Interspeech 2024
1-5 September 2024, Kos, Greece

Leveraging Phonemic Transcription and Whisper toward Clinically Significant Indices for Automatic Child Speech Assessment

Yeh-Sheng Lin[1,2], *Shu-Chuan Tseng*[1], *Jyh-Shing Roger Jang*[2]

[1]Institute of Linguistics, Academia Sinica
[2]National Taiwan University
yehsheng@gate.sinica.edu.tw, tsengsc@gate.sinica.edu.tw, jang@csie.ntu.edu.tw

Abstract

Diagnosing speech sound disorders (SSD) in children requires professional assessment by speech-language pathologists. Detecting and diagnosing a medical condition takes time and is usually expensive in terms of labor. However, early identification and treatment are essential for subsequent care. ASR-based child speech assessment prioritizes semantic understanding over phonetic accuracy, making it unsuitable for pronunciation assessment. This study uses phonemic transcription available in a normative dataset and utilizes pre-trained speech models to develop an automatic phoneme recognition model with a Phoneme Error Rate (PER) as low as 3.76%. Clinically relevant indices calculated from the model prediction are highly correlated with those from the original normative data. We regard these experimental results as solid evidence that validates the feasibility of our evaluation workflow for practical application in early screening for phonological development delays.

Index Terms: phoneme recognition, child speech assessment, norm-referenced indices

1. Introduction

1.1. Assessment of phonological development delay

Speech sound disorders (SSD) encompass a range of difficulties involving speech perception, motor production, or phonological representation of speech sounds and phonotactics[1]. SSD are one of the most common disorders among preschool and school-age children. The prevalence of SSD in young children is 8 to 9 %[2] with a rising trend [1]. For children with SSD, not only their language, psychological and social development are affected, it is likely that they experience feelings of isolation, depression and low self-esteem [2]. Speech-language-therapists conduct professional evaluation to determine whether a child has a delay in phonological development by referring to normative data of developing children. Early and intensive intervention is crucial for success and efficiency for treating children with SSD [3]. However, speech therapy demands medical resources and the process from caregivers' awareness of speech problems until clinical diagnosis usually takes time [4].

Percentage of Consonants Correct (PCC) and Percent Vowels Correct (PVC) [5, 6] are well-accepted indicators for evaluating SSD in children and are also widely used in clinical settings. The testing personnel manually transcribe recorded speech at the level of phonemes. Then the proportion of correctly pronounced consonants and vowels are calculated [7]. In this respect, automatic speech assessment models can serve as core techniques that facilitate rapid screening of speech problems. Automatic tools of these kinds are easy for parents and caregivers to use. Once severe delay is detected, clinical treatment can be sought as soon as possible. Moreover, the quantitative evaluation results generated by the automatic models can be used as an objective supplementary reference [4]. So far, non-automated computer-assisted speech analysis tools have enhanced speech therapists' efficiency. [8]. What we aim to achieve in this study is to develop an automated diagnostic assessment process that outputs near-expert phonemic transcription, reducing the workload of speech therapists. By so doing, they can dedicate themselves more to treatment than to diagnosis.

1.2. Automatic child speech assessment

Automatic child speech assessment has clear benefits, but it is also challenging. First of all, child speech data are difficult to acquire because they are restricted resource due to privacy protections. Reports are often based on small data with insufficient annotation granularity [4]. Conventional Automatic Speech Recognition (ASR) architectures often incorporate a language model that fundamentally prioritizes semantic understanding to optimize output results. Grapheme-to-phoneme conversion is then conducted to obtain phoneme sequences [9]. This method has an important caveat for our purposes. As it is not a direct speech-to-phoneme model, confusion caused by near-matches of words affects the precision of phoneme sequences pronounced by children. For speech evaluation, precise pronunciation variants, instead of optimized word sequences that reflect the expressed speech content are required. The applicability of evaluation results using such an ASR architecture would be limited [10]. For example, mispronunciation caused by distortion or substitution may not be recognized, as the language model is likely to perform auto-correction and output text that does not align with the acoustic features of the pronounced speech content.

