



# Automated evaluation of children’s speech fluency for low-resource languages

Bowen Zhang<sup>1,2</sup>, Nur Afiqah Abdul Latiff<sup>1</sup>, Justin Kan<sup>1</sup>, Rong Tong<sup>1</sup>, Donny Soh<sup>1</sup>, Xiaoxiao Miao<sup>1,3</sup>, Ian McLoughlin<sup>1</sup>

<sup>1</sup>ICT Cluster, Singapore Institute of Technology, Singapore

<sup>2</sup>College of Computing & Data Science, Nanyang Technological University, Singapore

<sup>3</sup>Division of Natural and Applied Sciences, Duke Kunshan University, China

bowen009@e.ntu.edu.sg, {nurafiqah.abdullatiff, justin.kan, tong.rong, donny.soh, xiaoxiao.miao, ian.mcloughlin}@singaporetech.edu.sg

## Abstract

Assessment of children’s speaking fluency in education is well researched for majority languages, but remains highly challenging for low resource languages. This paper proposes a system to automatically assess fluency by combining a fine-tuned multilingual ASR model, an objective metrics extraction stage, and a generative pre-trained transformer (GPT) network. The objective metrics include phonetic and word error rates, speech rate, and speech-pause duration ratio. These are interpreted by a GPT-based classifier guided by a small set of human-evaluated ground truth examples, to score fluency. We evaluate the proposed system on a dataset of children’s speech in two low-resource languages, Tamil and Malay and compare the classification performance against Random Forest and XG-Boost, as well as using ChatGPT-4o to predict fluency directly from speech input. Results demonstrate that the proposed approach achieves significantly higher accuracy than multimodal GPT or other methods.

**Index Terms:** speech recognition, human-computer interaction, computational paralinguistics

## 1. Introduction

In a multi-ethnic, multi-lingual society such as Singapore, or in mixed-race households, children may not be exposed to their “mother tongue” at home. To preserve cultural heritage and promote bilingualism, the Singapore Ministry of Education ensures that all primary school students learn their mother tongue, namely Mandarin, Malay, or Tamil, alongside English, which is the primary language of instruction and daily interaction. However, traditional methods of assessing oral fluency in these mother tongues often rely on human evaluation, which can be time-consuming, subjective, and inconsistent. Significant research into areas such as Computer-Assisted Language Learning (CALL) [1], offers potential solutions by integrating technology into language teaching and learning. CALL systems provide self-paced, interactive learning experiences and automated evaluation, with immediate feedback on pronunciation, fluency, and accuracy.

Among various assessments, researchers have developed several approaches to deriving fluency scores, which are key metrics to reflect the smoothness, naturalness, and efficiency of speech production. The methods are broadly categorised into non-ASR and ASR-based methods. The former aims to assess fluency without relying on ASR. For example, in [2], a generative model was trained on the marginal distribution of speech, where fluency is assessed by comparing raw speech without recognising specific words or phonemes. Similarly, self-supervised speech representation schemes, such as wav2vec [3], have been explored. These methods leverage pre-trained acous-

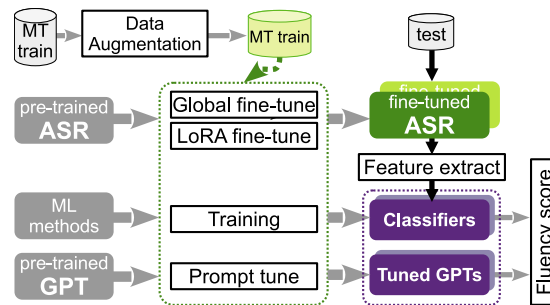


Figure 1: Proposed automatic fluency scoring framework showing adaptation of pre-trained ASR and GPT models using highly augmented low-resource mother tongue (MT) language data.

tic models to infer fluency-related metrics directly from speech, such as rhythm, pause patterns, and speech rate. While non-ASR models can be computationally efficient and less reliant on annotated data, they often lack the granularity to provide detailed insights into specific fluency issues, such as word-level disfluency or pronunciation errors.

