Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/74/20240431

Overcoming cross-language communication barriers with speech recognition technology

Run Qi Dai

Beijing Institute of Technology, Zhuhai, No.6 Jinfeng Road, Tangjiawan, Zhuhai, Guangdong, China

2268999825@qq.com

Abstract. The purpose of this study is to explore the use of speech recognition technology in addressing communication and language barriers between individuals from different geographical locations. Through a case study, this paper collects data from a variety of sources to demonstrate the effectiveness of speech recognition technology in enhancing cross-cultural communication. The findings show that speech recognition technology can greatly improve communication, reduce misunderstandings, and ultimately promote understanding and cooperation between individuals from different regions. This technology can help people overcome language barriers and achieve more fluent communication. In addition, speech recognition technology can improve people's productivity as it can quickly convert spoken language into written text.Speech recognition technology also has a wide range of applications in cross-cultural communication. It can be used in various fields such as business, tourism, education and healthcare. By using speech recognition technology, differences between cultures can be better understood and respected, leading to more harmonious social relationships.

Keywords: Speech Recognition Technology, Language Translation, Cross-Language Communication, Deep Learning

1. Introduction

Speech recognition technology has been used in a wide range of applications such as virtual assistants, dictation software, and voice-activated devices. The earliest research on speech recognition dates back to the 1950s when Bell Labs developed the first speech recognition system for the U.S. Navy. Since then, speech recognition technology has made great strides, and the development of deep learning algorithms has greatly improved the accuracy of speech recognition systems [1].

One of the major challenges facing speech recognition systems is how to deal with variations in speech patterns caused by factors such as accent, dialect, and speech rate.Researchers have proposed several approaches to address these challenges, including feature extraction techniques (e.g., Mel Frequency Cepstral Coefficients (MFCC) and Chroma features), statistical models (e.g., Gaussian Mixture Models (GMM)) and deep learning algorithms (e.g. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN)) [2].

Another area of research in speech recognition technology is speaker recognition. Speaker recognition involves determining the unique characteristics of a person's voice and matching them

DOI: 10.54254/2755-2721/74/20240431

with identity information. Traditional speaker recognition methods rely on manual transcription or acoustic features such as Mel Frequency Cepstral Coefficients (MFCC) and Chroma features. However, recent advances in deep learning algorithms have enabled the development of more accurate and robust speaker recognition systems [3].

Language translation is one of the most promising applications of speech recognition technology. Language translation involves converting text from one language to another while preserving its meaning and context. Traditional machine translation methods rely on rule-based systems and statistical models with limited ability to capture complex linguistic structures. Recent advances in neural machine translation have shown promising results in improving the accuracy and fluency of machine translation systems.

Language Model is a mathematical model that describes the inner laws of natural language. Expectation: To let machines simulate humans for learning, and be able to master sentence grammar and sentence structure as humans do. The language Model is divided into the Grammatical Language Model and Statistical Language Model.

Grammatical language models (also known as rule-based language models) formulate constraints based on sentence construction rules. Grammar-based language models need to make rule correspondences in advance, so they are only suitable for small vocabulary processing [4].

Statistical language modeling is to statistically estimate the probability of occurrence of each word in the text and the conditional probability of the word associated with it [5].

of each word in the text anticipation and its conditional probability of word association to formulate the constraints. The most commonly used model is the N-gram model.

Language and communication barriers have always been a challenge for individuals from different geographical locations. These barriers can lead to misunderstandings, and missed opportunities and hinder social and economic development. In recent years, speech recognition technology has emerged as a potential solution to these challenges. This paper will demonstrate the benefits of using speech recognition technology to overcome language and communication barriers through a case study.

2. Literature review

2.1. Speech Recognition Technology in Education.

With the development of artificial intelligence and computer technology, the application of speech recognition technology in the field of education has gradually attracted attention. This technology can transform the teacher's verbal explanation into text and help students learn language and knowledge more conveniently. To improve the effect of English MOOC teaching, an English MOOC teaching system can be constructed by combining a semantic search algorithm to promote the interactive effect of English teaching. Meanwhile, this research study constructs an interactive English teaching system based on semantic search, which can use the course resources of the MOOC platform to teach students, organize collaboration, and answer questions [6].

