ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Research on Innovative Ways of Chinese Dialect Protection and Cultural Inheritance Assisted by digital Technology

Enhui Wang[1,*]

[1] Hunan Urban Construction College, Xiangtan 411101, Hunan, China

SUMMARY: *A digital modeling method for dialect protection and cultural inheritance was proposed to solve the problems of dispersive resources of Chinese dialects, difficulties in phonetic transcription and insufficient cultural semantic correlation. The CDCC-2025 Chinese dialect culture corpus is constructed, which contains six types of dialect areas, 120 hours of speech, 36000 speech samples, 18000 dialect texts, 6200 cultural entities and 14800 relation triples. Method The acoustic coding model is used to complete dialect speech recognition, and the two-tower semantic alignment algorithm is used to establish the mapping between dialect words and Mandarin paraphrases. The regional cultural knowledge graph and attention fusion mechanism are combined to realize multimodal resource retrieval. Experimental results show that the dialect recognition accuracy of the proposed method reaches 94.2%, the F1-score is 93.5%, and the WER is reduced to 8.9%. Semantic retrieval Recall@10 reaches 91.6% and MRR reaches 0.846. Triple F1 of knowledge graph achieves 89.3%. The results show that the proposed method can improve the effect of dialect speech recognition, cultural semantic organization and digital retrieval.*

KEYWORDS: *Chinese dialect protection; Cultural inheritance; Speech recognition; Knowledge graph*

1 Introduction

Chinese dialects carry regional phonetic systems, lexical expressions, folk narrative structures and local cultural memories, which are important objects for the protection of language resources and cultural inheritance. The protection of traditional dialects mainly relies on manual recording, text sorting and paper archives preservation, which has problems such as long collection cycle, insufficient consistency of annotation, and limited resource retrieval and reuse ability. With the development of speech recognition, deep learning, knowledge graph and multi-modal retrieval technology, dialect resource protection has gradually shifted from static records to digital processing modes that can be calculated, searchable and interactive. Existing studies have systematically sorted out the task of Chinese dialect speech recognition, and pointed out that dialect acoustic differences, sample scarcity, cross-regional pronunciation variation and inconsistent labeling standards are still important factors restricting the performance of the model [1]. Focusing on the construction of speech data sets for specific dialects such as Cantonese and Hokkien, related studies have constructed basic data resources for automatic speech recognition from the aspects of corpus collection, audio segmentation, text transcription and pronunciation tagging [2, 3]. These studies provide a data basis for the digital protection of dialects, but the modeling of the association between dialect words,

*19607329710@163.com
https://doi.org/10.65102/is20261064
Published: 10 June, 2026

regional cultural scenes, folk custom narratives and inherited knowledge is still insufficient.

At the model level, self-supervised speech representation learning provides a new technical support for low-resource language and dialect processing. Related studies have shown that self-supervised pre-training models can learn acoustic structures from large-scale unlabeled speech and improve the feature generalization ability of downstream recognition tasks [4]. The end-to-end speech recognition model reduces the isolated dependence between traditional acoustic models, pronunciation dictionaries and language models, and makes the mapping process from dialect speech input to text output more unified [5]. WavLM, XLS-R and other models further improve the ability of cross-language, cross-accent and multi-task speech processing, and provide a transferable model basis for dialect acoustic feature coding [6, 7]. The weakly supervised large-scale speech recognition model shows strong robustness in complex noise environment and cross-scene speech recognition, which can provide reference for dialect audio processing under real acquisition conditions [8]. However, a single speech recognition model mainly solves the problems of "hearing" and "transcribing", which is difficult to fully express the regional cultural semantics behind dialects.

Cultural inheritance not only needs to preserve dialect speech and text, but also needs to establish a structured association between dialect expression and regional cultural knowledge. Knowledge graph technology can organize people, places, folk activities, object names, dialect words and cultural events into nodes and relationships, which provides support for semantic retrieval, association reasoning and knowledge services of dialect cultural resources [9]. Research on cultural heritage knowledge graph and semantic portal shows that structured knowledge modeling can enhance the ability of associated display, path discovery and cross-platform sharing of cultural resources [10, 11]. Pre-trained language models and cue learning methods also provide new implementation paths for semantic alignment, cultural semantic annotation and text retrieval between dialect words and Mandarin paraphrasing [12]. It can be seen that the digital protection of Chinese dialects should not stop at the level of speech recognition, but also include acoustic features, text semantics, cultural labels and knowledge relations into a unified technical framework.

Based on the above research basis, this paper focuses on the needs of Chinese dialect protection and cultural inheritance, and constructs a technical route of "dialect speech, text and cultural scene data collection - acoustic feature coding and semantic alignment - regional cultural knowledge graph construction - multi-modal fusion retrieval - system implementation and experimental result analysis". The research focuses on the structural processing of dialect cultural corpus, dialect speech recognition and acoustic feature coding, semantic alignment of dialect words and Mandarin, construction of regional cultural knowledge graph, design of multi-modal fusion retrieval algorithm, and implementation of digital inheritance system. Through dataset construction, comparative experiment, ablation experiment and system performance test, the effectiveness of the proposed method in dialect identification, semantic retrieval, knowledge organization and system service is verified, which provides technical solutions for the intelligent protection and engineering inheritance of Chinese dialect resources.

2 Collection and structural processing of Chinese dialect culture corpus

2.1 Data collection of dialect speech, text and cultural scene

Chinese dialect cultural corpus collection takes "voice recording-text transcription-cultural

scene association" as the basic data link, and integrates dialect pronunciation, vocabulary expression, folk custom narrative, regional objects and inheritance scenes into the same collection unit [13]. Each sample not only saves the audio file, but also synchronously records the speaker number, collection location, dialect area, transcribed text, Mandarin interpretation, cultural label and scene source, to avoid the problem of disconnection between speech data and cultural semantics in subsequent model training. In the process of collection, speech data mainly come from vocabulary reading, short sentence expression and natural narrative recording, text data includes manual transcription text, Mandarin interpretation text and local culture description text, and cultural scene data includes folk custom activities, traditional artifacts, place names, festival ceremonies and local life scenes [14]. To ensure that the data can be used for subsequent acoustic modeling, semantic alignment and knowledge graph construction, the collection results are organized into multi-field sample sets according to uniform numbering rules, and the form can be expressed as follows.

$$D=\{x_i \mid x_i=(a_i,t_i,m_i,r_i,c_i,s_i),i=1,2,\ldots,N\} \quad (1)$$

where, D represents the corpus collection of dialect culture; Let xi denote the ith sample; ai represents dialect speech data. ti represents the dialect transliteration text; mi represents Mandarin paraphrase text; ri denotes the locale label. ci stands for culture category label. si denotes the cultural scene source; N denotes the total number of samples. This structure enables speech recognition, text alignment, and cultural knowledge modeling to invoke the same data index.

In order to form a traceable data structure for speech files, transcribed texts, and cultural scene information in the collection stage, this paper divides the collection fields into four categories: dialect speech, dialect text, cultural scene, and metadata, and the detailed configurations are shown in Table 1.

Table 1: Configuration Table of Dialect Culture Corpus Collection Types and Data Fields

Data Type	Collection Content	Field Configuration
Dialect Speech	Vocabulary reading, short sentence expression, natural narrative recording	Audio ID, Speaker ID, Region, Duration
Dialect Text	Dialect transcription text, Mandarin interpretation text	Dialect Text, Mandarin Text, Token ID
Cultural Scene	Folk activities, traditional artifacts, place names, festival rituals	Scene Type, Cultural Tag, Entity Label
Metadata	Collection time, collection location, dialect area, speaker information	Time, Location, Dialect Zone, Speaker Attribute

After the field configuration is determined, the multi-source data needs to go through uniform numbering, field verification and format conversion before entering the subsequent feature extraction, semantic alignment and knowledge modeling processes. The overall collection and storage process is shown in Figure 1.

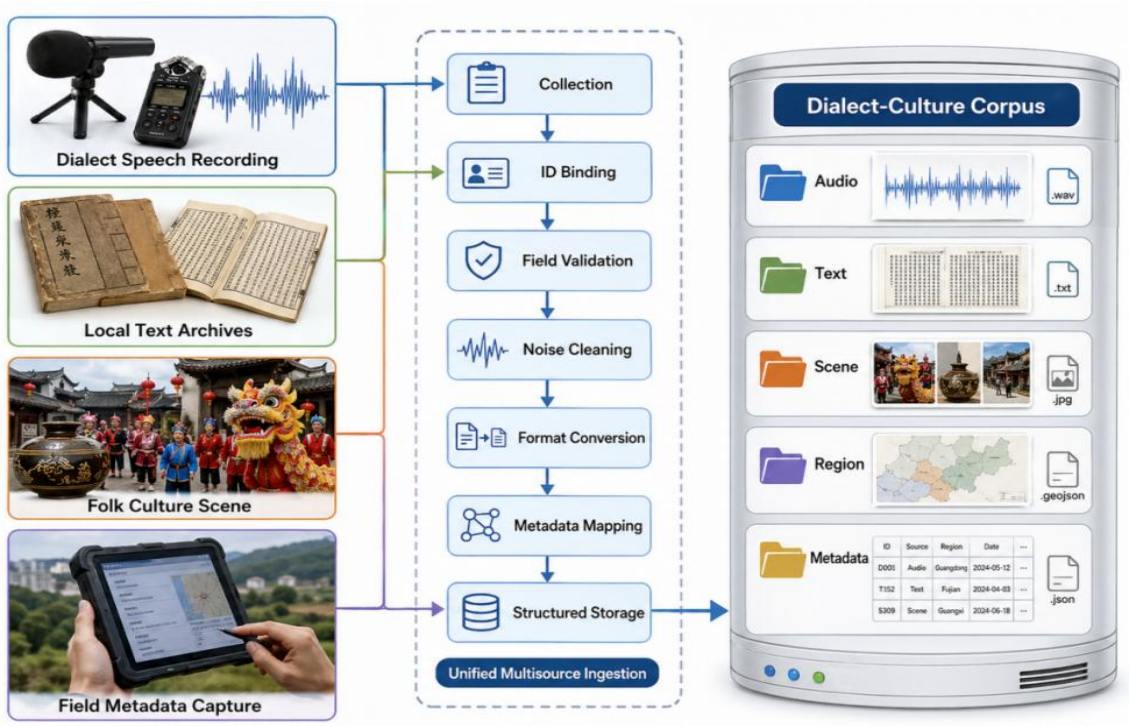

Figure 1: Flow chart of multi-source data collection and storage of Chinese dialect culture

Through the above collection methods, dialect speech, text interpretation and cultural scene are no longer stored as isolated materials, but form linked data objects with uniform sample numbers. The subsequent acoustic feature extraction can directly call the audio field, the semantic alignment model can call the dialect text and Mandarin interpretation field, and the knowledge graph construction can call the cultural label, regional label and entity field, so as to ensure that the two tasks of "dialect protection" and "cultural inheritance" have a common data basis from the data entrance [15].

