

# CDSD: Chinese Dysarthria Speech Database

Yan Wang<sup>#1,2</sup>, Mengyi Sun<sup>#1,2</sup>, Xinchun Kang<sup>3</sup>, Jingting Li<sup>1,2</sup>, Pengfei Guo<sup>4</sup>, Ming Gao<sup>5</sup>, Su-Jing Wang<sup>\*1,2</sup>

<sup>1</sup>CAS Key Laboratory of Behavioral Science, Institute of Psychology; <sup>2</sup>Department of Psychology, University of the Chinese Academy of Sciences; <sup>3</sup>Beijing Union University; <sup>4</sup> Jiangsu University of Science and Technology; <sup>5</sup>University of Science and Technology of China

wangsujiang@psych.ac.cn

## Abstract

Dysarthric speech poses significant challenges for individuals with dysarthria, impacting their ability to communicate socially. Despite the widespread use of Automatic Speech Recognition (ASR), accurately recognizing dysarthric speech remains a formidable task, largely due to the limited availability of dysarthric speech data. To address this gap, we developed the Chinese Dysarthria Speech Database (CDSD), the most extensive collection of Chinese dysarthria data to date, featuring 133 hours of recordings from 44 speakers. Our benchmarks reveal a best Character Error Rate (CER) of 16.4%. Compared to the CER of 20.45% from our additional human experiments, Dysarthric Speech Recognition (DSR) demonstrates its potential in significant improvement of communication for individuals with dysarthria. The CDSD database will be made publicly available at <http://melab.psych.ac.cn/CDSD.html>.

**Index Terms:** Database, Dysarthria, Dysarthric Speech Recognition, Mandarin, Cerebral Palsy

## 1. Introduction

Speech conversation is the fundamental to human communication [1, 2], and more proficient communication skills often reflect higher sociability [3]. Dysarthria adversely affects individuals suffering from it by hampering their social relationships [4] and negatively impacting their mental health [5]. Studies have shown that dysarthria may place adolescents at increased risk for mental and physical health [6]. This suggests that fluent speech plays a pivotal role in fostering human social interactions and supporting psychological health.

Dysarthria, characterized by impaired speech due to motor deficits, is primarily caused by conditions such as cerebral palsy, Parkinson’s disease [7], amyotrophic lateral sclerosis [8], and stroke [9]. Depending on the etiologies cause, individuals with dysarthria may exhibit varying degrees of speech issues, such as slowed speech rate, imprecise pronunciation, irrational pauses, and decreased clarity [10, 11]. The speech of individuals with dysarthria poses challenges in their social communication and human-computer interaction (HCI). Especially, traditional automated speech recognition (ASR) systems often struggle to identify dysarthric speech.

ASR offers a more convenient operation and has many applications, including smart home and voice assistants, primarily serving the needs of healthy individuals. Meantime, disabled people need such applications even more. Specifically, considering the etiologies of dysarthria, such as cerebral palsy, individuals with such disorders frequently encounter mobility challenges. They may require a more convenient form of interaction with computers [12, 13], with speech interaction serving as an ideal solution to this need.

However, ASR for individuals with dysarthria still presents challenges [14, 15]. Despite the high accuracy of ASR in interacting with healthy individuals, it serves individuals with dysarthria less effectively.

The Dysarthric Speech Recognition (DSR) system, designed to recognize speech in individuals with dysarthria, can significantly enhance their quality of life. However, the development of DSR is still faced with formidable challenges. First, there are significant variability in dysarthric speech due to different degrees of severity or diverse etiologies, requiring adaptive modeling approaches. Second, enhancing the performance of the DSR system also demands extensive dysarthric speech data. Furthermore, in DSR under data sparsity, the challenge of multi-feature and multi-modal data fusion remains.

