Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Why is children's ASR so difficult? Analyzing children's phonological error patterns using SSL-based phoneme recognizers

Koharu Horii[1], *Naohiro Tawara*[1], *Atsunori Ogawa*[1], *Shoko Araki*[1]

[1]NTT Corporation, Japan

koharu.horii@ntt.com

Abstract

Children's automatic speech recognition (child ASR) is generally more challenging than adult ASR, as children's speech differs from that of adults and continuously evolves with a child's development. To investigate the causes of this difficulty and leverage the findings to improve the child ASR performance in the future, this study examines phoneme recognition error patterns using self-supervised learning (SSL)-based phoneme recognizers (PRs). First, we apply several SSL-based child PRs to English children's speech from kindergarten to grade 10, confirming that the WavLM-based model performs best. Then, using this model, we examine whether misrecognitions are primarily due to child mispronunciations or model errors. Furthermore, analyzing the results obtained by the WavLM-based model, we investigate how error patterns change along with age. The results show that mispronunciations persist longer than previously reported, particularly in certain manners of articulation.

Index Terms: children's automatic speech recognition, self-supervised learning, phoneme recognition, phonological process, error analysis

1. Introduction

Accurate children's automatic speech recognition (child ASR) is essential for systems that support childcare and education at home and school more efficiently, such as social robots [1] and interactive educational applications [2]. However, child ASR is known to be more challenging than adult ASR. Even with a highly accurate ASR model like Whisper [3], child ASR accuracy remains significantly lower than that for adult ASR [4]. The authors in [5] analyzed the relationship between children's development and their changes in ASR accuracy, highlighting the difficulty of using ASR models for younger children.

One main reason for this difficulty in child ASR is the greater significant variability in children's speech with that in adults' speech [6]. The authors of [7] categorized the causes of this variability into two factors. The first is due to *acoustic* characteristics in children's speech, including a higher fundamental frequency [8] and a slower speaking rate [9] compared with adults. These characteristics vary significantly with the children's development? and individual differences, resulting in substantial within-speaker and between-speaker variations [6].

The second factor is due to *linguistic* (or *phonological*) characteristics specific to children's speech, including mispronunciations, disfluencies, and hesitations, particularly among younger children. Notably, studies on children's phonological acquisition report the existence of child-specific mispronunciation patterns, known as phonological processes [10, 11, 12]. These are believed to stem from underdeveloped articulation and speech-motor control abilities at various developmental stages [13], which gradually improve with development, typically converging toward adult speech by around nine years of age [10]. However, this development varies significantly among children, resulting in substantial between-speaker variations.

To improve child ASR accuracy in light of these child-specific speech variabilities, various strategies have been explored, including data augmentation techniques that convert adult speech into child speech [14, 15] and fine-tuning adult ASR models using limited child speech data [16, 17, 18]. Despite these approaches, efforts, a significant performance gap between adult and child ASR remains [4], raising a question: why does child ASR remain so difficult?

To answer this question, this paper presents an in-depth analysis of the errors produced by current ASR systems. In particular, it is essential to assess the extent to which acoustic and linguistic factors have been addressed and to determine which factor remains dominant even with current ASR systems. to this end, we conduct automatic phoneme recognition experiments on children's reading speech and categorize misrecognized phonemes into the following two cases:

A : Phonemes that the child pronounced correctly but were misrecognized by the model due to child-specific acoustic variations (i.e., errors related to *acoustic* variations).

B : Phonemes that the child could *not* pronounce correctly, leading to misrecognition likely due to incorrect pronunciation (i.e., errors related to *linguistic* variations).

Our results indicate that our phoneme recognizer (PR) significantly reduces the phoneme recognition error rate compared with those reported in prior studies [7, 19]. However, acoustic-related errors (Type A errors) are still predominant across all ages groups, suggesting that the recognizer still has room for improvement in handling acoustic variations. On the other hand, linguistic-related errors (Type B errors) are more common in younger children's speech but decrease rapidly with age, suggesting that these errors follow phonological development processes. These findings indicate that the challenges in recognizing younger children's speech stem from linguistic difficulties in pronouncing phonemes. Therefore, we provide a deeper analysis focusing on linguistic-related errors to better understand the challenges in child ASR. To analyze the relationship between age- and linguistic-related error patterns, we investigate the age at which these errors diminish and compare our findings with previous developmental research. Finally, we identify the most frequent substitution error patterns observed in our experiments and discuss their implications in the context of existing existing child development research.

