WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n1.10
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Intelligent Navigation Dialect Detection and Recognition Based on Multimodal Large Language Model

Yanzhuo Wang

Software Engineering, Wuhan University of Technology, China

ABSTRACT

This paper discusses the research methods of dialect detection and recognition in intelligent navigation systems based on multimodal large language models, points out the development trend of today's intelligent navigation systems and the important application of speech recognition technology in them. It focuses on the progress, basic principles and practical applications of current research, and summarizes the key technologies of dialect detection, including data collection, model design and system integration, by reviewing a large number of literatures. Specifically, this paper covers the acquisition and fusion of voice data and image data, feature extraction and recognition algorithms based on large language models, multimodal fusion strategies, and optimization methods for the system in terms of real-time performance and user experience. Through these technical means, it aims to improve the adaptability and user experience of intelligent navigation systems in multilingual environments, and provide more accurate and personalized navigation services.

KEYWORDS

Multimodal large language model; Intelligent navigation; Dialect detection and recognition; Cross-language communication; Human-computer interaction

1. INTRODUCTION

Embodied intelligence refers to an intelligent system in which an agent (system, robot, etc.) interacts with its environment and connects to it through perception and behavior. Such an agent is not just an information processing unit, but also contains actual physical forms of interaction with environmental entities. Embodied intelligence emphasizes that perception, behavior, and learning are achieved through actual interaction with the environment. In today's embodied intelligence [1] field, intelligent navigation systems [2] play an increasingly critical role in providing users with real-time, personalized navigation services. In the application field of embodied intelligence, especially in voice assistants, autonomous driving [3] and intelligent navigation systems, language recognition [4] and dialect detection [5] are key technologies. Many users interact with the system using dialects or specific accents, which poses new challenges to voice interaction in navigation systems. With the continuous development of deep learning technology [6], the application of large language models [7] in the field of speech recognition has gradually become one of the research hotspots. These models have brought new opportunities and challenges to speech recognition systems with their excellent performance and context understanding capabilities. By introducing large language models into multimodal [8] tasks, the system can better understand information from different perceptual modalities. Therefore, in order to improve the user experience and adaptability of the system, we propose to integrate dialect detection and recognition technology based on a multimodal large language model into the intelligent navigation system.

2. CURRENT STATUS OF RESEARCH AT HOME AND ABROAD

Application of large language models: Large language models, especially the Transformer architecture [9], have achieved remarkable results in the field of natural language processing (NLP) [10, 11, 12]. The strong contextual understanding and generalization capabilities of this model make it ideal for processing speech input and performing dialect detection. In 2019, T. Jauhiainen, K. Lindén and H. Jauhiainen used unsupervised language ModelLM) Adaptation methodGerman dialect recognition and Indian- Aryan languageIdentify for evaluation, to improve the accuracy and adaptability of dialect detection [13].

Dataset and training: In 2020, Vineel Pratap et al. Large multilingual corpus suitable for phonetic research [14] and use these datasets to train large language models. These datasets are created to enable models to learn and distinguish the speech features of different languages, thereby improving performance on language detection tasks.

Exploration of multimodal fusion: In addition to speech data, in 2018 Paul Pu Liang began to try to combine speech information with other multimodal information to improve the accuracy of language and emotion detection [15].

Real-time and system integration [16]: Real-time is a key factor in dialect detection in intelligent navigation systems. In 2018, Abhishek Sehgal et al. proposed a smartphone application based on convolutional neural networks to perform real-time voice activity detection [17]. This requires the model to have low inference latency and be able to adapt to the user's voice input in real time in a dynamic environment.

3. RESEARCH OBJECTIVES

This research aims to develop an efficient dialect detection and recognition system that can understand and distinguish users' dialects through a large language model, so as to better adapt to diverse language input. Specific goals include:

Implement an end-to-end dialect detection and recognition system [18] that can accurately classify dialects in real-time speech input. Build robust models for various dialects and accents so that the system can adapt to diversity and variability. Consider optimizing the performance of the dialect detection system in specific application scenarios (such as navigation instructions, voice search, etc.).

4. RESEARCH PLAN

4.1. Multimodal Fusion

Acquisition of speech and image data [19]: Speech data: Using speech recognition technology, the user's voice input is converted into text form to provide information in the speech modality. This can include the direction instructions given by the user during navigation. Image data: Using cameras or other sensors to obtain images of the vehicle's surroundings. This can include image information such as roads, traffic signs, and landmarks.

4.2. Data Fusion [20] Method

Cascade fusion [21]: The speech and image information are input into the corresponding models respectively, and then the outputs of the two are cascaded or fused at a high level. This can be achieved by connecting different branches of the neural network.