In the realm of language learning, ASR-based methods have become dominant in recent years, where speech first transcribed using an ASR model can be analyzed for linguistic accuracy, fluency, and pronunciation [4, 5]. Key metrics include WER (Word Error Rate) for transcription accuracy, phoneme alignment for pronunciation, speech rate and pauses for fluency, and prosodic features (pitch, intensity, rhythm) for naturalness [6]. Though ASR-based solutions show promising performance for adult speech in high-resource languages such as English and Chinese, they face major challenges related to the variable nature of children’s speaking patterns, leading to low baseline accuracy for children’s speech. The scarcity of high-quality annotated children’s speech data, compounded by ethical concerns around data collection, pose additional obstacles. The challenges are even greater for low-resource languages.

In this paper, the challenges of developing low resource automated assessment systems for children’s mother tongue speech fluency are addressed specifically for the two lowest resource MLs in Singapore: Malay and Tamil. We propose an ASR-based language proficiency assessment system as illustrated in Fig. 1 and evaluate the proposed system for Malay and Tamil, demonstrating very good performance. These findings are expected to be useful for speech evaluation in other low-resource language scenarios. Specifically, to tackle training data scarcity, we adapt and fine-tune a high-quality multilingual adult ASR engine using a small set of labeled child speech data. From the fine-tuned ASR, we extract metrics such as phonetic and word error rates, speech rate, and speech-pause duration ratios. A GPT network is then prompt-tuned to generate fluency

prediction scores based on these metrics. The system is optimized to align with a limited sized human-evaluated ground truth training dataset (discussed in more detail in Section 3).

The system is effective because it not only successfully harnesses the established capability of the pre-trained multilingual adult ASR engine, but also makes use of powerful generalisation and flexible adaptation properties of the GPT network that can capture the nuances of human-like assessment decisions.

## 2. Related works

Speech fluency evaluation is inherently subjective and challenging to automate. This section revisits related work on dominant ASR-based evaluation models, which lead the field due to their capacity for fine-grained speech analysis.

These ASR-based methods transcribe learner speech into text, and also enable extraction of detailed metrics such as word error rate (WER), phoneme error rate (PER), and pause duration. For example, Goodness of Pronunciation (GOP) [7] is widely used to detect mispronunciations, which can significantly impact fluency. Recent advancements, such as Transformer models with multi-task learning from GOP features [8], have achieved promising results in fluency assessment for non-native speakers. Additionally, ASR-based systems can capture prosodic information like pitch, intensity, and rhythm, which are essential for evaluating the naturalness of speech. With the emergence of Whisper [9], more automated DNN driven speech evaluation methods have been developed. For example, MultiPA [10], a multitask pronunciation assessment model, extracts phone alignment features from Whisper and word embeddings from HuBERT [11] to enhance assessment accuracy. MOSA-Net+ [12] enhances MOSA-Net by incorporating Whisper acoustic features and self-supervised learning (SSL) embeddings for more accurate speech quality prediction.

While ASR-based models excel in adult speech evaluation for common languages, they face significant challenges when applied to children’s speech, or to low-resource languages. This is because both the front-end ASR model, which extracts fine-grained features for fluency evaluation, and the back-end DNN-based classifiers, which assess fluency scores based on these fine-grained features, require large-scale high quality labelled data for training. In the case of children’s speech, the variability in pitch, pronunciation, and speech patterns differs significantly from adult speech, making it difficult for models trained on adult data to generalize effectively. To address these challenges, researchers have explored various strategies. For instance, a multitask learning framework was proposed in [13] to model the characteristics of both adult and child speech, improving recognition accuracy for children. A FBDS-based automatic acoustic measure was proposed for Korean children speech fluency prediction [14]. Data augmentation techniques, such as pitch, speed, tempo, volume perturbation [15], have also been employed to simulate children’s speech and expand the available training data. Low-resource languages also lack sufficient labelled data, hindering the ability of ASR systems to learn the necessary linguistic features and patterns, leading to suboptimal performance. These challenges highlight the need for specialised datasets and tailored approaches to improve ASR accuracy and robustness in these domains. Several adaptation approaches have been tried, for example in low-resource Chinese children ASR [16]. Fine-tuning, by adapting the final layers of Whisper or Wav2vec2 models with children’s speech data, has shown promise at improving ASR performance for low-resource languages [17], so we also adopt this method.

