Artificial Intelligence Review (2024) 57:25
https://doi.org/10.1007/s10462-023-10668-0

Chinese dialect speech recognition: a comprehensive survey

Qiang Li[1] · Qianyu Mai[1] · Mandou Wang[1] · Mingjuan Ma[2]

Published online: 31 January 2024

Abstract

As a multi-ethnic country with a large population, China is endowed with diverse dialects, which brings considerable challenges to speech recognition work. In fact, due to geographical location, population migration, and other factors, the research progress and practical application of Chinese dialect speech recognition are currently at different stages. Therefore, exploring the significant regional heterogeneities in specific recognition approaches and effects, dialect corpus, and other resources is of vital importance for Chinese speech recognition work. Based on this, we first start with the regional classification of dialects and analyze the pivotal acoustic characteristics of dialects, including specific vowels and tones patterns. Secondly, we comprehensively summarize the existing dialect phonetic corpus in China, which is of some assistance in exploring the general construction methods of dialect phonetic corpus. Moreover, we expound on the general process of dialect recognition. Several critical dialect recognition approaches are summarized and introduced in detail, especially the hybrid method of Artificial Neural Network (ANN) combined with the Hidden Markov Model(HMM), as well as the End-to-End (E2E). Thirdly, through the in-depth comparison of their principles, merits, disadvantages, and recognition performance for different dialects, the development trends and challenges in dialect recognition in the future are pointed out. Finally, some application examples of dialect speech recognition are collected and discussed.

Keywords Chinese dialect · Dialect corpus · Dialectal acoustic modeling · Automatic speech recognition · Deep neural network · End-to-end

✉ Qianyu Mai
nmu.csq@gmail.com

Qiang Li
dr.qiangli@gmail.com

Mandou Wang
nxmd.wang@gmail.com

Mingjuan Ma
mamingjuan2009@163.com

[1] School of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China

[2] School of Economics, North Minzu University, Yinchaun 750021, China

1 Introduction

During the past decade, with the popularity of smart devices and the rapid development of deep learning, Mandarin speech recognition (MSR) systems have been widely applied in all aspects of work and life. As the basic component of the modern Chinese language, mandarin has the largest number of speakers. Together with the standard Chinese characters, it constitutes the common Chinese language (Xie 2011). However, in the reality scene, the speakers have acoustic features with dialect characteristics due to the influence of factors such as regional customs, migration, educational background, cultural heritage, etc. The dialect acoustic features are often difficult to effectively acquire by the MSR system, which greatly affects the recognition effect and wider application. The key issue behind this phenomenon, dialect speech recognition (DSR), has attracted extensive attention in academia and engineering.

As a variant of Chinese, the Chinese dialect has a complete pronunciation system and strong local color (Li 2018). In most cases, dialects are often used in specific areas and, together with Mandarin constitute modern Chinese (Shao and Ma 2020). Compared with MSR, Chinese DSR refers to the feature processing of the speaker's dialect speech sequences, further acquiring the speech sequences with dialect characteristics, and finally combining the dialect dictionary, language model (LM), etc., to return the text content corresponding to the dialect (Li et al. 2006). At present, according to statistics, the number of recorded dialects in China is more than 100 (Fan and Xiao 2022). Chinese dialects can be roughly divided into seven categories based on tone value, loudness, the composition of initials and finals, and geographical locations, namely Northern Mandarin, Wu dialect, Gan dialect, Hunan dialect, Fujian dialect, Cantonese dialect, and Hakka dialect (Social Sciences 2012; Yuan 1960). However, the research progress of DSR in different regions is distinct, such as specific recognition approaches, speech corpus, and so on.

The construction of a dialect corpus is the crucial basis for dialect recognition (Bolia et al. 2000), but at present most corpora are dominated by Mandarin, and there are relatively few specific dialect corpora. So this paper organizes the literature on Chinese DSR and summarizes the dialect corpora constructed in recent years. Starting from the dialect division, we analyze the basic composition of the dialect corpus concretely. In addition, a general construction method of the specific dialect corpus is proposed by referring to the construction method of the general speech corpus and the components of the dialect corpus. It includes four steps designing the dialect text corpus, recording audio, dialect speech annotation, and data storage.

Similar to the process of MSR, Chinese DSR is inseparable from dialect speech signal processing, feature extraction, and recognition model building (Ma et al. 2006). Dialect signal processing removes the ambient noise and useless signals from the input dialect speech. The acoustic feature extraction further processes the frequency domain information of the speech signal, which aims to use feature vectors to represent the key information. Because of the complexity of Chinese dialect speech, its acoustic characteristics easily affect the recognition result. The tonal value of the dialect, the number of initials, vowels, voiced and unvoiced sounds jointly determine the acoustic characteristics of a dialect. Therefore, the study of the acoustic characteristics of a dialect is one of the key tasks in the process of dialect recognition. To obtain dialect speech signals that are closer to human hearing, most current methods utilize cepstral features or spectral coefficients to represent speech signal features according to the acoustic characteristics of the dialect (Guntur et al. 2022). The composition of the dialect recognition model adds a dialect dictionary based on the acoustic model (AM) and the LM (Ali et al. 2022). The phonetic feature sequences are transcribed with a dialect dictionary when dialect text is output.

As a result of the limited ability of computer data processing in the early days and the lack of Chinese dialect data, DSR models are mainly trained to recognize small isolated words through template matching methods and HMM (Zissman et al. 1996). Later, the computing power of the graphics processing unit increased significantly, and the scale of specific dialect corpus gradually expanded (Malmasi et al. 2015). Therefore, Chinese DSR approaches based on deep neural network (DNN) are widely used, especially for continuous dialect recognition with good performance (Wan et al. 2022). In recent years, E2E has been progressively applied in the Chinese DSR and achieved remarkable results. As a research hotspot in the current dialect recognition field, this method can map the input dialect speech sequence to the text sequence, optimize the overall components of the dialect recognition model, and effectively improve the dialect recognition effect.

The rest of the paper is organized as follows. Sect. 2 explains the literature sources of Chinese DSR and summarizes the research focus of scholars. In Sect. 3, we introduce the pronunciation characteristics of Chinese dialects, enumerate the existing corpora for Chinese dialect recognition tasks, and summarize the general construction process of dialect corpus. Sect. 4 explains the overall process of Chinese DSR, specifically describes the key components, and compares related recognition techniques. Sect. 5 presents a detailed illustration of the diverse challenges recently faced by Chinese DSR. Sect. 6 lists the application examples of Chinese dialect recognition. Sect. 7 specifically introduces some prospective research directions of Chinese DSR. Finally, Sect. 8 presents the discussion and conclusion of this survey.

2 Review methodology

For a long time, Chinese DSR has been an ignited topic that has attracted extensive attention from scholars at home and abroad. It is not only an effective way of communication in life but also diversifies the ways of human-computer interaction. Therefore, we extract information and summarize the current literature about Chinese DSR through multiple online academic research and discovery platforms, which include China National Knowledge Infrastructure (CNKI), Web of Science, IEEE Xplore, ACM Digital Library (ACM DL), and Springer Link. They contain high-quality research resources collected from full-text articles published by selected publishers. Firstly, we utilize “Chinese DSR” as the main retrieval object and retrieved extensively with words or strings in the title, keywords, and abstract. The search terms also contain “deep learning”, “neural networks”, “Chinese accent recognition”, “MSR”, “Chinese dialect identification”, “Chinese dialect corpus” and so on. In the subsequent search process, Boolean operators are used to refining the preliminary retrieval results, such as “Chinese DSR” AND ($\wedge$) “deep learning”. The time frame of the retrieved literature is set from January 2006 to September 2022. After removing some duplicate and irrelevant papers, 246 articles were obtained from conference proceedings, journals, and discussion articles provided by the above platforms. As shown in Fig. 1, the number of Chinese DSR research papers is displayed by time in different colored bars, with the corresponding specific values at the top. It also shows the number of published papers based on deep learning.

Based on the above work and the aggregated data, this review is mainly conducted from five aspects: (1) the classification of Chinese dialects according to the pronunciation

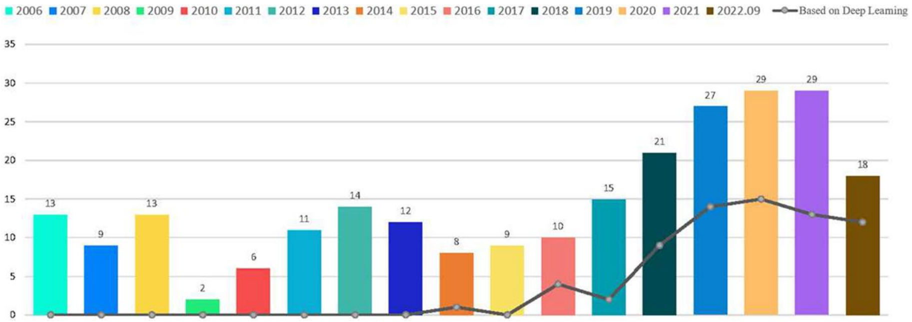


Fig. 1 The count of publication papers of Chinese DSR from 2006.01 to 2022.09

characteristics, (2) the aggregation of published Chinese dialect corpora, (3) the general construction procedure of dialect corpus, (4) the process of Chinese DSR and related models, especially the models based on deep learning, (5) the recent related applications of Chinese DSR.

3 Pronunciation characteristics of Chinese dialects and related speech corpora

The phonetic features of dialects are a necessary condition for distinguishing different dialects and a key factor for dialect recognition technology. Influenced by dialect geography, most Chinese DSR studies focus on region-specific phonetic characteristics by utilizing dialect maps or dialect geographic divisions. These features are crucial to the corresponding dialect corpus construction process. This chapter divides into three parts: (1) an overall analysis of the phonetic features of Chinese dialects, (2) a comprehensive survey of current Chinese dialects corpora, and (3) the general construction process of the Chinese dialect corpus.

3.1 Phonetic characteristics of Chinese dialects

As an internationally renowned scholar in linguistics, (Haugen 1966) pointed out that dialects are a set of specially written languages used for expression in daily life and interact with standard languages. Zaharia et al. (2021) argued that dialects and common languages have the same computer representation method in speech recognition. However, the pronunciation granularity of dialects is more delicate, such as accent, pitch value, etc. Shon et al. (2018) also emphasized that dialects are special cases of languages that could represent the similarities and differences between the two by computer quantitatively.

Since the 1920s, the study of Chinese dialectology has been established (Zhan 2000) and experienced different stages of development. Among them, the *Language Atlas of China* and its second edition (Social Sciences 2012) proposed two basic standards for classifying Chinese dialects, namely the evolution of ancient entering tone characters and ancient voiced initials. Their contributions laid the foundation for the research work on Chinese DSR and also played an active role in promoting the follow-up research work. In recent years, the Chinese have received more and more attention from all over the world.

Naturally, Chinese dialects have also become the focus of concern, for instance, the popularity of specific dialect tones and vocabulary. Sun (2020) proposed a prosodic-acoustic topic model, which verified that the acoustic features of Chinese dialects could be obtained through the unsupervised learning method. List (2015) pointed out that Chinese dialects can be regarded as a whole and promote each other with Mandarin. Gong et al. (2011) found that the frequency of dialect usage is related to an organization or community's acceptance and recognition process.

In addition, the issue of Chinese dialect classification has also attracted the attention of scholars. Since modern times, the earliest Chinese scholar to study dialect division has been (Zhang 1909). On this basis, the research point of view has gradually shifted from difference to considering commonality. Generally, the Chinese dialects can be divided into ten categories (Social Sciences 2012), i.e., Northern Mandarin dialect (Mandarin Dialect), Jin dialect, Wu dialect, Hui dialect, Pinghua dialect, Gan dialect, Xiang dialect, Min dialect, Cantonese and Hakka dialect. According to the scope of use, the predominant number of users, and their regions, the dialects in China are also classified into seven categories that include Northern Mandarin dialect (Mandarin Dialect), Northern Mandarin, Wu dialect, Gan dialect, Xiang dialect, Min dialect, Cantonese and Hakka dialect. Many scholars support this point of view (Yuan 1960; Yue 2003) and the seven categories of Chinese dialects are studied carefully in this paper.

Studies have shown that the phonetic characteristics of Chinese dialects are directly reflected through Chinese syllable structure and tone changes (Hu 2013). The syllables of Chinese characters are usually composed of initials and finals. The initial is the consonant part at the beginning of a Chinese character. According to whether the vocal cords are vibrating, the consonants can be divided into voiceless and voiced consonants (Chen et al. 2019). The finals utilize themselves as independent Chinese syllables. But in most cases, the finals are located behind the initials and together with the initials to form a compound syllable (Social Sciences 2012). Moreover, tone changes in dialect determine the pronunciation variations of the corresponding syllables. Although many Chinese dialect words have the same Pinyin syllables, different tones may finally express completely different meanings.

Ye (2011) investigated and counted the number of initial and final dialects in various regions of China. On this basis, we present the corresponding results of the above seven categories of Chinese dialects, as shown in Table 1. For instance, Luoyang is one of the representative areas of Northern Mandarin with 23 initials. Throughout the seven categories of dialects, the number of finals is generally more than the corresponding number of

Table 1 Representative areas of Chinese dialects and their number of initials and finals

Dialect type	Representative area	References	Number of initials	Number of finals
Northern Mandarin	Luoyang	Tang (2013)	23	37
Wu Dialect	Changzhou	Qian (2016)	28	44
Gan Dialect	Nanchang	Zhang (2007)	19	49
Xiang Dialect	Changsha	Tian (2009)	20	35
Min Dialect	Chaoyang	Zhang (1981)	18	90
Cantonese	Guangzhou	Rao (2007)	22	53
Hakka Dialect	Meixian	Liao (1994)	18	72

initials. The fundamental reason is that the number of finals contains more vowels. Shi (2006) studied 40 vowel patterns in Chinese dialects with the method of phonetic experiment. Chen et al. (2019) studied the phonetic characteristics of dialects and quantified the vowels. Their proposed quantization system can efficiently verify Mandarin dialect partitioning using statistical methods. Besides initials and finals, there are also tone differences among Chinese dialects. Zhao (1980) used the 5° notation method to record tone values, which required two or three digits between 1 and 5 to represent the value.

Although Mandarin is the standard language in China, people often use dialects and Mandarin together daily. Naturally, the pronunciation of most speakers will be mixed with dialect accents to some extent, such as Mandarin with Cantonese accent (Liu and Fung 2006). The accent is a pronunciation style that acquires intonation and phonology from dialects (Zhang et al. 2021). Surveys show that some people speak Mandarin with multiple dialect accents (Li 2012), such as Wu dialect, Cantonese, and so on, as a common challenge in ASR. Among them, their differences manifest in the tonal information of Chinese characters, a common challenge in ASR. Processing accented Chinese dialect speech sequences is crucial to improve the efficiency of the ASR system (Zheng et al. 2005), which is also one of the key issues of the Chinese dialect detection research (Wang et al. 2021).

