Bridging the Gap in Children's Speech Recognition: Zero-Speech Approaches with Speech Modifications and ASR architectures

Abhijit Sinha[1], Mittul Singh[2], Sudarsana Reddy Kadiri[3], Hemant Kumar Kathania[1], and Mikko Kurimo[4]
[1]*National Institute of Technology Sikkim, India*, [2]*AMD Silo AI, Helsinki, Finland*
[3]*University of Southern California, Los Angeles, USA*, [4]*Aalto University, Finland*

***Abstract*—Pretrained end-to-end (E2E) automatic speech recognition (ASR) models, such as Wav2Vec2, HuBERT and WavLM, have achieved near-human performance on adult speech in zero-resource settings. However, their performance in children's speech remains poor in zero-resource scenarios. To substantially improve performance in children ASR fine-tuning with little in-domain data is required, which might be untenable given the lack of labeled data. In this context, we wonder *how without using children's speech can we bridge the performance gap*? In this work, we address this challenge by (1) reviewing modifications applicable in zero-resource scenarios, (2) leveraging in-domain text resources for adaptation, and (3) comparing both E2E ASR architectures and hybrid HMM/DNN Kaldi-based systems. Our observations serve as important takeaways for building children ASR with minimal resources.**

***Index Terms*—end-to-end, ASR, HMM/DNN, children speech**

I. Introduction

The study of automatic speech recognition (ASR) for children's speech has led to great improvements in the areas of language learning for kids [1], [2], voice search [3], diagnosis and remedial therapy for pathological speech [4], and even toys and games [5]. The availability of pretrained models has further democratized the race to improved performance in such low-resource scenarios. Fine-tuning these pretrained models on children's speech data has been shown to significantly enhance ASR performance [6]–[9]. However, this approach often requires labeled datasets for optimal results, presenting a challenge given the limited availability of children's speech corpora.

On the other hand, adult speech is seeing improved performance even without fine-tuning, where zero-resource application of pretrained models has achieved near-human performance [10]. However, zero-resource ASR for children's speech has not seen similar effects, thus, building children ASR systems expensive. In this work, we explore bridging the performance gap between the zero-resource and fine-tuning based usage (low-resource) of pretrained models. Specifically, we review techniques that require zero-children speech but need in-domain text for decoding in children's speech recognition.

To this end, we make the following contributions: Firstly, we compare pretrained model performance on children's speech test set altered using speech modifications and their unaltered version. These comparisons tell us the relevance of speech modification as a zero-resource tool. Secondly, we use in-domain text to decode with an ASR model on children's speech, as obtaining in-domain text for scenarios like language assessment and voice search easier than labeled speech. These experiments help us understand the limits of using in-domain text. Thirdly, we compare these observations against fine-tuning with labeled children speech and hybrid HMM/DNN trained without children speech. Given speech modifications and decoding with language models, the former comparison gives the state of current gap for pretrained end-to-end models and the latter comparison presents these gains in context of different types of ASR models.

II. Related Work

This work distinguishes itself from prior work by investigating the use of pretrained models, speech modifications and ASR architectures for children ASR without using any children's speech. In this section, we discuss these dimensions in further detail.

A. Using Pretrained Models for Children ASR

Prior work [10], [11] has effectively used Whisper, a large pretrained model, to improve ASR performance on children's speech. These works have focused on fine-tuning with small amounts of children datasets effectively to show great improvements over zero-resource scenarios. These studies clearly highlight the substantial performance gap between zero-resource and low-resource approaches in children's ASR. On the other hand, zero-resource based improvements of Wav2Vec2 models for children ASR have also been studied using speech modifications [12]. In contrast to this work, we focus on bridging this gap by improving the children's ASR performance without using any children's speech, where we do not stay within the constrains of zero-resource and employ in-domain text for improving ASR.

B. Speech Modifications for Children ASR

Speech modifications [13]–[24] have been a popular set of methods for improving performance on children's speech. Intuitively, these methods focus on minimizing the mismatch between adult and children's speech to improve ASR performance. Recently, they have also shown promise in zero-resource scenarios [12], where application of speech modification techniques on children's speech test set can help improve

Wav2Vec2 performance on modified speech. Our work aims to study these improvements in combination with other factors like language modeling and ASR architectures like hybrid HMM/DNN and E2E models.

