Regional Identity Shifts In The Chinese Language And Culture AI Dialect Restoration Project

Long Liu

Department of Party Organization, Shangqiu Institute of Technology, shangqiu,476000, China

Corresponding author. E-mail: longliu62@outlook.com, liulong8201@163.com

Received: Nov. 24, 2025; Accepted: Jan. 03, 2026

The preservation of regional dialects in China has become increasingly important due to the dominance of Mandarin, driven by urbanization and globalization, which threatens local languages and cultures. This paper addresses the challenge of dialect restoration by leveraging advanced AI models, including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, to restore and maintain regional dialects in the digital era. The objective of this research is to develop a scalable, real-time framework for dialect restoration that ensures both linguistic accuracy and cultural preservation. The proposed method utilizes FastSpeech2 for text-to-speech synthesis and HiFi-GAN for high-fidelity speech generation, overcoming the limitations of traditional models. The framework also integrates Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction and the Arithmetic Optimization Algorithm (AOA) for efficient optimization, which improves both accuracy and processing speed. The results demonstrate the effectiveness of the proposed method, achieving an accuracy of 0.9812, precision of 0.9811, recall of 0.9808, and F1-score of 0.9805, indicating high performance in dialect classification and restoration. The Mean Opinion Score (MOS) for speech quality ranges from 4.52 to 4.72, with Mandarin achieving the highest score of 4.72. The Word Error Rate (WER) is between 1.58% and 2.45%, with Mandarin showing the lowest error rate of 1.58%. These results confirm the potential of the proposed framework in preserving regional dialects and ensuring their continued cultural significance.

Keywords: Regional dialects, AI-based restoration, Convolutional Neural Networks, Long Short-Term Memory, FastSpeech2, HiFi-GAN

http://dx.doi.org/10.6180/jase.202608_31.018

1. Introduction

The necessity for the preservation of regional dialects is becoming more critical with urbanization and globalization in China, which are major factors that will continue to shrink its already diverse linguistic landscape [1]. The widespread use of Mandarin has, to a large extent, silenced and even eliminated some of the regional languages and, consequently, the cultures linked to them [2]. The "Regional Identity Shifts in the Chinese Language and Culture AI Dialect Restoration Project" is a project that tackles this issue directly by using the latest AI technologies not only to restore but also to maintain these dialects, thus ensuring they will be recognized as the digital-age elements of China's vibrant cultural heritage [3], [4]. By utilizing state-of-the-art technology, this framework intends to come up with a long-term solution for dialect preservation [5].

A number of reasons have led to the decrease of regional dialects in China [6]. The massive use of Mandarin in schools, media, and government policies has greatly lessened the use of local dialects, especially among the young people [7]. Moreover, the movement of people and the

growth of cities have strengthened the position of Mandarin not only in business but also in social life [8]. The change in linguistic environment has been from one that was very diverse in terms of dialects to one that was very uniform and this has ousted their use [9]. Thus, the challenge of coming up with creative strategies for the protection and revival of these regional languages has become imperative [10].

Previously used techniques to maintain dialects primarily depended on human recordkeeping and conventional linguistic approaches like Hidden Markov Models (HMM) and LSTM networks [11]. Although these techniques have found their way into speech recognition, phonetic transcription, and dialect classification, they have major drawbacks [12], [13]. Particularly, they have problems with scalability which is critical when it comes to handling the changing and varied characteristics of regional dialects [14]. In addition, these systems do not support the restoration of dialects in real time and are likely to produce incorrect results when encountered with complicated linguistic variations [15].

The framework that has been proposed does not just get rid of the disadvantages of the existing methods but also makes use of a Hybrid CNN + LSTM model for dialect recognition and FastSpeech2 in combination with HiFi-GAN for producing high-quality speech and dialect restoration in real-time. MFCCs are used for feature extraction in the system and the AOA is used for the selection of features, which results in better accuracy and shorter processing time. The novelty of the methodology is in its capacity to restore dialects not only accurately but also quickly according to the specific linguistic input's context, which in turn guarantees that the revived speech is both culturally and linguistically distinct. The new system is one that is easily scalable and efficient for the preservation of dialects thus giving a big advantage over the old models that were limited in terms of scalability and real-time processing.

In rapidly urbanizing linguistic environments, the ability to restore and recognize regional dialects in real time has become increasingly significant from both social and cultural perspectives. Urban migration, digital communication, and centralized language policies often limit opportunities for intergenerational dialect transmission, causing regional languages to fade from daily use. Real-time dialect restoration enables immediate recognition, translation, and synthesis of dialectal speech during natural conversations, digital interactions, and multimedia content creation, ensuring that dialects remain actively used rather than passively archived. By supporting live communication and instant linguistic feedback, real-time restoration strengthens cultural identity, facilitates inclusive communication for dialectspeaking communities, and preserves linguistic diversity in fast-evolving urban societies. Therefore, prioritizing real-time dialect restoration is not only a technical advancement but also a socially driven necessity for sustaining regional linguistic heritage under conditions of accelerated urbanization.

Key Contributions are:

- Create an efficient framework for restoring regional dialects using the Chinese Dialect Speech-to-English Dataset. The goal is to preserve these dialects by applying advanced AI techniques.
- Use CNN and LSTM networks to accurately recognize and classify different regional dialects from speech data.
- Apply FastSpeech2 for generating high-quality speech from recognized dialects, ensuring that the output is linguistically and culturally accurate.
- Incorporate HiFi-GAN to turn the text generated by FastSpeech2 into realistic, fluent speech that maintains the unique features of the dialects.
- Enhance feature extraction and selection using MFCCs and AOA, improving processing speed and accuracy.

1.1. Structure of the Paper

The structure of this paper is as follows: A comprehensive literature review is given in Section 2. The method of the suggested approach is explained in Section 3. The results achieved are discussed exhaustively in Section 4. Finally, Section 5 draws a conclusion that underlines the key discoveries and their importance.

2. Theory and formula

Li et al.[16] presents the wide range of Chinese dialects makes speech recognition difficult because of the differences in areas where the dialects are spoken and the movement of people. The main areas of research are classifying dialects, identifying acoustic features, and creating phonetic corpora. The use of hybrid ANN-HMM methods and End-to-End techniques are among the popular ones. Wang & King,[17] describes the role of parental language ideologies in determining kids' language skills in three Chinese cities. It discovers four major ideological positions concerning Mandarin and local dialects, thus presenting the support for language use as per the regions. The research concludes that parental opinions are decisive for

kids' skill in regional dialects, with stronger ones for Mandarin associating with weaker ones in the case of regional dialect.

Duan et al.[18] present the traditional villages in multicultural areas facing globalization, industrialization, and urbanization challenges. It takes the Jiangxi-Anhui border area as a case and investigates the cultural environment of traditional villages and their houses, taking into account nature, space, architecture, and history. Zhang et al.[19]describes a research project that utilizes a sign analysis of the Chinese language in the linguistic landscape (LL) examines the linguistic behaviors in Singapore's Chinatown from the perspective of Chinese Singaporeans and the newly arrived Chinese immigrants. It reveals the competing claims of truth and self-identity projection of the two groups, with the new immigrants denying the authenticity that Chinese Singaporeans profess.

Kuang et al.[20] presents this project looks into problems accompanying urbanization and economic development that the traditional Chinese rural construction has to face. It takes the Goulou cluster of Yue dialects in Guangxi as a reference and studies the systematic elimination of traditional architectural features through GIS spatial analysis and typological identification. Privitera et al.[21] presents the research provides a comprehensive examination of the impact of artificial intelligence in the domain of aphasiology, primarily scrutinizing the case of generative AI and Natural Language Processing (NLP) models via the analysis of multilingual speech data and probing the effect of brain injury on language skills.

Qian et al.[22] describes the diverse languages used by the public to express their opinions about rain and floods, thereby showing the connection between the observation of precipitation and the use of different terms. The analysis of social media microblogs leads to the introduction of a new algorithm that classifies expressions related to rain and flood and, at the same time, uncovers the trends in the use of words, the timing of delays, and geographical differences. Szromek & Bugdol,[23] presents this article discusses the significance of open innovations in the process of cultural heritage transfer to cultural identity and education. The authors knowingly acknowledge that although previous research indicated the classification and benefits of the case, the systematic method of rebuilding cultural identity through heritage has not been that much explored. Among other things, the paper points out the generational shift in the perception of knowledge that extends particularly to the people born during the digital revolution.

Pizarro Contreras et al.[24] present study analyses the main effects of generative AI systems which are not able to create any differences or generate questions. The authors assert that the availability of these models is representative of a technological ontology of one kind only, thus, the futures would be non-diverse. The case of the LatamGPT project is brought in as an example of resistance to such a technological scenario which, by its very nature, reinstates diversity, imagination and politics in the ecosystem and thus, undermines the power of the streaming. J. Li et al.[25] describes the digital transformation of the intangible cultural heritage (ICH) from an anthropological standpoint focuses primarily on the aspects of user identity and sentiment analysis. The innovative analytical framework, dubbed Identity and SentimentCentered Framework for ICH (ISC-ICH), showed and described in the paper, unveils the relationship of the various public with ICH museums in Eastern Sichuan, China.

The previous discussions in the literature review have pointed out the transformation of ICH (intangible cultural heritage) preservation practices from old-fashioned ways to high-tech modern ones. One of the main results is the establishment of the Identity and SentimentCentered Framework (ISC-ICH), which incorporates sentiment analysis and user identity classification. The research is about the visitors of the ICH museums in Eastern Sichuan, China, and it distinguishes three elements emotional involvement, cultural property, and communication. Furthermore, the investigators have considered the technological challenges and suggested methods for preserving the cultural identity of ICH museums, especially in remote areas.