To sum up, the phonetic characteristics of dialects are not only one of the primary factors in distinguishing different dialects but also the core element of dialect recognition technology (Kim et al. 2017). The phonetic features of dialects are obtained by studying the pronunciation characteristics of dialects, which usually include acoustic features, phonemic features, and prosodic features (Fukuda and Nitta 2004). Dialect acoustic features can directly obtain spectrograms or related acoustic parameters from speech signals, which are the most commonly used features in DSR. The phoneme feature is the smallest unit of acoustic feature, which divides the input audio into phonemes of each frame. The prosodic feature is the macroscopic expression of the tone value of the dialect (Etman and Louis 2015). Assuming that the prosodic structure of each dialect is distinct, combining fundamental frequency, energy, and other features could achieve the Chinese DSR model.

3.2 Chinese dialects speech corpora review

For most DSR tasks, high-quality speech data is not only of great significance to the training and optimization of the recognition system but also affects the deployment and promotion of related practical applications. Generally, a dialect speech corpus (Bu et al. 2017) is constructed from a series of audio files through segmentation, processing, and text annotation of the original speech. In the 1990 s, relevant Chinese research institutions and universities (Wang and Li 2003; Li et al. 2001) began to build Chinese speech corpora. The work of the Chinese dialects corpora has also been carried out in succession. As shown in Table 2, the related Chinese dialects speech corpora from 2000 to 2022 cover a relatively comprehensive type of Chinese dialects. The composition of the dialect corpus will also be introduced in detail below, including the number of speakers, recording types of equipment, sources of recorded texts, etc.

(1) THUYG-20 Corpus Rouzi et al. (2017): As a public Uyghur speech dataset, THUYG-20 corpus was co-published in the *Journal of Tsinghua University* by Aes-Karrouz et al. The corpus was recorded in 30 prefectures in Xinjiang, with a total audio duration of 21 h, covering more than 45,000 words in the vocabulary, including morphemes, syllables, and characters. The number of participants in the recording

Table 2 Composition and related information of the existing Chinese dialects corpora. S denotes sentences and UP presents unpublished

Dialects	Refences	Duration	Recording place	Participants	Text type
THUYG-20	Rouzi et al. (2017)	23.63 h	Quiet room	595	Reading text
Tibetan Ando	Han and Yu (2010)	48.75 h	Quiet room	UP	Reading text
GCDC	Xu et al. (2018)	131.5 h	Radio cut	UP	Conversational
Datong Dialect	Liu et al. (2020)	12.21 h	Quiet room	UP	Reading and conversation
Hunan Dialect	Wang et al. (2009)	240 s	Quiet room	12	Spoken language
MDCC	Yu et al. (2022)	73.6 h	Audio books	UP	Reading text
Uyghur	Qimike et al. (2015)	4466 s	Quiet room	UP	Conversational
Lanzhou Dialect	Yang et al. (2009)	408 min	Recording studio	4	Reading text
Chongqing Dialect	Zhang et al. (2018)	20 words	Quiet room	UP	Reading isolated word
Mixed Accent Mandarin	Yang and Hu (2021)	200 h	Telephone recording	6300	Spoken language
Fujian Accent	Pan et al. (2005)	UP	Quiet room	23	Reading text
RASC863	Li et al. (2004)	8800 s	Quiet room	800	Reading text
Common Voice	Ardila et al. (2019)	2500 h	Quiet room	50,000	Reading text

is 348. Recorders utilized the sound card of an IBM Lenovo desktop computer and an external microphone to record and read literary works, news reports, and other materials in a quiet environment. The audio sampling rate is 16 kHz, and recorded in mono mode. Since the THUYG-20 corpus is mainly based on reading materials, the performance of isolated word recognition is impressive, but the continuous dialogue recognition effect is not ideal.

(2) Tibetan Ando Dialect Corpus Han and Yu (2010): Han Qinghua and others jointly constructed the Tibetan Ando corpus and it is currently serving the laboratory. The text content of this corpus consists of ten monosyllabic words in Ando dialect, namely "I", "warm", "person", "head", "dog", "walk", "water", "beauty", "two", and "body". In a quiet environment, 40 volunteers participated in the recording, including 20 males and 20 females. They read the Ando dialect words 12 times in a turn. The audio sampling rate is 16 kHz, and recorded in binaural mode. The Ando dialect corpus is suitable for dialect isolated word recognition. However, the number vocabulary is small, and the acoustic characteristics are insufficient, which would affect the robustness of the recognition model to a certain extent.

(3) GCDC Corpus Xu et al. (2018): The GCDC corpus was co-published in an international academic conference by XU Fang et al. The dialect in the GCDC corpus belongs to the Gan dialect. The total duration of the corpus is 131.5 h, of which 69 h are Gan dialect pronunciation and 62.5 h are Mandarin. The text content of the recorded audio comprises six genres: news reports, novels, announcements, poems, letters, and prose, with 310 documents. The corpus includes 19 sub-regions of Gan dialects, but their paper does not mention specific recording information, such as the audio sampling rate.

(4) Datong Dialect Corpus Liu et al. (2020): The Datong dialect corpus was published in the *Journal of North University of China (Nature Edition)* by Liu et al. This corpus collects the Datong dialect of Shanxi Province, with a total audio duration of 12 h, 21 min, and 13 s, including reading and daily spoken texts. There are a total of 8894 pieces of audio data, and the ratio of the training set and test set is 7:3. The recorded text in the Datong dialect corpus was relatively balanced. It can reflect the acoustic characteristics of the Datong dialect. It is suitable for training and testing speech recognition in small dialects, but the number and gender of recorders are not quantified and analyzed.

(5) Uyghur Corpus Qimike et al. (2015): Uyghur corpus was designed and constructed by Mick Batsy et al. The corpus contains 4466 phonetic sentences, the ratio of female to male participants is 1:1, and the ratio of the test set to the training set is 2:8. In the process of phonetic annotation, the research team adopted the Unicode encoding method to convert the 26 phonemes of Uyghur into the corresponding English representation.

(6) Lanzhou Dialect Corpus Yang et al. (2009): Lanzhou dialect corpus was jointly designed and constructed by Yang Hongwu et al. The corpus belongs to Lan–Yin Mandarin, and the audio duration is 408 min. A total of four volunteers participated in the recording, two males and two females. The reading text is designed according to the characteristics of the Lanzhou dialect, including syllables, two-character words, and sentences. High-fidelity microphones and external sound cards recorded all audio in a professional recording studio. The audio sampling rate is 44.1 kHz and recorded in mono mode. In dialect speech annotation, it used the endpoint detection method to segment the syllables of recorded audio. The text content of this corpus is mainly

reading materials, but the geographical scope of the collected dialect speech is small, and the duration is relatively limited.

(7) Chongqing Dialect Corpus Zhang et al. (2018): Chongqing dialect corpus was designed and established by Zhang Ce et al. They selected 20 colloquial Chongqing words and recorded them in monophonic. The audio sampling rate is 16 kHz. However, the number of participants and the speech duration were not specified.

(8) Hunan Dialect Corpus Wang et al. (2009): Hunan dialect corpus was designed and built by Wang Qixue et al. This corpus consists of Mandarin and three regional dialects: Changsha, Shaoyang, and Hengyang. The number of volunteers participating in the recording is 12. In a quiet recording environment, they read aloud 240 sentences from the text content of the corpus. But the corpus does not specify the audio duration and the recording device's parameters. Although the Hunan dialect corpus contains dialect sounds of the three regions, which can excavate the hidden acoustic features of the Hunan dialect, the amount of data is comparatively limited.

(9) Mixed Accent Mandarin Corpus Yang and Hu (2021): The mixed accent Mandarin corpus is an open source recorded by DataHall. The full length of recorded audio is 200 h. A total of 6300 volunteers participated in the recording, with a male-to-female ratio of 1:1, covering 34 provinces such as Guangdong and Fujian. The speakers randomly selected accented Mandarin conversation topics and recorded them using Android microphones in a quiet environment. The audio sampling rate is 16 kHz. The speech data in this corpus contains multiple places and covers a wide range of accents, which is suitable for large-scale accent speech recognition systems.

(10) Fujian Accent Corpus Pan et al. (2005): Fujian accent corpus was jointly constructed by Pan Fuping et al. The corpus consists of 23 sets of data. Each set of audio content contains 260 personal names, but it does not introduce the specific recording time. A total of 23 volunteers participated in the recording, and the audio sampling rate was 8 kHz. Since the text form and content of the accent corpus are relatively simple, it is suitable for speech recognition in specific fields.

(11) RASC863 Corpus Li et al. (2004): RASC863 (Regional Accented Speech Corpus funded by National 863 Project) corpus was designed and built by the Institute of Linguistics, Chinese Academy of Social Sciences. The first batch of data consists of 4 regional accents from Shanghai, Guangzhou, Xiamen, and Chongqing, representing the dialect of Wu, Yue, Min, and Southwest Mandarin, respectively. It was completed and published in 2004. The number of volunteers who participated in the recording was 800, using USB sound cards and computer microphones. There were 200 people in each region, with a balanced ratio of men and women. The reading text includes two materials: reading and naturally spoken sentences. The second batch of the RASC863 corpus comprises 6 other regional accents: Changsha, Luoyang, Nanchang, Taiyuan, Nanchang, and Wenzhou. RASC863 contains rich dialect accent content, which can effectively support the construction and application of large-scale speech recognition systems.

(12) MDCC Corpus Yu et al. (2022): MDCC corpus is a multi-domain Cantonese dataset established by Yu et al. The total audio duration of the corpus is 73.6 h, of which the training set is 57.53 h. The ratio of male to female recording audio is approximately balanced, with 50.29% males and 49.71% females. The speech corpus is collected from audiobooks in Cantonese, covering a wide range of topics such as philosophy, education, culture, etc. The sampling rate is 16 kHz, and the duration of each utterance ranges from 0.22 to 15.00 s. Annotation was performed in two phases, i.e. after automatically annotating utterances using the Google Cloud Speech-to-Text API,

some native Cantonese speakers were hired to manually improve the quality of the annotated transcripts.

(13) Common Voice Corpus Ardila et al. (2019): The Common Voice corpus is a large-scale speech dataset designed and published by Ardila et al. The corpus makes use of the content of Wikipedia as text material, and more than 50,000 people participated in the reading and recording work. A total of 29 languages were recorded with a duration of 2500 h. It includes 12 h of Chinese Cantonese dialect audio with 288 participants. The audio sampling rate is 48 kHz. This corpus contains various types of languages and is suitable for medium and large-sized speech recognition systems.

With the improvement of storage device performance and the popularity of streaming media, voice-based materials emerge in an endless stream, such as audiobooks, dialect dramas, etc. These materials may provide more abundant resources for the construction of dialect corpus and further improve the generalization and overall performance of the recognition model. Each dialect contains a unique cultural value. Our country makes the national norms of transcribing audio marking to protect and inherit the dialect culture. It provides literal language translation, transliteration, and corpus attribute annotation for guiding transliteration corpora (Zhiyun 2015) and encourages the development of multiple types of dialect applications based on normative standards. The design and construction of dialect corpus is a necessary foundation work of DSR. In addition to the Chinese dialect corpora mentioned above, there are also large cross-language corpora. For example, Facebook recently released a large-scale open multilingual speech corpus VoxPopuli (Wang et al. 2021), which provides 400,000 h of unlabeled speech data in 23 languages. Naturally, large-scale multilingual and mixed dialect corpora will be one of the important development trends of the DSR corpus in the future.

3.3 General construction process of Chinese dialect corpus

The dialect corpus aims to provide reliable speech data for dialect recognition and other related tasks (Bouamor et al. 2018), such as annotated text transcripts and corresponding audio of dialect utterances. Constructing a corpus is significant and can promote the development and research of related disciplines. The construction process of open source corpus and related annotation criteria provides an essential basis for the construction of Chinese dialect corpus. For example, WSJCAMO (Robinson et al. 1995) uses the Wall Street Journal as the recorded text. For a single dialect corpus, this process typically requires four steps: selecting and preparing dialect text materials, recording audio utterances, annotating dialect speech, and data storage. Figure 2 shows the general construction process of the Chinese dialect corpus. What's more, it also includes three judgment criteria.

The first judgment is the principle of text design because the source of Chinese dialects is text reading and new media methods, such as telephone recordings, local films, and television works. Therefore, the division of recording files can be standardized by dialect text. The second judgment is the principle of recording, which makes a certain standard about the environment and equipment to reduce outdoor noise. The last principle is dialect audio annotation. The process of audio annotation is relatively complex because it not only needs to label the tone value of the Chinese dialect but also aligns the speech information with the text content. However, it is easy to lose some prosodic information in this process, so dialect phonetic annotation has attracted extensive attention.

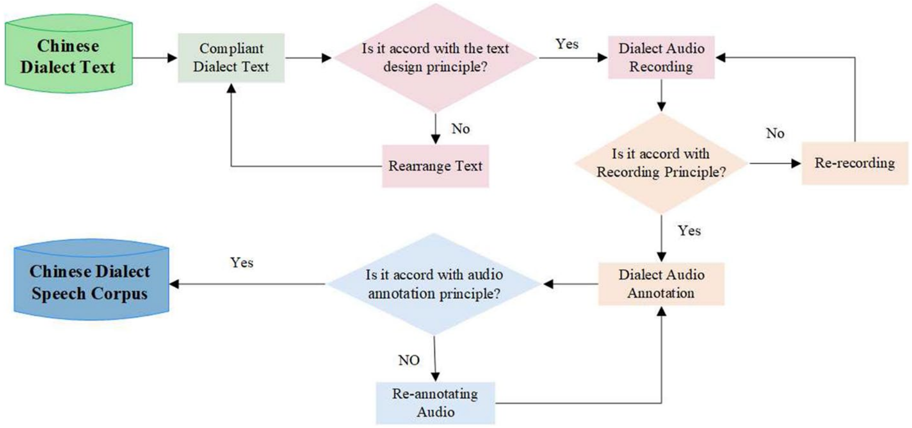

Chinese Dialect Text
Compliant Dialect Text
Is it accord with the text design principle?
Yes
Dialect Audio Recording
No
Rearrange Text
Is it accord with Recording Principle?
No
Re-recording
Yes
Chinese Dialect Speech Corpus
Yes
Is it accord with audio annotation principle?
Dialect Audio Annotation
NO
Re-annotating Audio

Fig. 2 General construction process of Chinese dialect corpus.

Dialect text design is the basis of dialects corpora, which can highlight the phonetic characteristics of dialects and reflect natural language phenomenon to the greatest extent (Haugen 1966). The objects of dialect text design can generally be divided into two categories: reading text and spoken language (Yang et al. 2017). Various text corpora are extracted in designated proportions to reflect the language and pronunciation characteristics of the dialect as much as possible. The reading texts have different genres: news reports, local chronicles, and literary books. The speakers need to directly read and record in a quiet environment. Since the scholar has designed reading content in advance, the speech annotation task can be completed quickly with the help of automated tools, such as the iFLYTEK or Baidu Speech-to-Text API. Manual proofreading by local people proficient in the dialect is necessary, mainly to modify the wrong or missing annotation results. Reading some given texts through the dialect will weaken the performance of dialect pronunciation characteristics in daily communication to a certain extent.

