APP_NAME - A mobile app for practising Finnish pronunciation

Anonymous Author
Affiliation
email@domain

Anonymous Author
Affiliation
email@domain

Anonymous Author
Affiliation
email@domain

Abstract

Learning a new language is often difficult, especially practising it independently. The main issue with self-study is the absence of accurate feedback from a teacher, which would enable students to learn unfamiliar languages. In recent years, with advances in Artificial Intelligence and Automatic Speech Recognition, it has become possible to build applications that can provide valuable feedback on the users' pronunciation. In this paper, we introduce the APP_NAME[1] app explicitly developed to aid students in practising their Finnish pronunciation on handheld devices. Our app is a valuable resource for immigrants who are busy with school or work, and it helps them integrate faster into society. Furthermore, by providing this service for L2 speakers and collecting their data, we can continuously improve our system and provide better aid in the future.

1 Introduction

Proper pronunciation is needed to build confidence in second language (L2) learners and is essential for effective communication and language acquisition (Gilakjani, 2012). L2 adult learners, who might not have regular exposure to the target language during their everyday life, may lack sufficient opportunities to practise and receive corrective feedback.

With recent advances in Automatic Speech Recognition (ASR) technologies, computer-assisted pronunciation training (CAPT) apps have become more and more effective in helping L2 learners. These apps can immediately give the users feedback on their pronunciation at their convenience. However, while popular languages such as English have many pronunciation applications (Kholis, 2021; Fouz-González, 2020; Wellocution, 2023), there are fewer resources available for Finnish L2 learners. To the best of our knowledge, there was no similar app for CAPT in Finnish before this work.

The main challenge in developing CAPT applications for Finnish and other low-resource languages is the lack of data from L2 speakers. Furthermore, if the L2 corpus is not annotated at the phoneme level, it makes developing an app for mispronunciation detection (MD) more complicated. We designed our APP_NAME app to function as well as possible using all available data and add the possibility of collecting users' data after the pilot phase (figure 1). Such information will help evaluate the app's effectiveness for language training and improve our model's performance to better address students' needs in later versions.

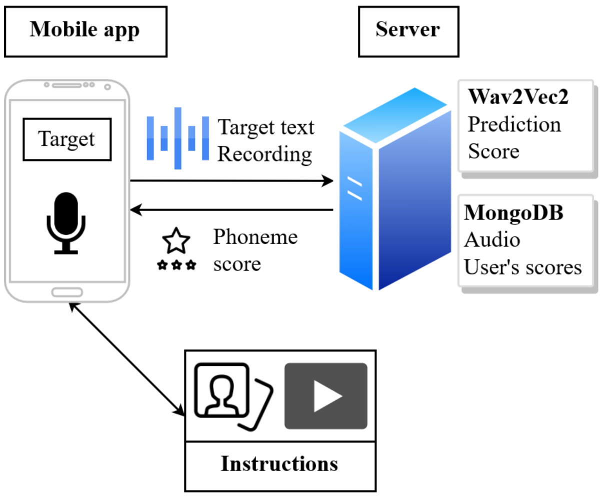

Figure 1: APP_NAME app processing flowchart

Recent works from Wu et al. (2021) and Xu et al. (2021) have demonstrated the effectiveness of end-to-end systems with Transformer-based architectures for English MD. While we focus more

[1] We hide the app name for anonymous reason

on practicality, we use a similar approach without a detailed annotation dataset for Finnish.

2 Dataset

One of the major challenges that we needed to overcome was the limited data at our disposal. We should note that for the English language, several datasets are available with phoneme level annotation (Zhao et al., 2018; Zhang et al., 2021; Weinberger, 2015). Unfortunately, no such public Finnish resources exist. Thus we opted to use the data collected during the Digitala project (Al-Ghezi et al., 2023) as our primary corpus. This dataset includes ratings from language experts on pronunciation, fluency, lexical, grammatical and the holistic overall level for each audio file, but it does not have phoneme level information.

The Digitala corpus consists of free-form and read-aloud speech, from which we selected 768 short read-aloud samples as those matched our intended scenario most closely. This gave us approximately 60 minutes of audio with the overall pronunciation ratings ranging from 1 to 4, with 4 being the best. The rating is for the whole pronunciation task and not individual phonemes. The lowest pronunciation level (1) contains approximately 2,200 phonemes, the highest one (4) has only 576 phonemes, while the remaining 14,000 phonemes are split almost equally between levels 2 and 3. The corpus was also transcribed by third parties who were not language experts.

The small size of the Digitala corpus and the lack of phoneme annotation meant it was not suitable for training or finetuning for the MD task. However, as there were no better alternatives, we used the Digitala read-aloud transcript as a replacement for the evaluation set. Consequently, we needed another dataset to train our models. After some preliminary experiments, we selected the Finnish Parliament corpus (Kielipankki, 2022), a publicly available corpus without any statistically significant use of dialects (Virkkunen et al., 2022). By training our models for the ASR task with suitably chosen native speakers' samples, we expected the models could learn the features of native Finnish speech and have the potential to identify deviations made by L2 speakers. As a first step, we filtered the most suitable portion of the data, by selecting speeches with low or average speaking rates (which is the most similar to how L2 learners speak). As an additional step, we also restricted the data by excluding older (50+) speakers, since our target audience is generally younger immigrants. The last step in data preparation was the splitting of the 281 hours of data into 75% for training, and 25% for tuning hyperparameters and evaluating the speech recognition models. We should note that we also used two publicly available reference models, called Finnish-NLP[2] and Finnish-NLP-S [3]. Both have been trained with 228 hours of Finnish Parliament data and approximately 47 hours of data from other sources.

