SCSMT: A Multilingual Children's Speech Corpus for Singapore's Mother Tongues

Bowen Zhang*†, Nur Afiqah Abdul Latiff*, Rong Tong*, Donny Soh* and Ian McLoughlin*
* Singapore Institute of Technology, Singapore
E-mail: {bowen.zhang, nurafiqah.abdullatiff, tong.rong, donny.soh, ian.mcloughlin}@singaporetech.edu.sg
† Nanyang Technological University, Singapore
E-mail: bowen009@e.ntu.edu.sg

***Abstract*—This paper presents the Singapore Children's Mother Tongue Speech Dataset, a multilingual corpus containing around 30 hours of speech from 659 children aged 7-9 years across three languages: Mandarin, Malay, and Tamil. The dataset includes read speech from structured text and spontaneous speech from picture-based question responses. We detail corpus design, participants, data acquisition, preprocessing, and annotation procedures. Comprehensive phonological and semantic analyses were conducted to examine pronunciation patterns and response content quality. We identify common linguistic challenges specific to each language, providing insights for mother tongue education. This dataset is a valuable resource for children's automatic speech recognition and speech assessment tools, while contributing to understanding multilingual language acquisition in young learners.**

I. Introduction

Singapore is a highly multilingual society where English serves as the primary medium of communication. Mandarin, Malay, and Tamil are designated as Mother Tongue languages to preserve cultural heritage and linguistic diversity [1]. Despite government efforts to promote bilingual education, many students face challenges in mastering their Mother Tongue due to the dominance of English in daily communication [2]. As language acquisition in childhood plays a crucial role in long-term linguistic proficiency [3], developing resources to support the teaching and learning of these languages are essential.

ASR plays an important role in Computer-Assisted Language Learning (CALL) [4], but developing an ASR system to recognize children's speech presents unique challenges. Due to their smaller vocal folds and shorter vocal tracts, children's speech differs acoustically from adult speech, typically exhibiting higher fundamental (pitch) and formant frequencies [5], [6]. In addition, children's speech tends to vary more in pronunciation, articulation, and fluency [7]. As existing ASR models are trained on adult datasets, these models have difficulty in recognizing children's speech accurately [5]. Additionally, multilingual speech processing faces phonetic and syntactic variations across languages, further complicating model development. The lack of large-scale, high-quality speech datasets for children, especially in Mandarin, Malay, and Tamil, worsens these issues. It hinders advancements in speech technology and language learning applications. Therefore, there is a pressing need for specialized, well-annotated speech resources that accurately capture the natural speech patterns of children, especially within Singapore's multilingual environment.

This study aims to address these challenges by developing a Singaporean children's speech dataset covering the three Mother Tongue languages. The dataset includes around 30 hours structured speech recordings from primary school students, incorporating reading and picture description tasks to capture both scripted and spontaneous speech. By providing high-quality, annotated speech data, this study contributes to the enhancement of ASR systems for children, facilitates speech evaluation tools for education, and supports linguistic research on children's language acquisition. The dataset serves as a valuable resource for advancing speech technology in multilingual settings and promoting the effective teaching of Mother Tongue languages in Singapore.

II. Dataset Development

This section summarizes the dataset creation workflow, from participant recruitment and demographics to data processing and annotation. It describes the recording setup and protocols for capturing high-quality speech, followed by the design of reading and picture description tasks to elicit both scripted and spontaneous speech. Finally, it outlines the data processing and annotation pipeline, combining automated segmentation, ASR transcription, and human refinement to ensure a robust dataset for downstream applications.

A. Participants

To develop speech resources for Mother Tongue education in Singapore, speech samples were collected from 10 primary schools covering three linguistic communities. Participants were selected based on their enrollment in Mother Tongue language subjects (Mandarin, Malay, or Tamil), age ranging from 7 to 9 years old, with parental and school consent. Efforts were made to ensure diversity in linguistic backgrounds, gender, and proficiency levels.