The other is the spoken language corpus, which represents the content where speakers randomly select topics for communication and involves many dialect spoken words. At the same time, characteristics such as regional word usage habits, speech emotion, and intonation can also be recorded naturally. However, due to the flexibility of spoken language, the difficulty of annotating spoken language is greater than that of reading text, and the collection process is likely to involve the speaker's privacy. Therefore, most Chinese dialect recording materials mainly use reading texts. For example, (Nurmemet and Wushour 2013) chose reading text to construct the Uyghur corpus. In addition, several dialect corpora use the speech data provided by local film and television works or local radio programs, which is another convenient and quick method to construct a dialect corpus. For instance, Fu et al. (2020) selected a series of film and television works as the text content of the Chengdu dialect.

Dialect recording quality is also one of the critical elements in corpus construction. Using professional microphones to directly collect dialect sounds in a quiet indoor environment can reduce noise interference and avoid problems such as distortion. The high sample rate, along with a high bit depth and multiple channels, will ensure the better audio quality of speech recordings, and vice versa (Rabiner and Juang 1993). The 16 kHz sampling rate with 16-bit depth is a common choice. For example, Li and Zhao (2017) recorded the

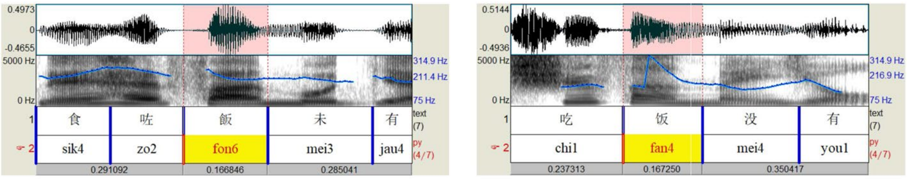
Hengyang dialect corpus in such condition, the RASC863 corpus, etc. There is a complete range of multimedia devices, and smartphones or tablets are also acceptable tools for speech recording and collection. The Datatang used mobile phones to record part of dialect speech data.

In general, dialect speech annotation information includes the speaker information, recording characteristics, speech data-related information (e.g., the duration), corresponding text information, speech annotation details (e.g., the tone or accent labels) and so on (Xu et al. 2017; Zhou et al. 2010). Although the recording personnel and equipment are usually selected and determined simultaneously as the text content design, the actual situation may affect the final result due to some accidents, such as the recording personnel getting sick or the temporary failure of the equipment. Praat software is widely used to label the tone value and Pinyin of the text concerning the dialect dictionary in speech annotation.

As shown in Fig. 3, there is an example of using Praat to label Cantonese and Mandarin speech. We used speech synthesis technology to synthesize the sentence "Have you eaten yet" in Mandarin and Cantonese, respectively, then we manually annotated the sound boundaries and Pinyin on a character level. It can directly show the difference in tone, resonance peak, and wavelength of the same sentence in dialect and Mandarin. Figure 3a presents the labeling result of Cantonese. The first line is characters, and the second line gives the phonetic syllables corresponding to dialects Pinyin refers to the local dialect dictionary. Furthermore, Fig. 3b shows the Mandarin annotation map. The first line contains the transcription using Chinese characters, and the second is the canonical syllable in Pinyin. Moreover, it also has detailed annotations that include identifying speech start position, speech segment, and syllable position. The more comprehensive and accurate the annotation information, the better dialect recognition effect. The speech data is frequently stored in WAV format and named with a digital serial number to facilitate subsequent updates.

4 Chinese DSR system

Chinese DSR systems are designed to recognize and transcribe dialects. They thus can be considered as a specialized type of general Chinese speech recognition system (e.g., the MSR system) (Malik et al. 2021). These systems utilize general speech recognition approaches but focus more on the dialect's unique phonetic features, tonal characteristics, vocabularies, and other specific factors. By incorporating these dialect-specific features into the recognition process, Chinese DSR systems efficiently accomplish recognition tasks, such as Cantonese or Datong dialect recognition systems (Yu et al. 2022; Liu et al. 2020). Therefore, the Chinese DSR system can be regarded as a subset of the Chinese



(a) Cantonese annotation map.

(b) Mandarin annotation map.

Fig. 3 The Cantonese and Mandarin speech are annotated by Praat.

speech recognition system. This section starts with dialect signal pre-processing and analyzes relevant feature extraction techniques and critical modules. Then we further sort out and summarize the Chinese DSR approaches, including state-of-art E2E methods.

4.1 Signal pre-processing

Speech signal preprocessing is performed before extracting the Chinese dialect feature. Chinese dialect signal preprocessing is similar to a general speech signal. Its goal is to retain the main content of the speech signal to the greatest extent. However, the dialect speech signal is a one-dimensional analog signal with continuous changes in time and dimension, and computers can only process the digital signal. Therefore the speech signal must make an analog-to-digital conversion (ADC), namely digital processing, before signal preprocessing.

The speech signal digital processing (Keerio et al. 2009) follows in these ordered steps. In the first step, sampling uses the sampling theorem to discretize the continuous dialect speech signal (Florescu and Bhandari 2022). It expresses the dialect signal in function form and values the signal within a certain time interval. After that, quantifying is the discretization of waveform amplitude of speech signal (Bhatia et al. 2023). It divides the dialect signal amplitude value into several intervals and represents the samples whose amplitude value is in this interval with the same value. Lastly, the quantized signal is encoded by different encoding methods, such as Manchester Encoding (Badea et al. 2019). The encoding can be transmitted over a digital communication channel or stored. After the signal digital processing, dialect signals as the input of preprocessing. As shown in Fig. 4, it presents the flow chart of the dialect speech digital processing and preprocessing.

The purpose of preprocessing dialect speech is mainly to clean signals from environmental noises and prepare them for feature extraction. It includes the following steps (Labied et al. 2022): Voice Activity Detection (VAD), pre-emphasis, framing, and windowing. VAD can not only be used to mark the starting point and end point of dialect speech signals but also to distinguish between dialect noise and sound areas. In addition, VAD technology can deal with conversion problems effectively (Guntur et al. 2022) when processing speech signals of Chinese dialects. Therefore, Liu et al. (2021) presented a dialect and Mandarin interaction system with the VAD method to eliminate the speech problem of both.

Dialect pre-emphasis can enhance the high-frequency resolution of the speech signal and increase signal-to-noise. Because the high-frequency part of dialect speech contains

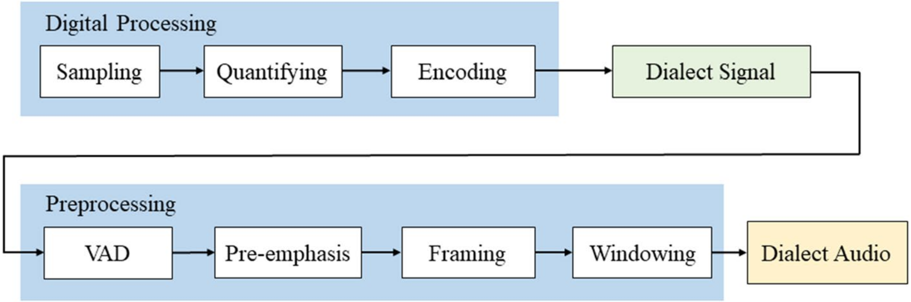


Fig. 4 The basic process of Chinese dialect speech signal digital processing and preprocessing.

rich information, the pre-emphasis technology enhances the high-frequency information of dialect. Wu and Liu (2013) used pre-emphasis to reduce the Uyghur high-frequency signal by 6 db to improve the audio distribution. However, the dialect speech signal is not stable. Therefore, framing divides the continuous speech signal into short-term and equal-length segments. It aims to give short-time stationarity to the speech signal (Labied et al. 2022). Generally, in the framing process, add windowing operations to effectively avoid signal interruption due to spectral distortion. The windowing operation ensures the integrity of the audio information. It includes several window functions. For example, Nisar and Tariq (2018) used Hamming window to process the low resource language signal. What's more, the signal preprocessing could provide reliable and effective data support for dialect recognition (Akçay and Oğuz 2020).

4.2 Feature extraction

Dialect feature extraction aims to obtain the feature vector sequence by processing the input speech audio. Thus the practical information of dialect speech (Prabakaran and Shyamala 2019) can be preserved. Acoustic features for dialect recognition mostly use auditory-based signal representation methods (Ramırez et al. 2004). There are three common methods to extract Chinese dialect acoustic features, which are the Mel Frequency Cepstral Coefficient (MFCC), the Linear Predictive Cepstral Coefficient (LPCC), and Perceptual Linear Predictive Coefficients (PLP). The feature extraction process is shown in Fig. 5.

MFCC is one of the most popular acoustic feature extraction techniques. And it can simulate the audio perception of the human ear at different frequencies and map the linear spectrum to the Mel nonlinear spectrum (Logan 2000). Zhang et al. (2018) used MFCC to extract speech features of the Chongqing dialect. Li and Zhao (2017) showed the Hengyang isolated word dialect recognition system that used MFCC to extract features. PLP is a feature parameter based on an auditory model that matches the features of the human auditory system by modifying the spectral features (Hermansky 1990). For instance, (Xie et al. 2022) analyzed four low-resource Chinese dialects with PLP and pitched parameters to extract acoustic features. They concluded that PLP could be helpful for low-resource Chinese DSR.

LPCC is a representation of Linear Predictive Coding (LPC) in the cepstrum domain (Wong and Sridharan 2001). It uses the linear method to obtain acoustic information. Deqing (2010) proposed a specific Tibetan speech recognition system that discussed the experimental results of LPCC and MFCC. The system showed that both had a good performance for Tibetan acoustic features. Kethireddy et al. (2022) optimized the acoustic feature

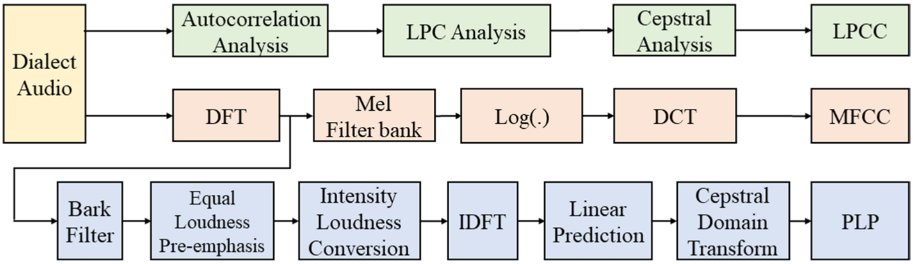


Fig. 5 The process of speech feature extraction in Chinese dialects.

extraction method. They proposed the frequency domain linear prediction method of cepstral coefficients that combined the characteristics of LPCC and PLP. And compared to the previous two methods, the suggested method performed better regarding recognition.

As shown in Table 3, there are merits and demerits about the three feature extraction methods. According to the above conclusion, MFCC is the most common method to extract dialect features. It could extract dialect speech signals in different frequency bands to maximize the inclusion of valid dialect speech information. Therefore, scholars further optimized the MFCC method. (Honnavalli and Shylaja 2021) construct continuous MFCC features for speech frames and adopt supervised learning techniques. (Dua 2023) combined MFCC and Constant Q Cepstral Coefficient (CQCC) to build an integrated feature extraction method, which achieves good recognition performance for low-resource speech.

4.3 Dialect dictionary, language model, acoustic model

Throughout the development and advancement of Chinese DSR, numerous systems have incorporated elements and frameworks from standard speech recognition models, encompassing components such as dialect-specific dictionaries, LM, and AM. However, notable differences exist between various dialects and Mandarin, including phonetics, tones, and characters. It is crucial to optimize these systems by taking into account the unique characteristics of each dialect to enhance the accuracy and efficiency of Chinese DSR. For instance, improving the AM is often a valid and popular strategy. Concurrently, optimizing the dictionary and LM by incorporating and adapting to the specific linguistic differences of each dialect can also improve the overall performance of the recognition system.

The primary role of dialect dictionary is to realize the mapping between phonemes and glyphs, and sometimes it is also called a pronunciation dictionary (Ali et al. 2022). The dialect dictionary could provide rhythm and tone references for dialect text transcription. For example, Li et al. (2004) labeled the tones of four regional dialects by referring to the Chinese dialect dictionary. Because most Chinese dialect texts usually consist of one or more Chinese characters, Chinese expressions use punctuation to break sentences. Unlike English sentences, there is no blank in dialect sentences, so they need dialect dictionaries to cut the text when training the LM (Lai 2022). On the other hand, the seven Chinese dialect partitions are recognized in the partition plan. According to the existing dialect corpora research, some regions of dialects still lack specific explanations. Therefore, different dialects need to build corresponding dialect dictionaries. With the deepening of Chinese dialects, the tone, rhythm, and text information of dialects are becoming more and more abundant. At present, in order to facilitate finding and using relevant information, some online Chinese dialect pronunciation dictionaries have been developed.

The LM can be distributed by calculating the probability of predicting a Chinese character or word and determining whether the text sequence obtained is an average sentence (Zhang et al. 2022). In short, LM uses statistical methods or neural networks to detect the connection between each word in the sentence. LM in the Chinese DSR system usually consists of a lexicon, a search space, and a search technology (Gu et al. 2022). The N-gram model (Gu and Xia 2008) is a commonly used method, which uses the first *n-1* Chinese character or word to predict the possibility of the *nth* Chinese character or word. It iterates to search for the correct word sequence, which is intuitive and easy to operate. In general, LM and dialect pronunciation dictionaries jointly act in Chinese DSR to encode the sequence output from AM training. This coding could improve the accuracy of Chinese DSR and reduce the burden on an AM to a certain extent.

Table 3 The merits and demerits of the three feature extraction methods in Chinese DSR

Method	Dialect example	Merits	Demerits
MFCC	Sichuan Dialect (Shi and Huang 2016) Chongqing Dialect (Zhang et al. 2018) Hengyang Dialect (Li and Zhao 2017)	MFCC close to the human auditory system. The correlation between the coefficients is small, and it could extract low frequency features	Less robust against noise. Due to linear calculation, it is easy to delete high-frequency information
LPCC	Tibetan (Deqing 2010)	LPCC is good for describe vowels and simple calculation	Easy to overlook key dialect features
PLP	Northern Dialect (Xie et al. 2022)	Processing the short spectrum of speech	The calculation is more complicated

HMM is a commonly used classical method to construct the AM (Shivaprasad and Sadanandam 2021). Dialect recognition leveraging HMM can model the smallest unit of a dialect and generate a corresponding observation state to calculate the phonetic unit set of the hidden state repeatedly and iteratively. Qimike et al. (2015) employed Gaussian Mixture Model (GMM)-HMM to build an AM for the Uyghur. However, due to the limitation of GMM, such as lack of temporal modeling, independence assumption, scalability issues, etc., neural networks were introduced to replace GMM and enhanced dialect recognition performance.

As research advances, the combination of neural networks with HMM in Chinese DSR has encountered difficulties. One of the challenges is that the forced alignment of speech data and independent modules makes global optimization difficult. Consequently, an increasing number of researchers have shifted their focus to E2E, which can overcome these issues and enhance the accuracy of dialect recognition models to a certain degree. In contrast to hybrid dialect recognition models based on HMM, E2E models employ a unified encoder-decoder framework that directly maps speech sequences to text sequences. This approach eliminates the need for constructing LM, thus reducing the overall complexity of the model. The following section will delve into the details of various Chinese DSR methodologies.

