Re-Sonance: A Dysarthric Asynchronous Real-Time Speech Conversion System Based on a Three-Stage Cascaded ASR-LLM-TTS Architecture

Yuxuan Wu, Yifan Xu, Junkun Wang, Jiayong Jiang, Xin Zhao, and Zhaojie Luo*

[1] State Key Laboratory of Digital Medical Engineering, Southeast University, Nanjing, China
[2] School of Biological Science & Medical Engineering, Southeast University, Nanjing, China

Abstract. Individuals with dysarthria face significant challenges in professional speaking scenarios such as conferences, presentations, and meetings, where real-time communication is crucial. While existing Augmentative and Alternative Communication (AAC) systems provide basic support, they often fail to meet the demands of professional speaking environments due to high latency and unnatural speech patterns. This paper presents Re-Sonance, a novel LLM-enhanced speech-driven AAC system designed for real-time professional speaking scenarios. By integrating Whisper ASR, Qwen LLM, and CosyVoice TTS, Re-Sonance achieves improved speech intelligibility and naturalness while maintaining real-time performance. Both subjective and objective evaluations using a Mandarin dysarthric speech dataset demonstrate that our speech reconstruction approach significantly improved intelligibility while preserving semantic coherence for speakers with mild to moderate dysarthria. Although performance remains limited for severe dysarthria cases, our findings validate the potential of LLM-based methods for enhancing speech-driven AAC systems, paving the way for more effective and accessible communication technologies.

Keywords: Dysarthria · Augmentative and Alternative Communication · Large Language Models · Real-time Communication · Assistive Technology.

1 Introduction

Dysarthria, a group of congenital or trauma-induced neuromotor disorders, impairs articulation and communication[22], significantly affecting patients' linguistic abilities, social interactions, and quality of life. Although advances in

* Corresponding author, email: luozhaojie@seu.edu.cn

Augmentative and Alternative Communication (AAC) technologies have improved expression for individuals with dysarthria, recent research has largely emphasized speech replacement, aiming to reduce user input and streamline workflows[24]. However, these methods disrupt natural communication patterns and face limitations in comprehensibility and real-time performance.

The unnatural interaction imposed by speech replacement can undermine users' self-identity, social recognition, and motivation to use assistive systems[1, 10]. Additionally, their operational complexity and latency hinder use in real-time settings like video conferences and public speaking. To address these issues, we propose an AAC system tailored for real-time meetings and presentations, enabling natural participation through intuitive, minimally intrusive interaction.

Building on recent models, our system ensures broad accessibility across devices and languages. The preprocessing module uses lightweight speech recognition for initial interpretation of dysarthric speech, followed by large language models (LLMs) to refine output and mitigate recognition errors through prompt engineering. The output module employs lightweight text-to-speech (TTS) to produce natural, intelligible speech.

This work bridges key gaps in AAC systems and advances speech-driven assistive communication, contributing to global efforts toward communication equity for individuals with disabilities.

2 Related Works

2.1 Human-Computer Interaction for Individuals with Disabilities

Research on Augmentative and Alternative Communication (AAC) has long explored tools such as eye-tracking[3] and touch-based keyboards with features like word prediction and phrase storage[16, 25, 17]. These are effective for severe impairments but less suitable for dysarthria, as replacing speech reduces communication speed[12] and may undermine self-identity or rehabilitation opportunities[5].

2.2 Speech-Driven AAC Systems

Speech provides a natural, high-bandwidth channel[18]. Existing work falls into two main approaches: signal processing and ASR-TTS frameworks[30].

Signal Processing Methods Deep learning methods enhance or convert speech, e.g., accent[11], emotion[15, 19], or dysarthric-to-healthy patterns[4, 28]. Tools like Wesper convert whispered to normal speech[21]. However, these approaches often demand high resources and may not improve intelligibility in practice.

ASR-TTS Framework ASR-TTS systems are widely applied in AAC[29, 13], but inconsistent ASR errors remain a barrier[9, 23]. Personalized systems like Google's Project Relate address this[1], though challenges remain in language coverage and domain-specific terms[6].