979-8-3315-7206-8/25/$31.00

TABLE I
PARTICIPANTS BY LANGUAGE AND PRIMARY LEVEL GROUPS. P1: PRIMARY ONE; P2: PRIMARY TWO.

	Mandarin		Malay		Tamil		Total
	P1	P2	P1	P2	P1	P2	
Female	94	72	96	84	16	22	384
Male	65	101	46	40	9	14	275
Total	332		266		61		659

A total of 1,318 recording sessions were conducted, involving 659 students, with 275 male (44%) and 384 female (56%) participants, as shown in Table I. The dataset reflects Singapore's multilingual education landscape and provides valuable insights into children's speech patterns.

B. Data Collection Setup

Each recording session lasted approximately 30 minutes, with 3 to 6 participants per session depending on classroom size. To minimize external disturbances, participants were seated in separate areas within a 15-square-metre space, ensuring sufficient distance between them. Each participant was accompanied by a member of the research team or designated student helper, who provided instructions, encouraged engagement, and managed the recording equipment to ensure a smooth data collection process.

The recordings were captured at 48 kHz sampling rate using microphones (Logitech Yeti and Elgato Wave:3), positioned approximately 30 centimeters from the participant's mouth to ensure high-quality audio. Inevitably, some background noise, such as prompts from the accompanying person, school bells, and classroom reverberations was also recorded. However, these unwanted noises were addressed in post-processing to enhance data quality.

C. Composition of Tasks

The data collection process consisted of two tasks: Reading Task and Picture Description Task[1]. In the Reading Task, participants were provided with four images and sets of sentences describing these images. The reading task covers a mix of vocabularies to ensure that all necessary phonological sounds and alphabets are present for each mother tongue language as seen in Fig. 1. Participants were required to read each sentence aloud, with pinyin provided for the Mandarin script to aid students.

For the Picture Description Task, participants were given six different images and had three minutes to observe and interpret each image before responding to a set of 4 to 7 questions. The questions are designed following the 5W1H (Who, What, When, Where, Why, How) framework in English as shown in Fig. 2. They were curated based on general vocabulary availability in Primary 1 textbooks and educational resources, ensuring alignment with the curriculum. The questions were translated into respective Mother Tongue languages, recorded, and played for students, to simulate oral practice in schools.

[1]https://github.com/zbowen0225/SG-children-dataset

Fig. 1. Example of reading texts for Mother Tongue languages.

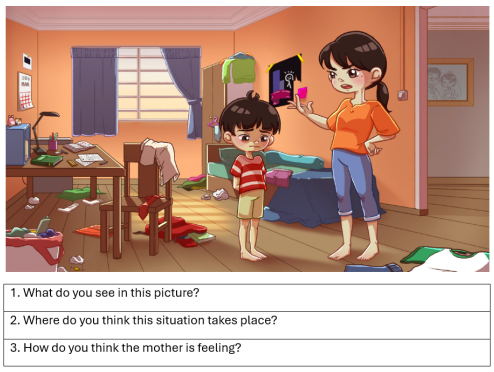

Fig. 2. Example 5W1H framework picture description task.

D. Data Processing and Annotation

To streamline data collection and processing, custom applications were developed: one to record the data collection session, and the other to automate key post-processing steps such as speech segmentation, transcription and annotation. After each data collection session, the system generates a recording file along with a corresponding JSON metadata file, which stores general session information, speaker labels, and timestamps of children's speech segments. These files are then processed to extract individual speech utterances for further analysis.

During data processing, the application automatically segments children's speech based on the labels and timestamps recorded in the JSON file. Additionally, it integrates Whisper [8], a state-of-the-art ASR model, to generate automatic transcriptions of each segment. Annotators can manually adjust the segmentation boundaries, correct transcription errors, and assign a fluency score for each spoken sentence. This semi-automated workflow ensures both efficiency and accuracy in data processing, providing high-quality annotated speech data for future research.

III. Phonological Analysis

This section examines language-specific patterns, including syllable or character structure, prevalent pronunciation trends, and common student errors, providing a detailed analysis of mistakes unique to each language.

