PHOENIXDSR: PHONEME-GUIDED AND LLM-ENHANCED DYSARTHRIC SPEECH RECOGNITION

Yuxuan Wu[1], *Yifan Xu*[1], *Junkun Wang*[1], *Xin Zhao*[1], *Jiayong Jiang*[1], *Zhaojie Luo*[1,2]

[1] State Key Laboratory of Digital Medical Engineering, Southeast University
School of Biological Science & Medical Engineering, Southeast University
[2] Shenzhen Loop Area Institute

ABSTRACT

Automatic speech recognition (ASR) still struggles on dysarthric speech due to data scarcity and speaker heterogeneity. We present PhoenixDSR, a phoneme-mediated framework that decouples acoustic variability from linguistic decoding. A Wav2Vec2-CTC recognizer trained on healthy speech provides stable phoneme sequences. From limited dysarthric alignments we estimate a weighted confusion probability matrix that fuses global and speaker-specific patterns. A lightweight LLM decoder is trained on five tasks—bidirectional text–phoneme mapping, dysarthric-to-healthy normalization, phoneme-to-text decoding, and edit-operation prediction—to enable context-driven repair of systematic phoneme errors. On CDSD, PhoenixDSR attains 18.3% CER and 13.7% PER, outperforming end-to-end fine-tuning and LLM post-editing; ablations verify the importance of phonotactic pretraining and confusion priors. Few-shot personalization updates only the prior, yielding additional gains without further gradients. By combining interpretable phoneme-level priors with context-aware decoding, PhoenixDSR achieves data-efficient and robust recognition.
Code: github.com/wyxuan721/PHOENIXDSR

***Index Terms*—** dysarthria, automatic speech recognition, phoneme confusion, large language models.

1. INTRODUCTION

Dysarthria is a neuromotor speech disorder arising from congenital conditions or acquired injury. It is characterized by impaired articulation and reduced intelligibility, with both segmental and suprasegmental cues deviating substantially from typical speech [1, 2]. These deficits profoundly hinder everyday communication [3].

Despite rapid progress in automatic speech recognition (ASR), recognition accuracy for dysarthric speech lags far behind that for healthy speech. A well-established cause is data scarcity: prolonged speaking is difficult for many individuals with dysarthria, limiting corpus collection and thus training resources [4, 5, 6]. Insufficient data prevents deep models from learning robust pathological acoustic characteristics. Compounding the challenge, dysarthria exhibits marked heterogeneity across etiologies and severities, which makes learning invariances difficult and further constrains cross-speaker generalization [7, 8].

To improve generalization under scarce supervision, prior work has explored self-supervised pretraining and meta-learning to enhance domain adaptation and speaker transfer [9, 10]. Data augmentation and generative approaches—e.g., synthetic speech or personalized perturbations via VAE/GAN—have also been investigated [11, 12]. While promising, a substantial gap to healthy-speech ASR remains.

In contrast to the difficulties exhibited by AI systems, interlocutors who are close to individuals with dysarthria—such as family members or partners—can more readily understand what they intend to communicate [5]. This rapid adaptation relies on two mechanisms: (i) discovering speaker-specific phoneme confusions (e.g., consistently realizing /c/ as /ch/), and (ii) leveraging linguistic context to infer intended words [13]. If an ASR system could emulate these processes, capturing phoneme-level deviations and using context to repair errors, it could enable fast and interpretable personalization.

Phoneme-level modeling has been repeatedly identified as crucial for dysarthric speech. For instance, Lee et al. propose a dynamic phoneme-level contrastive learning (DyPCL) framework to improve phoneme representations [14]; Yeo et al. show that phoneme-level articulatory features aid automatic severity classification in Korean dysarthria [15]; and Almadhor et al. develop a spatiotemporal dysarthric ASR (DASR) system that learns phoneme shapes using SCNN and multi-head attention transformers [11]. These studies underscore the value of explicit phoneme modeling.