10.21437/Interspeech.2025-2645

Table 1: *Details of the two speech corpora used in this study.*

Corpus	Subset	# Speakers	# Hours	# Utterences	Grade (age)
LibriSpeech [21]	Train	5,466	961.1	281,241	Adult
	Dev	188	10.5	5,567	
OGI Kids [20]	Train	777	45.0	46,890	0 (five)
	Dev	106	6.4	6,449	$\sim$ 10 (15)
	Test	233	13.7	13,970	

Table 2: *PER comparison of the different model architectures.*

Model architecture	SSL method	Denoising	PER
Wav2Vec2.0 [22]	Contrastive		18.50
HuBERT [23]	Predictive		18.25
WavLM [24]	Predictive	✓	**17.84**

2. Related works and our contribution

Focusing on Type B errors, the authors of [7, 19] investigated phoneme error patterns in a child-specific automatic phoneme recognition model based on Gaussian mixture model-hidden Markov models (GMM-HMM).

However, their research has the following three limitations, as noted in their conclusion in [19]. First, they adopted GMM-HMM-based PRs which are limited in performance. As a result, a large number of phoneme recognition errors, mainly caused by acoustic variability, could be present. Second, they analyzed 2,200 utterances from 60 speakers (i.e., twelve speakers per age group), which may be insufficient to account for individual differences among children. Third, it was unclear how these errors evolve as children grow since the age range studied was limited to five to nine years old.

To overcome these limitations, we designed new experiments covering a broader age range and including hundreds of speakers. Specifically, we introduce an end-to-end (E2E) PR built on a self-supervised learning (SSL)-based feature extractor to investigate phoneme error patterns, achieving performance comparable with state-of-the-art models. In addition, we utilize the OGI Kids speech corpus [20], a children's speech corpus that includes a wide range of age groups, from kindergarten to grade 10[1], and a relatively large number of speakers (1,100). Using this corpus enables a more general analysis of error patterns as children grow up. By conducting a detailed analysis by grade and phoneme, we can reveal clearer patterns.

In summary, this research makes two key contributions. First, we analyze error patterns in child ASR using the latest SSL-based PRs, categorizing errors into acoustic- and linguistic-related types. Second, we conduct a comprehensive error analysis on a large dataset spanning a wide age range, from kindergarten to grade 10, revealing the relationship between age and error patterns.

3. Speech corpora and models

3.1. Speech corpora

Table 1 shows an overview of the two speech corpora used in this study [21, 20]. Both the LibriSpeech and OGI Kids corpora contain read speech of given texts (prompts). We used LibriSpeech to train adult PRs and OGI Kids to train child-specific PRs. For LibriSpeech, we used the official training and development data splits. In contrast, OGI Kids does not provide official data splits, so we divided the corpus into training, development, and test sets as described in the following.

OGI Kids contains 67,309 somewhat noisy utterances produced by 1,166 children in grades 0 (five years old, i.e., kindergarteners) to 10 (15 years old, i.e., junior high school students). There are 317 different prompts consisting of isolated words, sentences, or digit strings. Each speaker uttered approximately 60–70 given prompts. We divided the corpus to minimize overlap between the prompts included in the training and test sets. By using OGI Kids, we can conduct a more detailed analysis as described in Section 2.

[1]Grade refers to school years in the United States education system, typically correspondings to ages 5 to 16, with September enrollment.

3.2. SSL-based child phoneme recognizer

We constructed child PRs using three SSL models: Wav2Vec2.0 [22], HuBERT [23], and WavLM [24]. These models have been used as the base models for phoneme recognition in both adult and child speech [25, 26, 27] and children's word-level ASR [17], showing high accuracy. All the models are based on Transformers [28] and are pre-trained on LibriSpeech, but have several differences. Wav2Vec2.0 uses contrastive training criteria, while HuBERT and WavLM use discriminative training criteria. In addition, only WavLM incoporates denoising during its training. All models have approximately the same number of parameters (94M).

Training SSL-based child PRs using only OGI Kids due to its size limitation. Inspired by [4], we used the following two-step procedure by exploiting LibriSpeech, i.e., a large-scale adult speech corpus.