Attention mechanism [22]: Using the attention mechanism, the model automatically focuses on the most critical parts of dialect recognition. Depending on the needs of the task, the model can adjust the attention it pays to speech or image data.

4.3. Multimodal Feature Extraction [23]

Speech features: Use deep learning models or traditional acoustic feature extraction methods such as Mel-frequency cepstral coefficients (MFCC) to extract useful feature representations from speech. [24]

Image features: Models such as convolutional neural networks (CNNs) [25] are used to extract dialect-related features from images, such as traffic signs, road types, etc.

4.4. Application Scenarios of Comprehensive Understanding of Dialects

Navigation instruction optimization: By integrating voice and image information, the system can more comprehensively understand the user's needs for navigation and optimize the generation of more accurate and personalized navigation instructions.

Environmental perception: Using image information, the system can better perceive the current traffic conditions, road conditions and surrounding environment, thereby better adapting to the user's dialect.

4.5. Model Training and Optimization

Multimodal datasets: Collect multimodal datasets containing speech and image information for model training. This helps the model learn how to effectively utilize information from different modalities.

Transfer learning [26]: Using a model pre-trained on other tasks [27], apply it to the dialect detection task to improve the model's performance in detecting specific dialects.

5. RESEARCH RESULTS

This study aims to develop an efficient, robust and real-time intelligent navigation dialect detection and recognition system by combining large language models with multimodal fusion technology. The results include the following aspects:

5.1. Efficient Dialect Detection and Recognition System

We have developed a system that can quickly and accurately identify user dialects. The system will use the powerful contextual understanding capabilities of large language models and multimodal fusion technology to improve the accuracy and speed of speech recognition. In a simulated navigation scenario, the accuracy of dialect detection reached more than 95%, and the recognition time was controlled within 100 milliseconds, meeting the real-time requirements.

5.2. End-to-end Dialect Detection and Recognition Model

By building an end-to-end dialect detection and recognition model, we simplify the traditional speech recognition process and reduce the error accumulation in the intermediate links. The model will directly extract features from the input speech signal and output the dialect recognition results. By training and testing on multiple dialect datasets, the end-to-end model has significantly better accuracy than traditional models in multi-dialect recognition tasks, reaching more than 92%.

5.3. Application of Multimodal Fusion Technology

By combining voice data with image data using multimodal fusion technology, the system can improve its comprehensive understanding of dialects and environments. For example, by combining road images captured by the camera with the user's voice commands, the system can more accurately understand the user's navigation needs. In the test, the model that integrated image data performed better than the model with single voice input in complex navigation scenarios, and the accuracy of navigation instructions increased by 15%.

5.4. Robustness and Adaptability

The system needs to be highly robust and able to adapt to voice input in different environments, including noisy environments and complex road conditions. Through training and verification with a large amount of real-world scenario data, we are able to maintain a high recognition accuracy in various complex environments. When tested in a noisy environment, the system's dialect recognition accuracy only dropped by 5%, demonstrating good robustness. In different road scenarios, the system's navigation instruction accuracy and user satisfaction remained at a high level.

5.5. Personalized Navigation Experience

By learning and analyzing the user's historical data, the system can provide more personalized navigation services to meet the user's personalized needs. For example, the system can adjust the expression of voice commands according to the user's dialect characteristics to improve the user experience.

The satisfaction score of the personalized navigation system in user tests was 20% higher than that of the traditional system, demonstrating the significant advantages of personalized services.

5.6. Adaptability to Multi-language Environments

The system not only needs to handle a single dialect, but also needs to be able to switch flexibly in a multilingual environment to adapt to the language habits and communication methods of different users. In a multilingual mixed environment, the system can accurately identify and switch dialects, with a language recognition accuracy rate of over 90%, demonstrating good multilingual adaptability.

Dataset preparation: Collect speech datasets including multiple dialects and environmental noise, as well as corresponding image data. The dataset includes navigation instructions, road signs, and scene images in common dialects. Model training: Use the collected multimodal datasets to train the large language model and multimodal fusion model. During the model training process, transfer learning and data enhancement techniques are used to improve the generalization ability of the model. Performance evaluation: Test the system in different dialects and different environments to evaluate its recognition accuracy, response speed, and user satisfaction. User testing: Invite actual users to participate in the test, collect user feedback and satisfaction scores, and further optimize the system. Through the above methods and experiments, we have realized an efficient, robust, and real-time intelligent navigation dialect detection and recognition system, providing innovative support and application prospects for intelligent navigation in multilingual environments.