2.2 Speech preprocessing and dialect acoustic feature extraction

After dialect speech collection is completed, the original audio usually contains environmental noise, silent segments, volume fluctuations and sampling differences caused by different devices, and direct input into the recognition model will reduce the stability of acoustic representation [16]. In this paper, the audio is uniformly converted into mono signal, and resampled according to a fixed sampling rate. Then the endpoint detection is used to remove the long silence interval, so that the effective speech segments are concentrated in the modelable range. Let the original speech sequence be x(n), and the MTH speech frame after windowing can be expressed as follows.

$$x_m(n)=x(n+mH)w(n), \quad 0\le n<L \tag{2}$$

where, xm(n) represents the MTH windowed speech; H represents frame shift; L denotes the frame length; w(n) is the window function. This processing segments continuous speech into local stationary segments, which provide input for subsequent frequency domain analysis.

Aiming at the background noise in the acquisition scene, this paper uses spectral subtraction to correct the speech amplitude spectrum. Let Ym(k) be the spectrum of the noisy speech at the MTH frame and KTH frequency point, and N(k) be the estimated noise spectrum, then the enhanced amplitude spectrum is as follows.

$$\hat{S}_m(k)=\max\left(|Y_m(k)|-\alpha|\hat{N}(k)|,\beta|Y_m(k)|\right) \tag{3}$$

where, $\hat{S}_m(k)$represents the speech amplitude spectrum after noise reduction; α is the noise suppression coefficient; β represents the spectral lower bound coefficient, which is used to avoid excessive weakening of weak speech components.

After noise reduction, the system maps the speech spectrum to Mel filter banks to extract acoustic features that can reflect variations in dialect initials, vowels, tone, and articulatory intensity. The energy of the MTH frame on the BTH Mel filter can be expressed as follows.

$$F_m(b)=\log\left(\sum_{k=1}^{K}|\hat{S}_m(k)|^2 M_b(k)+\varepsilon\right) \tag{4}$$

where, $F_m(b)$ represents the Mel acoustic feature of the MTH frame; $M_b(k)$is the weight of the BTH Mel filter at frequency k. K denotes the number of frequency points; Let ε be the smoothing term. The acoustic matrix composed of the features of each frame can be used as the input of the subsequent dialect speech recognition model.

To ensure that the preprocessing results can be stably transferred to the acoustic encoding stage, the original audio, valid speech segments, enhanced spectrum and Mel feature matrix are organized as continuous processing links in this paper, and the detailed process is shown in Figure 2.

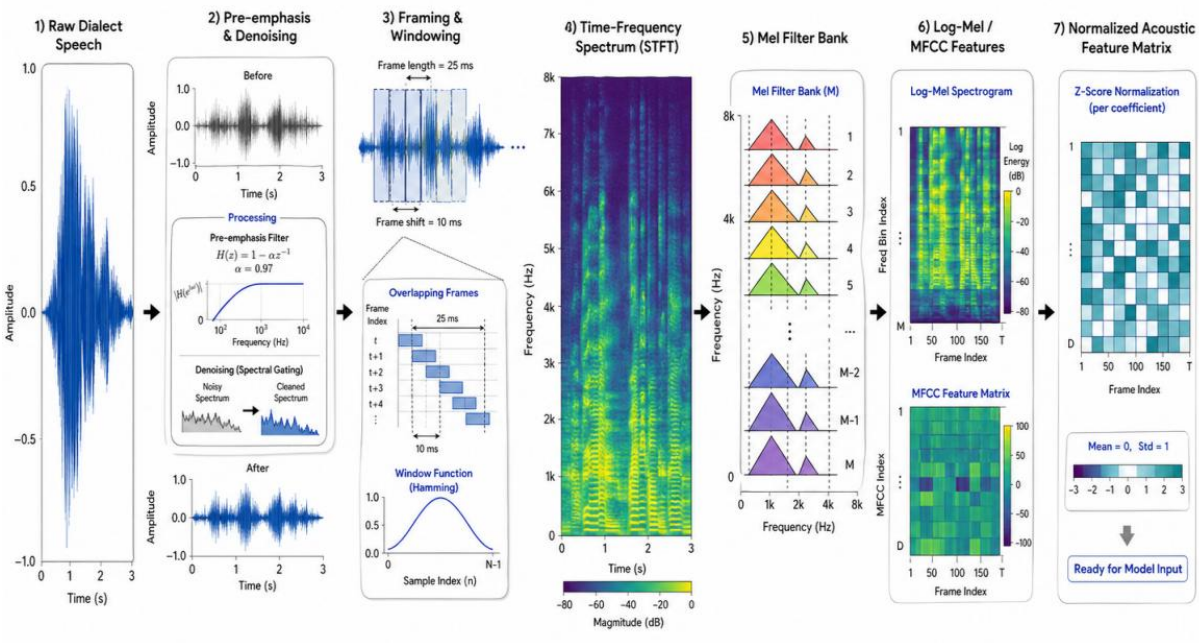

Figure 2: Flow chart of dialect speech preprocessing and acoustic feature extraction

Through this process, the original dialect speech is transformed into an acoustic feature matrix with uniform dimension and stable spectrum structure, which not only retains the differences in tone, prosody and phoneme in dialect pronunciation, but also reduces the interference of environmental noise and device differences on subsequent recognition models.

2.3 Annotation and unified coding of dialect culture corpus

After the collection and preprocessing of dialect culture corpus, it is necessary to incorporate

speech, text, regional and cultural information into a unified labeling system, so that data from different sources can be co-invoked by the model [17]. In this paper, the annotation objects are divided into speech layer, text layer, culture layer and region layer. The speech layer records syllables, tones, pauses and pronunciation variants. In the text layer, dialect words, Mandarin paraphrases, word segmentation results and semantic categories were recorded. The cultural layer records folk activities, traditional artifacts, place names, festival ceremonies and other entities and their relationships. The regional level records the collection location, dialect area and cultural scene source. Let the labeling result of the sample in article i be yi, whose structure can be expressed as follows.

$$y_i=(p_i,q_i,e_i,g_i), \quad i=1,2,\ldots,N \tag{5}$$

where, pi represents the annotation result of speech layer; qi represents the text-level annotation result; ei represents the annotation results of cultural entities and relations; gi denotes the locational label. N denotes the total number of samples. The annotation structure binds dialect pronunciation, lexical interpretation and cultural object under the same sample number, which provides the basis for subsequent semantic alignment, relation extraction and multi-modal retrieval.

In order to reduce the format differences between different annotation layers, this paper maps speech features, text features, cultural labels and geographical labels into a unified coding space. Let the speech encoding vector be v_i^a, the text encoding vector be v_i^t, the cultural label encoding vector be v_i^c, and the region encoding vector be v_i^r. The unified sample encoding can be expressed as follows.

$$z_i=\phi(W_a v_i^a+W_t v_i^t+W_c v_i^c+W_r v_i^r+b) \tag{6}$$

where, zi represents the unified coding representation of the ith sample; Wa, Wt, Wc, Wr represent the projection matrices of speech, text, cultural, and regional features, respectively. b represents the bias term; Let $\phi(\cdot)$ denote the nonlinear mapping function. Through unified coding, data from different sources are converted into vector representations with consistent dimensions, and subsequent models can complete recognition, matching and retrieval in the same feature space.

In order to ensure that the annotation results can directly serve model training, semantic retrieval and cultural knowledge modeling, this paper sets the annotation objects and coding rules according to the information hierarchy, as detailed in Table 2.

Table 2: Dialect Culture Corpus Annotation System and Coding Rules

Annotation Level	Annotation Object	Coding Rule
Speech Layer	Syllables, tones, pauses, pronunciation variants	Phoneme ID, Tone ID, Pause ID
Text Layer	Dialect vocabulary, Mandarin interpretation, word segmentation results, semantic categories	Token ID, Alignment ID, Segment ID
Cultural Layer	Folk activities, traditional artifacts, place names, festival rituals	Entity ID, Relation ID, Scene ID
Regional Layer	Collection location, dialect area, cultural scene source	Region ID, Dialect Zone ID, Source ID

After unified labeling and coding, the dialect culture corpus is transformed into a computable sample from the original collection material. The annotation at the speech layer supports acoustic feature learning, the annotation at the text layer supports the alignment between dialect words and Mandarin paraphrasing, and the annotation at the culture layer and region layer supports cultural entity extraction, knowledge relationship modeling, and multi-modal retrieval, so as to provide structured input for subsequent intelligent modeling.