Table 1: Number of utterances used to fine tune ASR models

Split	Number of utterances	Malay	Tamil
Train	Children	5853	2037
	Children + <i>AugA</i>	7359	6031
	Children + <i>AugB</i>	18329	6837
Test	Reading	178	114
	Picture	200	210

## 3. Children speech evaluation

To overcome the poor generalization of adult ASR for children’s speech, we collected a dataset of children’s Malay and Tamil speech for fine-tuning an adult ASR model. We leveraged our fine-tuned model to extract various objective metrics. These then served as input features for comparing traditional machine learning models and the proposed GPT-based approach to assess children’s speech fluency in both languages. The following sections discuss further.

### 3.1. Dataset collection and annotation

To address the lack of publicly available children’s speech datasets for Malay and Tamil, we collected an in-house dataset. What is unusual is the inclusion of both reading and question-answer (Q&A) tasks for each language. The dataset comprises speech samples from 218 male and 436 female primary one and primary two students in Singapore. In total, 654 recording sessions were conducted in real classroom environments. Data collection was approved by the Institutional Review Board of our university and by the Ministry of Education.

The reading task involved students reading 27 sentences aloud line by line. The Q&A task involved a picture description section, where students observed 6 pictures and answered 4 to 5 related questions per picture within a 3-minute timeframe. Each session was recorded, capturing both reading and conversational responses. Due to the real classroom settings, the recordings contain considerable ambient background noise.

Recordings were segmented into utterances, which were then transcribed by human experts to ensure accurate annotations. Each sentence was also assigned a fluency score on a scale of 1 to 4, with 4 being the highest. Since fluency scores of 1 and 2 were underrepresented in the collected data, we merged them into a single category labeled “low”. Fluency score 3 is labeled “medium”, and fluency score 4 is labeled “high”. To avoid nonsense input, we selected the medium and high fluency sentences only for ASR fine-tuning, but used all for fluency model training and classification.

### 3.2. Data augmentation

To address the scarcity issue, we employed two data augmentation methods. *AugA* utilised publicly available adult speech data to which we applied vocal tract length normalisation (VTLN) to simulate children’s vocal characteristics [18]. Children have shorter vocal tracts, resulting in higher formant frequencies. Bandpass filtering, formant shifting, pitch shifting (to 150Hz for male, 250Hz for female samples), and speed change (0.8) were also applied as in [15]. *AugB* varied the speech rate, pitch and intensity for both children and adults’ data to increase the volume of data for fine-tuning. Table 1 summarises the utterances per language, where the “Children” subset contains only medium and high fluency utterances, and either *AugA* or *AugB* substantially increases the data quantity. There was no speaker overlap between train and test sets.

### 3.3. Low resource Children’s fine-tuned ASR

To obtain the fluency-related metrics, we used the multilingual Whisper [9] large ASR model as a baseline. It was trained on 680,000 hours of weakly supervised speech and supports over 100 languages, making it a strong foundation for our task. For better performance on Malay and Tamil, we also utilized the open sourced pre-finetuned Whisper models for these two languages respectively.