Large language models (LLMs) are natural candidates for context-based correction because they encode strong linguistic priors and have shown potential for repairing ASR outputs [16, 17, 18]. However, to our knowledge, no published work has targeted LLM-based correction specifically for dysarthric ASR. Existing LLM correction pipelines typically assume that most tokens are already correct and rely on sparse, localized edits, an assumption that breaks down when dysarthric ASR exhibits high error rates.

Our approach. Motivated by these insights, we propose a *phoneme-mediated* dysarthric ASR pipeline that (i) uses a phoneme recognizer trained on abundant healthy speech to map acoustics into a stable and interpretable phoneme sequence, thereby decoupling acoustic variability from linguistic decoding; (ii) estimates a weighted phoneme *confusion probability matrix* from limited dysarthric alignments by fusing global patterns with patient-specific deviations for few-shot personalization; and (iii) decodes phonemes to text with a lightweight LLM via a two-phase fine-tuning strategy: Phase I learns canonical phoneme↔text mappings from healthy data, and Phase II adapts to dysarthric realizations under the phoneme-confusion prior. This design concentrates scarce pathological supervision on confusion modeling while leveraging rich healthy data for phonotactics and decoding.

Our contributions are summarized as follows.

- We propose a *phoneme-mediated* dysarthric ASR framework that maps speech into a robust phonemic space learned from healthy data and leverages a two-phase multi-task LLM decoder to bridge phonemes and text with normalization and context-aware repair.

Corresponding author: luozhaojie@seu.edu.cn

- We introduce a lightweight and interpretable adaptation mechanism based on a weighted confusion probability matrix that integrates global confusions with patient-specific deviations and conditions the LLM for few-shot personalization.
- We conduct comprehensive experiments on the public CDSD corpus, demonstrating consistent improvements over strong baselines, strong data efficiency in few-shot personalization, and ablations that validate each component.

2. METHODOLOGY

As shown in Fig. 1, **PhoenixDSR** consists of a phoneme recognizer trained on healthy speech and a multi-task LLM decoder, with a phoneme confusion matrix serving as an intermediate prior for adaptation.

2.1. Phoneme Recognition Model

We first train a phoneme recognition model on large-scale healthy speech. This strategy reduces reliance on large amounts of pathological acoustic data: the recognizer, learned from abundant healthy speech, maps the signal—analogous to human perception—into a finite, interpretable, and stable phonemic representation.

Specifically, we adopt a Wav2Vec2-CTC architecture [19], where the objective is to maximize the conditional likelihood of phoneme sequences given the input audio:

$$\mathcal{L}_{\text{CTC}} = -\log P(\mathbf{y} \mid \mathbf{x}; \theta), \tag{1}$$

where $\mathbf{x}$ is the acoustic feature sequence, $\mathbf{y}$ is the target phoneme sequence, and θ are model parameters. After training, this model produces reliable phoneme sequences $\hat{\mathbf{p}}$ for healthy speech. However, when applied to *dysarthric speech*, it inevitably introduces systematic errors, which we later exploit to build a confusion prior.

2.2. Phoneme Confusion Matrix

Dysarthric speech often deviates systematically from canonical phonemes. To capture these patterns, we construct a **phoneme confusion matrix** that quantifies error regularities and supplies phoneme-level priors for decoding.

2.2.1. Global Phoneme Confusion Matrix

Given a dysarthric utterance, the recognizer outputs

$$\hat{\mathbf{p}}^{(d)} = (\hat{p}_1, \ldots, \hat{p}_m),$$

which we align to the ground-truth sequence

$$\mathbf{p}^{(gt)} = (p_1, \ldots, p_n)$$

using weighted dynamic programming with operations *match* (M), *substitution* (S), *deletion* (D), and *insertion* (I).

