Are Speech Language Models More Noise Robust for Zero Shot Children's Speech Recognition

Arindam Das, Abhijit Sinha and Hemant Kumar Kathania
Department of Electronics and Communication Engineering, NIT Sikkim, India
b220046@nitsikkim.ac.in; phec230023@nitsikkim.ac.in; hemant.ece@nitsikkim.ac.in

***Abstract*—Children's speech recognition remains challenging due to variability in pronunciation and sensitivity to acoustic conditions. In this work, we analyze the performance of self-supervised learning (SSL) models and speech-language models (SLMs) for children's automatic speech recognition under noisy conditions. Experiments are conducted on the PFSTAR and CMU Kids datasets using a zero-shot setup with direct decoding. The results show that model performance in clean conditions is dataset-dependent, with different SSL and SLM models performing best across datasets. Under noisy conditions, SLMs tend to demonstrate improved robustness, particularly at lower signal-to-noise ratio (SNR) levels, while SSL models show greater sensitivity to noise. This advantage is especially pronounced on CMU Kids, where the SLM outperforms the SSL model across all noise types and SNR levels. Additionally, we observe that moderate noise can, in some cases, improve recognition performance for SLMs, suggesting that alignment between training and evaluation conditions plays an important role. Error analysis further reveals that SLMs maintain more stable and complete transcriptions under noise, with errors dominated by substitutions rather than deletions. Overall, the results highlight the benefit of incorporating linguistic context for robust children's speech recognition in noisy environments, particularly in challenging zero-shot scenarios.**

***Index Terms*—Children's speech, Self-supervised learning, Speech-language models, Noise, SNR.**

I. Introduction

Automatic speech recognition (ASR) has improved significantly in recent years, largely driven by self-supervised learning (SSL) models trained on large-scale unlabeled speech data. Models such as Wav2Vec2 [1], HuBERT [2], Data2Vec [3], and WavLM [4] learn representations directly from raw audio and have demonstrated strong performance across a wide range of speech tasks [5]–[9]. These models reduce the dependence on labeled data and generalize well across speakers, accents, and recording conditions.

Despite these advances, recognizing children's speech remains a challenging problem. One of the primary reasons is the mismatch between training and evaluation data. Most SSL models are trained predominantly on adult speech, whereas children's speech differs in several fundamental aspects, including higher pitch, greater variability in pronunciation, and less stable articulation [10]–[13]. In addition, children often exhibit inconsistent speaking patterns, which further increases the variability in the acoustic signal. These differences make it difficult for models trained on adult data to generalize effectively to children's speech [14]–[17].

The challenge becomes more pronounced in the presence of noise. In real-world scenarios, speech signals are rarely clean and are often affected by various types of noise, such as babble noise from multiple speakers, factory noise, and white noise [18], [19]. These distortions degrade the acoustic signal and reduce the reliability of the extracted representations. While SSL models perform well under clean conditions [20], their performance degrades significantly as the noise level increases [21]. This behavior highlights a key limitation of SSL-based approaches, which rely heavily on acoustic information.

Another important aspect is that SSL models primarily focus on learning acoustic representations from speech. Although deeper layers capture more abstract features, the representations remain closely tied to the input signal [22]–[25]. As a result, when the acoustic signal is corrupted, the model has limited ability to recover from such distortions, leading to increased recognition errors. This suggests that relying solely on acoustic information may not be sufficient for robust speech recognition in challenging conditions.

To address these limitations, recent work has explored speech-language models (SLMs) [26], [27], where speech representations are combined with large language models. By integrating acoustic and linguistic information, these models can utilize contextual cues during decoding, which can help resolve ambiguities when the acoustic signal is degraded. This is particularly relevant for children's speech, where variability in pronunciation can make purely acoustic modeling less reliable.

However, despite their potential, the effectiveness of SLMs for children's speech recognition under noisy conditions has not been systematically studied. It remains unclear whether incorporating linguistic context can consistently improve robustness across different datasets and noise conditions.

In this work, we investigate the noise robustness of SSL models and SLMs for children's speech recognition in a zero-shot setting. We evaluate multiple SSL models and compare them with speech-language models under different noise types and SNR levels. All models are evaluated without any fine-tuning on children's speech data, making the setup both realistic and challenging.