3 Intelligent modeling method for dialect protection and cultural inheritance

3.1 Dialect speech recognition and acoustic feature coding model

The dialect speech recognition model takes the acoustic feature matrix generated in Chapter 2 as input, and aims to extract the context representation with regional differences, tone variations and phoneme combination features from continuous speech frames, and output the corresponding dialect transcription sequence. Considering that Chinese dialects have differences in initial and final combinations, tone contour variations, and cross-regional pronunciation shifts, we adopt a modeling structure of "acoustic feature input-convolutional downsampling-Transformer context encoder-CTC sequence decoding". The convolutional layer is used to compress local redundant frames and enhance short-time spectral patterns, the Transformer encoder is used to capture long-distance speech context, and the CTC decoding layer is used to deal with the problem that the speech frame sequence is inconsistent with the length of the text label [18].

Let the ith speech sample be preprocessed to obtain the acoustic feature matrix Xi, whose input form can be expressed as follows.

$$X_i=[f_{i,1},f_{i,2},\ldots,f_{i,T_i}],\quad f_{i,t}\in R^{d_a} \tag{7}$$

where, Xi represents the acoustic feature matrix of the ith speech sample; fi,t denotes the acoustic features of the t-th frame; Ti represents the frame sequence length; da represents the acoustic feature dimension. The model first reduces the frame-level input length through the convolutional downsampling module, and then sends the result to the Transformer encoder. The hidden state update process of the encoder at layer l can be expressed as follows.

$$H^{(l)}=\mathrm{FFN}\left(\mathrm{MHA}\left(H^{(l-1)}\right)+H^{(l-1)}\right) \tag{8}$$

where, H(l) represents the coding output of layer l; MHA(·) denotes multi-head self-attention computation; FFN(·) denotes the feedforward network mapping. The proposed structure is able to establish global dependencies between acoustic frames, enabling the model to focus not only on local phoneme changes, but also to exploit prosodic and contextual information in complete short sentences.

Aiming at the problem that the speech frame sequence and the dialect paraphrase text cannot be aligned frame by frame, this paper introduces the CTC decoding mechanism. Given an output label sequence y, the set of all paths that collapse into y is denoted B−1(y), and its conditional probability can be expressed as follows.

$$P(y|X)=\sum_{\pi\in B^{-1}(y)}\prod_{t=1}^{T}P(\pi_t|H_t) \tag{9}$$

where, π denotes CTC candidate path; B(·) is the collapse function to remove whitespace and duplicate labels. Ht represents the ttt frame coding representation; P(πt ∣ Ht) represents the probability of output label in the current frame. In the training phase, the negative log-likelihood is used as the recognition loss, and the model optimization objective is as follows.

$$L_{asr}=-\sum_{i=1}^{N}\log\ P(y_i|X_i) \tag{10}$$

where, Lasr represents dialect speech recognition loss; yi denotes the dialect transcription label corresponding to the ith speech sample; N denotes the number of training samples. By minimizing this loss, the model is able to learn the mapping relationship between acoustic frames and dialect text sequences.

In order to ensure that the model training parameters are consistent with the configuration of subsequent experiments, the main modules, parameter Settings and roles of the acoustic coding model are uniformly configured in this paper, as detailed in Table 3.

Table 3: Dialect Speech Recognition Model Parameters and Training Configuration

Module	Parameter Setting	Function Description
Acoustic Feature	80-dim Mel Filter Bank	Construct frame-level acoustic input matrix
CNN Subsampling	2-layer CNN	Compress frame length and enhance local spectral features
Transformer Encoder	6 layers	Extract long-distance speech context representation
Hidden Size	768	Control acoustic encoding vector dimension
Attention Heads	8	Establish multi-subspace speech dependency relationship
Decoder	CTC Layer	Complete non-aligned speech sequence decoding
Batch Size	32	Maintain gradient estimation stability
Learning Rate	1e-4	Control model parameter update step size

The process of dialect speech recognition needs to connect acoustic features, context encoding and sequence decoding into a complete network structure, and the internal data transmission relationship of the model is shown in Figure 3.

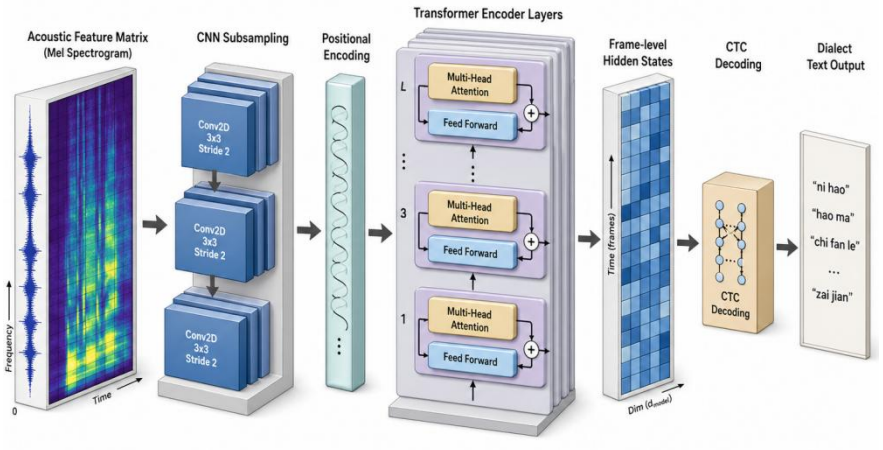

Figure 3: Structure diagram of dialect speech recognition and acoustic feature encoding model

The short-time spectrum difference, tone variation and contextual pronunciation pattern in dialect speech are uniformly encoded into frame-level hidden representation, and then CTC is used to complete dialect text output. Compared with the traditional separated acoustic modeling method, this structure reduces the dependency on artificial pronunciation dictionary, and can adapt to the common pronunciation variants, speech rate changes and local missing phenomena in dialect speech, which provides a stable text input for the subsequent semantic alignment of dialect words and cultural knowledge modeling.

3.2 Semantic alignment algorithm between dialect words and Mandarin

The transliterated text output from dialect speech recognition still retains strong regional expression characteristics. If you directly enter the retrieval or knowledge modeling process, it is easy to have problems such as synonymous with different forms, the same word with different meanings, and the lack of Mandarin paraphrasing. In order to establish a stable mapping relationship between dialect words and Mandarin semantics, this paper adopts a double-tower semantic alignment structure. The dialect vocabulary text and Mandarin paraphrases text are input into independent encoders respectively, and the matching degree of the two types of expressions is calculated in the shared semantic space. The method does not require that dialect words and Mandarin paraphrasing are completely consistent in literal form, but learns the semantic correspondence between them through context encoding and similarity constraints.

Let the ith dialect lexical sequence be di=(di,1,di,2,... ,di,n), which is mapped by the dialect text encoder to obtain the semantic representation:

$$u_i=\text{Encoder}_d(d_i) \tag{11}$$

where, ui represents the semantic vector of dialect words; Encoderd(·) denotes the dialect text encoder; di denotes the ith dialect word or phrase sequence. The Mandarin paraphrase sequence is denoted as mj=(mj,1,mj,2,... ,mj,l), which is mapped by the Mandarin paraphrasing encoder.

$$v_j=\text{Encoder}_m(m_j) \tag{12}$$

where, vj represents the semantic vector of Mandarin paraphrase; Encoderm(·) stands for Mandarin text encoder; mj represents the JTH Mandarin paraphrase sequence. The two types of encoders can be initialized by a shared pre-trained language model, and then fine-tuned by the paired samples of dialect word-paraphrasing, so that the dialect expressions, Mandarin interpretation and cultural semantic labels are gradually closer in the vector space.

In the semantic matching stage, this paper uses the normalized cosine similarity to calculate the correspondence strength between the dialect word vector and the Mandarin paraphrase vector. The higher the similarity, the closer the two are in the semantic space, which is calculated as follows.

$$s_{ij}=\frac{u_i^T v_j}{\|u_i\|_2\|v_j\|_2} \tag{13}$$

where, sij represents the semantic similarity between dialect word di and Mandarin paraphrasing mj; $\|.\|_2$ is the two-norm. In order to enhance the model's ability to distinguish synonyms, cultural proper names and regional appellations, the real paraphrases are used as positive samples in the training phase, and other paraphrases in the same batch are used as

negative samples to construct contrastive learning objectives. For the ith dialect word, the semantic alignment loss can be expressed as follows.

$$L_{align}=-\sum_{i=1}^{B}\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp(s_{ij}/\tau)} \tag{14}$$

where, Lalign represents the semantic alignment loss; B is the number of samples in the batch. sii represents the similarity between the ith dialect word and its true Mandarin definition. sij denotes the similarity between the ith dialect word and the JTH candidate paraphrase; Let τ denote the temperature coefficient. The loss promotes the positive sample pairs to be aggregated in the vector space, and the negative sample pairs to maintain the separable distance, so as to improve the accuracy of dialect word interpretation retrieval and Mandarin mapping.

The alignment process between dialect words and Mandarin paraphrases needs to maintain the continuous transitive relationship between bilateral encoding, similarity calculation and contrastive learning constraints, as detailed in Figure 4.

