

Corpus-based research and digital preservation of endangered Lingnan dialects

Guiying Kong^{1,*}

¹ Xijiang River Valley Folk Literature Research Center, Wuzhou University, Wuzhou, Guangxi, 543002, China

Corresponding authors: (e-mail: camelliared1973@163.com).

Abstract To establish and preserve a corpus of endangered Lingnan dialects, this paper combines convolutional neural networks and gated recurrent unit technology to build a CNN-CTC acoustic model, proposing a Lingnan dialect recognition model that achieves mapping recognition from Lingnan dialects to Mandarin. Taking 160 audio files and approximately 43 hours of raw audio corpus as the research object, a special topic analysis was conducted and the storage and presentation forms of the corpus data were presented. The results show that the highest word frequency of the corpus "My Hometown" is "ge", which is 95 times, with a frequency of 0.0362, followed by "shi", "ah", "di", etc. Approximately half of the 0.1% class symbols in the Lingnan dialect spoken language corpus correspond to character symbols. In the usage test of the Lingnan dialect corpus, the average SUS value was 82.40, which can drive the continuous optimization of corpus design and user experience, thereby achieving its digital preservation.

Index Terms CNN-CTC, dialect recognition, Lingnan dialect, corpus

I. Introduction

Dialects, as unique cultural forms, convey the local accent and serve as an important link for linguistic communication and emotional connection among the people of a region. They are the key to the inheritance of regional cultural and artistic traditions and an important vehicle for fostering cultural confidence [1]-[3]. The protection and development of dialects are crucial for formulating relevant policies, developing new cultural industries, and safeguarding the cultural genes of ethnic groups [4]. Unfortunately, dialects have been gradually declining in recent years. In the digital age, characterized by the continuous innovation and development of emerging technologies such as big data and artificial intelligence, conducting integrated research and development applications on dialect resources that embody the distinctive features of traditional Chinese culture using digital humanities technology holds significant practical significance [5]-[8].

With the widespread adoption of the internet and digital technologies, the field of digital humanities research has expanded to encompass multiple traditional humanities disciplines, including linguistics and sociology, as well as broader interdisciplinary fields [9], [10]. Its research subjects are various digitizable resources in the field of humanities and social sciences. Digitization is a cultural process of reinterpretation and recombination that is open to everyone. Digital humanities provide traditional humanities researchers with new research perspectives and ways of thinking. People have begun to consider reconstructing the context and content of humanities knowledge from a data perspective and constructing contemporary knowledge systems and cognitive methods from a new technological perspective [11]-[13]. The construction of a dialect corpus repository is a powerful tool to address the increasingly severe issue of dialect loss [14]. The purpose of this repository is to comprehensively and authentically document all aspects of dialect culture. By collecting dialect materials in various forms, including images, text, audio, and video, it provides strong support for the protection and inheritance of dialects [15]-[18]. With the support of digital technology, the dialect repository can serve as a bridge connecting tradition and modernity, reality and the future, enabling Lingnan dialect culture to thrive with new vitality in the new media era.

Numerous studies have shown that some universities and relevant civil society organizations have begun to conduct research and undertake conservation efforts for local dialects. Reference [19] utilized Cooper's statistical method to explore the role of cultural heritage in regional dialect conservation projects and proposed a new framework for sustainable language resource conservation, providing decision-making support for the formulation of language conservation and inheritance policies. Literature [20] indicates that the linguistic ecological context of dialects is facing a severe crisis. Therefore, it proposes a series of work recommendations from the perspective of linguistic ecology, focusing on policy, education, publicity, and resource development. Literature [21] investigated the linguistic ideology and practices of migrant families toward dialect culture, finding that the most common

linguistic practice is the combined use of dialects and Mandarin, while language transmission between grandparents and children is the primary form of dialect inheritance. Literature [22] proposes strategies for the protection of ethnic language resources in Northeast China, including the implementation of multilingual subtitles in publicity and the conduct of bilingual education in education. The proposed strategies strengthen the protection and inheritance of regional ethnic language resources. The above studies demonstrate society's deep concern for linguistic diversity and the regional cultures it carries.

Additionally, with the advent of the digital age, the rapid development of digital technology has brought unprecedented opportunities for the protection and preservation of dialects. Reference [23] established a bidirectional long short-term memory-conditional random field (BLSTM-CRF) model framework for dialect recognition, integrating it with dialect product design to provide a new digital development pathway for the protection and dissemination of dialect culture. Reference [24] explored the shaping role of digital communication technology on the linguistic context of minority languages, emphasizing the integration and protective functions of digital tools, thereby offering insights for language protection in the digital world. Literature [25] proposes a minority language protection strategy that integrates big data technology, using data mining techniques to analyze the current state of ethnic language resources, thereby formulating targeted protection strategies to disseminate and inherit ethnic culture. However, compared to mainstream language systems with abundant speech data resources, the aforementioned technologies often face the challenge of insufficient corpora during the implementation of dialect protection tasks. This means that for dialects with scarce resources, establishing a corpus is a prerequisite for implementing digital protection.

This paper leverages digital platforms and digital technology to establish a Lingnan dialect recognition model based on the GRU-CTC algorithm for the establishment and digital preservation of endangered Lingnan dialect corpora, enabling the recognition of continuous Lingnan dialect speech under a large vocabulary. First, Mel-frequency cepstral coefficients are used to extract feature parameters of endangered Lingnan dialects. These feature parameters are then input into a gated linear unit for training and optimization, yielding prediction labels for the entire input sequence. The speech of endangered Lingnan dialects is mapped into text to construct the corpus. A total of 160 audio files, equivalent to approximately 43 hours of audio data, were collected for application research to validate the reliability of the methods proposed in this paper.

II. Corpora and digital preservation

II. A. Corpora and digital preservation

Corpus construction refers to the process of collecting, organizing, storing, managing, and utilizing language materials to facilitate the observation, analysis, and research of linguistic phenomena by linguists, computer language processing researchers, and other researchers in related fields.

