# Speech recognition for different dialects and accents

*Anrui* Wang*

Jinshan World Foreign Language School, 200000, Shanghai, China

**Abstract.** China is a populous nation made up of many different ethnic groups. Specifically, due to multiple factors, dialects in different regions of China exhibit notable differences in speech characteristics, intonation, and vocabulary. As a result, research progress and practical applications in dialect speech recognition face an imbalanced situation. Therefore, exploring specific recognition methods, establishing diverse dialect corpora, and investigating regional heterogeneity within different dialects are crucial for enhancing the accuracy and applicability of Chinese speech recognition. This paper sorts out the key technologies related to dialect speech model, integrates the deep neural network and Supervised learning of the model. In addition, data enhancement and adaptation methods of various model techniques, attention mechanism, end to end System are also introduced. These techniques can effectively improve the performance of the model in different dialect environments. Moreover, the current limitations of the field will be discussed, such as the lack of accuracy in identifying certain dialects and the challenges in data collection and processing. By analyzing these issues, this research aims to propose potential solutions for the further development of dialect speech recognition technology, offering valuable reference material for researchers and developers.

## 1 Introduction

China is a country with many accents and dialects. Due to the differences in pronunciation methods and vocabulary spellings, dialects differ significantly from Mandarin. Accents cause variations in the pronunciation of Chinese words, altering initials and finals. Therefore, speech recognition is often less accurate when recognizing dialects [1].

Dialects are defined as variations in language that differ by geographic regions and social groups. These variations can be distinguished by phonological, grammatical, and lexical features [2]. There are many reasons why dialects lead to inaccuracies in speech recognition. First, there are differences in acoustic features; the pronunciation and tone of dialects often differ from standard language, resulting in changes to acoustic characteristics. Acoustic features are extracted from audio signals and serve as the foundational data for training speech recognition models. Dialects can alter the audio features of pronunciation, such as frequency, intensity, and prosody, creating challenges for model training. There may also be phonemes that differ from those in the standard language, making it difficult for the model

* Corresponding author: fengshanwei@ldy.edu.rs

to recognize speech. Additionally, the intonation and prosody of dialects may differ from Mandarin, affecting rhythm and stress patterns in speech. Moreover, datasets for dialects and accents are usually relatively small, leading to an imbalance in samples when training speech recognition models. The different vocabulary, grammar, and expressions used in dialects may also result in the standard language model being unable to effectively handle these dialectal characteristics. Speech recognition models are typically optimized on training data, which may perform well in standard language contexts but lack generalization ability when dealing with dialects, and this is another reason affecting recognition accuracy. The existence of various dialects presents significant challenges for speech recognition work. In fact, due to factors such as geographic location and population migration, research and practical applications of speech recognition for Chinese dialects are at different stages. Therefore, exploring specific recognition methods and their effectiveness, as well as the significant regional heterogeneity of dialect corpora, is of great importance for Chinese speech recognition [3, 4, 5].

This paper first systematically introduces the key technologies of dialect speech recognition, from basic technologies to advanced applications, covering deep neural networks, supervised learning, data augmentation and adaptation, attention mechanisms, end-to-end systems, and so on. Then, the existing problems in the field of dialect speech recognition and their corresponding solutions are deeply discussed. Finally, a conclusion is drawn. It is hoped that this paper will serve as a useful resource for pertinent researchers and developers, aid in the advancement of dialect speech recognition technology, and advance the fields of artificial intelligence and natural language processing.

## 2 Key Technologies in dialect speech recognition

### 2.1 Deep neural network

Automatic speech recognition accuracy has greatly improved with the advent of hybrid deep neural network (DNN) models. Building these models usually requires a number of different components, including language, vocabulary, and acoustic models. While this complex architecture can improve recognition precision, it also introduces cumbersome and varied training processes. Each component must be individually fine-tuned and optimized to ensure optimal performance within the overall system. This requirement for independent adjustments not only extends development time and increases resource consumption but can also lead to cumulative errors during the integration of the different components. These cumulative errors can negatively impact the final recognition outcomes and may compromise the system's stability and robustness in certain environments. Therefore, although hybrid DNN models offer new possibilities for automatic speech recognition technology, effectively coordinating and optimizing these components remains a critical challenge in practical applications.