For each reference phoneme t, we collect observed outputs o (including $\langle\text{DEL}\rangle$ for deletions) and estimate

$$P(o \mid t) = \frac{\text{Count}(t \to o)}{\sum_{o'} \text{Count}(t \to o')}. \tag{2}$$

To improve robustness with limited counts, we apply hierarchical smoothing:

$$\tilde{P}(o \mid t) = \rho_t \, P_{\text{MLE}}(o \mid t) + (1 - \rho_t) \, P_{\text{backoff}}(o), \tag{3}$$

where P_{backoff} is a class-level (initial/final) or global distribution and

$$\rho_t = \frac{N_t}{N_t + \beta}$$

is a confidence weight based on sample size N_t and shrinkage β. We denote the resulting global phoneme confusion distribution as

$$C_g(o \mid t) \coloneqq \tilde{P}(o \mid t).$$

It encodes dysarthria-specific error patterns and guides LLM decoding.

2.2.2. Personalized Phoneme Confusion Matrix

The global matrix captures population-level tendencies, but personalization requires speaker-specific modeling. We treat the global distribution as a Bayesian prior and update it with individual statistics.

For speaker s, let $n_s(t \to o)$ be the count of substitutions of t into o, with

$$N_{s,t} = \sum_o n_s(t \to o).$$

We define a smoothed distribution

$$\tilde{C}_s(o \mid t) = \lambda_t \, P_s^{\text{MLE}}(o \mid t) + (1 - \lambda_t) \, C_g(o \mid t), \tag{4}$$

where

$$P_s^{\text{MLE}}(o \mid t) = \frac{n_s(t \to o)}{N_t}, \qquad \lambda_t = \frac{N_{s,t}}{N_{s,t} + \kappa}.$$

To balance global stability and personalization, we introduce a gating mechanism:

$$\widehat{C}_s(o \mid t) = (1 - \gamma_{s,t}) \, C_g(o \mid t) + \gamma_{s,t} \, \tilde{C}_s(o \mid t), \tag{5}$$

with

$$\gamma_{s,t} = \left(\frac{N_{s,t}}{N_{s,t} + \tau} \right)^{\alpha},$$

where τ and α control smoothness and sharpness. With even a few samples, $\widehat{C}_s$ emphasizes personalized statistics while reverting to C_g for under-sampled phonemes.

The final personalized matrix integrates population-level regularities with speaker-specific deviations, yielding a compact prior that enables rapid and robust adaptation in dysarthric ASR.

2.3. Multi-task Large Language Model

We fine-tune an LLM with multi-task objectives to transform dysarthric phonemes into fluent healthy text. We therefore adopt a two-phase design. For parameter efficiency, the base LLM parameters Θ are frozen; we train lightweight adapters.

2.3.1. Phase I: Healthy Speech Supervision

Since the LLM can only process symbolic semantics rather than acoustic signals, Phase I is designed to learn bidirectional mappings and phonotactics between *text* and *canonical phonemes* using healthy data. Let $\mathbf{t}^{(h)}$ denote healthy text and $\mathbf{p}^{(h)}$ its canonical phoneme sequence. We use two complementary seq2seq tasks:

$$\mathcal{T}_1 : \mathbf{t}^{(h)} \to \mathbf{p}^{(h)}, \qquad \mathcal{T}_2 : \mathbf{p}^{(h)} \to \mathbf{t}^{(h)}. \tag{6}$$