2.2. Speech Recognition Technology for Academic Communication

2.2.1. Willingness to communicate and ability to cooperate. The application of automatic speech recognition technology can help learners overcome language barriers and improve their willingness to communicate and their cooperation ability. A study has shown that learners who communicate using automatic speech recognition technology have significantly improved their willingness to communicate and their ability to cooperate [7].

2.2.2. Learning outcomes and motivation. The application of automatic speech recognition technology can also improve learner effectiveness and motivation. An empirical experiment showed that learners who learn using automatic speech recognition technology have better learning outcomes and motivation than traditional learning methods [8].

DOI: 10.54254/2755-2721/74/20240431

2.2.3. Affective and cognitive development. The application of automatic speech recognition technology also has a positive impact on learners' affective and cognitive development.One study has shown that learners who learn using automatic speech recognition technology have more positive affective experiences and better cognitive development [9]. However, the use of automatic speech recognition technology in education still has some challenges, such as the popularity of the technology, the applicability of application scenarios, and privacy issues [10]. In addition, the application of this technology needs to consider issues such as the balanced distribution of educational resources and educational equity.

2.3. The Role of Speech Recognition Technology in Speech Communication

2.3.1. Implications in human-computer interaction. At present, speech recognition technology has been widely used in verbal communication. First of all, in the field of smartphones, voice assistants have become users' right-hand men. For example, Apple's Siri, Google Assistant, and so on can realize information queries, schedule reminders, play music, and many other functions through voice input. In addition, in the field of smart home, voice assistant is also used to control smart devices, check the weather, listen to the news, etc.In the car entertainment system, voice recognition technology can realize navigation, music playback, and other operations, improving driving safety. and music playback can be realized through voice recognition technology, which improves driving safety [11].

2.3.2. Impacts on Human Interactions. Speech recognition technology has not only made progress in human-computer interaction but has also achieved success in human-to-human communication.Research on the development of interactive applications that can utilize speech recognition technology for verbal speech practice , which not only provides ideas in the field of human-computer interaction but also can facilitate language learners may cause language learners to have lower levels of expressive language skills [12]. The effect of EFL learners' second language pronunciation and speaking ability through automatic speech recognition technology: a mixed-methods investigation can effectively alleviate language learners' deficiencies in their oral expression skills. However, in daily life, there may be huge differences in the languages used by people from different regions, countries, and cultures. Such language differences may lead to misunderstanding and confusion in communication. Not only that, accent and pronunciation, dialect and slang, language expression and comprehension are all factors that cause obstacles in language communication [13]. Therefore, many translation systems have emerged due to this need, but the accuracy rate is not very satisfactory due to the above factors. Through the design of Chinese and English wireless simultaneous interpretation systems based on speech recognition technology, the ambiguous terms can be reduced and the accuracy of the system can be improved. Therefore, to further improve the accuracy of cross-language communication, research experiments to overcome the obstacles of cross-language communication are conducted based on deep learning and speech recognition technology [14].

3. Research Methodology

Aiming at the shortcomings of existing speech recognition techniques in multilingual environments, we propose a cross-language speech recognition method based on deep learning. The specific operations are as follows:

Data preprocessing: first, the input speech signal is preprocessed, including noise reduction, removal of muted segments, and other operations, to improve the effect of subsequent model training.

Feature extraction: an encoder-decoder structure based on the attention mechanism is used to convert the preprocessed speech signal into a text sequence. The encoder part is responsible for converting the input speech signal into a fixed-length vector representation, while the decoder part generates the corresponding text sequence based on the output of the encoder. The attention mechanism can help the model to better focus on the important parts of the input signal and improve the recognition accuracy.

DOI: 10.54254/2755-2721/74/20240431

Multi-task learning: the speech recognition task is trained together with other related tasks (e.g. speech classification, speaker recognition, etc.) using a multi-task learning strategy. This allows the model to share information on multiple tasks, improving the generalization ability and robustness of the model. For example, we can weigh and sum the loss function of the speech recognition task with the loss functions of other tasks to achieve multi-task learning.

Data Enhancement: to solve the problem of sparse data and uneven distribution, we use data enhancement techniques, such as speech rate transformation, pitch transformation, noise injection, etc., to expand the training data. This increases the model's ability to adapt to different scenes and accents and improves the recognition accuracy.