### 2.2 Supervised learning

Supervised learning relies on labeled data for classification or regression, identifying patterns by establishing functional relationships between inputs and outputs. However, its limitation lies in learning from a finite set of labeled data, failing to fully utilize abundant unlabeled data. Self-supervised learning, on the other hand, is a more efficient way to pre-train seque-to-sequence (seq2seq) models by extracting relevant and broadly applicable underlying representations from large amounts of unlabeled data. This approach sets auxiliary tasks that

allow the network to gain representations beneficial for downstream tasks, thus uncovering latent knowledge in unsupervised data.

The shortage of labeled data is a common problem in dialect speech recognition, so how to effectively utilize supervised learning techniques is also an aspect that needs attention. Multi-scale feature fusion and multi-view self-supervised learning are used in an inventive end-to-end voice recognition model presented by Zhao et al. [6]. This new model uses a hybrid training approach that combines state-of-the-art self-supervised techniques with classic supervised learning in order to improve the model's voice data representation. The fundamental principle of the concept is to leverage inter-layer information from a shared encoder to improve feature representation. The model is better able to convey and capture the intricate features of speech data by making use of the diversity of this data. Furthermore, multi-view self-supervised learning is integrated to increase data consumption efficiency and improve model robustness. To be more precise, this is accomplished by developing several shared encoder sub-models, each of which excludes specific data during training in order to facilitate the use of multi-perspective data and further optimize model performance. Multiple conformer blocks make up the shared encoder, which is capable of efficiently learning both local and global characteristics from the input voice sequence. In order to create the final output representation, the outputs of different conformer blocks are given different weights, which are then combined by the multi-scale feature fusion module (MFF). Eventually, each conformer block's outputs are concatenated to create a comprehensive speech representation, guaranteeing the successful integration of multi-layer features.The resulting output representation is processed by the model using an Attention decoder and Connectionist Temporal Classification (CTC) during the decoding stage. The research team used the WeNet speech recognition program as a benchmark and carried out methodical training and testing on the Aishell-1 data set in order to assess the performance of the suggested model. After that, an additional assessment was conducted on the WSJ English corpus. According to the experimental findings, the model dramatically decreased the character error rate (CER) and showed appreciable gains in speech recognition accuracy after implementing four distinct decoding strategies.These outcomes not only attest to the efficacy of the suggested end-to-end speech recognition paradigm but also showcases its immense potential in enhancing speech recognition accuracy and overall performance, marking a significant advancement in this field. Through this innovative design, future speech recognition technologies are expected to achieve higher accuracy and reliability in more complex application scenarios.

### 2.3 Data enhancement and adaptive

In speech recognition, data augmentation and domain adaptation are crucial techniques for improving system performance and generalization ability. They help the model more effectively handle speech data under various real-world conditions and enhance its adaptability to specific domains or environments. Through the conversion of speech signals into text, Automatic Speech Recognition (ASR) technology is essential to the facilitation of human-computer interaction [7]. Advancements in deep learning-based ASR systems have been substantial in recent years, leading to major improvements in their processing power and accuracy. These systems can better understand natural language and perform real-time recognition in various environments, significantly enhancing user experience. However, as the demand for ASR models continues to grow, users' expectations for accuracy, stability, and robustness have also increased. This means that ASR technology must maintain high recognition performance under different accents, background noise, and speaking styles. Additionally, handling recognition challenges related to multiple languages and dialects remains a major research hurdle. Therefore, despite the immense potential for ASR

technology, comprehensive satisfaction of diverse user needs still requires in-depth exploration and research in areas such as algorithms, datasets, and model optimization.

### 2.4 Attention mechanism

Attention mechanism plays an important role in improving the performance of speech recognition and dealing with dialect variants, especially when modeling long-distance dependencies. Liu et al. [8] proposed the integration of attention mechanisms within a two-layer Convolutional Neural Network (CNN) to model the acoustic features, resulting in a Datong dialect speech translation model. Zhang et al. [9] created a far-field speech recognition model that integrates attention Long Short-Term Memory (LSTM) network with multi-task learning. This model incorporates attention mechanisms into the LSTM framework, enabling it to automatically learn the weights of context features while optimizing the prediction of three-factor state posterior probabilities for far-field speech recognition. Research indicates that attention-based acoustic models can yield effective results in speech recognition.

