Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/52/20241518

A comparative study of machine learning-based Chinese dialect speech recognition

Xinrui Xi

College of Letters and Science, University of Wisconsin-Madison, Madison, Wisconsin, 53715, United States

xxi9@wisc.edu

Abstract. As the official language of China, Mandarin is used on all formal occasions. However, due to China's rich history and vast population, dialects have become the primary language spoken by many people. While not taught in school, dialects serve as unofficial languages in China and are passed down through word of mouth. In recent years, the use of dialects has significantly declined as people have adapted to using official Mandarin in all settings. However, among the elderly, dialects remain the primary language of communication. The great need to use dialects as the primary mode of communication has laid the foundation for the importance and high demand for dialect speech recognition. Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, research on this topic is further analysed and compared. The results suggest that there has been a certain degree of development and progress in dialect speech recognition technology. However, further research is needed to overcome the limitations of low-resource language speech recognition.

Keywords: Chinese Dialects, Machine Learning, Speech Recognition.

1. Introduction

Due to multiple factors such as growth environment and societal development, dialects have replaced Mandarin as the first communication language for many elderly people in China. Recently, especially in the era of rapid technological development, public information transmission, and communication have been shared in the form of Mandarin [1]. Under normal circumstances, elderly people can satisfy their needs using their comfort languages and having real-life communication. However, during the three years of the epidemic, face-to-face communications were not suggested but transferred to online devices. Reaching out for help became difficult for them, with some supplement problems of technology or personal issues, meeting requirements of medical and food supply were less possible.

Even in the post-epidemic period when offline activities were retrieved, due to the development of science and technology, public resources functions are gradually transferred online, bringing convenience to more people. However, circumstances are different for Chinese elders. Reaching help through customer services can't efficiently solve their problems. Customer service robots were invented to replace the original button transfer function. People are required to contact the robots first to be transferred to live agents. Unlike human, robots' built-in Mandarin voice datasets are not able to precisely recognize dialects, non-standard spoken tones of Mandarin, or abnormal recording environments, which reduces the chance of meeting their needs [2].

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/52/20241518

Artificial intelligence technology has been integrated into people's lives, and derivative technologies such as machine learning have been widely used in various real-life scenarios.

Automatic speech recognition is a technology that converts audio into written text for further analysis, processing, and storage [3]. This technology has been applied in various public and private resources applications, such as family, medical, security, education, entertainment, etc. Automated Speech Recognition consists of acoustic modeling (AM) and language modeling (LM) The NN-FIMM framework is its production standard [4]. Cutting-edged accuracy and efficiency in real-time tasks were shown in hybrid ASR system and its standard test like LibriSpeech [4]. Nonetheless, the accuracy tends to decline when the provided speech is accented or foreign language, especially in the code-switching (CS) scenario.