4.4 Chinese DSR approaches

Due to the richness of daily dialect speakers and the diversity of self-media, the continuous expansion of the scale of various dialect corpora provides the necessary data support for training dialect-specific acoustic models. Different approaches have been employed to process dialect acoustic features and generate corresponding text output. In this section, we will primarily discuss and analyze the evolution of Chinese DSR approaches. Dynamic Time Warping (DTW), HMM, and GMM are classified as traditional recognition approaches. This categorization primarily follows three factors: first, these methods require a relatively small amount of data for training; second, their computational complexities dictate that the training process typically does not necessitate the support of substantial computational power, e.g., GPUs or supercomputers; and finally, their actual recognition performance in common scenarios.

4.4.1 The traditional approaches

DTW, HMM, and GMM have played significant roles in the development of Chinese DSR, which employ dialect speech templates or statistical models to compare input dialect words in the training stage, allowing the system to learn the probability distribution of dialect words effectively. However, their performance and capabilities differ from those of DNN-based approaches in several ways, e.g., the ability to model complex relationships and capture higher-level features in dialect speech signals.

(1) Dynamic time warping

DTW is particularly useful for template matching (Berndt and Clifford 1994), which detects similar shapes at different locations by "elastic" transformations of the time series (Senin 2008), aims to find the lowest distance path. It has been widely used for Chinese dialect isolated word recognition. The Chinese DSR based on DTW extracts dialect acoustic features to generate training vectors firstly. Then it matches the test words with the given templates. The final output text is the shortest alignment distance between the input Chinese dialect and the template. The distance is calculated as shown in Eq. (1). (i, j) presents a pair of time series and $D(i, j)$ is used to measure distance between features of dialect input sample and saved template.

$$D(i,j) = \min \left[D(i-1,j-1), D(i-1,j), D(i,j-1) \right] + d(i,j) \tag{1}$$

Yao et al. (2009) proposed a DTW model for Tibetan isolated word recognition. They used MFCC as feature extraction due to the characteristics of Tibetan multi-syllabic words. Wu (2012) presented Uyghur isolated word recognition system that employed DTW as the training model. The system could recognize ten common words with a high accuracy rate. However, the training time of DTW will increase as the amount of data increases, which is arduous to satisfy the demand for continuous DSR. At present, the DTW approach is suitable for limited resources of speech recognition or isolated word recognition.

(2) Hidden Markov Model

For a long time, the HMM has been considered the mainstream for Chinese DSR. Because it has an accurate mathematical model that can calculate the training model parameter from the speech data. Additionally, the size and structure of the model can be adapted to specific speech, providing flexibility and ease of use. In general, HMM is a probability model that includes observable and hidden states, describing the next state's stochastic generation process using Markov chains (Juang and Rabiner 1991). The HMM-based Chinese DSR constructs a probabilistic distribution model from the input dialect speech sample to calculate the probability values of observed sequences. When the observed probability is maximum, it is the classification result. Finally, the Chinese dialect dictionary codes the text sequences corresponding to the dialect phonetics.

Han and Yu (2010) constructed a Tibetan DSR with HMM. They used the MFCC to extract the Ando feature, then trained in HMM model to form feature templates. Finally, the model could obtain the most significant similarity text by comparing the above feature template library. Zhang et al. (2018) presented a Chongqing DSR system based on HMM. Meanwhile, according to the pronunciation characteristics, they also constructed a Chongqing pronunciation dictionary to improve recognition accuracy. Chinese DSR based on HMM could use known states to calculate the probability of hidden states. However, the stochastic process is constrained by the time of the first order and can only retain the state information of speech sequence at the current moment and the next moment.

(3) Gaussian Mixture Model

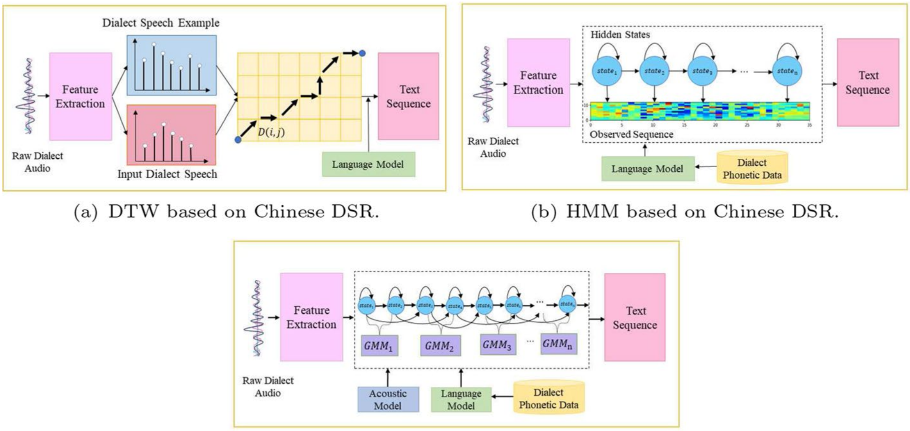
The GMM makes Gaussian distribution as a parameter model, using an expectation-maximization algorithm for parameter estimation. When the probability value of a class is maximum, it is the final result. In general, GMM and HMM are combined to construct a Chinese DSR system (Reynolds 2009), the linear combination of multiple GMM, and estimate the probability density distribution of the sample data. The Chinese DSR model of GMM/HMM includes three main components: AM, LM, and dialect dictionary. The AM inputs dialect feature vectors probabilistically and GMM/HMM calculates the probability distribution between dialect features and phonemes. Moreover, LM models the structure, grammar, and so on based on the existing dialect speech data. Finally, coding the output of the AM with the dialect dictionary obtained the dialect text information.

Wang (2001) presented an Uyghur recognition system with a large vocabulary based on GMM-HMM, in which the feature matrix is improved using the central distance normal distribution. This method could obtain more acoustic features of Uyghur. The Chinese DSR based on HMM-GMM performs well for a large vocabulary recognition system. Li and Meng (2012) built a Tibetan Lhasa recognition system based on GMM-HMM. They used the Tibetan phonemes and vowels as acoustic modeling units. The experiment result showed that this model has a better recognition accuracy for vowels. However, GMM struggle with modeling complex, nonlinear relationships and dynamic variations in dialect speech signals.

As shown the Fig. 6, there are Chinese DSR traditional approaches' frameworks. Due to the GMM/HMM hybrid model's fast training speed and the use of probability values to determine the corresponding word level, this approach is suitable for large-scale Chinese dialect vocabulary recognition when compared to Chinese DSR based on DTW. Despite its high recognition accuracy, the GMM-HMM model has some limitations, including its complex calculation process and difficulty acquiring additional dialect information. With the development of deep learning technology, an ANN replaced GMM to build the Chinese DSR system.

4.4.2 Deep neural network

DNN consists of input, hidden, and output layers, in which the output layer is computed by weighted summation with a nonlinear activation function (Pan et al. 2012). Because of their ability to process nonlinear data and capture high-level features in dialect speech signals, DNN-based methods have become a powerful alternative to traditional Chinese DSR approaches. Therefore, for constructing the continuous Chinese DSR system, DNN combined with HMM builds the training model, in which the DNN calculates the posterior probability of the HMM state. As shown in Fig. 7, there is the structure comparison of Chinese DSR based on DNN-HMM and GMM-HMM. For the observed sequence, the



(a) DTW based on Chinese DSR.

(b) HMM based on Chinese DSR.

(c) GMM/HMM based on Chinese DSR.

Fig. 6 Comparison of the Chinese DSR traditional approaches.

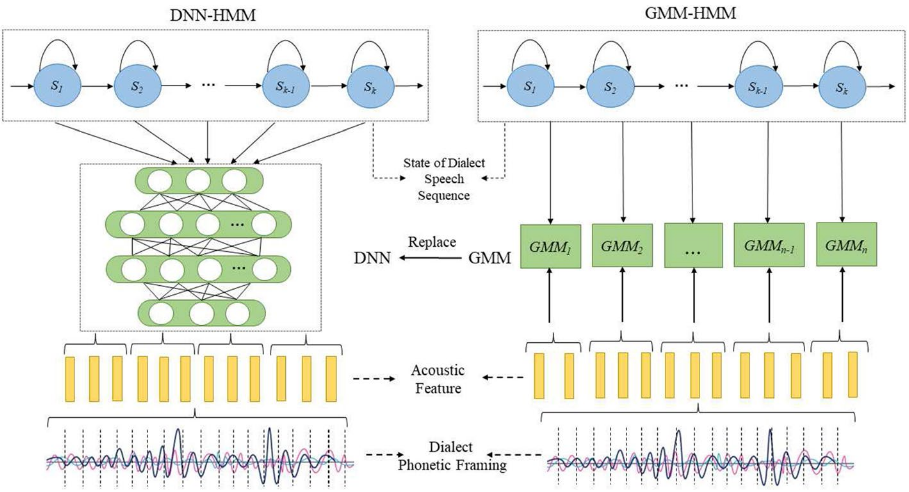


Fig. 7 The structure comparison of Chinese DSR based on DNN-HMM and GMM-HMM.

state value of GMM corresponds to the input sequence value one by one. The DNN and the input sequence are in a one-to-many mapping relationship. Some advantages of DNN-based methods over GMM-HMM include improved performance and better representation learning. DNN has shown superior performance in Chinese DSR tasks, outperforming traditional approaches like HMM and GMM. Moreover, DNNs can learn hierarchical representations of input data, allowing them to capture complex patterns and structures in dialect speech signals.

Tuerxun and Dai (2015) presented a continuous Uyghur speech recognition system evaluated against DNN-HMM, BN-GMM-HMM, and GMM-HMM in a series of trials. The experiment results showed that the performance of DNN-HMM was better than other models for a continuous Uyghur speech recognition system. Moreover, Qimike et al. (2015) also proposed an Uyghur speech recognition based on DNN-HMM. They extracted mono photons as acoustic features and used Kaldi as a testing platform. Shi and Huang (2016) introduced a Sichuan DSR system based on DNN-HMM. In order to improve the recognition accuracy of the Sichuan dialect, they constructed a Sichuan dialect pronunciation dictionary. However, DNN based approaches often require large amounts of data and substantial computational power for training, which is a drawback compared to traditional approaches. Another issue in training Chinese DSR models is the need to force the alignment of dialect sequences to generate reference state labels for each frame.

4.4.3 Convolutional neural network

Convolutional neural network (CNN) is one of the deep neural-network architectures. It can effectively process time series data and performs well in speech recognition. The main reason for the excellent effect is the network architecture of CNN, which includes the convolution and pooling layers (O'Shea and Nash 2015). For Chinese DSR based on CNN, convolution is applied over windows of acoustic frames to obtain more stable acoustic feature classes (Abdel-Hamid et al. 2014) when processing Chinese dialectal speech data. In training the model, the weights are learned and shared with other network layers to improve the robustness of the model. Compared with the dialect AM based on DNN, CNN uses local relevance to overcome the instability of dialect signals and obtain more dialect information.

Ai and Fei (2019) proposed a Guizhou dialect recognition model based on CNN, and they built a Guizhou dialect corpus containing six dialect areas. In order to determine the categories of dialects, they added a classified storage layer and competitive output layer to the network structure of CNN. The experimental results indicated that CNN has good robustness and generalization ability for the Guizhou dialect. Iminjan et al. (2021) constructed an Uyghur speech recognition system based on CNN. This model combines local connection, weight sharing, and pooling to minimize model training. All in all, their recognition performance is better when using a neural network to build a Chinese DSR system than the traditional approaches. The CNN has shortcomings in obtaining long-term correlation features combined with context. Namely, the ability to model speech signals is limited.

4.4.4 Long short-term memory

Long short-term memory (LSTM) is a Recurrent Neural Network (RNN) with improved memory function (Graves et al. 2013). It could obtain long-term correlation features combined with context to compensate for the shortcomings of CNN's insufficient understanding of the context. The advantage of LSTM comes from its network structure, which includes in-gates, forgetting gates, out-gates, and hidden states for each cell. This network can not only find the time connections between messages but also extract and store vectors within a long time.

Compared with Chinese DSR based on GMM/HMM, LSTM can directly learn its internal model from data, while GMM-HMM uses probability density distribution to optimize training parameters. On the other hand, dialect speech model construction also depends on the corpus size. Ying et al. (2020) proposed a Sichuan DSR system based on HMM-LSTM. Then they also discussed other models which used DNN to train the Sichuan dialect, which showed that LSTM could obtain more context information from pronunciation and have better accuracy than DNN. Connected to the related studies, some scholars rely on the pronunciation characteristics of dialects to optimize the LSTM model and construct a dialect recognition system. Ye et al. (2019) proposed a dialect recognition system that used NOAA and LSTM to recognize six Chinese dialects. Then the results indicate that the improved model outperformed the single LSTM recognition model.

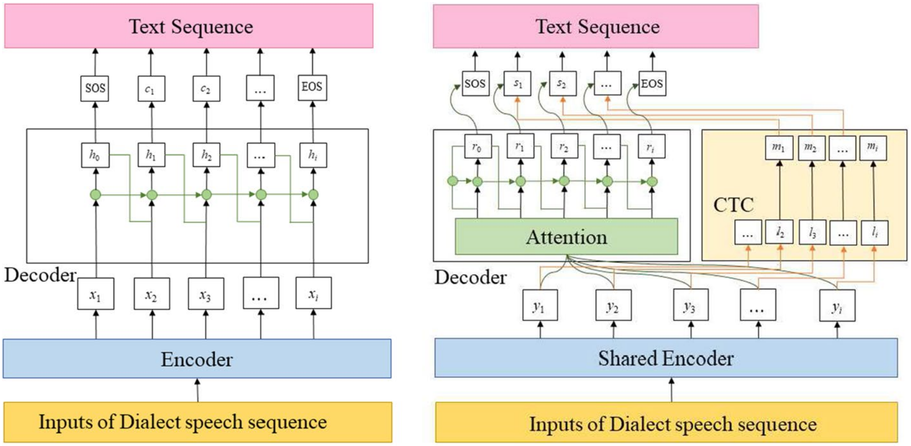
While the Chinese DSR based on LSTM has proven effective in many applications, it could not process data in reverse and has some semantic errors. Science the Bidirectional LSTM (BiLSTM) could obtain the current speech state and the last moment's speech state (Passricha and Aggarwal 2020). Therefore, researchers have utilized BiLSTM (Schuster and Paliwal 1997) to train the Chinese DSR system. Zhao et al. (2019) proposed a Tibetan multi-dialect recognition system based on Connectionist Temporal Classification (CTC)-BiLSTM, which used latent regression Bayesian network to extract dialectal features. Experimental results show that CTC-BiLSTM has a good performance in multi-Tibetan dialects recognition. The above research found that the Chinese DSR based on neural networks can complete the task of dialect speech transcription into text. According to different network structures, they learn deep acoustic features from each layer of network learning, optimize the training model, and effectively improve recognition accuracy.