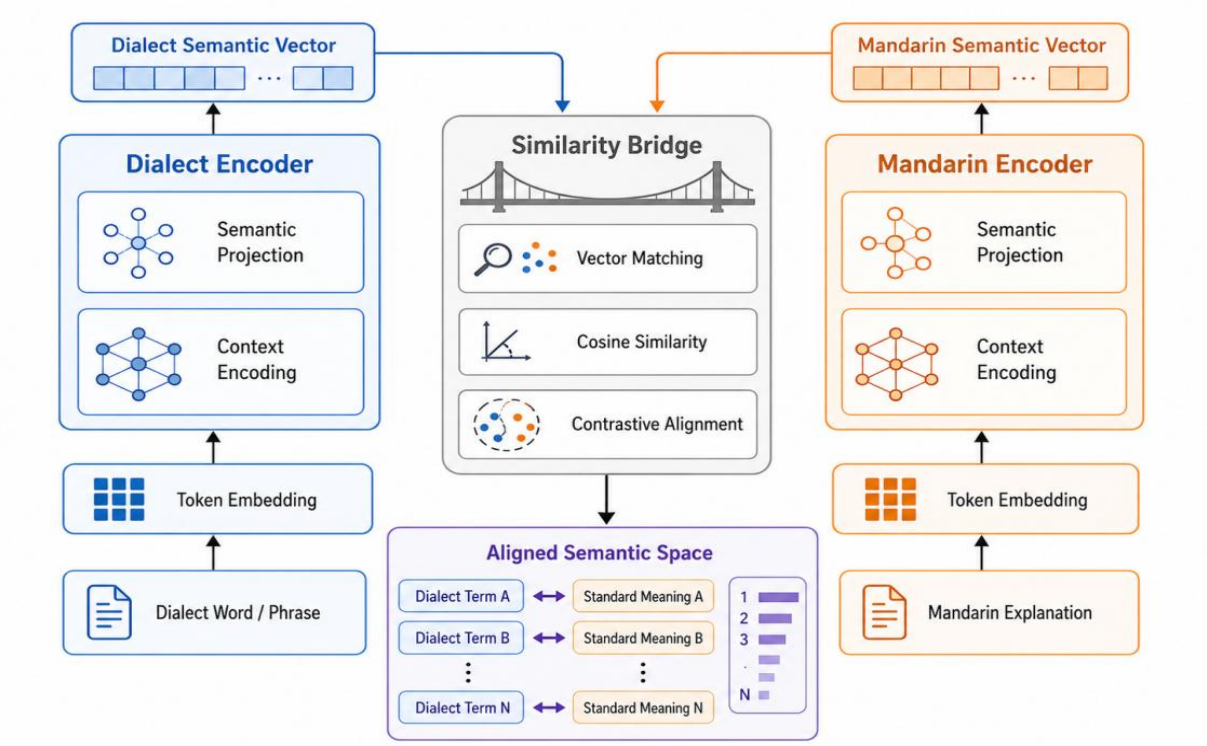

Figure 4: Semantic alignment structure diagram of dialect words and Mandarin paraphrases with two towers

With the above algorithms, the dialect words are able to obtain retrivable and comparable representation results in the Mandarin semantic space. For synonymous and heteromorphic dialect expressions, the model can establish stable matching based on context semantics and cultural labels. For words with regional cultural meaning, the model can combine the Mandarin paraphrase vector to preserve their cultural orientation. The alignment results can provide a text semantic basis for subsequent cultural entity recognition, knowledge graph construction and multimodal resource retrieval [19].

3.3 Regional cultural knowledge graph construction and relation extraction model

The regional cultural knowledge graph is used to organize dialect words, regional locations, folk activities, traditional artifacts, character titles and narrative fragments into a query-able and reasonable structured knowledge network. After the processing of speech recognition and semantic alignment, the system has obtained dialect text, Mandarin interpretation, regional label and cultural label, but these information still exist in the form of sample fields. It is difficult to directly express the correlation relations such as "what kind of cultural scene a certain dialect word belongs to", "where a certain folk custom activity is popular" and "how to call a certain object name in different dialect areas" [20]. To this end, this paper focuses on entity recognition, relation extraction and triple construction, transforms the dialect cultural corpus into graph structure data, and further obtains computable knowledge representation through the graph embedding model.

Let the regional cultural knowledge graph be G, where the set of entities is E, the set of relations is R, and the set of triples is T. Each cultural knowledge triple can be expressed as follows.

$$T=\{(h_k,r_k,o_k) \mid h_k,o_k \in E,r_k \in R,k=1,2,\ldots,K\} \tag{15}$$

where, hk represents the head entity of the KTH triple; ok represents the tail entity. rk represents the relationship between entities. K denotes the number of triples. For example, "a dialect word -- reference -- traditional utensils", "folk custom activity -- popular in -- regional location", "dialect appellation -- source -- narrative fragment" can be stored according to this structure. The triplet structure can transform the cultural knowledge in the scattered corpus into nodes and edges, so that the subsequent retrieval is not limited to keyword matching, but can use the entity relationship to expand the query.

In the entity representation stage, the model maps dialect words, cultural objects and geographical labels into a unified vector. Let the e-th entity be composed of the text description vector te, the type embedding ce and the region embedding ge. Its fusion representation can be written as follows.

$$z_e=\sigma\left(W_t t_e+W_c c_e+W_g g_e+b_e\right) \tag{16}$$

where, ze represents entity fusion vector; Wt, Wc, Wg denote text, genre, and regional feature projection matrices, respectively. be represents the bias term; Let $\sigma(\cdot)$ denote the nonlinear activation function. This representation can preserve cultural semantics and regional differences at the entity level, so that the expression differences of the same cultural object in different dialect regions can be captured by the model.

The relation extraction stage takes dialect text segments and entity pairs as input to determine whether there is a preset relation type between entities. Let the entity pair (h,o) be denoted as cho in the context semantics, and the classification probability of relation type r can be expressed as follows.

$$P(r|h,o)=\mathrm{softmax}\left(W_r c_{ho}+b_r\right) \tag{17}$$

where P(r | h,o) represents the probability that the entity pair (h,o) belongs to relation r; Wr represents the relational classification weight matrix; br stands for relation classification bias; cho represents the entity pair semantic representation obtained by the context encoder. The model generates candidate triples according to the maximum probability relationship, and

then filters invalid relationships through entity type constraints and regional label consistency verification to reduce the problems of cultural entity misconnection and relationship over-generalization.

In order to support knowledge completion and graph retrieval, we further use a translational graph embedding scoring function to measure the reliability of triples. Given a triple (h,r,o), the score function is

$$f(h,r,o)=-\|z_h+z_r-z_o\|_2 \tag{18}$$

where zh, zr and zo represent the embedding vectors of head entity, relation and tail entity respectively. f(h,r,o) represents the triple confidence score. A higher score indicates that the head entity is closer to the tail entity after the relation vector transformation, and the triple is more likely to hold. This mechanism can be used to discover missing relationships, such as inferring the potential association between dialect words and folk activities according to "dialect word-reference-instrument" and "instrument-belonging-folk activity".

The construction process of regional cultural knowledge graph needs to complete entity extraction, relationship judgment, triple storage and graph embedding update at the same time. The data transmission relationship between each module is shown in Figure 5.

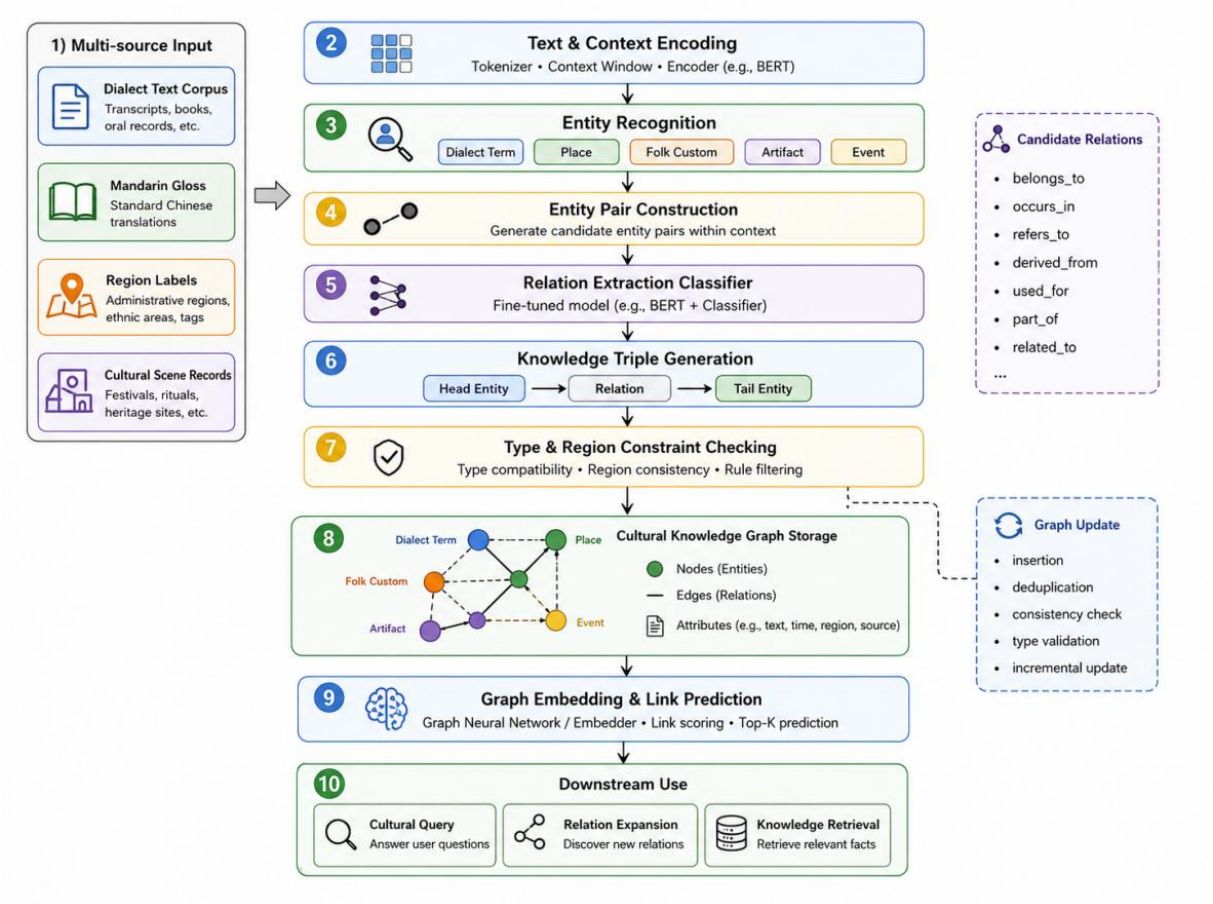

Figure 5: Model diagram of regional cultural knowledge graph construction and relation extraction

Through the above modeling process, the words, regions and cultural objects in the dialect culture corpus are transformed into a knowledge graph with an associated structure. The graph can not only support the cultural paraphrase query of dialect words, but also support the

relation expansion retrieval of regional cultural resources. Compared with simple text index, knowledge graph can express multi-hop relationships and semantic paths between entities, so that dialect protection is no longer limited to voice and word preservation, but further forms a structured knowledge organization method for cultural inheritance.