Traditional speech recognition systems primarily relied on HMM-GMM architectures, which serve as classical baselines due to their probabilistic temporal modeling but are limited in capturing complex non-linear and dialect-specific speech variations. Recent studies have shifted toward neural and hybrid models such as CNN-, LSTM-, and CNN-LSTM-based architectures, which learn richer acoustic representations and demonstrate significantly improved performance. In this context, the proposed framework aligns with modern hybrid speech processing paradigms and advances beyond HMM-GMM baselines in both recognition accuracy and speech reconstruction quality.

2.1. Problem Statement

The obstacles that regional dialects' preservation and restoration confront in China, namely, rapid urbanization [17], globalization, and the dominance of Mandarin as a standard language, are nonetheless very challenging [19]. All these changes have the same effect of pointing to the disappearance of the different regions' cultural and linguistic identities[21]. The traditional methods used for

dialect preservation are not sufficient most of the time as they do not consider the rapid changes brought about by technology [23]26. Within the "Regional Identity Shifts in the Chinese Language and Culture AI Dialect Restoration Project" context, AI-based dialect restoration methods are a promising solution, with DL models being the ones that best capture and reconstruct dialects. The project's main strategy for achieving the goal of reviving and maintaining regional dialects through the use of Large Language Models and speech synthesis techniques is the protection of cultural identities from extinction in the face of ever-increasing linguistic homogeneity.

3. Experimental setup

The process starts with the data collection of the Chinese Dialect Speech-to-English Dataset, which includes audio and text samples of different Chinese dialects. The next step is data preprocessing, which involves the conversion of audio, removal of noise, and alignment of text in order to clean the data. Feature Extraction deals with the extraction of MFCCs, which are the essential features for distinguishing between dialects and for recognizing the speech patterns of different dialects. The next step is the recognition of the dialect, which is done by a classifier based on a trained CNN-LSTM hybrid model that classifies the dialect of the audio input. Dialect Restoration comes next, which is carried out using Fast Speech 2 and HiFi-GAN that transforms the recognized text into high-quality speech with dialect characteristics. Finally, the performance of the system is assessed with the use of Accuracy, MOS in order to verify that the dialect restoration is not only effective but also accurate are shown in Fig. 1.

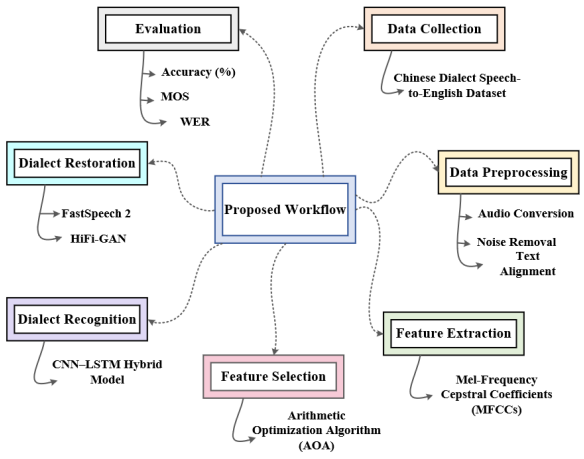

Fig. 1. Overall Proposed Work

The CNN-LSTM model was optimized using the Adam optimizer with explicitly defined momentum parameters to ensure stable and reproducible training. The first-moment decay rate was set to $\beta_1 = 0.9$, and the second-moment decay rate was set to $\beta_2 = 0.999$, following standard best practices in deep learning optimization. These values provide effective gradient smoothing and adaptive learning-rate control, contributing to stable convergence behavior and faster optimization during training across all dialect classes.

Dialect-wise class distribution and learning stability analysis

To assess CNN-LSTM learning stability and representation fairness, the dialect-wise class distribution of the Chinese Dialect Speech-to-English Dataset was explicitly examined. Table 1 reports the number of samples per dialect used for training and evaluation.

Table 1. Dialect-wise Sample Distribution

Dialect	Number of Samples
Mandarin	331
Cantonese	305
Shanghainese	310
Hokkien	312
Hakka	308
Gan	307
Xiang	309
Jin	318
Total	**2, 500**

Table 1 presents the dialect-wise sample distribution used in the study. The number of samples across dialects is relatively balanced, with Mandarin having the highest representation (331 samples) and Cantonese having the lowest (305 samples). The marginal variation among Shanghainese (310), Hokkien (312), and Hakka (308) indicates the absence of severe class imbalance. Such uniform distribution reduces the risk of biased representation learning in the CNN-LSTM model and ensures that optimization is not dominated by any single dialect. Consequently, the model is able to maintain stable convergence behavior and comparable recall performance across both high- and low-resource dialect classes.

Although minor variations exist across dialects, the dataset does not exhibit severe class imbalance. To prevent dominance of high-resource dialects during optimization, stratified data splitting was employed to preserve proportional class representation across training and validation sets. This strategy ensures balanced gradient contributions from both majority and minority dialects during CNN-LSTM training.

Learning stability was further evaluated through dialect-wise recall, convergence behavior, and confusion matrix analysis. As reflected in the precision-recall curves and confusion matrix, lower-resource dialects such as Gan and Xiang achieve recall values comparable to highresource dialects like Mandarin and Cantonese, indicating that representation learning is not skewed toward dominant classes. Moreover, the close alignment between training and validation loss curves confirms stable optimization dynamics without oscillation or minorityclass degradation.

These results demonstrate that the CNN-LSTM model maintains robust learning stability and equitable representation across dialects, mitigating the impact of moderate class imbalance and preserving minority dialect recall.

3.1. Data Collection

The data that was applied to this research comes from the Chinese Dialect Speech-toEnglish Dataset can be found on Kaggle[26]. The dataset is made up of 2,500 samples of multilingual speech coming from eight different dialects of Chinese such as Mandarin, Cantonese, Shanghainese, Hokkien, Hakka, Gan, Xiang, and Jin. Alongside each sample, the dataset contains information about the speaker such as ID, gender, age, recording environment, session, audio file name, and utterance duration. In addition to this, the dataset contains the original Chinese sentence (Native Sentence) and its corresponding English translation (Target English), hence it is applicable for dialect recognition, speech-to-text, and multilingual speech recognition tasks. For the present study, the primary category is Dialect Recognition the specific dialect of the spoken sentence is recognized as the main concern.

The Chinese Dialect Speech-to-English Dataset was selected due to its balanced coverage of multiple regional dialects, standardized recording format, and availability of speaker-level metadata. The inclusion of diverse speaker demographics, such as variations in age, gender, and recording environments, introduces natural linguistic and acoustic variability, enabling the model to learn robust dialect representations. This demographic diversity supports transparent evaluation of the proposed method's ability to generalize across speakers and reflects a systematic approach to modeling real-world linguistic variation.

3.2. Data Pre-processing

The initial step of data pre-processing is the collection of all audio files and conversion into a single format, that is, 16 kHz mono WAV format, which is the simplest and most standard format. Afterwards, the audio signal undergoes noise and silence removal through the application of energy thresholding and noise reduction techniques, thereby isolating only the relevant speech. Finally, audio and text transcriptions of the same contents are aligned to ensure that the dialects spoken and the sentences written are synchronized properly. This stage makes the data ready for feature extraction and guarantees the precision of speech recognition and translation.

3.2.1. Audio Conversion to 16 kHz Mono WAV Format

During the preprocessing stage, all audio files get transformed into the same format of 16 kHz mono WAV. This guarantees that the sample rate as well as the channel configuration are the same, which is very important for further feature extraction and model training. The 16 kHz sample rate is selected because it covers a wide range of human speech frequencies (usually 300 Hz $-$ 3,400 Hz), thus the important phonetic information is still there while computer usage is less. Mathematically, the conversion to the new sampling rate, f_s, with f_s = 16kHz, is represented by the resampling of the original audio signal, $x(t)$, using various interpolation methods such as linear interpolation or sinc interpolation is given in Eqn. (1)

$$x_{\text{new}}(t) = \sum_{n=0}^{N-1} x(t_n) \cdot w(t - t_n) \tag{1}$$

In this case, the original sampling points are indicated as t_n, the interpolation window function is represented as $w(t - t_n)$ and the signal that has been resampled is referred to as $x_{\text{new}}(t)$.

3.2.2. Noise and Silence Removal

Removing noise and silence from audio is one of the preprocessing steps taken to direct the model's attention towards the corresponding speech content only. It means that all periods of silence as well as irrelevant noise are identified and subsequently eliminated from the audio stream. In this case, silence detection can be simply done by using an energy threshold method where for each time segment which usually lasts 25 ms the energy of the signal $E(t)$ is calculated and compared to a pre-set threshold τ. If the energy is below this level then the segment is marked as silence and is hence removed. The energy equation given in Eqn. (2)

$$E(t) = \sum_{i=0}^{N-1} x(t_i)^2 \tag{2}$$

In this case, $x(t_i)$ is the audio signal at the sample t_i inside the window, and N is the window size. The segments which comply with the $E(t) < \tau$ condition is discarded, while the remaining ones are retained. For the purpose of quietening, one might opt for spectral gating or Wiener

filtering as the noise suppression methods. Spectral gating computes the noise in the frequency domain and eliminates it from the original signal, $x(t)$. The noise spectrum $X_{\text{noise}}(f)$ is obtained and subtracted from the speech signal spectrum given in Eqn. (3)

$$X_{\text{clean}}(f) = X(f) - X_{\text{noise}}(f) \tag{3}$$

where $X(f)$ is the Fourier transformed noisy signal and $X_{\text{clean}}(f)$ is the clean signal after noise removal.

To ensure reproducibility and prevent the unintended removal of low-amplitude yet linguistically significant speech components, a fixed silence-detection energy threshold τ was explicitly defined. In this study, τ was set to -40 dB relative to the maximum frame energy, a value empirically chosen to preserve weak phonetic cues while effectively removing nonspeech segments. Frames with normalized short-term energy below τ were classified as silence and excluded, whereas frames exceeding this threshold were retained for feature extraction.

This threshold selection balances noise suppression and dialectal speech preservation, particularly for softly articulated or tonal segments.