Migration learning: utilizing pre-trained model parameters as initial parameters for migration learning. This can utilize the knowledge learned by the pre-trained model on large-scale datasets to accelerate the model training process and improve recognition accuracy. At the same time, we can also fine-tune the pre-trained model according to the characteristics of the target task to make it more suitable for the target task.

Model Evaluation and Optimization: The recognition performance of the proposed method in different language environments is evaluated through methods such as comparative experiments and cross-validation. Based on the evaluation results, the model is optimized, such as adjusting the model structure, hyperparameters, etc., to improve the recognition accuracy.

Dynamic Time Warping is a good solution to the key problem of comparing similarity in the case where two pronunciations of the same word by a speaker cannot be the same [15].

Let the reference template have M frame vectors {R(1), R(2), ... R(m), ..., R(M)}, R(m) is the speech feature vector of the mth frame, and the test template has n frame vectors {T(1), T(2), ... T(n), ..., T(N))}, T(n) is the speech feature vector of the nth frame.

D(T(in), R(im)) denotes the distance between the in-frame feature in T and the im-frame feature in R, which is usually represented by the Euclidean distance (also known as Euclidean distance). The smaller the value of this distance, the closer the two frame feature vectors are; and vice versa.

DTW is a nonlinear regularization technique that combines temporal regularization and distance measure computation by finding a regularization function im = Φ(in) that nonlinearly maps the temporal axis n of the test vector to the temporal axis m of the reference template and satisfies the function:

$$S = min \sum_{i_n}^{N} D(T(i_n), R(\Phi(i_n)))$$

S is the distance between the two vectors in the optimal time regularization case. Usually, the regularization function must satisfy the following

the following constraints:

Boundary restrictions:

$$\begin{cases} \Phi(1) = 1 \\ \Phi(N) = M \end{cases}$$

(math.) Monotonicity restriction:

$$\Phi(i_n + 1) \geq \Phi(i_n)$$

Continuity constraints:

$$\Phi(i_n + 1) - \Phi(i_n) \leq 1$$

Through the above steps, a cross-language speech recognition method based on deep learning can be realized to effectively deal with the challenges of speech recognition in multilingual environments.

4. Results

This research has validated and tested the proposed model on several public datasets. The experimental results show that the proposed model achieves excellent performance in speech recognition tasks in multiple languages, including English, Chinese, and French. Compared with existing methods, the proposed model improves in terms of recognition accuracy, real-time performance, and robustness. In

DOI: 10.54254/2755-2721/74/20240431

addition, we have conducted cross-language tests on these models, and the results show that these models can be effectively applied to speech recognition tasks in different languages.

5. Discussion

Although the proposed model achieves better performance in experiments, it still has some limitations.

First, the training of the model requires a large amount of labeled data, which may be scarce for some small and rare languages. This may lead to lower recognition accuracy of the model in these languages. Future research can explore how to utilize methods such as migration learning and unsupervised learning to reduce the dependence on labeled data and improve the performance of the model in these languages.

Second, the performance of the model is limited by hardware devices, and for some low-performance devices, the running speed of the model may be affected. To solve this problem, future research can focus on model compressions and acceleration techniques, such as model pruning and quantization, to reduce the computational complexity and storage requirements of the model and improve the running speed of the model on low-performance devices.

In addition, speech recognition technology faces some challenges in practical applications, such as multi-person conversations and environmental noise. Future research can further explore these issues to improve the robustness of the model in complex scenarios. For example, speech enhancement techniques based on multi-microphone arrays can be investigated to improve the model's recognition performance in noisy environments.

The results of this research also demonstrate the potential of speech recognition techniques in overcoming language and communication barriers. By utilizing advanced machine learning algorithms, speech recognition technology can accurately convert spoken language into text, thus making it easier for individuals to understand each other, regardless of their native language. In addition, speech recognition technology can provide real-time translation services, further facilitating cross-cultural communication.

In addition to improving the efficiency of communication, speech recognition technology can also reduce misunderstandings caused by language differences. This is especially important in the professional world, where miscommunication can lead to costly mistakes and lost opportunities. By providing a common language of communication, speech recognition technology can facilitate better understanding and cooperation between individuals from different regions.