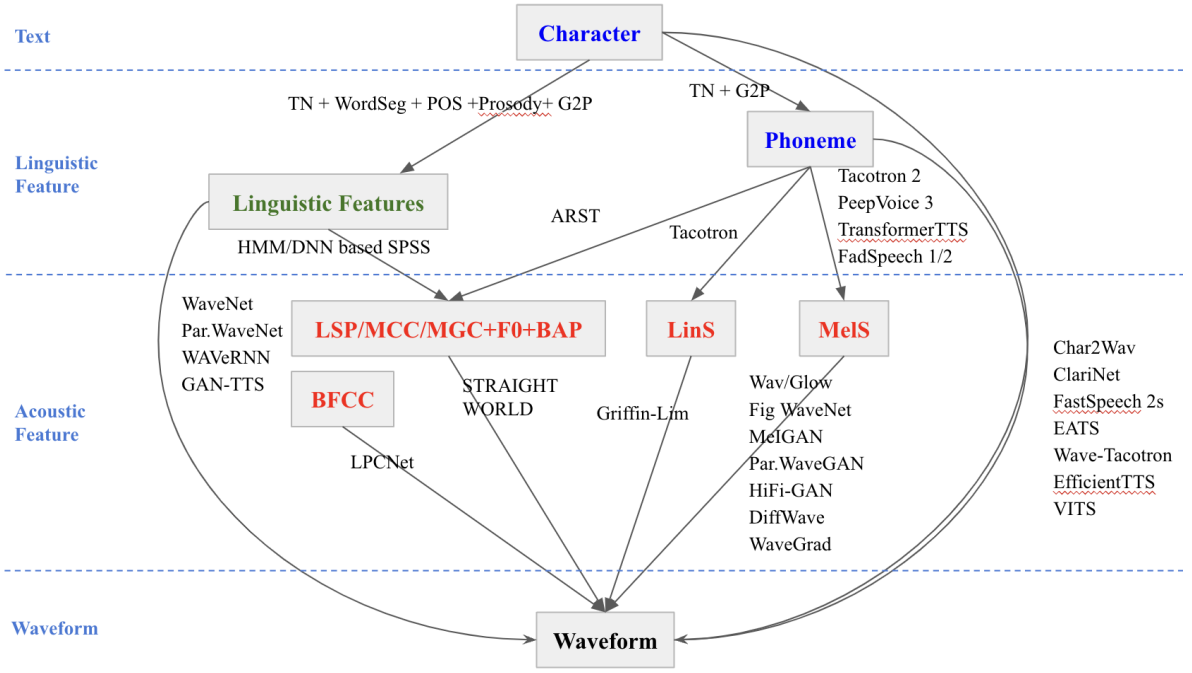

Figure 1. Techniques of text-to-speech synthesis models (Figure Credits: Original).

Figure 1 depicts a more complex text-to-speech synthesis architecture. The system starts from "text" and extracts language features through methods such as "TN + WordSeg + POS + G2P", resulting in two different paths: "characters" and "phonemes". These representations undergo another layer of transformation, focusing on acoustic features. These features are extracted using methods such as “LSP/MCC/MGC+FO+BAP” and “BFCC”. The next stage involves converting these acoustic signatures into "waveforms" or sound waves. Several modern technologies and techniques, including "WaveNet", "MelGAN" and "Tacotron", among others, are depicted, demonstrating various approaches to this transformation. The entire process from text to waveform illustrates the layered complexity and multiple approaches involved in modern text-to-speech systems.

This article provides an overview of various recognition techniques, offering a comprehensive description of their operational principles, and demonstrating their real-world applications in learning and education through illustrative examples. Furthermore, the paper investigates the methodologies of relevant studies, encompassing participant selection, data collection procedures, and study designs, and delves into the advantages and drawbacks of these identification techniques. This information is highly valuable for educators and researchers, particularly those lacking a technical background, as it facilitates a deeper understanding of the inner workings of identification technologies, their practical utility, and research possibilities, as well as their strengths and limitations. Moreover, the insights presented in this

DOI: 10.54254/2755-2721/52/20241518

review paper serve as a valuable resource for steering future research and development efforts in recognition technology within the education sector.

One of the primary challenges in developing dialectal speech conversion technology is the scarcity of adequate data. China boasts 737 distinct dialects [5], Cheng classifies these dialects into two broad categories. Northern Mandarin (291 out of 318) was found to have four tones, while southern dialects, Wu, Xiang, Gan, Hakka, Hokkien, and Cantonese (145 out of 173) had five to ten tones [5].

Despite the significant advancements in modern technology within the realms of natural language processing and speech recognition, the acquisition and processing of data about China's myriad dialects remains an arduous undertaking. The disparities between the northern and southern Chinese dialects are notably pronounced. Northern dialects, typically epitomized by Mandarin, are distinguished by their precise level of enunciation with reduced stress. Conversely, the southern dialects exhibit more diverse pronunciation, accentuated stress patterns, and a rhythm characterized by cadence. Compounding this complexity, each dialect region may encompass multiple subdialects, amplifying the intricacy of data acquisition.

2. Method Analysis

When studying and modeling a new dialect, having an adequate speech database is crucial to building an effective language model. During this process, the careful screening, filtering, and selection of data become particularly important. The quality of data directly influences the quality of the training model and impacts computational costs. Several methods are used to process the filtering and selection.

The Klakow method scores each text paragraph by separately removing it and calculating the logarithmic probability of the development data, compared to the logarithmic probability of a language model trained on all the training data. This approach is suitable for situations with very limited development data because all models are estimated from training data, which alleviates the problem of out of vocabulary vocabularies or subwords. The method is devised to enhance the model's clarity (or logarithmic probability) when applied to the filtered data during the evaluation on the development dataset. In a straightforward implementation, each text paragraph undergoes an assessment by removing it from the training data, training a language model, and computing the logarithmic probability concerning the development dataset. This computed logarithmic probability is then juxtaposed with the logarithmic probability derived from the language model trained on the complete training dataset, with the difference serving as the score for the text paragraph [6].