Hernandez et al. [16] demonstrated that speech representations pre-trained on extensive unlabeled data by models such as Wav2vec could significantly enhance ASR performance for dysarthric speech. Baskar et al. [17] delved into integrating Wave2vec with either fMLLR features or x-vectors during the fine-tuning process. Meanwhile, Violeta et al. [18] explored the efficacy of self-supervised learning frameworks, namely Wav2vec 2.0 and WavLM. Their comparative analysis revealed that the optimal Word Error Rate (WER) for severe dysarthric speech could reach 51.8%. However, the significant disparity between normal and dysarthric speech acts as a limiting factor for performance enhancement. To bridge this gap, it is imperative to identify an effective approach. By pre-training on extensive speech corpora, neural networks can effectively learn prior knowledge from the training data. This strategy highlights the potential of utilizing large-scale speech data to refine representational models. Consequently, by fine-tuning based on specific data in downstream tasks, the capability of ASR systems to address the challenges presented by dysarthric speech can be significantly improved.

In addition to acoustic features, vision constitutes another modality in human speech perception, indicating a bi-modal process [19]. Furthermore, visual features remain unaffected by any degradation in acoustic signals and can thus provide valuable compensatory information for ASR systems. In this context, the integration of visual information to enhance DSR performance, has been the subject of recent research efforts. For instance, the utilization of Bayesian gated control facilitates a strong integration of audio and visual modalities [20]. Furthermore, the multi-modal integration also help to address challenges such as severe voice quality degradation and the pronounced mismatch between dysarthric and normal speech patterns. For instance, Liu et al. [21] proposed a novel strategy involving the generation of cross-domain visual features. Yu et al. [22] developed a novel multi-stage fusion framework to improve the effectiveness of DSR systems.

Compared to the typical speech databases, the sizes of dysarthric speech databases are pretty small. However, individuals with dysarthria have difficulty speaking, so collecting speech data from individuals with dysarthria is challenging. Moreover, annotating speech data from individuals with dysarthria is particularly challenging, further increasing the difficulty of constructing databases for dysarthric speech. Therefore, dysarthric speech databases are relatively scarce compared to typical speech databases.

The constructed databases for disordered speech, such as Whitaker [23], UA-Speech [24], Torgo [25], EasyCall [26], and Euphonia corpora [27], are predominantly English-speaking disordered speech databases. The distinct pronunciation, morphology, and syntax of Chinese, along with its many homophones and polyphones, mean that English dysarthric speech databases fall short for Chinese DSR tasks. Clearly, advancing Chinese DSR urgently requires a substantial Chinese dysarthria speech dataset for training.

Existing Chinese dysarthria speech databases include the Cantonese Dysarthric Speech Corpus [28] and the Mandarin Subacute Stroke Dysarthric Multimodal database [29]. However, the size of these existing Chinese dysarthric speech databases does not exceed 10 hours, which is relatively small. To address these resource gaps, we are committed to developing a Chinese dysarthric speech database for DSR training, focusing on individuals with cerebral palsy.

To our knowledge, the Chinese Dysarthria Speech Database (CDSD) is not only among the rare Chinese dysarthria speech databases available but also stands as the largest in scale compared to any existing databases of its kind. With a total duration of 133 hours, CDSD represents a more extensive resource, offering a broader sample for the study of challenging speech in the Chinese language context. Furthermore, through comparative experiments between human and computer, DSR has shown the potential to improve social communication of individual with dysarthria. Finally, we conducted a preliminary exploration into the optimal data collection length for speaker-dependent DSR by comparing training sets of various sizes.

## 2. Database construction

The CDSD database has collected speech data from 44 speakers, including 124 hours of audio data and 9 hours video data. Based on the two different durations of each subject’s recording, we divided the database into two parts. Specifically, Part A includes 44 hours of audio data from all 44 speakers, 1 hour of recording each. Additionally, it includes video recordings that are synchronized with the audio from the nine speakers. Part B consists of 80 hours of audio data, 10 hours recorded by each of the 8 speakers in Part A. The detail of CDSD is listed in Table 1.