2.3 LLM-Driven Assistive Systems

Large Language Models (LLMs) enable new assistive applications. Rambler summarizes ASR outputs[14], AscleAI refines clinical transcripts[8], and Valencia et al. show improved AAC interaction via speech macros[26]. These works highlight LLMs' potential to correct ASR errors and enhance clarity in real-time communication.

3 Method

3.1 Re-Sonance Prototype

The Re-Sonance system is designed to convert pathological speech from individuals with dysarthria into healthy speech. The technical framework of Re-Sonance consists of three primary components (Fig. 1), which function collaboratively to transform dysarthric speech into more intelligible output.

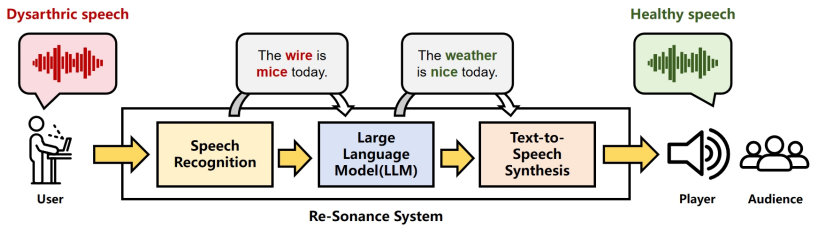

Fig. 1. Technical framework and application scenarios of Re-Sonance

For speech recognition, we employ Whisper, one of the most advanced open-source ASR models developed by OpenAI[20]. Released in 2022, Whisper supports 99 languages and can be deployed on low-performance devices, demonstrating robustness and accuracy in recognizing healthy speech. While some fine-tuned versions of Whisper exist for dysarthric speech, this study utilizes the unmodified Whisper-Turbo model. For the large language model, we use Qwen-Plus, an enhanced large-scale language model developed by Alibaba Cloud[2]. Qwen-Plus excels in language comprehension and generation, supports multiple languages, and optimally balances performance, speed, and cost, making it suitable for diverse applications. For TTS synthesis, we employ CosyVoice, an advanced speech synthesis tool that generates natural-sounding speech with support for multiple languages and personalized adjustments[7]. Additionally, CosyVoice offers voice cloning capabilities, enabling highly realistic speech synthesis from a few seconds of audio samples.

3.2 Asynchronous Real-Time Processing

To achieve low-latency interaction, Re-Sonance adopts an asynchronous real-time design. Here, "asynchronous" refers to the non-blocking pipeline architecture, where different modules (ASR, LLM, and TTS) operate in a streaming and overlapping manner rather than strictly sequential execution. Specifically, the Whisper ASR continuously transcribes dysarthric speech into partial text segments. These segments are immediately passed to the Qwen LLM for correction and refinement, without waiting for the entire utterance to finish. In turn, the corrected segments are incrementally fed into CosyVoice TTS, which can synthesize speech outputs on-the-fly. This overlapping process effectively reduces end-to-end latency compared with synchronous pipelines, ensuring that the user experiences smooth and near-instantaneous communication.

3.3 Interface

The Re-Sonance interface consists of input and output control components, optimized for both mobile and desktop platforms[3]. Fig. 2 presents the user interface of the Re-Sonance system.

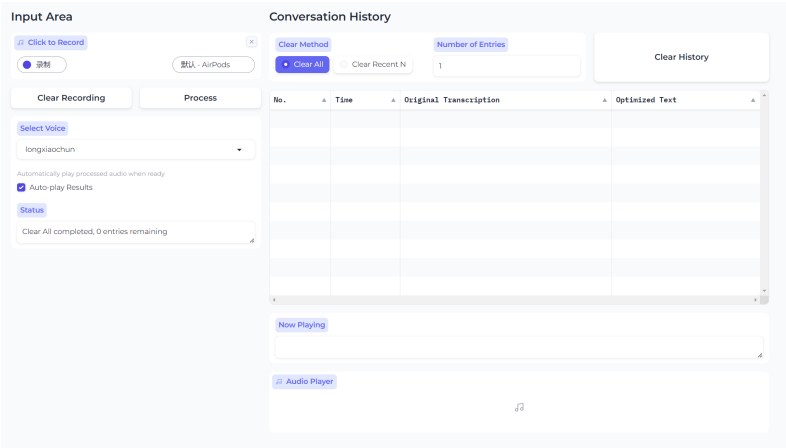

Fig. 2. Interface of the Re-Sonance

3.4 Experimental Procedure

To investigate the potential of LLM-based speech-driven AAC systems in dysarthric speech recognition, we conducted simulation experiments using the Chinese Dysarthric Speech Dataset (CDSD)[27]. The evaluation comprised both subjective and objective assessments.