The main contributions of this work are as follows:

- We provide a comparative analysis of SSL models and SLMs for children's ASR under multiple noise conditions.
- We analyze the impact of different noise types and SNR levels on recognition performance.
- We observe that SLMs tend to exhibit greater noise robustness than SSL models in a zero-shot setup, par-

ticularly under severe noise conditions.
- We highlight the role of linguistic context in improving ASR performance when the acoustic signal is degraded.

II. Related Work

Children's speech recognition remains a challenging problem due to differences in pitch, pronunciation, and speaking patterns compared to adult speech. Models trained on adult speech data often exhibit degraded performance when applied to children's speech, and this mismatch becomes more pronounced under noisy conditions.

SSL models such as Wav2Vec2, HuBERT, Data2Vec and WavLM have significantly improved speech representation learning by leveraging large-scale unlabeled data [28]. These models learn rich acoustic representations that generalize well across tasks and domains [29]–[38]. However, their performance is still closely tied to the quality of the input signal, making them sensitive to noise and other distortions, particularly in mismatched conditions such as children's speech.

Noise robustness in ASR has traditionally been addressed through data augmentation, multi-condition training, and speech enhancement techniques. While these approaches improve robustness, they typically require additional training data, model adaptation, or preprocessing, which may not be feasible in zero-shot scenarios. As a result, the robustness of pretrained SSL models under noisy conditions, especially for children's speech, remains limited.

More recently, SLMs have been proposed to integrate acoustic representations with large language models. By incorporating linguistic context during decoding, these models can better handle ambiguities in the acoustic signal. This approach has shown promise in improving recognition performance, particularly in challenging conditions where acoustic information alone is insufficient.

Despite these developments, existing work has largely addressed children's speech, noise robustness, and self-supervised learning as separate problems. A systematic comparison between SSL models and SLMs for children's speech recognition under noisy conditions is still lacking. In particular, it is not well understood how these models behave under different noise types and SNR levels in a zero-shot setting.

This work aims to address this gap by providing a detailed analysis of noise robustness in SSL models and SLMs for children's speech recognition.

III. Experimental Setup

A. Datasets

We evaluate all models on two children's speech datasets.

The PFSTAR [39] dataset contains British English children's speech with an age range of 4 to 14 years. It includes read speech recordings from multiple speakers. We use the standard test set consisting of approximately 1.1 hours of speech from around 60 speakers.

The CMU Kids [40] dataset consists of American English children's speech, where children read prompted sentences. The dataset includes recordings from children aged 6 to 11 years. We use the standard test set containing approximately 2.8 hours of speech.

All audio is sampled at 16 kHz.

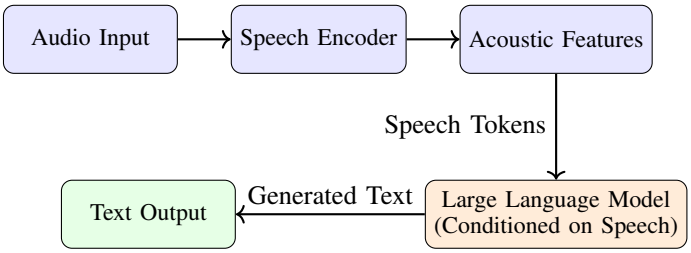

Fig. 1. Speech-language model (SLM) architecture. The acoustic pipeline (top row) extracts features from speech, which are provided to a large language model (bottom row). The language model generates context-aware transcriptions, improving robustness under degraded acoustic conditions.

B. Models

We evaluate multiple pretrained SSL models, including Wav2Vec2, HuBERT, Data2Vec and WavLM. These models are trained on large-scale speech corpora and are designed to learn general-purpose acoustic representations. We also evaluate two speech-language models: Granite 3.3 (8B) and Granite 4.0 (1B). These models integrate speech representations with a large language model, enabling the use of contextual information during decoding. As illustrated in Fig. 1, the input speech is first processed by a speech encoder to extract acoustic features, which are then provided to the language model. The language model generates the final transcription by leveraging both acoustic information and contextual knowledge. Granite 3.3 represents a larger model with higher capacity, while Granite 4.0 is a more compact model designed for efficient inference. This setup enables a direct comparison between purely acoustic SSL models and both large and lightweight SLMs under identical evaluation conditions.