### 2.5 End-to-end system

The end-to-end concept streamlines the processing procedure and aids in raising the precision and effectiveness of dialect voice recognition as compared to the conventional modular approach. Li et al. [10] developed a system for Chinese dialect voice recognition (DSR). These systems can be thought of as a particular kind of broad Chinese speech recognition system because they are made to recognize and transcribe different dialects with accuracy. These systems make use of standard speech recognition technology, but they give more weight to the distinctive phonetic traits, tonal qualities, lexicon, and other pertinent aspects of regional accents. Chinese DSR systems are able to efficiently accomplish some recognition tasks by integrating these unique dialect elements into the recognition process. As a result, Chinese voice recognition systems can be thought of as a subset of Chinese DSR systems. Dialect feature extraction is used in this model, The aim of dialect feature extraction is to analyze the input audio to obtain a sequence of feature vectors, thereby retaining practical information from the dialect. In dialect recognition, the primary acoustic features rely on auditory signal representations. Common methods for extracting acoustic features of Chinese dialects include three main techniques: Mel-frequency cepstral coefficients (MFCC), linear predictive cepstral coefficients (LPCC), and perceptual linear prediction coefficients (PLP). In end-to-end speech recognition systems, the CTC-Attention model combination has emerged as a critical technique. This model trains with both CTC and Attention targets, utilizing a multi-task learning framework. An attention decoder, a CTC layer, and a shared encoder make up the architecture. In order to improve the model's capacity to extract pertinent information, the shared encoder makes use of transformer or conformer technology to efficiently learn the local and global properties of input speech sequences. Nevertheless, there are issues with extracting supervised information from massive volumes of unsupervised data for current end-to-end speech recognition models. These models ignore the information from intermediate levels and concentrate mostly on the output properties of the encoder's final layer. This restriction offers space for enhancing the robustness, data use, and features of the model. Research in these areas is ongoing and offers potential to increase recognition performance across a variety of applications and strengthen and optimize models that can better exploit unstructured data and progress end-to-end speech recognition systems.

# 3 Existing limitations and future prospects

## 3.1 Model training difficulty

In order to explore the current limitations of models and their future development directions, this analysis uses the Transformer as an example. The Transformer has garnered interest in end-to-end voice recognition architectures recently because to its exceptional performance in a variety of natural language processing workloads. Research indicates that deep Transformer architectures can significantly enhance the performance of speech recognition models. However, when constructing speech recognition models based on deep Transformers, training difficulties often arise, leading to models getting trapped in local optima.

The end-to-end encoder-decoder speech recognition architecture utilizes the Transformer model with self-attention mechanisms. This approach improves the accuracy of speech recognition and accelerates training speed through multi-head self-attention mechanisms and parallel computing. Nevertheless, this architecture still relies on a large number of parameters. As the number of layers in the model increases, the computational overhead of the self-attention mechanism also rises. This situation renders deployment on portable devices less practical and significantly reduces inference speed. Therefore, despite the substantial progress made by Transformers in the field of speech recognition, challenges related to model complexity must still be addressed in practical applications. Enhancing adaptability and practicality in resource-constrained environments, such as mobile devices, remains crucial. In order to fully leverage Transformers' potential in a wider range of scenarios, future research can concentrate on optimizing model topologies, lowering the number of parameters and computational complexity, as well as enhancing inference efficiency.