To adapt to children’s speech, we fine-tuned the model parameters using the target children’s speech dataset described in section 3.1. Given Whisper’s strong latent language representations, fine-tuning enables the model to better capture pronunciation variations and fluency patterns in the target speech. To reduce training complexity, we employed Low-Rank Adaptation (LoRA) [19]. LoRA introduces trainable adapter layers while keeping the original model weights frozen, significantly reducing the number of trainable parameters. Instead of updating the entire weight matrix, LoRA factorizes it into smaller matrices, allowing efficient adaptation to children’s speech.

### 3.4. Fluency scoring metrics

Fluency scoring evaluates how smoothly and naturally a speaker produces language and typically is highly subjective. Aiming to automate the assessment, we extracted metrics from the fine-tuned ASR model output as summarised in Table 2

Table 2: Metrics for automatic fluency assessment

Primary metrics	
Language	Malay or Tamil
Task type	Reading (R) or Picture Q&A (P)
WER	word error rate
CER	character error rate
PER	phoneme error rate
Pause duration	duration between words
Total duration	total duration of speech
Num pauses	number of pauses in the sentence
Secondary metrics	
Speed	$\text{num\_words}/\text{total\_duration} \times 60$
Pause ratio	$\text{pause\_duration}/\text{total\_duration}$

This included four granularities for accuracy at utterance, word, character, and phoneme level. Errors in a single syllable can significantly impact both the character error rate (CER) and word error rate (WER), especially in phonetically based languages like Tamil, where small syllable changes can radically alter word meanings. The phoneme error rate (PER) is computed by converting words into their International Phonetic Alphabet (IPA) equivalent [20], to provide a more precise evaluation of pronunciation accuracy.

Speech rate is another key indicator of fluency. A consistently slow rate may suggest difficulty with language retrieval or articulation, while an overly fast rate can indicate disfluency due to unclear articulation or skipped words.

Speech pauses provide insight into cognitive processing and articulatory control. While natural pauses aid comprehension, excessive or poorly timed pauses may indicate fluency challenges. A high pause ratio may suggest language difficulties or lack of confidence. To capture these patterns, we measure the proportion of time spent pausing relative to total speaking time, along with speech rate and error rates at different linguistic levels (word, character, phoneme).

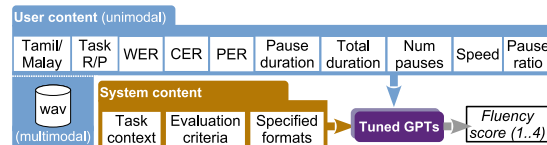


Figure 2: Automatic scoring using tuned GPTs. The fixed “system content” defines the task, the context and the input/output formats. “user content” contains per-instance input data.

### 3.5. Fluency score prediction model

Due to the scarcity of children’s data, especially in low-resource languages, training a deep neural network (DNN) adequately for speech evaluation is challenging. However, given our set of metrics and small but well labelled dataset, we can leverage traditional machine learning methods, such as random forest and XGBoost. We also compared performance against a multi-modal LLM, which can input both raw speech and text prompts. In this case we used prompt tuning to obtain fluency scores from metrics plus the raw speech.

#### 3.5.1. Traditional machine learning methods

We evaluated two machine learning methods for automatic fluency scoring: Random Forest and XGBoost. They both rely on ensemble learning techniques to improve prediction accuracy.

#### 3.5.2. GPT-based models

We also compare two GPT-assisted approaches. The first is a multimodal LLM, gpt-4o-audio-preview<sup>1</sup>, with direct speech inputs and a prompt-tuned fluency scoring task.

The second is gpt-4o-mini<sup>2</sup> with text-only input, prompt-tuned as a meta-evaluator based on our extracted metrics, context, and prototypes. Both models were pretrained on vast amounts of language data, which allows them to understand patterns in pronunciation, rhythm, and speech rate. The audio-input model is additionally trained on a large selection of speech samples, including speech in both languages.

