Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/115/2025.18486

Research and Application Analysis on Key Problems of Automatic Speech Recognition for Dysarthria

Jie Yi[1,a,*]

[1]Institute of International Education, Zhengzhou University of Light Industry, Zhengzhou, China
a. yijie2005666@ldy.edu.rs
**corresponding author*

Abstract: At present, automatic speech recognition technology has developed rapidly, and the recognition accuracy of Automatic Speech Recognition (ASR) based on deep learning has been very high. Therefore, the speech recognition problem of patients with dysarthria has been paid more attention in recent years. However, due to the particularity of dysarthria patients and the strong variability of their speech, the relevant available datasets are very scarce, and it is difficult to adapt to the current general recognition model. In order to promote the development and progress of this field, based on a large number of literature research, this paper summarizes the construction of data sets, the solution of speech variability, and the key issues and development trends of multi-feature speech data fusion. At the same time, it lists the current application of automatic speech recognition for dysarticulation in some fields. It is hoped that the need of social communication and intelligent life of patients with dysarthria who exist as a minority group can be solved as soon as possible.

Keywords: Dysarthria, Automatic Speech Recognition, Speech Variability, Speech Data Fusion.

1. Introduction

Dysarthsia is a kind of speech disorder. It can be divided into two categories: organic disorder and developmental phonological abnormality. Dysarthria can occur at any age — in children, dysarthria can be caused by delayed speech development, abnormalities of the oral structure (such as cleft lip, cleft palate), weak oral muscle movement, poor control, or impaired hearing. It can also occur in people with brain diseases or trauma. However, due to the small number of dysarthria patients, it is very difficult to collect a large number of speech data of dysarthria patients, and the available data sets are very scarce. Moreover, the pronunciation characteristics of dysarthsia speech and normal speech are inconsistent in the probability distribution of acoustic feature vectors, and there is strong variability in pronunciation, which cannot be well matched with the general speech recognition model [1]. Both at home and abroad are facing the same dilemma—there is a lack of sufficient data sets for model training, which makes it difficult to construct a suitable speech acoustic model. Based on a large number of literature surveys, this literature aims to summarize and analyze the key issues in some of its fields. At the same time, it focuses on the needs of social communication and intelligent life of patients with dysarthria who have impaired language ability and limited self-expression ability, which seriously affect the quality of life. So far, the number of review articles on speech recognition of dysarthsia is small, and it is urgent to summarize and compare and analyze the construction methods and advanced technologies in this field. This paper is dedicated to provide convenience for researchers entering the field who want to quickly obtain this knowledge.

In recent years, with the progress of deep learning and speech recognition technology, the speech recognition models of dysarthsia have also been developed rapidly. Here is a brief overview of the development of Dysarthric Speech recognition models—Artificial Neural Networks(ANN) - based Automatic Speech Recognition (ASR) for Dysarthric Speech was the first acoustic model for dysarthric speech [2], Produced by Jayaram and Abdelhamied in 1995. In 2007, Hawley et al. developed a vocabulary limited and speaker dependent ASR, the Speech-Controlled Environmental Control System[3], which is very reliable in dealing with Speech changes. The accuracy on the training dataset improves from 88.5% to 95.4%, and even for speakers with severe dysarthasis, the average accuracy of word-level recognition still reaches 86.9%. In 2013, Sharma and Hasegawa-Johnson proposed an interpolation-based method. The experimental results show that the interpolation-based method achieves 8% absolute improvement and 40% relative improvement in recognition accuracy compared with the baseline MAP adaptation method. In 2021, Shahamiri proposed a new ASR system for dysarthric Speech, Speech Vision[4]. Experimental results on the UASpeech database show that the accuracy is improved by 67%, and the improvement is the largest for severe dysarthric speech. The rising accuracy of dysarthria recognition is also a process of continuous optimization of acoustic models.

However, there are still many problems and challenges in this process. The following will summarize and analyze the speech variability, multi-feature speech data fusion, and data sets. In the following section, the second section is about the analysis of speech variability and solution suggestions, the third section is about multi-feature speech data fusion, the fourth section is about the construction of the data set, the fifth section is about the application analysis of automatic speech recognition technology for dysarthsia, and finally the future prospects and challenges and summary of this paper are proposed.