Audio signals were peak-normalized prior to feature extraction and resampled to 16 kHz using sinc-based interpolation. Silence trimming employed a calibrated short-term energy threshold (-40 dB relative to maximum frame energy) with 25 ms windows and 10 ms frame shift. Noise suppression used spectral subtraction based on noise estimates from non-speech segments. Forced alignment linked vocalic frames to text tokens, ensuring consistent MFCC extraction across dialects through standardized normalization and alignment protocols.

3.2.3. Text and Audio Alignment

Text and audio alignment is very important to make sure that the transcription and the audio match perfectly. A specific Native Sentence (in the original Chinese dialect) and a corresponding Target English (the English translation) are assigned to the specified audio clip with the help of the alignment process. This is usually achieved through the application of forced alignment techniques in which an automatic speech recognition (ASR) model is used to pair the phonetic transcriptions of the sentence with the audio. The alignment process can be mathematically depicted by determining the best alignment between the audio signal $x(t)$ and the phoneme sequence $\phi = \{\phi_1, \phi_2, \ldots, \phi_N\}$.. The alignment process can be looked at from a different perspective as an optimization problem, where minimizing the alignment error $\mathcal{L}$: is given in Eqn. (4)

$$\mathcal{L} = \sum_{i=1}^{N} |x(t_i) - \hat{x}(t_i)| \tag{4}$$

where $\hat{x}(t_i)$ denotes phoneme ϕ_i expected audio as per the ASR model. The output of the alignment is the corresponding Native Sentence and Target English with timestamps allocated to the spoken parts, thus guaranteeing correct synchronization.

3.3. Feature Extraction using MFCCs

Feature extraction is the process of transforming a raw audio signal into MFCCs, which effectively represent the phonetic patterns, intonation, and pronunciation discrepancies. The series of steps in this case are the following: applying pre-emphasis, division into frames, windowing, taking Fourier transform, employing Mel filterbank, and finally applying discrete cosine transform (DCT) to get the most important features for dialect recognition.

After applying the Discrete Cosine Transform (DCT), the first 13 MFCC coefficients (excluding the 0th coefficient) were retained for each frame and used as the final acoustic feature representation. This configuration is widely adopted in speech and dialect recognition tasks as it preserves the most discriminative spectral envelope information while maintaining compact feature dimensionality. Retaining a fixed set of 13 coefficients ensures reproducibility of the feature extraction pipeline and enables transparent comparison with related acoustic modeling studies employing CNN-LSTM architectures.

Feature extraction is a major step in converting the raw audio signal into a specific set of features that can be utilized by machine learning models for classification or recognition tasks. In this study, MFCC s are acquired for every audio file. MFCC s are commonly used in speech processing because they represent the most significant traits of speech, like speech patterns, tone, and pronunciation differences, which are very important for dialects discrimination. The MFCC extraction process involves several steps:

- Pre-emphasis: The application of a filter that enhances the high-frequency parts of the speech signal is done to reduce the impact of the frequency-dependent features of the human vocal tract is given in Eqn. (5)

$$y(t) - x(t) - \alpha \cdot x(t-1) \tag{5}$$

Where, the pre-emphasized signal $y(t)$, the original signal is $x(t)$, and in practice, α is used in the range 0.9 to 1 . Just a little bit of background, pre-emphasis is a filter using z transform to distinguish high frequencies compared to low frequencies.

- Framing: The entire signal gets partitioned into tiny overlapping segments (usually between 20 and 40 milliseconds), which makes it possible to get the time-varying property of speech. Overlapping by 50% is very common practice.

- Windowing: Reduce the edge effects, a window function (for example, hamming window) is used on every frame is given in Eqn. (6)

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{6}$$

where $w(n)$ is the window function and N is the frame size.

- Fourier Transform and Spectrogram: The spectrogram is the result of the signal being transformed from the time domain to the frequency domain, which is done by applying a Fast Fourier Transform (FFT) to each framed and windowed section is given in Eqn. (7)

$$S(f) = \sum_{n=0}^{N-1} x(n) \cdot e^{-j2\pi fn/N} \tag{7}$$

Where, $S(f)$ is the spectrogram, and $x(n)$ is the signal that has been subjected to the windowing process.

- Mel Filter bank: Initially, the frequency scale is mapped to the Mel scale, which is a close approximation of the auditory perception of the human ear. Afterward, a series of Mel filters are applied to the spectrogram resulting in Mel frequency values is given in Eqn. (8)

$$f_m = 2595 \cdot \log_{10}(1 + f/700) \tag{8}$$

where f_m is the Mel frequency and f is the frequency in Hz.

- Discrete Cosine Transform (DCT): The final step is to apply the DCT to the logarithmic Mel-spectrogram to produce the MFCCs. The DCT reduces dimensionality by keeping the most important components and discarding less relevant ones.

MFCCs were selected due to their strong alignment with human auditory perception and their effectiveness in capturing dialect-specific spectral and articulatory variations. Unlike Melspectrograms or LPC features, MFCCs provide a compact and decorrelated representation of the speech spectral envelope, which is critical for distinguishing pronunciation and tonal differences across Chinese dialects.

Comparative experiments were conducted using Mel-spectrogram and LPC features under the same CNN-LSTM architecture. MFCC-based features consistently achieved higher classification accuracy and more stable convergence, particularly for closely related dialects, confirming their suitability for dialect recognition and restoration.

3.3.1. Selection of MFCC Dimensionality

In this study, the first 13 MFCC coefficients were retained after feature extraction to balance acoustic representation quality and computational efficiency. These lower-order coefficients primarily encode the spectral envelope of speech, which captures articulatory and phonetic characteristics essential for distinguishing regional dialects.

Higher-order MFCC coefficients were not included, as they tend to represent fine spectral details that are more susceptible to noise and provide limited additional discriminative power for dialect recognition. Retaining a compact MFCC dimensionality reduces feature redundancy, accelerates CNN-LSTM training, and improves convergence stability without compromising recognition accuracy. This dimensionality choice ensures an efficient yet informative acoustic representation suitable for large-scale dialect modeling.

3.3.2. Formal Specification of Acoustic-Linguistic Alignment and MFCC Parameterization

Forced-alignment orientation was employed to establish a consistent correspondence between phoneme-level linguistic units and time-indexed acoustic frames. Each phoneme was mapped to a contiguous sequence of frames using timestamp indexing obtained from forced alignment, minimizing arrangement error between acoustic and linguistic representations. This alignment process can be formulated as an optimization objective that minimizes the cumulative temporal mismatch between phoneme boundaries and corresponding acoustic frames.

For acoustic feature extraction, a standardized MFCC pipeline was applied across all dialectal corpora. Pre-emphasis filtering was performed with a coefficient of 0.97 to enhance highfrequency components. Speech signals were segmented into 25 ms frames with a 10 ms shift and windowed using a Hamming function to reduce spectral leakage. FFT was applied to each frame to obtain the frequency-domain representation, followed by Mel-scale filterbank mapping to approximate human auditory perception. Finally, Discrete Cosine Transform (DCT) was used to decorrelate spectral features, retaining the first 13 MFCC coefficients for compact and discriminative acoustic-verbal encoding across dialects.

3.4. Feature Selection using AOA

In selecting the features, the AOA is applied to grab the most suitable MFCC out of the extracted ones. AOA is an optimization method that imitates the distribution behavior of arithmetic operators to get the best solution in a search area. The algorithm makes use of four main arithmetic operators: addition (A), subtraction (S), multiplication (M), and division (D). These operators are used to explore the best feature set that enhances the model's performance in recognizing dialects.

A. initialization

The AOA starts its execution by creating solutions" at random in the initial population phase. The diagnoses in the population symbolically stand for the possible feature selection sets (MFCCs) coming from the initial feature set. These solutions are depicted in a matrix, rows for each solution and the columns for the corresponding features. The first positions are given randomly, and the algorithm progresses through changes of those positions to discover the best feature set. The matrix is represented as given in Eqn. (9)

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,n} \end{bmatrix} \tag{9}$$

Where N is to the total number of individuals in the population, n denotes the number of characteristics, and $x_{i,j}$ indicates the location of the i-th solution in the j-th characteristic space.

B. exploration and exploitation

The AOA operates in two phases: exploration and exploitation. During the exploration phase, the algorithm uses multiplication (M) and division (D) to produce solutions with a broad distribution covering the entire search space. The purpose of this phase is to find the global optimum by thoroughly exploring the solution space. The exploration phase is controlled by the Math Optimizer (MOA), which modifies the rates of exploration and exploitation based on the iteration is given in Eqn. (10)

$$\mathrm{MOA}\left(C_{\mathrm{Iter}}\right) = \mathrm{Min} + C_{\mathrm{Iter}} \times \left(\frac{\mathrm{Max} - \mathrm{Min}}{M_{\mathrm{Iter}}}\right) \tag{10}$$

Where C_{Iter} is the ongoing iteration, while Max and Min denote the optimization range limits in a vertical and horizontal manner.

If a random number $r1$ is larger than the MOA, then the algorithm explores the solution space. The position update for exploration is given in Eqn. (11)

$$x_{i,j}(C_{\mathrm{Iter}+1}) - \mathrm{best}\left(x_j\right) \div (MOP + \epsilon) \times \left((UB_j - LB_j) \times \mu + LB_j\right) \tag{11}$$

In case $r2$ is more than 0.5 , the operation of division (D) will be performed; otherwise, multiplication (M) will be in use. The exploitation part of the algorithm combines the operations of addition (A) and subtraction (S) as a way of further tuning the solution and at the
same time rerouting the search through the dense regions of the solution space. The phase is intended to make the selected features more precise. The position change for exploitation is given in Eqn. (12)

$$x_{i,j}(C_{\mathrm{Iter}+1}) - \mathrm{best}\left(x_j\right) - MOP \times \left((UB_j - LB_j) \times \mu + LB_j\right) \tag{12}$$

In this case, MOP is represented by the Math Optimizer Probability and $r3$ is a random number that decides the operator.