3.4 Multi-modal fusion and retrieval algorithm of dialect cultural resources

After speech recognition, semantic alignment and knowledge graph construction, dialect cultural resources form speech features, text semantic features, cultural entity features and regional label features. A single modality can only reflect local information, speech modality describes pronunciation differences and tone changes, text modality expresses word interpretation, knowledge graph modality preserves cultural object relationships, and regional modality describes dialect areas and scene sources. In order to support dialect resource retrieval, cultural knowledge query and cross-modal matching, this paper maps multi-class features into a unified vector space, and controls the contribution of different modalities through attention weights.

Let the speech representation, text representation, knowledge graph representation and geographical representation of dialect cultural resources in Article i be z_i^a, z_i^t, $z_i^k$and z_i^r respectively, and the multi-modal fusion input can be expressed as follows.

$$M_i=[z_i^a;z_i^t;z_i^k;z_i^r] \tag{19}$$

where, Mi represents the multi-modal splicing representation of the ith resource; [;] represents the vector concatenation operation. This structure compresses dialect pronunciation, lexical interpretation, cultural relations and regional information into the same candidate sample representation, avoiding relying only on text keywords in the retrieval stage.

Different retrieval requests do not have the same degree of dependence on modalities. Voice input queries are more dependent on acoustic vectors, cultural entity queries are more dependent on knowledge graph vectors, and regional scene queries are more dependent on regional labels and cultural relations. Let the MTH modality feature bez_i^m, and its attention weight can be expressed as follows.

$$\alpha_i^m=\frac{\exp(w^T \tanh(W_m z_i^m+b_m))}{\sum_{m=1}^{4}\exp(w^T \tanh(W_m z_i^m+b_m))} \tag{20}$$

where, α_i^m represents the attention weight of the MTH modality; Wm represents the modal projection matrix; bm represents the bias term; w represents the attention score vector. The higher the weight, the greater the contribution of the modality to the current resource representation.

After obtaining the modal weights, the system performs weighted fusion of multi-source features to generate the final resource index vector. The fusion process can be expressed as follows.

$$h_i=\phi\left(\sum_{m=1}^{4}\alpha_i^m W_f^m z_i^m+b_f\right) \tag{21}$$

where, hi represents the fusion index vector of the ith dialect cultural resources; W_f^mrepresents the fusion projection matrix of the MTH mode. bf represents fusion bias; Let

$\phi(\cdot)$ denote the nonlinear mapping function. The fused vectors are bound with resource numbers, text paraphrases, cultural entities, audio paths, and atlas node numbers to form a searchable multimodal resource unit.

In the retrieval phase, user queries can come from dialect words, Mandarin paraphrases, speech segments, or cultural entities. The system encodes the query content into a query vector q, calculates the similarity with the candidate resource index vector, and returns the top K results according to the score. The Top-K ranking function can be expressed as follows.

$$R_K(q)=\mathrm{TopK}_{i\in D}\left(\frac{q^T h_i}{\|q\|_2\|h_i\|_2}+\lambda f_g(q,h_i)\right) \tag{22}$$

where RK(q) represents the top K retrieval results corresponding to the query vector; D denotes the set of candidate resources; The cosine term represents the semantic similarity between the query vector and the resource fusion vector. fg(q,hi) represents the knowledge graph relation matching score; Let λ denote the atlas score adjustment coefficient. This ranking method combines vector similarity with knowledge relation constraints, so that the system can return association results such as dialect words, Mandarin paraphrases, regional scenes and cultural objects.

Through multi-modal fusion and retrieval algorithm, dialect cultural resources are transformed from single audio, text or label records into vectorized objects that can be uniformly indexed. The speech retrieval task calls the acoustic similarity, the text query task calls the semantic alignment results, and the cultural heritage query uses the knowledge graph relationship for extended recall, which provides the evaluation basis for the subsequent Recall@K, MRR, response delay and ablation experiments.

4 Implementation of digital inheritance system and analysis of experimental results

4.1 Dialect culture Resource System Architecture and experimental environment Configuration

The digital inheritance system of dialect cultural resources is constructed around "data access, model calculation, knowledge storage, semantic retrieval and interactive service". The data access layer is responsible for receiving dialect speech, transcribed text, Mandarin interpretation, cultural scene text and regional labels. Acoustic coding, semantic alignment, relation extraction and multi-modal fusion modules were deployed in the model computing layer. In the knowledge storage layer, cultural entities, regional nodes, dialect words and relationship triples were written into the graph database. The semantic retrieval layer completes similarity recall and Top-K ranking based on vector index. The interactive service layer provides users with dialect word query, speech recognition, cultural entity association retrieval and regional cultural resources display functions. Each module of the system is connected by a uniform resource number, which ensures that audio path, text interpretation, cultural label and graph node can be called in the same retrieval link.

In order to ensure that model training, graph query and vector retrieval can be completed in the same experimental environment, the hardware, software, deep learning framework and storage components are configured uniformly in this paper, as detailed in Table IV.

Table 4: Experimental environment and system operation configuration table

Category	Project	Configuration Value
Hardware	CPU	Intel Xeon Silver 4314 × 2
Hardware	GPU	NVIDIA RTX 4090 24GB
Hardware	Memory	128 GB
Software	Operating System	Ubuntu 22.04
Framework	Deep Learning Framework	PyTorch 2.2
Storage	Graph Database	Neo4j 5.12
Retrieval	Vector Index	FAISS 1.7.4
Training	Batch Size	32
Training	Learning Rate	1e-4
Evaluation	Metrics	Accuracy, F1-score, WER, Recall@K, MRR

During the operation of the system, the model service processes voice and text input by batch reasoning, the graph database is responsible for entity relationship query, and the FAISS index is responsible for high-dimensional vector recall. The speech recognition results are converted into Mandarin paraphras vector by semantic alignment module, and then enter the fusion retrieval module together with cultural entity vector and regional label vector. In order to show the data transmission relationship between different levels in the system, Figure 6 shows the overall architecture of the digital inheritance system of dialect cultural resources.

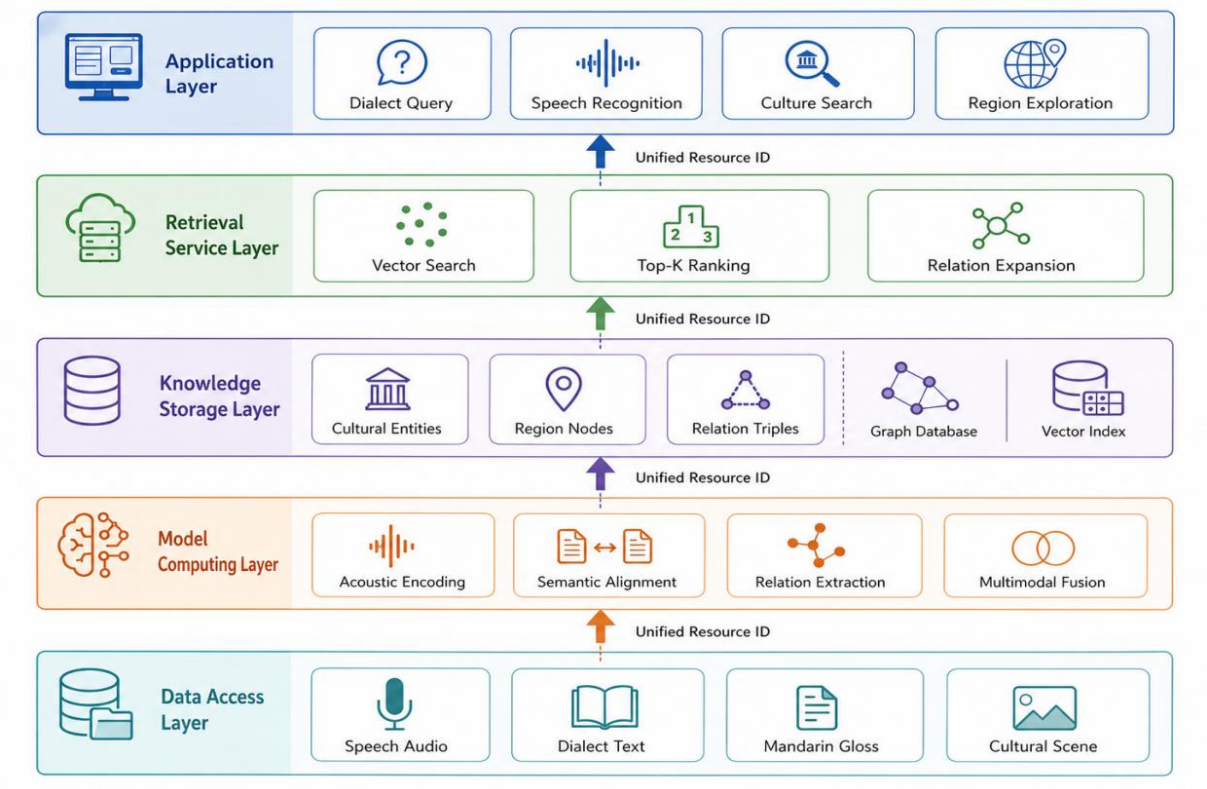

Figure 6: Overall architecture of digital inheritance system of dialect cultural resources

The system architecture connects the speech recognition task in dialect resource protection and the knowledge organization task in cultural inheritance into the same engineering link. After entering the model calculation module from the underlying collection end, the data are transformed into acoustic vectors, semantic vectors and graph entity relations, and then the

retrieval service is supported by the vector index and the graph database, which provides a unified experimental platform for subsequent dataset experiments, comparative tests and system performance analysis.