4.4.5 Connectionist temporal classification

The Chinese DSR system based on a neural network combined with HMM has a good recognition performance. However, some limitations remain, such as forced data alignment and independent assumption of model composition, which eventually lead to time-consuming model training and slow convergence. Therefore, the Chinese DSR system based on the E2E could solve the above problems (Li 2022). As shown in Fig. 8a, the framework of Chinese DSR based on E2E includes input dialect sequence, encoder, decoder, and output dialect character sequence. The x_*, h_*, and c_* represent the encoder output vectors, decoder vectors, and output vectors, respectively. It could realize the direct mapping from dialect speech data to text, simplifying the complex modeling process. There are two types of E2E architectures for DSR: CTC and Attention Mechanism.

CTC could solve the temporal data classification tasks (Hori et al. 2017), which calculates the error between the input data and the given output after passing through the neural network. The Chinese DSR system based on CTC could realize automatic alignment, which uses CTC as the loss function of a neural network to fully model the sequence. Compared with the hybrid model based on a neural network, it does not require pre-segmentation of training data or postprocessing label sequences extracted from the network. When all the label sequences are in the correct order, CTC can predict the label through the network at any point in the input sequence. Generally, CTC combined with different neural networks performs well in Chinese DSR.

Fu et al. (2020) investigated various neural network methods for Chengdu DSR. They used CNN, DNN, and CTC/CNN to extract semantic and morpheme features from the Chengdu dialect. The results showed that CTC/CNN performed better than others since CTC could align characters and map the corresponding speech sequence. Additionally, Nan et al. (2019) introduced an E2E recognition system for Tibetan based on CTC and BLSTM. They trained the 512-dimensional BLSTM layer and used CTC to calculate the output sequence posterior probability, and they confirmed that this system is an effective



(a) E2E framework.

(b) Hybird CTC/Attention framework.

Fig. 8 The basic E2E architecture and hybrid CTC/Attention architecture are used to build Chinese DSR model. “SOS” and “EOS” represent the start and end of an Chinese sentence.

sequence labeler. Going deeper, some studies have discussed that the CTC could eliminate the problem of data alignment and can build an E2E model using a multi-layer network to directly map speech sequences to text. But each sequence is an independent output, and it cannot effectively utilize contextual information to determine whether the Chinese characters of the output dialect conform to logic.

4.4.6 Attention mechanism

The attention mechanism can selectively attend to different aspects of information (Santana Correia and Colombini 2022), enhancing the perceptual properties of biometric systems. The framework of the attention mechanism has added a layer based on E2E, and it mainly includes an encoder module, attention module, and decoder module (Bahdanau et al. 2016). Chinese DSR based on the attention mechanism can implicitly learn the soft alignment between input and output sequences. The decoder network uses an attention mechanism to find alignment between the dialect speech input and the text sequence produced by the encoder network.

Xu et al. (2021) proposed Gan Dialect and Hakka DSR system based on the attention mechanism, which used CNN and BiLSTM to extract dialect features, respectively. The experiment results showed that the performance based on the attention dialect recognition model outperformed the DNN model. Besides, Liu et al. (2020) presented the Datong dialect speech translation system based on the attention mechanism. When they compared the attention mechanism with the cascade model, it was shown that the proposed Chinese DSR not only saved time cost but also effectively improved accuracy. The attention mechanism is often combined with neural networks to identify Chinese dialects. However, due to the non-linear alignment of the attention mechanism, it is easily affected by noise in acquiring information.

4.4.7 Hybrid CTC/attention

The above two E2E DSR approaches can complete dialect speech-to-text transcription, but they need a lot of training data or LM support. Therefore, scholars combined the advantages of CTC and attention mechanism (Kim et al. 2017) that proposed a Chinese DSR based on hybrid CTC/Attention. As shown in Fig. 8b, y_*, r_*, s_* represent the shared encoder output vectors, attention decoder vectors and output vectors, respectively. l_* and m_* are the parameters of CTC. During training, it uses a multiobjective learning framework to improve robustness. In the decoding process, a joint decoding method combines attention and CTC scores in the search algorithm to eliminate irregular alignment. The loss function is calculated as shown in Eq. (2), which adds the loss functions of CTC and Attention.

$$L_{MOL} = \partial L^{ctc} + (1 - \partial) L^{att}. \tag{2}$$

∂ for linear interpolation and $0 < \partial < 1$. when ∂ equals 0, the decoding layer is equivalent to the Attention model. When ∂ is equal to 1, the decoding layer is equivalent to the CTC model. Figure 8 shows the Chinese DSR approaches based on E2E and hybrid CTC/Attention. The hybrid model adds attention mechanism and CTC to the basic E2E framework, analyzing the input Chinese dialect sequences.

Yang and Hu (2021) proposed a dialect accent recognition system trained by hybrid CTC/Attention. The work improves recognition accuracy by reducing the CTC weights and deepening the number of encoder layers. Comparing the work and traditional models, it is clear that hybrid CTC/Attention outperforms the traditional models in recognition rate. What's more, several research scholars proposed optimization E2E recognition methods (Hussein et al. 2022; Gong et al. 2022). They used a multi-head attention mechanism for dialect recognition.

In general, we summarize the commonly used approaches for Chinese DSR in Table 4, which shows the dialect, approaches, accuracy, and description of the Chinese DSR. The recognition method based on HMM has two types: GMM/HMM and ANN/HMM. GMM/HMM is utilized to create a dialect recognition model, which laid the critical foundation for the research of modern dialect recognition. Additionally, the model based on GMM/HMM could calculate the maximum probability of the input speech sequence corresponding to the labeled dialect. Later, GMM was replaced with ANN due to its superior computational capabilities. At that time, the HMM/ANN-based Chinese DSR had advanced to a state-of-the-art level. Its success can be attributed to the fact that Chinese DSR systems have independent modules to train the data and that more dialect features are obtained by designing the layers of neural networks. From the table, ANN/HMM method could build a robust recognition model.

The DSR recognition based on E2E are widely used by ANN/CTC, attention mechanism, and hybrid CTC/Attention. CTC combined with different deep neural networks to optimize the Chinese DSR approaches, which the performance is better than ANN/HMM with the same data, such as Gan dialect (Xu et al. 2021). Moreover, the Chinese DSR based on hybrid CTC/Attention could automatically align the input speech sequence and does not require an independent LM or dialect dictionary, which is the most frequently used method. After comparing the recognition method based on HMM and E2E, the performance of E2E is slightly lower than the traditional approaches when the amount of dialect speech data is small. And the E2E recognition model relies on training data. Most researchers continually supplement dialect corpora and construct DSR systems using an E2E framework. To increase the precision of dialect identification, they further optimize the structure.

4.5 Evaluation

Evaluation is a guideline for assessing DSR models, which formulas to precisely determine the critical performance of the model. Since researchers focused on standard language recognition in early times, dialects were studied as specific languages. Bahari et al. (2013) used P_acc, the correct rate of language identification as an evaluation index.

However, with the gradual enrichment of the dialects data, the evaluation for DSR has been further improved. At present, the primary evaluation uses Character Error Rate (CER), Word Error Rate (WER), and Sentence Error Rate (SER). The calculation method of the CER and WER is similar. CER takes the basic unit of Chinese dialect "character" as the analysis object and measures the performance of the dialect recognition model by calculating the error level of the character. The calculation is shown in Eq. (3). S denotes the dialectal characters replaced in the sentence sequence, D denotes the deleted dialectal characters, I indicates the inserted dialectal characters, and N denotes the total number of Chinese characters.

$$CER = \frac{S + D + I}{N}. \tag{3}$$

WER takes Chinese dialect "words" as the analysis object and is also one of the critical evaluations for the speech recognition system. The calculation is shown in Eq. (4).

Table 4 Summary of Chinese DSR approaches, include including dialect category, method description, and accuracy

Method	Dialects	Approaches	Accuracy (%)	Description
Recognition method based on HMM	Northern Mandarin (Yang and Hu 2021)	HMM-GMM	87.8	HMM calculates the probability of speech sequence, and GMM performs dialect recognition.
	Wu Dialect (Pan et al. 2005)	HMM-GMM	84.32	
	Xiang Dialect (Wang et al. 2009)	GMM	74	GMM calculates the probability of hidden states of dialect speech features.
	Cantonese (Li et al. 2019)	GMM-HMM	62.1	
	Gan Dialect (Xu et al. 2021)	DNN-HMM	65.3	DNN obtains contextual information through successive dialect frames and trains status sequence and dialect label.
	Uyghur (Qimike et al. 2015)	DNN-HMM	80.43	
	Tibetan (Li et al. 2019)	DNN-HMM	85.72	
	Hakka Dialect (Yu 2019)	CNN-HMM	84.64	CNN has the characteristics of pooling and sharing so that could obtain acoustic features.
	Uyghur (Iminjan et al. 2021)	CNN-HMM	82.9	
Recognition method based on E2E	Chengdu Dialect (Fu et al. 2020)	CNN-CTC	96.7	CNN and CTC are combined to build a dialect recognition model
	North Mandarin (Yang et al. 2022)	DFCNN-CTC	87.2	It is built by referring to Transformer's connection method and combining with CTC
	Gan Dialect (Xu et al. 2021)	LSTM-CTC	90.8	LSTM is used to extract the dialect features, and CTC trains the recognition model.
	Tibetan (Wang et al. 2017)	LSTM-CTC	81.29	
	Hakka Dialect (Xu et al. 2021)	Attention	90.7	Optimizing attention mechanism and using self-attention to train recognition model.
	Uyghur (Ding et al. 2020)	Attention	85.3	
	Uyghur (Lu et al. 2021)	CTC-Attention	91.35	The advantages of CTC and attention are combined to model the context of dialect sequences
	Tibetan (Nan et al. 2019)	BiLSTM-CTC	87.95	BiLSTM is used as initial training network, and CTC is used to align the sequence of dialect

Assuming that the word sequences and transcription sequences are consistent, S denotes the number of errors in the replaced words, D denotes the number of deleted words, I indicates the inserted dialect words, and N represents the sum of words.

$$WER = \frac{S + D + I}{N}. \tag{4}$$

SER represents the ratio of misidentified sentences to the total dialect sentences. Equation (5) is a calculation formula in which N represents the total number of all sentences, and E represents the number of sentences with at least one Chinese character error in the data.

$$SER = \frac{E}{N}. \tag{5}$$

5 Challenges of Chinese DSR

The discussion above has comprehensively analyzed the advancements in Chinese dialect speech signal processing, feature extraction, and recognition methods. Chinese DSR technologies have evolved from traditional GMM/HMM-based approaches to advanced deep learning methods, yielding significant improvements in recognition performance. However, due to the inherent characteristics of Chinese dialects, their identification process still faces numerous challenges, particularly in the following aspects:

(1) *Corpus Scale, standardization, and diversity:* The scale of speech data for different dialects varies significantly, especially for some dialects with a low usage range, where available speech data tends to be limited and restricts the development and application of recognition systems. For instance, the sample size of the Lanzhou dialect corpus constructed in literature (Yang et al. 2009) is relatively small, mainly due to its limited usage and consequently fewer recording participants. Additionally, some dialect corpora lack a unified standardization process, with differences in text sources, recording frequencies, and other settings across dialect corpora, which further affects the usability of speech data in model training processes. For example, the Gan dialect corpus utilizes local news and broadcast audio content as its dialect text sources (Xu et al. 2018), while the Datong dialect corpus involves recording everyday conversations in a studio (Liu et al. 2020), resulting in differences in textual information and speech frequencies. Moreover, factors such as the educational background, work experience, and life experiences of dialect speakers lead to inconsistent annotation levels, increasing the difficulty of corpus standardization. Lastly, the comprehensiveness and diversity of corpora also pose challenges, which should include speakers of various genders, ages, social backgrounds, and educational levels to better capture the unique phonetic features inherent in dialects.

(2) *Diversity and evolution of dialectal phonetic features:* The formation and development of Chinese dialects are dynamic, continuously changing over time under the influence of social environments and cultural exchanges. Even within the same dialect, pronunciation can vary significantly across different regions, including differences in phonemes, tones, and rhythm (Gong et al. 2011). For example, the specific differences in the Uyghur dialects of the Hetian and Luobu regions are manifested in vowels (Sun

et al. 2019). This increases the complexity of speech recognition systems and makes extracting speech features difficult. Furthermore, in multilingual environments, people often communicate employing a mix of several dialects or a hybrid language combining dialects with Mandarin (Yang and Hu 2021), presenting additional challenges for recognition models.

(3) *Accent variation:* Another significant challenge faced in dialect recognition is the diversity of accents. Even within the same dialect, accents can vary significantly from region to region. These variations often manifest themselves in the use of specific vocabulary, the speed or intonation of speech, or other factors. For instance, both Chengdu and Chongqing dialects are categorized under the Sichuan dialect. However, the falling tone in the Chengdu accent owns a distinct falling-rising intonation (Fu et al. 2020), while in the Chongqing accent, the falling tone is more obviously rising (Zhang et al. 2018). The diversity and dynamism of these characteristics necessitate that recognition systems be capable of understanding and processing various pronunciation patterns and scenarios. Especially with the development of society and population mobility, pronunciation habits will also constantly change, thus forming new accents, which brings additional challenges to dialect recognition.

(4) *Limitations of recognition methods:* Existing Chinese DSR methods still have some limitations. Firstly, traditional GMM/HMM methods are sensitive to noise in speech signals (Ouisaadane and Safi 2021), which means that their recognition accuracy will decrease in a noisy environment. Secondly, although deep learning methods are more effective in dealing with complex speech patterns, they typically rely on a large amount of training data (Yu et al. 2020). For dialects with a smaller usage range, the limitation of the corpus scale will lead to poor recognition performance. Additionally, in real-world speech scenarios, factors such as environmental noise, the speaker's emotions, variations in speech rate, and other conditions will also influence recognition performance, which poses higher requirements on the robustness of the models. Therefore, improving the robustness of the model and effectively handling small-scale dialect corpora have become critical challenges in the field of Chinese DSR.

(5) *Updates of Language Models:* The language model plays a vital role in Chinese DSR (Ren et al. 2019). These models not only provide grammatical and lexical support for speech recognition but also enhance the system's ability to comprehend and transcribe dialect content accurately. However, the study of language habits and vocabulary usage in specific dialects still faces some challenges. Specifically, new vocabulary and expressions are constantly emerging, which requires language models to be able to adapt to these changes and continuously update and learn.

In summary, within the current context of deep learning, the primary challenge facing Chinese DSR systems is to develop more flexible and accurate models to address the diversity, dynamism, and complexity of dialects. To this end, not only is a high-quality dataset required to support the development and training of models but also to achieve interdisciplinary integration and innovation, including but not limited to the fields of linguistics, speech recognition, and noise processing. Furthermore, in practical applications, the generalizability of the model and its handling of non-standard or abnormal speech are equally crucial, further ensuring the system's stability and reliability.