IV. Methodology

In this work, we evaluate SSL models and SLMs for children's speech recognition in a zero-shot setting using direct decoding.

A. Decoding and Text Processing

All models are evaluated using a unified decoding pipeline. Audio waveforms are first resampled to 16 kHz when required. For SLMs, the input speech is combined with a text prompt using a chat-based template of the form: "$<|audio|>$ *Transcribe the speech. Do not convert numbers into digits; keep them in word form and add appropriate spacing.*"

The combined input is processed using the model-specific processor, and transcription is generated using greedy decoding without sampling, with a maximum of 200 generated tokens. In the case of SLMs, the acoustic features extracted from speech are provided to the language model, which generates the final transcription conditioned on both acoustic and contextual information, as illustrated in Fig. 1.

The generated transcriptions are post-processed before evaluation. All text is converted to uppercase and non-alphanumeric characters are removed. Additionally, numeric

digits in the output are converted into their corresponding word forms to ensure consistency with reference transcriptions.

B. Noise Generation

To evaluate robustness under noisy conditions, clean speech signals are corrupted with additive noise from the NOISEX-92 dataset. For a given clean speech signal x and noise signal n, the noisy signal is generated by scaling the noise to a target SNR level and adding it to the clean speech.

Noise is added at multiple SNR levels (0 dB, 5 dB, and 10 dB), representing increasingly challenging acoustic conditions. This process is applied independently to each utterance across different noise types, including babble, factory, and white noise, enabling a systematic evaluation of model performance under varying noise characteristics.

V. Results and Discussion

This section presents a comparative evaluation of SSL models and SLMs on children's speech under both clean and noisy conditions. We first analyze performance in clean conditions to identify the best-performing models, which are then used for detailed noise robustness analysis across different SNR levels and noise types.

TABLE I
WER (%) ON CLEAN CHILDREN'S SPEECH FOR SSL AND SLM MODELS.

Model	PFSTAR	CMU Kids
SSL Models		
Wav2Vec2	10.65	**22.63**
HuBERT	10.67	24.24
Data2Vec	**9.82**	22.85
WavLM	25.24	82.38
SLM Models		
Granite 3.3 (8B)	**9.85**	13.28
Granite 4.0 (1B)	15.20	**11.89**

A. Performance in Clean Conditions

Table I presents the WER results on clean children's speech for all models. Among the SSL models, Data2Vec achieves the best performance on PFSTAR with a WER of 9.82%, followed closely by Wav2Vec2 and HuBERT. In contrast, on CMU Kids, Wav2Vec2 achieves the lowest WER among SSL models, with Data2Vec showing comparable performance.

For the SLMs, Granite 3.3 achieves a WER of 9.85% on PFSTAR, which is comparable to the best-performing SSL model. On CMU Kids, Granite 4.0 achieves the best overall performance with a WER of 11.89%, followed by Granite 3.3.

These results indicate that model performance is strongly dataset-dependent. On PFSTAR, SSL and SLM models show comparable performance, suggesting that acoustic representations are sufficient for this dataset. In contrast, on CMU Kids, SLMs significantly outperform SSL models, indicating that incorporating linguistic context provides a clear advantage.

Based on these observations, Data2Vec and Granite 3.3 are selected for further analysis on PFSTAR, while Wav2Vec2 and Granite 4.0 are selected for CMU Kids.

B. Noise Robustness

Tables II and III present the WER results under different noise conditions and SNR levels for the selected models.

TABLE II
WER (%) UNDER DIFFERENT NOISE TYPES AND SNR LEVELS ON PFSTAR FOR BEST SELECTED MODELS.

Noise Type	Model	10 dB	5 dB	0 dB
Babble	Data2Vec	10.12	15.19	27.40
	Granite 3.3 (8B)	8.23	14.18	27.96
Factory	Data2Vec	10.93	16.97	38.64
	Granite 3.3 (8B)	8.67	13.68	22.57
White	Data2Vec	11.84	16.89	31.35
	Granite 3.3 (8B)	10.19	19.05	22.45

On PFSTAR, both models exhibit increasing WER as the SNR decreases, particularly at 0 dB. However, an interesting trend is observed for Granite 3.3 at moderate noise levels, where the WER can be comparable to or even lower than the clean condition. For example, under babble noise at 10 dB, Granite 3.3 achieves a lower WER than in the clean setting. In contrast, Data2Vec follows a more consistent degradation pattern as noise increases.