The theory behind corpus construction spans multiple fields, including linguistics, computer science, and information science. First, corpus design involves various aspects such as data sources, collection methods, data formats, and storage methods, all of which must be considered in light of the corpus's intended use and research requirements. Second, corpus annotation refers to the process of adding tags such as part-of-speech, syntactic structure, and semantic information to the text within a corpus, to meet the needs of corpus construction research. Corpus annotation methods include manual annotation and automatic annotation. Corpus management involves tasks such as corpus storage, backup, retrieval, and updating, which require consideration of factors such as corpus size, nature, and usage requirements. Finally, corpus utilization refers to the process of using corpora for linguistic and computational language processing research, including applications such as text mining, information retrieval, and natural language processing. Finally, corpus evaluation refers to the process of assessing the quality of a corpus, which requires consideration of multiple aspects such as accuracy, completeness, diversity, and timeliness. Therefore, corpus construction involves multiple disciplines and requires comprehensive consideration of various factors to ensure the quality and practicality of the corpus [26].

II. B. Principles for corpus construction

The construction of a corpus should follow certain principles, including:

(1) Language balance principle: A corpus should include various text types, styles, and topics to ensure its representativeness and diversity. For example, a bilingual corpus used for machine translation should include various text types, such as news reports, agricultural information, literary works, and scientific articles, to cover different language usage scenarios.

(2) Data quality principle: The corpus should avoid low-quality texts, such as those containing spelling or grammatical errors. Additionally, the corpus should adhere to standardized writing conventions as much as possible, such as formal writing styles and standard language norms, to ensure the accuracy and reliability of the data.

(3) Copyright Principle: The construction of a corpus must comply with copyright regulations, including obtaining authorization and respecting the rights of the original copyright holders. In practice, researchers often choose publicly available text data, such as news reports, blog posts, and social media content from the internet.

(4) Diversity Principle: A corpus should include text data from different regions, social groups, and time periods to reflect different cultural backgrounds and language usage scenarios. For example, a corpus used for agricultural information research should include text from different regions and different agricultural technologies to reflect different language variants and linguistic variation.

(5) Annotation and labeling principle: The construction of a corpus requires the annotation and labeling of text data to support various natural language processing tasks. For example, when constructing a corpus for named entity recognition, the text data must be annotated with entity types.

II. C. Dialect Number Protection

In regions such as Lingnan, digital platforms and technologies are being utilized to digitally preserve regional dialects from a long-term perspective. This not only ensures the permanent preservation of dialects as intangible cultural heritage but also integrates them into social life through comprehensive utilization, thereby preventing collective amnesia through cultural transmission. The primary objective of digital preservation efforts for dialects across various regions is to rescue endangered languages and promote the continuation of traditional culture. Additionally, the value-added objectives of dialect digital preservation have also been realized in practice. Digital archives position their functions as “preserving traditional culture and fostering new culture,” aiming to promote the establishment of a fair society through digital initiatives, provide materials for educational reforms at all levels, offer tool support for linguistic research, and to some extent, stimulate local economic and industrial development. This seeks to achieve the comprehensive utilization of dialects across industries, education, academic research, daily life, and leisure activities.

III. Principles of dialect recognition

Dialect automatic recognition generally refers to the construction of a mapping relationship between Lingnan dialects and standard languages (such as Mandarin Chinese) based on information technology methods. Researchers both domestically and internationally have conducted extensive studies using high-quality dialect corpora. Addressing the scarcity of research on Lingnan dialect speech recognition, the author proposes an improved GRU neural network speech recognition method, as shown in Figure 1. This method first extracts feature information from the original speech using MFCC, then combines it with the CTC model to map it to text. The model was trained and tested on a Lingnan dialect corpus. It was implemented using the Paddle framework, based on GRU and CTC. The model consists of four main modules: MFCC speech feature extraction, CTC-GRU model training, CTC maximum path decoding, and prediction evaluation [27].

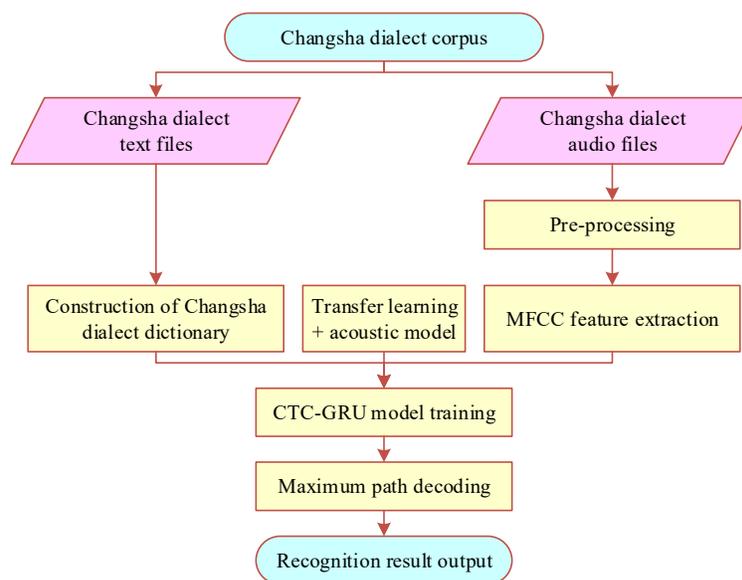


Figure 1: Changsha language recognition process

III. A. Feature parameter extraction

Mel-frequency cepstral coefficients (MFCCs) are cepstral parameters extracted in the Mel-scale frequency domain, describing the nonlinear characteristics of human ear frequencies. Since this feature does not depend on the nature of the signal, makes no assumptions or restrictions on the input signal, and utilizes the research results of auditory models, it has good robustness. Therefore, this study uses MFCC as the feature parameter for dialect recognition. The relationship between Mel frequency and speech signal frequency f is as follows:

$$Mel(f) = 1125 \log_{10} \left(1 + \frac{f}{700} \right) \quad (1)$$

The MFCC extraction process is as follows:

(1) Pre-emphasis. Pass the speech signal through a high-pass filter to flatten the signal spectrum.