The xe-diff method is an approach proposed by Moore and Lewis that uses two language models, one estimated from development data and the other from an equal amount of unfiltered training data. The score of the text paragraph is the difference in cross-entropy between the two models. This method only needs to calculate the probability of two language models for each text paragraph, so the running time is proportional to the number of words in the unfiltered training data [6].

The core idea of the devel-re method proposed by Sethy et al. is to filter the distribution of data and develop data through relative entropy matching. First, a language model is estimated from the development data, and then the same amount of unfiltered training data is used to initialize the model of the selected set. Then, the text paragraphs are processed sequentially to calculate how much relative entropy changes would occur if a paragraph were included in the selection set relative to developing the data model. If the change is negative, include the text paragraph and update the selection set model. The running time of this method is proportional to the number of words in the unfiltered training data [6].

Speech synthesis technology plays a vital role in addressing the diversity of Chinese dialects. When network and database resources are limited or inaccurate, speech synthesis technology can effectively collect a vast amount of data. Hidden Markov models, a type of dynamic Bayesian network, are employed to generate speech parameter sequences, allowing for the flexible synthesis of speech in dialects with limited resources [7]. Additionally, the conversion of written characters into their corresponding phonetic representations, often referred to as Grapheme-to-Phoneme (G2P) conversion, is a task that involves translating a series of written characters into their phonetic counterparts. Automated G2P tools frequently facilitate this conversion process [8]. Mandarin's phonetic transcription

DOI: 10.54254/2755-2721/52/20241518

system encompasses various schemes, including Chinese Pinyin, Bopomofo, Mandarin note second form, Witoma Pinyin, and more. These schemes provide different means of representing Mandarin pronunciation. Furthermore, some schemes incorporate Latinization, using the Latin alphabet to represent the pronunciation of Chinese characters. To enhance the efficiency and interoperability of the symbol system, simplified or complementary symbols may be employed. This diversity can simplify the understanding and communication of pronunciation information among different individuals or systems, although it may occasionally lead to confusion, particularly for non-native speakers. In the context of G2P used in the Chinese language, the development of rule-based text frontends is necessary. A "Chinese rule-based Text Frontend" has been established based on the syllabic inventory of Chinese Pinyin, encompassing 21 initial consonants and 41 vowels. This facilitates the accurate conversion of written text into phonetic representations. Polyphony and continuity are the most important phonological features in Chinese language. In Chinese, the combination of syllables and tones often involves multi-syllable words and continuous speech, which brings challenges to speech recognition. These rules and models will take into account variations in tone, syllable combinations, and continuous speech to more accurately capture Chinese pronunciation and meaning [9].