This behavior for Granite 3.3 can be attributed to the training strategy of modern speech-aware models. As reported in [26], such models are trained with extensive data augmentation, including noise at a wide range of SNR levels. As a result, moderate noise during evaluation may better align with the training conditions. Additionally, noise can reduce sensitivity to fine-grained acoustic variations in children's speech, leading to more stable decoding in certain cases.

Across noise types, Granite 3.3 demonstrates better robustness than Data2Vec under more challenging conditions. For instance, under factory noise at 0 dB, Granite 3.3 achieves a WER of 22.57% compared to 38.64% for Data2Vec, with similar trends observed for white noise. This indicates that the SLM maintains stronger performance when the acoustic signal is severely degraded.

TABLE III
WER (%) UNDER DIFFERENT NOISE TYPES AND SNR LEVELS ON CMU KIDS FOR BEST SELECTED MODELS.

Noise Type	Model	10 dB	5 dB	0 dB
Babble	Wav2Vec2	24.41	29.89	46.35
	Granite 4.0 (1B)	13.43	17.82	19.34
Factory	Wav2Vec2	25.32	31.16	50.74
	Granite 4.0 (1B)	13.46	16.06	19.34
White	Wav2Vec2	25.78	33.27	52.81
	Granite 4.0 (1B)	15.61	13.41	19.89

In contrast, on CMU Kids, Granite 4.0 consistently outperforms Wav2Vec2 across all noise types and SNR levels, with the performance gap becoming more pronounced as the SNR decreases. For example, under babble noise at 0 dB, Granite 4.0 achieves a WER of 19.34%, whereas Wav2Vec2

degrades significantly to 46.35%. Similar behavior is observed for factory and white noise, where Wav2Vec2 shows a sharp increase in WER, while Granite 4.0 maintains relatively stable performance.

Overall, the results highlight a clear distinction between SSL models and SLMs under noisy conditions. SSL models are more sensitive to noise due to their reliance on acoustic representations, whereas SLMs demonstrate improved robustness, particularly at lower SNR levels. The observation that moderate noise can occasionally improve performance further suggests that alignment between training and evaluation conditions plays an important role in children's speech recognition.

C. Error Analysis using Substitutions, Deletions, and Insertions

Tables IV and V present the distribution of substitution (S), deletion (D), and insertion (I) errors at 0 dB SNR across different noise types. We focus on 0 dB as it represents the most challenging acoustic condition, where the impact of noise on model behavior is most pronounced. Analyzing error patterns under this setting provides clearer insight into how different models handle severely degraded speech.

On PFSTAR, Granite 3.3 exhibits distinct error patterns depending on the noise type. Under babble noise, insertion errors are dominant, indicating a tendency to generate additional tokens in highly interfering conditions. In contrast, for factory and white noise, insertion errors reduce significantly, while substitutions become the primary source of error. This shift suggests that as the noise changes from highly interfering to more stationary, the model transitions from over-generation to incorrect word prediction. Deletion errors remain comparatively lower, indicating that the overall structure of the utterance is largely preserved.

On CMU Kids, Granite 4.0 shows a more consistent error profile across noise conditions. Substitution errors dominate in all cases, indicating that the model continues to produce complete transcriptions but with incorrect word choices. Deletion errors are moderate, suggesting that fewer words are dropped even under severe noise, while insertion errors, although present, remain lower than substitutions.

A comparison with SSL models further highlights these differences. On PFSTAR, Data2Vec exhibits a higher number of substitution and deletion errors, particularly under factory and white noise, indicating difficulty in maintaining both word accuracy and sequence continuity. In contrast, Granite 3.3 produces fewer deletions and more balanced error distributions. On CMU Kids, Wav2Vec2 shows a sharp increase in substitution and deletion errors across all noise types, whereas Granite 4.0 maintains a relatively controlled error profile.

TABLE IV
DISTRIBUTION OF SUBSTITUTION (S), DELETION (D), AND INSERTION (I) ERRORS AT 0 DB SNR ON PFSTAR.