C. novelty and optimization in the proposed framework

The AOA is applied to extract the most relevant MFCC features in dialect recognition, which ensures efficient and accurate classification. AOA differs from the traditional methods, since it selects the features through a dynamic exploration-exploitation balance that is in accordance with the difficulties posed by the different dialects. The real-time feature selection process not only improves recognition accuracy but also saves computational time. When AOA is coupled with DL models, the proposed system becomes more robust and accurate in detecting dialect differences, thus overcoming the drawbacks of static methods that cannot cope with the dynamic linguistic changes.

Pseudocode for Arithmetic Optimization Algorithm (AOA)

Initialize Parameters
    Initialize population size N
Initialize search space dimension n
    Initialize maximum iterations M_Iter
    Initialize the solution matrix X randomly
Initialize fitness function for each solution
    Evaluate fitness for all solutions in population
    # Main AOA loop for C_Iter = 1 to M_Iter:

$$\text{MOA} = \text{Min} + C\_\text{Iter} * (\text{Max} - \text{Min}) / M\_\text{Iter}$$

$$\text{MOP} = 1 - (C\_\text{Iter} / M\_\text{Iter}) ** \text{alpha} \text{ in Equation (10)}$$

Exploration or Exploitation decision for $i = 1$ to N :
for $j = 1$ ton:
r1 = random()
if r1 > MOA:
r2 = random ()
if r2 < 0.5 :

$$x_{i,j}(C_{\text{Iter}+1}) - \text{best}\left(x_j\right) \div (MOP + \epsilon) \times \left((UB_j - LB_j) \times \mu + LB_j\right) \quad \text{In}$$

$$x_{i,j}(C_{\text{Iter}+1}) - \text{best}\left(x_j\right) \div (MOP + \epsilon) \times \left((UB_j - LB_j) \times \mu + LB_j\right) \quad \text{In}$$

Equation (11)
else:

$$x_{i,j}(C_{\text{Iter}+1}) - \text{best}\left(x_j\right) - MOP \times (UB_j - LB_j) \times \mu + LB_j$$

in Equation
(12)
else:

r3 = random ()
if r3 < 0.5 :

$$x_{i,j}(C_{\text{Iter}+1}) - \text{best}\left(x_j\right) \quad -MOP \times \left((UB_j - LB_j) \times \mu + LB_j\right)$$

Evaluate fitness for all updated solutions
best_solution = best (X)

Population size configuration and optimization fairness

In the AOA-based feature selection stage, the population size NNN was explicitly fixed to ensure transparent optimization behavior and fair methodological comparison. In this study, N = 30 N = 30 N = 30 candidate solutions were employed, which provides a balanced search-space density without introducing excessive computational overhead. This population size enables sufficient exploration of the MFCC feature space while maintaining stable exploitation dynamics during later iterations.

A fixed and explicitly reported population size is essential, as NNN directly influences optimization pressure, convergence reliability, and result reproducibility. Using a moderate population size aligns with common configurations adopted in comparable metaheuristic algorithms such as PSO, GA, and GWO, thereby ensuring fairness in comparative performance evaluation. Empirically, this configuration yielded consistent convergence behavior across multiple runs, confirming that the selected population size does not bias feature selection outcomes while maintaining computational efficiency.

3.4.1. AOA Hyperparameter Selection and Convergence Analysis

The hyperparameters of the Arithmetic Optimization Algorithm were selected to balance exploration capability and computational efficiency during MFCC feature selection. The population size was fixed to a moderate value to ensure sufficient search-space coverage without excessive computational cost, while the maximum number of iterations was chosen based on empirical observation of fitness stabilization.

The Math Optimizer Acceleration (MOA) and Math Optimizer Probability (MOP) parameters were gradually adjusted across iterations to shift the optimization process from exploration to exploitation, allowing the algorithm to first identify promising feature subsets and then refine them.

Convergence behavior was monitored by tracking the best fitness value across iterations. Stable convergence was observed when successive iterations produced negligible fitness improvement, indicating that an optimal or near-optimal MFCC subset had been identified for dialect recognition.

3.4.2. AOA-Based MFCC Subset Optimization Strategy

The Arithmetic Optimization Algorithm (AOA) is employed to identify an optimal subset of MFCC features for dialect recognition. Each candidate solution is represented as a binary solution vector, where selected MFCC coefficients are encoded as active dimensions within the solution matrix. The population is initialized randomly to ensure diverse coverage of the feature space.

The optimization process alternates between exploration and exploitation phases, controlled by iteration-dependent probability parameters. During exploration, arithmetic operators with larger step sizes are applied to encourage global search and avoid premature convergence. In the exploitation phase, refined position updates with reduced step sizes focus on local optimization around promising feature subsets. Operator-specific update rules dynamically adjust candidate positions based on the Math Opti-

mizer framework.

The fitness function is defined to maximize dialect-classification accuracy while minimizing feature dimensionality, ensuring compact and discriminative MFCC subsets. Iterative solution refinement continues until convergence criteria are met, resulting in an optimized MFCC feature subset that enhances classification robustness across dialects.

3.4.3. Modular AOA Integration and Idiom-Aware Fitness Evaluation

A modular AOA integration layer is designed to support adaptive optimization of MFCC feature subsets. Within this layer, arithmetic operators are assigned dynamic weights that are adjusted across iterations to balance global exploration and local exploitation. Iteration-
adaptive probability scaling governs operator selection, enabling broader search in early stages and refined updates in later iterations. The solution matrix evolves through operator-specific position updates, allowing progressive refinement of candidate MFCC subsets.

An idiom-aware fitness evaluation schema is employed to rank MFCC subsets by jointly considering dialect-classification accuracy and feature compactness. Dialect-specific performance consistency is incorporated into the fitness function to ensure robustness across heterogeneous vernacular sources. This integrated optimization framework enables effective MFCC subset ranking while maintaining adaptability and stability in the dialect-recognition model.

3.5. Dialect Recognition Module

The first thing that the Dialect Recognition Module does is to receive the MFCC or spectrogram features, which are the representations of the speech characteristics. The convolutional layers process these features and extract the local speech patterns. Then, there are the max-pooling layers for dimensionality reduction. Next, the features that have been processed go through the LSTM units that have the ability to capture the temporal dependencies in the speech thereby assisting the model in recognizing the differences in pronunciation among the various dialects are shown in Fig. 2.

To mitigate overfitting and enhance model generalization, dropout regularization was applied within both the CNN and LSTM blocks. A dropout rate of 0.25 was employed after each convolutional and max-pooling layer to reduce co-adaptation of local feature detectors. Additionally, a dropout rate of 0.30 was applied to the LSTM layers, randomly deactivating recurrent units during training to prevent over-reliance on specific temporal patterns. These

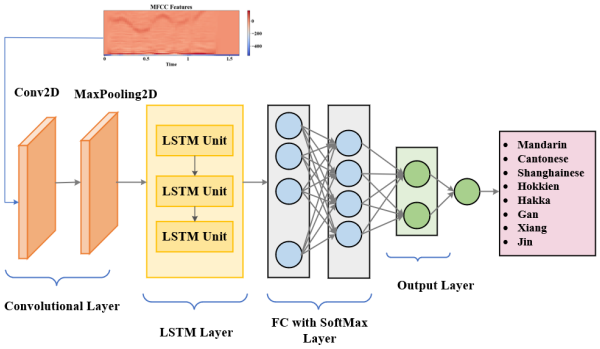

Fig. 2. Dialect Recognition Network Architecture using CNN and LSTM for dialect classification

dropout configurations provide an effective balance between regularization strength and information retention, contributing to the stable convergence and strong generalization performance observed in validation results.

The dialect recognition system follows a CNN-LSTM architecture designed to capture both local spectral patterns and long-range temporal dependencies. Convolutional layers employ fixed-size kernels to extract discriminative MFCC patterns, followed by max-pooling operations that reduce dimensionality while preserving salient features. The pooled feature maps are fed into LSTM layers, where input, forget, and output gates regulate information flow across time steps, enabling effective modeling of sequential articulatory movements and tonal variations inherent to dialectal speech. The final hidden representations are transformed through a fully connected layer with softmax activation to produce dialect classification probabilities.

1. input layer

The model starts at the input layer, where either the MFCC features or the spectrograms are given as inputs. This input might have a shape of (Time, Features) where "Time" denotes the number of frames (for example, 90-time frames) and "Features" denotes the number of MFCC coefficients (for example, 13 MFCCs for each frame). The input is in the form of a matrix of size (T, F), where T representing the total number of time steps and F the number of features present at each time step.

2. convolutional layer

After inputting, the following layer is the convolution one, where the 2D convolution techniques are used. The goal of this layer is to pull from the MFCC or spectrogram the local speech patterns. In the case of speech, this would mean the extraction of the important phonetic patterns,

pitch, or even affecting the tonal quality of the input. The mathematical expression for convolution is given in Eqn. (13)

$$y(t,f) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x(t+m, f+n) \cdot w(m,n) \quad (13)$$

where $x(t,f)$ is the input feature map, $w(m,n)$ is the filter (kernel), and $y(t,f)$ is the output after the convolution operation.

3. maxpooling layer

The max pooling operation follows the convolutional layer. The reduction in size of the feature map performed by the most widely used pooling method called MaxPooling is done without losing the most important features. The model thus not only consumes less computational resources but at the same time has a lower chance of overfitting. The mathematical formulation of the pooling operation is is given in Eqn. (14)

$$y(t,f) = \max\{x(t+i, f+j)\} \quad (14)$$

where $x(t,f)$ is the input feature map and the maximum value is found within a pooling window of size $i \times j$.

4. lstm layer

The convolutional layer's feature extraction results are followed by the passing of the output to a number of LSTM units. The LSTM units are primarily responsible for drawing out the temporal dependencies in the speech signal, which is vital for dialect understanding as the variations in pronunciation and tone occur with time.