6 Application of Chinese DSR

Dialect is an indispensable part of the regional culture (Hu et al. 2022), which carries the life wisdom and cultural genes of the people in the region, such as the characteristics of language emotional expression, way of thinking, and values. As the carrier of communication, dialects help people deeply understand the region's cultural characteristics and historical heritage, which is conducive to promoting cultural diversity and cross-cultural understanding. Nowadays, Chinese DSR is widely used in many aspects. There are various application fields for a particular use. For example, healthcare, voice assistant, and so on. Figure 9 represents the related technologies and Chinese DSR applications.

With the increasing popularity of various high-performance recorders and microphones, dialect audio information in daily life can be quickly recorded and saved. At the same time, using cloud computing, artificial intelligence, and other technologies, the development of corresponding DSR applications or related services can provide personalized dialect speech services for all walks of life. The DSR applications will not only satisfy people's inner needs for the enjoyment of life and culture but also further promote economic development and social prosperity. The combination of dialects and artificial intelligence (AI) technology is significant for preserving and inheriting dialect culture, especially endangered dialects and key regional dialects. Therefore, we should actively use AI and other technical means to protect, promote and optimize dialect culture so that more people can understand and feel the charm of regional culture.

In the field of healthcare, in order to provide efficient and accurate information exchange for medical workers during the COVID-19 pandemic, universities, and research institutions jointly developed the "Hubei Dialect Guide for Fighting the Epidemic" (Wang et al. 2020), which covers nine dialect areas in Hubei. Moreover, the software includes several application versions, which provide a convenient solution for dialect speech communication, such

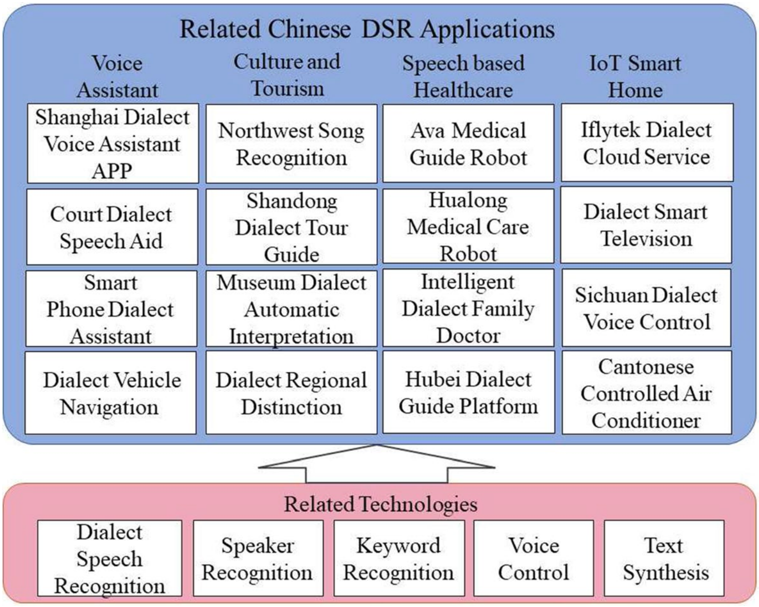

Related Chinese DSR Applications
Voice Assistant
Culture and Tourism
Speech based Healthcare
IoT Smart Home
Shanghai Dialect Voice Assistant APP
Northwest Song Recognition
Ava Medical Guide Robot
Iflytek Dialect Cloud Service
Court Dialect Speech Aid
Shandong Dialect Tour Guide
Hualong Medical Care Robot
Dialect Smart Television
Smart Phone Dialect Assistant
Museum Dialect Automatic Interpretation
Intelligent Dialect Family Doctor
Sichuan Dialect Voice Control
Dialect Vehicle Navigation
Dialect Regional Distinction
Hubei Dialect Guide Platform
Cantonese Controlled Air Conditioner
Related Technologies
Dialect Speech Recognition
Speaker Recognition
Keyword Recognition
Voice Control
Text Synthesis

Fig. 9 The related technologies and application fields of Chinese DSR.

as the WeChat version, online version, integrated media pocketbook, instant translation software, online dialect service, and other forms. In addition, there are also commercial home robots (Chen et al. 2017), such as Ava Medical Guide Robot and Hualong Medical Care Robot, which are based on dialect voice service. The intelligent medical speaker developed and promoted by Spectrum can also understand dialects and heavy accent Mandarin, which is a good choice for family doctors.

In culture and tourism, dialect recognition can give users a special sensory experience. On the one hand, the DSR application plays an important role in dialect songs, movies, and TV plays. For example, the development of dialect song speech recognition can help people understand dialect culture through songs (Islam et al. 2015) and promote its inheritance and development. On the other hand, DSR applications can help tourists better understand and integrate into the local culture to improve their travel experience. For example, a dialect guide application (Ogundokun et al. 2021) is developed to make tourists understand the local dialect tour guide's explanation and the local culture personally.

In the high-tech or voice assistant field, DSR is applied to dialect speech search and intelligent customer service. Through dialect speech search (Chiang 2017), users can find the required information faster and improve the experience. By using dialect intelligent customer service, users can communicate with enterprises more conveniently, which improves service quality and efficiency. For example, Chinese dialect voice commands are used to operate mobile applications (Ma 2014) and handle related business, etc. DSR is also widely used in the court system to improve work efficiency and serve the public better. Due to the different litigants' accents and dialects, the Voice Cloud Service deployed by the People's Supreme Court can recognize 29 different dialects (Wang 2020).

In the field of IoT smart home, through voice control access, hardware devices have become the target of enterprise manufacturing intelligent products. For example, Midea air conditioners, water heaters, ovens, and other household appliances support a variety of dialects of speech command recognition (Cantonese, Sichuan, Shandong, Shanghai, etc.). By identifying keywords such as "turn on the light", "turn off the light", and other words (Ni et al. 2019) to realize the dialect voice control of home appliances. Other types of smart household appliances can also effectively combine DSR technology. In other ways, Iflytek adopts iFlytek dialect Cloud service (Zhu 2019), which has supported the recognition and synthesis of 24 dialects to provide a full range of dialect recognition technology for home appliance manufacturers. And the accuracy rate of Cantonese, Sichuan, Northeastern, Henan, Tianjin, Shandong, and Ningxia dialects is over 90%.

With the advancement of deep learning techniques such as transfer learning and self-supervised learning (SSL), multilingual automatic speech recognition (ASR) systems have made significant progress in recent years. This process often requires SSL pre-training on a large amount of unlabeled multilingual audio data, which can more effectively capture the characteristics and structures of different languages by constructing language-aware encoders and adapter modules (Yadav and Sitaram 2022). In addition, studies also indicate that compared with monolingual models, some multilingual models show better performance when trained with the same amount of data, especially in low-resource languages (Miao et al. 2022). The acoustic model modeling methods for multilingual ASR models primarily include those based on model parameter sharing and those based on language classification information for multilingual acoustic modeling (Cheng and Yan 2022), suitable for language recognition in different resource situations. For model parameter sharing models, all languages share the same set of acoustic model parameters, while models based on language classification information introduce language-specific information into the acoustic model. To develop scalable multilingual models, there is a need for large, diverse, and easily expandable multilingual corpora (Salesky et al. 2021). Currently, Meta provides a large-scale corpus of over 1100 languages Pratap et al. (2023) and has pre-trained the wav2vec 2.0 model, enabling multilingual ASR and speech synthesis. Multilingual ASR models have great potential in practical applications, such as the multilingual air traffic control system (Lin et al. 2021), which can recognize Chinese and English speech in real-time, thereby understanding control intentions and improving control efficiency. Moreover, multilingual ASR is attempting to build high-quality recognition models using lifelong learning methods (Li et al. 2022) to achieve effective recognition of speech in various languages, domains, and environments. Cross-language transfer leverages multiple resource-rich languages to build models for resource-poor target languages (Hou et al. 2021), providing new pathways for speech recognition in low-resource languages.

The rapid development of multilingual ASR provides crucial references for advancing Chinese multi-dialect recognition technology. At the same time, due to the diversity and richness of Chinese dialects, which positively impact cultural exchange and regional economic development, there is an escalating demand for applications such as multi-dialect speech recognition and cross-dialect interaction, with broad development prospects. Currently, the technology for Chinese multi-dialect speech recognition mainly adopts E2E methods. For example, Dan et al. proposed an E2E multi-task learning framework, which combines soft parameter sharing and Speech-Transformer, specifically for handling multi-dialect speech recognition tasks (Li et al. 2022) and (Dan et al. 2022). In addition, Zhao et al. proposed an E2E model that integrates WaveNet-CTC with multi-task learning, effectively achieving multi-dialect speech recognition and dialect detection for Tibetan, with experimental results demonstrating its significant potential in processing multi-dialects, especially resource-constrained dialects (Zhao et al. 2019). However, the acoustic features and languages differences of various Chinese dialects, such as intonation, grammatical structure, and vocabulary, further increase the complexity of multi-dialect speech recognition tasks (Li et al. 2022) and (Dan et al. 2022). In terms of practical applications, numerous IT companies have shown strong interest in Chinese multi-dialect speech recognition technology and have developed APIs or SDKs to support real-time or offline dialect recognition functions. For instance, the input method of iFlytek can recognize 23 different Chinese dialects (Zhu 2019), while Microsoft has developed intelligent speech applications that support the Wu, Cantonese, and Southwestern Mandarin dialects. Additionally, ByteDance has integrated support for eight major Mandarin dialects into its video captioning feature. Overall, the rapid development of deep learning technology in multilingual ASR has provided innovative ideas and methods for multi-dialect recognition and cross-dialect interaction, encouraging the global popularization of speech recognition technology and the expansion of its application scenarios. In the future, the advancement of research in dialect large language models will further promote the application of multi-dialect speech recognition and interaction in various fields, such as medical care, safety production, and more.

In conclusion, the development potential of DSR in different fields is huge, which can provide better services and experience for various industries, and also contribute to the protection and inheritance of dialect culture. Dialect plays a vital role in the development of regional culture. And it is irreplaceable in regional culture's inheritance, protection, and promotion.

7 Future research scope on Chinese DSR

Although Chinese DSR technology has demonstrated promising application performance in multiple fields, its development potential still needs to be fully explored. Compared with Mandarin, most dialects are still low-resource languages, and there are differences in the specific research progress of each dialect recognition model. Nevertheless, they encounter similar developmental problems. Consequently, this article recommends that future research should primarily focus on the following areas:

(1) *Model optimization:* Considering the uniqueness of dialect speech, exploration of more suitable model structures for processing these characteristics is necessary. An effective practical approach is to adopt a transfer learning strategy (Wang et al. 2021), which involves initially pre-training a model on a large-scale general speech dataset and then fine-tuning it on the data of a specific dialect. For instance, Wang et al. (2022) constructed a cross-language recognition model for Lhasa and Tibetan based on transfer learning with limited data and achieved excellent results. Furthermore, Generative Adversarial Networks (GANs) (Chen et al. 2020) can also generate speech samples specific to certain regions, thereby effectively managing the differences between various dialects. Multi-task learning is another effective strategy, allowing the training of models to recognize different dialects simultaneously. For example, Dan et al. (2022) applied multi-task learning to the task of low-resource multi-dialect speech recognition. By sharing soft parameters and setting up two task streams, they learned the commonalities between different dialects, enhancing the model's generalization ability. Finally, integrating model fusion techniques (Escobar-Grisales et al. 2022) can further improve recognition accuracy. Through fusion strategies, a deeper analysis of the representation methods of dialect speech and text is facilitated, optimizing the final output.

(2) *Data enhancement:* Since Chinese dialect corpora are generally small in scale, it is worthwhile to explore data augmentation methods to expand the dataset (Das et al. 2021). This approach can enhance data diversity and overall quality through techniques such as variable speed playback and adding noise. In addition to these technical approaches, strategies based on feedback mechanisms can also be explored to expand the dataset (Kusherbaeva and Zhou 2022) and facilitate iterative improvements according to the collected results. For example, Ballard et al. (2019) combined speech recognition and feedback mechanisms, collecting feedback from specific user groups in actual application scenarios through mobile applications to continuously optimize and adjust the recognition system. At the same time, these data can also serve as additional training resources.

(3) *Noise suppression:* Background noise in speech signals is a common problem for general speech recognition tasks. However, the complexity of dialect phonetic features makes them more susceptible to interference, affecting recognition speed and accuracy. Consequently, it becomes particularly crucial to deeply explore how to suppress various types of noise more effectively during the dialect recognition process. This includes developing and applying more advanced noise suppression algorithms (Reddy et al. 2020), enabling more precise identification and elimination of various types of noise. Additionally, using deep learning technologies to recognize and separate human voices from background noise (Al-Barhan et al. 2021) is also a noteworthy direction. Furthermore, future research should also focus on the characteristics of noise in different environments (such as outdoors, inside vehicles, public places, etc.) (Dubey et al.

2022), and explore how to adjust noise suppression strategies adaptively to suit these specific environments.

(4) *Dialect speech emotion recognition:* Investigating the emotional information contained in dialect speech is essential for creating personalized services and improving user experience (Aljuhani et al. 2021). It can not only augment the understanding of dialect culture but also further perceive users' emotions. Additionally, this has significant value in building natural human-computer interaction systems (Wani et al. 2021). Therefore, future research can focus on analyzing the special emotional features in dialect speech, building emotional speech datasets, and optimizing emotion recognition algorithms. For instance, Cheng et al. (2021) constructed a Henan dialect speech emotion database, which includes five emotions: happiness, surprise, sadness, anger, and neutrality. The effectiveness of the dataset was verified using different network models. The experimental results showed that the average recognition rate of the Learning Vector Quantization (LVQ) neural network was the highest. By improving the accuracy and practicality of Chinese dialect emotion recognition, we can further promote the development of related applications in terms of intelligence and humanization.

(5) *Multi-dialect recognition:* With the rapid economic development and significant population mobility, there is a growing demand for speech recognition systems capable of processing multiple dialects. Consequently, future research should focus on developing multi-dialect speech recognition systems to support the voice wake-up and recognition of various dialects. For instance, Wan et al. (2022) analyzed 10 Chinese dialects and constructed a multi-dialect speech recognition system based on deep neural networks. The system adopts a multi-task learning model assisted by dialect region, which reduces network complexity and enhances recognition accuracy compared to a single-task model. Moreover, in numerous application scenarios, such as customer service systems and real-time translation, besides the need for support of multiple dialects, it is also required to the speech recognition system maintains efficient real-time processing capabilities (Chen et al. 2021a). Future research can focus on reducing the system's latency and increasing the speed of multi-dialect recognition, while maintaining accuracy. This may involve algorithm optimization, the application of hardware acceleration technologies, and the full utilization of cloud computing resources (Tyagi et al. 2023).

Currently, Chinese DSR is in a rapid development stage, showing broad prospects of application. In light of the current overarching trends and dynamics in Chinese DSR, this article provides a comprehensive analysis of its future research directions. This not only broadens the current research perspective from multiple aspects but also provides valuable reference and guidance for the further development of this field.

