SUMMARY OF LOW-RESOURCE DYSARTHRIA WAKE-UP WORD SPOTTING CHALLENGE

Ming Gao[1], *Hang Chen*[1], *Jun Du*[1,*], *Xin Xu*[2], *Hongxiao Guo*[2], *Hui Bu*[2], *Ming Li*[3], *Chin-Hui Lee*[4]

[1]University of Science and Technology of China, China
[2]Beijing AISHELL Technology Co., Ltd., China
[3]Wuhan University, China
[4]Georgia Institute of Technology, USA

ABSTRACT

In recent years, the rapid advancement and widespread adoption of speech technology have made smart home systems a common feature in many households. However, individuals with dysarthria face difficulties using these technologies due to inconsistent speech patterns. This paper summarizes the Low-Resource Dysarthria Wake-Up Word Spotting (LRDWWS) Challenge at SLT 2024, which aimed to develop effective voice wake-up systems for individuals with dysarthria. The challenge attracted 25 teams from 4 countries, with 7 teams submitting results and 5 providing detailed system descriptions. This paper presents an overview of the dataset, evaluation metrics, and key innovations from participating teams. Our findings highlight the potential of these systems to enhance the accessibility and usability of smart home technologies for individuals with dysarthria. The challenge results underscore the importance of developing specialized solutions to meet the unique needs of this user group.

Index Terms— Dysarthria, Wake-up Word Spotting, Speaker-dependent Systems, Speech Disorder

1. INTRODUCTION

In recent years, speech technology has rapidly advanced, making smart home systems common. While most people find it easy to manage smart devices, individuals with dysarthria face challenges due to inconsistent speech patterns. Dysarthria, often associated with conditions like cerebral palsy, Parkinson's disease, ALS, and stroke, affects articulation, fluency, volume, clarity, and speech rate, making it hard for commercial smart devices to understand commands. These individuals may also have motor impairments, further complicating device use. Voice control offers an empowering solution, enabling greater independence, with voice wake-up being a key step for access.

Before this challenge, significant research had already been conducted on wake-up spotting (WWS) for the people without dysarthria. With the rise of deep learning, many end-to-end neural network architectures emerged, enhancing the accuracy and efficiency of voice wake-up technologies[1, 2, 3]. Recent studies have focused on handling complex scenarios and difficult conditions, such as far-field WWS[4], speech wake-up amidst background noise[5], WWS with few-shot learning[6], and real-time processing[7]. Furthermore, there is a growing interest in voice wake-up systems that incorporate multimodal information, combining audio with visual or tactile inputs to improve accuracy and user experience[8, 9, 10]. On the other hand, in the field of dysarthric speech research, existing studies primarily focus on database construction[11, 12, 13], speech recognition[14, 15, 16], data augmentation[17, 18, 19], voice conversion[20, 21, 22] and speech severity classification[23, 24, 25]. To the best of our knowledge, there has been no specific research on wake-up spotting systems for people with dysarthria.

Our challenge seeks to solve the speaker-dependent wake-up spotting task by utilizing a small amount of audio from the specific person. This research has the potential to not only enhance the quality of life for individuals with dysarthria, but also facilitate smart devices in better accommodating diverse user requirements, making it a truly universal technology. Given that individuals with dysarthria from a minority group with limited research conducted on them, we propose the Low-Resource Dysarthria Wake-up Word Spotting (LRDWWS) Challenge on SLT 2024 grand challenges, separate from the regular conference, to foster collaboration among researchers who share our interests. We hope that this will raise awareness about dysarthria and encourage greater participation in related research endeavors.

This challenge is distinguished by the following three key features.

- **Innovative focus.** It is difficult for existing wake-up word spotting systems to deal with dysarthric speech. This represents the first focus on mandarin wake-up

[*]Corresponding author

word spotting for individuals with dysarthria, a pioneering effort in the research field. This not only highlights the uniqueness of the project but also emphasizes the recognition of the need for speech recognition technology for this particular group.