(2) Frame splitting with windowing. To increase the continuity of the left and right ends of each frame after frame splitting, each frame after frame splitting is multiplied by a Hamming window. Assuming that the signal after frame splitting is $S(n), n = 0, \dots, N-1$ and N is the frame size, then after multiplication by the Hamming window, the form of $S(n) = S(n) \times W(n)$ and $W(n)$ is:

$$W(n, a) = (1-a) - a \times \cos \left(\frac{2\pi n}{N-1} \right) \quad 0 \leq n \leq N-1 \quad (2)$$

(3) Fast Fourier Transform (FFT) calculates the spectrum. The spectrum is calculated by performing a fast Fourier transform on N point of each frame of the preprocessed signal. Therefore, the discrete Fourier transform (DFT) of the speech signal is:

$$S_i(k) = \sum_{n=0}^{N-1} s_i(n) e^{-j2\pi kn/N} \quad 0 \leq k \leq N \quad (3)$$

In the equation, $s_i(n)$ represents the input speech signal, and N represents the number of points in the Fourier transform.

(4) Calculate the power spectrum (periodic diagram), and take the modulus square of the spectrum obtained from equation (2) to obtain the spectral line energy of the speech signal:

$$P = \frac{|S_i(k)|^2}{N} \quad (4)$$

(5) The Mel filter eliminates harmonics, highlights the resonant peaks of the original voice, smoothes the spectrum, and reduces the amount of computation. Pass the energy obtained from equation (3) through the Mel filter to calculate the energy of the voice signal after passing through the $m(1 \leq m \leq M)$ st filter. H_m and M are the number of filters. The frequency response of the triangular filter is defined as:

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \leq k < f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \leq k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases} \quad (5)$$

Here $\sum_{m=1}^M H_m(k) = 1$.

(6) Logarithmic operation. Logarithms are better suited to describe the human ear's perception of sound due to their nonlinear relationship. Calculate the logarithmic energy of the output of the two filter groups M :

$$s(m) = \ln \left(\sum_{k=0}^{N-1} |S_i(k)|^2 H_m(k) \right) \quad 1 \leq m \leq M \quad (6)$$

(7) MFCC coefficients can be obtained by discrete cosine transform (DCT):

$$C(n) = \sum_{m=0}^{N-1} s(m) \cos \frac{\pi n(m-0.5)}{M} \quad n = 1, \dots, L \quad (7)$$

III. B. CTC model

III. B. 1) Definition of CTC

For an input sequence of length T , x , define a GRU with m inputs, n outputs, and a weight vector of w to obtain a continuous mapping, $N_w : (G_m)^T \rightarrow (G_n)^T$. Let $y = N_w(x)$ be the network output sequence, and use y_t^k to represent the probability of observing k at time t .

Define the finite character set of the input sequence as L . To align the input and output, consider inserting blank characters in missing positions. Define the character set for inserting blank labels as: $L' = L \cup \{blank\}$.

Redefine a one-to-many mapping $B: L^T \rightarrow L'^T$, where L' can be transformed into the original L via B . Clearly, there is a maximum length $|L| \leq T$ for L . This can be achieved by simply removing all spaces and duplicate tags from the path. Finally, use B to define the conditional probability of a given tag $l \in L'^T$ as the sum of the probabilities of all corresponding paths:

$$p(l|x) = \sum_{\pi \in B^{-1}(l)} p(\pi|x) \quad (8)$$

Among them, $\pi \in B^{-1}(l)$ represents all paths π that are l after undergoing transformation B . Therefore, for any path π , we have:

$$p(\pi|x) = \prod_{t=1}^T y_{\pi_t}^t \quad \forall \pi \in L^T \quad (9)$$

(9) The condition for the validity of the formula is that, given the internal state of the network, the outputs of the network at different times are independent of each other. This is achieved by ensuring that there are no feedback connections from the output layer to itself or to the network.

III. B. 2) Constructing a decoder

The decoder's output should be the most probable label for the input sequence. This study adopts best path decoding based on the assumption that the most probable path corresponds to the most probable label:

$$\begin{aligned} h(x) &= \arg \max_{l \in L'^T} p(l|x) \approx B(\pi) \\ \pi^* &= \arg \max_{\pi \in N^T} p(\pi|x) \end{aligned} \quad (10)$$

Since π^* is the most active output in each time step, the optimal path decoding calculation is very simple. However, optimal path decoding cannot guarantee finding the most probable label.

III. B. 3) Forward-backward algorithm

Due to the existence of blank label *blank*, the calculation of conditional probability $p(l|x)$ can be solved using a dynamic programming algorithm, similar to the forward-backward algorithm of HMM. The key idea is to decompose the sum along the path corresponding to the label into iterations corresponding to the prefix (or suffix) of the label, and then use a recursive approach to efficiently iterate and calculate the forward and backward variables.

For label l , define $\alpha_t(s)$ as the forward recursive probability of $l_{1:s}$ at time t :

$$\alpha_t(s) \stackrel{def}{=} \sum_{\substack{\pi \in L^T \\ B(\pi_{1:t})=l_{1:s}}} \prod_{t'=1}^T y_{\pi_{t'}}^{t'} \quad (11)$$

To retain the blank label b in the output path, a modified character label sequence l' was considered, with blank labels b added at the beginning and end, and blank labels b inserted between each pair of character labels. Therefore, the length of l' is $2|l|+1$. When calculating the forward probability and l' , all transitions between blank labels b and character labels are allowed, as well as any transitions between any two different characters.