Standard error measurement metrics obtained from the ASR output were used to measure the accuracy of student's speech i.e., Word Error Rate (WER), Character Error Rate (CER), and Phoneme Error Rate (PER). WER and CER were calculated based on the number of substitutions, deletions, and insertions made by students, relative to the total number of words or characters in the reference text respectively [9]. PER, phoneme error rate, was measured with respect to the International Phonetic Alphabet (IPA) standard for respective languages [10].

A. Mandarin

A Mandarin pinyin word consists of an initial, final and a lexical tone. Mandarin has four lexical tones: high-level (Tone 1), high-rising (Tone 2), falling-rising (Tone 3) and high-falling (Tone 4) [11]. Table II shows the initial and final pinyin components in Mandarin syllables. The initial component refers to the consonant that begins the syllable, while the final consists of the vowels that follow [12]. Finals are categorized into three groups: simple vowels, compound vowels, and nasal vowels.

A change in tone can lead to a change in the meaning of a syllable, even when the vowel and consonant components remain the same. Therefore, the error metric Tone Error Rate (TER) is applicable specifically to the Mandarin language to measure the number of tone errors between the actual and reference syllables in a given utterance.

ASR errors shown in Table III, indicate CER decreasing from 16.8% in Primary 1 (P1) to 10.1% in Primary 2 (P2). A similar pattern can also be observed for PER and TER. Although the decrease in error rates is relatively small, the initial error rates in P1 were notably low.

The largest proportion of errors made by Mandarin students were tone errors, accounting for 34.1% of the total. Final errors, accounted approximately 11%, while 3.6% of the errors were due to initial consonant mistakes. The second most common type of error was a combination of tone and final errors, compromising approximately 20% of the total errors made. Around 16.5% of the errors were due to mistakes in all components of the Mandarin syllable.

Fig. 3 presents the Mandarin tone confusion matrices for P1 and P2 students, illustrating the percentage of tone substitution errors in their pronunciation. In P1, notable confusion is observed between Tone 2 and Tone 1 (35.8%) and between Tone 4 and Tone 3 (36.1%), indicating early-stage difficulty in differentiating rising and high-falling tones. Additionally, Tone 1 is frequently mispronounced as Tone 4 (34.6%), suggesting instability in tone perception. By P2, the overall distribution of errors shifts, with a notable improvement in Tone 1 and Tone 2 accuracy. However, Tone 3 continues to be frequently substituted as Tone 2, and Tone 4 continues to be frequently substituted as Tone 3, highlighting a persistent challenge in producing high-falling tones. These differ significantly to predominant tonal confusion patterns of adult Mandarin [13].

TABLE II
Mandarin Pinyin Initials and Finals

Initial	b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, r, z, c, s, y, w
Final	a, ai, an, ang, ao, e, ei, en, eng, er, i, ia, ian, iang, iao, ie, in, ing, iong, iu, o, ong, ou, u, ua, uai, uan, uang, ue, ui, un, uo, v

TABLE III
P1 and P2 students Character (CER), Phoneme (PER) and Tone (TER) error rates for Mandarin.

Mandarin	CER	PER	TER
P1	16.8%	12.9%	13.7%
P2	10.1%	8.3%	8.8%

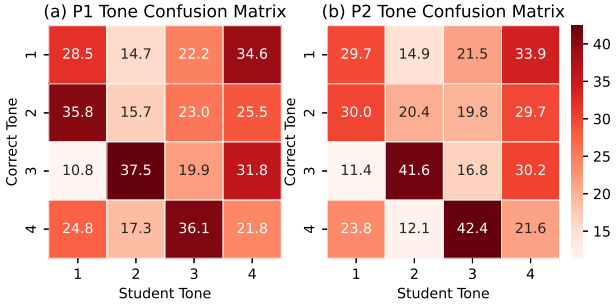

Fig. 3. Mandarin tone confusion matrices for Primary 1 (P1) and 2 (P2) students, showing percentage tone substitution errors. Rows indicate target tones, columns show actual spoken tones, with percentages representing the distribution of student responses.