C. ASR Architectures for Children ASR

For children's ASR, end-to-end (E2E) models like RNN-T [25]–[27] have incorporated integrated language models (LMs) due to their architectural design. In contrast, decoding with external LMs is often omitted for other E2E models, such as Wav2Vec2 [8], [12], as part of the ASR pipeline. These design choices have also contributed to the declining focus on hybrid HMM/DNN models for zero-resource children's ASR scenarios.

To address the performance gap between zero- and low-resource children's ASR, this work investigates the zero-speech performance of hybrid HMM/DNN models, leveraging an in-domain language model for decoding.

III. Methods to Improve Children ASR

In this section, we explore various approaches aimed at improving ASR performance for children's speech. These approaches target the acoustic and linguistic mismatches between adult and children's speech, which are significant factors behind the suboptimal performance of ASR systems primarily trained on adult speech data. We classify these methods into three key strategies: speech modifications, decoding with children-specific text, and fine-tuning using transcribed children's speech data.

A. Speech Modifications

Children's speech exhibits notable differences from adult speech in terms of acoustic and linguistic features such as pitch, speaking rate, and formant positions, which can adversely affect ASR performance.

Pitch varies with age and gender [16], [17], [19]; for example, adult females usually have a pitch between 200-250 Hz, adult males between 100-150 Hz, whereas children generally have a higher pitch, averaging around 250-350 Hz. To adjust for these differences, we employ a time-domain pitch modification algorithm called Real-Time Iterative Spectrogram Inversion with Look-Ahead (RTISI-LA) [20]–[22]. RTISI-LA reconstructs a high-quality time-domain signal from the speech spectrogram, allowing for precise pitch alterations while preserving essential signal characteristics.

In addition to pitch, speaking rate also differs with adults typically speaking faster than children [16], [17], [19], [28]. We utilize the same RTISI-LA method for modifying speaking rate by adjusting a speed factor, thereby changing the duration of the speech signal per unit time.

Finally, formant locations vary between adult and child speech [16], [17], [29]–[31]. To compensate for these differences, we explore a linear prediction (LP) based method for formant modification. This approach involves warping the LP spectrum via an all-pass filter, using a warping factor α (where $-1 < \alpha < 1$). When α is positive, formant frequencies are shifted lower; when negative, they are shifted higher.

In our study, we concentrate on proven techniques to mitigate each of these variations by leveraging both conventional HMM/DNN, TDNN systems and readily available self-supervised learning (SSL) frameworks.

B. Decoding with Children's Text

Traditionally, language models (LMs) have relied on text data tailored to adult speech, which often fails to capture the linguistic patterns and vocabulary specific to children. By incorporating children's text, the system learn to predict and decode child-specific speech.

This adaptation is particularly effective for hybrid HMM/DNN models, which combine the temporal modeling capabilities of Hidden Markov Models (HMMs) with the feature extraction strengths of Deep Neural Networks (DNNs). Within this framework, children's text data is utilized in two critical ways:

- updating the dictionary to include child-specific lexical items and their phonetic representations, and
- training the LM on children's text data to better reflect their unique language usage patterns.

The dictionary update addresses challenges in recognizing child-specific pronunciations and vocabulary items that may be absent in standard adult-focused datasets. Meanwhile, the LM benefits from text tailored to children's linguistic habits, enhancing the accuracy of predicted word sequences. By integrating these updates, the hybrid HMM/DNN system becomes more robust in handling the phonetic variability and language patterns characteristic of children's speech.

C. Fine-tuning with Children's Transcribed Speech

Fine-tuning ASR systems with transcribed children's speech is one of the most effective methods to address the acoustic and linguistic mismatches between adult and children's speech. This process involves adapting a pre-trained ASR model using a smaller, manually annotated dataset of children's speech. Fine-tuning allows the acoustic model to learn child-specific phonetic and acoustic patterns while retaining the general knowledge gained from larger adult speech datasets.