Table 1: *CDSD overview. # indicates the amount of the speakers, and D/P represents the recording duration per speaker. \*The speakers in Part B were also involved in Part A.*

CDSD	#	D/P	Total audio	Total video
Part A	44	1 h	44 h	9 h
Part B	8*	10 h	80 h	\
All	44	\	124 h	9 h

### 2.1. Data collection

**Participants:** In the preparatory phase of data collection, each speaker signed informed consent prior to the recording. A total of 44 speakers were recruited, 39 speakers over 18 years old, and the remaining 5 speakers were younger than 18. For those minor speakers, parental or guardian consent was obtained, along with the minors’ assent. Additionally, the 9 speakers recorded on video were provided with additional informed consent forms. Speakers were informed of their right to discontinue participation at any point without consequence.

To protect speakers’ privacy, speakers’ names were not collected in CDSD. Instead, after speech data collection, each data was assigned a speaker representative serial number. In addition, we have collected information about the speakers’ etiologies of dysarthria, recording environments, ages, accents, and other speaker-related details. The overall information of the speakers in the CDSD is shown in Table 2.

Table 2: *The overall information of the speakers.*

Factor	Categorization	Overall (N=44)
Sex	Female	18
	Male	26
Age	Adults	39
	Children	5
Etiology	Cerebral Palsy	33
	Other Disease	11
Recording devices	Smartphone	39
	ZOOM F8n	5

Notably, the speaker cohort for this database includes two authors of this article, both diagnosed with dysarthria. Their dual roles as researchers and speakers enriched the study with nuanced insights, fostering a research environment characterized by inclusivity. For example, prior to the formal recording, we asked one of the authors to record for 10 hours as the first speaker. He/She finds it challenging to speak continuously for long duration, as it may affect pronunciation accuracy. Therefore, speakers were allowed to submit recordings in segments based on their individual conditions. Furthermore, sentences reported by the first speaker as challenging to articulate were excluded from subsequent data collection phases.

**Text pool:** The text pool for constructing the CDSD consists of two types of texts. In particular, the first type is sourced from the AISHELL-1 database [30]. This text pool cover various domains of language diversity, providing a better representation of daily speech usage. Additionally, it encompass a wide range of commonly used Chinese words and characters, enhancing the database’s universality and improving model robustness. Furthermore, texts are purged of inappropriate content related to sensitive political issues, user privacy, pornography, violence, etc. Meantime, the second type comprises elementary and middle school speeches and fairy tales. This is because that a small portion of our speakers are children. To accommodate children’s reading habits and match children’s level of literacy knowledge, we extracted speeches and fairy tales from the internet as the second type of text.

**Recording devices:** Two types of recording devices were used to collect data: the ZOOM F8n field recorder and smartphone. Specifically, some speakers recorded their audio in

a recording studio of approximately 10 square meters using professional recording equipment — the ZOOM F8n field recorder. The ZOOM F8n has high-quality preamplifier analog input channels, producing superior audio recording quality. This high-quality recording allows for the minimization of potential interference factors in subsequent data analysis. Meanwhile, to accommodate speakers with motor impairments, speakers could record their audio comfortably using Smartphones within the tranquil confines of their home environment. Using smartphone recording effectively simulates the speakers’ daily speech recognition tasks, enhancing the ecological validity of the audio data and improving the robustness of the speech recognition model. Additionally, video data of some speakers were recorded using smartphone, synchronized with their corresponding audio recordings. The acquisition of multimodal data could facilitate the improvement of speech recognition accuracy.

All recording devices underwent rigorous testing before formal recording to ensure efficiency and compliance with the required recording standards. This process ensured the audio data collection with high quality and without any noticeable loss or distortion. Additionally, audio data recorded in different scenarios and with different devices were stored in WAV format and analyzed using the same recognition algorithm.