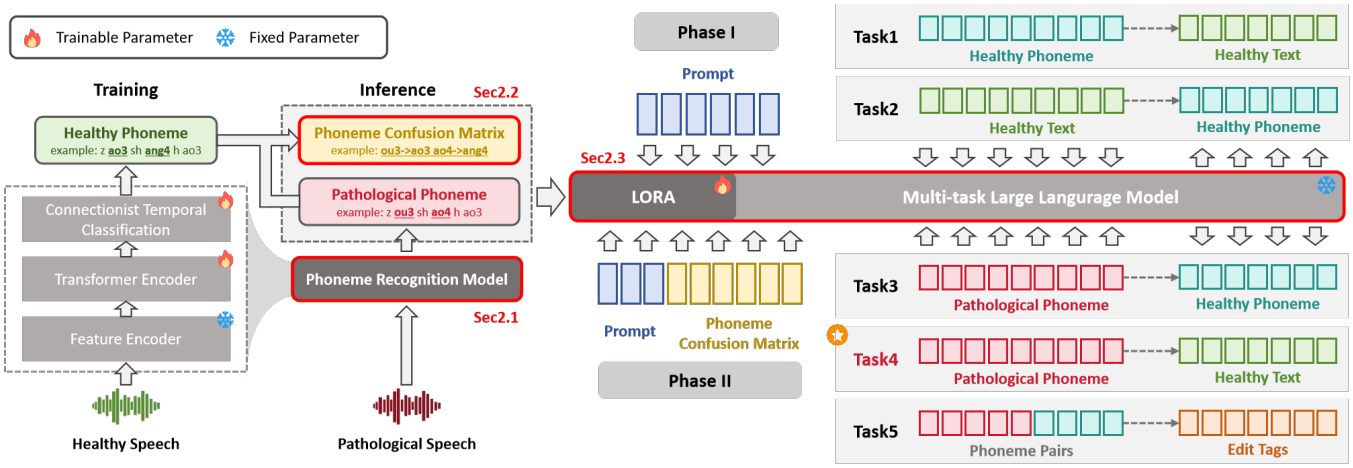

Fig. 1. PhoenixDSR framework.

Both are optimized with token-level cross-entropy under teacher forcing. Denoting the losses by $\mathcal{L}_{\mathcal{T}_1}$ and $\mathcal{L}_{\mathcal{T}_2}$, the Phase I objective is

$$\mathcal{L}^{(\mathrm{I})} = \lambda_1 \mathcal{L}_{\mathcal{T}_1} + \lambda_2 \mathcal{L}_{\mathcal{T}_2}. \tag{7}$$

The base LLM Θ is frozen; only lightweight adapters and task heads are trained, yielding $A^{(\mathrm{I})}$ for subsequent initialization.

2.3.2. *Phase II: Dysarthric Speech Adaptation*

Phase II adapts the model to dysarthric inputs by conditioning on the phoneme-confusion prior through three complementary tasks: dysarthric-to-healthy phoneme normalization and phoneme-to-text mappings to handle pronunciation variability, and edit-operation prediction to provide explicit correction signals. Given a dysarthric phoneme sequence $\mathbf{p}^{(d)}$ from the recognizer, we construct

$$U = \big[\texttt{OBS} = \mathbf{p}^{(d)};\ \texttt{PRIOR} = \mathcal{P}\big],$$

where $\mathcal{P}$ serializes top-k canonical candidates with probabilities from the phoneme confusion prior $P(\text{true} \mid \text{obs})$. Phase II comprises

$$\mathcal{T}_3 : \mathbf{p}^{(d)} \xrightarrow{U} \mathbf{p}^{(h)} \quad \text{(phoneme normalization)}, \tag{8}$$

$$\mathcal{T}_4 : \mathbf{p}^{(d)} \xrightarrow{U} \mathbf{t}^{(h)} \quad \text{(core decoding)}, \tag{9}$$

$$\mathcal{T}_5 : \big(\mathbf{p}^{(d)}, \mathbf{p}^{(h)}\big) \xrightarrow{U} \mathbf{e} \quad \text{(edit ops)}, \tag{10}$$

where $\mathbf{e} \in \{\texttt{M}, \texttt{S}, \texttt{D}, \texttt{I}\}$. Generation tasks use the standard negative log-likelihood

$$\mathcal{L}_{\text{gen}}(\mathcal{T}) = -\sum_t \log P_{\Theta, A^{(\mathrm{II})}}\big(y_t \mid y_{<t}, U\big), \tag{11}$$

and sequence labeling uses token-level cross-entropy. We initialize adapters with $A^{(\mathrm{II})} \leftarrow A^{(\mathrm{I})}$ and optimize

$$\mathcal{L}^{(\mathrm{II})} = \lambda_3 \mathcal{L}_{\text{gen}}(\mathcal{T}_3) + \lambda_4 \mathcal{L}_{\text{gen}}(\mathcal{T}_4) + \lambda_5 \mathcal{L}_{\mathcal{T}_5}, \tag{12}$$

with a larger weight on the core task $\mathcal{T}_4$. Training proceeds strictly *Phase I* $\rightarrow$ *Phase II* (no mixing), and early stopping is based on dysarthric validation.