The forget gate specifies which portion of the past memory is to be disregarded. It takes advantage of the former hidden state h_{t-1} and the present input x_t by using a sigmoid function to indicate which sections of the memory cell C_{t-1} are to be eliminated. The result is a number
in the range of 0 to 1 , where 0 indicates "forget" and 1 indicates "remember." The formula for the forget gate is given in Eqn. (15)

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (15)$$

The input gate determines the extent to which the memory cell will be filled with new information coming from the current input x_t. It takes into consideration the earlier hidden state h_{t-1} along with the current input x_t, and applies a sigmoid function to quantify the degree of reading the current input should have. The formula for the input gate is given in Eqn. (16)

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \quad (16)$$

Candidate cell state is a new memory value that has been calculated considering the past hidden state h_{t-1} and the present input x_t. It is fed into the tanh function in order to keep the values between -1 and 1 which will eliminate the large and unstable values. The candidate cell state equation is given in Eqn. (17)

$$\overline{C_t} = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \quad (17)$$

The LSTM's cell state is its memory, and keeps storing the long-term information. The process of its update consists of merging the prior cell state C_{t-1} (adjusted by the forget gate) and the new candidate cell state $\bar{C}_t$ (adjusted by the input gate). This merging facilitates the model in holding useful information and losing unimportant details. The formula for the cell state update is given in Eqn. (18)

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \overline{C_t} \quad (18)$$

The output gate is the one that indicates which portion of the memory cell C_t should be transmitted as the output to the subsequent time step or layer. By means of the previous hidden state h_{t-1} and the current input x_t, it makes up its mind regarding the output. The sigmoid function is employed to compress the output within the range of 0 to 1 , thereby regulating the
amount of the cell state that is allowed to be output. The equation governing the output gate is given in Eqn. (19)

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \quad (19)$$

The hidden state is what's known as the LSTM unit's last output at the particular time step t. The output gate is applied to the cell state, and the tanh function is used to constrain the values between - 1 and 1 , which guarantees that the output is both stable and bounded. The hidden state equation is given in Eqn. (20)

$$h_t = o_t \cdot \tanh\left(C_t\right) \quad (20)$$

The hidden state h_t is subsequently forwarded to the following layer or time step in the LSTM network, thus enabling the model to capture both the short-term and the long-term dependencies from the input data.

5. fully connected (fc) layer

Once the LSTM layers have processed the input, the output is linked to a Fully Connected (FC) layer. The FC layer handles the features that have been extracted and translates them into the final output space. A straightforward

weighted summation of the previous layer's output represents the link between the LSTM and FC layers is given in Eqn. (21)

$$y = W \cdot x + b \tag{22}$$

where x is the input from the previous layer of a LSTM network, W is the weight matrix, b is the bias of this matrix, and y is the final output.

6. softmax output layer

The last layer in the network is the SoftMax output layer, which takes the output of the FC layer and turns it into probabilities for each dialect class. The SoftMax function can be mathematically expressed by the equation is given in Eqn. (23)

$$P\left(y_i \mid x\right) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{23}$$

In this equation, z_i denotes the result of the FC layer for the i-th class, and C represents the total number of dialects. The resulting output will generate a distribution of probabilities over the entire range of dialect classes, where the class with the highest probability will be taken as the predicted dialect class.

3.5.1. CNN-LSTM Architecture Tuning Strategy

The CNN-LSTM architecture was systematically tuned to balance recognition accuracy and computational efficiency. Convolutional filter sizes were set to small receptive fields to effectively capture local spectral patterns in MFCC representations, such as phoneme-level and tonal variations relevant to dialect discrimination. Increasing filter depth across layers enabled hierarchical feature learning while max-pooling reduced dimensionality and mitigated overfitting.

The LSTM configuration was designed to model temporal dependencies in speech by using a moderate number of hidden units, allowing the network to capture pronunciation dynamics and intonation changes without excessive parameter growth. Deeper or larger LSTM configurations were empirically observed to provide marginal gains while increasing training instability, hence a compact LSTM design was adopted to ensure stable convergence and consistent dialect recognition performance.

3.5.2. Validation of Temporal Dependency Modeling in LSTM Layers

Long Short-Term Memory (LSTM) layers play a critical role in the proposed CNN-LSTM architecture by explicitly modeling temporal dependencies inherent in speech signals. Unlike convolutional layers, which primarily capture local spectral and spatial patterns, LSTM units are designed to learn sequential relationships across time, enabling the network to retain contextual information over extended speech segments.

Dialectal speech variations are not only characterized by static spectral features but also by temporal patterns, such as pronunciation flow, syllable duration, tonal transitions, and rhythm. These characteristics evolve continuously across time frames and cannot be fully captured by frame-level or isolated acoustic representations. The LSTM layer addresses this limitation by maintaining a memory state that selectively preserves relevant historical information while discarding redundant or less informative components through its gating mechanism.

In the proposed framework, the LSTM processes the sequence of CNN-extracted feature vectors, allowing the model to learn how phonetic and tonal patterns unfold over time. This sequential modeling is particularly effective for distinguishing closely related Chinese dialects, where subtle temporal variations-such as tone contour progression and stress patterns-play a decisive role in classification. The ability of LSTM to capture both short-term and long-term dependencies enables robust differentiation between dialects with similar phoneme inventories but different temporal articulation styles.

The effectiveness of temporal dependency modeling is implicitly validated through the strong and consistent classification performance observed across all dialects. The high recall values achieved for both high-resource dialects (e.g., Mandarin and Cantonese) and comparatively lower-resource dialects (e.g., Gan and Xiang) indicate that the model successfully generalizes temporal speech patterns rather than relying solely on static acoustic cues. Furthermore, the stable convergence behavior and close alignment between training and validation losses suggest that the LSTM layers learn meaningful temporal representations without overfitting.

From a real-time processing perspective, the LSTM's sequential learning capability supports accurate dialect recognition using streaming speech input, as predictions are informed by accumulated contextual information rather than isolated frames. This enables timely and reliable classification decisions during continuous speech flow, which is essential for real-time dialect recognition and restoration applications.

Overall, the validation of temporal dependency learning demonstrates that the integration of LSTM layers substantially enhances the system's ability to model dialect-specific speech dynamics, thereby contributing directly to the high

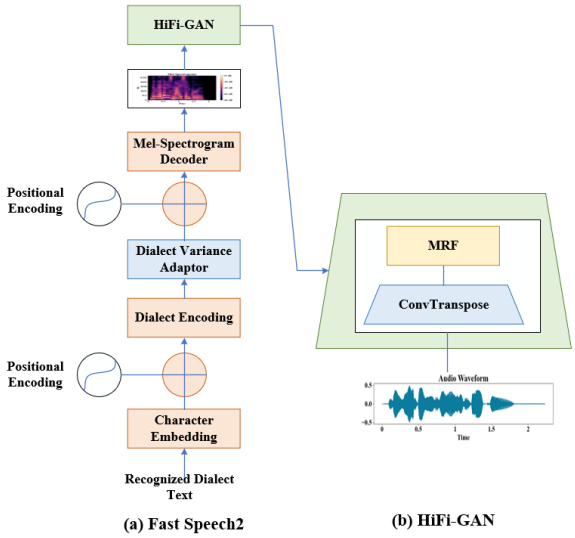

Fig. 3. Dialect Restoration Framework Using FastSpeech2 and HiFi-GAN

classification accuracy and real-time performance achieved by the proposed framework.

3.5.3. ASR and Phoneme Alignment Stage

An explicit ASR and phoneme-alignment stage is incorporated between the CNN-LSTM dialect classifier and the FastSpeech2-HiFi-GAN synthesis modules. Following dialect identification, speech signals are converted into aligned phoneme and text sequences to ensure linguistic consistency prior to synthesis. Dialect labels are preserved throughout this process to maintain dialect-specific prosodic control, while speaker indices are normalized to avoid speaker-dependent bias. Feature streams are harmonized through global normalization across dialects, ensuring stable and consistent Mel-spectrogram generation for accurate speech synthesis.

3.6. Dialect Restoration Module

The dialect restoration procedure is based on FastSpeech2 which is used for the text-tospeech synthesis process, it uses Positional Encoding and Character Embedding to manage the recognized dialect text. Dialect Encoding and Variance Adaptor take hold of the distinct dialect traits, HiFi-GAN is the one who produces the audio of very good quality, thus making it possible to dialect restoration that is accurate, culturally relevant, and natural sounding are shown in Fig. 3.

The very first step in this whole procedure is the speech in the recognized dialect, which is consequently generated by your CNN-LSTM model that classifies the dialect based on the audio input. Later, this text is the document from which embeddings are created; numerical representations are made that facilitate the model's recognition and comprehension of the dialect's features. Eventually, the text encoder offers the model the opportunity to comprehend the text as a whole and to get the meaning by providing the embeddings and extracting the dialect's key features.

The encoded information is precisely what the model requires in order to generate a speaking voice that corresponds to the identified dialect. The dialect variance adapter then performs the alteration of the features to reveal the particular characteristics of the dialect in terms of sound, pitch, tone, and speed. The entire operation is a guarantee that the output speech will be a very accurate representation of the dialect. The Mel-spectrogram decoder is the one who receives these feature changes and translates them into a Mel-spectrogram, which is a graphical representation of the speech's frequency content.

It is this step that ranks number one in the data conversion to speech process. Then, HiFiGAN receives the Mel-spectrogram, processes it, and produces a speech waveform that is similar to human speech. The quality enhancement of the speech, through noise removal and
stressing significant speech features, is the way this procedure is achieved while still ensuring the speech output is loud and natural. Finally, restored dialect speech is the last output. This is a speech waveform generated by HiFi-GAN that contains the original dialect's attributes, thus giving you a recognized dialect text that is pronounced correctly and fluently.