Step 1 : We first constructed an adult PR. We added a connectionist temporal classification (CTC) head [29] to an SSL model and trained it using LibriSpeech with phoneme-level transcriptions.

Step 2 : We then constructed a child PR by adapting the aforementioned adult PR using OGI Kids with phoneme-level transcriptions.

We conducted our experiments using Hugging Face Transformers [30], leveraging the pre-trained models wav2vec2-base [31], hubert-base-ls960 [32], and wavlm-base [33] as our three SSL-based models. Using g2p_en [34], we converted the word-level prompts into phoneme-level transcriptions (on the basis of a 39-phoneme system) to adapt the SSL models in accordance with the above two-step procedure. The training hyperparameters largely followed those in the tutorial [35], with several modifications to prevent overfitting. Specifically we used smaller learning rates : 10^{-5} for Step 1 and 10^{-6} for Step 2. During training, we saved models regular intervals. After training, we use mergekit [36] to combine the ten models that achieved the lowest development loss values, resulting in a robust final model.

4. Results and analysis

We conducted five experiments to investigate the challenges of child ASR.

4.1. Model comparison

We compare the three SSL-based models described in subsection 3.2 to determine the most suitable architecture for children's phoneme recognition tasks. Since each model uses a different training strategy, identifying the best-performing model

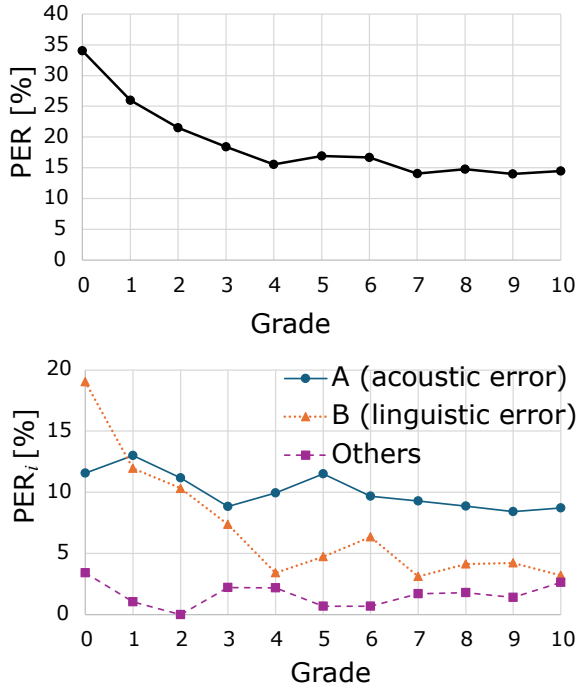


Figure 1: *(Top) PER for each grade. (Bottom) Ratio of A: acoustic-related, B: linguistic-related, and Others: neither errors to the overall errors for each grade.*

provides insight into which characteristics are most effective for recognizing children's phoneme.

Table 2 shows the phoneme error rate (PER) obtained by each model. While the PERs do not differ significantly, the WavLM [24]-based model achieves the lowest PER. This is likely because the speech in OGI Kids is noisy, so the denoising task in WavLM proves effective. In the following sections, we utilize the WavLM-based model to analyze the recognition results.

4.2. PER changes with children's development

We analyze how children's mispronunciations change with age by comparing PERs across grades. Figure 1 (top) plots the PER for each grade. Result shows that PER decreases as children grow. In grade 0, the PER exceeds 30%, but it rapidly declines with age and stabilizes around 15% starting from grade 4, indicating that younger children's speech is more challenging to recognize. The best PER remains around 15% due to the datasets' relatively high noise level.

4.3. Error type analysis in each grade

Figure 1 (top) includes both errors caused by acoustic and linguistic factors, as described in Section 1. To identify which error type is dominant at each grade level, an annotator conducted a listening test. Specifically, we first randomly selected 50 misrecognized utterances from each grade. The annotator then listened to the audio and judged whether the sound was different from the prompted phoneme (A), the same as the prompted phoneme B), or neither (Others). We then calculated the ratio of the number of samples n_i categorized into each error type $i \in \{\mathrm{A}, \mathrm{B}, \mathrm{Other}\}$ to the total number of misrecognized samples n, yielding PER_i for each error type i as,

$$\mathrm{PER}_i = \mathrm{PER} \cdot \frac{n_i}{n}. \quad (1)$$

Figure 1 (bottom) shows the ratio of each error type to the overall errors (PER_i) by grade. This result highlights two important findings.

