

Exploring Generative Error Correction for Dysarthric Speech Recognition

Original

Exploring Generative Error Correction for Dysarthric Speech Recognition / La Quatra, Moreno; Koudounas, Alkis; Mario Salerno, Valerio; Marco Siniscalchi, Sabato. - (2025), pp. 3284-3288. (Interspeech 2025 Rotterdam (NL) 17-21 August, 2025) [10.21437/Interspeech.2025-1553].

Availability:

This version is available at: 11583/3002219 since: 2025-07-29T14:27:05Z

Publisher:

ISCA

Published

DOI:10.21437/Interspeech.2025-1553

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)



Exploring Generative Error Correction for Dysarthric Speech Recognition

Moreno La Quatra*[♦], Alkis Koudounas*[♣], Valerio Mario Salerno[♦], Sabato Marco Siniscalchi[♣]

[♦]Kore University of Enna, Italy, [♣]Politecnico di Torino, Italy, [♣]Università di Palermo, Italy
moreno.laquatra@unikore.it, alkis.koudounas@polito.it, valerio.salerno@unikore.it,
sabatomarco.siniscalchi@unipa.it

Abstract

Despite the remarkable progress in end-to-end Automatic Speech Recognition (ASR) engines, accurately transcribing dysarthric speech remains a major challenge. In this work, we proposed a two-stage framework for the Speech Accessibility Project Challenge at INTERSPEECH 2025, which combines cutting-edge speech recognition models with LLM-based generative error correction (GER). We assess different configurations of model scales and training strategies, incorporating specific hypothesis selection to improve transcription accuracy. Experiments on the Speech Accessibility Project dataset demonstrate the strength of our approach on structured and spontaneous speech, while highlighting challenges in single-word recognition. Through comprehensive analysis, we provide insights into the complementary roles of acoustic and linguistic modeling in dysarthric speech recognition¹.

Index Terms: dysarthric speech recognition, speech accessibility, generative error correction

1. Introduction

Speech recognition technology has become essential in daily life, enabling natural interaction with devices and applications [1–4]. Voice-controlled systems now support tasks ranging from simple commands to complex conversations. However, these advances have not benefited all users equally [5–7]. Individuals with dysarthric speech face significant challenges when using conventional Automatic Speech Recognition (ASR) systems [8,9]. Dysarthric speech, resulting from motor speech disorders, exhibits irregular articulation, atypical prosody, and inconsistent speaking rates. These speech characteristics are often associated with various medical conditions, as speech patterns serve as important biomarkers for neurological disorders like Parkinson’s disease and voice disorders [10–13]. While modern ASR systems have achieved notable improvements through self-supervised learning and large-scale training data [14–17], their error rates on dysarthric speech often exceed 30% [18].

Modern ASR systems using beam-search-based decoding strategies are able to generate multiple possible transcriptions (hypotheses) ranked by confidence scores. While the top-ranked hypothesis might be incorrect for dysarthric speech, more correct alternatives may appear among lower-ranked candidates (see Section 4.3 and Table 4 for more details). This suggests that ASR models may capture the acoustic information but struggle with the proper ranking of alternatives. The Speech Accessibility Project (SAP) addresses these challenges by providing a dataset of 140+ hours of dysarthric speech [18]. Unlike

previous collections, the SAP dataset offers greater scale and detailed speech annotations through contextualized information and verbatim transcriptions. In this work, we investigate two fundamental questions about dysarthric speech recognition:

1. Do general-purpose ASR models possess sufficient acoustic modeling capacity to capture dysarthric speech patterns, even if they struggle with producing correct transcriptions?
2. Can large language models leverage their linguistic knowledge to identify correct transcriptions by analyzing patterns across multiple ASR hypotheses while ensuring grammatical and contextual coherence?