As shown in equation (11), $\alpha_t(s)$ can be derived recursively from $\alpha_{t-1}(s)$ and $\alpha_{t-1}(s-1)$:

$$\alpha_t(s) = \begin{cases} \bar{\alpha}_t(s) y_{l'_s}^t & l'_s = b \text{ Or } l'_{s-2} = l'_s \\ (\bar{\alpha}_t(s) + \alpha_{t-1}(s-2)) y_{l'_s}^t & \text{Other} \end{cases} \quad (12)$$

Among them, $\bar{\alpha}_t(s) \stackrel{def}{=} \alpha_{t-1}(s) + \alpha_{t-1}(s-1)$.

Assuming that all forward variables are allowed to begin with a blank label *blank*(b) or the first character of a character label (l_k), then $\alpha_t(s)$ has the following properties:

$$\begin{aligned} \alpha_1(1) &= y_b^1, l_1 = b \\ \alpha_1(2) &= y_{l_1}^1, \pi_1 = l_1 = l'_2 \\ \alpha_1(s) &= 0 \quad \forall s > 2 \end{aligned} \quad (13)$$

Similarly, we can define the sum of probabilities at time t on $l_{s:|l|}$, i.e., the reverse recursive probability sum $\beta_t(s)$:

$$\beta_t(s) \stackrel{\text{def}}{=} \sum_{\substack{\pi \in L^T \\ B(\pi_{1:T})=l_{s:|l|}}} \prod_{t'=t}^T y_{\pi_{t'}}^t \quad (14)$$

$$\begin{aligned} \beta_T(|l|) &= y_b^T, \beta_T(|l|-1) = y_{l_{|l|}}^T, \beta_T(s) = 0 \forall s < |l|-1 \\ \beta_t(s) &= \begin{cases} \bar{\beta}_t(s) y_{l'_s}^t & l'_s = b \text{ Or } l'_{s+2} = l'_s \\ (\bar{\beta}_t(s) + \beta_{t+1}(s+2)) y_{l'_s}^t & \text{Other} \end{cases} \end{aligned} \quad (15)$$

Among them $\bar{\beta}_t(s) \stackrel{\text{def}}{=} \beta_{t+1}(s) + \beta_{t+1}(s+1)$.

III. B. 4) Maximum likelihood training

The goal of maximum likelihood training is to simultaneously maximize the probability of all correctly classified records in the training set, which means minimizing the following objective function:

$$O^{ML}(S, N_w) = - \sum_{(x,l) \in S} \ln(p(l|x)) \quad (16)$$

Among them: S is the finite character set of the training set; N_w is the output sequence of the *GRU* neural network; l is one of the character sequences. The network is trained using the gradient descent method. Since the training examples are independent, the network output can be considered separately:

$$\frac{\partial O^{ML}(\{(x,l)\}, N_w)}{\partial y_k^t} = - \frac{\partial \ln(p(l|x))}{\partial y_k^t} \quad (17)$$

For label l , the product of the forward and backward variables at points s and t is the sum of the probabilities of all paths through symbol s at time t , which can be derived from equations (11) and (13):

$$\alpha_t(s) \beta_t(s) = \sum_{\substack{\pi \in B^{-1}(l) \\ \pi_t = l'_s}} y_{l'_s}^t \prod_{t=1}^T y_{\pi_t}^t \quad (18)$$

$$\frac{\alpha_t(s) \beta_t(s)}{y_{l'_s}^t} = \sum_{\substack{\pi \in B^{-1}(l) \\ \pi_t = l'_s}} p(\pi|x) \quad (19)$$

It is known that $p(l|x)$ is the total probability. Since these paths take l'_s time t , at any time t , summing all results, we have:

$$p(l|x) = \sum_{s=1}^{|l|} \frac{\alpha_t(s) \beta_t(s)}{y_{l'_s}^t} \quad (20)$$

To distinguish it from y_k^t , we only need to consider the path of time t passing through tag k . Note that the same tag (or blank) may be deleted multiple times as a single tag l . Define the set of locations where tag k appears as $lab(l,k) = \{s : l'_s = k\}$. This set may be empty. Therefore, according to equation (20), we obtain:

$$\frac{\partial p(l|x)}{\partial y_k^t} = \frac{1}{y_k^{t^2}} \sum_{s \in lab(l,k)} \alpha_t(s) \beta_t(s) \quad (21)$$

Obviously:

$$\frac{\partial \ln p(l|x)}{\partial y_k^t} = \frac{1}{p(l|x)} \cdot \frac{\partial p(l|x)}{\partial y_k^t} \quad (22)$$

Finally, to propagate the gradient backward through layer *soft max*, we differentiate the output result u_k^t of the objective function. Since:

$$y_k^t = \frac{e^{u_k^t}}{\sum_k e^{u_k^t}} \quad (23)$$

Therefore:

$$\frac{\partial y_j^t}{\partial u_k^t} = y_j^t (\delta_{jk} - y_k^t) \quad (24)$$

Among them, $\delta_{jk} = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}$, so:

$$\frac{\partial O^{ML}(\{(x,l)\}, N_w)}{\partial u_k^t} = y_k^t - \frac{1}{y_k^t Z_t} \sum_{s \in \text{lab}(l,k)} \hat{\alpha}_t(s) \hat{\beta}_t(s) \quad (25)$$

Among them:

$$\begin{aligned} Z_t &= \sum_{s=1}^{|\mathcal{I}|} \frac{\hat{\alpha}_t(s) \hat{\beta}_t(s)}{y_t^s}; C_t = \sum_s \hat{\alpha}_t(s); \\ \hat{\alpha}_t(s) &= \frac{\alpha_t(s)}{C_t}; D_t = \sum_s \beta_t(s); \hat{\beta}_t(s) = \frac{\beta_t(s)}{D_t} \end{aligned} \quad (26)$$