1.3. Our contribution

In response to the challenges posed by the scarcity of children's speech datasets, a normative dataset of Mandarin-speaking preschool children is constructed [11]. A wordlist of 70 multisyllabic words with a balanced design of consonants, vowels, and tones, was read by 798 normally hearing, developing children aged 3 to 6 years old. The primary objective of this dataset construction was to create a scientific foundation that supports norm-referenced tests (NRT) for child speech evaluation. It was also designed to offer well-annotated

[1] American Speech-Language-Hearing Association (ASHA), https://www.asha.org/practice-portal/clinical-topics/articulation-and-phonology/

[2] National Institute on Deafness and Other Communication Disorders (NIDCD), https://www.nidcd.nih.gov/health/statistics/quick-statistics-voice-speech-language

10.21437/Interspeech.2024-1887

speech materials for sophisticated speech analysis for phonological development research.

Using the phonemic annotation of this normative dataset, we utilize the pre-trained speech model Whisper released by OpenAI [12] to build a customized Phoneme Recognition Model, for which we will show that the model performance is more than satisfactory. We have achieved a Phoneme Error Rate (PER) of 3.76%. We will also show that quantitative indices computed from our model output are highly correlated with the trends of the normative dataset. This model is likely to serve as a core technique that can be further developed into an automatic child speech assessment tool. This will enable early screening of delays in phonological development. A similar model structure using normative data of children can be used for L2 learning and for languages other than Mandarin. To the best of our knowledge, this is the first approach to training a PRM model on phonemic transcription of normative data and using the model to generate objective indicators for phonological development assessment.

2. Methodology

2.1. The normative dataset

798 Mandarin-speaking preschool children were recorded in a picture-naming task with the Sinica Child Balanced Wordlist [11]. Figure 1 shows the number of subjects by age and gender.

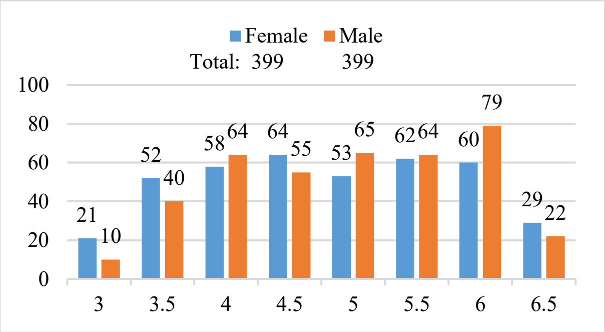

Figure 1: *Number of subjects.*

None of the children had known, diagnosed diseases related to language, hearing and cognitive development. The word list consists of 62 disyllabic and 8 trisyllabic words. A total of 55,860 words, equivalent to 118,104 syllables, were recorded. All onsets and rhymes eligible for composing Chinese syllables occur in both the first and second syllable positions. To facilitate easy reading and repetition for children, the word list was specifically designed to encompass semantic fields familiar to young children. The speech data were digitized at a sampling rate of 16 kHz. Signal-aligned boundary information for syllables was obtained using the ILAS phone aligner and manually post-edited in PRAAT [13].

For obtaining phonemic transcription, a two-stage annotation procedure was adopted. Accuracy and acceptability at the levels of “words”, “syllables”, and “tones” were first labelled by two annotators with high agreement rates as reported in [14]. “Correct syllables”, defined as those whose labels in word, syllable, and tone are all correct, were automatically converted to standard phoneme sequences, without entering into the second stage of the transcription process. The remaining 18,695 syllables (15.8% of the overall dataset) have at least one label among word, syllable, and tone annotated as “incorrect”. These syllables were phonetically transcribed using the International Phonetic Alphabet (IPA) by an experienced phonetician using PRAAT.

2.2. Phoneme recognition model

Mandarin Chinese syllables have at most four segments. The onset inventory consists of six plosives /p p^{h} t t^{h} k k^{h}/, six fricatives /f s ʂ ɕ x ʐ/, six affricates /ts ts^{h} tʂ tʂ^{h} tɕ, tɕ^{h}/, two nasals /m n/ and one lateral /l/, or it can be vacant. Only /n/ or /ŋ/ is allowed in the coda position. There are two glides /j w/ and 15 vowels including mono- and diphthongs /i ɨ ɯ u y a o ə e ɚ ai ei au ou ye/. Phoneme Recognition Model (PRM) serves as the core component of automatic speech assessment. Instead of manual phonemic transcription, a PRM outputs phoneme sequences pronounced by children. In recent years, significant breakthroughs have been achieved in speech applications with the development and release of pre-trained large-scale speech models. In this study, we adopted Whisper for our task. Whisper utilizes a transformer-based sequence-to-sequence model, which includes an encoder-decoder architecture and follows an end-to-end training approach. The input speech is segmented into 30-second fragments, transformed into log-Mel spectrograms, exhibiting superior noise resistance characteristics. These segments are then processed through an encoder, representing a deep understanding process of the speech features. The decoder is trained to predict text content corresponding to the speech, serving as a language model capable of directly outputting text in different languages for various recognition tasks. Whisper is trained on a vast and diverse set of speech data, including 680,000 hours of speech, comprising 117,000 hours across 96 different languages, along with 125,000 hours of transcribed and translated data [12].

