of the preference nature of emotional expression in three age groups. In regard to lexical preference, extraction of high-frequency emotional words in each of the three age groups employing TF-IDF preference vectors brought out that the young generation prefers the use of modal particles and novel expressions, the elderly generation favors traditional dialect words, and the middle-aged generation has an intermediate form of lexical preference in between. In regard to the outcome of lexical diversity, it was found that TTR in the young generation was 0.72, in the middle-aged generation was 0.68, and in the elderly generation was 0.61, reflecting an orderly decrease in the value of lexical diversity with an increase in age.

The question whether there is a further manifestation at the lexical level among the aforementioned differences in overall emotional tendencies awaits verification on the basis of sentiment category distribution. The proportion of positive, negative, and neutral sentiments for each group was calculated and tested for differences using Chi-square tests, with the results in Figure 2.

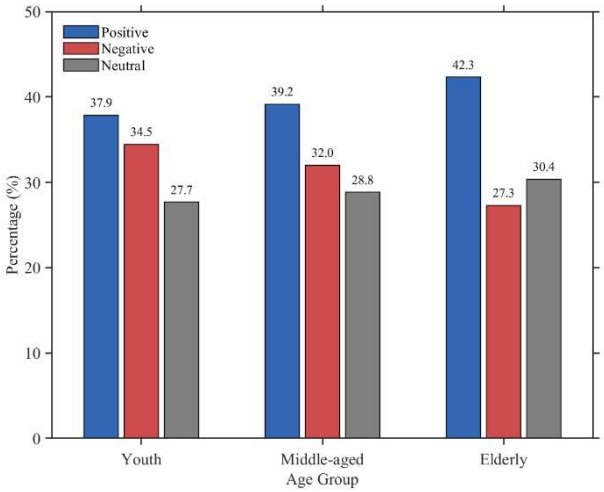


**Figure 2.** Sentiment Category Distribution Across Age Groups

Diagram 2 shows that the trend for the ratio of positive sentiment expressed rises with age, together with the ratio of negative sentiment expressed declining, with differences between groups significant at the statistical level ($\chi^2$=6.82, $p<0.05$), thus indicating that age-related factors significantly affect the expression of dialect emotion sentiment.

## 4. Discussion

The hybrid model developed in this paper has reached an accuracy rate of 76.9% and a macro average F1 value of 73.6% on the Nanjing Wu dialect sentiment classification task, which has

improved the F1 value by 3.1% compared to the BERT-BiLSTM model without word frequency features. Existent studies have confirmed that the multi-channel CNN-BiLSTM model has stronger context expression abilities in the process of sentiment analysis on Chinese texts[9]. Furthermore, in this paper, the range of the covered feature space has also been extended through the addition of TF-IDF word frequency features and age embedding vectors. The ablation test has proved that the weight of the word frequency component is higher than that of the age embedding component. The weight assignment method through the attention mechanism[10] and the potential for the application of the BERT-BiLSTM-Attention model[11] may serve as possible clues for the optimization of the later model.

In terms of age-specific emotion expression preference, data revealed that the proportion of positive emotion preferences increased with age, as did the proportion of preference for negative emotions, which decreased with age. The results are well aligned with one of the primary hypotheses of social emotional choice theory, which states that older people express preferences for positive emotions to keep emotional stability[12]. The youth group prefers interjections and emerging terms, while older people favor traditional dialect terms, which illustrate language conservatism due to age factors[13]. The lexical diversity index indicated that TTR for youth group was 0.72, while for older people, it declined to 0.61. The graded level corresponds to results stating that lexical statistical levels are associated with emotional expression[14]. The speech recognition engine indicates that CER for older people increased to 23.1%, which considerably differs from 15.2% CER of the youth group. These age-specific differences conform to challenges posed by research done on dialect-based deep learning modeling with low resource levels[15].

In addition, there are still some limitations in this study. First, the publicly available dialect speech datasets are found to have some imperfections in the coverage of Wu dialects, resulting in a smaller scale of the corpus and thus an insufficient generalization capability of deep learning. Furthermore, there might also be potential subjective biases brought by the subjective assignment of sentiment. In extension, there are two possible aspects that future studies may focus on. On the one hand, the study may extend to different points of Wu dialects, such as Suzhou dialect and Shanghai dialect, in order to test the inter-dialect migration capability of the framework. On the other hand, it may incorporate acoustic prosodic parameters and facial expressions in an attempt to build a multimodal framework for dialect sentimental analysis.