[3] We have created a demo: https://demo-resonance.hai-lab.cn/.

Subjective Evaluations Speech samples from 10 individuals with varying degrees of dysarthria were selected from the CDSD dataset to reflect real-world conditions. Twenty native Chinese speakers participated in the evaluation, assessing three dimensions: intelligibility, naturalness, and semantic relevance—the last of which measures the alignment between generated speech and the speaker's original intent. The evaluation covered three types of speech: original dysarthric speech, baseline ASR-TTS output, and LLM-enhanced ASR-TTS speech (Re-Sonance system). As shown in Table 1, participants reviewed a slide presentation with embedded audio samples and completed an online questionnaire. All evaluators underwent prior online training with annotated scoring examples to ensure rating consistency.

Table 1. Rating Scale for Subjective Evaluation

Score	Intelligibility Description	Naturalness Description
5	Fully intelligible, all information is clear	Highly natural, indistinguishable from human speech
4	Mostly intelligible, minor blurring	Generally natural, minor discontinuities
3	Partially intelligible, significant effort needed	Moderately natural, noticeable issues
2	Difficult to understand, most information unclear	Less natural, significant quality issues
1	Completely unintelligible	Highly unnatural, severely distorted

Objective evaluations The objective assessment focused on transcription accuracy and latency measurements of the Re-Sonance system, evaluating LLM optimization effectiveness using CDSD samples. For accuracy evaluation, speech samples from 30 speakers were categorized into mild, moderate, and severe dysarthria groups. Accuracy metrics included Word Error Rate (WER), Match Error Rate (MER), and Word Information Lost (WIL), analyzed across severity levels. Latency evaluation involved processing 200 approximately 10-second clips through Re-Sonance, measuring ASR latency, LLM optimization latency, and speech generation latency. Additionally, we computed the ratio of total latency to speech duration, a crucial metric for clip-based interaction in the Re-Sonance system.

4 Result

4.1 Subjective Evaluation

We recruited 20 native Mandarin speakers to evaluate the Re-Sonance system. Each participant rated 10 speech samples from the CDSD dataset on intelligibility, naturalness, and semantic association (Fig. 3).

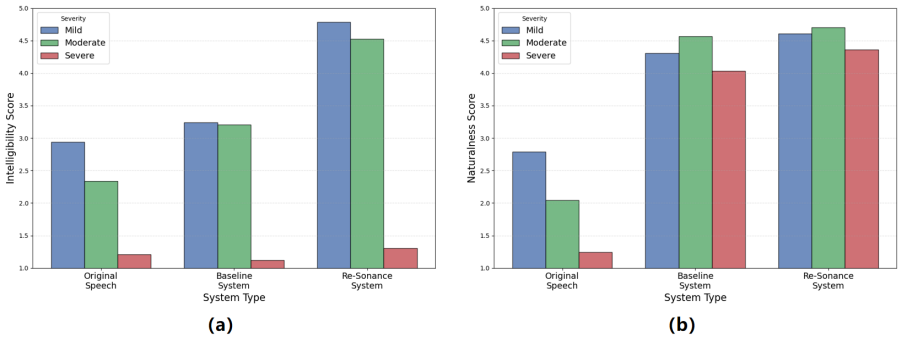

Fig. 3. Distribution of (a) Intelligibility and (b) Naturalness scores across system types