We propose a two-stage framework that combines ASR with generative error correction. Our approach first uses ASR models to generate multiple transcription hypotheses, capturing different interpretations of the acoustic input. A large language model then analyzes these hypotheses collectively, using its understanding of language patterns and context to produce more accurate transcriptions. While previous approaches to dysarthric speech recognition have focused on adapting acoustic models [19,20], we also explore error correction through text-to-text mapping using language modeling. We evaluate this system as part of the Speech Accessibility Challenge at INTERSPEECH 2025, focusing on developing effective ASR solutions for dysarthric speech. By combining acoustic and linguistic approaches, our work contributes to making speech technology more accessible for individuals with speech impairments.

2. Methodology

Our work explores (i) whether general ASR systems can capture useful information about dysarthric speech in their N-best hypotheses, and (ii) if language models can effectively distill this information. To investigate these research questions, we develop a two-stage framework that combines ASR with Generative Error Correction (GER) (Figure 1). Our framework separates the acoustic processing from linguistic error correction, allowing us to analyze the contribution of each component.

2.1. ASR Stage

The first stage of our framework uses Whisper [16] for speech recognition, implemented within the WhisperX framework [21] for efficient processing of long-form audio. We investigate two distinct approaches to handling dysarthric speech: zero-shot evaluation and fine-tuning. The zero-shot approach tests how well a general-purpose ASR model handles dysarthric speech without any adaptation. For adaptation, we fine-tune the entire model to specialize it for dysarthric speech recognition. For each audio input, we generate N transcription hypotheses through beam search. This N-best generation allows to capture

¹<https://github.com/MorenoLaQuatra/GER4Dys>

* Both authors contributed equally to this work.

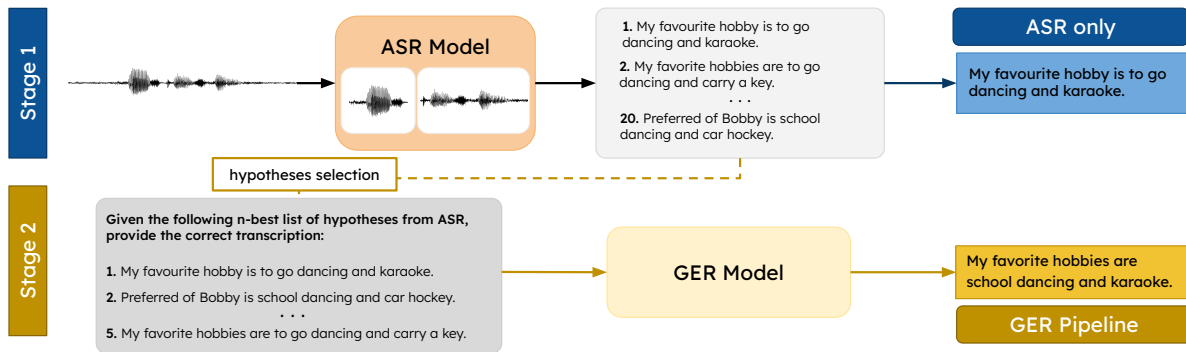


Figure 1: Overview of our two-stage framework for dysarthric speech recognition. Stage 1 uses the ASR model to generate 20-best hypotheses from the input audio. Stage 2 selects diverse hypotheses and employs the GER model to analyze them collectively, producing a refined final transcription.

different interpretations of unclear or challenging speech segments. For long recordings, WhisperX handles segmentation using integrated voice activity detection [22,23]. When processing these multi-segment recordings (i.e., utterances longer than 30 seconds), we maintain consistency by concatenating corresponding ranked hypotheses across segments (e.g., all top-1 hypotheses are concatenated to form the complete top-1 hypothesis).

Hypothesis Selection To ensure meaningful variation while maintaining computational efficiency, we implement a diversity-based selection algorithm that chooses 5 hypotheses from the initial pool of 20. It follows these steps:

1. Retain the top-scoring hypothesis to preserve the highest confidence transcription
2. Calculate normalized edit distances between all remaining hypotheses
3. Iteratively select hypotheses that maximize the minimum distance to previously selected ones

This approach is designed to select hypotheses that represent genuinely different interpretations rather than minor variations of the same transcription.