This gives us the gradient descent formula for CTC:

$$\begin{aligned} u_k^{t'} &= u_k^t - \lambda \frac{\partial O^{ML}(\{(x,l)\}, N_w)}{\partial u_k^t} \\ &= u_k^t - \lambda \left(y_k^t - \frac{1}{y_k^t Z_t} \sum_{s \in \text{lab}(l,k)} \hat{\alpha}_t(s) \hat{\beta}_t(s) \right) \end{aligned} \quad (27)$$

IV. Corpus construction and application research

IV. A. Current Status of Multimodal Corpus Construction for Chinese Dialects

As linguistic research continues to evolve, the field of linguistics has increasingly adopted a “visualization” research paradigm. This new trend significantly incorporates multidimensional elements such as discourse, gestures, eye contact, facial expressions, and emotions into the scope of research, clearly highlighting the “multimodal” direction of linguistic research. This provides us with a more comprehensive and multidimensional perspective for understanding and analyzing linguistic phenomena. In the study, 11 multimodal corpora from abroad were analyzed in depth, revealing the breadth and diversity of their content selection, which offers important insights for the development of a multimodal corpus resource for the Leizhou dialect.

China began introducing multimodal corpus research in 2007, with early focus primarily on ethnic minority languages and literature, aiming to provide effective means for preserving endangered languages. In recent years, with the rapid advancement of multimedia technology and speech engineering research, breakthroughs in technology and methodology have made natural speech, particularly multimodal research, feasible. Scholars such as Cao Lei have successfully constructed a dialect corpus, which enables automatic word segmentation and keyword extraction of annotated corpora.

IV. B. Endangered Lingnan Dialect Corpus Resources

Corpus resources provide an important foundation for research into linguistic phenomena. This section analyzes the development of Suzhou dialect and other domestic and international dialect corpora, and proposes further development strategies.

Endangered Lingnan Dialect Corpus: This corpus was compiled through scientific sampling and processing of linguistic materials from the actual use of endangered Lingnan dialects, ultimately forming a large-scale electronic text corpus. Currently, the construction of corpora for endangered Lingnan dialects is primarily focused on the development of phonetic corpora, with no related literature yet established. Among these, iFlytek's Endangered Lingnan Dialect Preservation Project (currently under expansion) and DataTang's annotated Endangered Lingnan Dialect Mandarin Speech Recognition Corpus are two representative phonetic corpora for endangered Lingnan dialects. The Guangdong Language Resource Audio Database also includes audio materials of endangered Lingnan dialects. The main parameters of the corpus are shown in Table 1, including corpus size, number of participants, age distribution, and collection methods.

From the data in the table, it can be seen that existing corpora of endangered Lingnan dialects are primarily focused on the phonetic level. On the one hand, they only annotate the pronunciation files of endangered Lingnan dialects and their corresponding Mandarin texts. Annotation resources for pinyin notation of the pronunciation of endangered Lingnan dialects, dialect words, and parallel text corpora of Mandarin are still relatively scarce. On the other hand, the collection and annotation of some corpora are still unknown, and it is very likely that the stored information does not meet academic research standards. Additionally, constructing these phonetic corpora requires

significant human and material resources. According to relevant statistics, annotating one hour of phonetic data requires 15 hours of phonetic recordings; accordingly, the larger the corpus, the higher the cost.

Table 1: Dialect corpus

Corpus	Scale	Number	Age distribution	Collection mode	Area
Special plan for dialect protection	1550274	743598	18~30;31~45;46~60	Mobile phone	Lingnan
Dialect mandarin voice recognition corpus	178h	252	16~30;31~45;46~55	Mobile phone	Lingnan
Language resource audio database	325	426	Unknown	Unknown	Lingnan

IV. C. Corpus-based analysis of Lingnan dialect topics

Up to now, a total of 160 audio files have been collected for the research. The total duration of the raw audio corpus is approximately 43 hours, with a total of about 480,000 words. The dialect recognition technology was utilized for the investigation, and the content of the investigation data was statistically analyzed. The statistical results are shown in Table 2. Since it is based on the frequency of character combinations, n is often set as the number of syllables or the corresponding Chinese character to count word forms and their frequencies. In this way, through analysis, all adjacent combinations of Chinese characters in the corpus can be presented, and the frequency of occurrence of each combination can be counted. The most commonly used case in n-gram is when n is from 1 to 3. The following takes a single corpus ("My Hometown", totaling 2,700 words) as an example to briefly explain the process of n-gram. Firstly, by analyzing the 1-gram of "characters", a word frequency table of this corpus can be obtained. The content of the table shows that the word with the highest frequency in this corpus is "ge", which appears 95 times with a frequency of 0.0362. The other words with the highest frequency are "shi", "ah", "di", etc.

Table 2: Statistical result

Gram	Ge	Shi	A	Pa	Ni	Ya	Jiu	You	Yi	Wo
N	95	70	64	63	61	54	48	43	44	35
Frequency	0.0362	0.257	0.0245	0.0236	0.0228	0.0196	0.0178	0.0164	0.0164	0.0128

Taking the character "ge" as an example for 2-gram analysis, high-frequency two-character combinations starting with "ge" can be obtained, and the results are shown in Table 3. The frequency shown in this table is that although the content sample size of a single corpus is not large, it can be initially inferred from the above table. For example:

(1) The frequency of "ge" followed by "(pause)" is very high. After a simple analysis, it is found that the usage of "ge" as a modal particle "de" is extensive.

(2) Some high-frequency words appear, such as "ge di" (those), "ge bian" (over there), "ge zhi" (that), etc. In these examples, "ge" is mainly used as an demonstrative pronoun. In addition, the word "ge" appears. The pronunciations and meanings of the two "ge" are different. They are temporarily transcribed as "ge¹ ge²" to distinguish them. The pronunciation of "ge¹" is [kɔ⁴⁴], with the Yin rising, and it is a demonstrative pronoun "na". "Ge²" [kɔ⁴²³], in a Yin tone, is a measure word.

(3) Although other combinations occur more frequently, it is impossible to determine whether a two-character combination forms a word just by looking at it. Therefore, a broader grammatical context needs to be considered.