For mild articulation disorders (n=3), Re-Sonance showed significant improvements in intelligibility (M=4.79, SD=0.48) and naturalness (M=4.61, SD=0.95) versus baseline (M=3.24, SD=0.92; M=4.30, SD=0.63). The semantic association ratio increased from 78.8% to 98.3%. With moderate disorders (n=4), the system achieved higher intelligibility (M=4.55, SD=0.97) and naturalness (M=4.70, SD=0.59) compared to baseline (M=3.14, SD=1.02; M=4.57, SD=0.62), with semantic association improving from 68.2% to 88.6%. However, for severe disorders (n=3), improvements were limited. Re-Sonance showed minimal gains in intelligibility (M=1.24, SD=0.60) and naturalness (M=4.36, SD=0.48) versus baseline (M=1.09, SD=0.51; M=4.03, SD=0.67). Semantic association remained low, increasing only from 3.0% to 6.1%, due to inherent speech recognition challenges with severely impaired speech.

4.2 Objective evaluation

Transcription accuracy The Re-Sonance system yielded notable improvements in speech recognition metrics for mild (G1) and moderate (G2) dysarthria groups (Fig. 4). For G1, relative improvements were observed in WER (-7.84%), MER (-10.66%), and WIL (-18.00%). Similar enhancements were achieved in G2, with improvements in WER (-5.82%), MER (-10.18%), and WIL (-13.54%). However, for patients with severe dysarthria (G3), the system exhibited suboptimal performance, as evidenced by increases in WER (+0.63%) and MER (+3.46%). Only WIL shows slight optimization (-0.51%). Changes in speech recognition accuracy indicators after speech reconstruction for various levels of dysarthria are shown in Table 2.

Latency Performance analysis demonstrated that Re-Sonance achieves a Real-Time Factor of 0.8189 (Mdn = 0.7800, SD = 0.2396), confirming its efficiency and reliability in speech transcription and optimization tasks. The distribution of specific latency parameters is shown in Figure. 5.

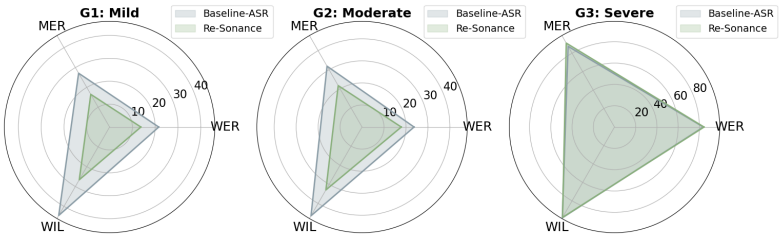

Fig. 4. Comparison of Baseline-ASR and Re-Sonance for Different Severity Levels

Table 2. Transcription Accuracy Results of the Re-Sonance System

Severity Level	Count	WER			MER			WIL		
		ASR	Re-Son.	Δ	ASR	Re-Son.	Δ	ASR	Re-Son.	Δ
G1: Mild	10	21.58	13.74	-7.84	27.14	16.48	-10.66	44.59	26.59	-18.00
G2: Moderate	10	23.70	17.88	-5.82	31.87	21.69	-10.18	46.47	32.93	-13.54
G3: Severe	10	83.77	84.40	+0.63	87.50	90.96	+3.46	98.40	97.89	-0.51

5 Discussion and Conclusion

5.1 Limitations

This study has several limitations: First, while CDSD provides diverse dysarthric speech samples, the system may not generalize to all patient profiles, particularly severe cases or those with atypical speech patterns. Our results showed reduced performance for users with more severe dysarthria. Second, evaluations were conducted solely in Mandarin Chinese. Although Re-Sonance integrates multilingual components (Whisper ASR, Qwen LLM, Cosyvoice), its cross-linguistic effectiveness remains unverified due to potential language-specific dysarthric variations. Third, despite real-time optimizations, latency from model inference and network transmission limits precise speech synchronization.

5.2 Potential of LLMs in Correcting Speech Interaction Errors

LLMs show strong potential in enhancing speech interaction for individuals with dysarthria. Even without ASR fine-tuning, integration of LLMs significantly improved accuracy and comprehensibility through contextual inference and automatic correction. LLMs not only correct ASR errors but also infer missing semantic content, especially in complex or unclear speech. Effectiveness depends on input quality, model configuration, and prompt design—structured prompts notably improved alignment with user intent. Objective evaluations confirm that well-optimized prompts enhance correction accuracy and reduce semantic drift, underscoring LLMs' value in future AAC systems.