4.2 Experimental data set construction, sample division and evaluation index design

The experimental data set uses the self-built Chinese dialect culture corpus CDCC-2025. The data sources include dialect vocabulary reading, short sentence expression, natural narrative recording, manual transcription text, Mandarin paraphrase text and regional cultural scene description. The data collection covers six types of dialect areas, and a total of 120 hours of dialect speech, 36000 speech samples, 18000 dialect texts, 18000 Mandarin interpretations, 5600 cultural scene texts, 6200 cultural entities and 14800 relationship triples are formed. Each sample is bound with speech file, dialect transcription, Mandarin paraphrase, regional label, cultural label and entity relationship annotation, so that dialect recognition, semantic retrieval and knowledge graph experiments can call the same batch of sample indexes.

The samples were divided into training set, validation set and test set with the ratio of 70% : 10% : 20%. The training set contains 84 h speech, 25200 speech samples, 12600 dialect texts and 10360 relation triples for model parameter updating. The validation set contains 12 h speech, 3600 speech samples, 1800 dialect texts and 1480 relation triples, which are used for hyperparameter selection and early stop judgment. The test set contains 24 h speech, 7200 speech samples, 3600 dialect texts, and 2960 relation triples, which are only used for the final performance evaluation. The CDCC-2025 dataset establishes a unified number among four types of task objects: speech, text, cultural entity and relation triplet, and the specific scale and division are shown in Table 5.

Table 5: Scale and Sample Division of Chinese Dialect Culture Experimental Dataset

Data Item	Total Amount	Training Set	Validation Set	Test Set
Dialect Area/Class	6	6	6	6
Speech Duration/h	120	84	12	24
Speech Samples/Items	36000	25200	3600	7200
Dialect Texts/Items	18000	12600	1800	3600
Mandarin Interpretations/Items	18000	12600	1800	3600
Cultural Scene Texts/Items	5600	3920	560	1120
Cultural Entities/Items	6200	4340	620	1240
Relation Triples/Items	14800	10360	1480	2960

Accuracy, F1-score and WER were used to evaluate the classification judgment and transliteration output ability of the dialect recognition task. Accuracy measures the accuracy of dialect category judgment, F1-score comprehensively reflects the precision and recall rate, and WER evaluates the word-level error degree between the transcribed text and the manually annotated text. The comprehensive recognition evaluation index can be expressed as follows.

$$S_{rec}=\eta_1 \text{Accuracy}+\eta_2 F1+\eta_3(1\text{-WER}) \tag{23}$$

where, Srec represents the comprehensive score of dialect recognition; η1, η2, η3 represent the weights of different indicators, and satisfy η1+η2+η3=1. A lower WER indicates a better transcribing effect, so 1−WER is used to convert to a positive index.

The semantic retrieval and cultural knowledge query tasks used Recall@K and MRR to

evaluate the matching degree between the returned results and the real annotation results. Recall@K reflects whether the true relevant resource appears in the top K returned results, and MRR reflects the ranking position of the first correct result. The retrieval evaluation function can be expressed as follows.

$$S_{ret}=\mu_1 Recall@K+\mu_2 MRR \tag{24}$$

where, Sret represents the comprehensive score of semantic retrieval; μ1 and μ2 represent the retrieval index weights and satisfy μ1+μ2=1. Through the above data set division and evaluation index design, dialect identification, semantic alignment, knowledge graph retrieval and system operation testing can be completed on a unified experimental sample, which ensures that the subsequent comparison experiments are comparable with ablation experiments.

4.3 Comparative experiments of dialect identification, semantic Retrieval and knowledge graph

In order to verify the comprehensive performance of the proposed method in the task of dialect protection and cultural inheritance, this paper sets up three comparative experiments on the CDCC-2025 test set, including dialect recognition, semantic retrieval and knowledge graph relationship modeling. HMM-GMM, CNN-CTC, BiLSTM-CTC and Transformer-CTC are selected for dialect recognition task to compare with the proposed method. For semantic Retrieval tasks, TF-IDF, BM25, BERT Similarity and Multimodal Retrieval are selected to compare with the proposed method. TransE, DistMult, ComplEx, R-GCN are selected for knowledge graph tasks to compare with the proposed method. All methods used the same training set, validation set and test set partition, the speech recognition model evaluated Accuracy, F1-score and WER, the semantic retrieval model evaluated Recall@10 and MRR, and the knowledge graph model evaluated Triple F1, Hits@10 and MRR.

The experimental results of different methods on three types of tasks are shown in Table 6, where a lower WER for the recognition task indicates a smaller transcribe error rate, and a higher value for the remaining indicators indicates a better model performance。

Table 6: Summary of Comparative Experimental Results of Different Methods

Task Type	Method	Main Result
Dialect Recognition	HMM-GMM	Accuracy 78.4%, F1-score 76.9%, WER 21.7%
	CNN-CTC	Accuracy 84.6%, F1-score 83.2%, WER 17.5%
	BiLSTM-CTC	Accuracy 87.3%, F1-score 86.1%, WER 15.8%
	Transformer-CTC	Accuracy 90.8%, F1-score 89.6%, WER 12.4%
	Proposed Method	Accuracy 94.2%, F1-score 93.5%, WER 8.9%
Semantic Retrieval	TF-IDF	Recall@10 68.5%, MRR 0.612
	BM25	Recall@10 72.8%, MRR 0.657
	BERT Similarity	Recall@10 81.4%, MRR 0.734
	Multimodal Retrieval	Recall@10 86.7%, MRR 0.781
	Proposed Method	Recall@10 91.6%, MRR 0.846
Knowledge Graph	TransE	Triple F1 74.2%, Hits@10 78.6%, MRR 0.692
	DistMult	Triple F1 76.8%, Hits@10 80.4%, MRR 0.715
	ComplEx	Triple F1 79.5%, Hits@10 83.2%, MRR 0.746
	R-GCN	Triple F1 84.1%, Hits@10 87.6%, MRR 0.793
	Proposed Method	Triple F1 89.3%, Hits@10 92.4%, MRR 0.851

In order to quantify the improvement of the proposed method over the optimal baseline model, let the result of the proposed method on index j be Pj, and the optimal baseline result be Bj. The performance improvement rate can be expressed as follows.

$$\Delta_j = \frac{P_j - B_j}{B_j} \times 100\% \tag{25}$$

where, Δj represents the relative improvement rate of the JTH index; Pj denotes the experimental results of the proposed method; Bj denotes the optimal baseline result in the comparison method. For negative metrics such as WER, we use (Bj−Pj) to represent the decrease in error rate.

In order to intuitively present the differences in Accuracy and word error rate of dialect recognition models, Figure 7 shows the recognition results of different models in the form of bar line combination, where the bar represents accuracy and the broken line represents WER.

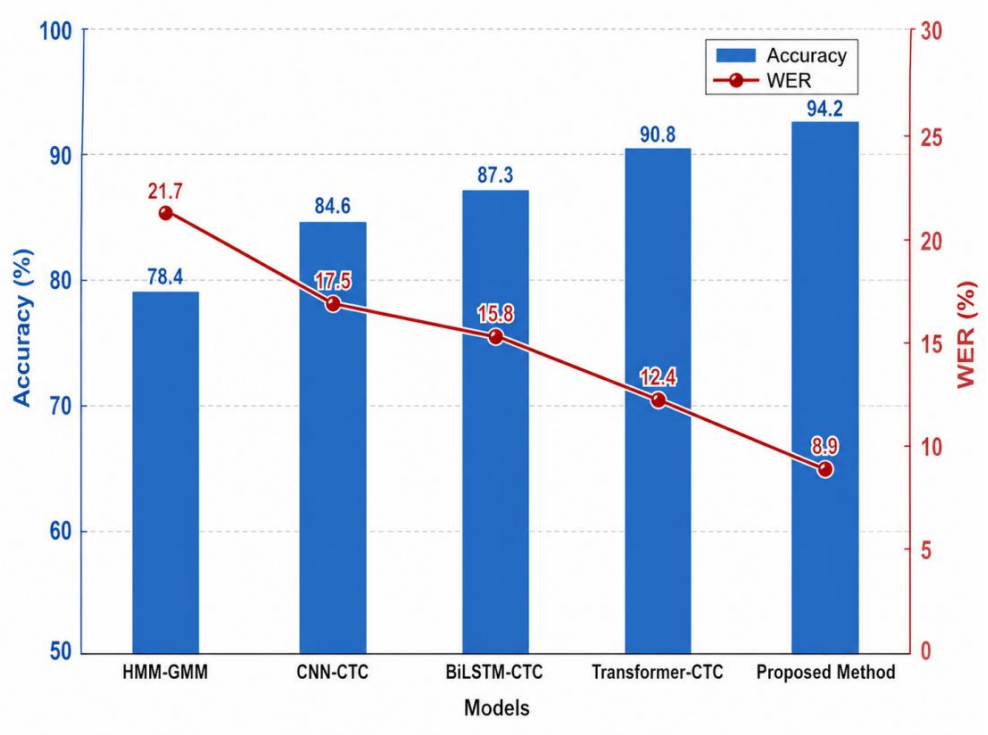

Figure 7: Comparison of dialect recognition accuracy and word error rate under different models

The recognition results show that the traditional HMM-GMM is weak in adapting to dialect tone, prosody and cross-regional pronunciation changes, with an Accuracy of 78.4% and a WER of 21.7%. CNN-CTC can extract local spectral patterns, and BiLSTM-CTC can further exploit frame dependence, both of which are improved compared with HMM-GMM. Transformer-CTC enhances the long-distance speech context modeling ability through the self-attention mechanism, and the Accuracy is improved to 90.8%. Under the combined effect of convolutional down-sampling, context coding and multi-modal semantic constraints, the Accuracy of the proposed method reaches 94.2%, the F1-score reaches 93.5%, and the WER is reduced to 8.9%. Compared with Transformer-CTC, the Accuracy is improved by 3.7%. The WER decreases by 28.2%, indicating that the proposed model can recognize complex dialect speech more stably.