Although the availability of transcribed children's data is often limited, fine-tuning can be performed with relatively small datasets, making it feasible in resource-constrained scenarios. This method has consistently shown significant improvements in reducing word error rates (WERs) for children's ASR systems across diverse datasets.

IV. Methodology

A. Datasets

Three speech corpora were used in this study: two British English datasets (WSJCAM0 and PFSTAR) and one American English dataset (LibriSpeech).

- **WSJCAM0** [32]: A British English adult speech corpus containing recordings from 140 speakers. The training subset consists of 15.5 hours of data from 92 speakers.

This dataset was used to train hybrid HMM/DNN and TDNN models for analyzing in-dialect (British adult → British children) generalization.

- **PFSTAR** [19], [31], [33]: A British English children's speech corpus with recordings from children aged 4-14 years. The training set includes 8.3 hours of speech from 122 speakers, while the test set comprises 1.1 hours of read speech from 60 speakers (32 male, 28 female). PFSTAR was used for SSL fine-tuning and evaluating all ASR systems.
- **LibriSpeech** [34]: A 960-hour American English adult speech corpus with recordings from 2,484 speakers. SSL models (Wav2Vec2, HuBERT, Data2Vec, WavLM) pretrained on LibriSpeech were used to study cross-dialect (American adult → British children) and cross-domain (adult → child) mismatches.

PFSTAR's test set served as the benchmark for all evaluations. To isolate domain/dialect effects:

- **In-dialect models**: Kaldi systems trained on WSJCAM0 (British adults) with PFSTAR LM adaptation.
- **Cross-dialect models**: SSL models pretrained on LibriSpeech (American adults) tested directly on PFSTAR.

B. Kaldi ASR Configuration

The Kaldi toolkit was used to train hybrid HMM/DNN and TDNN acoustic models. Speech signals were analyzed using overlapping Hamming-windowed frames (10 ms frame shift) to compute 13-dimensional MFCC features, augmented with delta and delta-delta coefficients. A 40-channel mel filterbank was employed for MFCC computation. Cepstral feature space maximum likelihood linear regression (fMLLR) was applied for normalization, with transformations derived using speaker adaptive training (SAT). The hybrid HMM/DNN system utilized fMLLR-normalized features with time-splicing. The DNN comprised eight hidden layers (1,024 nodes each), trained with a minibatch size of 256. The initial learning rate of 0.015 was reduced to 0.002 after 10 epochs, followed by 5 epochs of fine-tuning. For the TDNN acoustic model, training involved linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT), and SAT-based GMM alignments. Speaker adaptation was performed using i-vectors, with an initial learning rate of 0.0005 reduced to 0.00005 during training.

C. Language Model Integration

The decoding stage in Kaldi employs a two-pass strategy to improve accuracy. In the first pass, a trigram language model (LM) generates lattices of initial hypotheses. These lattices are then re-scored in the second pass using a more complex 4-gram LM trained on transcripts from the PFSTAR dataset (excluding the test set). This domain-specific LM incorporates child-specific lexical items (e.g., simplified vocabulary, repetitions) and phonetic variations, enabling the hybrid HMM/DNN system to better recognize children's speech patterns. The integration of in-domain text resources ensures contextual relevance during decoding, addressing linguistic mismatches between adult-trained models and children's speech.

D. Pretrained SSL Models

We evaluated three state-of-the-art self-supervised learning (SSL) models pretrained on large-scale speech corpora:

- **Wav2Vec2** [35]: Pretrained on 60k hours of unlabeled Libri-Light audio using contrastive learning, and fine-tuned on 960 hours of labeled LibriSpeech.
- **HuBERT** [36]: Trained on the same 60k-hour Libri-Light corpus using masked prediction with offline clustering to generate targets.
- **WavLM** [37]: Pretrained on 94k hours from Libri-Light, VoxPopuli, GigaSpeech, and MLS, with a noise-robust training objective. Fine-tuned on LibriSpeech.

All models share a common architecture comprising 25 hidden layers with a feature size of 1024. The initial layer extracts CNN-based features, while the remaining 24 layers employ transformers to model long-range dependencies and improve contextual understanding.