- **Open-source dataset and sample diversity.** The development and open-sourcing of the first mandarin dysarthria wake-up word dataset, coupled with normal samples parallel to dysarthria samples, promote sharing and collaboration within the research community. This approach ensures that the study is based on a wide and diverse set of voice samples, increasing the universality and reliability of the research outcomes.
- **Innovative system framework.** Given the relative scarcity of data resources for mandarin-speaking individuals with dysarthria, the proposal introduces an innovative training strategy. This strategy involves pre-training using a large-scale mandarin corpus, followed by fine-tuning with a smaller set of speaker-aware dysarthria corpora. This method could effectively enhance the performance and adaptability of the model, especially in resource-limited scenarios.

The paper is structured as follows. Section 2 discusses the dataset for this challenge and examines the inherent difficulties of dysarthric speech data. Section 3 delineates the task of this challenge, provides an overview of the baseline code, and describes the evaluation metrics. It also outlines the special rules and phases of the competition. Section 4 introduces the participants and their methodologies, followed by a thorough discussion and analysis of their approaches. The paper concludes in Section 5. Further details are available on the challenge website[1].

2. DATASET

2.1. Statistical information

We utilized our previously released Mandarin Dysarthria Speech Corpus (MDSC) [26] as the foundational training dataset for this challenge. MDSC includes 18,630 recordings totaling 17 hours, of which 10,125 are from non-dysarthric recordings (Control) totaling 7.6 hours, and 8,505 are from dysarthric recordings (Dysarthria) totaling 9.4 hours. We record utterances from 21 speakers with dysarthria (12 females, 9 males) and 25 speakers without dysarthria (13 females, 12 males). The recordings consist of 10 wake-up words repeated five times at varying speeds. MDSC also includes 355 non-wake-up words, encompassing fixed command words, free command words, household instructions, and other phrases. Each person's text list has 295 non-repeated sentences. The recordings, sampled at 16kHz, take place in a quiet indoor environment, with the participants positioned approximately 20cm away from the mobile microphone.

Table 1. Statistical information of every subset in MDSCv2. "Control" stands for individuals without dysarthria while "Dys." represents individuals with dysarthria.

Subset	**Train**		**Dev**	**Test-A**	**Test-B**
	Control	**Dys.**	**Dys.**	**Dys.**	**Dys.**
Duration(h)	7.6	7.4	2	4.9	4.3
Speaker	25	17	4	10	10

Additionally, we recorded extra speech data from 20 new speakers with dysarthria, referred to as MDSC-Eval, as the test set for this competition. MDSC-Eval includes 8,760 recordings totaling 9.2 hours. The recording method for the MDSC-Eval is consistent with the MDSC, with the difference being that each person in the set has 11 additional confounding negative words to increase the challenge of the competition.

Table 1 delineates the distribution of Control and Dysarthria across the training, development, test-A and test-B sets, providing specifics on the duration and the number of speakers. There is no overlap in speakers among the data in each subset. Test-A and test-B together constitute MDSC-Eval, representing the test data for Leaderboard-A and Leaderboard-B of the competition, respectively.

During the data collection phase, we followed some principles to recruit volunteers with dysarthria for recording: (1) native mandarin speakers; (2) aged between 18 and 48; (3) diagnosed with dysarthric conditions. All volunteers signed informed consent forms, ensuring they were fully aware of the recording's purpose, usage scope, and privacy protection measures.

During the data processing phase, we meticulously reviewed all data. Each dysarthric speech recording was annotated and verified by five experts. All previously released data from the MDSC were assigned to the training set and development set, ensuring that the test set comprised entirely unreleased data, thereby maintaining the fairness of the competition.

In addition, we compiled a summary of other dysarthric speech databases and non-dysarthric wake-up word databases, including download instructions, for researchers to reference and utilize.

2.2. Challenges in dysarthric speech

Dysarthric speech data has several inherent challenges that significantly impact speech recognition systems. These challenges stem from the unique characteristics of dysarthric speech, which include variability in articulation, irregular speech rates, and low intelligibility.