1. Acoustic-related errors (type A) still dominate across most grades, even when using the best-performing SSL-based PRs. This suggests that further improvements are needed to address child-specific acoustic variations.
2. The dominance between acoustic- and linguistic-related errors changes with grade. Specifically, linguistic-related errors (type B) occur frequently in younger children (grades 0 to 3) but rapidly decrease as the children grow. In contrast, acoustic-related errors (type A) remain relatively stable across all grades. This suggests that linguistic-related errors are more prominent in younger children.

The high PER in younger child ASR is particularly problematic. Our findings provide evidence that this issue primarily stems from their linguistic difficulty in correctly pronouncing phonemes. In the following sections, we focus on linguistic-related errors (type B) and provide further analysis to identify when these difficulties are diminish as children grow.

Table 3: *Grade at which the phoneme error rate drops below 20% for each phoneme. Vowels are underlined.*

Grade	0	1	2	3	4	5	7	9	10, Adult
Phone	a͡ʊ, ð	a͡ɪ, d, e͡ɪ, h, i, n, ɹ, s, w, z	ʌ, b, f, k, p, v	æ, t͡ʃ, l, m, t	ɝ, θ	ʃ	ŋ, u	ɪ, ɔ͡ɪ, ʊ, j, ʒ	a, ɔ, ɛ, d͡ʒ, o͡ʊ, g,

4.4. Grade in which the PER of each phoneme decreases

We calculated the PER for each phoneme and each grade, investigating the grade at which The PER for each phoneme decreases below 20% and remains below this threshold. This enables us to determine the grade at which each phoneme can be correctly pronounced In addition, we focused on consonants' place and manner of articulation to examine the specific factors that make them difficult for child ASR.

Table 3 shows the grades at which the PER for each phoneme decreases. Many phonemes become correctly recognized as children grow. The variation in error patterns by grade suggests the importance of age-specific approaches for child ASR. The phonemes underlined are vowels. Previous studies [10, 7] that vowels are typically acquired at ages 2–3 gave them little attention. However, our results show vowels (excluding /a͡ʊ/) have a PER exceeding 20% at grade 0 (age 5). While the PER for half the vowels decreases during elementary school (Grades 1–6), that for the other half did not decrease even as the children grew up.

Focusing on consonants, almost all show a high PER at grade 0, and this decreasing significantly in grades 1–3. To determine the factors behind these results, we investigated the relationship between consonant phonemes articulated by place and manner and the grades at which PER decreases, as shown in Table 4. Each column represents a place of articulation, and each row represents a manner of articulation. Bilabial sounds are articulated with both lips, labiodental sounds with the upper

Table 4: *Relationship between the age at which the PER decreases to below 20% and the place/manner of articulation. Phonemes whose PER decreases to below 20% at grades 0–3 are in red; at grades 4–7 in blue; and after 9 in black.*

Manner \ Place	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Palatal	Velar	Uvular	Glottal
Plosive	p b			t d			k ɡ		
Affricate				tʃ dʒ					
Fricative		f v	θ ð	s z	ʃ ʒ				h
Nasal	m			n			ŋ		
Approximant				ɹ		j		w	
Lateral approximant				l					

Table 5: *Top 10 most frequent substitution errors made by children in grade 0.*

Substitution error rate [%]	Substitution pattern	Example	Phonological process [7]	Grade where the error rate halves
12.50	dʒ → d	**j**ive→**d**ive	Stopping	not halved
10.53	ʒ → ʃ	vi**s**ion→vi**sh**on	-	not halved
10.53	ʒ → z	vi**s**ion→vi**z**on	Fronting	not halved
8.03	ŋ → n	si**ng**→si**n**	Fronting	9
7.41	θ → t	**th**ank→**t**ank	Stopping	1
6.25	dʒ → ʒ	**j**ive→**ʒ**ive	Deaffrication	10
6.19	ɡ → k	**g**uy→**k**i	-	not halved
6.19	w → m	**w**ild→**m**ild	-	not halved
6.12	tʃ → ʃ	**ch**ai→**sh**y	Deaffrication	7
5.88	ʃ → tʃ	**sh**y→**ch**ai	-	not halved

teeth and lower lip, glottal sounds with the vocal cords, and others with the tongue and various articulators listed in the column. Each phoneme is color-coded in accordance with the degree in which the PER decreases: red indicates early acquisition, blue indicates later acquisition, and black indicates phonemes that remain difficult even as children grow older.