Table 3: The frequency statistics of the combination of the two words

Gram	Ge	Gepa	Geshi	Gecun	Gebian	Gewu	Gesan	Gedi	Gezhi	Getao	Gege
Frequency	0.2532	0.0784	0.0563	0.0442	0.0447	0.0442	0.0331	0.0338	0.0221	0.0218	0.0218

The processing of 2-gram can visually present any adjacent two-character combination, and then use the subsequent characters to perform a complete 2-gram algorithm search. After integrating the nodes with strong correlations, the frequency relationship of the following gram can be plotted. Theoretically, it can present all possible combinations of adjacent two-characters and the frequencies at which they occur. Due to space limitations, the data in the paper is not fully displayed. Only the top few with the highest frequencies are shown for illustration purposes. The content in the text box is word form. The direction of the arrow indicates the combinations of the "word form" that appear along the direction of the arrow, and the value on the arrow is the frequency of that combination. For instance, "ge" points to "zhi" with a frequency of 0.022, indicating that the two-character combination "ge Zhi" appears at a frequency of 0.022. Since the corpus is scanned in full from front to back, in the process of class symbol determination, the reverse frequency determination method is usually adopted. For instance, the frequency

at which "kiln" points to "address" in the figure is 1.0000, indicating that there is only one case of "kiln address" in the combination of "kiln address". This suggests that "kiln" and "address" are closely combined. Given the large volume of data in the corpus, it is reasonable to infer that "kiln address" is a class symbol. The statistical results of Chinese character combinations and frequencies are shown in Figure 2.

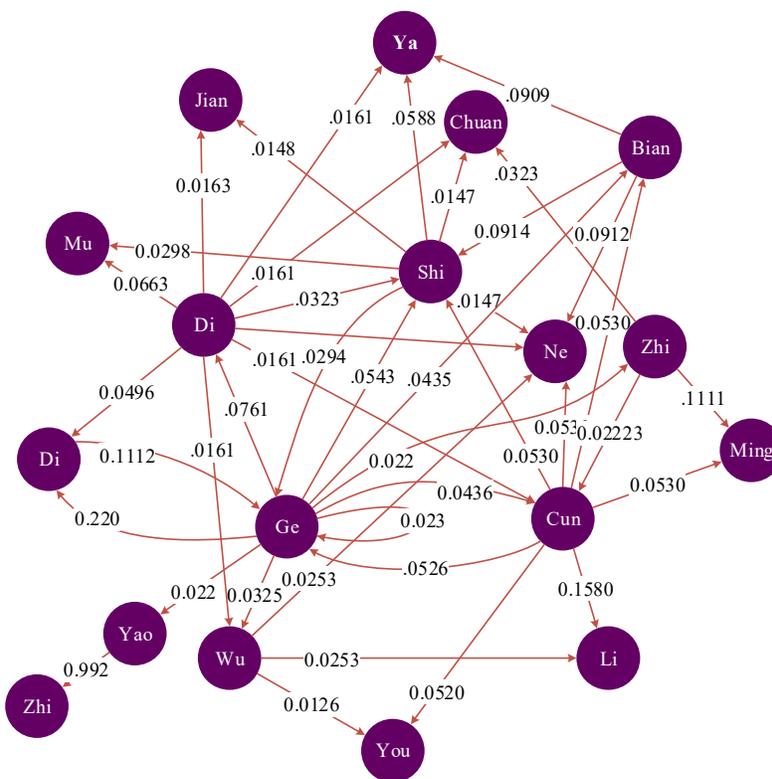


Figure 2: Chinese character combination and frequency statistics sketch

This section presents a list of tokens obtained from the corpus, with the results shown in Table 4. A statistical analysis of the morphological tokens in the Cantonese spoken corpus was conducted in two ranges: multilingual and monolingual. By using the corpus to process the transcribed text of the Cantonese spoken corpus, this section only displays the top 50 tokens.

If the cases of homomorphic synonyms are ignored and only the word forms are examined, the total frequency of the first 30 items in the above table (from "ah" to "in") is approximately 50.22%, while the total number of class symbols is 3,259. This indicates that approximately 0.1% of the class symbols in the Lingnan dialect oral corpus of this study account for nearly half of the corresponding symbol frequencies. Looking at the top 50 words in terms of word form frequency alone, the high-frequency words in Lingnan dialect mostly consist of the following parts:

- (1) Particles, including those that express tone, such as "ah," "ne," "sheng (like this)," "en," "ma," "bo," and "ge," those that express the completion of an action, such as "kai," and those that express comparison, such as "sheng sheng (like this, like that)."
- (2) Pronouns, such as personal pronouns "you," "I," and "he," demonstrative pronouns "ge [kɔ⁴⁴]" (near) and "ya" (far), and adverbial pronouns "dàn (then)" and "dan wen (then)," etc.
- (3) Adverbs, such as negative adverbs like "shou (not; no)" and degree adverbs like "hao."
- (4) Unmarked quantifiers, such as "ge [kɔ⁴²³]," "lu," "di", etc.
- (5) Commonly used basic vocabulary, among which there are many noun-meaning verbs, such as "is", "see [t ɐ i⁴⁴](see)", "speak (say); "Speak", "talk", "chicken", "tie". Among them, "is" and "tie" are both judgment verbs, but "is" is used much more frequently than "tie".