When using LLMs, prompt design is crucial, and this is shown in Fig. 2<sup>3</sup>. The input prompt is composed of two main components, the “system” and “user content”. The first outlines the task context, specifies the evaluation metrics, defines the input-output format, and most importantly, provides prototypes and boundaries for the evaluation process. The second, “user content” is structured as a JSON object that encapsulates all necessary features for the speech evaluation. For the unimodal system, this included the metrics extracted from our fine-tuned Whisper models for each language, and contained speech utterance files for the multimodal system. The output is a quantitative fluency score for each utterance.

## 4. Experiment

We conducted extensive experiments to obtain an optimal ASR model. Whisper-medium for Malay and Whisper-small for Tamil were adopted for the full fine tuning, sized to match the data availability (much more training data for Malay than Tamil). We also leveraged two pre-finetuned Whisper models:

<sup>1</sup><https://platform.openai.com/docs/models#gpt-4o-audio>

<sup>2</sup><https://platform.openai.com/docs/models#gpt-4o-mini>

<sup>3</sup><https://github.com/zbowen0225/children-fluency-assessment>

Table 3: ASR performance on various models

Language/WER(%)		Malay task		Tamil task	
Model	Training	R	P	R	P
Whisper-sm	NA	71.16	72.28	86.33	86.67
Whisper-med	NA	60.96	66.30	80.09	78.94
Whisper-lg_v3	NA	55.81	59.21	79.12	76.96
Pre-finetuned	NA	53.62	46.14	40.48	50.43
Global fine-tune	Children	7.46	13.86	40.43	75.07
	+AugA	7.68	<b>13.70</b>	29.12	52.02
	+AugB	8.00	<b>13.70</b>	<b>24.41</b>	48.46
LoRA fine-tune	Children	9.54	15.28	39.11	52.19
	+AugA	10.09	16.06	30.75	45.59
	+AugB	<b>7.24</b>	14.31	32.90	<b>40.06</b>

Mesolitica-medium<sup>4</sup> for Malay and Vasista-medium<sup>5</sup> for Tamil as base models prior to the LoRA fine tuning. For fine-tuning, we used the datasets described in section 3.

The best ASR model in terms of picture task WER (prioritised for its broader vocabulary and realistic context) was adopted to extract the metrics described in Section 3.4. The machine learning speech evaluation models<sup>6,7</sup> were trained on metrics extracted from all utterances using default parameters.

#### 4.1. ASR model fine-tuning performance

First we assess the ASR performance we were able to achieve for both Malay and Tamil in Table 3. A lower WER indicates better model performance. R is the reading task and P is the picture Q&A task. The performance of the fine-tuned models is shown using different datasets and augmentations (Aug). All base models performed poorly on children’s speech, with Tamil performing worst. This is unsurprising because Tamil ASR has long been challenging due to limited training data, and children’s Tamil ASR even more so. Results show that both global fine tuning and LoRA fine tuning significantly improve performance, with Malay enjoying the greatest gain, likely due to having more fine tuning data.

For Tamil, global fine tuning was more effective on the R task, while the LoRA-finetuned model performed better on the P task. This may be because reading data has a more limited vocabulary, causing full fine tuning to overfit and become less generalised. Overall, data augmentation improved model performance, with one outlier being the LoRA-fine tuned model for Tamil task R. By analysing specific recognition results, we found that the LoRA model fine tuned on Children+AugB sometimes incorrectly combined or separated words.

#### 4.2. Fluency assessment

Next we assess our automated scoring framework via Pearson correlation, balanced accuracy [21, 22], and weighted F1 score (needed due to the imbalanced, multi-class nature, of the training data). Table 4 presents fluency prediction results. Among machine learning methods, XGB achieved the best performance for both Malay and Tamil.