[1] https://lrdwws.org

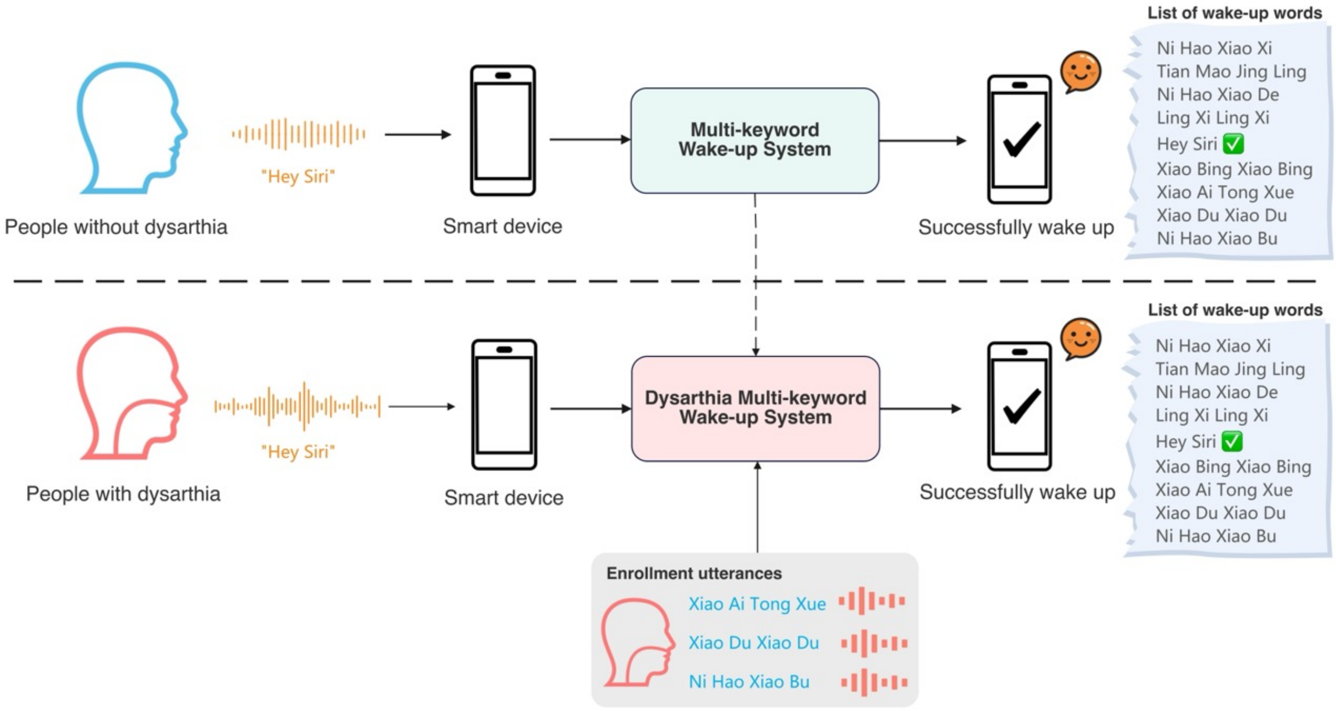

Fig. 1. The wake-up spotting architecture for both users with dysarthria (top) and users without dysarthria (bottom). For users without dysarthria, a multi-keyword wake-up system is trained to enable voice-activated functionalities. For users with dysarthria, due to the variability in their speech and significant differences between individuals, a small amount of enrollment speech is used for each person to train a speaker-dependent wake-up word system. This challenge focuses solely on developing systems for individuals with dysarthria.

Articulation Variability. Dysarthric speech shows significant articulation variability due to inconsistent muscle control, causing unpredictable phoneme pronunciation. This makes it hard for automated systems to recognize patterns, as the same word can sound different each time.

Speech rate and rhythm irregularities. Dysarthria often results in irregular speech rates and rhythms, with speech being slow or fast, and having unexpected pauses and fluctuations. This challenges traditional speech recognition systems, which are trained on consistent speech patterns.

Prosody and intonation issues. Dysarthric speech lacks natural variations in pitch, tone, and stress (prosody), making it hard for systems to distinguish between different utterances (e.g., statements vs. questions) and understand the emotional context.

Low speech intelligibility. Dysarthric speech often has low clarity, making it difficult for both humans and machines to understand. This is due to weak articulation, breathy or strained voice quality, and nasalization.

Co-occurring conditions. Many individuals with dysarthria have other motor impairments or cognitive challenges, affecting their ability to produce clear and consistent speech, adding complexity to speech data collection and processing.

3. CHALLENGE DESCRIPTION

3.1. Challenge

As an initial foray into dysarthric speech, we have formulated guidelines to define the problem scope and focus on the unique aspects of the dysarthric wake-up word spotting task.