Semantic retrieval and knowledge graph experiments need to compare recall ability, ranking effect and relationship modeling ability at the same time, Figure 8 uses radar chart to show the comprehensive differences of indicators such as Recall@10, MRR, Triple F1 and Hits@10.

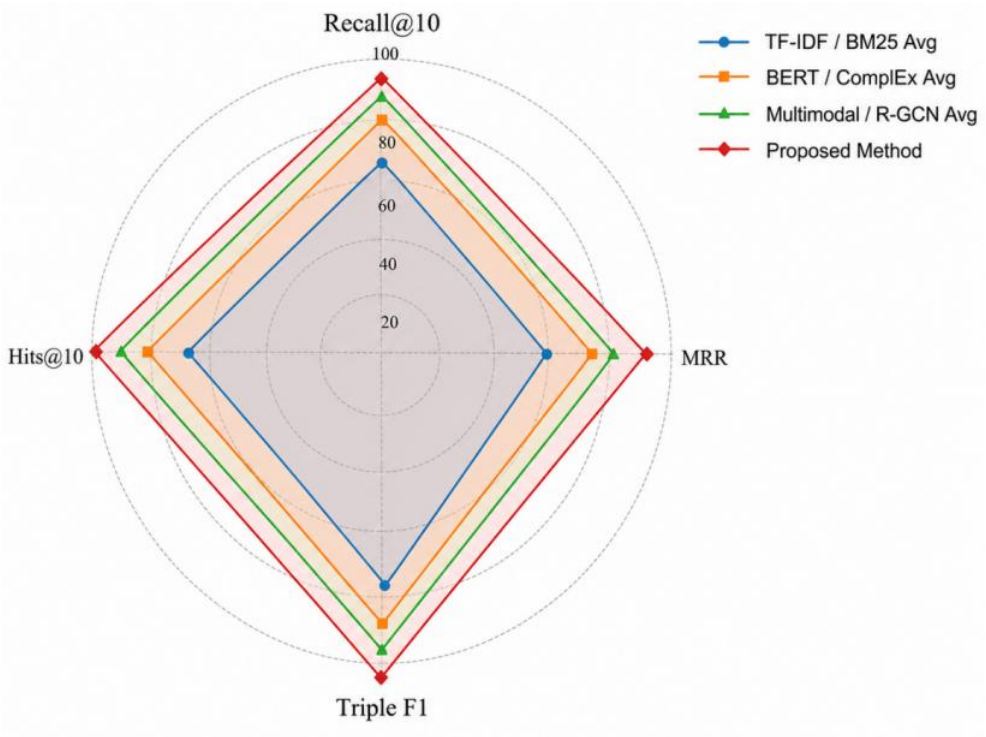

Figure 8: Radar chart of comprehensive performance of different retrieval and knowledge graph methods

According to the retrieval results, TF-IDF and BM25 rely on word plane matching, and the processing of dialect synonymous and cultural proper name variants is insufficient, Recall@10 is 68.5% and 72.8% respectively. After the introduction of semantic representation by BERT Similarity, the MRR is improved to 0.734, but the cultural entity relationship constraint is missing. Multimodal Retrieval achieves 86.7% Recall@10 after speech, text and label fusion. The method in this paper further adds knowledge graph relation matching, and the Recall@10 is increased to 91.6%, and the MRR reaches 0.846. In the knowledge graph experiment, Triple F1 reaches 89.3% and Hits@10 reaches 92.4%, both of which are higher than R-GCN, indicating that regional labels, cultural entity relationships and dialect lexical semantic alignment can jointly improve the organization and retrieval effect of dialect cultural knowledge.

4.4 Multi-modal fusion ablation experiment and analysis of system operation results

In order to test the actual contribution of each module in the multimodal fusion structure, this paper carries out ablation experiments on the CDCC-2025 test set. The experiment takes the complete model in 4.3 as the benchmark, keeps the partition of training set, validation set and test set unchanged, sets the Batch Size to 32, the learning rate to 1e-4, the maximum training round to 80, and adopts the same early stopping strategy. Five groups of models were set up in the ablation experiment: Full Model, w/o Speech Branch, w/o Text Branch, w/o Knowledge Graph, and w/o Attention Fusion. Only one module was removed from each group of

experiments, and the rest of the network structure, training parameters, index configuration and evaluation indicators were kept consistent to ensure that the results of different models were comparable.

To observe the difference in the original index values between the complete model and each ablation model, Figure 9 shows the comparison results of the five groups of models in the four indicators of Accuracy, F1-score, Recall@10 and MRR.

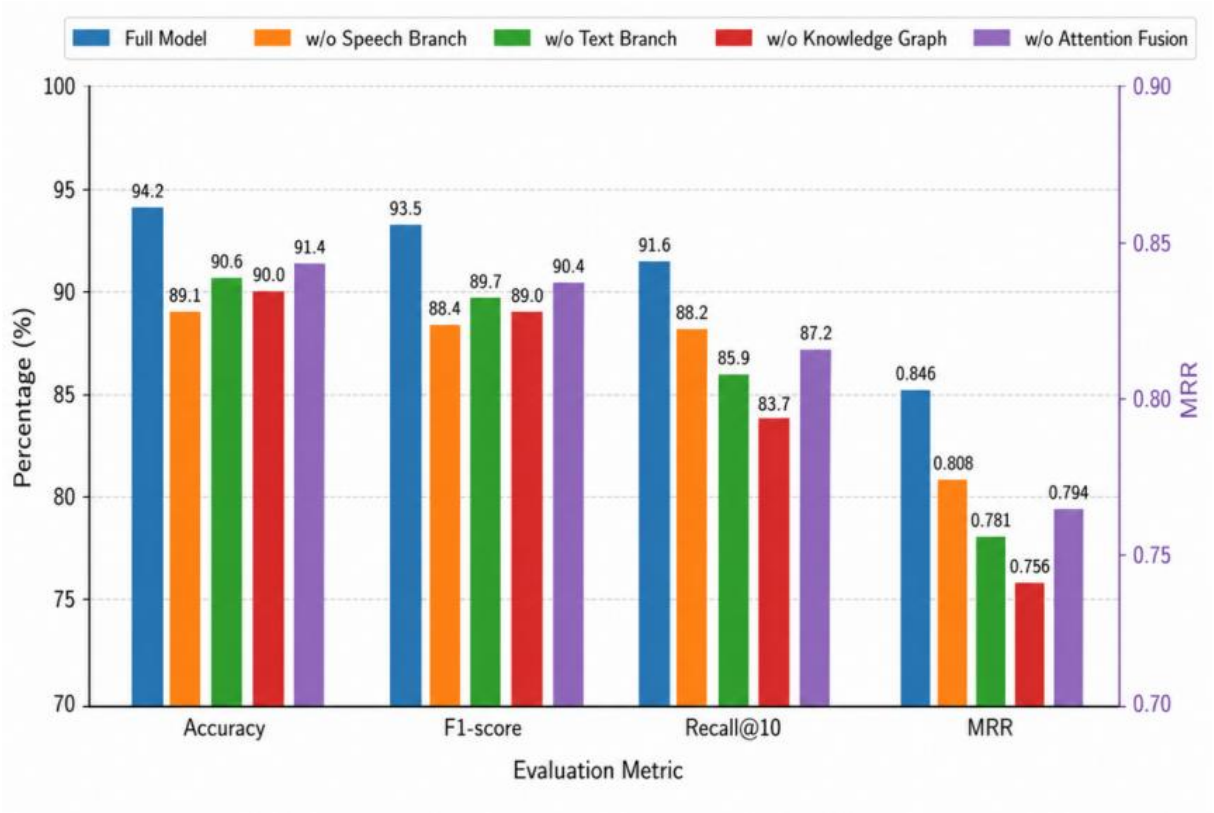

Figure 9: Comparison of key indicators between the complete model and the ablation model

The full model maintains the highest level in all four indicators. After removing the speech branches, the Accuracy and F1-score are reduced to 89.1% and 88.4%, respectively, which indicates that the acoustic features have a direct support role in the identification of dialect pronunciation differences. After removing text branches, Recall@10 and MRR decrease to 85.9% and 0.781, respectively, indicating that semantic alignment between dialect words and Mandarin paraphrases affects retrieval recall and ranking quality. After removing the branches of the knowledge graph, Recall@10 is reduced to 83.7%, and MRR is reduced to 0.756, indicating that cultural entity relationships have an obvious effect on the recall of regional cultural resources expansion.

On the basis of the original index comparison, Figure 10 further calculates the performance degradation range of each ablation model compared with the complete model, which is used to judge the contribution strength of different modules to the overall performance.