2. Analysis and Resolution of Speech Variability

2.1. Speaker adaptive connotation

This chapter first introduces several factors that constitute speech variation [5]. The first is that a person's physical characteristics and lifestyle habits, such as smoking and drinking, can affect the tone and pitch of the voice; The second is the individual's deliberate voice adjustment, such as emphasis or doubt; The third is the change in the pronunciation of words. For dysarthra, some pathological features of dysarthra, such as discontinuous pronunciation, slow speech speed and improper pause, will cause differences between dysarthra and normal people, and it is difficult to match speech data with acoustic models. At present, the mainstream solution to speech variability is to introduce speaker adaptation [1]. The definition of speaker adaptation is as follows: Speaker adaptation is a series of speech acoustic feature vectors $X = \{x1, x2, \ldots, xt, \ldots, xT\}$ using speaker data to obtain feature parameters θE, thus, the acoustic features are mapped onto the word sequence W. The model can be defined as: $yt = f(xt; \theta; \theta E)$ where, $yt = f(xt; \theta; \theta E)$ is the model θ with parameters θE and parameters, and yt is the output label of frame t. The advantage of this method is that it can use speaker data to adjust the speaker independent model, solve the mismatch problem caused by speaker differences, and improve the accuracy of speech recognition system.

2.2. Dysarthria speaker adaptation

Then there are two ideas for the current adaptation of dysarthsia speakers [6]: (1) processing the acoustic feature parameters of dysarthsia speakers so that they can be applied to the original model;

(2) The dysarthria speech is used to train the model, and the model parameters are changed to adapt to the variability of dysarthria speech. According to the different ways of parameter adjustment in the adaptation idea, the adaptation of dysformant speakers can be divided into feature-domain adaptation and model-domain adaptation. However, there are still three major challenges in dysarthria adaptation. The first one is the modeling of speech variability and variation uncertainty in different degrees of dysarthria. Second, the balance of probability distribution between source domain and target domain in model adaptation. The third is the adaptive problem of dysarthria identification with sparse data.

3. Fusion of multi-feature speech data

3.1. Advantages of data fusion

Data fusion has the following advantages: first, it can improve the accuracy of data and reduce the case of misidentification; Second, reduces the cost of data processing, by integrating multiple data sources, researchers can reduce the dependence on a single high-cost device, thereby reducing the overall cost of the system, realizing resource optimization and simplifying management. Third, the robustness of the system is enhanced to improve the adaptability of the system to different environments and speakers. Figure 1 below shows a mind map that demonstrates the advantages of data fusion.

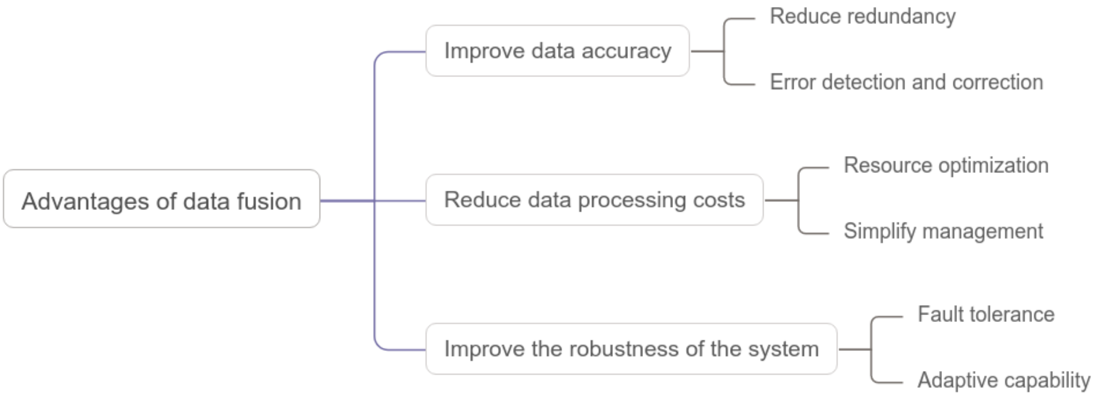
Advantages of data fusion
Improve data accuracy
Reduce redundancy
Error detection and correction
Reduce data processing costs
Resource optimization
Simplify management
Improve the robustness of the system
Fault tolerance
Adaptive capability