8 Conclusion

This paper provides a systematic overview of existing research on Chinese DSR and its related applications. First, we detailed the primary acoustic features of Chinese dialects, including pronunciation and intonation. For further analysis, we categorize the available Chinese dialect corpora according to the distinct divisions of Chinese dialects. This classification considers various vital components such as dialect type, audio duration, recording environment, number of recorders, and text type, which are crucial for a comprehensive understanding of the dialect corpus data. The structural composition and construction methods of various dialect corpora are compared, and we summarize and propose a basic construction process for a particular dialect corpus. Finally, regarding the research status of available Chinese speech recognition systems, this paper comprehensively analyzes and summarizes the standard approaches and related applications of Chinese DSR. Furthermore, with the development of multicultural communication, there are still many challenges for Chinese DSR. Although a single dialect currently dominates the dialect corpus, there is an increasing demand for multi-dialect corpora. As a result, the continuous optimization of speech storage devices and technologies provides a platform for the construction of multi-dialect corpus or mixed dialect corpus (Li et al. 2019; Luo et al. 2022). Chinese dialects are mainly tonal characters, especially in low-resource dialects (Xu et al. 2021) with similar pronunciations. Additionally, the speech annotation algorithm needs to be further optimized to avoid the deviation of speech and text information when annotating dialect corpus audio. Moreover, in the Chinese DSR based on E2E, some scholars add LM to achieve optimal speech matching to text. In conclusion, advancements in Chinese DSR technology are beneficial for promoting diverse and personalized intelligent voice services. Additionally, this technology can help to preserve and share phonetic data resources for Chinese dialects, ultimately contributing to the protection and inheritance of Chinese dialect culture.

Acknowledgements Research reported in this paper was partially supported by the Provincial Natural Science Foundation of Ningxia (Grant No.2022AAC05040); the National Natural Science Foundation of China (Grant No.61962002 and NO.61402017). These supports are greatly acknowledged.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Abdel-Hamid O, Mohamed A-R, Jiang H, Deng L, Penn G, Yu D (2014) Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio Speech Lang Process 22(10):1533–1545

Ai H, Fei L (2019) Identification of Guizhou dialect based on improved convolutional neural network. Mod Inform Technol 3(1):5–10

Akçay MB, Oğuz K (2020) Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun 116:56–76

Al-Barhan HA, Elyass SM, Saeed TR, Hatem GM, Ziboon HT (2021) Modified speech separation deep learning network based on hamming window. In: IOP conference series: materials science and engineering, vol 1076, IOP Publishing, p 012059

Ali MH, Jaber MM, Abd SK, Rehman A et al (2022) Harris hawks sparse auto-encoder networks for automatic speech recognition system. Appl Sci 12(3):1091

Aljuhani RH, Alshutayri A, Alahdal S (2021) Arabic speech emotion recognition from Saudi dialect corpus. IEEE Access 9:127081–127085

Ardila R, Branson M, Davis K, Henretty M, et al (2019) Common voice: a massively-multilingual speech corpus. arXiv:1912.06670

Badea A, Halunga S, Berceanu M et al (2019) Influence of Manchester encoding over spreading codes used in multiple access techniques for IoT purposes. figshare. https://doi.org/10.1109/SIITME47687.2019.8990780

Bahari MH, Saeidi R, Van Leeuwen D, et al (2013) Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech. In: 2013 IEEE international conference on acoustics, speech and signal processing, IEEE, pp 7344–7348

Bahdanau D, Chorowski J, Serdyuk D, Brakel P, Bengio Y (2016) End-to-end attention-based large vocabulary speech recognition. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 4945–4949

Ballard KJ, Etter NM, Shen S, Monroe P, Tien Tan C (2019) Feasibility of automatic speech recognition for providing feedback during tablet-based treatment for apraxia of speech plus aphasia. Am J Speech Lang Pathol 28(2S):818–834

Berndt DJ, Clifford J (1994) Using dynamic time warping to find patterns in time series. In: KDD Workshop, vol 10 Seattle, WA, pp 359–370

Bhatia S, Kumar A, Reddy T, Varshney N, Basheer S (2023) Matrix quantization and LPC vocoder based linear predictive for low-resource speech recognition system. ACM Trans Asian Low Resour Lang Inform Process 16(04):18–21

Bolia RS, Nelson WT, Ericson MA, Simpson BD (2000) A speech corpus for multitalker communications research. J Acoust Soc Am 107(2):1065–1066

Bouamor H, Habash N, Salameh M, et al (2018) The MADAR Arabic Dialect Corpus and Lexicon. Paper presented at LREC

Bu H, Du J, Na X, Wu B, Zheng H (2017) Aishell-1: an open-source mandarin speech corpus and a speech recognition baseline. figshare. https://doi.org/10.1109/ICSDA.2017.8384449

Chen L, Sun R, Liu Y, Chen J, Li Z (2019) Quantitative model of phonetic differences among Chinese dialects. J Beijing Normal Univ 20(103–110):8

Chen M, Wang L, Xu C-Z, Li R (2017) A novel approach of system design for dialect speech interaction with NAO robot. figshare. https://doi.org/10.1109/ICAR.2017.8023652

Chen Y-C, Yang Z, Yeh C-F, Jain M, Seltzer ML (2020) Aipnet: generative adversarial pre-training of accent-invariant networks for end-to-end speech recognition. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 6979–6983

Chen X, Wu Y, Wang Z, Liu S, Li J (2021) Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp. 5904–5908

Chen X, Wu Y, Wang Z, Liu S, Li J (2021) Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 5904–5908

Cheng G, Yan Y (2022) Latest development of multilingual speech recognition acoustic model modeling methods. Comput Sci 49:47–52

Cheng Z, Li Y, Jiu M, Ge J (2021) Speech emotion recognition based on henan dialect. In: International conference in communications, signal processing, and systems, Springer, pp 199–206

Chiang C-Y (2017) Cross-dialect adaptation framework for constructing prosodic models for Chinese dialect text-to-speech systems. IEEE/ACM Trans Audio Speech Lang Process 26(1):108–121

Dan Z, Zhao Y, Bi X, Wu L, Ji Q (2022) Multi-task transformer with adaptive cross-entropy loss for multi-dialect speech recognition. Entropy 24(10):1429

Dan Z, Zhao Y, Bi X, Wu L, Ji Q (2022) Multi-task learning with auxiliary cross-attention transformer for low-resource multi-dialect speech recognition. In: CCF international conference on natural language processing and Chinese computing, Springer, pp 107–118

Das N, Chakraborty S, Chaki J, Padhy N, Dey N (2021) Fundamentals, present and future perspectives of speech enhancement. Int J Speech Technol 24:883–901

Deqing Z (2010) The research on the Tibetan speech feature parameter based on speaker-dependent small vocabulary. PhD thesis, Tibet University

Ding F, Guo W, Sun J (2020) Research on end-to-end speech recognition system for Uyghur. J Chin Comput Syst 41(1):19–23

Dua M (2023) Gujarati language automatic speech recognition using integrated feature extraction and hybrid acoustic model. figshare. https://doi.org/10.1007/978-981-19-7753-4_4

Dubey H, Gopal V, Cutler R, Aazami A, Matusevych S, Braun S, Eskimez SE, Thakker M, Yoshioka T, Gamper H, et al (2022) Icassp 2022 deep noise suppression challenge. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 9271–9275

Escobar-Grisales D, Rios-Urrego C, Gallo-Aristizabal J, López-Santander D, Calvo-Ariza N, Nöth E, Orozco-Arroyave J (2022) Colombian dialect recognition from call-center conversations using fusion strategies. In: Workshop on engineering applications, Springer, pp 54–65

Etman A, Louis A (2015) American dialect identification using phonotactic and prosodic features. In: Paper presented at 2015 SAI intelligent systems conference (IntelliSys), 2015

Fan J, Xiao Z (2022) The classification of Chinese languages and the definition of language code set. Jinan J Philos Soc Sci 44(3):9

Florescu D, Bhandari A (2022) Unlimited sampling with local averages. figshare. https://doi.org/10.1109/ICASSP43922.2022.9747127

Fu J, Li Y, Tao W, Luo J, Li W (2020) Chengdu dialect recognition based on convolutional neural network. J China West Norm Univ Nat Sci 41(4):440–444

Fukuda T, Nitta T (2004) Orthogonalized distinctive phonetic feature extraction for noise-robust automatic speech recognition. IEICE Trans Inf Syst 87(5):1110–1118

Gong Y, Chow IH-S, Ahlstrom D (2011) Cultural diversity in china: dialect, job embeddedness, and turnover. Asia Pac J Manag 28:221–238

Gong Y, Chow IH, Ahlstrom D (2011) Cultural diversity in china dialect, job embeddedness, and turnover. Asia Pac J Manag 28(2):221–238

Gong X, Lu Y, Zhou Z, Qian Y (2022) Layer-wise fast adaptation for end-to-end multi-accent speech recognition. arXiv:2204.09883

Graves A, Jaitly N, Mohamed A-R (2013) Hybrid speech recognition with deep bidirectional LSTM. In: 2013 IEEE workshop on automatic speech recognition and understanding, IEEE, pp 273–278

Gu M-J, Kang S-G (2022) A study on the improvement of chinese automatic speech recognition accuracy using a lexicon. In: 2022 international conference on artificial intelligence in information and communication (ICAIIC), IEEE, pp 313–317

Gu M, Xia Y (2008) Chinese dialect identification using clustered support vector machine. 2008 international conference on neural networks and signal processing, 7–11 June 2008

Guntur RK, Ramakrishnan K, Vinay Kumar M (2022) An automated classification system based on regional accent. Circuits Syst Signal Process 41(6):1–21

Han Q, Yu H (2010) Research on speech recognition for Ando Tibetan besed on HMM. Softw Guide 9(7):173–175

Haugen E (1966) Dialect, language, nation-1. Am Anthropol 68(4):922–935

Hermansky H (1990) Perceptual linear predictive (PLP) analysis of speech. J Acoust Soc Am 87(4):1738–1752

Honnavalli D, Shylaja S (2021) Supervised machine learning model for accent recognition in English speech using sequential MFCC features. Figshare. https://doi.org/10.1007/978-981-15-3514-7_5

Hori T, Watanabe S, Hershey JR (2017) Joint CTC/attention decoding for end-to-end speech recognition. figshare. https://doi.org/10.18653/v1/P17-1048

Hou W, Zhu H, Wang Y, Wang J, Qin T, Xu R, Shinozaki T (2021) Exploiting adapters for cross-lingual low-resource speech recognition. IEEE/ACM Trans Audio Speech Lang Process 30:317–329

Hu Z (2013) A course in linguistics. PEKING UNIVERSITY PRESS, Beijing

Hu H, Yu G, Xiong X, Guo L, Huang J (2022) Cultural diversity and innovation: an empirical study from dialect. Technol Soc 69:101939

Hussein A, Watanabe S, Ali A (2022) Arabic speech recognition by end-to-end, modular systems and human. Comput Speech Lang 71:101272

Iminjan M, Hamdulla A, Mijit A (2021) Uyghur speech recognition based on CNN-HMM and RNN. Mod Electron Tech 44(17):5

Islam R, Xu M, Fan Y (2015) Chinese traditional opera database for music genre recognition. In: Paper presented at 2015 International Conference Oriental COCOSDA held jointly with 2015 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE), pp 38–41

Juang BH, Rabiner LR (1991) Hidden Markov models for speech recognition. Technometrics 33(3):251–272

Keerio A, Mitra BK, Birch P, Young R, Chatwin C (2009) On preprocessing of speech signals. Int J Signal Process 5(3):216–222

Kethireddy R, Kadiri SR, Gangashetty SV (2022) Exploration of temporal dynamics of frequency domain linear prediction cepstral coefficients for dialect classification. Appl Acoust 188:108553

Kim S, Hori T, Watanabe S (2017) Joint CTC-attention based end-to-end speech recognition using multi-task learning. In: Paper presented at 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2017

Kusherbaeva V, Zhou N (2022) Multiobjective data-driven production optimization with a feedback mechanism. IEEE Trans Industr Inf 19(4):5456–5464

Labied M, Belangour A, Banane M, Erraissi A (2022) An overview of automatic speech recognition preprocessing techniques. In: 2022 international conference on decision aid sciences and applications (DASA), IEEE, pp 804–809

Lai Y (2022) Application of the artificial intelligence algorithm in the automatic segmentation of mandarin dialect accent. Mob Inf Syst 2022(12):1–7

Li Y (2012) Problems in contemporary Chinese language life. Soc Sci China 9(201):150–156

Li L (2018) On the history of Chinese dialect partition and its methods. Chin J Lang Policy Plan 3(2):38–49

Li J (2022) Recent advances in end-to-end automatic speech recognition. APSIPA Trans Signal Inform Process 11(1):1–64

Li G, Meng M (2012) Research on acoustic model of large-vocabulary continuous speech recognition for Lhasa Tibetan. Comput Eng 38(5):189–191

Li A, Yin Z, Wang T, Fang Q, Hu F (2004) RASC863-A Chinese speech corpus with four regional accents. ICSLT-o-COCOSDA, New Delhi

Li R, Zhao Z (2017) Isolated word recognition of Hengyang dialect. Comput Syst Appl 26(5):247–252

Li J, Zheng TF, Byrne W, Jurafsky D (2006) A dialectal Chinese speech recognition framework. J Comput Sci Technol 21(1):106–115

Li B, Wang X, Beigi H (2019) Cantonese automatic speech recognition using transfer learning from mandarin. arXiv:1911.09271

Li B, Pang R, Zhang Y, Sainath TN, Strohman T, Haghani P, Zhu Y, Farris B, Gaur N, Prasad M (2022) Massively multilingual ASR: a lifelong learning solution. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 6397–6401

Li A, Yin Z, Wang M (2001) Chinese annotated dialogue and conversation corpus. In: Paper presented at the 5th national conference on modern phonetics

Liao G (1994) Annals of Meixian. Guangdong People Publishing House, Guangzhou

Lin Y, Yang B, Li L, Guo D, Zhang J, Chen H, Zhang Y (2021) Atcspeechnet: a multilingual end-to-end speech recognition framework for air traffic control systems. Appl Soft Comput 112:107847

List JM (2015) Network perspectives on Chinese dialect history chances and challenges. Bull Chin Ling 8(1):27–47

Liu X, Song W, Yu B, Huan J, Chen X, Li Z (2020) Research on attention-based speech translation model of Datong dialect. J North Univ China 41(3):238–243

Liu Y, Fung P (2006) Multi-accent Chinese speech recognition. In: Paper presented at the 9th international conference on spoken language processing(ICSLP), 2006

Liu Z, Lei L, Huang X, Li X, Liu H (2021) Design and realization of dialect interaction system based on VAD. In: 2021 international conference on culture-oriented science and technology (ICCST), IEEE, pp 72–76

Logan B (2000) Mel frequency cepstral coefficients for music modeling. In: In international symposium on music information retrieval. Citeseer

Lu K, Wu C, Liang Y et al (2021) An End-to-End Chinese speech recognition algorithm integrating language model. Acta Electonica Sin 49(11):2177

Luo J, Wang J, Cheng N, Zheng Z, Xiao J (2022) Adaptive activation network for low resource multilingual speech recognition. In: 2022 International joint conference on neural networks (IJCNN), IEEE, pp 1–7

Ma H (2014) Iflytek released a number of new voice power smart home field. Comput Netw 40(16):32–33