## 5. Conclusion

This research aims at the problem of computational modeling of age-stratified emotional expression preference in Nanjing Wu dialect, and it builds a hybrid approach based on speech recognition, word frequency analysis, and BERT-BiLSTM. The approach can achieve dialect speech transcription by transferring the pre-trained language model Wav2Vec2.0, and it can also obtain the vectors of vocabulary preference for each age group through TF-IDF. By using the cascading structure of BERT-BiLSTM, it can extract the deep semantic characteristics and expression sequences. Experiment results show that, compared to the base model BERT-BiLSTM, the suggested combined framework can increase the value of F1 by 3.1 percentage points, achieving a total accuracy of 76.9% and a macro-F1 value of 73.6% for the Nanjing Wu dialect. Age stratification analysis also found a significant rule that the proportion of positive emotions rises and the proportion of negative emotions reduces with the increase of age. The vocabulary diversity index has a gradient characteristic presenting a reducing trend from youth to the elderly.

Theoretical contribution: The contribution of the research theory lies in the integration of the age factor into the model framework of dialect sentiment computing. This adds a new dimension to the research approach on the interrelation of sociolinguistics and natural

language processes. Practical contribution: The hybrid approach can serve as technical assistance in the digital preservation of dialects. Simultaneously, the age-related emotional preference data found by the research can provide new empirical evidence for creating elderly-friendly intelligent voice assistance systems.

## References


[1] Q. Li, Q. Mai, M. Wang, and M. Ma, "Chinese dialect speech recognition: a comprehensive survey," Artificial Intelligence Review, vol. 57, no. 2, p. 25, 2024.

[2] M. Hejná and A. Jespersen, "Ageing well: Social but also biological reasons for age-grading," Language and Linguistics Compass, vol. 16, no. 5-6, p. e12450, 2022.

[3] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face," Entropy, vol. 25, no. 10, p. 1440, 2023.

[4] W. Lai and Y. Zheng, "Speech recognition of south China languages based on federated learning and mathematical construction," Electronic Research Archive, vol. 31, no. 8, 2023.

[5] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," Expert Systems with Applications, vol. 237, p. 121692, 2024.

[6] X. Yue, L. Miao, and J. Ding, "Research on Wu Dialect Recognition and Regional Variations Based on Deep Learning," Applied Sciences, vol. 15, no. 18, p. 10227, 2025.

[7] D. E. Cahyani and I. Patasik, "Performance comparison of tf-idf and word2vec models for emotion text classification," Bulletin of Electrical Engineering and Informatics, vol. 10, no. 5, pp. 2780-2788, 2021.

[8] X. Li, Y. Lei, and S. Ji, "BERT-and BiLSTM-based sentiment analysis of online Chinese buzzwords," Future Internet, vol. 14, no. 11, p. 332, 2022.

[9] C. Gan, Q. Feng, and Z. Zhang, "Scalable multi-channel dilated CNN–BiLSTM model with attention mechanism for Chinese textual sentiment analysis," Future Generation Computer Systems, vol. 118, pp. 297-309, 2021.

[10] Y. Wu, S. Zhang, and P. Li, "Multi-modal emotion recognition in conversation based on prompt learning with text-audio fusion features," Scientific Reports, vol. 15, no. 1, p. 8855, 2025.

[11] X. Li, L. Chen, B. Chen, and X. Ge, "BERT-BiLSTM-Attention model for sentiment analysis on Chinese stock reviews," 2024.

[12] P.-Y. Zeng and S.-L. Yeh, "Exploring semantic expression disparities in intragenerational and intergenerational communication: A novel perspective on socioemotional selectivity theory," Psychology and Aging, 2025.

[13] Y. Lin et al., "Category-sensitive age-related shifts between prosodic and semantic dominance in emotion perception linked to cognitive capacities," Journal of Speech, Language, and Hearing Research, vol. 67, no. 12, pp. 4829-4849, 2024.

[14] J. Xu, "A natural language processing based technique for sentiment analysis of college english corpus," PeerJ Computer Science, vol. 9, p. e1235, 2023.

[15] Z. Qi, F. Li, and H. Long, "Research on optimal deep learning modeling in HaiNan dialect recognition," Scientific Reports, vol. 15, no. 1, p. 31735, 2025.