Figure 1: Mind map of the advantages of data fusion

3.2. Execution ideas

In the execution flow of automatic speech recognition, feature extraction is a very important link, such as Mel-frequency Cepstral Coefficient (MFCC), Linear Predictive Coding (LPC), etc. [7]. These features can effectively represent the audio characteristics of speech and be used for subsequent recognition. In order to extract richer linguistic features, a multi-modal self-supervised learning framework is considered because the multi-modal self-supervised learning framework has the following advantages—Self-supervised learning does not need to rely on expensive manual labeling and a large amount of data, which can effectively save costs. In addition, it can make good use of unlabeled data [8] and reduce the dependence on labeled data. Self-supervised learning can also improve the generalization ability of the model by learning richer feature representations. Human perception of language is inherently multimodal, including visual and auditory. Then by introducing multimodality, researchers can improve its ability to understand language in noisy environments and provide a means of communication for dysarthria. Finally, the extracted language features were fused to create the acoustic model. However, whether it is self-supervised learning or processing multi-modal data, there are some problems that need to be solved, which are also difficulties for scholars to overcome in the future. Figure 2 shows the conception of this paper for the future execution of automatic speech recognition for dysarthra.

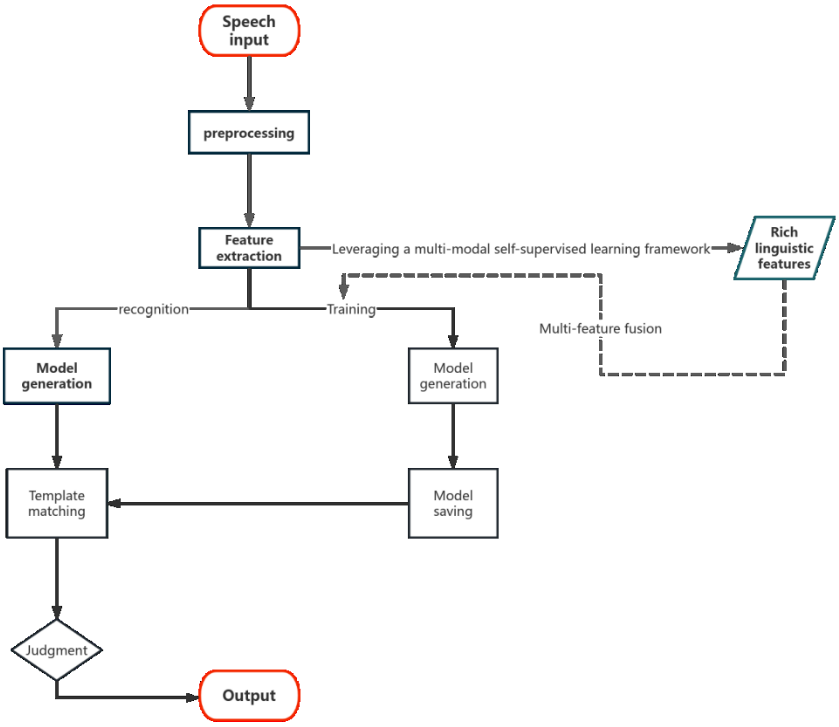

Figure 2: A conception of the future execution process of automatic speech recognition for dysarthria

4. Building the dataset

In this part, this paper briefly introduces the development of standard data sets for dysarthria. Whitaker data set is the first data set produced in 1993 [9], which aims to provide standard dysarthria speech data for researchers to improve the performance of speech recognition models; In 1996, Nemours dataset was produced by dupont Institute of AI in Wilmington, Texas, USA [9], which aims to test the intelligibility of dysarthria speech enhanced by various signal processing methods; Torgo dataset produced by the University of Toronto in 2012 [9], which laid a solid foundation for the performance improvement of speech recognition algorithms for dysarthsia; The UK EPSRC project produced HomeService data set [9] in 2016, which aims to realize speech recognition in the home environment. The data obtained is more natural and more consistent with the real scene of speech recognition; In 2020, the speech laboratory dataset of Shantou University was produced [10], which aims to help patients with aphasia to conduct rehabilitation training through systematic language guidance, and also plays the role of assisted living; In 2021, the University of Pisa produced the IDEA dataset [11], which aims to provide the basic corpus for speech recognition of dysarthra, while conducting research on the acoustic characteristics of patients with different pathologies.