TABLE I
SPECIFICATIONS OF PRETRAINED SSL MODELS USED IN THE STUDY. THE TABLE INCLUDES MODEL SIZE (IN MILLIONS OF PARAMETERS), PRETRAINING DURATION (IN HOURS), AND FINE-TUNING DURATION (IN HOURS) FOR LARGE-SCALE SSL ARCHITECTURES: WAV2VEC2, HUBERT, AND WAVLM

Model	**Size**	**Pretraining (h)**	**Fine-tuning (h)**
Wav2Vec2-large-960h-lv60-self	317M	60,000	960
HuBERT-large-ls960-ft	316M	60,000	960
WavLM-large	343M	94,000	960

E. Fine-Tuning Setup

The SSL models were fine-tuned on the PFSTAR children's speech dataset (8.3 hours of training data). Training used a fixed learning rate of 1×10^{-4}, weight decay of 0.005, and 20 epochs with gradient checkpointing to prevent overfitting. The vocabulary included all characters from PFSTAR transcriptions, and Connectionist Temporal Classification (CTC) loss aligned speech-to-text sequences. Decoding used greedy search without external language models to isolate the impact of acoustic model adaptation.

V. RESULTS

This section evaluates the performance of hybrid (Kaldi) and SSL-based ASR systems on the PFSTAR children's speech test set. We analyze baseline models, speech modifications, combined adaptations, and fine-tuned SSL models.

A. Baseline Performance

Table II compares baseline WERs for hybrid and SSL models. The Kaldi TDNN system trained on adult British English speech (WSJCAM0) achieved 83.17% WER when decoded with the WSJCAM0 adult LM. Replacing the LM with PFSTAR's child-specific text reduced WER to 14.16%, demonstrating the necessity of in-domain linguistic adaptation. SSL models pretrained on LibriSpeech (American English

TABLE II
Baseline WER (%) on PFSTAR test set. The table compares traditional Kaldi-based systems (TDNN, DNN) with domain-specific language models (LMs) and modern self-supervised architectures (Wav2Vec2, HuBERT, WavLM) trained without LM integration

Model	**LM**	**WER (%)**
Kaldi TDNN	WSJCAM	83.17
Kaldi TDNN	PFSTAR	14.16
Kaldi DNN	PFSTAR	19.58
Wav2Vec2-base-100h	-	36.50
Wav2Vec2-base-960h	-	21.95
Wav2Vec2-large-960h	-	14.09
Wav2Vec2-large-960h-lv60-self	-	10.65
HuBERT-large-ls960-ft	-	10.67
WavLM-large	-	25.42

adults) showed cross-domain/cross-dialect gaps: Wav2Vec2-large-960h-lv60-self achieved 10.65% WER, while WavLM struggled (25.42% WER) struggled the most among the SSL models. We excluded WavLM-large from further experiments due to its comparatively poor performance among the evaluated SSL models.

B. Impact of Speech Modifications

Table III evaluates the effect of pitch modification (PM), speaking rate modification (SR), and formant modifications (FM). For hybrid systems, formant adjustments yielded the largest gains (14.16% → 12.37%), aligning children's higher formant frequencies (e.g., F1/F2 shifts) with adult-trained acoustic models. SSL models showed mixed results: smaller architectures like Wav2Vec2-base improved substantially with FM (36.50% → 32.71%), while larger models like Wav2Vec2-large saw marginal gains (10.65% → 10.41%).

C. Combined Modifications

Table IV shows cumulative improvements from combining PM, SR, and FM. Hybrid HMM/DNN systems achieved comprehensive gains, reducing WER to 8.87% (a 36.5% relative improvement over the TDNN baseline). In contrast, SSL models exhibited limited adaptability: for example, Wav2Vec2 plateaued at 10.16% WER (from 10.65%), underscoring its reliance on pretrained acoustic invariance. HuBERT demonstrated performance degradation, with WER increasing from 10.67% to 10.99%.