- **Controlled environment.** Given the intricate speech patterns exhibited by individuals with dysarthria, meticulous efforts were made to minimize potential interferences during the audio recording process. Specifically, the recordings were conducted in a controlled environment carefully chosen to be devoid of any background noise.

- **Speaker-dependent.** Inaccurate pronunciation and reduced fluency exhibited by individuals with dysarthria lead to substantial variability in their speech. Even among individuals with the same type of dysarthria, the pronunciation of a single word can vary under different conditions. Hence, this challenge aims to address the speaker-dependent wake-up word spotting, offering a promising solution to voice wake-up problem among speakers with dysarthria.

- **Limited samples.** Individuals with dysarthria often have trouble in recording lengthy audio due to physical limitations. From the perspective of practical application, this challenge allows the use of only a limited number of wake-up word audio samples from a specific individual with dysarthria. However, we provide additional wake-up word audio samples and non-wake-up audio samples from both individuals with and without dysarthria as supplementary data.

The architecture of this challenge is illustrated in Figure1. The objective of the Low-Resource Dysarthria Wake-Up Word Spotting (LRDWWS) Challenge is to develop a speaker-dependent wake-up word spotting system specifically designed for individuals with dysarthria. Participants are required to create systems that accurately detect wake-up words spoken by individuals with dysarthria using enrollment utterances from target speakers (about 3 minutes per person, including both wake-up and non-wake-up words) and supplementary speech data from other individuals with dysarthria. The challenge encourages the use of other relevant open-source datasets, provided they are explicitly stated in the final report or technical documentation.

3.2. Baseline

The baseline speaker-dependent dysarthria wake-up word system uses a multi-stage fine-tuning process with a CNN architecture based on the WEKWS[27] framework. Initially, a CNN-based model is trained on the Control dataset of MDSC. This model is then fine-tuned using the Dysarthria training set to adapt to the general characteristics of dysarthric speech. Finally, the model undergoes a second fine-tuning phase with the enrollment speech specific to each individual speaker with dysarthria. This multi-stage approach, combined with data augmentation techniques such as adding background noise and varying pitch and speed, enhances the system's robustness and accuracy in recognizing wake-up words for speakers with dysarthria. The code can be found on https://github.com/greeeenmouth/LRDWWS.

3.3. Evaluation metrics

The evaluation metric of this challenge comprehensively considers both the False Acceptance Rate (FAR) and the False Rejection Rate (FRR), which is defined as follows:

$$FRR = \frac{N_{FR}}{N_{wake}}, FAR = \frac{N_{FA}}{N_{non-wake}} \tag{1}$$

where N_{wake} and $N_{non-wake}$ denote the number of samples with and without wake-up words in the evaluation set, respectively. N_{FR} denotes the number of samples containing the wake-up word while not recognized by the system. N_{FA} denotes the number of samples containing no wake words while predicted to be positive by the system.

Specifically, we define the $Score$ for the i-th wake-up word as the sum of FAR_i and FRR_i. To obtain the overall performance measure, we calculate the $Score$ as the average of the scores across all wake-up words. The $Score$ is computed using the following formula:

$$Score = \frac{1}{N} \sum_{i=1}^{N} (FAR_i + FRR_i) \tag{2}$$

where N is the total number of wake-up words, which is 10 in this challenge. The lower $Score$, the better the system performance.

3.4. Challenge rules and phases

We established a set of rules to ensure a fair and competitive environment for all participants. The key rules are as follows:

Registration. Each team must register through the official challenge website. Only one member from each team needs to register, providing necessary details such as team name, affiliation, and email address.

Data Usage. Participants are allowed to use the provided dysarthric speech datasets and may incorporate other open-source datasets, provided these additional datasets are explicitly disclosed in their final report.

Submission. Participants can submit results up to 3 times a day in the prescribed format. After the leaderboard freeze, they must submit a system description for algorithm compliance review.