Noise Type	Model	S	D	I	WER
Babble	Data2Vec	1088	109	169	27.40
	Granite 3.3	424	189	803	27.96
Factory	Data2Vec	1399	458	69	38.64
	Granite 3.3	558	273	312	22.57
White	Data2Vec	1302	169	98	31.35
	Granite 3.3	492	109	536	22.45

TABLE V
DISTRIBUTION OF SUBSTITUTION (S), DELETION (D), AND INSERTION (I) ERRORS AT 0 DB SNR ON CMU KIDS.

Noise Type	Model	S	D	I	WER
Babble	Wav2Vec2	3699	1114	651	46.35
	Granite 4.0	1148	516	791	19.34
Factory	Wav2Vec2	3997	1469	528	50.74
	Granite 4.0	1175	517	763	19.34
White	Wav2Vec2	4368	1356	559	52.81
	Granite 4.0	1214	443	868	19.89

Overall, the error distribution reveals a key distinction in model behavior. SSL models tend to suffer from higher deletion and substitution errors under noise, reflecting their sensitivity to degraded acoustic signals. In contrast, SLMs maintain more complete transcriptions, with errors dominated by substitutions. This behavior suggests that SLMs rely on contextual information to preserve transcription continuity even when the acoustic signal is unreliable.

D. Limitations

This study is limited to zero-shot evaluation on English datasets with synthetically generated noise. The comparison between SSL and SLM models is not fully controlled, as the two model classes differ in decoding strategy and training data. Future work will explore fine-tuning, multilingual settings, and real-world noise conditions

VI. CONCLUSION

In this work, we investigated the noise robustness of SSL models and SLMs for children's speech recognition in a zero-shot setting. The results show that model performance in clean conditions is dataset-dependent, with SSL and SLM models performing comparably on PFSTAR, while SLMs achieve significantly better performance on CMU Kids. Under noisy conditions, a clear distinction emerges between the two model types. SSL models exhibit significant degradation as the noise level increases, reflecting their reliance on acoustic representations. In contrast, SLMs demonstrate improved robustness, particularly at lower SNR levels, by leveraging contextual information during decoding. An interesting observation is that moderate noise can, in some cases, improve performance for SLMs. This suggests that alignment between training and evaluation conditions plays an important role, especially for models trained with diverse acoustic variations. Overall, the results suggest the potential benefit of integrating linguistic context for robust children's speech recognition, particularly in challenging acoustic environments.

REFERENCES

[1] 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.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, 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*, vol. 29, pp. 3451–3460, 2021.

[3] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[4] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[5] K. Radha, M. Bansal, and R. B. Pachori, "Speech and speaker recognition using raw waveform modeling for adult and children's speech: A comprehensive review," *Engineering Applications of Artificial Intelligence*, vol. 131, p. 107661, 2024.

[6] P. G. Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Computer Speech & Language*, vol. 72, p. 101289, 2022.

[7] R. Fan and A. Alwan, "Draft: A novel framework to reduce domain shifting in self-supervised learning and its application to children's asr," in *Interspeech*, 2022.

[8] T. Rolland, A. Abad, C. Cucchiarini, and H. Strik, "Multilingual transfer learning for children automatic speech recognition," in *International Conference on Language Resources and Evaluation*, 2022.

[9] J. Thienpondt and K. Demuynck, "Transfer learning for robust low-resource children's speech asr with transformers and source-filter warping," in *Interspeech*, 2022.

[10] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 603–616, 2003.

[11] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, no. 10-11, pp. 847–860, 2007.

[12] G. Yeung and A. Alwan, "On the difficulties of automatic speech recognition for kindergarten-aged children," *Interspeech*, 2018.

[13] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[14] L. L. Koenig, J. C. Lucero, and E. Perlman, "Speech production variability in fricatives of children and adults: Results of functional data analysis," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3158–3170, 2008.

[15] T. Tran, M. Tinkler, G. Yeung, A. Alwan, and M. Ostendorf, "Analysis of disfluency in children's speech," *Interspeech*, 2020.

[16] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Computer speech & language*, vol. 63, p. 101077, 2020.

[17] Ankita and S. Shahnawazuddin, "Developing children's asr system under low-resource conditions using end-to-end architecture," *Digital Signal Processing*, vol. 146, p. 104385, 2024.