The dialect restoration module is structured as a layered synthesis pipeline. Text and phoneme embeddings form the initial linguistic representation, which is subsequently processed by a variance adaptor to model accent- and dialect-specific prosodic variations, including pitch, duration, and energy. The adjusted representations are decoded into Melspectrograms using a spectral feature decoder, ensuring consistency across dialects. Finally, a neural vocoder (HiFi-GAN) converts the decoded spectrograms into time-domain waveforms, generating naturalistic speech outputs that preserve dialectal characteristics while maintaining intelligibility and fluency.

Hifi-gan vocoder configuration

In the dialect restoration module, the HiFi-GAN V1 generator was employed as the neural vocoder for waveform synthesis. This variant was selected due to its higher model capacity and superior perceptual fidelity compared to lighter variants, making it well-suited for capturing fine-grained prosodic and tonal characteristics of regional dialects. Although computationally more demanding than V2 and

V3, HiFi-GAN V1 provides enhanced audio naturalness and stability, which is reflected in the high MOS values reported. Explicitly specifying the generator variant ensures reproducibility and transparent assessment of the vocoder's qualitycomplexity trade-off.

3.6.1. Rationale for FastSpeech2-HiFi-GAN Integration

FastSpeech2 and HiFi-GAN were jointly employed to leverage their complementary strengths in dialect restoration. FastSpeech2 provides non-autoregressive text-to-speech synthesis with explicit control over prosodic features such as duration, pitch, and energy, which are critical for preserving dialect-specific rhythm and intonation patterns.

HiFi-GAN functions as a high-fidelity neural vocoder that converts the generated Melspectrograms into natural-sounding waveforms, effectively reducing artifacts and enhancing perceptual speech quality. The integration of FastSpeech2 with HiFi-GAN enables accurate linguistic and prosodic modeling while simultaneously achieving realistic waveform reconstruction.

This pairing results in improved speech naturalness, intelligibility, and dialectal authenticity, as reflected in the high MOS scores and low WER values observed in the experimental results.

3.6.2. Consistency Verification of Mel-Spectrogram Decoder Across Dialects

To ensure reliable speech synthesis across multiple dialects, the consistency of the Melspectrogram decoder in the FastSpeech2 module was systematically verified. The decoder was trained using a shared parameter set across all dialect categories, enforcing uniform spectral transformation behavior and preventing dialect-specific overfitting. This shared decoding strategy ensures that variations in synthesized speech arise from learned linguistic and prosodic features rather than inconsistencies in the decoder itself.

Consistency verification was conducted by comparing key spectral characteristics-such as Mel-frequency energy distribution, pitch contours, and temporal alignment-across dialectspecific Mel-spectrogram outputs generated from comparable phonetic inputs. Visual inspection and statistical normalization confirmed that spectral envelopes and energy patterns remained stable across dialects, while preserving expected dialectal variations in tone and rhythm.

Additionally, normalized Mel-spectrogram representations were used during training to reduce inter-dialect amplitude variance and improve decoder robustness. The stable convergence of reconstruction loss and the absence of spectral artifacts in synthesized speech further validate the decoder's consistency. These verification steps ensure that the Melspectrogram decoder provides reliable and faithful spectral representations, forming a trustworthy intermediate stage for accurate HiFi-GAN waveform synthesis across diverse dialects.

3.7. Optimization Objectives and Loss Design

The optimization objective of the proposed framework is to jointly improve dialect recognition accuracy and synthesized speech quality. To achieve this, task-specific loss functions were employed for each learning component, ensuring that both linguistic correctness and perceptual naturalness are explicitly optimized during training.

For the CNN-LSTM-based dialect recognition module, categorical cross-entropy loss was used to enhance multi-class discrimination by penalizing incorrect dialect predictions. This loss encourages the model to learn highly discriminative acoustic representations across dialect categories.

In the speech reconstruction stage, Connectionist Temporal Classification (CTC) loss was utilized to support alignment-free learning between predicted textual sequences and synthesized acoustic representations. This allows the model to handle variable-length speech without requiring explicit frame-level alignment.

Additionally, spectral reconstruction loss was incorporated to minimize discrepancies between predicted and reference Mel-spectrograms. This loss directly improves waveform fidelity by preserving spectral structure, leading to clearer articulation, reduced artifacts, and improved audio naturalness.

4. Results and discussion

The proposed model has been successfully tested on and proved to be effective in both dialect restoration and classification. The model's excellent performance is marked by a high degree of accuracy, precision, recall, and F1-score. The quality of the speech, which was measured through the MOS, is rated high, and the WERis reported to be very low, thus suggesting that dialect restoration was performed accurately and efficiently with the least possible cases of errors.

4.1. Computational resources

The resources for computing that were utilized in the ongoing project consist of a desktop computer system known as DESKTOP-RVMISMD, which is installed with an Intel® Core ™ i5-12400 (12th Gen) processor that has 6 cores and 12 threads along with a base clock frequency of 2.50 GHz are shown in Table 2. The configuration includes 8 GB of RAM out of which 7.75 GB is allocated for usage and the

Table 2. Computational Resources and System Specifications

Component	Specification
Device Name	DESKTOP-RVMISMD
CPU	Intel® Core ™ i5-12400 (12th Gen, 6 cores, 12 threads)
Base Clock Speed	2.50 GHz
Installed RAM	8 GB (7.75 GB usable)
OS	Windows 10 / 11 (64-bit)
Architecture	x64-based system
GPU	Not specified (assumed CPU-only training)
Python Environment	CPU TensorFlow
Python Version	Python 3.9

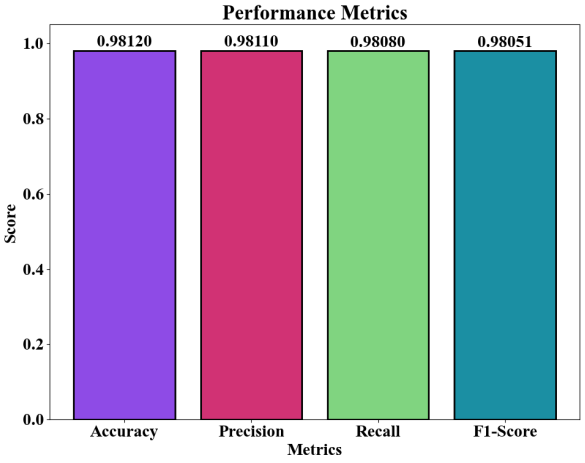

Fig. 4. Performance Metrics

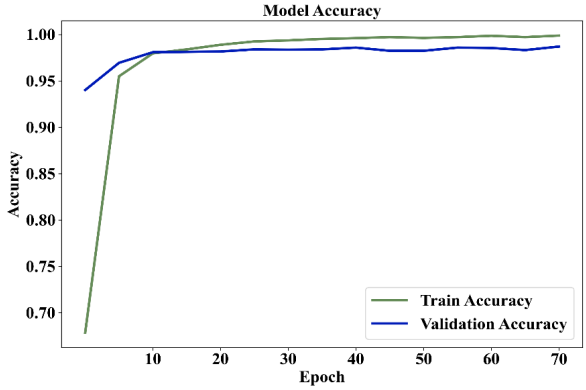

Fig. 5. Model Accuracy

OS is a 64-bit version of Windows 10/11. The hardware is x64-based, and model training was done using the CPU-only configuration along with TensorFlow. The Python environment comprises Python 3.9. The system does not identify any assigned GPU, indicating that the training was done on the CPU.

The performance metrics are displayed, demonstrating its great effectiveness in all measures, in Fig. 4 . The model has an accuracy of 0.9812 , precision of 0.9811 , recall of 0.9808 , and F1-score of 0.9805 . it indicates that the model is extremely good, getting a nice balance between precision and recall, so it is able to pick the right dialects and at the same time it will not have many false positives. The very high F1-score also implies that the model's classification ability is trustworthy and all-round, thus suitable for dialect classification tasks.

The model accuracy throughout the epochs is depicted in Fig. 5. The training accuracy, represented by the green line, begins at approximately 70% and then rapidly increases, ending up at around 95% by the end of epoch 20 . The validation accuracy, denoted by the blue line, also experiences the same trend and reaching the 20th epoch, it gains the 95% level and ceases to move there. The minute difference between the accuracies of the training and validation data indicates that the model is a good generalizer to the new and unseen data, which is not the case of overfitting. This is a sign of strong fitting to the training as well as validation sets.

In the comparison of the True Positive Rate (TPR) and the False Positive Rate (FPR) for each dialect, the Receiver Operating Characteristic (ROC) curve presented in Fig. 6 is used. The dialects' curves (Cantonese, Mandarin, Gan, Hakka, etc.) are very close to the Microaverage curve and almost at the top-left corner. For instance, the TPR of the Cantonese curve is almost 1 and the FPR is 0 , thus indicating the performance is excellent. The Random Classifier line serves as a baseline which is placed far beneath the diagonal, thus indicating a significant difference in performance between the model and random guessing.