Table 1. CDSD speaker split (speaker-independent, 8:1:1).

Split	#Speakers	Share	Notes
Train	36	≈80%	Global modeling
Dev	4	≈10%	Model selection
Test	4	≈10%	All reporting
Total	44	100%	–

3. EXPERIMENTS

3.1. Datasets

Healthy speech. We use AISHELL-1 (Mandarin) [20] as healthy supervision to learn text–phoneme bidirectional mappings. The corpus is split into train/dev/test with an 8:1:1 ratio.

Dysarthric speech. We experiment on CDSD [5] with a speaker-independent split at an 8:1:1 ratio.Table 1 reports the speaker counts per split; all experiments strictly use speakers unseen in training.

We adopt a tone-aware Mandarin phoneme inventory that models initials and finals separately, where each final is bound with its tone to form a distinct phoneme (e.g., *f*, *an1*, *an2*, ...). The phoneme recognizer is trained on AISHELL under this inventory and transferred to dysarthric speech, and all phoneme-based experiments use this same inventory.

3.2. Experimental setups

Phoneme recognizer. We initialize with chinese-wav2vec2-large [21]and train a CTC phoneme recognizer on AISHELL-1 (AdamW, lr=2×10^{-4}, 200k steps, warmup 10k, SpecAugment). This model is *frozen* in Phase II and used to produce dysarthric phoneme sequences $\hat{\mathbf{p}}^{(d)}$ and alignments.

Confusion priors. From alignments on CDSD we estimate a smoothed global phoneme confusion matrix and a personalized matrix per test speaker using the gated interpolation described in Section 2. At inference we serialize, for each observed phoneme, the top-k canonical candidates and log-probabilities as conditioning tokens.

LLM decoder. We use Qwen3-4B-Instruct-2507 with LoRA adapters (rank 16, α=32, dropout 0.05) applied to attention and MLP projection layers; base weights are frozen. *Phase I (pinyin)* trains on AISHELL-1 to learn explicit text↔canonical-phoneme mappings, producing initialization $A^{(\mathrm{I})}$. *Phase II (cli)* initializes from $A^{(\mathrm{I})}$ and

Table 2. Main results on CDSD (test set).

System	CER (%)	PER (%)
CDSD strong baseline	22.4	19.8
Whisper-FT	34.4	27.9
LLM-Post (Qwen3-4B)	30.0	27.1
PhoenixDSR (global-conf)	20.2	16.7
PhoenixDSR (personalized, K=100)	**18.3**	**13.7**

continues on CDSD with dysarthric objectives (phoneme normalization, phoneme→text, edit-operation prediction), conditioned on the serialized confusion prior. Adapters are optimized with AdamW (lr 1×10^{-4}); early stopping is based on dev CER.

Personalization. Few-shot adaptation is realized by *only* updating the personalized confusion prior and its gate, without any gradient update to the recognizer or LLM. We consider $K \in \{0, 50, 100, 200\}$ utterances per speaker (each 2–8 s), sampled uniformly from the speaker's personalization pool disjoint from test prompts.

Baselines and systems. (i) **CDSD strong baseline (reported).** Fbank front-end with WenetSpeech pre-training.
(ii) **Whisper-FT.** Whisper-Large-v3 fine-tuned end-to-end on CDSD.
(iii) **LLM-Post.** Whisper-FT hypotheses are post-edited by LoRA-tuned Qwen3-4B-Instruct-2507 (no phoneme mediation).
(iv) **PhoenixDSR (global-conf).** Our full pipeline with global confusion prior only.
(v) **PhoenixDSR (personalized-conf).** Our full pipeline with personalized prior (default K=100 unless otherwise noted).