**Recording process:** First, the quietude of the recording environment was imperative, regardless of whether the setting is a recording studio or a home. Then, during the audio and video recording processes, speakers were instructed to ensure that the microphone was approximately 20-40 centimeters away from their mouth. Additionally, they were required to maintain stable recording equipment and consistent volume levels throughout the recording process. Throughout the video recording process, speakers were instructed to maintain their facial images centrally aligned on the screen, ensuring both shoulders and the full movement of their lips were clearly visible.

## 2.2. Data annotation

Due to the substantial differences between the speech of individuals with dysarthria and healthy individuals, as well as the differences in the speech of each individual with dysarthria, annotating the speech of dysarthric speakers posed particular challenges. Speakers might have made reading errors or skipped words when reading text, requiring annotators to confirm the speech’s accuracy repeatedly. Additionally, some speakers had severe dysarthria, resulting in very unclear pronunciation. And some other speakers had noticeable regional accents. Both issues made speech recognition more complicated for annotators.

The audio annotation task was conducted by five proficient annotators using the AIBIAOKE annotation platform<sup>1</sup>. Prior to annotation, all annotators underwent standardized training and adhered to uniform editing standards and procedures. Then, annotators reviewed the quality of all audio data and contacted the speaker for re-recording if necessary. After the audio data passed inspection, it was imported into the AIBIAOKE annotation platform. Annotators transcribed the text verbatim based on the original audio content to ensure consistency between the audio and the text. A 0.1-second buffer was placed before and after each marked speech waveform to prevent any “clipping” phenomenon. Non-speech segments of human voice lasting  $\geq 0.5$  seconds, such as static noise, laughter, breathing, coughing, and singing, were marked as NOISE.

<sup>1</sup><http://124.243.239.193:8081/#/login>

## 3. Experiments

Experiments are conducted on Part A and Part B of CDS. In particular, Part A comprises 44 speakers, with each contributing roughly one hour of data, amounting to a total of approximately 44 hours. The data of Speaker #2 in Part B was removed from the experiment because it was incorrectly annotated. The correct version has now been updated to the published database. Therefore, Part B remains 7 speakers for performance comparison, with each contributing roughly ten hours of data, amounting to a total of approximately 70 hours. The data is segmented into training, development, and test sets in an 8:1:1 ratio.

Two kinds of feature: Fbank and Wav2vec, are utilized in our experiments. The Fbank serves as a traditional spectrum feature, while Wav2vec, a self-supervised learning representation, significantly enhances DSR tasks [16, 17, 18]. The Wav2vec is modeled on WenetSpeech<sup>2</sup>.

An end-to-end model is trained as the baseline model based on the ESPnet toolkit. Specifically, Conformer [31] is used as the ASR config and RNN is used as the inference config. We adopted two training approaches for our models. Initially, we directly train models on Part A or B of CDS. Separately, we first train models on the AISHELL-1 [30], AISHELL-2 [32], and WenetSpeech [33] databases to obtain pre-trained models respectively, which we then fine-tune on CDS Part A or B. Table 3 lists character error rates (CERs). Due to computational resource limitations, the model was not trained on WenetSpeech database using Wav2vec features.

The scales of Part A / B of CDS, AISHELL-1, AISHELL-2, and WenetSpeech are 40+ / 80+, 100+, 1000+ and 10000+ hours, respectively. As the data scale increases (referring to the pre-training), the CERs of using Fbank as features decreases, and shows a significant reduction with pretraining on WenetSpeech. Meanwhile, the Wav2vec feature has better performance on the case of modelling directly on the CDS without any pre-trained models. This proves that Wav2vec can effectively represent the strong variability of dysarthric speech. However, this performance did not surpass that achieved using the Fbank acoustic features in AISHELL-1 and AISHELL-2, attributing the limitation to the significant disparity between normal and dysarthric speech.