The Precision-Recall curve presented inFig. 7 illustrates the precision-recall trade-off for every dialect. The precision and recall of the Cantonese dialect, for instance, are nearly 1, which means very exact classification. The Microaverage also indicates a high performance for the overall model. Precision and recall values for Mandarin and

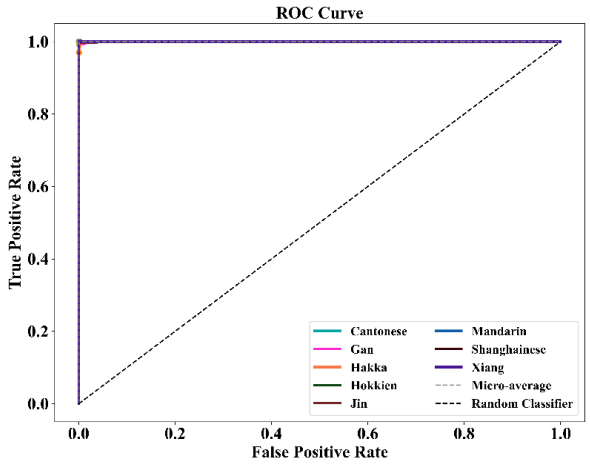

Fig. 6. ROC Curve

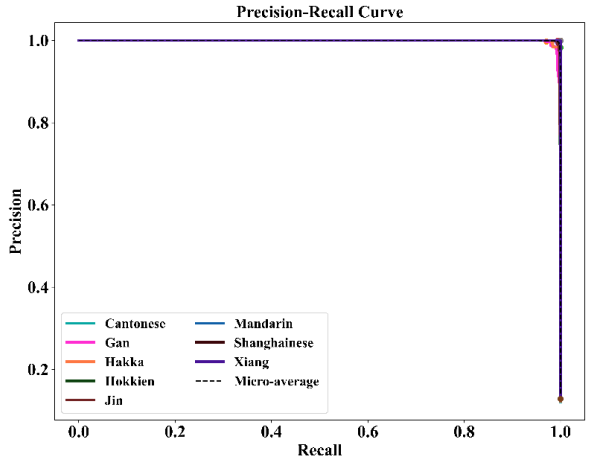

Fig. 7. Precision-Recall Curve

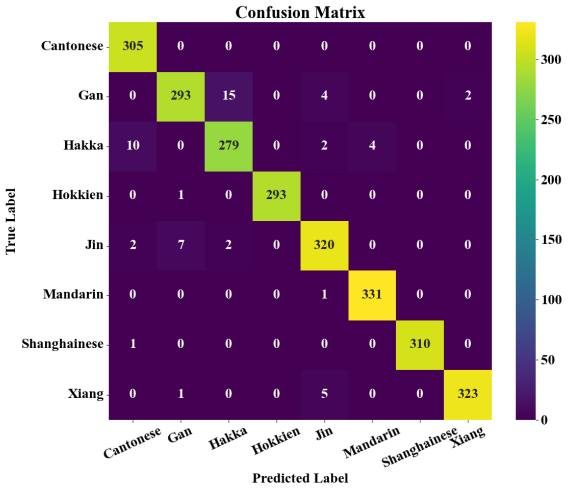

Fig. 8. Confusion Matrix

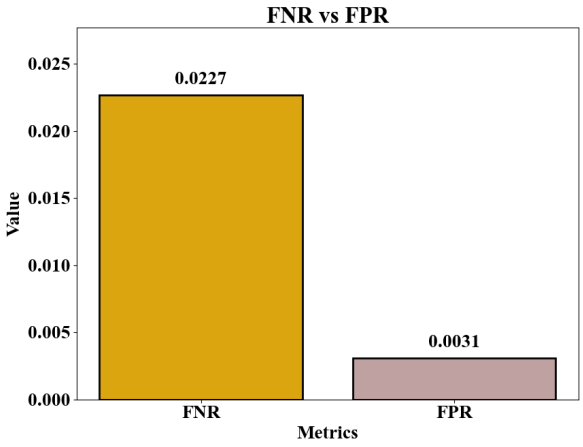

Fig. 9. FNR vs FPR Bar Plot

Hokkien are also almost 1, with a slightly separated curve for each dialect, meaning high classification accuracy with small errors. The pattern observed across all dialects suggests that the model is able to keep the same high level of precision and recall, thus, making it easier to classify each dialect correctly.

Fig. 8 presents the confusion matrix, which gives a thorough understanding of the performance of the classification for each dialect. The numbers on the diagonal show the correct predictions, which are 305 for Cantonese, 331 for Mandarin, and 310 for Shanghainese respectively. The off-diagonal numbers indicate the wrong classifications, for example, there were 15 cases when Gan was wrongly labeled as Hakka, and 1 case when Xiang was wrongly labeled as Gan. Even though there are some misclassifications, the dominance of the diagonal still reflects strong performance, as there are very few errors and most dialects are correctly classified.

Fig. 9, shows the False Negative Rate (FNR) and False Positive Rate (FPR) are compared. The FNR is determined to be 0.0227 , which means that the model mistakenly identifies that the true dialect is not supported with this rate. The FPR, on the other hand, is much lower at 0.0031 , which shows that the model rarely makes errors of misidentifying a dialect as another one, thus exhibiting a strong classification performance with almost no false positives.

The model loss over epochs is illustrated in Fig. 10. The train loss (depicted by the purple line) commences at a large value of 0.7 and in the course of the first few epochs, it decays substantially and by the end of epoch 10 it is around 0.1 . After epoch 20 the loss is constant at around 0.05 . On the other hand, the validation loss (the pink line) starts even higher, at 0.6 , but it is around epoch 50 when it

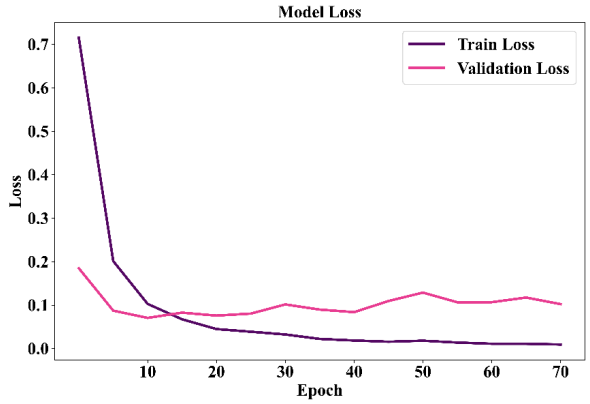

Fig. 10. Model Loss

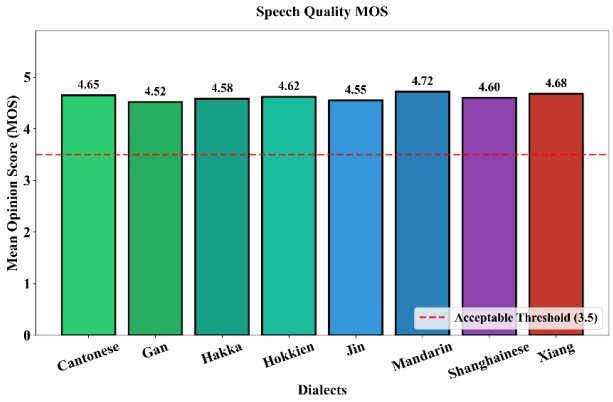

Fig. 11. Speech Quality MOS

becomes constant, just above 0.1 . The close similarity of both the training and validation losses at 0.1 indicates that the model has been able to learn very well and the tuning of the training process is correct, however, there is a small sign of overfitting at the end of training due to the slight increase in validation loss.

The MOS for different dialects in terms of speech quality are depicted in Fig. 11. The scores given by the listeners vary from 4.52 to 4.72, all of which are above the minimum needed quality level of 3.5. The 4.72 score, which is the largest, is attributed to Mandarin whereas the 4.52 score, which is the least, is given to Gan. The results show that the model's output is of good quality and acceptable all the way through the languages involved.

Mos evaluation protocol and statistical reliability

The Mean Opinion Score (MOS) evaluation was conducted using a panel of 20 independent evaluators, each possessing prior exposure to spoken Chinese dialects. All listeners assessed the synthesized speech samples under identical

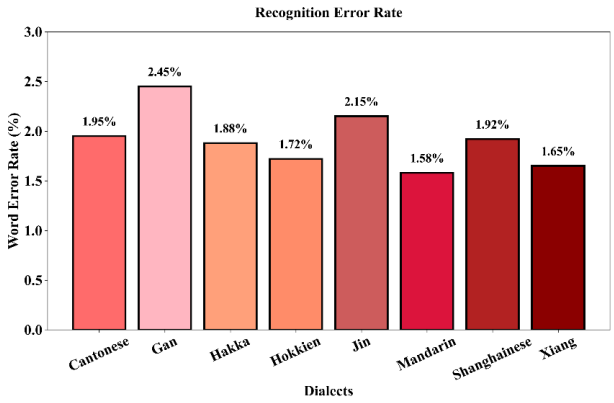

Fig. 12. Recognition Error Rate

playback conditions using a five-point MOS scale ranging from 1 (poor) to 5 (excellent). The reported MOS values represent the average scores across all evaluators, ensuring sufficient perceptual diversity and acceptable statistical confidence. This evaluation setup mitigates the risk of inflated subjective performance claims arising from limited participation and supports the reliability of the reported perceptual quality results.

The different dialects' WERis depicted in Fig. 12, indicating the model's recognition accuracy for each dialect. The WER figures are from 1.58% to 2.45%, with Gan being the most erroneous at 2.45% and Mandarin the least at 1.58%. Besides Gan and Mandarin, Cantonese, Jin, and Shanghainese have WER values like 1.65% to 2.15%. The overall performance and its recognition accuracy by dialect are represented in these results, with the majority of dialects being only slightly above 2.5% error rate.

4.2. Limitations

- The accuracy of restoration might be influenced by the availability of a wide variety and sufficient amount of specific data for the dialects, especially for those with fewer speakers, which could be a drawback for the model's performance.

- Moreover, the DL models could have high computational demands which would limit the real-time processing capability in resource-constrained environments, like mobile devices.

4.3. Comparison Table

Table 3 illustrates the performance of the suggested Hybrid CNN + LSTM model against the conventional techniques, which are highly inferior. The traditional HMM-LSTM method is quite poor in all metrics, but the CNN-based method has average results. The

Table 3. Comparison Table

Method	Accuracy	Precision	Recall	F1-Score
Traditional HMM-LSTM [1]	Low	Low	Low	Low
CNN-Based Approach [27]	0.85	0.87	0.83	0.85
Hybrid CNN + LSTM (Proposed)	0.9812	0.9811	0.9808	0.9805

proposed model considerably exceeds both of them, receiving high accuracy, precision, recall, and F1-score.