In this paper, by comparing the earliest Whitaker dataset with Euphonia dataset of Google, it can be found that the dataset of disease type dysarthria has been greatly improved and developed, both in terms of data volume and corpus content. Therefore, the development direction of the dataset in the future can consider these three points [10]—First, dataset recording is closer to the actual use case.

The direction suggests two ways to do this — the first is to use words that focus on specific domains, such as life, work, and play-related scenarios, the second way is to record audio directly using the recorder's mobile phone in the daily life environment, avoiding the tension of the recorder caused by the change of the recording environment, and laying the foundation for the collection of a large number of data. Second, the post-processing of the dataset is more refined. Researchers can discard the invalid data with large noise and incorrect format, so as to avoid the interference of the recognition effect of the whole data set. Third, form a more comprehensive type of data set. With the increasing amount of data in the dysarthra dataset, speech labels will certainly develop more comprehensively, and at the same time, the factor labels that affect speech recognition will be increased. As shown in Figure 3 below, the bar chart of the comparative analysis of the two data sets is shown.

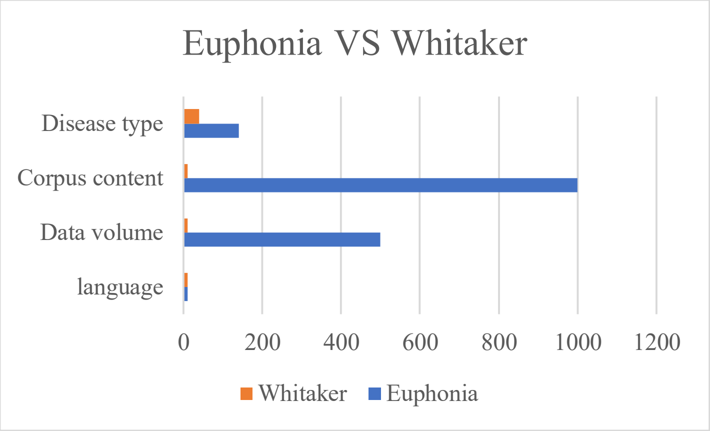

Figure 3: Comparative analysis of Whitaker dataset and Euphonia dataset

5. Application Analysis

5.1. Medical field

First, it provides detailed speech features to assess the severity of dysarthria. ASR models are used to analyze the pronunciation and intonation of dysarthria individuals, such as pitch, volume, speech rate, etc., so that doctors can carry out more in-depth diagnosis and treatment; Second, realize real-time feedback and correction of pronunciation training, so as to help patients correct errors in time and gradually improve pronunciation; Third, personalized intervention and treatment are provided. Speech therapists can design targeted treatment plans based on the results of automatic speech recognition analysis to help patients carry out intensive training on specific pronunciation. Online platforms can also incorporate this technology to allow patients to perform treatment remotely.

5.2. Daily life

First, it can help patients with dysarthria to communicate. The dysarthria acoustic model can be used to develop communication assistance tools for patients with severe dysarthria, thereby helping patients who cannot pronounce clearly to communicate. Second, to improve the public awareness of such diseases as dysarthria, to help patients and their families better understand some of the manifestations and pathological characteristics of supply disorders. Third, to realize cross-lingual communication, automatic speech models can be trained according to the characteristics of different languages and dialects, and then help patients worldwide to obtain better communication experience and communication opportunities.

5.3. Research and technology development

First, large-scale data analysis can be carried out, ASR technology is able to process a large amount of speech data, which provides support for the research of dysarthria. Researchers can explore the causes and treatment of dysarthria by analyzing the articulatory characteristics of different populations of dysarthria. Secondly, a standardized and comprehensive data set is constructed. The acoustic model of some dysarthria can generate speech with the characteristics of the specified speaker according to a small number of speech audio samples provided, so as to effectively solve the problem of data set scarcity. Third, identify the articulation patterns of different dysarthres, and train the ASR model of dysarthres to identify the articulation patterns of different types of dysarthres, so as to provide data support for early diagnosis and intervention.