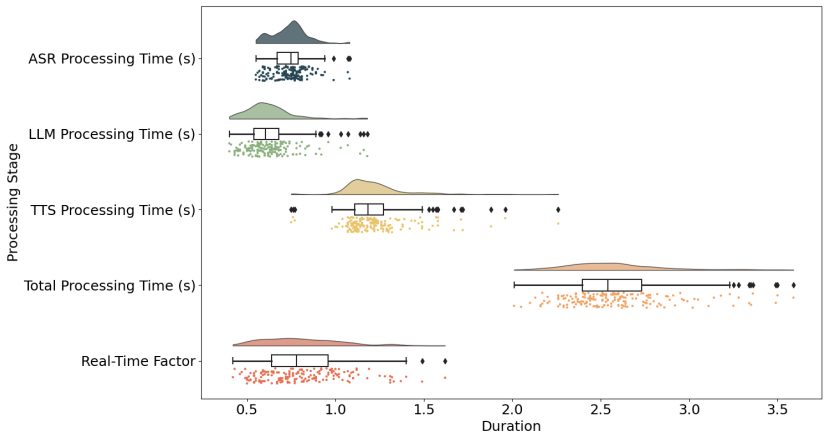

Fig. 5. Distribution of Processing Times Across Different Stages

5.3 Future Work and Research Agenda

This study lays the foundation for a more accessible AAC system and highlights key research directions: First, improving intelligibility for users with severe dysarthria remains essential. Future work will explore ASR fine-tuning and personalization to address this gap. Second, longitudinal user studies are needed to assess real-world usage patterns and social impacts, which may not be captured in lab settings. Third, cross-linguistic validation is necessary. While Re-Sonance supports multiple languages, systematic evaluation across linguistic and cultural contexts is required, including adaptations for language-specific dysarthria traits. These directions aim to advance universally accessible AAC technologies and promote communication equity for individuals with dysarthria worldwide.

References

1. Ayoka, G., Barbareschi, G., Cave, R., Holloway, C.: Enhancing communication equity: Evaluation of an automated speech recognition application in ghana. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–16 (2024)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Cai, S., Venugopalan, S., Tomanek, K., Kane, S., Morris, M.R., Cave, R., Macdonald, R., Campbell, J., Casey, B., Kornman, E., et al.: Speakfaster observer: long-term instrumentation of eye-gaze typing for measuring aac communication. In: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–8 (2023)
4. Chu, M., Yang, M., Xu, C., Ma, Y., Wang, J., Fan, Z., Tao, Z., Wu, D.: E-dgan: an encoder-decoder generative adversarial network based method for pathological to normal voice conversion. IEEE Journal of Biomedical and Health Informatics **27**(5), 2489–2500 (2023)