However, both GPT systems outperformed traditional methods for both Malay and Tamil, which is impressive given that the GPTs are essentially performing zero-shot inference un-

<sup>4</sup><https://huggingface.co/mesolitica/malaysian-whisper-medium>

<sup>5</sup><https://huggingface.co/vasista22/whisper-tamil-medium>

<sup>6</sup><https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

<sup>7</sup><https://xgboost.readthedocs.io/en/stable/parameter.html>

Table 4: Results for RF and XGB; Corr: Pearson correlation; Acc: accuracy (%); F1:weighted F1 (%).

Methods	Malay			Tamil		
	Corr	Acc	F1	Corr	Acc	F1
RF	0.72	0.71	0.70	0.43	0.52	0.59
XGB	0.75	0.74	0.73	0.54	0.65	0.70
gpt-audio <sup>†</sup>	0.65	0.71	0.69	0.43	0.52	0.51
gpt-meta	<b>0.92</b>	<b>0.91</b>	<b>0.91</b>	<b>0.66</b>	<b>0.75</b>	<b>0.74</b>

Table 5: Ablation study of metrics input to the tuned GPT.

Excluded features	Malay			Tamil		
	Corr	Acc	F1	Corr	Acc	F1
wer	0.85	0.81	0.80	0.61	0.67	0.67
cer	0.91	0.90	0.90	0.60	0.70	0.68
per	0.92	0.91	0.90	0.63	0.72	0.71
pau duration	0.89	0.87	0.87	0.64	0.73	0.73
tot duration	0.91	0.90	0.90	0.63	0.73	0.72
num pauses	0.91	0.89	0.89	0.64	0.74	0.74
speed	0.91	0.90	0.90	0.64	0.74	0.73
pause ratio	0.89	0.88	0.88	0.59	0.74	0.74
base	0.92	0.91	0.91	0.66	0.75	0.74

like machine learning models that required training. The superior performance for Malay over Tamil can be directly attributed to its more effective ASR model, which yields more accurate acoustic and tempo metrics than the Tamil model. The multimodal LLM, gpt-4o-audio-preview<sup>†</sup>, which directly inputs speech waveforms, performed less well indicating that current LLM models may not be well trained for Tamil or Malay speech. This is likely to be true as long as both remain low-resource languages. Our proposed system (gpt-meta), outperformed all other methods by a significant margin. For Malay speech it achieved a highly impressive F1 score of 0.91, and it also outperformed all other systems for Tamil speech.

Finally, comprehensive ablation experiments analyse the impact of every metric on model performance. In Table 5, we remove each metric in turn and re-evaluate. Clearly, WER is the most important metric for Malay while CER and pause ratio are more important for Tamil. This may reflect the stress-timed syllabic nature of Tamil compared to the syllable-timed nature of Malay. Interestingly, no feature appears to be redundant.

## 5. Conclusion

This paper has explored the automated fluency evaluation for low resource languages. We propose an automated assessment system for Malay (low resource) and Tamil (very low resource), two important languages for education in Singapore. We adapt the powerful Whisper multilingual ASR model to extract metrics from utterances in each language and use machine learning models to provide fluency predictions from the metrics. Our Whisper model is extensively fine-tuned with speech highly augmented by state-of-the-art methods to significantly improve baseline accuracy in each language. For fluency evaluation, we compare two GPT systems and two machine learning methods. One GPT system is a state-of-the-art multimodal LLM directly inputting speech samples, while the other is our proposed text LLM acting as a meta-evaluator of extracted metrics. The proposed fluency evaluation system was found to perform well in both languages, and achieves greater than 90% accuracy in Malay. More importantly, the proposed technique has potential to be applied to other very low resource languages in future.

## 6. Acknowledgement

This research / project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG award no. AISG2-GC-2022-004) and with Infocomm Media Development Authority under its Trust Tech Funding Initiative (project no. DTC-RGC-07). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, Infocomm Media Development Authority or AI Singapore. The authors would also like to acknowledge the contributions of the many student helpers for their assistance in this project, the staff and the schools in which testing was conducted.