The challenge is divided into two phases: leaderboard-A and leaderboard-B. Leaderboard-A utilizes test-A data, which includes released labels for both enrollment and evaluation data. Participants have access to these labels but are strictly prohibited from using them to enhance their models during the training phase. The primary objective of leaderboard-A is to familiarize participants with the submission process and allow them to adjust their models based on the feedback. Importantly, scores from leaderboard-A are not included in the final ranking. Leaderboard-B employs test-B data, which consists of labeled enrollment data and unlabeled evaluation data. Participants are permitted to use all the data from leaderboard-A during the training phase for leaderboard-B. However, they must submit their results for leaderboard-B within a short time frame, leaving insufficient time for major model adjustments. This structure ensures that the final evaluation phase reflects the participants' genuine performance without extensive post-evaluation tuning.

4. PARTICIPANTS AND SUBMITTED SYSTEMS

4.1. Participants

This challenge attracted registrations from 25 teams across 4 countries, including China, USA, Canada, and Australia. A

Table 2. Results of all submissions of Leaderboard-A and Leaderboard-B. The teams are ranked based on their scores in Leaderboard-B. "Multi" denotes the use of multiple speech datasets. T007 and T017 did not submit system description papers, making it impossible to determine their methodologies; thus, a "-" is used to denote this unknown information. The "/" symbol indicates that no pre-trained model or external dataset was used in those cases.

Team ID	**Pre-trained**	**External Datasets**		**Leaderboard-A**			**Leaderboard-B**		
	Model	**Non-dysarthria**	**Dysarthria**	**FAR↓**	**FRR↓**	**Score↓**	**FAR↓**	**FRR↓**	**Score↓**
T011	Data2vec2	Multi(60kh)	/	0.0183	0.0825	0.1008	0.0032	0.0050	0.0082
T021	Hubert	WenetSpeech(10kh)	/	0.0050	0.0325	0.0375	0.0048	0.0050	0.0098
T018	Data2vec	Multi(60kh)	CDSD(34h)	0.0042	0.0225	0.0267	0.0030	0.0075	0.0105
T007	-	-	-	0.0032	0.2825	0.2857	0.0033	0.0075	0.0108
T017	-	-	-	0.0042	0.1125	0.1167	0.0029	0.0192	0.0221
T019	Wav2vec2	Multi(50kh)	/	0.0302	0.2325	0.2627	0.0203	0.0790	0.0993
T008	/	/	/	0.0595	0.1900	0.2495	0.0389	0.0738	0.1127
Baseline	/	/	/	0.0387	0.2725	0.3112	0.0286	0.1017	0.1303

total of 7 teams from 3 countries submitted results during the evaluation phase. Among them, 5 teams submitted system description papers to introduce their methods.

4.2. Submitted systems

Table 2 shows the performance results for all submissions on Leaderboard-A and Leaderboard-B. Each team is identified by their Team ID. Notably, in the usage of external data, non-dysarthria datasets were utilized for training the pre-trained models rather than being directly employed. The high score of the baseline is attributed to the challenging articulation of speakers with dysarthria. Additionally, the baseline system did not utilize any external datasets or pre-trained models. It is also notable that the best result of the Leaderboard-B achieved was 0.0082, surpassing the performance of the system across all teams.

The T011 system introduced an end-to-end Pretrain-based Dual-filter Dysarthria Wake-up word Spotting (PD-DWS) system, featuring a 2branch-d2v2 model and a dual-filter strategy. The pre-trained data2vec2 (d2v2)[28] model, trained on 60,000 hours of speech data, handles both ASR and WWS tasks through multi-task fine-tuning. The model uses max pooling loss for WWS and CTC loss for ASR, with dynamic data augmentation. The dual-filter strategy includes a threshold filter for initial WWS probability filtering and an ASR filter for secondary filtering using outputs from the WWS branch and a fine-tuned Paraformer[29] model. The PD-DWS system achieved a FAR of 0.0032 and a FRR of 0.005, resulting in a total score of 0.0082 on the test-B evaluation set, securing first place in the competition.

Team T021 employed a Chinese Hubert[30] model, pre-trained on 10,000 hours of WenetSpeech[31] data and adapted for mandarin. They used only the challenge-provided dysarthric dataset and followed the three-stage fine-tuning process, same as the baseline. During inference, features from enrollment speech created class prototypes, and test speech features were compared using cosine similarity to classify wake-up words. The method achieved a score of 0.0098, with a FAR of 0.0048 and a FRR of 0.005, securing second place in the competition.