[18] N. Djeffal, D. Addou, H. Kheddar, and S. A. Selouani, "A robust framework for noisy speech recognition using frequency-guided-swin transformer," *Computer Speech Language*, vol. 98, p. 101907, 2026.

[19] S. Dang, T. Matsumoto, Y. Takeuchi, and H. Kudo, "A comprehensive study on supervised single-channel noisy speech separation with multi-task learning," *Speech Communication*, vol. 167, p. 103162, 2025.

[20] A. Sinha, M. Singh, S. R. Kadiri, M. Kurimo, and H. K. Kathania, "Effect of speech modification on wav2vec2 models for children speech recognition," in *International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2024, pp. 1–5.

[21] A. A. Attia, D. Demszky, T. Ògúnremí, J. Liu, and C. Espy-Wilson, "Cpt-boosted wav2vec2.0: Towards noise robust speech recognition for classroom environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[22] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[23] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[24] Li, Jialu and Hasegawa-Johnson, Mark and McElwain, Nancy L., "Analysis of self-supervised speech models on children's speech and infant vocalizations," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 550–554.

[25] J. Li, M. Hasegawa-Johnson, and N. L. McElwain, "Towards Robust Family-Infant Audio Analysis Based on Unsupervised Pretraining of Wav2vec 2.0 on Large-Scale Unlabeled Family Audio," in *Interspeech 2023*, 2023, pp. 1035–1039.

[26] G. Saon, A. Dekel, A. Brooks, T. Nagano, A. Daniels, A. Satt, A. Mittal, B. Kingsbury, D. Haws, E. Morais, G. Kurata, H. Aronowitz, I. Ibrahim, J. Kuo, K. Soule, L. Lastras, M. Suzuki, R. Hoory, S. Thomas, S. Novitasari, T. Fukuda, V. Sunder, X. Cui, and Z. Kons, "Granite-speech: open-source speech-aware llms with strong english asr capabilities," 2025. [Online]. Available: https://arxiv.org/abs/2505.08699

[27] G. Saon, S. Thomas, T. Fukuda, T. Nagano, A. Dekel, and L. Lastras, "Self-speculative decoding for llm-based asr with ctc encoder drafts," 2026. [Online]. Available: https://arxiv.org/abs/2603.11243

[28] A. Sinha, H. K. Kathania, and M. Kurimo, "Beyond traditional speech modifications : Utilizing self supervised features for enhanced zero-shot children asr," in *INTERSPEECH*, 2025.

[29] 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.

[30] 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.

[31] A. Sinha, H. Kumar, M. Joshi, H. K. Kathania, S. Narayanan, and S. R. Kadiri, "Layer-wise analysis of self-supervised representations for age and gender classification in children's speech," in *Proc. WOCCI 2025*, 2025, pp. 46–50.

[32] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "Superb: Speech processing universal performance benchmark," in *Interspeech*, 2021, pp. 1194–1198.

[33] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Interspeech 2021*, 2021, pp. 3400–3404.

[34] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7967–7971.

[35] R. Fan, N. Balaji Shankar, and A. Alwan, "Benchmarking children's asr with supervised and self-supervised speech foundation models," in *Interspeech 2024*, 2024, pp. 5173–5177.

[36] T. Rolland and A. Abad, "Exploring adapters with conformers for children's automatic speech recognition," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 747–12 751.

[37] A. Sinha, H. K. Kathania, S. R. Kadiri, and S. Narayanan, "Can layer-wise ssl features improve zero-shot asr performance for children's speech?" *IEEE Signal Processing Letters*, vol. 32, pp. 3759–3763, 2025.

[38] S. Kutum, P. Sapkota, A. Sinha, H. K. Kathania, and M. C. Govil, "Empowering dysarthric communication: a self-supervised approach to keyword spotting," *Signal, Image and Video Processing*, vol. 19, no. 17, p. 1429, 2025.

[39] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, and M. Wong, "The PF_STAR children's speech corpus," in *Proc. INTERSPEECH*, 2005, pp. 2761–2764.

[40] M. Eskenazi, J. Mostow, and D. Graff, "The cmu kids corpus," *Linguistic Data Consortium*, vol. 11, 1997.