3.3. Evaluation metrics

We use **CER** (character error rate) as the primary metric and **PER** (phoneme error rate) as a complement. Both follow the standard edit-distance formulation

$$\mathrm{ER} = \frac{S + D + I}{N} \times 100\%,$$

where S, D, I denote substitution, deletion, and insertion counts, and N is the number of reference units (characters for CER, phonemes for PER). Error types are further broken down in personalization analysis.

3.4. Main results

Table 2 summarizes development and test results on CDSD. The literature *CDSD strong baseline* is listed for horizontal reference; Whisper-FT and LLM-Post represent competitive end-to-end and text-only post-editing pipelines. Our **PhoenixDSR** surpasses both, and adding *personalized* confusion priors (with K=100) yields further gains.

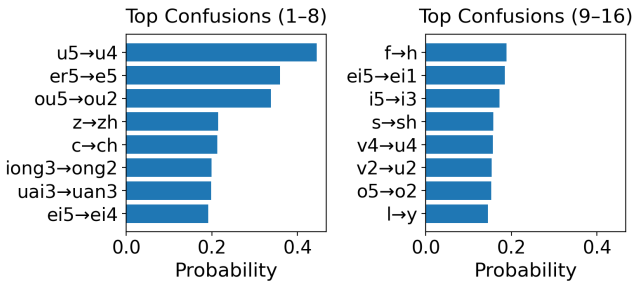

Fig. 2. Top confusion phonemes pairs.

Table 3. Ablations on CDSD (Test set).

Variant	CER (%)	PER (%)
PhoenixDSR (personalized, K=100)	**18.3**	**13.7**
w/o Phase I pretraining	25.9	30.6
w/o confusion prior	21.9	18.0

Table 4. Few-shot personalization on CDSD (Test set). Updating only the personalized confusion prior.

K (utt/spk)	CER (%)	PER (%)
0	20.2	16.7
50	18.9	14.6
100	**18.3**	**13.7**
200	18.3	13.6

Beyond overall averages, we further analyze error patterns. As visualized in Fig. 2, the dominant errors are highly structured rather than random: (i) tone substitutions within the same final (e.g., *u5→u4*, *er5→e5*, *ou5→ou2*); (ii) alveolar to retroflex sibilant confusions (*z→zh*, *c→ch*, *s→sh*); and (iii) vowel rounding or nasal shifts and medial reduction (*v→u*, *iong3→ong2*, *uai3→uan3*). PhoenixDSR alleviates these patterned substitutions by conditioning the decoder on phoneme-level confusion candidates, while standard LLM post-editing without phoneme mediation often leaves such systematic errors unresolved.

3.5. Ablation study

We ablate PhoenixDSR on CDSD (Table 3), using the personalized model (K=100) as reference. Dropping Phase I sharply worsens CER/PER—Phase I teaches the LLM explicit phoneme–text mappings that link pronunciations to text for reliable correction. Removing the confusion prior also degrades accuracy; a global-only prior recovers partially but still lags, underscoring the need for speaker-specific priors.

Few-shot personalization efficiency. We vary the number of per-speaker adaptation utterances $K \in \{0, 50, 100, 200\}$ (2–8 s each), updating only the personalized confusion prior and its gate. Table 4 shows monotonic CER/PER reductions with increasing K.

Overall, these studies confirm that (i) phoneme mediation with a learned confusion prior is critical under high error regimes; (ii) healthy speech pretraining (Phase I) provides transferable phonotactics; and (iii) a handful of per-speaker utterances suffice to capture idiosyncratic dysarthric patterns for robust decoding.

4. CONCLUSION

We introduced **PhoenixDSR**, a phoneme-mediated approach that maps speech into a robust phoneme space learned from healthy data and conditions a multi-task LLM on a fused global–personalized confusion prior for context-aware, interpretable correction. On CDSD, PhoenixDSR achieves **18.3%** CER and **13.7%** PER, surpassing end-to-end fine-tuning and text-only post-editing; ablations confirm the importance of phonotactic pretraining and the confusion prior. Few-shot personalization is realized by updating only the prior without gradient updates.