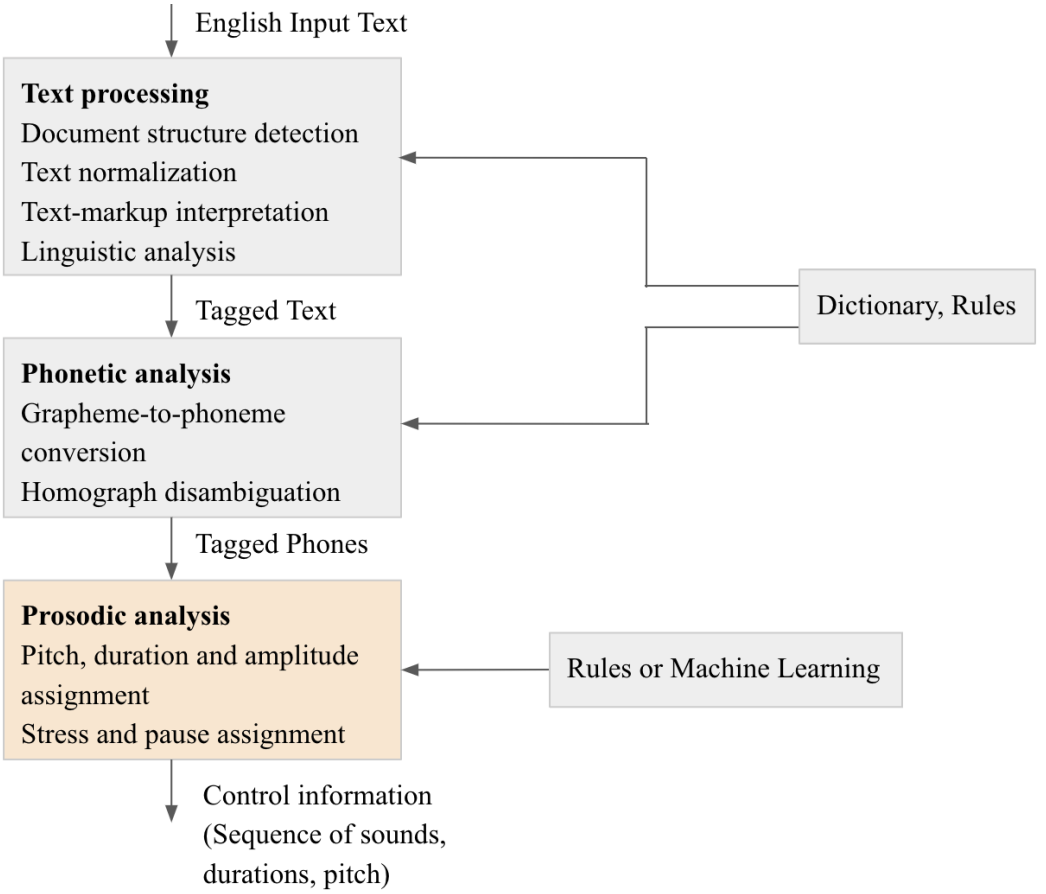

Figure 2. Representative pipeline of converting English to audible speech (Figure Credits: Original).

The task of the Text to Speech (TTS) backend is to transform various features generated by the TTS frontend, including language, phonemes, rhythm, emotional labels, and potentially speaker information in the case of multi-speaker models, into acoustic features (e.g., mel spectrograms) through model learning. The design of the corresponding model, the quantification and fusion of these features, and the timing of their incorporation all significantly affect the quality of the synthesis. Notably, the introduction of components like the Reference Encoder and global style token (GST) has greatly improved the effectiveness of multi-lingual, multi-speaker, and multi-emotion models. Due to the disparity in duration

DOI: 10.54254/2755-2721/52/20241518

between text and speech features, current deep learning models for speech synthesis fall into two categories. One involves autoregressive models, such as Tacotron, which employs an attention mechanism in a seq2seq architecture. The other category, represented by models like FastSpeech, uses a feedforward non-autoregressive network that predicts phoneme durations and expands text features into speech [10].

Figure 2 outlines a structured process for converting English text input into audible speech. Starting from "English input text", it first goes through a series of "Text Processing" steps. These steps include detecting the document structure, normalizing the text, interpreting any text markup, and performing linguistic analysis. Once processed, the text is tagged and enters the "Phonetic analysis" stage. This stage emphasizes grapheme-to-phoneme conversion and disambiguation of homophones, resulting in "Tagged Phones." After this, the "prosodic analysis" stage comes, where pitch, duration, amplitude, stress, and pauses are assigned to the speech content. The final output is "Control Information" detailing sound sequences, durations, and pitches, which are guided and influenced by external factors such as "Dictionaries, Rules" and potentially "Rules or Machine Learning".

Recognition of foreign proper names is another great challenge for researchers. Foreign Proper Name (FPN) contains proper nouns that are unconventional or irregular compared to the target language, including personal names, geographical locations, and brand names. Improving the accuracy of FPN recognition plays an important role in promoting the research of dialect speech recognition technology. The difficulty in recognizing FPNs primarily stems from challenges in pronunciation modeling and language models. To address this, a two-stage adaptation framework is implemented. Initially, it screens potential candidates for foreign words. Then, it employs a combination of letter n-gram models and topic similarity to determine the most likely foreign names, thereby improving the accuracy of FPN recognition [6].

Consequently, the development of dialectal speech recognition systems necessitates the availability of extensive speech data to ensure their accuracy and adaptability. Furthermore, the geographic dispersion of dialects in China poses a substantial challenge, as certain regional dialects may solely be prevalent within specific localized areas. Therefore, comprehensive data collection efforts are indispensable, encompassing both urban and rural regions across the nation, to facilitate the creation of speech recognition systems capable of accommodating the diverse array of dialects found throughout China.

3. Discussion

Speech recognition technologies in international languages have benefited many people, and these technologies have spread to all aspects of life. Dialect speech recognition needs further development. The large number of dialects, the lack of sound and speech data sets, especially audio sets for minority languages and different accents, background culture, and the miniaturization of information all contribute to the lag in innovation of public functions. Cheng emphasized the extensive coverage of Chinese dialects across Asia, underscoring the necessity for additional quantitative investigations [5]. This endeavor necessitates the collective involvement of multiple researchers, as conducting a comprehensive examination of all 737 dialects by a single individual is not a feasible undertaking. Further scholarly inquiry is vital to enhancing the comprehension of these linguistic variations.