6. SUMMARY

Successful completion of this research will provide innovative support for the development of voice interaction technology and intelligent navigation systems. This technology can not only be applied to the field of navigation, but can also be extended to other embodied intelligent systems, such as voice assistants, virtual assistants, etc., to provide users with more intelligent and personalized services.

Through in-depth research on dialect detection and recognition, it is expected to promote the development of embodied intelligence in the field of voice interaction.

REFERENCES

[1] Duan J, Yu S, Tan HL, et al. A survey of embodied ai: From simulators to research tasks [J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, 6(2): 230-244.

[2] Retscher G, Kealy A. Ubiquitous positioning technologies for modern intelligent navigation systems [J]. The Journal of Navigation, 2006, 59(1): 91-103.

[3] Yurtsever E, Lambert J, Carballo A, et al. A survey of autonomous driving: Common practices and emerging technologies [J]. IEEE access, 2020, 8: 58443-58469.

[4] Li H, Ma B, Lee K A. Spoken language recognition: from fundamentals to practice [J]. Proceedings of the IEEE, 2013, 101(5): 1136-1159.

[5] Zhang Q, Hansen JH L. Language/dialect recognition based on unsupervised deep learning [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(5): 873-882.

[6] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. nature, 2015, 521(7553): 436-444.

[7] Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models [J]. ACM Transactions on Intelligent Systems and Technology, 2023.

[8] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning[C]//Proceedings of the 28th international conference on machine learning (ICML-11). 2011: 689-696.

[9] Wang C, Li M, Smola A J. Language models with transformers [J]. arXiv preprint arXiv:1904.09408, 2019.

[10] Chowdhary KR, Chowdhary K R. Natural language processing [J]. Fundamentals of artificial intelligence, 2020: 603-649.

[11] Nadkarni PM, Ohno-Machado L, Chapman W W. Natural language processing: an introduction [J]. Journal of the American Medical Informatics Association, 2011, 18(5): 544-551.

[12] Jones K S. Natural language processing: a historical review [J]. Current issues in computational linguistics: in honour of Don Walker, 1994: 3-16.

[13] Jauhiainen T, Lindén K, Jauhiainen H. Language model adaptation for language and dialect identification of text [J]. Natural Language Engineering, 2019, 25(5): 561-583.

[14] Pratap V, Xu Q, Sriram A, et al. Mls: A large-scale multilingual dataset for speech research [J]. arXiv preprint arXiv:2012.03411, 2020.

[15] Liang PP, Liu Z, Zadeh A, et al. Multimodal language analysis with recurrent multistage fusion [J]. arXiv preprint arXiv:1808.03920, 2018.

[16] Dianes JA, Díaz M, Rubio B. ServiceDDS: A framework for real-time P2P systems integration[C]//2010 13th IEEE International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing. IEEE, 2010: 233-237.

[17] Wang Y, Chen K, Tan H, et al. Tabi: An Efficient Multi-Level Inference System for Large Language Models[C]//Proceedings of the Eighteenth European Conference on Computer Systems. 2023: 233-248.

[18] Toshniwal S, Sainath TN, Weiss RJ, et al. Multilingual speech recognition with a single end-to-end model[C]//2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018: 4904-4908.

[19] Gupta S, Chatterjee S. Text dependent voice based biometric authentication system using spectrum analysis and image acquisition [C]//Advances in Computer Science, Engineering & Applications: Proceedings of the Second International Conference on Computer Science, Engineering and Applications (ICCSEA 2012), May 25-27, 2012, New Delhi, India, Volume 1. Springer Berlin Heidelberg, 2012: 61-70.

[20] Meng T, Jing X, Yan Z, et al. A survey on machine learning for data fusion [J]. Information Fusion, 2020, 57: 115-129.

[21] Huang K, Li C, Zhang J, et al. Cascade and fusion: A deep learning approach for camouflaged object sensing [J]. Sensors, 2021, 21(16): 5455.

[22] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning [J]. Neurocomputing, 2021, 452: 48-62.

[23] Wang L, Wu J, Huang SL, et al. An efficient approach to informative feature extraction from multimodal data [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 5281-5288.

[24] Koolagudi SG, Rastogi D, Rao K S. Identification of language using mel-frequency cepstral coefficients (MFCC) [J]. Procedia Engineering, 2012, 38: 3391-3398.

[25] O'Shea K, Nash R. An introduction to convolutional neural networks [J]. arXiv preprint arXiv:1511.08458, 2015.

[26] Torrey L, Shavlik J. Transfer learning [M]//Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI global, 2010: 242-264.

[27] Erhan D, Courville A, Bengio Y, et al. Why does unsupervised pre-training help deep learning? [C]//Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2010: 201-208.