T018 developed a system using Paraformer, which is trained on 60,000 hours of mandarin speech for feature extraction and a wake-up word detection model. They evaluated MDTC[32], Squeezeformer[33] and E-Branchformer[34], selecting E-Branchformer for its performance. They used additional datasets, including CDSD and over 2000 hours from WenetSpeech, in different fine-tuning stages. Their training involved: pre-training on WenetSpeech, CDSD, and MDSC using the data2vec[35] algorithm; fine-tuning on CDSD and MDSC; and final fine-tuning on MDSC for keyword spotting using max-pooling loss and cosine annealing learning rate decay. The system achieved a leaderboard-B score of 0.0105, ranking 3rd, and 1st in leaderboard-A with a score of 0.0267.

T019 replaced traditional FBank features with those extracted using the pre-trained Wav2vec[36] model from HuggingFace[37] to reduce error rates across all speakers. They used contrastive learning, pairing keyword samples from control and speakers with dysarthria, and computing contrastive loss based on Euclidean distance. This, combined with cross-entropy loss, improved robustness, with the best results at λ = 0.005. The system also integrates voice and prosodic features, capturing dysarthria-affected characteristics, such as speaking rate, fundamental frequency, harmonic-to-noise ratio, jitter, shimmer, cepstral peak prominence, and envelope modulated spectrum. These features, combined with neural network embeddings, enhance keyword classification performance. The final system showed improved FAR and FRR with contrastive learning.

T008 introduced two novel layers to enhance the Temporal Convolutional Network (TCN) performance under low-resource conditions: Frame-Wise Normalization (FWN) and Learned Feature Augmentation (LFA). Unlike other teams, T008 did not use pre-trained models or external datasets. The

FWN layer normalizes each frame individually, addressing variability in dysarthric speech. The LFA layer samples affine transformation parameters from Gaussian distributions within the BN layer, enhancing generalization and preventing overfitting. This approach, replacing BN with FWN and LFA in the baseline's three training stages, significantly improved performance, outperforming the baseline by 0.0176 in the final score on the Leaderboard-B.

4.3. Discussion and analysis

In analyzing the methodologies and results of the five teams, several commonalities and differences emerge, leading to meaningful conclusions about handling dysarthric speech in low-resource settings.

Most teams utilized large pre-trained models as the backbone of their systems and most of them implemented a multi-stage training process, typically including pre-training on general datasets, fine-tuning on dysarthric speech datasets, and further adaptation to specific speakers or keywords. These help transfer knowledge from large, generic datasets to the specific task of dysarthric speech recognition. To handle variability in dysarthric speech, some teams incorporated data augmentation and normalization techniques; for instance, T011 used dynamic data augmentation, T019 employed contrastive learning with prosodic features, and T008 introduced FWN and LFA layers. Enhancing model robustness was a primary focus, with T011's dual-filter strategy, T019's contrastive learning, and T008's feature augmentation layers all aimed at improving the system's ability to handle the inconsistencies and distortions typical in dysarthric speech.

The differences in model architectures and innovations were evident among the teams. T011 introduced a dual-filter strategy that combined wake-up word spotting probabilities and automatic speech recognition outputs; T021 utilized a cosine similarity-based classification approach along with a three-stage fine-tuning process; T018 evaluated multiple models, including MDTC, Squeezeformer, and E-Branchformer, ultimately selecting E-Branchformer for its superior performance; T019 focused on replacing traditional FBank features with those derived from Wav2vec2, integrating a range of voice and prosodic features; and T008 proposed novel layers, FWN and LFA, to enhance temporal convolutional network performance by addressing normalization and feature augmentation specifically for dysarthric speech. While all teams used the competition-provided dysarthria dataset, T018 supplemented their training with additional dysarthria datasets CDSD to enhance performance under low-resource conditions.

Regarding evaluation metrics and performance, T011 achieved the lowest FAR and FRR, securing the first place with a total score of 0.0082. T021 followed with a score of 0.0098. T018 ranked third on on leaderboard-B with 0.0105 and ranked first on leaderboard-A with 0.0267. T019 and T008 also reported significant improvements over the baseline.