Ma B, Zhu D, Tong R (2006) Chinese dialect identification using tone features based on pitch flux. In: 2006 IEEE International conference on acoustics speech and signal processing proceedings, vol 1, IEEE

Malik M, Malik MK, Mehmood K, Makhdoom I (2021) Automatic speech recognition: a survey. Multim Tools Appl 80(9):9411–9457

Malmasi S, Refaee E, Dras M (2015) Arabic dialect identification using a parallel multidialectal corpus. In: Conference of the pacific association for computational linguistics, vol 593, Springer, pp 35–53

Miao L, W, J, Behre P, Chang S, Parthasarathy S (2022) Multilingual transformer language model for speech recognition in low-resource languages. In: 2022 Ninth international conference on social networks analysis, management and security (SNAMS), IEEE, pp 1–5

Nan C, Cai R, Du G (2019) Tibetan speech recognition based on BLSTM-CTC. J Qinghai Norm Univ Nat Sci Ed 35(4):26–33

Ni R, Zhang Y, Ren Z, Chen R (2019) Development of intelligent home appliance control system with embedded multi-language speech recognition. Instrum Technol 1(8):17–20

Nisar S, Tariq M (2018) Dialect recognition for low resource language using an adaptive filter bank. Int J Wavel Multiresolut Inf Process 16(04):1850031

Nurmemet Y, Wushour S (2013) Research on large vocabulary continuous speech recognition for Uyghur. Comput Eng Appl 49(9):115–119

Ogundokun RO, Awotunde JB, Misra S, et al (2021) An android based language translator application. In Journal of Physics: Conference Series, vol 1767, IOP Publishing, p 012032

Ouisaadane A, Safi S (2021) A comparative study for Arabic speech recognition system in noisy environments. Int J Speech Technol 24(3):761–770

O'Shea K, Nash R (2015) An introduction to convolutional neural networks. arXiv:1511.08458

Pan F, Zhao Q, Yan Y (2005) Pronunciation dictionary adaptation based accent modeling for large vocabulary continuous speech recognition. Comput Eng Appl 41(23):4–6

Pan J, Liu C, Wang Z, Hu Y, Jiang H (2012) Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why DNN surpasses GMMs in acoustic modeling. In: Paper presented at the 8th international symposium on Chinese spoken language processing, 5–8 December 2012 (2012)

Passricha V, Aggarwal RK (2020) A hybrid of deep CNN and bidirectional LSTM for automatic speech recognition. J Intell Syst 29(1):1261–1274

Prabakaran D, Shyamala R (2019) A review on performance of voice feature extraction techniques. figshare. https://doi.org/10.1109/ICCCT2.2019.8824988

Pratap V, Tjandra A, Shi B, Tomasello P, Babu A, Kundu S, Elkahky A, Ni Z, Vyas A, Fazel-Zarandi M, et al (2023) Scaling speech technology to 1,000+ languages. arXiv:2305.13516

Qian H (2016) A description of the phonetic system of Jintan dialect. J Wuxi Inst Commer 16(5):105–112

Qimike B, Huang H, Wang X (2015) Uyghur speech recognition based on deep neural network. Comput Eng Des 36(8):2239–2244

Rabiner L, Juang B-H (1993) Fundamentals of Speech Recognition. Prentice-Hall Inc, Hoboken

Ramırez J, Segura JC, Benıtez C, De La Torre A, Rubio A (2004) Efficient voice activity detection algorithms using long-term speech information. Speech Commun 42(3–4):271–287

Rao B (2007) Guangzhou Sound Dictionary. Guangzhou dictionary

Reddy CK, Gopal V, Cutler R, Beyrami E, Cheng R, Dubey H, Matusevych S, Aichner R, Aazami A, Braun S, et al (2020) The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. arXiv:2005.13981

Ren Z, Yang G, Xu S (2019) Two-stage training for chinese dialect recognition. arXiv:1908.02284

Reynolds DA (2009) Gaussian mixture models. Encycl Biom 741:659–663

Robinson T, Fransen J, Pye D, Foote J, Renals S (1995) WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition. figshare. https://doi.org/10.1109/ICASSP.1995.479278

Rouzi A, Shi Y, Zhang Z, Wang D, Hamdulla A, Zheng F (2017) THUYG-20: A free Uyghur speech datanase. J Tsinghua Univ 57(2):182–187

SUN L (2020) Using prosodic and acoustic features for Chinese dialects identification. In: 2020 2nd international conference on image processing and machine vision, vol 6, ACM, pp 118–123

Salesky E, Wiesner M, Bremerman J, Cattoni, R, Negri M, Turchi M, Oard DW, Post M (2021) The multilingual tedx corpus for speech recognition and translation. arXiv:2102.01757

Santana Correia A, Colombini EL (2022) Attention, please! a survey of neural attention models in deep learning. Artif Intell Rev 55(8):6037–6124

Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673–2681

Senin P (2008) Dynamic time warping algorithm review. Inform Comput Sci Dep Univ Hawaii Manoa Honolulu USA 855(1–23):40

Shao X, Ma H (2020) The functions of dialects and its English translation based on Gao Xing by Jia Pingwa. J Xi'an Int Stu Univ 28(02):104–109

Shi X (2006) A systematic representation of the vowel patterns of Chinese dialects. Dialect 26(4):323–331

Shi J, Huang W (2016) Sichuan dialect speech recognition based on deep neural network. Mod Comput 2016(9):3–6

Shivaprasad S, Sadanandam M (2021) Dialect recognition from Telugu speech utterances using spectral and prosodic features. Int J Speech Technol 4(23):1–10

Shon S, Ali A, Glass J (2018) Convolutional neural networks and language embeddings for end-to-end dialect recognition. arXiv:1803.04567

Social Sciences CA (2012) Chinese language atlas. Commercial Press, Shanghai

Sun J, Wushouer S, Reyiman T, Zhang J (2019) Acoustic analysis and language recognition of Uygur. Acta Acust 06(44):1083–1092

Tang M (2013) Phonological investigation of luoyang dialect. Youth Literator. 2013(11X):2

Tian F (2009) Two striking books in Changsha dialect: exegetical harmonics and Xiang Yin Jian Zi. Lexicogr Stud 9(1):136–144

Tuerxun T, Dai L (2015) Deep neural network based Uyghur large vocabulary continuous speech recognition. J Data Acquis Process 30(2):365–371

Tyagi H, Kumar V, Danish M, Agarwal G, Mishra P (2023) Speech Recognition Intelligence System for Desktop voice Assistant by using AI &IoT. International Journal of Intelligent Systems and Applications in Engineering, 11(5s): 266-272.

Wan M, Ren J, Ma M, Li Z, et al (2022) Deep neural network based chinese dialect classification. In: 2021 Ninth international conference on advanced cloud and big data (CBD), vol 25, IEEE, pp 207–212

Wan M, Ren J, Ma M, Li Z, Cao R, Gao Q (2022) Deep neural network based chinese dialect classification. In: 2021 ninth international conference on advanced cloud and big data (CBD), IEEE, pp 207–212

Wang K (2001) Uighur speaker-independent speech recognition based on cdcpm. J Comput Res Dev 38(10):1242–1245

Wang Q, Guo W, Xie C (2017) Towards end to end speech recognition system for Tibetan. Pattern Recognit Art Intell 30(4):359–364

Wang G, Pang B, Li C, Yang D (2020) An evaluation of Xunfei speech input software in the COVID-19 pandemic prevention. Chin J Lang Policy Plan 5(5):48–56

Wang Q, Qian S, Zhao X (2009) Hunan dialects identification based on GMM and difference speech feature. Comput Eng Appl 45(35):129–131

Wang D, Ye S, Hu X, Li S, Xu X (2021) An end-to-end dialect identification system with transfer learning from a multilingual automatic speech recognition model. Figshare. https://doi.org/10.21437/Interspeech.2021-374

Wang Z, Zhao Y, Wu L, Bi X, Dawa Z, Ji Q (2022) Cross-language transfer learning-based lhasa-tibetan speech recognition. CMC Comput Mater Continua 73(1):629–639

Wang, T., Li, A. (2003). Design of continuous Chinese speech recognition corpus. In: Paper presented at the 6th national conference on modern phonetics vol 2, pp 18–20

Wang N (2020) "Black Box Justice": Robot Judges and AI-based Judgment Processes in China's Court System. Paper presented at 2020 IEEE international symposium on technology and society (ISTAS), 12–15 November 2020 (2020)

Wang D, Ye S, Hu X, Li S, Xu X (2021) An end-to-end dialect identification system with transfer learning from a multilingual automatic speech recognition model. In: Interspeech, pp 3266–3270

Wang C, Riviere M, Lee A, et al (2021) Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. arXiv:2101.00390

Wani TM, Gunawan TS, Qadri SAA, Kartiwi M, Ambikairajah E (2021) A comprehensive review of speech emotion recognition systems. IEEE Access 9:47795–47814

Wong E, Sridharan S (2001) Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification. In: Proceedings of 2001 international symposium on intelligent multimedia, video and speech processing. ISIMP 2001 (IEEE Cat. No. 01EX489), IEEE, pp 95–98

Wu G (2012) Tuerhongjiang: research and implementation of speech recognition about Uyghur in southern Xinjiang. J Tarim Univ 24(3):51–55

Wu G, Liu F (2013) Research of pre-processing about Uyghur in Southern Xinjiang for speech recognition. figshare. https://doi.org/10.1109/ICCIS.2013.73

Xie J et al (2011) The survey of the current situation of putonghua popularization. Appl Linguis 79(3):2–10

Xie X, Sui X, Liu X, Wang L (2022) Investigation of deep neural network acoustic modelling approaches for low resource accented mandarin speech recognition. arXiv:2201.09432

Xu F, Dan Y, Yan K, Ma Y, Wang M (2021) Low-resource language discrimination toward Chinese dialects with transfer learning and data augmentation. Trans Asian Low Resour Lang Inform Process 21(2):1–21

Xu F, Yang J, Yan W, Mingwen W (2021) An end-to-end dialect speech recognition model based on self attention. J Signal Process 37(10):1–14

Xu F, Wang M, Li M (2018) Building parallel monolingual Gan Chinese dialects corpus

Xu B, Hong Q, Li B, Zhen D, Li L, Xiao L (2017) The design and transcription of corpus based on the technology of speech recognition for southern min dialects. In: Proceedings of the 14th national conference on man-machine speech communication (NCMMSC'2017)

Yadav H, Sitaram S (2022) A survey of multilingual models for automatic speech recognition. arXiv:2202.12576

Yang L, Guo W, Han F (2022) Chinese automatic speech recognition based on DFCNN-CTC and transformer. Fire Control and Command Control 47(3):16–21

Yang W, Hu Y (2021) Hybrid CTC/attention architecture for end-to-end multi-accent Mandarin speech recognition. Appl Res Comput 38(03):755–759

Yang J, Li H, Zhang X (2017) On the construction of a Bai speech corpus. J Dali Univ 2(12):21–24

Yang H, Ling Q, Guo W, Li J, Chen L (2009) A Lanzhou dialect corpus for speech engineering. J Northwest Norm Univ 45(6):54–59

Yao X, Li Y, Shan G, Yu H (2009) Research on Tibetan isolated-word speech recognition system. J Northwest Univ Natl Nat Sci 30(1):29–36

Ye S, Li C, Zhao R, Wu W (2019) NOAA-LSTM: A new method of dialect identification. In: International conference on artificial intelligence and security, Springer, pp 16–26

Ye X (2011) A typology study of Chinese dialect phonetics. PhD thesis, Fudan University

Ying W, Zhang L, Deng H (2020) Sichuan dialect speech recognition with deep LSTM network. Front Comp Sci 14(2):378–387

Yu C, Kang M, Chen Y, Wu J, Zhao X (2020) Acoustic modeling based on deep learning for low-resource speech recognition: an overview. IEEE Access 8:163829–163843

Yu, L. (2019) Speech recognition of Hakka dialect based on deep learning. Master's thesis, South China University of Technology

Yu T, Frieske R, Xu P, Cahyawijaya S, Yiu CT, et al (2022) Automatic speech recognition datasets in cantonese: A survey and new dataset. Paper presented at the 13th language resources and evaluation conference, 2022

Yuan J (1960) Outline of Chinese dialects. Language and Culture Press, Beijing

Yue AO (2003) Chinese dialects: grammar. In: Thurgood G, Lapolla RJ (eds) The Sino–Tibetan languages. Routledge London, New York, pp 84–125

Zaharia GE, Avram AM, Cercel DC, Rebedea T (2021) Dialect identification through adversarial learning and knowledge distillation on romanian bert. In: Proceedings of the Eighth Workshop on NLP for Similar languages, varieties and dialects, association for computational linguistics, Kiyv, Ukraine, pp 113–119

Zhan B (2000) A review on the studies of Chinese dialects in the past two decades. Fangyan (Dialect) 4(8):317–324

Zhang B (1909) Chinese new dialects. Zhejiang Publishing House, Hangzhou

Zhang S (1981) The phonetic system of Chaoyang dialect. Dialect 11(1):13

Zhang Y (2007) A Study of Nanchang Dialect. CHINESE NATIONAL ACADEMY OF ARTS, Hangzhou

Zhang C, Wei P, Lu X, Shi X (2018) Design and implementation of speech recognition system in Chongqing dialect. Comput Meas Control 26(1):256–259. https://doi.org/10.16526/j.cnki.11-4762/tp.2018.01.063

Zhang S, Zhao F, Huang J, Liu Q (2021) The influence of Mandarin accent on the listener's attitudes and behaviors in ethnic minority areas. J Res Educ Ethn Minor 21(3):111–118

Zhang F, Xie X, Quan X (2022) Chinese Dialect Speech Recognition Based on End-to-end Machine Learning. Paper presented at 2022 international conference on machine learning, control, and robotics (MLCR), October 2022

Zhao Y (1980) A system of "Tone-Letters''. Fangyan (Dialect) 11(2):81–83

Zhao Y, Yue J, Song W (2019) Others: Tibetan multi-dialect speech recognition using latent regression Bayesian network and End-to-End mode. J Internet Things 1(1):17

Zhao Y, Yue J, Song W, Xu X, Li X, Wu L, Ji Q (2019) Tibetan multi-dialect speech and dialect identity recognition. Comput Mater Contin 60(3):1223–1235

Zheng Y, Sproat R, Gu L, et al (2005) Accent detection and speech recognition for shanghai-accented mandarin. In: Paper presented at the 9th European conference on speech communication and technology, 4–8 Septermber 2005

Zhiyun C (2015) On the orientations, objectives and missions of the project for protecting language resources of China. Appl Linguis 15(4):10–17

Zhou K, Li A, Yin Z, Zong C (2010) CASIA-CASSIL: a Chinese Telephone Conversation Corpus in Real Scenarios with Multi-leveled Annotation. LREC, May 2010

Zhu X (2019) CaSe i: iflytek: a technology innovator's journey from intelligent speech to artificial intelligence. In: Emerging champions in the digital economy: new theories and cases on evolving technologies and business models, Springer, Singapore, pp 67–89

Zissman MA, Gleason TP, Rekart DM, Losiewicz BL (1996) Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech. In: 1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings, vol. 2, IEEE, pp 777–780

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.