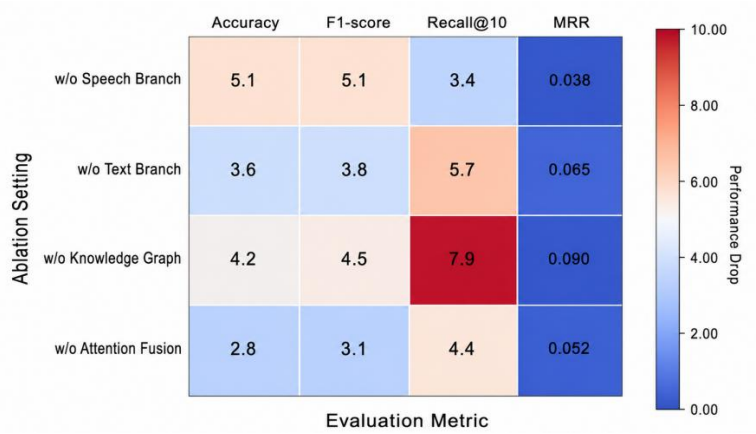

Figure 10: Heat map of performance degradation in ablation experiments with multimodal fusion

After voice branch removal, the Accuracy and F1-score both decreased by 5.1 percentage points, which was the most obvious ablation term for recognition indicators. After the removal of knowledge graph branches, Recall@10 decreased by 7.9 percentage points and MRR decreased by 0.090, which was the most obvious ablation term for retrieval indexes. After removing the text branch, Recall@10 decreases by 5.7 percentage points, indicating that Mandarin paraphrase alignment can improve the semantic matching ability of dialect word retrieval. All four indicators decline after removing the attention fusion module, indicating that dynamic weight fusion is more suitable for multi-source dialect cultural resources modeling than simple splicing.

The system running performance test was carried out after the ablation experiment, and the test objects were the complete model and its corresponding vector index database and graph database. The experiment set five groups of concurrent access scale of 10, 50, 100, 200 and 400, and each group sent 1000 consecutive query requests, covering dialect word retrieval, speech recognition result retrieval, cultural entity query and regional relationship expansion query. The system records the average response delay, query throughput, and graph query time, and the results are shown in Figure 11.

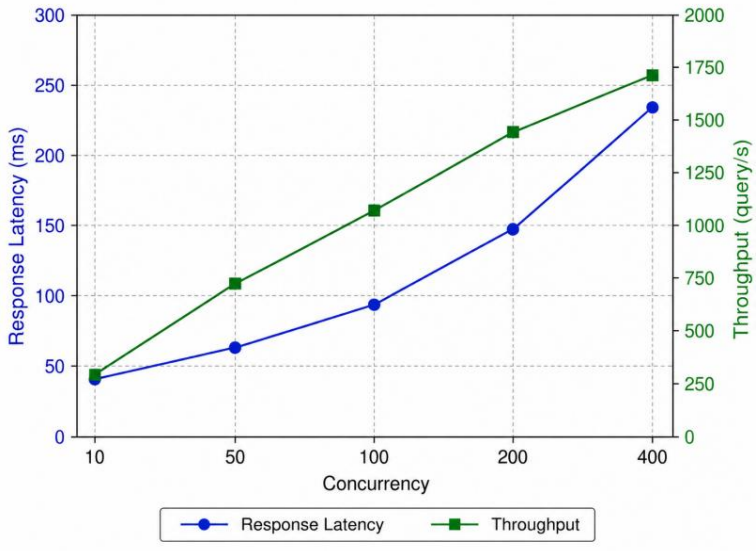

Figure 11: System response time and throughput change diagram under different concurrent accesses

When the number of concurrency increases from 10 to 400, the average response delay of the system increases from 42 ms to 236 ms, the throughput increases from 238 query/s to 1715 query/s, and the graph query time increases from 18 ms to 96 ms. When the concurrent scale is enlarged, the response time and graph query time rise synchronously, but the average response time is still controlled within 250 ms under 400 concurrent conditions, which indicates that vector index recall and graph database relationship expansion can maintain stable cooperation. Comprehensive ablation experiments and system operation results show that the performance improvement of the proposed method comes from the joint effect of speech branch, text branch, knowledge graph branch and attention fusion module, and the system has the engineering operation ability to support dialect speech recognition, cultural resource retrieval and regional knowledge association query.

5 Discussion

Multimodal information plays different roles in the task of dialect protection and cultural inheritance. Acoustic features have a more direct impact on dialect category recognition and transliteration output, while the relationship between text semantics and knowledge graph plays a more important role in cultural resource retrieval, Mandarin paraphrasing matching and regional culture association query. In the ablation results, the decline of recognition indicators is more obvious after removing the speech branch, and the decline of retrieval indicators is more prominent after removing the knowledge graph branch, indicating that dialect protection focuses on "accurate recognition", and cultural inheritance focuses on "relationship organization" and "semantic expansion".

From the comparative experiment, a single speech recognition model can complete the basic transliteration, but it is difficult to explain the cultural orientation behind the dialect words. Single text retrieval method can deal with explicit word-surface matching, but it is easy to ignore the implicit relationship between local artifacts, folk activities and place names. The addition of knowledge graph and multi-modal fusion mechanism enables the system to establish continuous mapping between recognition results, Mandarin paraphrases and cultural entities.

The system performance also reflects the bottleneck in engineering implementation. When the number of concurrent accesses increases, the response delay and graph query time rise synchronously, indicating that multi-modal retrieval is not only affected by vector recall speed, entity relationship expansion and result reordering also affect the overall efficiency. The subsequent optimization can focus on graph caching, index compression and query path clipping to improve the real-time retrieval ability of the system in large-scale dialect cultural resources.

6 Conclusion

Focusing on the digital processing requirements in Chinese dialect protection and cultural inheritance, this paper constructs a technical solution that integrates corpus collection, acoustic recognition, semantic alignment, knowledge graph construction, multi-modal fusion retrieval and system implementation. Dialect speech recognition is completed through the acoustic coding model. The two-tower semantic alignment algorithm is used to establish the mapping between dialect words and putongning interpretations. The technical framework connects the tasks of speech preservation and transcribe in dialect protection, semantic association and knowledge organization in cultural inheritance into the same engineering link,

which provides a realizable technical support for the intelligent preservation, structured management and digital dissemination of Chinese dialect resources.

Funding

This paper is the research achievement of the 2026 Excellent Project of Ideological and Political Work in Hunan Provincial Institutions of Higher Education: "Fangcao Intelligent Study Tour" — Construction and Practice of an Immersive Intelligent Agent Platform for Huxiang Red Culture Research and Study from the Perspective of Higher Vocational Ideological and Political Education.

References

[1] LI Q, MAI Q Y, WANG M D, et al. Chinese dialect speech recognition: a comprehensive survey[J]. Artificial Intelligence Review, 2024, 57(2): 25.

[2] YU T Z, FRIESKE R, XU P, et al. Automatic speech recognition datasets in Cantonese: a survey and new dataset[C]//Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille: European Language Resources Association, 2022: 6487-6494.

[3] LIN J Y, LU S H, HUANG H K, et al. MinSpeech: a corpus of Southern Min dialect for automatic speech recognition[C]//Proceedings of Interspeech 2024. Kos Island: ISCA, 2024: 2330-2334.

[4] MOHAMED A, LEE H Y, BORGHOLT L, et al. Self-supervised speech representation learning: a review[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1179-1210.

[5] PRABHAVALKAR R, HORI T, SAINATH T N, et al. End-to-end speech recognition: a survey[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 325-351.

[6] CHEN S Y, WANG C Y, CHEN Z Y, et al. WavLM: large-scale self-supervised pre-training for full stack speech processing[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505-1518.

[7] BABU A, WANG C H, TJANDRA A, et al. XLS-R: self-supervised cross-lingual speech representation learning at scale[C]//Proceedings of Interspeech 2022. Incheon: ISCA, 2022: 2278-2282.

[8] RADFORD A, KIM J W, XU T, et al. Robust speech recognition via large-scale weak supervision[C]//Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR, 2023, 202: 28492-28518.

[9] PENG C Y, XIA F, NASERIPARSA M, et al. Knowledge graphs: opportunities and challenges[J]. Artificial Intelligence Review, 2023, 56(11): 13071-13102.

[10] PELLEGRINO M A, SCARANO V, SPAGNUOLO C. Move cultural heritage

knowledge graphs in everyone’s pocket[J]. Semantic Web, 2023, 14(2): 323-359.

[11] HYVÖNEN E. Digital humanities on the Semantic Web: Sampo model and portal series[J]. Semantic Web, 2023, 14(4): 729-744.

[12] LIU P F, YUAN W Z, FU J L, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023, 55(9): 1-35.

[13] BAEVSKI A, HSU W N, XU Q T, et al. data2vec: a general framework for self-supervised learning in speech, vision and language[C]//Proceedings of the 39th International Conference on Machine Learning. Baltimore: PMLR, 2022, 162: 1298-1312.

[14] ERICSSON L, GOUK H, LOY C C, et al. Self-supervised representation learning: introduction, advances, and challenges[J]. IEEE Signal Processing Magazine, 2022, 39(3): 42-62.

[15] ORKEN M, DINA O, KEYLAN A, et al. A study of transformer-based end-to-end speech recognition system for Kazakh language[J]. Scientific Reports, 2022, 12(1): 8337.

[16] CHEN Y, GE X K, YANG S L, et al. A survey on multimodal knowledge graphs: construction, completion and applications[J]. Mathematics, 2023, 11(8): 1815.

[17] CAO J H, FANG J Y, MENG Z Q, et al. Knowledge graph embedding: a survey from the perspective of representation spaces[J]. ACM Computing Surveys, 2024, 56(6): 1-42.

[18] CHEN J, HUANG Y, DING H, et al. Knowledge graphs for the life sciences[J]. Transactions on Graph Data and Knowledge, 2023, 1(1): 5:1-5:47.

[19] PAN J Z, RAZNIEWSKI S, KALO J C, et al. Large language models and knowledge graphs: opportunities and challenges[J]. Transactions on Graph Data and Knowledge, 2023, 1(1): 2:1-2:38.

[20] MAREE M, SPYROU E, MAKRIS C, et al. Quantifying relational exploration in cultural heritage knowledge graphs[J]. Data, 2025, 10(4): 52.