2.2. Generative Error Correction

The second stage of our framework uses a sequence-to-sequence model to analyze the selected hypotheses collectively and generate a refined transcription. We employ the FlanT5 model [24] which has demonstrated strong capabilities in error correction tasks [25]. Like other sequence-to-sequence models, FlanT5 can process variable-length input sequences and generate coherent text while maintaining semantic meaning. The model receives the hypotheses through a structured prompt as shown in Figure 2.

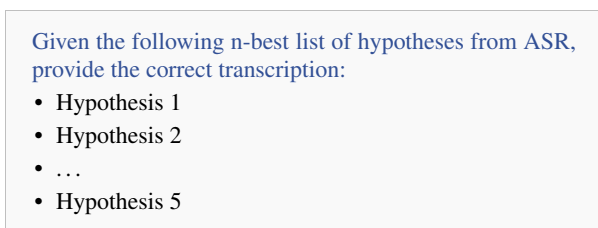


Figure 2: GER model prompt.

We train the GER model using LoRA [26] to maintain efficiency while specializing in transcription correction. This approach allows the model to leverage patterns across hypotheses and its own language understanding to resolve ambiguities in the ASR output. The GER stage serves multiple purposes: (i) analyzes patterns across multiple hypotheses to identify consistent elements, (ii) leverages linguistic knowledge to correct grammatical errors and resolve ambiguities, and (iii) produces a final transcription that balances acoustic evidence with linguistic plausibility.

Training Approach We train our framework in two distinct phases. First, we either use the ASR model in zero-shot mode or fine-tune it on our dataset. Second, we train the GER model using pairs of N-best hypotheses and ground truth transcriptions from the training set. This two-phase approach allows each component to specialize in its role while maintaining complementary capabilities.

3. Experimental Setup

Our experimental evaluation primarily uses the Speech Accessibility Project Challenge (SAPC) dataset [18], containing 105.76 hours of dysarthric speech in the training set and 39.56 hours in the development set. The dataset consists of multiple utterance types: Digital Assistant Commands (DAC), Sentences from Novels (SN), Spontaneous Speech (SS), and Non-spontaneous Speech which includes single-word utterances (SW).

To enhance model robustness and generalization capabilities, we augment the SAPC dataset with additional speech sources. We incorporate the TORGO dataset [27] (5.71 hours of natural dysarthric speech and 4.42 hours of synthetic speech) to increase exposure to diverse dysarthric speech patterns. Additionally, we include VoxPopuli [28] (500+ hours of native and non-native English speech) to ensure the model maintains strong performance on general English speech patterns while adapting to dysarthric speech characteristics.

3.1. Data Processing and Augmentation

To address the specific challenges of dysarthric speech recognition, we implement a data augmentation pipeline inspired by CrisperWhisper [29]. Our approach focuses on creating realistic variations in speech timing and background conditions, which are particularly relevant for dysarthric speech processing:

1. Following CrisperWhisper’s positional augmentation strategy, we inject background noise from MUSAN dataset [30] at either the beginning or end of each audio segment with 50%

Table 1: Performance comparison across ASR and GER configurations.

ASR Model	FT	GER	Dev		TEST-1		TEST-2	
			WER	SemScore	WER	SemScore	WER	SemScore
Lv3	✗	✗	11.60	83.91	11.39	84.82	14.49	78.83
Lv3	✗	3B	7.91	89.42	10.99	86.45	13.87	79.88
Lv3	✗	11B	7.34	90.32	10.63	86.96	13.64	80.4
Lv2	✓	✗	7.17	91.94	10.90	87.24	13.04	81.89
Lv2	✓	3B	6.40	92.47	10.67	87.51	12.89	82.16

Table 2: Ablation study on N-best list size.