Table 4: The lexical frequency statistics of the material

Class character	Ranking	Frequency	Number of languages	Standard frequency ratio	Standard representation
A	1	10864	60	77086.551	0.96
Ge	2	8852	62	62857.06	0.976
Ni	3	5147	60	36522.202	0.96
Jiang	4	5000	59	3.35513E7	0.943
Lai	5	3689	59	26192.857	0.943
Pa	6	3452	57	24388.418	0.911
Shi	7	3420	62	24281.857	0.976
Jiu	8	3219	60	22846.831	0.96
Ganwen	9	2730	58	19351.47	0.895
Ju	10	2645	57	18776.188	0.911
Mo	11	1964	55	13924.095	0.895
You	12	1533	60	10869.336	0.96
Hao	13	1430	62	10151.823	0.976
You	14	1425	58	10094.991	0.911
Yapa	15	1342	54	9505.351	0.847
Suoyi	16	1224	58	610.236	0.927
Ren	17	1132	60	8013.493	0.96
Luo	18	1108	55	7992.18	0.879
Ni	19	1100	54	7800.37	0.863
Hua	20	994	59	7033.138	0.928
Yao	21	967	59	6848.432	0.928
Youpa	22	920	50	6514.54	0.799
Yi	23	913	58	6471.915	0.928
De	24	895	58	6336.937	0.928
Qu	25	887	58	6273.001	0.928
Wo	26	830	43	5868.068	0.654
Dao	27	819	58	5782.819	0.928
Xi	28	819	54	5782.819	0.864
Do	29	810	59	5747.298	0.944
Zai	30	674	60	4759.83	0.961
Zuo	31	542	58	3814.986	0.912
Zhong	32	536	51	3786.569	0.815
Yingwei	33	528	55	3708.425	0.88
Ya	34	496	52	3438.469	0.815
Hui	35	471	57	3282.179	0.912
Chu	36	469	56	3267.971	0.896
Ma	37	460	45	3253.763	0.719
Duo	38	450	56	3182.722	0.88
Shang	39	443	50	3125.889	0.799
Bo	40	438	44	3104.577	0.703
Yong	41	436	48	3076.161	0.767
Kai	42	428	54	3012.224	0.848
Bijiao	43	428	56	3012.224	0.864
Di	44	425	44	2998.016	0.703
En	45	411	43	2912.767	0.686
Lu	46	394	48	2777.789	0.767
Gansheng	47	389	42	2720.956	0.67
Chi	48	386	39	2706.748	0.606
Yiban	49	374	50	2614.395	0.783
Da	50	362	56	2543.354	0.864

V. Testing and evaluation of the Lingnan dialect corpus

The primary focus is on the effectiveness evaluation of the Lingnan dialect corpus. From a user-centered perspective, this study explores whether the corpus effectively supports users' reading learning and dialect preservation. The user experience test evaluation adopted the SUS (System Usability Scale) as the key assessment tool. Q1: I would be willing to read this corpus frequently. Q2: I find the functions of this corpus too complex. Q3: I find the operations of this corpus easy to use. Q4 I need professional assistance to use this corpus. Q5 I think the various functions of this corpus are well integrated. Q6 I think there are too many inconsistent aspects of this corpus. Q7 I think most people can learn to use this corpus quickly. Q8 I find this corpus difficult to use. Q9 I feel confident when using this corpus. Q10 I need to learn a lot to be able to use this corpus.

The experiment recruited 11 participants to test the experimental system, with participant IDs ranging from "A" to "K," to conduct a comprehensive evaluation of the system. Subsequently, we carefully compiled the scoring results of these 11 participants and organized them into the following table, with the test results shown in Table 5. According to the summary results, the average SUS score was 82.40. This study not only helps us gain a deeper understanding of users' actual needs and pain points during corpus usage but also provides important references and basis for corpus improvements and upgrades, contributing to the continuous optimization of corpus design and user experience.

Table 5: SUS scale topic

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	SUS
A	4	2	5	1	3	2	4	2	3	1	78.3
B	4	3	4	2	4	2	3	3	4	4	68.3
C	5	3	3	1	4	3	4	3	3	2	73.3
D	4	2	4	3	3	3	5	3	4	4	68.3
E	4	1	5	1	5	2	5	2	4	1	90.8
F	5	2	5	1	4	1	5	2	5	1	95.8
G	5	4	3	3	5	3	4	3	4	2	65.8
H	5	1	5	2	5	2	5	3	5	1	95.8
I	3	3	4	2	5	3	5	1	5	1	80.8
J	4	1	5	1	3	1	5	1	5	1	95.8
K	5	1	4	1	4	1	5	2	5	1	93.3

As shown in Figure 3, the overall SUS score falls within the range of "B" to "A." A detailed analysis of the experimental data reveals significant differences among the subjects in their assessment of ease of use.

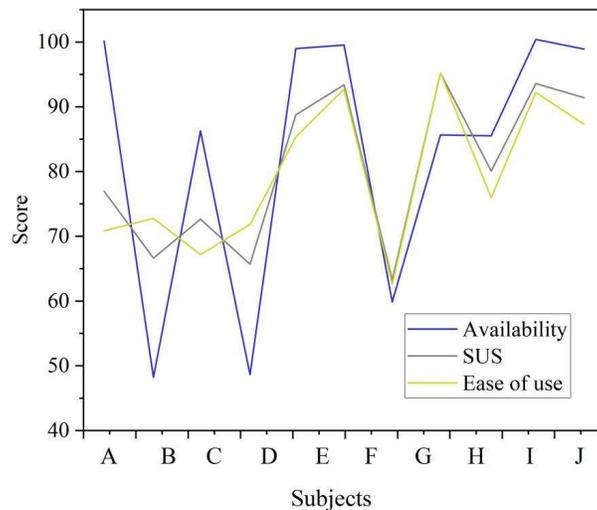


Figure 3: SUS scale score line diagram

VI. Conclusion

This paper utilizes four major modules of the CNN-CTC acoustic model, namely MFCC speech feature extraction, CTC-GRU model training, CTC maximum path method decoding and prediction evaluation, to achieve mapping

recognition from Lingnan dialect to Mandarin, and establish a corpus to provide digital protection for the endangered Lingnan dialect. Empirical research shows that, taking a single corpus ("My Hometown", a total of 2,700 words) as an example, "ge" appears most frequently, followed by "shi", "ah", "di" and other examples. The frequency of the two-character combination "ge zhi" is 0.0220. By evaluating the effect of the Lingnan dialect corpus, it can be found that the average value of SUS is 82.40, and it shows obvious differences in terms of learnability.