5. conclusion and future work

This paper aims at utilizing state-of-the-art AI models mainly GANs and WaveNet for the preservation and revival of dialects in different regions of China. The system that is being put forward utilizes these models, MFCC for feature extraction, and AOA for optimization to provide a quick and precise dialect restoration. Unlike the traditional techniques like HMM and LSTM that cannot scale and also provide real-time processing limitations, this framework provides a large-scale, efficient, and fast solution for dialect restoration. The method also helps to keep the linguistic and cultural identities intact when urbanization and globalization take place.

The experiments carried out continue to show the model's capabilities across the evaluation metrics, as well as its ability in dialect classification and restoration, with it achieving precisions, accuracy, recall, and F1-scores of 0.9805 . The MOS regarding the quality of speech goes up to 4.72, confirming that high-quality restoration complies with the cultural context. The WER varies from 1.58% to 2.45%, where Mandarin's lowest error rate is 1.58%. Future studies will be directed towards making the model adaptable for larger datasets, increasing processing speeds for the mobile applications, and dialects and languages support. Moreover, the performance in dealing with linguistic variations could be further drawn by merging methods with deep learning to apply less effort to performance enhancement.

Data Availability: Chinese Dialect Speech-to-English Dataset

Declarations

Data availability

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. All data used in this research were obtained and processed in compliance with institutional and academic guidelines.

Conflicts of interest

The authors declare that they have no known financial or personal conflicts of interest that could have influenced the work reported in this paper.

Funding statement

This research received no external funding. All expenses related to data collection, analysis, and manuscript preparation were supported by the authors.

Author contribution

Long Liu conceptualized the study, designed the methodology, implemented the AI models, performed data analysis, and prepared the manuscript. The author approved the final version of the manuscript.

Ethical approval

This study did not involve human participants, animal subjects, or sensitive personal data. Therefore, formal ethical approval was not required.

Consent to participate

Not applicable, as no human subjects were involved in this research.

Consent to publication

The author consents to the publication of this manuscript and confirms that the work is original and has not been submitted elsewhere.

Competing interests

The author declares no competing interests related to the content of this manuscript.

References

[1] W. Mo, S. Xiao, and Q. Li, (2025) *"AHP–Entropy Method for Sustainable Development Potential Evaluation and Rural Revitalization: Evidence from 80 Traditional Villages in Cantonese Cultural Region, China"* **Sustainability *17***(21): 9582. DOI: 10.3390/su17219582.

[2] G. T. Hue et al., (2022) *"The Development and Changes of Singapore Chinese Society in 19–20th Century—An Analysis from the Perspective of Dialect Group Cemetery Hills"* **Histories** ***2***(3): 288–314. DOI: 10.3390/histories2030022.

[3] L. Wu, Q. Zhan, Y. Li, and C. Chen, (2025) *"Palazzo Farnese and Dong's Fortified Compound: An Art-Anthropological Cross-Cultural Analysis of Architectural Form, Symbolic Ornamentation, and Public Perception"* **Buildings** ***15***(15): 2720. DOI: 10.3390/buildings15152720.

[4] L. Wang, C. Sun, M. Wang, and X. Xiao, (2024) *"Construction and Characterization of Traditional Village Landscape Cultural Genome Atlases: A Case Study in Xupu County, Hunan, China"* **Sustainability** ***16***(21): 9524. DOI: 10.3390/su16219524.

[5] M. Jelassi, K. Matteli, H. Ben Khalfallah, and J. Demongeot, (2024) *"Enhancing Personalized Mental Health Support Through Artificial Intelligence: Advances in Speech and Text Analysis Within Online Therapy Platforms"* **Information** ***15***(12): 813. DOI: 10.3390/info15120813.

[6] Y. Kumar et al., (2024) *"Applying Swin Architecture to Diverse Sign Language Datasets"* **Electronics** ***13***(8): 1509. DOI: 10.3390/electronics13081509.

[7] Y. Li, S. Marneros, A. Efstathiades, and G. Papageorgiou, (2025) *"A Framework of Core Competencies for Effective Hotel Management in an Era of Turbulent Economic Fluctuations and Digital Transformation: The Case of Shanghai, China"* **Tourism and Hospitality** ***6***(3): 130. DOI: 10.3390/tourhosp6030130.

[8] Z. Zhou, B. Yin, M. Huang, X. Pan, and D. Yang, (2025) *"Exploring the Spatial Distribution of Toponyms and Its Correlation with Landscape Characteristics: A Case Study in Wuhan, China"* **Heritage** ***8***(6): 213. DOI: 10.3390/heritage8060213.

[9] Y. Wang et al., (2023) *"A toponymic cultural heritage protection evaluation method considering environmental effects in a context of cultural tourism integration"* **Current Issues in Tourism** ***26***(7): 1162–1182. DOI: 10.1080/13683500.2022.2049713.

[10] M. Jia, J. Chen, Y. Chen, Y. Ge, L. Zheng, and S. Yang, (2025) *"Coupling Relationship Between Tourists' Space Perception and Tourism Image in Nanxun Ancient Town Based on Social Media Data Visualization"* **Buildings** ***15***(9): 1465. DOI: 10.3390/buildings15091465.

[11] W. Lan, J. Li, J. Wang, Y. Wang, and Z. Lei, (2025) *"Cultural Diversity Conservation in Historic Districts via Spatial-Gene Perspectives: The Small Wild Goose Pagoda District, Xi'an"* **Sustainability** ***17***(5): 2189. DOI: 10.3390/su17052189.

[12] M. Hu, J. Suh, and C. Pedro, (2023) *"An Integrated Framework for Preservation of Hawaii Indigenous Culture: Learning from Vernacular Knowledge"* **Buildings** ***13***(5): 1190. DOI: 10.3390/buildings13051190.

[13] Y. Guo, Z. Li, and X. Chen, (2025) *"Sustainable Disorder: The Hybrid Logic of 'Sense of Place' Construction in Tourist Spaces—A Case Study of Harbin Morning Market"* **Sustainability** ***17***(21): 9675. DOI: 10.3390/su17219675.

[14] G. Alkhateeb, J. Storie, and M. Külvik, (2024) *"Post-Conflict Urban Landscape Storytelling: Two Approaches to Contemporary Virtual Visualisation of Oral Narratives"* **Land** ***13***(4): 406. DOI: 10.3390/land13040406.

[15] L. Chen, Y. Song, X. Niu, X. Luan, L. Yang, and S. Qin, (2025) *"Epitome of the Region—Regional Nostalgia Design Based on Digital Twins"* **Behavioral Sciences** ***15***(1): 12. DOI: 10.3390/bs15010012.

[16] Q. Li, Q. Mai, M. Wang, and M. Ma, (2024) *"Chinese dialect speech recognition: a comprehensive survey"* **Artificial Intelligence Review** ***57***(2): 25. DOI: 10.1007/s10462-023-10668-0.

[17] L. Wang and K. King, (2024) *"Language ideologies, language policies, and shifting regional dialect proficiencies in three Chinese cities"* **Journal of Multilingual and Multicultural Development** ***45***(6): 2166–2182. DOI: 10.1080/01434632.2022.2044339.

[18] Y. Duan, M. Chen, Y. Liu, Y. Wang, and L. Zhang, (2025) *"Research on the Cultural Landscape Features and Regional Variations of Traditional Villages and Dwellings in Multicultural Blending Areas: A Case Study of the Jiangxi-Anhui Junction Region"* **Applied Sciences** ***15***(4): 2185. DOI: 10.3390/app15042185.

[19] H. Zhang, M. F. Seilhamer, and Y. L. Cheung, (2023) *"Identity construction on shop signs in Singapore's Chinatown: a study of linguistic choices by Chinese Singaporeans and New Chinese immigrants"* **International Multilingual Research Journal** ***17***(1): 15–32. DOI: 10.1080/19313152.2022.2080445.

[20] Y. Kuang, F. Zheng, C. Lin, and Y. Hu, (2025) *"Research on Chinese Traditional Architectural Culture and Inheritance Strategy: A Case Study of the Goulou Cluster of Yue Dialects in Guangxi"* **Buildings** ***15***(3): 489. DOI: 10.3390/buildings15030489.

[21] A. J. Privitera, S. H. S. Ng, A. P.-H. Kong, and B. S. Weekes, (2024) *"AI and Aphasia in the Digital Age: A Critical Review"* **Brain Sciences *14***(4): 383. DOI: 10.3390/brainsci14040383.

[22] J. Qian et al., (2024) *"Quantifying Urban Linguistic Diversity Related to Rainfall and Flood across China with Social Media Data"* **ISPRS International Journal of Geo-Information *13***(3): 92. DOI: 10.3390/ijgi13030092.

[23] A. R. Szromek and M. Bugdol, (2024) *"Sharing Heritage through Open Innovation—An Attempt to Apply the Concept of Open Innovation in Heritage Education and the Reconstruction of Cultural Identity"* **Heritage 7**(1): 193–205. DOI: 10.3390/heritage7010010.

[24] R. Pizarro Contreras, Z. Zhao, and G. Zhang, (2025) *"The Reproduction Without Alterity of AI: LatamGPT as a Form of Dissent and Technological Reappropriation"* **NanoEthics *19***(3): 16. DOI: 10.1007/s11569-025-00480-1.

[25] J. Li, M. He, Z. Yang, and K. W. M. Siu, (2025) *"Anthropological Insights into Emotion Semantics in Intangible Cultural Heritage Museums: A Case Study of Eastern Sichuan, China"* **Electronics *14***(5): 891. DOI: 10.3390/electronics14050891.

[26] *Chinese Dialect Speech-to-English Dataset*. Kaggle Dataset. Accessed Nov. 18, 2025. 2025. eprint: https://www.kaggle.com/datasets/zyan1999/chinese-dialect-speech-to-english-dataset.

[27] X. Yue, L. Miao, and J. Ding, (2025) *"Research on Wu Dialect Recognition and Regional Variations Based on Deep Learning"* **Applied Sciences *15***(18): 10227. DOI: 10.3390/app151810227.