ASR Model	N	Dev	
		WER	SemScore
Lv3	1	12.37	83.07
Lv3	5	11.71	83.79
Lv3	10	11.62	83.89
Lv3	20	11.60	83.91

probability, while maintaining the original transcription. This approach prevents overfitting to specific temporal positions in the encoder output while simulating real-world acoustic conditions.

- We introduce pure noise samples from MUSAN [30] during training with 1% probability and require empty predictions as output. This technique aims to mitigate hallucinations, a common issue in generative ASR methods like Whisper.
- With 25% probability, we apply either time stretching (0.85x-1.15x speed range) to address variable speaking rates in dysarthric speech, or SpecAugment [31] with frequency and time masking to enhance spectral robustness.

3.2. Transcription Format

The SAPC dataset provides detailed transcriptions that capture both the main speech content and disfluencies. For spontaneous speech utterances, transcriptions include the input question in square brackets, while prompted speech (i.e., digital assistant commands or read sentences) contains only the speaker’s response. We preserve all content in parentheses as they mark speech events with clear acoustic evidence in the recordings². For instance, partial word attempts or sound repetitions are transcribed as “(che- che-) checkout” or “(d-*) demo”. Additionally, round brackets mark contextual events through special tags: “(cs:...)” indicates interviewer speech, while “(ss:...)” denotes speaker utterances not related to the main prompt. In the target transcription for ASR we remove content within square brackets [. .] while retaining all content within round brackets (. .), as these segments have corresponding acoustic evidence in the speech signal.

3.3. Model Architecture and Training

Our framework implements a two-stage approach combining ASR and Generative Error Correction.

ASR Stage. We evaluate two Whisper configurations to analyze zero-shot capabilities and adaptation benefits. The first uses Whisper LARGE-V3³ without adaptation, while the second fine-tunes Whisper LARGE-V2⁴ on our combined dataset. We opted for Whisper LARGE-V2 fine-tuning based on preliminary experiments showing that LARGE-V3 exhibited significant instability when adapted to dysarthric speech. Specifically, fine-tuning LARGE-V3 led to severe degradation in transcription quality, characterized by repetitive pattern artifacts and incomplete utterances. This suggests that LARGE-V3’s pre-trained representations, while effective for general ASR, may either be

less suitable for domain adaptation to dysarthric speech patterns or require additional research to identify optimal fine-tuning strategies. For ASR fine-tuning, we use the AdamW optimizer [32] with a peak learning rate of 5×10^{-6} . The training runs for 5 epochs using a batch size of 64. We employ a linear learning rate schedule with 10% warmup steps followed by linear decay.

GER Stage. The GER stage uses two sizes of the FlanT5 model family: XL (3B parameters) and XXL (11B parameters). Given the large model sizes, we employ LoRA adaptation with rank $r = 16$, $\alpha = 32$, targeting all linear layers. This introduces approximately 1% additional trainable parameters while keeping the base model frozen. For GER training, we use AdamW with a peak learning rate of 1×10^{-4} and the same warmup-decay schedule as ASR fine-tuning. Training runs for 10 epochs with an effective batch size of 32 on 4 NVIDIA A100 80GB GPUs, the same hardware setup used for ASR fine-tuning.

3.4. Evaluation Protocol

The challenge provides three evaluation sets: development, TEST-1, and TEST-2. The development set enables model tuning and hyperparameter optimization, while TEST-1 provides ongoing feedback through a public leaderboard. TEST-2 remains hidden until the challenge closes, ensuring unbiased final evaluation. Performance is measured using two complementary metrics. Word Error Rate (WER) captures literal transcription accuracy through normalized string edit distance. The Semantic Score (SemScore) combines three normalized components with different weights: BERTScore [33] for semantic similarity, phonetic distance for pronunciation accuracy, and natural language inference for bidirectional entailment between hypothesis and reference. Each utterance is evaluated against two reference transcripts: one with disfluencies and one without. Following challenge guidelines, the system selects the reference with the lower normalized edit distance for scoring. This approach supports both verbatim and cleaned transcriptions, as each may be more appropriate depending on the application context.