As analyzed above, the following conclusions can be drawn:

- **Effectiveness of Pre-trained Models:** Pre-trained models significantly improve performance when fine-tuned for specific tasks like dysarthric speech recognition.
- **Importance of Multi-stage Training:** Multi-stage training allows models to gradually adapt from general to specific speech characteristics, improving accuracy and robustness.
- **Advanced Normalization and Augmentation:** Techniques like FWN, LFA, and contrastive learning effectively address variability in dysarthric speech, enhancing model resilience.
- **Use of Additional Datasets:** Incorporating additional datasets can improve performance in low-resource settings by providing diverse training examples.

Combining pre-trained models, innovative training techniques, and effective data augmentation is essential for advancing wake-up word spotting for dysarthric speech. These methods help mitigate challenges posed by variability and data scarcity, leading to more accurate and robust detection systems.

5. CONCLUSION

This paper presented the Low-Resource Dysarthria Wake-Up Word Spotting (LRDWWS) Challenge, a pioneering effort to address the significant challenges faced by individuals with dysarthria when interacting with voice-activated technologies. We provided a Mandarin Dysarthria Speech Corpus (MDSC) and its newly recorded evaluation subset MDSC-Eval, offering a robust dataset for training and evaluating wake-up word systems. Additionally, we proposed a comprehensive baseline system using convolutional neural networks and evaluation metrics designed to assess both false acceptance and false rejection rates. The innovative approaches of the participating teams highlighted novel model architectures and unique strategies for feature extraction and data augmentation, underscoring the potential of these systems to enhance the accessibility and usability of smart home technologies for individuals with dysarthria.

In the future, we intend to integrate diverse fields to investigate intriguing issues in pathology and speech of dysarthria, while simultaneously promoting social awareness and understanding of this condition. By doing so, we could eliminate discrimination and prejudice against people with dysarthria.

6. REFERENCES

[1] Ming Sun, David Snyder, Yixin Gao, Varun Nagaraja, Mike Rodehorst, Sankaran Panchapagesan, Nikko Ström, Spyros Matsoukas, and Shiv Vitaladevuni, "Compressed time delay neural network for small-footprint keyword spotting," 2017.

[2] Tara N Sainath and Carolina Parada, "Convolutional neural networks for small-footprint keyword spotting.," in *Interspeech*, 2015, pp. 1478–1482.

[3] Seungwoo Choi, Seokjun Seo, Beomjun Shin, Hyeongmin Byun, Martin Kersner, Beomsu Kim, Dongyoung Kim, and Sungjoo Ha, "Temporal convolution for real-time keyword spotting on mobile devices," *arXiv preprint arXiv:1904.03814*, 2019.

[4] Yixin Gao, Yuriy Mishchenko, Anish Shah, Spyros Matsoukas, and Shiv Vitaladevuni, "Towards data-efficient modeling for wake word spotting," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7479–7483.

[5] Iván López-Espejo, Zheng-Hua Tan, and Jesper Jensen, "A novel loss function and training strategy for noise-robust keyword spotting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2254–2266, 2021.

[6] Mete Ozay, "Joint embedding learning and latent subspace probing for cross-domain few-shot keyword spotting," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 6425–6429.

[7] Prakash Dhungana and Sayed Ahmad Salehi, "Rtkws: Real-time keyword spotting based on integer arithmetic for edge deployment," in *2024 25th International Symposium on Quality Electronic Design (ISQED)*. IEEE, 2024, pp. 1–7.

[8] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018.

[9] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan, "Recurrent neural network transducer for audio-visual speech recognition," in *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE, 2019, pp. 905–912.

[10] Haoxu Wang, Ming Cheng, Qiang Fu, and Ming Li, "The dku post-challenge audio-visual wake word spotting system for the 2021 misp challenge: Deep analysis," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas Huang, Kenneth Watkin, and Simone Frame, "Dysarthric Speech Database for Universal Access Research," 2008.

[12] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, Dec. 2012.

[13] Mengyi Sun, Ming Gao, Xinchen Kang, Shiru Wang, Jun Du, Dengfeng Yao, and Su-Jing Wang, "CDSD: Chinese Dysarthria Speech Database," Oct. 2023.