B. Malay

The Malay speech unit, the syllable, typically consists of a vowel or a combination of a vowel and a consonant [14]. Table IV shows the phonemes found in Malay syllables.

As shown in Table V, P1 Malay students made approximately twice as many mistakes as those in P2 students, with a WER of 47% compared to 24.9%. A similar pattern was observed for CER and PER. Significantly, student error rates approximately halved in the course of a year.

Syllabification was applied to the annotated Malay texts to segment them into syllables. After analysing the syllables, the errors were classified into three types: vowel mispronunciations, consonant mispronunciations, and the

TABLE IV
Smallest sound units (phonemes) of Malay syllables

Vowel	a, e, i, o, u
Consonant	b, c, d, f, g, h, j, k, l, m, n, p, q, r, s t, v, w, x, y, z, ny, ng, kh, sy, gh, th

TABLE V
P1 and P2 students. Word (WER), Character (CER) and Phoneme (PER) error rates for Malay.

Malay	WER	CER	PER
P1	47.0%	17.3%	16.1%
P2	24.9%	8.5%	8.7%

TABLE VI
Commonly mispronounced Malay reading task syllables

Syllable	Type	#	Syllable	Type	#
(ya, ye)	Vowel	179	(tik, tek)	Vowel	77
(ka, ke)	Vowel	67	(ga, gak)	Added	62
(un, on)	Vowel	50	(sya, sa)	Cons.	47
(sih, seh)	Vowel	38	(tak, ta)	Added	37
(sa, se)	Vowel	37	(pung, pong)	Vowel	33
(un, hun)	Added	33			

addition of extra letters to words. It was observed that the pattern of syllable error distributions were similar for both P1 and P2 students, with 60.5% of the total number of syllables having additional phoneme(s) added to them. 54.7% of the vowels and 38.3% of the consonants were mispronounced.

Table VI presents the syllables most frequently mispronounced by students. A common pattern was the overgeneralization of similar vowel sounds, such as 'ya' with 'ye', 'tik' with 'tek', and 'un' with 'on'. Substitutions often involved phonetically similar vowels e.g., 'a' or 'i with 'e', and 'u' with 'o'. These errors frequently appeared in common words such as *saya* ('I'), *itik* ('duck'), and *mereka* ('they'). Additionally, glottal stops were observed through insertion of final consonants such as 'k' to the end of the word e.g., *nasi* ('rice') pronounced as *nasik*. In addition, some complex consonants such as 'sy' and 'ng' were simplified. For example, 'sy' in the name *Syakir* from the reference text was pronounced as *Siakir* or *Sakir*. This may be due to the rarity of such phonemes in early education vocabulary. Such mispronunciations and glottal stops could be attributed to the influence of colloquial speech used in daily conversation, as opposed to formal Malay, Sebutan Baku [15]. A lack of exposure to this may also hinder students' ability to accurately pronounce words.

C. Tamil

A Tamil character may consist of a consonant only, a vowel only or a combination of both. There are 247 Tamil characters made from the vowel and consonents as shown in Table VII.

In this paper, Consonant Error Rate (ConER) and Vowel Error Rate (VowER) are additional error metrics specific for Tamil language only. ConER and VowER errors refer to the substitution, deletion, or addition of consonants and vowels in a sentence.

Table VIII compares the performance of Tamil students in Primary 1 and Primary 2. P2 students only performed slightly better than P1 students, with approximately 15% decrease in WER from 68.1% to 49.3%. A similar trend

TABLE VII
Smallest sound units (phonemes) in Tamil syllables, consisting of vowels and consonants

Vowel	அ (a), ஆ (aa), இ (i), ஈ (ii), உ (u), ஊ (uu), எ (e), ஏ (ee), ஐ (ai), ஒ (o), ஓ (oo), ஔ (au)
Consonant	க (ka), ங (nga), ச (ca), ஜ (ja), ஞ (nya), ட (tta), ண (nna), த (ta), ந (na), ன (nna), ப (ppa), ம (ma), ய (ya), ர (ra), ற (rra), ல (la), ள (lla), ழ (llla), வ (va), ஶ (sha), ஷ (ssa), ஸ (sa), ஹ (ha)