6. Challenges and Prospects

This section is about the challenges and prospects of automatic speech recognition for dysarthria. The first challenge is the current situation that it is too difficult to collect data from dysarthra recognizers. In the future, to solve this challenge, the paper can consider two directions: "How to make more speech data corpora of dysformant" and "how to further improve the speech recognition performance trained with little or no resource data". The second challenge is the poor generalization ability of acoustic models caused by the huge differences in dysarticulation. For the problem of speech diversity, current researchers mainly consider the introduction of speaker adaptation methods to solve the mismatch problem caused by speaker differences, so as to improve the accuracy of speech recognition systems, however, there are still relevant difficulties such as "speech variability of different degrees of dysarthria and the modeling of its variation uncertainty". In the future, the adaptive methods of dysarthria can be started from three directions: "speech variability analysis", "multi-feature and multi-modal data fusion", and "adaptation under small data amount" [1]. The third challenge is that due to the limited availability of speech data, it is difficult to achieve high accuracy of the acoustic model. In the face of the limited scale of dysarthria speech data, it is suggested to fuse more modal signals to solve the problem of too little resource data, or to use small sample learning to break through the bottleneck of dysarthria speech recognition research in the future.

The following is the outlook of this paper. First, the multi-modal self-supervised learning framework can be used to train acoustic models in the future. The reason is that multi-modal self-supervised learning is very suitable for the situation without a large amount of labeled data, which can effectively improve the learning ability and generalization ability of the model. Using this framework to train acoustic models of dysarthria can capture richer speech features. Second, a more representative public data set is constructed. So far, the existing data sets have problems such as relatively small data volume, unclear data classification, and few available data signals. It is imperative to establish a public data set for dysarthsia speech. Thirdly, multi-feature and multi-modal speech data fusion is carried out to improve its recognition accuracy. As the collection of multi-modal data required for speech recognition becomes more convenient and accurate [12], the speech information obtained by the speech recognition system will be more comprehensive, and the recognition accuracy will certainly be improved.

7. Conclusion

This paper reviewed a large number of literature, deconstructed the development process of the field of automatic speech recognition of dysarthsia, and the existing challenges, compared and summarized such as data set construction, speech recognition and other technologies. At the same time, this paper also puts forward an idea of the future execution process of automatic speech recognition of dysarthsia. Finally, this paper expresses the challenges of automatic speech recognition of dysarthria to be overcome in the future and the prospect of future development. At present, the speech recognition technology of dysarthsia still has a great development space. The lack of public data sets and the limitation of training resources are all urgent problems to be solved. In the future, if you want to improve the supply barrier speech recognition model, it is recommended to consider from the construction of datasets, the use of multi-modal self-supervised representation learning framework for feature extraction, and the training of acoustic models based on multi-feature fusion.

References

[1] Kang, X. C., Dong, X. Y., Yao, D. F., Zhong, J. H. (2024). Research progress and prospects of speaker adaptation for dysarthria. Computer Science, 51(8): 11-19.

[2] Shahamiri, S. R., Salim, S. S. B. (2014). Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach. Advanced Engineering Informatics, 28(1): 102-110.

[3] Hawley, M. S., Enderby, P., Green, P., et al. (2007). A speech-controlled environmental control system for people with severe dysarthria. Medical Engineering & Physics, 29(5): 586-593.

[4] Shahamiri, S. R. (2021). Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29: 852-861.

[5] Markl, N. (2022). Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition//Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 521-534.

[6] Turrisi, R., Badino, L. (2022). Interpretable dysarthric speaker adaptation based on optimal-transport. arXiv preprint arXiv:2203.07143.

[7] Geng, M., Xie, X., Ye, Z., et al. (2022). Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30: 2597-2611.

[8] Liu, X., Zhang, F., Hou, Z., et al. (2021). Self-supervised learning: Generative or contrastive. IEEE transactions on knowledge and data engineering, 35(1): 857-876.

[9] Qian, Z., Xiao, K. (2023). A survey of automatic speech recognition for dysarthric speech. Electronics, 12(20): 4278.

[10] Song, W., Zhang, Y. H. (2024). A review of research on speech recognition algorithms for dysarthria. Computer Engineering and Applications, 60(11): 62-74.

[11] Marini, M., Viganò, M., Corbo, M., et al. (2021). IDEA: an Italian dysarthric speech database//2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 1086-1093.

[12] Gandhi, A., Adhvaryu, K., Poria, S., et al. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, 91: 424-444.