Funding

General project of the National Social Science Fund "Character Rhyme Group Books Dialect Phonetic Research of Lingnan and Literature Pedigree Sorting" (Project number: 21BY130).

References

- [1] Huang, Y., & Fang, F. (2024). 'I feel a sense of solidarity when speaking Teochew': unpacking family language planning and sustainable development of Teochew from a multilingual perspective. *Journal of Multilingual and Multicultural Development*, 45(5), 1375-1391.
- [2] McCullough, E. A., Clopper, C. G., & Wagner, L. (2019). Regional dialect perception across the lifespan: Identification and discrimination. *Language and speech*, 62(1), 115-136.
- [3] Wichmann, S. (2020). How to distinguish languages and dialects. *Computational Linguistics*, 45(4), 823-831.
- [4] Wang, H. (2020, February). Inheritance and development of Chinese dialect culture. In 6th International Conference on Education, Language, Art and Inter-cultural Communication (ICELAIC 2019) (pp. 924-926). Atlantis Press.
- [5] Wang, C. (2018, January). Dialect Protection from the Perspective of Natural Characteristics Loss of Dialects. In 2017 5th International Education, Economics, Social Science, Arts, Sports and Management Engineering Conference (IEESASM 2017) (pp. 370-373). Atlantis Press.
- [6] Dodsworth, R. (2017). Migration and dialect contact. *Annual Review of Linguistics*, 3(1), 331-346.
- [7] Berruto, G. (2018). 18. The languages and dialects of Italy. W. Ayres-Bennett & J. Carruthers. (Eds.), *Manual of Romance sociolinguistics*, 494-525.
- [8] Jiang, B. (2023, March). The Preservation and Inheritance of Chinese Dialects: Reflections on the Phenomenon of Chinese Dialect Loss in the New Era. In 9th International Conference on Education, Language, Art and Inter-cultural Communication (ICELAIC 2022) (pp. 13-18). Athena Publishing.
- [9] Luqiu, L. R. (2018). Counter-hegemony: grassroots use of the Internet to save dialects in China. *Journal of Multilingual and Multicultural Development*, 39(8), 663-674.
- [10] Forkel, R., & Hammarström, H. (2022). Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web*, 13(6), 917-924.
- [11] Khered, A., Abdelhalim, I. A., & Batista-Navarro, R. T. (2022, December). Building an ensemble of transformer models for Arabic dialect classification and sentiment analysis. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)* (pp. 479-484).
- [12] Huang, Q. (2023). Current situation and path of foreign minority language protection based on Internet of Things from the perspective of ethnic identity. *Journal of Computational Methods in Science and Engineering*, 23(5), 2677-2686.
- [13] Li, Q., Mai, Q., Wang, M., & Ma, M. (2024). Chinese dialect speech recognition: a comprehensive survey. *Artificial Intelligence Review*, 57(2), 25.
- [14] Leemann, A., Kolly, M. J., & Britain, D. (2018). The English Dialects App: The creation of a crowdsourced dialect corpus. *Ampersand*, 5, 1-17.
- [15] Zhao, Y., Ding, Y., & Min, X. (2025). Construction of a multimodal dialect corpus based on deep learning and digital twin technology: A case study on the Hangzhou dialect. *Journal of Computational Methods in Sciences and Engineering*, 25(2), 1448-1460.
- [16] Szmrecsanyi, B., & Anderwald, L. (2017). Corpus-based approaches to dialect study. *The handbook of dialectology*, 300-313.
- [17] Dash, N. S., Ramamoorthy, L., Dash, N. S., & Ramamoorthy, L. (2019). Corpus and dialect study. *Utility and Application of Language Corpora*, 139-153.
- [18] Keränen, M. (2018). Language maintenance through corpus planning—the case of Kven. *Acta Borealia*, 35(2), 176-191.
- [19] Xu, J., Zhou, C., & Liu, H. (2024). Cultural heritage as a key motivation for sustainable language protection: a case study of the Suzhou dialect protection project. *Journal of Multilingual and Multicultural Development*, 1-18.
- [20] Han, Q. (2023). The preservation and transmission of dialect culture in the context of language ecology. *Journal of Humanities, Arts and Social Science*, 7(8).
- [21] Zou, C. (2022). Inter-generational language shift and maintenance: language practice observed in Guangzhou Hakka families. *Asian Ethnicity*, 23(2), 362-376.
- [22] He, Z., & Li, H. (2017, January). Research on Protection and Construction of Ethnic Language Resources in Northeast China. In 2016 2nd International Conference on Economics, Management Engineering and Education Technology (ICEMEET 2016) (pp. 1057-1062). Atlantis Press.
- [23] Han, M., Zhu, D., Wen, X., Shu, L., & Yao, Z. (2024). Research on dialect protection: interaction design of Chinese dialects based on BLSTM-CRF and FBM theories. *IEEE Access*, 12, 22059-22071.
- [24] Tan, Y., & Jehom, W. J. (2024). The Function of Digital Technology in Minority Language Preservation: The Case of the Gyalrong Tibetan Language. *Preservation, Digital Technology & Culture*, 53(3), 165-177.
- [25] Bai, A. (2022). Digitalization Mining and Protection of Security Language Resources in Big Data Environment. In *MATEC Web of Conferences* (Vol. 365, p. 01022). EDP Sciences.
- [26] Nikita Muravyev. (2025). The Emergence of Hierarchical Alignment in Northern Khanty: A Comparative Dialectal Corpus Study. *Linguistica Uralica*, 61(2), 106-130.
- [27] Nasir Tayyab & Malik Muhammad Kamran. (2024). Efficient CRNN: Towards end-to-end low resource Urdu text recognition using depthwise separable convolutions and gated recurrent units. *Information Processing and Management*, 61(1).