Whisper has demonstrated excellent performance in recent speech applications, e.g., the classification of dysarthria, aphasia and child speech recognition tasks [15-17]. While Whisper's decoder demonstrates powerful performance, it lacks the ability to recognize and output information at the phoneme level. Moreover, its output is subject to automatic correction based on contextual information. Therefore, we utilize only Whisper's encoder and train a dedicated output layer to discern phoneme categories. Following the approach of wav2vec2 [18], we append a linear classification layer to the encoder and train it using Connectionist Temporal Classification (CTC) [19]. CTC is an efficient framework for speech recognition tasks, primarily employed to align variable-length input sequences with variable-length output sequences without explicit alignment information. It achieves this by utilizing a softmax output layer, which generates probability distributions over a set of target labels or characters, enabling the model to predict variable-length sequences effectively. CTC offers end-to-end training capabilities, robustness to temporal variability, and effective handling of noisy data. By leveraging recurrent neural networks and softmax output layers, CTC captures temporal dependencies and generates probability distributions over target labels, making it well-suited for real-world speech recognition applications.

We opt for this approach because the set of phonemes is finite. The output set of PRMs is relatively small and limited compared to text. However, in the fine-tuning process, unlike wav2vec2, we will not freeze the encoder of Whisper. Instead, we will train it together with the output layer.

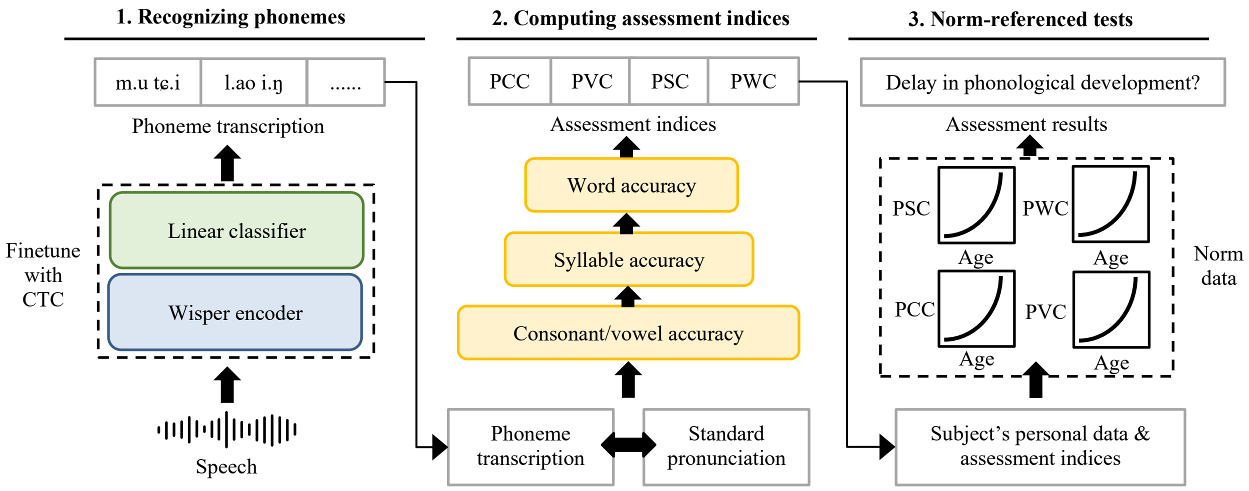

Figure 2: *Norm-referenced Child Speech Assessment Flowchart.*

2.3. Norm-referenced child speech assessment

We propose a workflow for constructing an automatic child speech assessment system utilizing the normative dataset as the norm and our Whisper-adapted PRM as the core technique, as shown in Figure 2. The PRM outputs predicted phoneme sequences for computing clinically crucial indicators that reflect the trends of phonological development of children. We consider consonants and vowels as well as syllables and words. Multi-level indices are proposed for two reasons. First, spoken language performance related to phonetic representation should be evaluated at various linguistic levels, e.g., segmental features, prosodic features and overall intelligibility [5, 6, 20]. In [20], prediction results of lexical tones in Mandarin-speaking children seem to be correlated with segment and syllable accuracy. Second, higher level information may suffice for early screening for severe phonological development delays.