TABLE VIII
P1 and P2 Tamil students Word (WER), Character (CER), Consonant (ConER) and Vowel (VowER) error rates

Tamil	WER	CER	ConER	VowER
P1	68.1%	38.6%	32.7%	39.6%
P2	49.3%	22.8%	19.0%	24.5%

was observed for CER, ConER and VowER. The slight decrease in error rates indicates that after the span of a year, students continued to repeat many of the mistakes made in P1. This may be due to students coming from English-speaking households [16].

Upon closer analysis of the substituted characters, the errors were classified into two types: vowel errors and consonant errors. Students made more of the former than the latter. Among P1 students, 45.1% of substituted errors were vowel errors, 16.0% were consonant errors and the remaining 38.9% were a combination of both. Similarly, for P2 students, 35.6% were vowel errors, 19.0% were consonant errors and 45.4% were a combination of both.

Table IX presents commonly mispronounced Tamil characters. Students frequently substituted the vowels 'u' and 'i' with the simpler 'a' sound. This led to oversimplification of the vowel pronunciation e.g., 'tha (த)' mispronounced as 'thu (து)', 'ki (கி)' as 'ka (க)', and 'ru(ரு)' as 'ra (ர)'. Commonly used words like 'மரக்கிளையில்' mispronounced as 'மரக்களையில்' (on the tree branch) and 'அருகில் ' as 'அரகில்' (nearby) reflect these patterns. Some students confused short for long vowels such as 'paa (பா)' pronounced as 'pa (ப)' as seen in familiar words such as 'பால்' (milk) mispronounced as 'பல்' (teeth).

In conclusion, students' reading capabilities improved across all three Mother Tongue languages. However, each language exhibited distinct characteristics in terms of

TABLE IX
Common Tamil reading task mispronunciation

Character	Transliteration	Error	Count
(த, து)	(tha, thu)	Vowel	34
(கி, க)	(ki, ka)	Vowel	23
(ன, ன்)	(na, nn)	Vowel	21
(பா, ப)	(paa, pa)	Vowel	17
(து, த)	(thu, tha)	Vowel	16
(து, தி)	(thu, thi)	Vowel	14
(த், ந்)	(th, nn)	Consonant	13
(யி, ய)	(yi, ya)	Vowel	12
(ல், து)	(ll, thu)	Consonant & Vowel	11
(ரு, ர)	(ru, ra)	Vowel	11

learning outcomes. Although Mandarin language showed a relatively low improvement rate of approximately 7%, it is important to note that its initial error rate was already low. In contrast, Malay language demonstrated substantial improvement, nearly doubling the performance. Tamil showed only a modest improvement. These results suggest that environmental factors significantly influence students' linguistic abilities, such as tone perception instability in Mandarin, over generalization of pronunciation and glottal stops influenced by colloquial Malay, and limited improvement observed in Tamil.

IV. Semantic Analysis

In this section, we examine the semantic properties of student responses from the picture description task detailed in Section II-C. Leveraging the cross-lingual processing capabilities of Large Language Models (LLMs), we implement an automated assessment framework that enhances scoring efficiency. Through systematic prompt tuning, we optimize the system to produce more coherent and linguistically valid scoring distributions, thereby increasing the reliability of our automated assessment methodology [17].

Students' spontaneous speech was evaluated on four dimensions: content relevance, answer completeness, vocabulary richness, and grammatical accuracy, each rated on a scale from 1 (poor) to 5 (excellent). Content relevance refers to the degree of relevance between a student's response and the provided sample answer, while answer completeness assesses the extent to which key ideas are included. Vocabulary richness evaluates the variety and appropriateness of the words used, and grammatical accuracy measures the frequency and severity of grammatical errors. The first two criteria evaluate the content of a response, while the latter two focus on linguistic quality. In recognition of the students' young age, the use of simple vocabulary and sentence structures was considered appropriate and expected.