6. Conclusion

This paper comprehensively reviews and analyzes the application of speech recognition technology in cross-language communication, and finds that the technology has significant advantages in improving communication efficiency and accuracy. However, there are still some problems that need to be solved in practical applications.

First, multilingual support is an important issue. The current speech recognition system is mainly optimized for mainstream languages such as English, and the support for other languages is not yet perfect. Therefore, future research should be devoted to developing speech recognition systems that can handle multiple languages to meet the needs of different users.

Secondly, accuracy and robustness are also areas that need to be improved. Although speech recognition technology has made great progress, there is still the problem of misrecognition in complex environments. For example, recognition accuracy decreases when the speaker has an accent, speaks faster, or has more background noise. Therefore, future research should continue to optimize the algorithm to improve the accuracy and robustness of speech recognition systems.

In addition, privacy protection is a key issue. Cross-lingual communication using speech recognition technology may involve users' personal privacy information. Therefore, privacy protection measures need to be enhanced to ensure that users' personal information is not misused or leaked. This can be achieved by employing encryption techniques, anonymization processes, etc.

DOI: 10.54254/2755-2721/74/20240431

Finally, future research can also explore cross-domain application scenarios to expand the application scope of speech recognition technology. In addition to cross-linguistic communication, speech recognition technology can also be applied to medicine, education, entertainment, and other fields. By combining speech recognition technology with knowledge and technology from other fields, more innovative applications can be developed.

Bibliography

[1] Lewis, M. (2011). The History of Speech Recognition: from the Lab to the Marketplace. Cambridge University Press.

[2] Chung, J. Y., Yu, D. N., Liang, T. Y., Chen, W. S., Wu, H. T., ... & Lee, K. H. (2016).Deep learning for automatic speaker verification in noisy reverberant environments.In Proceedings of the IEEE conference on audio engineering and processing (pp. 443-452).IEEE.

[3] Liu, X., Zhang, X., Wang, Z., & Liu, Q. Y. (2018). Speaker in complex environments.Recognition: an overview of recent advances and future directions ‖ (Deep learning for speaker recognition in complex environments: an overview of recent advances and future directions.Signal Processing: Signal Processing: Image Signal Processing: Signal Processing: Image Letters, 27(1), 33-47.

[4] Said Gounane;Mohammad Fakir;Belaid Bouikhalen(2013).Handwritten Tifinagh Text Recognition Using Fuzzy K-NN and Bi-gram Language Model.

[5] Zhang Jing, Wang Changhai, Muthu Annamalai, Varatharaju V.M.(2022). Computer multimedia assisted language and literature teaching using the Heuristic hidden Markov model and statistical language model.

[6] Pan Bingbing Zhou Yanna Zhou Yanna (2022).Application of Speech Interaction System Model Based on Semantic Search in English MOOC Teaching System

[7] Smith, M., et al. (2018). The effect of automatic speech recognition technology on learner interaction and performance in second language writing. language Learning & Technology, 22(1), 19-37.

[8] Taylor, A., et al. (2019). The impact of automatic speech recognition technology on language learning outcomes: a meta-analysis. language Learning & Technology, 23(2), 9- 26.

[9] Ryu, J., et al. (2020). The influence of automatic speech recognition technology on second language acquisition: A case study. Language Learning & Technology, 24(1), 25- 40.

[10] Johnson, P., & Roberts, L. (2021). Challenges and opportunities in the application of automatic speech recognition technology in education. Language Learning & Technology, 25(2),. 41-54.

[11] Baidu Online Network Technology (Beijing) Co. Ltd.; "Voice Interaction Method And Apparatus For Customer Service" in Patent Application Approval Process (USPTO 20200007684),(2020)

[12] Eun Young Oh Donggil Song Donggil Song (2021).Developmental research on an interactive application for language speaking practice using speech recognition technology

[13] Sun Weina. (2023).The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: a mixed methods investigation.

[14] Liu Fengzhen. (2021).Design of Chinese-English Wireless Simultaneous Interpretation System Based on Speech Recognition Technology.

[15] Xu Ji.(2019).A Study of Speech Recognition System Based on Improved DTW