## 7. References

- [1] C. Baur, A. Caines, C. Chua, J. Gerlach, M. Qian, M. Rayner, M. Russell, H. Strik, and X. Wei, "Overview of the 2019 spoken CALL shared task." ISCA, 2019.
- [2] S. Cheng, Z. Liu, L. Li, Z. Tang, D. Wang, and T. F. Zheng, "ASR-free pronunciation assessment," in *Proc. INTERSPEECH 2020 – 21<sup>th</sup> Annual Conference of the International Speech Communication Association*, 2020, pp. 3047–3051.
- [3] S. Bannò, K. M. Knill, M. Matassoni, V. Raina, and M. Gales, "Assessment of L2 oral proficiency using self-supervised speech representation learning." ISCA SLaTE Workshop, 2023.
- [4] R. Matsuura, S. Suzuki, M. Saeki, T. Ogawa, and Y. Matsuyama, "Refinement of utterance fluency feature extraction and automated scoring of L2 oral fluency with dialogic features," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 1312–1320.
- [5] Y. Wang, M. J. Gales, K. M. Knill, K. Kyriakopoulos, A. Malinin, R. C. van Dalen, and M. Rashid, "Towards automatic assessment of spontaneous spoken English," *Speech Communication*, vol. 104, pp. 47–56, 2018.
- [6] P. Bamdev, M. S. Grover, Y. K. Singla, P. Vafaee, M. Hama, and R. R. Shah, "Automated speech scoring system under the lens: evaluating and interpreting the linguistic cues for language proficiency," *International Journal of Artificial Intelligence in Education*, vol. 33, no. 1, pp. 119–154, 2023.
- [7] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.
- [8] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7262–7266.
- [9] 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*. PMLR, 2023, pp. 28 492–28 518.
- [10] Y.-W. Chen, Z. Yu, and J. Hirschberg, "Multipa: A multi-task speech pronunciation assessment model for open response scenarios," in *Proc. INTERSPEECH 2024 – 25<sup>th</sup> Annual Conference of the International Speech Communication Association*, 2024, pp. 297–301.
- [11] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, 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.
- [12] R. E. Zezario, Y.-W. Chen, S.-W. Fu, Y. Tsao, H.-M. Wang, and C.-S. Fuh, "A study on incorporating whisper for robust speech assessment," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.
- [13] R. Tong, N. F. Chen, and B. Ma, "Multi-task learning for mispronunciation detection on Singapore children's Mandarin speech." in *Interspeech*, 2017, pp. 2193–2197.
- [14] L. Fontan, S. Kim, V. De Fino, and S. Detey, "Predicting speech fluency in children using automatic acoustic features," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 1085–1090.
- [15] G. Chen, X. Na, Y. Wang, Z. Yan, J. Zhang, S. Ma, and Y. Wang, "Data augmentation for children's speech recognition—the" ethiopian" system for the slt 2021 children speech recognition challenge," *arXiv preprint arXiv:2011.04547*, 2020.
- [16] W. Liu, Y. Qin, Z. Peng, and T. Lee, "Sparsely shared lora on whisper for child speech recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 751–11 755.
- [17] R. Jain, A. Barcovschi, M. Yiwere, P. Corcoran, and H. Cucu, "Adaptation of whisper models to child speech recognition," *Proc. INTERSPEECH 2023 – 24<sup>rd</sup> Annual Conference of the International Speech Communication Association*, 2023.
- [18] M. Qian, I. McLoughlin, W. Quo, and L. Dai, "Mismatched training data enhancement for automatic recognition of children's speech using DNN-HMM," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2016, pp. 1–5.
- [19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.
- [20] I. V. McLoughlin, *Speech and Audio Processing: a MATLAB-based approach*. Cambridge University Press, 2016.
- [21] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 3121–3124.
- [22] J. D. Kelleher, B. Mac Namee, and A. D'arcy, *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press, 2020.