Table X presents the evaluation of picture-based question and answer responses using LLM-based scoring. Overall, students demonstrated strong content relevance in all three languages, with an average score of 4.14, where Mandarin leading at 4.35, followed by Malay at 4.07 and Tamil at 3.99. However, performance was lower for answer completeness, with an average score of 2.78 (Mandarin: 2.99, Malay: 2.68, Tamil: 2.69). Vocabulary richness received the lowest average score of 2.48, with Mandarin again highest (2.58), followed by Tamil (2.57) and Malay (2.30). Grammatical accuracy outperformed vocabulary, averaging 3.48 overall; Mandarin scored 3.73, Malay 3.56, and Tamil 3.17.

The results indicate that students' responses were generally relevant, with an average content relevance score of 4.14. However, answers often lacked detail, as reflected by a lower completeness score of 2.78, which may be linked to their limited vocabulary (average 2.48). Despite these limitations, students demonstrated solid grammatical proficiency, with an average score of 3.48.

TABLE X
Comparison of Mandarin, Malay, and Tamil students' performance based on LLM evaluation scores (1 to 5) for four content criteria. Higher is better.

Criteria	Mandarin	Malay	Tamil	Avg
Content Relevance	4.35	4.07	3.99	4.14
Answer Completeness	2.99	2.68	2.69	2.78
Vocabulary Richness	2.58	2.30	2.57	2.48
Grammatical Accuracy	3.73	3.56	3.17	3.48

However, a limitation of this method is that if a student provides a single-word answer, the LLM may still assign a relatively high score for vocabulary richness. This is reflected in the lower vocabulary richness score for Malay (2.30) compared to Tamil (2.57). Despite these limitations, the method has demonstrated reasonable effectiveness as an evaluation tool.

V. Downstream application

The collected dataset serves as a valuable resource for advancing speech technology, education, and linguistic research, particularly in addressing the challenges of children's speech processing within multilingual, non-native language environments.

A. Children's ASR

A key application of this dataset is the enhancement of Automatic Speech Recognition systems. By offering a high-quality corpus of children's speech in three languages, this dataset enables effective ASR model adaptation, fine-tuning, and robustness testing, ultimately improving recognition accuracy for children in educational settings. Data augmentation techniques such as pitch and speed perturbation and the addition of background noise can be used to further enrich dataset diversity.

Pretrained Openai Whisper Medium [8] ASR models were selected and fine-tuned for each languages. Model performance were evaluated using Character Error Rate (CER) for Mandarin language and Word Error Rate (WER) for Malay and Tamil languages. CER and WER were calculated based on the number of substitutions, deletions, and insertions between the predicted transcriptions and the reference (ground truth) speech.

As shown in Table XI, a significant decrease in error rates was observed, with the fine-tuned models achieving error rates that were, on average, four times lower than their respective baselines. For Mandarin, the CER decreased from 29.4% to 7.5%, followed by Malay, which improved from 49.9% to 10.1%, and Tamil, which decreased from 79.5% to 35%. Notably, the initial baseline CER for Mandarin was lower than that of Malay and Tamil, likely due to the greater availability of publicly accessible training data. Overall, the high-quality corpus has demonstrated strong potential in enhancing ASR model performance.

TABLE XI
PERFORMANCE METRICS COMPARING BASELINE AND FINE-TUNED WHISPER MEDIUM MODELS ON THE COLLECTED SPEECH DATASET.

Language	Baseline CER/WER	Fine-tuned CER/WER
Mandarin	29.4%	7.5%
Malay	63.6%	10.1%
Tamil	79.5%	35.0%

B. Education Evaluation

Beyond ASR, the dataset supports the development of educational tools, such as automated speech evaluation systems, language learning applications, and educational analytics. Human-annotated transcriptions and fluency scores assist in creation of tools to provide personalized feedback on pronunciation, fluency, and tone accuracy, helping students improve oral language skills [17].