For each place of articulation, the consonants /j/ and /ɡ/, articulated on the hard and soft palates, remain difficult to recognize correctly even in older children. For manners of articulation, affricates and fricates—produced by forming a narrow passage for airflow—show delayed improvement compared with other types.

4.5. Analysis of frequently observed substitution errors

We analyzed the frequently observed substitution errors among grade 0 children. Table 5 lists the ten most common consonant substitution errors in this group. The phonological processes responsible for these errors were classified into seven categories used by Fringi [7], which are known to cause mispronunciations. If the error rate for a particular substitution is certain at a particular grade remains low, it has been "halved". Six of the ten errors were predictable on the basis of known phonological processes, while the remaining four were errors subjectively reasonable. In particular, fricatives and affricates are often misrecognized. Many of these errors align closely with typical child mispronunciations, often reflecting common phonological processes, which indicates that these misrecognitions are similar to actual mispronunciations.

However, many substitution errors do not halve even as the child grows older, so, likely, they may be not unique to children. However, many substitution errors do not halve even as the child grows older, suggesting that several may not be unique to children. Notably, "Stopping"—the replacement of fricative with stops—halves by grade 1, indicating its specificity to the lower grades. In addition, "Deaffrication" halves as the child grows older. These results suggest that younger children tend to substitute difficult-to-pronounce affricates and fricatives with phonemes that are easier to pronounce.

5. Conclusion

We analyzed error patterns in child ASR using state-of-the-art SSL-based PRs. We first categorized these errors into acoustic-related and linguistic-related types on the basis of a listening survey conducted by a human annotator. The experimental results indicated that the primary difficulties in recognizing young children's speech stem from their own mispronunciations (i.e., linguistic-related errors) . We conducted a comprehensive error analysis using a large dataset spanning a wide age range, from kindergarten to grade 10, revealing how error patterns evolve with age. The results showed that most phonemes are recognized more accurately as children grow older. Notably, frequent recognition errors were observed for phonemes produced with the soft palate, hard palate, as well as fricatives and affricates in younger children. Since these substitution errors often aligned with known phonological processes in child speech, we conclude that the PR's errors reflect underlying phonological characteristics.

We believe these findings offer valuable insights for both engineering and scientific perspectives on children's speech. From an engineering perspective, the age-dependent nature of error patterns underscores the potential benefits of age-specific modeling approaches to improve child ASR performance. From a scientific perspective, this study suggests the effectiveness of automatic annotation, as the error patterns of the PR closely resembled those observed in the manual experiments.

6. References

[1] K. J. Rohlfing, N. Altvater-Mackensen, N. Caruana, R. van den Berghe, B. Bruno, N. F. Tolksdorf, and A. Hanulíková, "Social/dialogical roles of social robots in supporting children's learning of language and literacy—A review and analysis of innovative roles," *Frontiers in Robotics and AI*, vol. 9, 2022.

[2] E. Okur, R. Fuentes Alba, S. Sahay, and L. Nachman, "Inspecting Spoken Language Understanding from Kids for Basic Math Learning at Home," in *Proc. BEA*, 2023, pp. 692–708.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.

[4] R. Jain, A. Barcovschi, M. Yiwere, P. Corcoran, and H. Cucu, "Adaptation of Whisper models to child speech recognition," in *Proc. Interspeech*, 2023, pp. 5242–5246.

[5] P. Gurunath 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.

[6] J. Huttenlocher, H. Waterfall, M. Vasilyeva, J. Vevea, and L. V. Hedges, "Sources of variability in children's language growth," *Cognitive Psychology*, vol. 61, no. 4, pp. 343–365, 2010.

[7] E. Fringi, J. F. Lehman, and M. Russell, "Evidence of phonological processes in automatic recognition of children's speech," in *Proc. Interspeech*, 2015, pp. 1621–1624.