4. Results and Analysis

We evaluate our framework through a series of experiments examining both overall performance and specific components. Our analysis focuses on understanding how acoustic and linguistic modeling capabilities contribute to dysarthric speech recognition.

4.1. Overall System Performance

Table 1 shows the performance of different system configurations across development and test sets. The baseline Whisper LARGE-V3 model achieves 11.60% WER on the development set without any adaptation, demonstrating strong zero-shot capabilities for dysarthric speech. Performance varies consid-

²Examples provided are for illustration purposes only and do not represent actual data from the SAPC dataset.

³<https://huggingface.co/openai/whisper-large-v3>

⁴<https://huggingface.co/openai/whisper-large-v2>

Table 3: Performance breakdown by utterance category on development set. Categories: DAC (Digital Assistant Commands), SN (Sentences from Novels), SS (Spontaneous Speech), SW (Single Words).

Category	#	w/o GER		w/ GER	
		WER ↓	SemScore ↑	WER ↓	SemScore ↑
DAC	15,066	6.47	92.77	5.61 _{-0.86}	93.28 _{+0.51}
SN	3,746	5.74	92.96	5.21 _{-0.53}	93.41 _{+0.45}
SS	2,313	10.86	87.27	10.31 _{-0.55}	88.14 _{+0.87}
SW	130	63.08	49.84	63.08 _{+0.00}	49.46 _{-0.38}

erably across evaluation sets: this baseline achieves 11.39% WER on TEST-1 but degrades to 14.49% WER on TEST-2, suggesting different levels of speech complexity in the test sets. Adding generative error correction with the 11B parameter model reduces development set WER to 7.34% and TEST-1 WER to 10.63%, showing that linguistic knowledge can effectively disambiguate ASR hypotheses. The improvement in semantic scores (from 83.91 to 90.32 on development set) suggests that GER not only corrects transcription errors but also enhances semantic understanding. Fine-tuning Whisper LARGE-V2 provides substantial gains, achieving 7.17% WER and 91.94 SemScore without GER in the development set, indicating effective acoustic model adaptation for dysarthric speech. Combining fine-tuning with GER achieves the best overall performance (6.40% WER, 92.47 SemScore on development; 12.89% WER on TEST-2), demonstrating complementary benefits from acoustic and linguistic components. Our experiments with model scaling revealed interesting trade-offs between performance and computational requirements. Scaling the GER model from 3B to 11B parameters shows diminishing returns on test sets, suggesting that computational resources might be better invested in other system components. Using a single NVIDIA A100 GPU, our fine-tuned Whisper LARGE-V2 achieves an average inference time of 0.55s per sample, while the complete pipeline including GER averages 0.69s per sample – remaining practical for real-world applications despite the high computational demands of the models. The consistent performance gap between development and test sets (approximately 6 percentage points in WER) highlights the challenge of building robust systems for diverse speaking patterns.

4.2. Analysis of N-best List Generation

Table 2 demonstrates the impact of N-best list size on recognition accuracy using Whisper LARGE-V3 on the development set. Increasing the number of hypotheses from 1 to 20 provides diminishing returns, with most gains achieved by the first 5 hypotheses (WER reduction from 12.37% to 11.71%). This pattern suggests that the ASR model captures relevant acoustic information within its top hypotheses, supporting our first research question about acoustic modeling capacity. The modest improvements beyond 5 hypotheses (only 0.11% WER reduction from 5 to 20 hypotheses) indicate that additional candidates often represent minor variations rather than fundamentally different interpretations. Our diversity-based selection algorithm helps maintain meaningful variation while keeping computational costs manageable.

4.3. Performance Across Utterance Types

Table 3 presents a detailed breakdown of system performance across different utterance categories in the development set. The results reveal distinct patterns in how GER affects different speaking styles. For digital assistant commands (DAC),

Table 4: A qualitative example from SS, showing diverse N-best hypotheses with correct transcription in green and errors in red.