Percentages of correct consonants and vowels are computed directly from the phoneme sequence output. "Correct syllables" must have all their syllable components correctly pronounced. Likewise, "correct words" must contain only the correct syllables. Applying the computation proposed in [6] as shown in (1), we propose to use PCC, PVC, PSC (Percentage of Syllables Correct) and PWC (Percentage of Words Correct) as assessment indices.

$$
\begin{gathered}
Percetage\ of\ x\ correct = \frac{Number\ of\ correct\ x}{Total\ number\ of\ x} \times 100 \\
x\ =\ \{Consonatn, Vowel, Syllable, Word\}
\end{gathered}
\tag{1}
$$

In this paper, we focus on the construction of the Whisper-adapted PRM and its performance by evaluating whether the output trends in terms of PCC, PVC, PSC, and PWC are correlated with the developmental data of our normative dataset. If the results are highly correlated, it is strong evidence supporting the proposed workflow's applicability for clinical applications. More sophisticated and comprehensive evaluation reports can then be developed to help diagnose SSD. For children with obvious delays, error types and analyses can be explicitly summarized using phoneme prediction. This procedure and tool are expected to relieve the burden of manual transcription for speech therapists significantly. In addition, an effective early screening tool will facilitate accurate and rapid evaluation of speech development delays.

3. Experiments and Results

For constructing the Whisper-adapted PRM, we added special symbols, including syllable delimiter, unknown token, and end-of-speech token during the training phase. Please note that in our phoneme transcription convention, speech sounds that do not occur in Mandarin's phoneme inventory are labeled "unknown token". We utilized the Whisper module provided by HuggingFace[1] for fine-tuning, ranging from the smallest size of model "tiny" to the largest scale of model "large". The model architecture employed Whisper's encoder followed by a fully connected layer for phoneme classification using linear classification. CTC loss was computed for training. Training was carried out on National Center for High-performance Computing (NCHC)[2] wtih Nvidia V100 GPUs. Each model size was trained for 20 epochs.

The default values for all parameters were used. Except for retaining the best-performing model for 20 epochs, no other hyperparameters were searched. Additionally, due to the small number of subjects in each age-gender subgroup, only the training and testing sets were split. We evenly selected 20% of speakers from each age-gender subgroup in the normative dataset as the testing set, while the remaining 80% were used for training. Since the speech data were recorded word by word, each training audio sample was processed in the units of "words".

Table 1 illustrates the relationship between the size of various models and their performance after fine-tuning, including PER on all words in the test set and the correlation between the model output and the norm data, i.e., the test set, in terms of Pearson correlation coefficient. The model performance increases as the model size grows. The best PER we achieved is 3.76%. There are currently no benchmarks for Mandarin-speaking children. But our PRM has improved significantly, compared to previous PRMs [21]. Besides, high correlation coefficients are found in all four indices, PCC, PVC, PSC, and PWC. Our PRM seems to perform well in the sense

[1] https://huggingface.co/openai/whisper-base

[2] https://www.nchc.org.tw/

that the indices calculated by the model prediction and by the norm data are highly correlated. Among them, PCC, PSC, and PWC seem more stable than PVC.

Table 1: *Performance of model.*

Size	Para-meters	PER (%)	Pearson Correlation Coefficient			
			PWC	PSC	PCC	PVC
tiny	8 M	7.46	0.918	0.928	0.941	0.842
base	21 M	6.04	0.947	0.949	0.951	0.857
small	88 M	4.63	0.967	0.965	0.972	0.880
medium	307 M	4.29	0.968	0.966	0.972	0.877
large	637 M	3.76	0.970	0.967	0.975	0.892

Figure 3 illustrates the relationship between the normative dataset (Norm, in blue) and the model prediction (Model, in orange) in terms of subject-based distributions of the four linguistic components. They clearly show the correlation and difference between manual and automatic approaches from the perspective of individual subjects. Except for the obvious discrepancies in the three-year-old subgroup, for which the data size is really small, the remaining subgroups exhibit a high level of consistency.