We designed an experiment comparing human and computer on dysarthric speech recognition. The experiment included 10 participants, divided into two groups of five each based on whether they had experience communicating with individuals with dysarthric speech. All participants were asked to recognize and transcribe 17 dysarthric speech utterances within a 10-minute time limit. All utterances were randomly selected from the CDS Part A, were approximately 4 seconds in length, and were ensured to be clear and unambiguous.

The mean CER of the participants with experience in communicating with dysarthric speakers was 20.45%, and the inexperienced participants were 35.71%. Such comparisons show that sufficient experience is required to understand dysarthric speech accurately. Meantime, compared with the best result in Table 3, the computer’s speech recognition ability is superior to that of humans, highlighting the potential of DSR to enhance social interaction for individuals with dysarthria.

In Table 3, superior average speech recognition performance is demonstrated on Part A compared to Part B. Despite Part B having a longer data duration compared to Part A, Part A has a higher number of subjects (70 hours / 7 vs. 44 hours

<sup>2</sup>[https://github.com/TencentGameMate/chinese\\_speech\\_pretrain](https://github.com/TencentGameMate/chinese_speech_pretrain)

Table 3: CERs with Fbank and Wav2vec on various pre-trained models. [D/T] D and T mean the CER on the development set and the CER on test set. The arrow “→” indicates that pre-training is conducted on the databases of the former, followed by fine-tuning on the Part A or Part B of the CDSB database.

	Features	Part X	AISHELL-1→Part X	AISHELL-2→Part X	WenetSpeech→Part X
<b>Part A</b>	Fbank	24.9 / 24.9	20.9 / 21.2	20.5 / 20.7	16.5 / 16.4
	Wav2vec	22.0 / 22.2	25.6 / 25.6	26.9 / 27.1	No Answer
<b>Part B</b>	Fbank	31.9 / 30.2	28.4 / 26.8	26.3 / 24.7	23.7 / 22.2
	Wav2vec	30.4 / 28.7	30.8 / 29.2	30.5 / 29.0	No Answer

Table 4: CERs of computer and human transcription of each speaker in Part B.

Speaker ID	#1	#4	#6	#8	#9	#12	#20
Computer	11.9 / 11.6	29.3 / 27.0	37.9 / 37.1	21.7 / 19.4	20.1 / 19.8	11.1 / 10.5	8.5 / 9.4
Human	6.2	10.8	60.5	48.9	28.2	12.8	15.5

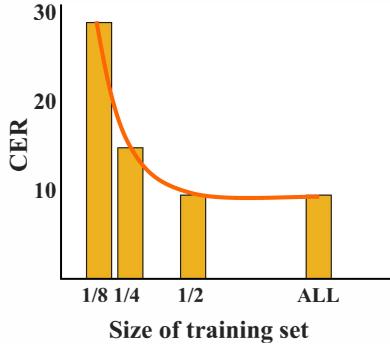


Figure 1: Optimal training data quantity determination.

/ 44). This indicates that under the limitation of subject quantity, the model exhibits poorer generalization ability, unable to learn a more diverse set of speech characteristics. This piqued our curiosity about the model’s performance in a single-subject speech environment. Consequently, we put Part B into speaker-dependent model training, using Fbank features, and calculated CER with each speaker’s sample in Part B as the development set and test set. The experimental results are shown in Table 4.

Besides, we compared the recognition result of computer with that of participants who had experience communicating with individuals with dysarthria. We randomly selected 10 utterances for each speaker in Part B. The participants were asked to recognize these utterances and their CERs were calculated.

More than half of the comparisons in Table 4 show that the computer’s DSR is better than the human’s. However, each dysarthric speaker has various acoustic features, which makes the DSR performance of the computer exhibits fluctuations. Compared to unimpaired speech, there are longer pauses in their expression and a lack of coherence between syllables. This characteristic has led to challenges in model training due to a limited database, possibly causing an over-reliance on pre-trained features. Hence, the model struggles to construct contextual relationships from clear speech punctuated by atypical pauses, often misclassifying the pronunciations from this speaker and resulting in a high CER.