Reference	My favorite pet is the one that sits on my lap.
ASR Output	My favorite play is the one that’s set on Monday .
GER Output	My favorite pet is the one that sits on my lap .
N-best Hypotheses:	
1.	My favorite play is the one that’s set on Monday .
2.	My favorite pet is the one that sits on my lap .
3.	My favorite player is the one that’s in Orlando .
4.	My favorite play is the ones that sit on the.
5.	My favorite pick is the one that said ” Wonder .”

which represent the majority of the dataset, GER reduces WER by 0.86% absolute while improving semantic scores by 0.51 points. Similar improvements appear in novel sentences (SN), with a 0.53% WER reduction and a 0.45-point increase in semantic accuracy, suggesting GER’s effectiveness on structured speech patterns. Spontaneous speech (SS) shows comparable WER reduction (0.55%) but achieves a larger semantic improvement (0.87 points), indicating that GER’s corrections are particularly meaningful for conversational speech.

Table 4 presents a case where our diversity-based hypothesis selection preserves the correct transcription. While standard beam search ranks hypotheses by confidence score alone, our approach explicitly selects diverse alternatives. This diversity is evident in the example, where despite errors in the top hypothesis (*play, set, Monday*), the correct transcription with *pet, sits, and lap* is preserved as the second hypothesis. By analyzing these diverse candidates, the GER model successfully identifies and selects the correct interpretation. However, single words (SW) transcriptions reveals a critical limitation, with extremely high WER (63.08%) and no improvement from GER. The analysis of SW errors shows that the system consistently misinterprets isolated words as short common phrases or expressions (e.g., “football” → “What law?”, “tape” → “Hey”) or longer phrases (e.g., “Birmingham” → “Lonnie Hill”), suggesting a strong bias towards generating complete utterances rather than single words. These findings highlight the impact of speaking style on dysarthric speech recognition, where our approach benefits both structured and conversational speech while isolated word recognition remains an open challenge.

5. Conclusions

Our investigation through the Speech Accessibility Project Challenge reveals both opportunities and challenges in dysarthric speech recognition. The strong performance of zero-shot Whisper (11.60% WER) answers our first research question, demonstrating that general-purpose ASR models can effectively capture dysarthric speech patterns. The consistent improvements from GER across speaking styles addresses our second question, validating language models’ ability to leverage multiple hypotheses, though with diminishing returns at larger scales. However, important limitations emerged: the poor performance on isolated words (63.08% WER) suggests the evaluated models may overly favor complete utterances, while the significant gap between development and test set performance indicates generalization challenges across different dysarthric speech patterns. These findings point to two promising research directions: investigating specialized architectures for isolated word recognition, and developing robust adaptation techniques to better handle the variability of dysarthric speech patterns.

6. Acknowledgements

This work has been partially supported by the "D.A.R.E. – Digital Lifelong Prevention" project (code: PNC0000002, CUP: B53C22006450001), co-funded by the Italian Complementary National Plan PNC-I.1 Research initiatives for innovative technologies and pathways in the health and welfare sector (D.D. 931 of 06/06/2022) and by the European Union – Next Generation EU under the National Recovery and Resilience Plan (PNRR) – M4 C2, Investment 1.1: Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN) - PRIN 2022 - "SHAPE-AD" (CUP: J53D23007240008).