[8] D. N. Sorenson, "A fundamental frequency investigation of children ages 6–10 years old," *Journal of Communication Disorders*, vol. 22, no. 2, pp. 115–123, 1989.

[9] I. S. B. Nip and J. R. Green, "Increases in cognitive and linguistic processing primarily account for increases in speaking rate with age." *Child development*, vol. 84, pp. 1324–1337, 2013.

[10] A. Smit, L. Hand, J. Freilinger, J. Bernthal, and A. Bird, "The Iowa Articulation Norms Project and Its Nebraska Replication," *The Journal of speech and hearing disorders*, vol. 55, pp. 779–798, 1990.

[11] W. Cohen and C. Anderson, "Identification of phonological processes in preschool children's single-word productions," *International journal of language & communication disorders*, vol. 46, no. 4, pp. 481–488, 2011.

[12] B. Dodd, A. Holm, H. Zhu, and S. Crosbie, "Phonological development: A normative study of British English-speaking children," *Clinical linguistics & phonetics*, vol. 17, pp. 617–643, 2004.

[13] M. E. Beckman, A. R. Plummer, B. Munson, and P. F. Reidy, "Methods for eliciting, annotating, and analyzing databases for child speech development," *Computer Speech & Language*, vol. 45, pp. 278–299, 2017.

[14] V. P. Singh, H. Sailor, S. Bhattacharya, and A. Pandey, "Spectral Modification Based Data Augmentation For Improving End-to-End ASR For Children's Speech," in *Proc. Interspeech*, 2022, pp. 3213–3217.

[15] Z. Fan, X. Cao, G. Salvi, and T. Svendsen, "Using Modified Adult Speech as Data Augmentation for Child Speech Recognition," in *Proc. ICASSP*, 2023, pp. 1–5.

[16] 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.

[17] R. Fan, N. Balaji Shankar, and A. Alwan, "Benchmarking Children's ASR with Supervised and Self-supervised Speech Foundation Models," in *Proc. Interspeech*, 2024, pp. 5173–5177.

[18] J. Li, M. Hasegawa-Johnson, and N. L. McElwain, "Analysis of Self-Supervised Speech Models on Children's Speech and Infant Vocalizations," in *Proc. ICASSPW*, 2024, pp. 550–554.

[19] E. Fringi, J. F. Lehman, and M. Russell, "Analysis of phone errors in computer recognition of children's speech," in *Proc. SLaTE 2015*, 2015, pp. 101–105.

[20] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The OGI kids² speech corpus and recognizers," in *Proc. ICSLP*, 2000, pp. 258–261.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[22] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Proc. NIPS*, 2020, pp. 12 449–12 460.

[23] 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.

[24] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "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.

[25] Q. Xu, A. Baevski, and M. Auli, "Simple and Effective Zero-shot Cross-lingual Phoneme Recognition," in *Proc. Interspeech*, 2022, pp. 2113–2117.

[26] A. Meghanani and T. Hain, "LASER: Learning by Aligning Self-supervised Representations of Speech for Improving Content-related Tasks," in *Proc. Interspeech*, 2024, pp. 2835–2839.

[27] L. Block Medin, T. Pellegrini, and L. Gelin, "Self-Supervised Models for Phoneme Recognition: Applications in Children's Speech for Reading Learning," in *Proc. Interspeech*, 2024, pp. 5168–5172.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 6000–6010.

[29] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, pp. 369–376.

[30] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP: System Demonstrations*, 2020, pp. 38–45.

[31] "Wav2Vec2-Base," https://huggingface.co/facebook/wav2vec2-base.

[32] "Hubert-Base," https://huggingface.co/facebook/hubert-base-ls960.

[33] "WavLM-Base," https://huggingface.co/microsoft/wavlm-base.

[34] "g2pe," https://github.com/Kyubyong/g2p.

[35] "Fine-Tune Wav2Vec2 for English ASR with Transformers," https://huggingface.co/blog/fine-tune-wav2vec2-english.

[36] C. Goddard, S. Siriwardhana, M. Ehghaghi, L. Meyers, V. Karpukhin, B. Benedict, M. McQuade, and J. Solawetz, "Arcee's MergeKit: A Toolkit for Merging Large Language Models," in *Proc. EMNLP: Industry Track*, 2024, pp. 477–485.