Majumdar, R. et al. (Eds.) (2025). Proceedings of the 1st International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education

Improve English Pronunciation at Sentence Level for Thai EFL Learners With Thai Automatic Speech Recognition Model

Narabodee RODJANANANT[a], Phurinat POLASA[b], Phonlaphat NA POMPECH[a], Thanadech SAENGCHAN[a], Kongpop BOONMA[a], & Nattapol KRITSUTHIKUL[c*]
[a]*Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Thailand*
[b]*Department of Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi, Thailand*
[c]*National Electronics and Computer Technology Center (NECTEC), Thailand*
*nattapol.kritsuthikul@{nectec.or.th, gmail.com}

Abstract: ASR (Automatic Speech Recognition) is favorably chosen as a learning technology, which is used for English pronunciation practice. However, the base ASR model usually does not recognize the accent of the EFL (English as a Foreign Language) learner, especially in the sentence level. This research aims to extend the platform to improve English pronunciation from word level to sentence level for Thai EFL learners using a fine-tuned ASR with Thai datasets to detect Thai accent mispronounced sounds. The sentences are classified with CEFR level to provide learning steps for learners. The twelve of Grade 12 Thai native students were selected as sampling process. The results show that 75% of the samples have improved their pronunciation according to the CEFR level after using our system. Furthermore, the verb's present and past forms are most problematic.

Keywords: English as a Foreign Language (EFL), Mispronunciation, Thai, Automatic Speech Recognition (ASR), Computer aided pronunciation training (CAPT)

1. Introduction

Pronunciation plays a crucial role in foreign language learners' communication competence as it is directly connected to speech comprehensibility among interlocutors. ASR (Automatic Speech Recognition) is used in several works that state that it does not detect speakers with heavily accented sounds, including Google Speech-to-Text (Dillon & Wells, 2023; Kritsuthikul et al., 2024), SpeechNote (Evers & Chen, 2022; Sun, 2023), Window Speech Recognition (McCrocklin, 2019) and OpenAI's Whisper ASR (Radford et al., 2022; Ballier et al., 2023). Recent publication (Aung et al., 2024; Tipaksorn et al., 2024; SCB 10X, 2024) provides new pretrained fined-tune Thai speech recognition models and can improve the system. In addition, practicing speaking at sentence level is also important for communication.

This paper intends to improve word mispronunciation detection in Kritsuthikul et al. (2024) by employing the most effective pretrained fine-tuned Thai ASR models to detect sentence level.

2. The most effective pretrained fine-tuned Thai ASR models

The OpenAI's Whisper model (Radford et al., 2022) is highly regarded and has set a strong benchmark for ASR , especially for open-source models, we select the derivative pretrained fine-tuned Thai ASR based on the Whisper model as follows: Pathumma Whisper (Large), Thonburian Whisper (Small, Medium, Large) (Aung et al., 2024), and Monsoon Whisper (Medium) (SCB 10X, 2024). We used WER (Word Error Rate) as an ASR effective

measurement on a constructed corpus; CEFR-based sentences corpus (~1,100 sentences) annotated with 11 difficulties in mispronunciation for Thai EFL, as a result shown in Table 1.

Table 1. *WER of each ASR models*

ASR Model	Overall	A1	A2	B1	C1
Google Speech-to-Text API	0.4287	0.2130	0.4352	0.5093	0.4417
Whisper (Small)	0.3782	0.2315	0.1574	0.4167	0.3333
Whisper (Medium)	0.3095	0.1389	0.1481	0.3981	0.2833
Whisper (Large)	0.2588	0.0648	**0.0370**	0.3796	0.2750
Pathumma Whisper (Large)	0.2676	0.1667	0.0926	0.4070	0.2083
Thonburian Whisper (Small)	0.3324	0.2778	0.2593	0.4537	0.2917
Thonburian Whisper (Medium)	0.2576	0.1111	0.1570	0.3611	0.2250
Thonburian Whisper (Large)	**0.1699**	**0.0278**	**0.0370**	**0.3333**	**0.1333**
Monsoon Whisper (Medium)	0.2708	0.2129	0.0648	0.3148	0.3417

According to the results, Thonburian Whisper (Large) was the best performer. Due to the lack of students achieving CEFR level B2, we do not include it.

3. Experiment

The experiments were conducted with 12 Grade 12 native Thai speaker students with CEFR level A1 to C1. Students first start with a pre-test. Then, they can practice speaking in the system. Finally, they have to do a post-test to evaluate differences after using the system.

The learning process in the system starts with a pre-test to determine which CEFR level they are at. The test consists of 4 sentences, one for each CEFR level from A1 to B2. After the pre-test, they can practice speaking with sentences in the system. To level up a level, they have to speak all sentences at the level correctly.

The record from pre-test is used to create the new Thai-English pronunciation corpus. This is used for our ASR models' evaluation to select the model with the lowest word error rate for the practice process. The corpus (48 sounds) is divided into 4 datasets by the CEFR level of the sentence; each has 12 sounds.