VI. CONCLUSION

This paper presented the Singapore Children Mother Tongue Speech Dataset, a comprehensive multilingual corpus spanning Mandarin, Malay, and Tamil. With 30 hours of speech from 659 children aged 7-9 years, this dataset addresses a critical gap in resources for children's speech technology and mother tongue education in Southeast Asia. The combination of reading and spontaneous speech provides a rich resource for understanding both controlled phonological production and natural language use among young multilingual learners.

Our phonological and semantic analyses revealed language-specific patterns and common errors that have direct implications for mother tongue education. These findings can inform curriculum development and teaching strategies tailored to each language's unique challenges. The dataset's careful design and annotation make it immediately applicable to developing children's ASR systems and automated speech assessment tools, which are increasingly important in educational technology.

Future work includes expanding the dataset to include more age groups and exploring cross-linguistic transfer effects in multilingual children. We aim to support further research in children's speech technology and to advance mother tongue education in multilingual societies.

ACKNOWLEDGMENT

This research is supported by the National Research Foundation, Singapore under AISG Award No: AISG2-GC-2022-004. We also acknowledge the contributions of the many student helpers for their assistance.

REFERENCES

[1] C. L. Lee and C. P. Phua, "Singapore bilingual education: One policy, many interpretations," *J. Asia-Pac. Comms.*, vol. 30, no. 1-2, pp. 90–114, 2020.

[2] J. Lo Bianco, S. Jones, and J. Loh, *English language education in singapore: Research, practice & implications*, 2021.

[3] J. Finders, E. Wilson, and R. Duncan, "Early childhood education language environments: Considerations for research and practice," *Frontiers in Psychology*, vol. 14, p. 1 202 819, 2023.

[4] K. Beatty, *Teaching & researching: Computer-assisted language learning*. Routledge, 2013.

[5] G. Chen, X. Na, Y. Wang, *et al.*, "Data augmentation for children's speech recognition–the" ethiopian" system for the slt 2021 children speech recognition challenge," *arXiv preprint arXiv:2011.04547*, 2020.

[6] R. Tong, L. Wang, and B. Ma, "Transfer learning for children's speech recognition," in *2017 International Conference on Asian Language Processing (IALP)*, IEEE, 2017, pp. 36–39.

[7] B. Munson, "Phonological pattern frequency and speech production in adults and children," 2001.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, 2022.

[9] A. C. Morris, V. Maier, and P. D. Green, "From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition.," in *Interspeech*, 2004, pp. 2765–2768.

[10] D. R. Mortensen, S. Dalmia, and P. Littell, "Epitran: Precision g2p for many languages," in *Proc. 11th Int. Conf. Lang. Res. and Eval. (LREC)*, 2018.

[11] L. Zhang, "Tone features of Chinese and teaching methods for second language learners," *Int. J. Chinese Language Education*, vol. 5, pp. 45–66, 2019.

[12] H. Třísková, "The structure of the Mandarin syllable: Why, when and how to teach it," *Archiv orientální*, vol. 79, no. 1, pp. 99–134, 2011.

[13] I. McLoughlin, "Subjective intelligibility testing of Chinese speech," *IEEE Tran. ASLP*, vol. 16, no. 1, pp. 23–33, 2008.

[14] I. Ramli, N. Jamil, N. Seman, and N. Ardi, "An improved syllabification for a better Malay language text-to-speech synthesis (tts)," *Procedia Computer Science*, vol. 76, pp. 417–424, 2015.

[15] M. A. Bakar and L. Wee, "Pronouncing the Malay identity: Sebutan Johor-Riau and sebutan baku," in *Multilingual S'pore*, Routledge, 2021, pp. 142–158.

[16] M. R. Kalaimani and M. S. Kaliamoorthy, "Using it to improve the oral and aural performance of Tamil language pupils by developing their metacognitive skills.,"

[17] B. Zhang, N. A. A. Latiff, J. Kan, *et al.*, "Automated evaluation of children's speech fluency for low-resource languages," arXiv, 2025. [Online]. Available: https://arxiv.org/abs/2505.19671.