Low-resource speech recognition aims to achieve optimal

performance using minimal data. To determine the optimal training data quantity for speaker-dependent DSR performance, we compared the DSR performance of the speaker with the lowest CER listed in Table 4 across varying training durations (in a proportional decrement). Meanwhile, the durations of the development and test sets remained constant at 1 hour. As illustrated in Fig. 1, when the training data duration is greater than or equal to 4 hours, the improvement in model performance becomes markedly modest. From this, it can be concluded that fine-tuning and training speech recognition for dysarthric speakers, using pretrained models from large-scale datasets like WenetSpeech, requires approximately 5 hours of data. This finding aids in future dysarthric speech data collection, reducing database construction costs, especially when acquiring and annotating such data is challenging.

## 4. Conclusion

We have presented the Chinese Dysarthria Speech Database (CDSB), which includes 133 hours of data collected from 44 speakers. As we know, CDSB is the largest database of dysarthria in Chinese. The samples were collected in different scenarios and devices to ensure the high quality of the audio and to better match the daily life of dysarthric speech. The construction of the CDSB aims to improve the performance of DSR for Chinese, empowering practical and convenient DSR technology for individuals with dysarthria in China.

## 5. Acknowledgements

Yan Wang and Mengyi Sun contributed equally to this research. We would like to express our gratitude to all the participants in our research. And we also extend our thanks to the Hangzhou Xiaoshan Noah Cerebral Palsy Service Center, the Angel House Rehabilitation and Educational Activity Center, Our Family China, and the Little Snail Family Support Center for People with Disabilities. This research was partially funded by 1) the National Natural Science Foundation of China (62276252, 62106256); 2) the Youth Innovation Promotion Association CAS.