TABLE III
WER (%) for ASR models under baseline and modified speech conditions. The table compares traditional Kaldi-based systems (DNN, TDNN) and self-supervised architectures (Wav2Vec2, HuBERT) across pitch modification (Pitch), speech rate adjustment (SR), and formant modification (FM)

Model	**Baseline**	**PM**	**SR**	**FM**
Kaldi DNN	19.58	12.68	16.68	14.22
Kaldi TDNN	14.16	12.55	13.11	12.37
Wav2Vec2-base-100h	36.50	35.08	36.56	32.71
Wav2Vec2-base-960h	21.95	22.50	22.74	21.08
Wav2Vec2-large-960h	14.09	15.11	15.37	13.85
Wav2Vec2-large-960h-lv60-self	10.65	10.33	11.09	10.41
HuBERT-large-ls960-ft	10.67	10.43	10.49	10.22

TABLE IV
WER (%) for ASR systems under baseline conditions and after combined speech modifications *(PM + SR + FM)*. The table compares traditional (Kaldi TDNN) and self-supervised (Wav2Vec2-large-960h-lv60-self, HuBERT-large-ls960-ft) models.

System	**Baseline**	**Combined**
Kaldi TDNN	14.16	8.87
Wav2Vec2-large-960h-lv60-self	10.65	10.16
HuBERT-large-ls960-ft	10.67	10.99

D. Fine-Tuning SSL Models

Fine-tuning SSL models on 8.3h of PFSTAR data (Table V) achieved state-of-the-art results. Wav2Vec2-large attained 7.70% WER (27.6% improvement over zero-resource Wav2vec2 baseline), surpassing hybrid systems. HuBERT also showed simialr improvements (10.67% → 7.84%) highlighting the potential of these largely pre-trained SSL models when training data is limited or scarce.

TABLE V
WER (%) for fine-tuned SSL models on the PFSTAR dataset (8.3 hours). The table compares the performance of Wav2Vec2-large-960h-lv60-self and HuBERT-large-ls960-ft after fine-tuning.

Model	**WER (%)**
Wav2Vec2-large-960h-lv60-self	7.70
HuBERT-large-ls960-ft	7.84

E. Key Insights

- **Hybrid Systems**: Achieve **8.87% WER** (Table IV) without child speech by combining in-domain text, speech modifications, and HMM/DNN flexibility.
- **SSL Models**: Require fine-tuning but dominate with labeled data (**7.70% WER** for Wav2Vec2-large in Table V).
- **Speech Modifications**: Most impactful for hybrid systems (14.16% → 8.87%) and smaller SSLs (Wav2Vec2-base: 36.50% → 32.71%).
- **Architectural Trade-offs**: Hybrid systems excel with text/signal processing resources; SSL models require labeled child speech.

VI. Conclusion

Pretrained end-to-end models have transformed ASR for adult speech, yet their application to children's speech remains challenging in zero-resource scenarios. This work explores strategies to bridge the performance gap between zero-resource and fine-tuned systems by integrating speech modifications, in-domain text adaptation, and architectural innovations. We demonstrate that while speech modifications alone show limited effectiveness for self-supervised models, hybrid HMM/DNN systems prove more adaptable, combining domain-specific language modeling and signal adjustments to achieve competitive performance without child speech data. Conversely, self-supervised models excel when minimal labeled children's speech is available, underscoring their dependency on targeted fine-tuning. These findings highlight the importance of architectural and resource-aware design: hybrid systems offer a pragmatic path for low-resource settings, while self-supervised models prioritize efficiency when labeled data is accessible. By addressing acoustic, linguistic, and structural mismatches, this work advances equitable ASR solutions for children's speech, encouraging future research into adaptive frameworks.

References

[1] P. Vogt, M. de Haas, C. de Jong, P. Baxter, and E. Krahmer, "Child-robot interactions for second language tutoring to preschool children," *Frontiers in Human Neuroscience*, 2017.

[2] R. Al-Ghezi, K. Vosboinik, Y. Getman, A. von Zansen, H. Kallio, A. Clara, M. Kuronen, A. Huhta, and R. Hilden, "Automatic speaking assessment of spontaneus l2 finnish and swedish," *Language Assessment Quarterly*, 2022.

[3] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Your word is my command: Google search by voice: A case study," in *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, 2010.

[4] K. J. Ballard, N. M. Etter, S. Shen, P. Monroe, and C. T. Tan, "Feasibility of automatic speech recognition for providing feedback during tablet-based treatment for apraxia of speech plus aphasia," *American Journal of Speech-Language Pathology*, 2019.