Furthermore, we averaged the model-assessed indices obtained from each subgroup and presented them in comparison with the norm data in Figure 4. The trends are mostly consistent. 3-year-olds have greater deviations than other subgroups. Nevertheless, the applicability of our proposed model-generated assessment indices is supported. Improvements can surely be achieved with more training data from 3-year-olds and other age groups.

4. Conclusion

We present in this study an automatic norm-referenced assessment workflow using a Whisper-adapted PRM trained on a normative dataset with phonemic transcription. We fine-tuned Whisper's encoder with a CTC algorithm to train a linear classifier for phonemes. Clinically relevant, multi-level assessment indices were generated from the model prediction and the norm data. They are highly correlated, supporting the applicability of our PRM and workflow as objective assessment for screening for phonological development delays. In addition to the low performance in the 3-year-old subgroup due to data scarcity, we also found that vowels may not serve as a suitable indicator for child speech assessment, speculatively due to indistinctive spectral features of vocalic qualities. For future work, more sophisticated workflows will be designed for developing comprehensive evaluation reports. Currently, we are testing the validity of our PRM on hearing-impaired children's data.

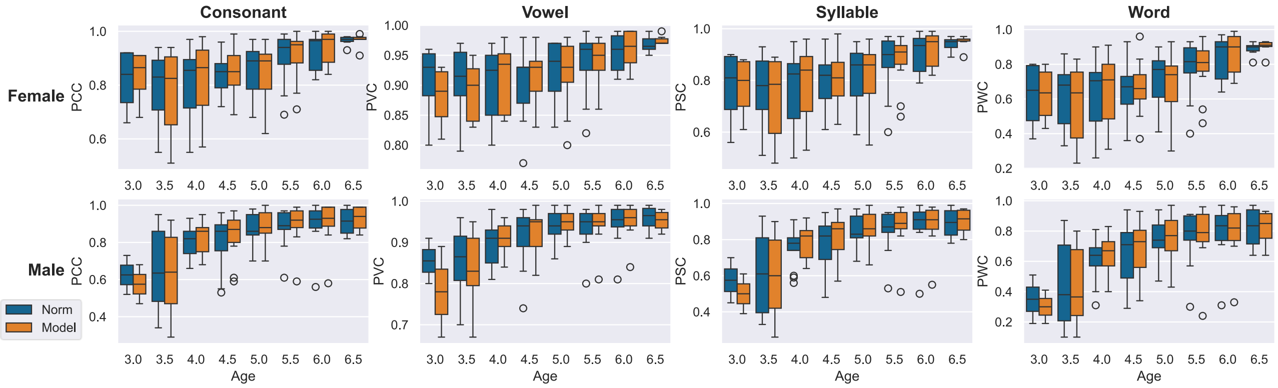

Figure 3: *Boxplots of subject-based assessment indices using Norm data and Model prediction.*