The main solution to this problem is the need to attract more tech-savvy individuals from different cultures and backgrounds to participate and collaborate across disciplines. At the same time, more extensive market research should be conducted to allow teams and technicians to understand user needs more deeply, and ultimately provide greater convenience for people. As more and more speech data is collected and utilized, we can expect significant improvements in the performance of dialect speech recognition systems. This will involve richer training data related to speech features and language models, enhancing the adaptability of the system to a variety of dialects and accents. In addition, the continuous development of deep learning and neural network technology will drive the progress of dialect speech recognition. More sophisticated acoustic and linguistic models will improve accuracy while reducing reliance on large-scale annotated data. Knowledge of models trained in one dialect can

DOI: 10.54254/2755-2721/52/20241518

be more easily transferred to other dialects, speeding up the development of new language-specific speech recognition systems. However, this approach still has certain limitations. A major challenge is the diversity of dialects, with some dialect variants showing marked differences in phonetic features, thus increasing the difficulty of recognition. To sum up, the development prospect of dialect speech recognition technology is broad, which can improve our understanding and application of various spoken languages. However, it must be acknowledged that limitations still exist and continuous efforts are needed to overcome them. Developments in this area will help enable broader multilingual speech recognition, providing enhanced speech technology support for different language groups.

4. Conclusion

In this work several speech recognition approaches are analysed and compared. Numerous dialects, limited datasets, and diverse cultural backgrounds hinder public function innovation. This motivates the need for additional quantitative investigations on Chinese dialects. To solve this problem, speech data collected and utilized, dialect speech recognition systems are improved. Continuous development of deep learning and neural network technology are promising solutions of these problems. Sophisticated acoustic and linguistic models could improve accuracy while reducing reliance on large-scale annotated data, enabling knowledge transfer between dialects. However, challenges remain due to dialect diversity, with some variants showing marked phonetic differences, increasing recognition difficulty. The development prospect of dialect speech recognition technology is broad, but limitations exist and require continuous efforts to overcome. Developments will enable broader multilingual speech recognition, providing enhanced speech technology support for different language groups.

References

[1] Lu, Y., Zheng, Y., & Lin, S. (2019). Mandarin Chinese teachers across borders: Challenges and needs for professional development. International Journal of Chinese Language Education, (6), 135-168.

[2] Soh, K. W., & Loo, J. H. Y. (2021). A review of Mandarin speech recognition test materials for use in Singapore. International Journal of Audiology, 60(6), 399-411.

[3] Gokay, R., & Yalcin, H. (2019). Improving low resource Turkish speech recognition with data augmentation and TTS. In 2019 16th International Multi-Conference on Systems, Signals & Devices, 357-360.

[4] Tan, Z., Fan, X., Zhu, H., & Lin, E. (2020). Addressing accent mismatch In Mandarin-English code-switching speech recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, 8259-826.

[5] Cheng, C. C. (1991). Quantifying affinity among Chinese dialects. Journal of Chinese Linguistics monograph series, (3), 76-110.

[6] Kurimo, M., Enarvi, S., Tilk, O., Varjokallio, M., Mansikkaniemi, A., & Alumäe, T. (2017). Modeling under-resourced languages for speech recognition. Language Resources and Evaluation, 51, 961-987.

[7] Ragni, A., Knill, K. M., Rath, S. P., & Gales, M. J. (2014). Data augmentation for low resource languages. In Interspeech 2014: 15th Annual Conference of the International Speech Communication Association, 810-814.

[8] Masmoudi, A., Bougares, F., Ellouze, M., Estève, Y., & Belguith, L. (2018). Automatic speech r ecognition system for Tunisian dialect. Language Resources and Evaluation, 52, 249-267.

[9] Chinese Rule-Based Text Frontend. URL: https://github.com/PaddlePaddle/PaddleSpeech/blob/develop/docs/source/tts/zh_text_frontend.md. Last Accessed 2023/11/01.

[10] Speech Synthesis Technology (Introduction to Deep Learning Methods). URL: https://www.cnblogs.com/jacen789/p/14260194.html. Last Accessed 2023/11/01.