4. Result

The nine students (75%) improved after using our system. Most of the students in the level between A1 to B2 can speak up to level B2 words. However, none of the students in A1 level pass the C1 level sentences.

We also found that the part of the sentence that students struggle with most is with the word from the 11 consonant sounds from previous study. In addition, present and past form of the verb is also problematic especially at pronunciation at the "s" and "ed". We found that this is an important part that prevents students from leveling further.

The evaluation result of ASR model with the corpus collected in the pre-test is the model with the lowest word error rate (WER) is Thonburian Whisper Large (16.99%). This is an improvement over the base Whisper model (25.88%) and Google Speech-to-Text API (42.87%). Thus, this model is selected for the usage in the practice and post-test part for students. Furthermore, the result with other models suggests that by incorporating Thai ASR model, the WER is lower than the base model and Google Speech-to-Text API.

5. Conclusion, Discussion, and Future Works

In this paper, we extend a personalized learning platform to improve English pronunciation from the word level to the sentence level by using Thai ASR model to detect mispronounced sounds and give correct pronunciation. Our study aims at sentences that use word with the 11 consonant sounds that Thai EFL learners have problems with (see Kritsuthikul et al., 2024)

Based on the experimental results, it can be concluded that using ASR to detect mispronounced sounds has the potential to improve pronunciation skills. Moreover, we also found that samples have problems with the present and past form of the word used in sentences the most. In addition, using ASR model fine-tuned to Thai language help improve speech recognition of sentence with some mispronunciation.

Since the WER of ASR models has been improved significantly with pre-trained Thai ASR models, the model size is still big. We intend to employ smaller ASR models that keep the low WER. We also intend to find more methods to provide learners with more insight into mispronunciation. Furthermore, samples and the amount of variation of sentences used in the experiment are still too low to provide reliable results. Therefore, in further research, we want to apply more samples and sentences to extend experimental data for reliable results. The corpus also needs to expand to provide proper accuracy.

Acknowledgements

We would like to thank all the 12 participants for participating in our experiment.

References

Aung, Z. H., Thavornmongkol, T., Boribalburephan, A., Tangsriworakan, V., Pipatsrisawat, K., & Achakulvisut, T. (2024). Thonburian Whisper: Robust Fine-tuned and Distilled Whisper for Thai. In M. Abbas & A. A. Freihat (Eds.), *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)* (pp. 149–156). Association for Computational Linguistics. https://aclanthology.org/2024.icnlsp-1.17/

Ballier, N., Meli, A., Amand, M., & Yunès, J.-B. (2023). Using Whisper LLM for Automatic Phonetic Diagnosis of L2 Speech, a Case Study with French Learners of English. In M. Abbas & A. A. Freihat (Eds.), *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)* (pp. 282–292). Association for Computational Linguistics. https://aclanthology.org/2023.icnlsp-1.30/

Dillon, T., & Wells, D. (2023). Effects of Pronunciation Training Using Automatic Speech Recognition on Pronunciation Accuracy of Korean English Language Learners. *English Teaching*, *78*(1), 3–23. https://doi.org/10.15858/engtea.78.1.202303.3

Evers, K., & and Chen, S. (2022). Effects of an automatic speech recognition system with peer feedback on pronunciation instruction for adults. *Computer Assisted Language Learning*, *35*(8), 1869–1889. https://doi.org/10.1080/09588221.2020.1839504

Kritsuthikul, N., Boonma, K., Muangprathub, J., Na Chai, W., & Supnithi, T. (2024). Improve English Pronunciation at Word Level for Thai EFL Learners in Southern Region Using End-to-End Automatic Speech Recognition. *International Conference on Computers in Education*. https://doi.org/10.58459/icce.2024.4917

McCrocklin, S. (2019). Learners' Feedback Regarding ASR-based Dictation Practice for Pronunciation Learning. *CALICO Journal*, *36*(2), 119–137. https://doi.org/10.1558/cj.34738

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning*, *202*, 28492–28518.

SCB 10X. (2024, August 13). *Monsoon - a scb10x Collection*. https://huggingface.co/collections/scb10x/monsoon-66bb68f00198033f32980bb6

Sharma, S., Mhasakar, M., Mehra, A., Venaik, U., Singhal, U., Kumar, D., & Mittal, K. (2024). Comuniqa: Exploring Large Language Models For Improving English Speaking Skills. *Proceedings of the 7th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, 256–267. https://doi.org/10.1145/3674829.3675082

Sun, W. (2023). The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: A mixed methods investigation. *Frontiers in Psychology*, *14*, 1210187. https://doi.org/10.3389/fpsyg.2023.1210187

Tipaksorn, P., Sommuang, W., Chatthong, O., & Thangthai, K. (2024). *Pathumma Whisper Large V3 (TH)*. Hugging Face. https://huggingface.co/nectec/Pathumma-whisper-th-large-v3