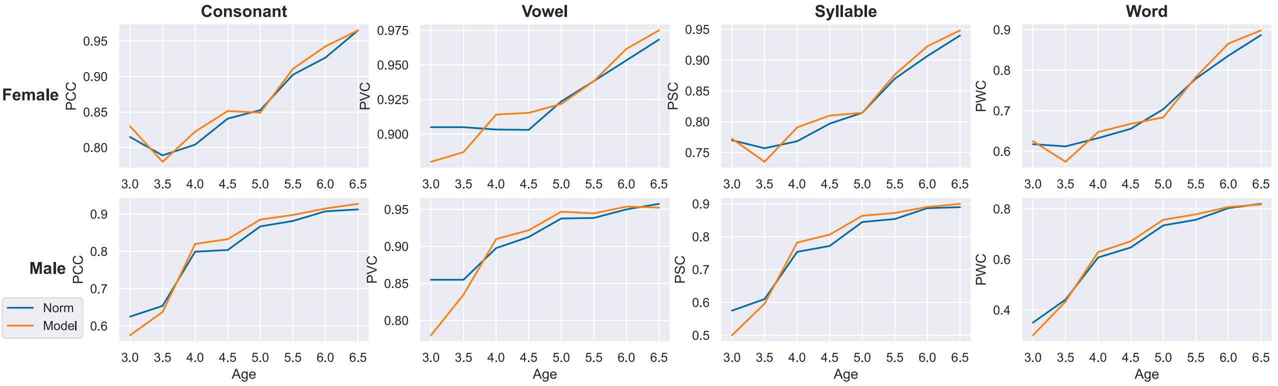

Figure 4: *Developmental patterns by Norm data and Model prediction.*

5. References

[1] S. K. Ravi, P. Sumanth, T. Saraswathi, M. A. B. Chinoor, N. Ashwini, and E. Ahemed, "Prevalence of communication disorders among school children in Ballari, South India: a cross-sectional study," *Clinical Epidemiology and Global Health,* vol. 12, p. 100851, 2021.

[2] S. N. d. Simoni, I. C. Leidow, D. L. Britz, D. A. d. O. Moraes, and M. Keske-Soares, "Impact of the speech sound disorders: family and child perception," *Revista CEFAC,* 2019.

[3] H. McFaul, L. Mulgrew, J. Smyth, and J. Titterington, "Applying evidence to practice by increasing intensity of intervention for children with severe speech sound disorder: a quality improvement project," *BMJ Open Quality,* vol. 11, no. 2, p. e001761, 2022.

[4] G. P. Usha and J. S. R. Alex, "Speech assessment tool methods for speech impaired children: a systematic literature review on the state-of-the-art in Speech impairment analysis," *Multimedia Tools and Applications,* pp. 1 - 38, 2023.

[5] L. Shriberg, D. Austin, B. A. Lewis, J. L. Mcsweeny, and D. L. Wilson, "The percentage of consonants correct (PCC) metric: extensions and reliability data," *Journal of speech, language, and hearing research : JSLHR,* vol. 40 4, pp. 708-22, 1997.

[6] L. Shriberg and J. Kwiatkowski, "Phonological disorders II: a conceptual framework for management," *The Journal of speech and hearing disorders,* vol. 47 3, pp. 242-56, 1982.

[7] R. E. Owens Jr, *Early language intervention for infants, toddlers, and preschoolers*. Pearson, 2017.

[8] N. B. John Bernthal, Peter Flipsen, *Articulation and Phonological Disorders: Speech Sound Disorders in Children*, 8 ed. Pearson, 2016.

[9] D. Towey *et al.*, "CHOCSLAT: Chinese Healthcare-Oriented Computerised Speech & Language Assessment Tools," *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC),* pp. 1460-1465, 2020.

[10] E. Cámara-Arenas, "Automatic pronunciation assessment vs. automatic speech recognition: A study of conflicting conditions for L2-English," 2023.

[11] S.-C. Tseng, "ILAS Chinese Spoken Language Resources," in *LPSS*, Taipei 2019, pp. 13-20.

[12] 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*, 2023: PMLR, pp. 28492-28518.

[13] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.,* vol. 5, no. 9, pp. 341-345, 2001.

[14] S.-C. Tseng, "Corpus-based research on speech acquisition and automatic assessment of Taiwan Mandarin-speaking children aged 3 to 6," in *Linguistic Diversity, but Unity in Research: Celebrating the Twentieth Anniversary of the Institute of Linguistics, Academia Sinica*, S.-C. T. a. E. Zeitoun Ed. Taipei: Institute of Linguistics, Academia Sinica, 2024, pp. 475-502.

[15] S. Rathod, M. Charola, A. Vora, Y. Jogi, and H. A. Patil, "Whisper Features for Dysarthric Severity-Level Classification," in *Proc. INTERSPEECH*, 2023, pp. 1523-1527.

[16] L. Wagner, M. Zusag, and T. Bloder, "Careful Whisper - leveraging advances in automatic speech recognition for robust and interpretable aphasia subtype classification," in *Proc. INTERSPEECH*, 2023, pp. 3013-3017.

[17] 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.

[18] 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. 12449-12460, 2020.

[19] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369-376.

[20] S.-C. Tseng, Y.-F. Liu, and X.-L. Lu, "Model-assisted Lexical Tone Evaluation of three-year-old Chinese-speaking Children by also Considering Segment Production," in *Proc. INTERSPEECH 2023*, 2023, pp. 3909-3913.

[21] M. Malakar and R. B. Keskar, "Progress of machine learning based automatic phoneme recognition and its prospect," *Speech Communication,* vol. 135, pp. 37-53, 2021/12/01/ 2021.