Future work will extend evaluation to other dysarthric corpora for cross-lingual robustness, broaden baselines, and report statistical significance. We also plan streaming and human-in-the-loop evaluations, finer-grained error analysis, and efficiency studies with smaller LLM backbones and lightweight prior estimation.

5. REFERENCES

[1] Frank Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 947–960, 2010.

[2] Shimon Sapir, "Multiple factors are involved in the dysarthria associated with parkinson's disease: a review with implications for clinical practice and research," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 4, pp. 1330–1343, 2014.

[3] Donald B Freed, *Motor speech disorders: Diagnosis and treatment*, plural publishing, 2023.

[4] 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, 2012.

[5] Mengyi Sun, Ming Gao, Xinchen Kang, Shiru Wang, Jun Du, Dengfeng Yao, and Su-Jing Wang, "Cdsd: Chinese dysarthria speech database," *arXiv preprint arXiv:2310.15930*, 2023.

[6] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon R Gunderson, Thomas S Huang, Kenneth L Watkin, Simone Frame, et al., "Dysarthric speech database for universal access research.," in *Interspeech*, 2008, vol. 2008, pp. 1741–1744.

[7] Siddharth Sehgal and Stuart Cunningham, "Model adaptation and adaptive training for the recognition of dysarthric speech," in *proceedings of SLPAT 2015: 6th workshop on speech and language processing for assistive technologies*. Association for Computational Linguistics, 2015, vol. 15, pp. 65–71.

[8] Seyed Reza Shahamiri, "Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 852–861, 2021.

[9] Shujie Hu, Xurong Xie, Zengrui Jin, Mengzhe Geng, Yi Wang, Mingyu Cui, Jiajun Deng, Xunying Liu, and Helen 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.

[10] Shiyao Wang, Shiwan Zhao, Jiaming Zhou, Aobo Kong, and Yong Qin, "Enhancing dysarthric speech recognition for unseen speakers via prototype-based adaptation," in *Proc. Interspeech 2024*, 2024, pp. 1305–1309.

[11] 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.

[12] Zengrui Jin, Mengzhe Geng, Jiajun Deng, Tianzi Wang, Shujie Hu, Guinan Li, and Xunying Liu, "Personalized adversarial data augmentation for dysarthric and elderly speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 413–429, 2023.

[13] Gifty Ayoka, Giulia Barbareschi, Richard Cave, and Catherine Holloway, "Enhancing communication equity: evaluation of an automated speech recognition application in ghana," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–16.

[14] Wonjun Lee, Solee Im, Heejin Do, Yunsu Kim, Jungseul Ok, and Gary Lee, "Dypcl: Dynamic phoneme-level contrastive learning for dysarthric speech recognition," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025, pp. 4701–4712.

[15] Eun Jung Yeo, Sunhee Kim, and Minhwa Chung, "Automatic severity classification of korean dysarthric speech using phoneme-level pronunciation features.," in *Interspeech*, 2021, pp. 4838–4842.

[16] Yuang Li, Xiaosong Qiao, Xiaofeng Zhao, Huan Zhao, Wei Tang, Min Zhang, and Hao Yang, "Large language model should understand pinyin for chinese asr error correction," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[17] Pranay Dighe, Yi Su, Shangshang Zheng, Yunshu Liu, Vineet Garg, Xiaochuan Niu, and Ahmed Tewfik, "Leveraging large language models for exploiting asr uncertainty," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12231–12235.

[18] Sheng Li, Chen Chen, Chin Yuen Kwok, Chenhui Chu, Eng Siong Chng, and Hisashi Kawai, "Investigating asr error correction with large language model and multilingual 1-best hypotheses," in *Proc. Interspeech*, 2024, vol. 2024, pp. 1315–1319.

[19] 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.

[20] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao 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)*. IEEE, 2017, pp. 1–5.

[21] Pengcheng Guo and Shixing Liu, "Chinese speech pretrain," 2022.