7. References

- [1] S. wen Yang *et al.*, "Superb: Speech processing universal performance benchmark," in *Interspeech 2021*, 2021, pp. 1194–1198.
- [2] A. Koudounas, M. La Quatra, L. Vaiani, L. Colomba, G. Attanasio, E. Pastor, L. Cagliero, and E. Baralis, "Italic: An italian intent classification dataset," in *Interspeech 2023*, 2023, pp. 2153–2157.
- [3] B. Lee, I. Calapodescu, M. Gaido, M. Negri, and L. Besacier, "Speech-massive: A multilingual speech dataset for slt and beyond," *Interspeech 2024*, 2024.
- [4] M. La Quatra, A. Koudounas, L. Vaiani, E. Baralis, L. Cagliero, P. Garza, and S. M. Siniscalchi, "Benchmarking representations for speech, music, and acoustic events," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2024, pp. 505–509.
- [5] A. Koudounas, E. Pastor, G. Attanasio, V. Mazzia, M. Giollo, T. Gueudre, L. Cagliero, L. de Alfaro, E. Baralis, and D. Amberti, "Exploring subgroup performance in end-to-end speech models," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [6] A. Koudounas, E. Pastor, G. Attanasio, V. Mazzia, M. Giollo, T. Gueudre, E. Reale, L. Cagliero, S. Cumani, L. de Alfaro, E. Baralis, and D. Amberti, "Towards comprehensive subgroup performance analysis in speech models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1468–1480, 2024.
- [7] A. Koudounas, F. Giobergia, E. Pastor, and E. Baralis, "A contrastive learning approach to mitigate bias in speech models," in *Proc. INTERSPEECH 2024*, 2024, pp. 827–831.
- [8] W.-Z. Leung, M. Cross, A. Ragni, and S. Goetze, "Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis," in *Interspeech*, 2024.
- [9] D. Shah, V. Lal, Z. Zhong, Q. Wang, and S. R. Shahamiri, "Dysarthric speech recognition: A comparative study," in *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2023, pp. 89–94.
- [10] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from parkinson's disease." in *Lrec*, 2014, pp. 342–347.
- [11] A. Koudounas, G. Ciravegna, M. Fantini, E. Crosetti, G. Succo, T. Cerquitelli, and E. Baralis, "Voice disorder analysis: a transformer-based approach," *Interspeech 2024*, 2024.
- [12] M. L. Quatra, M. F. Turco, T. Svendsen, G. Salvi, J. R. Orozco-Arroyave, and S. M. Siniscalchi, "Exploiting foundation models and speech enhancement for parkinson's disease detection from speech in real-world operative conditions," *Interspeech 2024*, 2024.
- [13] M. La Quatra, J. R. Orozco-Arroyave, and M. S. Siniscalchi, "Bilingual dual-head deep model for parkinson's disease detection from speech," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [14] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, 2021.
- [16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [17] A. Koudounas, M. La Quatra, S. M. Siniscalchi, and E. Baralis, "voc2vec: A foundation model for non-verbal vocalization," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.
- [18] M. Hasegawa-Johnson *et al.*, "Community-supported shared infrastructure in support of speech accessibility," *Journal of Speech, Language, and Hearing Research*, 2024.
- [19] H. Wang, Z. Jin, M. Geng, S. Hu, G. Li, T. Wang, H. Xu, and X. Liu, "Enhancing pre-trained asr system fine-tuning for dysarthric speech recognition using adversarial data augmentation," in *ICASSP*, 2024.
- [20] I.-T. Hsieh and C.-H. Wu, "Dysarthric speech recognition using curriculum learning and articulatory feature embedding," in *Interspeech 2024*, 2024, pp. 1300–1304.
- [21] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," in *Interspeech 2023*, 2023, pp. 4489–4493.
- [22] G. Gelly and J.-L. Gauvain, "Optimization of rnn-based speech activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 646–656, 2018.
- [23] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote.audio: Neural building blocks for speaker diarization," in *ICASSP*, 2020.
- [24] H. W. Chung *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.
- [25] M. La Quatra, V. M. Salerno, Y. Tsao, and S. M. Siniscalchi, "Flanec: Exploring flan-t5 for post-asr error correction," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 608–615.
- [26] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.
- [27] 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.
- [28] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," 2021.
- [29] M. Zusag, L. Wagner, and B. Thallinger, "Crisperwhisper: Accurate timestamps on verbatim speech transcriptions," *Interspeech 2024*, 2024.
- [30] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.
- [31] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019.
- [32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.
- [33] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.