[14] Ahmad Almadhor, Rizwana Irfan, Jiechao Gao, Nasir Saleem, Hafiz Tayyab Rauf, and Seifedine Kadry, "E2e-dasr: End-to-end deep learning-based dysarthric automatic speech recognition," *Expert Systems with Applications*, vol. 222, pp. 119797, 2023.

[15] Seyed Reza Shahamiri, Vanshika Lal, and Dhvani Shah, "Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[16] Chongchong Yu, Xiaosu Su, and Zhaopeng Qian, "Multi-stage audio-visual fusion for dysarthric speech recognition with pre-trained models," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1912–1921, 2023.

[17] Chitralekha Bhat, Ashish Panda, and Helmer Strik, "Improved asr performance for dysarthric speech using two-stage dataaugmentation.," in *INTERSPEECH*, 2022, pp. 46–50.

[18] Andrew Hu, Dhruv Phadnis, and Seyed Reza Shahamiri, "Generating synthetic dysarthric speech to overcome dysarthria acoustic data scarcity," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 6, pp. 6751–6768, 2023.

[19] TA Mariya Celin, P Vijayalakshmi, and T Nagarajan, "Data augmentation techniques for transfer learning-based continuous dysarthric speech recognition," *Circuits, Systems, and Signal Processing*, vol. 42, no. 1, pp. 601–622, 2023.

[20] Wei-Zhong Zheng, Ji-Yan Han, Chen-Kai Lee, Yu-Yi Lin, Shu-Han Chang, and Ying-Hui Lai, "Phonetic posteriorgram-based voice conversion system to improve speech intelligibility of dysarthric patients," *Computer Methods and Programs in Biomedicine*, vol. 215, pp. 106602, 2022.

[21] Wei-Zhong Zheng, Ji-Yan Han, Chen-Yu Chen, Yuh-Jer Chang, and Ying-Hui Lai, "Improving the efficiency of dysarthria voice conversion system based on data augmentation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 4613–4623, 2023.

[22] Hadil Mehrez, Mounira Chaiani, and Sid Ahmed Selouani, "Using starganv2 voice conversion to enhance the quality of dysarthric speech," in *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 2024, pp. 738–744.

[23] Bassam Ali Al-Qatab and Mumtaz Begum Mustafa, "Classification of dysarthric speech according to the severity of impairment: an analysis of acoustic features," *IEEE Access*, vol. 9, pp. 18183–18194, 2021.

[24] Guilherme Schu, Parvaneh Janbakhshi, and Ina Kodrasi, "On using the ua-speech and torgo databases to validate automatic dysarthric speech classification approaches," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[25] Amlu Anna Joshy and Rajeev Rajan, "Dysarthria severity classification using multi-head attention and multi-task learning," *Speech Communication*, vol. 147, pp. 1–11, 2023.

[26] Ming Gao, Hang Chen, Jun Du, Xin Xu, Hongxiao Guo, Hui Bu, Jianxing Yang, Ming Li, and Chin-Hui Lee, "Enhancing voice wake-up for dysarthria: Mandarin dysarthria speech corpus release and customized system design," 2024.

[27] Jie Wang, Menglong Xu, Jingyong Hou, Binbin Zhang, Xiao-Lei Zhang, Lei Xie, and Fuping Pan, "WeKws: A production first small-footprint end-to-end Keyword Spotting Toolkit," Oct. 2022.

[28] Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli, "Efficient self-supervised learning with contextualized target representations for vision, speech and language," in *International Conference on Machine Learning*. PMLR, 2023, pp. 1416–1429.

[29] Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," *arXiv preprint arXiv:2206.08317*, 2022.

[30] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[31] Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al., "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6182–6186.

[32] Jingyong Hou, Lei Xie, and Shilei Zhang, "Two-stage streaming keyword detection and localization with multi-scale depthwise temporal convolution," *Neural Networks*, vol. 150, pp. 28–42, 2022.

[33] Sehoon Kim, Amir Gholami, Albert Shaw, Nicholas Lee, Karttikeya Mangalam, Jitendra Malik, Michael W Mahoney, and Kurt Keutzer, "Squeezeformer: An efficient transformer for automatic speech recognition," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9361–9373, 2022.

[34] Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J Han, and Shinji Watanabe, "E-branchformer: Branchformer with enhanced merging for speech recognition," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 84–91.

[35] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[36] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[37] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.