## 3.2 Model data set acquisition

In addition to the requirements and limitations related to model training, the insufficient scale of data sets is also a significant issue currently faced. Chinese dialect corpora often have a modest scale. To increase the size of the data set, it is crucial to investigate data augmentation techniques. These methods can be implemented through various technical means, such as speed variation and the addition of background noise, which enhance the diversity and overall quality of the data. Investigating techniques built around feedback systems is also important. Based on the outcomes of the feedback that has been gathered, these tactics can successfully aid in expanding the data collection and facilitating iterative changes. Researchers such as Escobar-Grisales [11] have integrated feedback techniques with voice recognition. Through the use of mobile applications, they have gathered feedback from certain user groups in real-world application scenarios, enabling ongoing optimization and modification of the recognition system. The data collected can serve not only as additional training resources but also provide more accurate and diverse training information for the model. By integrating data augmentation techniques with feedback mechanisms, researchers can significantly enhance the adaptability and performance of the model. This approach can address the limitations of insufficient data set size and allow the model to perform better across different dialects and real-world usage environments. Future research can further explore how to effectively combine these techniques to achieve more precise and robust speech recognition systems.

## 4 Conclusions

Dialects are difficult to be accurately understood by speech recognition systems because of their unique pronunciation, vocabulary and grammatical structure. Dialects in different regions differ significantly in pronunciation, and common vocabulary differs greatly from that of Mandarin, leading to frequent misjudgments. Thus, research and optimization of dialect speech models are necessary to raise the system's recognition performance. This paper introduces some existing key technologies related to dialect speech recognition and the existing models developed by predecessors. According to the characteristics of different dialects, the data collection and model training are also summarized by this paper. However, the robustness of existing models is still limited in the face of dialect diversity and noisy environments. Looking forward to the future, this paper hopes that more researchers can improve the accuracy and applicability of dialect speech recognition by developing more extensive and diversified data sets and adopting more advanced algorithms, and train more adaptable models with more extensive databases to make speech recognition play a greater role in the field of dialects.

## References

1. R. Sproat, L. Gu, J. Li, et al., Dialectal Chinese speech recognition, in Proceedings of the CLSP Summer Workshop, Johns Hopkins University, Baltimore, (2004)
2. Q. Li, Q. Mai, M. Wang, et al., Chinese dialect speech recognition: a comprehensive survey. Artif. Intell. Rev. **57**, 25 (2024)
3. W. Du, Y. Maimaitiyiming, M. Nijat, et al., Automatic speech recognition for Uyghur, Kazakh, and Kyrgyz: An overview. Appl. Sci. **13**, 326 (2022)
4. A. Rahman, M.-M. Kabir, M.-F. Mridha, et al., Arabic speech recognition: Advancement and challenges. IEEE Access. **12**, 39689 - 39716 (2024)
5. S. Singh, M. Singh, V. Kadyan, Speech recognition transformers: Topological-lingualism perspective. arXiv preprint arXiv:2408.14991 (2024)
6. J. Zhao, R. Li, M. Tian, et al., Multi-view self-supervised learning and multi-scale feature fusion for automatic speech recognition. Neural Process Lett. **56**(4), 168 (2024)
7. X. Huang, A. Acero, H.-W. Hon, et al., Spoken language processing: A guide to theory, algorithm, and system development. (2001)
8. X. Liu, W. Song, B. Yu, et al., Research on attention-based speech translation model of Datong dialect. J. North Univ. China (Nat. Sci. Ed.). **41**, 238-243, 248 (2020)
9. Z. Yu, P. Zhang, Y. Yan, Long short-term memory with attention and multitask learning for distant speech recognition. J. Tsinghua Univ. Sci. Technol. **58**(3), 249-253 (2018)
10. Q. Li, Q. Mai, M. Wang, et al., Chinese dialect speech recognition: a comprehensive survey. Artif. Intell. Rev. **57**, 25 (2024)
11. D. Escobar-Grisales, C. Rios-Urrego, J. Gallo-Aristizabal, D. López-Santander, N. Calvo-Ariza, E. Nöth, J. Orozco-Arroyave, Colombian dialect recognition from call-center conversations using fusion strategies, in Proceedings of the Workshop on Engineering Applications, Springer, (2022), 54-65
12. D. Escobar-Grisales, C.-D. Rios-Urrego, J.-D. Gallo-Aristizabal, et al., Colombian dialect recognition from call-center conversations using fusion strategies, in Proceedings of the Workshop on Engineering Applications, Cham: Springer Nature Switzerland, (2022), 54-65