5. Dai, J., Moffatt, K., Lin, J., Truong, K.: Designing for relational maintenance: New directions for aac research. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–15 (2022)
6. Deshpande, R., Tuna, T., Subhlok, J., Barker, L.: A crowdsourcing caption editor for educational videos. In: 2014 ieee frontiers in education conference (fie) proceedings. pp. 1–8. IEEE (2014)
7. Du, Z., Chen, Q., Zhang, S., Hu, K., Lu, H., Yang, Y., Hu, H., Zheng, S., Gu, Y., Ma, Z., et al.: Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. arXiv preprint arXiv:2407.05407 (2024)
8. Han, J., Park, J., Huh, J., Oh, U., Do, J., Kim, D.: Ascleai: A llm-based clinical note management system for enhancing clinician productivity. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp. 1–7 (2024)
9. Hong, J., Findlater, L.: Identifying speech input errors through audio-only interaction. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2018)
10. Ibrahim, S.B., Vasalou, A., Clarke, M.: Design opportunities for aac and children with severe speech and physical impairments. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–13 (2018)
11. Jia, Z., Xue, H., Peng, X., Lu, Y.: Convert and speak: Zero-shot accent conversion with minimum supervision. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 4446–4454 (2024)
12. Kane, S.K., Morris, M.R., Paradiso, A., Campbell, J.: " at times avuncular and cantankerous, with the reflexes of a mongoose" understanding self-expression through augmentative and alternative communication devices. In: Proceedings of the 2017 acm conference on computer supported cooperative work and social computing. pp. 1166–1179 (2017)
13. Kim, W., Lee, S.: "i can't talk now": Speaking with voice output communication aid using text-to-speech synthesis during multiparty video conference. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–6 (2021)
14. Lin, S., Warner, J., Zamfirescu-Pereira, J., Lee, M.G., Jain, S., Cai, S., Lertvittayakumjorn, P., Huang, M.X., Zhai, S., Hartmann, B., et al.: Rambler: Supporting writing with speech via llm-assisted gist manipulation. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–19 (2024)
15. Luo, Z., Lin, S., Liu, R., Baba, J., Yoshikawa, Y., Ishiguro, H.: Decoupling speaker-independent emotions for voice conversion via source-filter networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing **31**, 11–24 (2022)
16. Mitchell, C., Cler, G., Fager, S., Contessa, P., Roy, S., De Luca, G., Kline, J., Vojtech, J.: Ability-based keyboards for augmentative and alternative communication: Understanding how individuals' movement patterns translate to more efficient keyboards: Methods to generate keyboards tailored to user-specific motor abilities. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts. pp. 1–7 (2022)
17. Mohan, A., Chakraborti, M., Eng, K., Kushaeva, N., Prpa, M., Lewis, J., Zhang, T., Geisler, V., Geisler, C.: A powerful and modern aac composition tool for impaired speakers. In: Proc. Interspeech 2024. pp. 991–992 (2024)
18. Munteanu, C., Penn, G.: Speech and hands-free interaction: Myths, challenges, and opportunities. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–4 (2018)

19. Qi, T., Zheng, W., Lu, C., Zong, Y., Lian, H.: Pavits: Exploring prosody-aware vits for end-to-end emotional voice conversion. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 12697–12701. IEEE (2024)
20. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International conference on machine learning. pp. 28492–28518. PMLR (2023)
21. Rekimoto, J.: Wesper: Zero-shot and realtime whisper to normal voice conversion for whisper-based speech interactions. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2023)
22. Rudzicz, F.: Articulatory knowledge in the recognition of dysarthric speech. IEEE Transactions on Audio, Speech, and Language Processing **19**(4), 947–960 (2010)
23. Tran, N., DeVries, P.S., Seita, M., Kushalnagar, R., Glasser, A., Vogler, C.: Assessment of sign language-based versus touch-based input for deaf users interacting with intelligent personal assistants. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–15 (2024)
24. Valencia, S., Cave, R., Kallarackal, K., Seaver, K., Terry, M., Kane, S.K.: "the less i type, the better": How ai language models can enhance or impede communication for aac users. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2023)
25. Valencia, S., Huynh, J., Jiang, E.Y., Wu, Y., Wan, T., Zheng, Z., Admoni, H., Bigham, J.P., Pavel, A.: Compa: Using conversation context to achieve common ground in aac. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–18 (2024)
26. Valencia, S., Pavel, A., Santa Maria, J., Yu, S., Bigham, J.P., Admoni, H.: Conversational agency in augmentative and alternative communication. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2020)
27. Wan, Y., Sun, M., Kang, X., Li, J., Guo, P., Gao, M., Wang, S.J.: Cdsd: Chinese dysarthria speech database. In: Proc. Interspeech 2024. pp. 4109–4113 (2024)
28. Wang, Y., Wu, X., Wang, D., Meng, L., Meng, H.: Unit-dsr: Dysarthric speech reconstruction system using speech unit normalization. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 12306–12310. IEEE (2024)
29. Wu, S., Li, J., Leshed, G.: Finding my voice over zoom: An autoethnography of videoconferencing experience for a person who stutters. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–16 (2024)
30. Zheng, W.Z., Han, J.Y., Cheng, H.L., Chu, W.C., Chen, K.C., Lai, Y.H.: Comparing the performance of classic voice-driven assistive systems for dysarthric speech. Biomedical Signal Processing and Control **81**, 104447 (2023)