[5] K. Matthes, R. Petrick, and H. Hain, "Lingunia world of learning," in *ISCA International Workshop on Speech and Language Technology in Education, SLaTE 2015, Leipzig, Germany, September 4-5, 2015*, 2015.

[6] 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.

[7] R. Fan, Y. Zhu, J. Wang, and A. Alwan, "Towards better domain adaptation for self-supervised models: A case study of child asr," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1242–1252, 2022.

[8] R. Jain, A. Barcovschi, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition," *IEEE Access*, vol. 11, pp. 46 938–46 948, 2023.

[9] J. Li, M. A. Hasegawa-Johnson, and N. L. McElwain, "Analysis of self-supervised speech models on children's speech and infant vocalizations," *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSP)*, pp. 550–554, 2024.

[10] R. Jain, A. Barcovschi, M. Yiwere, P. Corcoran, and H. Cucu, "Adaptation of whisper models to child speech recognition," in *INTERSPEECH*, 2023.

[11] R. Jain, A. Barcovschi, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition," *IEEE Access*, vol. 11, pp. 46 938–46 948, 2023.

[12] 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.

[13] J. Driedger, M. Müller, and S. Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Processing Letters*, 2014.

[14] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, 1999.

[15] H. K. Kathania, "Role of Prosodic Features and Prosody modification in Improving Children Mismatched ASR," Ph.D. dissertation, Department of ECE, National Institute of Technology Sikkim, India, October 2018.

[16] S. Lee, A. Potamianos, and S. Narayanan, "Analysis of children's speech: duration, pitch and formants," in *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 1997.

[17] S. Lee, A. Potamianos, and S. 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, March 1999.

[18] S. Ghai and R. Sinha, "Exploring the effect of differences in the acoustic correlates of adults' and children's speech in the context of automatic speech recognition," *EURASIP Journal on Audio, Speech and Music Processing*, 2010.

[19] S. Shahnawazuddin, N. Adiga, H. K. Kathania, and B. T. Sai, "Creating speaker independent ASR system through prosody modification based data augmentation," *Pattern Recognition Letters*, 2020.

[20] X. Zhu, G. T. Beauregard, and L. L. Wyse, "Real-time signal estimation from modified short-time fourier transform magnitude spectra," *IEEE Transactions on Audio, Speech, and Language Processing*, July 2007.

[21] H. K. Kathania, W. Ahmad, S. Shahnawazuddin, and A. B. Samaddar, "Explicit pitch mapping for improved children's speech recognition," *Circuits, Systems, and Signal Processing*, 2018.

[22] H. Kathania, M. Singh, T. Grósz, and M. Kurimo, "Data augmentation using prosody and false starts to recognize non-native children's speech," in *Proc. Interspeech 2020*, 2020.

[23] J. Laroche and M. Dolson, "New phase-vocoder techniques for real-time pitch shifting," 04 2000.

[24] R. Bristow-Johnson, "A detailed analysis of a time-domain formant-corrected pitch-shifting algorithm," *Journal of the Audio Engineering Society. Audio Engineering Society*, 1995.

[25] 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.

[26] T. B. Patel and O. Scharenborg, "Improving end-to-end models for children's speech recognition," *Applied Sciences*, 2024.

[27] 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.

[28] T. Nagano, T. Fukuda, M. Suzuki, and G. Kurata, "Data augmentation based on vowel stretch for improving children's speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

[29] J. E. Huber, E. T. Stathopoulos, G. M. Curione, T. A. Ash, and K. Johnson, "Formants of children, women, and men: The effects of vocal intensity variation," *The Journal of the Acoustical Society of America*, 1999.

[30] H. K. Kathania, S. R. Kadiri, P. Alku, and M. Kurimo, "Study of formant modification for children ASR," in *ICASSP*, 2020.

[31] H. K. Kathania, S. R. Kadiri, P. Alku, and M. Kurimo, "A formant modification method for improved asr of children's speech," *Speech Communication*, 2022.

[32] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *Proc. ICASSP*, vol. 1, May 1995, pp. 81–84.

[33] 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.

[34] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[35] 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.

[36] 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.

[37] 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.