## 6. References

- [1] B. Jowett *et al.*, *The politics of Aristotle: I.* Clarendon Press, 1887.
- [2] J. H. Manson, G. A. Bryant, M. M. Gervais, and M. A. Kline, "Convergence of speech rate in conversation predicts cooperation," *Evolution and Human Behavior*, vol. 34, no. 6, pp. 419–426, 2013.
- [3] R. E. Riggio, "Assessment of basic social skills," *Journal of Personality and Social Psychology*, vol. 51, no. 3, p. 649, 1986.
- [4] S. Braithwaite and J. Holt-Lunstad, "Romantic relationships and mental health," *Current Opinion in Psychology*, vol. 13, pp. 120–125, 2017.
- [5] A. D. Palmer, J. T. Newsom, and K. S. Rook, "How does difficulty communicating affect the social relationships of older adults? An exploration using data from a national survey," *Journal of communication disorders*, vol. 62, pp. 131–146, 2016.
- [6] M. L. Rice, M. A. Sell, and P. A. Hadley, "Social interactions of speech, and language-impaired children," *Journal of Speech, Language, and Hearing Research*, vol. 34, no. 6, pp. 1299–1307, 1991.
- [7] S. Scott and F. Caird, "Speech therapy for Parkinson's disease," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 46, no. 2, pp. 140–144, 1983.
- [8] T. Makkonen, H. Ruottinen, R. Puhto, M. Helminen, and J. Palmio, "Speech deterioration in amyotrophic lateral sclerosis (ALS) after manifestation of bulbar symptoms," *International journal of language & communication disorders*, vol. 53, no. 2, pp. 385–392, 2018.
- [9] P. Jerntorp and G. Berglund, "Stroke registry in malmö, sweden," *Stroke*, vol. 23, no. 3, pp. 357–361, 1992.
- [10] R. D. Kent, "Research on speech motor control and its disorders: A review and prospective," *Journal of Communication disorders*, vol. 33, no. 5, pp. 391–428, 2000.
- [11] H. P. Rowe, S. E. Gutz, M. F. Maffei, K. Tomanek, and J. R. Green, "Characterizing dysarthria diversity for automatic speech recognition: A tutorial from the clinical perspective," *Frontiers in Computer Science*, vol. 4, p. 770210, 2022.
- [12] S. Liu, M. Geng, S. Hu, X. Xie, M. Cui, J. Yu, X. Liu, and H. Meng, "Recent progress in the CUHK dysarthric speech recognition system," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2267–2281, 2021.
- [13] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt, A. Hassidim, and Y. Matias, "Personalizing ASR for dysarthric and accented speech with limited data," in *Proc. Interspeech*, 2019, pp. 784–788.
- [14] S. Hu, X. Xie, Z. Jin, M. Geng, Y. Wang, M. Cui, J. Deng, X. Liu, and H. Meng, "Exploring self-supervised pre-trained ASR models for dysarthric and elderly speech recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [15] M. Geng, X. Xie, Z. Ye, T. Wang, G. Li, S. Hu, X. Liu, and H. Meng, "Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2597–2611, 2022.
- [16] A. Hernandez, P. A. Pérez-Toro, E. Noeth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," in *Proc. Interspeech*, 2022, pp. 51–55.
- [17] M. K. Baskar, T. Herzig, D. Nguyen, M. Diez, T. Polzehl, L. Burget, and J. Černocký, "Speaker adaptation for Wav2vec2 based dysarthric ASR," in *Proc. Interspeech*, 2022, pp. 3403–3407.
- [18] L. P. Violeta, W.-C. Huang, and T. Toda, "Investigating self-supervised pretraining frameworks for pathological speech recognition," in *Proc. Interspeech*, 2022, pp. 41–45.
- [19] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.
- [20] S. Liu, S. Hu, Y. Wang, J. Yu, R. Su, X. Liu, and H. Meng, "Exploiting visual features using Bayesian gated neural networks for disordered speech recognition," in *Proc. Interspeech*, 2019, pp. 4120–4124.
- [21] S. Liu, X. Xie, J. Yu, S. Hu, M. Geng, R. Su, S.-X. Zhang, X. Liu, and H. Meng, "Exploiting cross-domain visual feature generation for disordered speech recognition," in *Proc. Interspeech*, 2020, pp. 711–715.
- [22] C. Yu, X. Su, and Z. 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.
- [23] J. Deller Jr, M. Liu, L. Ferrier, and P. Robichaud, "The Whitaker database of dysarthric (cerebral palsy) speech," *The Journal of the Acoustical Society of America*, vol. 93, no. 6, pp. 3516–3518, 1993.
- [24] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.
- [25] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, pp. 523–541, 2012.
- [26] R. Turrisi, A. Braccia, M. Emanuele, S. Giulietti, M. Pugliatti, M. Sensi, L. Fadiga, and L. Badino, "EasyCall Corpus: A Dysarthric Speech Dataset," in *Proc. Interspeech 2021*, 2021, pp. 41–45.
- [27] R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner, P. C. Nelson *et al.*, "Disordered speech data collection: Lessons learned at 1 million utterances from project Euphonia," in *Proc. Interspeech*, 2021, pp. 4833–4837.
- [28] K. H. Wong, Y. T. Yeung, E. H. Chan, P. C. Wong, G.-A. Levow, and H. Meng, "Development of a cantonese dysarthric speech corpus," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [29] J. Liu, X. Du, S. Lu, Y.-M. Zhang, H. An-ming, M. L. Ng, R. Su, L. Wang, and N. Yan, "Audio-video database from subacute stroke patients for dysarthric speech intelligence assessment and preliminary analysis," *Biomedical Signal Processing and Control*, vol. 79, p. 104161, 2023.
- [30] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.
- [31] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.
- [32] J. Du, X. Na, X. Liu, and H. Bu, "AISHELL-2: Transforming mandarin ASR research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.
- [33] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. 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.