3 Implementation

3.1 Server

The core technology inside our server is based on wav2vec 2.0 (Baevski et al., 2020), which was already proven to work exceptionally well even with very limited amount of data (Wu et al., 2021; Xu et al., 2021). We selected XLS-R (Babu et al., 2021) and Uralic, a subset of VoxPopuli (Wang et al., 2021), as our pre-train models, and use the state-of-the-art model in Finnsh ASR, Finnish-NLP, as our baseline. Except for entropy β, all models used the same hyperparameters, and there is no language model used for decoding.

Leveraging the phonetic nature of the Finnish language, where each phoneme is represented by exactly one grapheme[4], we can use graphemes as output units during the ASR training procedure. Once the ASR models were trained, we used the forced alignment algorithm for Connectionist Temporal Classification (CTC) from Kürzinger et al. (2020) to determine the success of pronunciation. This algorithm provides both time alignment and a probability score for each grapheme. Inspired by the traditional Goodness of Pronunciation method (Witt and Young, 2000), we use such information to generate feedback for the user.

One major issue we had to overcome was the overconfidence of the wav2vec 2.0 models. As it is well known, the CTC algorithm often results in spiky outputs (Zeyer et al., 2021), which in terms would mean that we can only provide binary (correct/incorrect) feedback to the user. Naturally, a good pronunciation training app should give more detailed information (Engwall and Bälter, 2007),

[2] https://huggingface.co/Finnish-NLP/wav2vec2-xlsr-1b-finnish-lm-v2

[3] https://huggingface.co/Finnish-NLP/wav2vec2-xlsr-300m-finnish-lm

[4] except "nk" [ŋk] and "ng" [ŋː]

Model	Vocabulary	Parameters	Entropy β	CER	Recall	Precision	F_1
Finnish-NLP	Grapheme	1bil	0%	**15.4%**	59.8%	**33.3%**	**42.8%**
Finnish-NLP-S		300mil		22.3%	**65.0%**	26.1%	37.2%
XLS-R	Grapheme	300mil	0%	20.9%	61.1%	26.7%	37.2%
XLS-R-5	Grapheme		5%	**19.5%**	63.1%	**30.0%**	**40.6%**
XLS-R-10	Grapheme		10%	21.2%	63.1%	29.4%	40.1%
XLS-R-10-P	Phoneme		10%	21.3%	63.2%	27.3%	38.1%
Uralic-10	Grapheme		10%	30.4%	64.3%	23.4%	34.3%
Uralic-10-P	Phoneme		10%	29.6%	**66.8%**	22.6%	33.8%

Table 1: Speech models' performance in ASR and MD on Digitala read-aloud set.

thus, reducing the peakedness of the outputs was important. To achieve this, we chose the negative maximum entropy regularization technique Liu et al. (2018) during training, which redistributes β% of the total probability mass uniformly to all outputs, ensuring the smoothness of the final predictions.

3.2 Mobile app

We use Unity (Juliani et al., 2018) as our development engine. With Unity we can simultaneously publish our APP_NAME app to multiple platforms: Android, iOS and Windows. Our app contains various study materials, and Unity Editor allows us to easily integrate those multimedia content into the app. We make use of the engine to visualize our pronunciation instructions with animations and limit the rest to simple UI, thus lowering the application's power consumption.

Arapakis et al. (2021) estimated a 7 seconds threshold where mobile (web search) users' experience decreases significantly. To maintain a reasonable response time, we use a manual VAD system to remove the silent parts from the recording: the users must press and hold the record button to record their audio samples.

The app supports two modes; the "Topic" mode supplies curated words and phrases for various topics, often along with English translation and audio samples from native speakers. On the other hand, the "Freestyle" mode enables users to practice any word or phrase by first prompting for the text that the user will attempt to pronounce.

The score for each phoneme is saved locally, enabling users to track their progress. The data is valuable in developing speech applications for L2 speakers. In the future, with the users' permission, we can collect their records to evaluate the app's effectiveness and other metadata.

APP_NAME also provides pronunciation instructions via sample audios, pictures, animations and videos, which are beneficial for users during self-practice (Engwall and Bälter, 2007). The audio, photo and animation materials are directly stored in the app, while the videos are accessible via a public, ad-free platform. We should note that external links would generally have an adverse effect on user experience, still we choose this solution to supply high-quality tutorial videos while keeping the size of the app reasonably small.

4 Results

To validate our models, we computed their character error rate (CER), Recall (percentage of mispronunciations correctly detected) and Precision (the ratio of detected mispronunciations actually being mispronunciation, according to a native Finnish listener) using the Digitala read-aloud corpus. The empirical results can be seen in Table 1. The first thing that we noticed is that the large Finnish-NLP produced significantly lower and the small Finnish-NLP-S higher CER compared to the majority of our models. Next, we compared the models in terms of MD and saw that Finnish-NLP yielded the highest overall F_1 score. However, the smaller XLS-R-5 and XLS-R-10 managed to achieve comparable results with the help of entropy regularization.

The benefit of entropy regularization is seen when we increase the value of β and note that both Recall and Precision also increase. From our experiment, we found that β between 5% and 10% produces the best result for MD task. Looking at the detailed breakdown in table 2, we also found that, the smaller XLS-R outperformed the Finnish-NLP in Recall for pronunciation level 1 samples, while slightly falling behind in Precision. The gap in Precision widens as the speakers' pronuncia-

Model	CER	Recall	Precision
Finnish-NLP	**26.9%**	72.6%	**38.7%**
XLS-R-5	31.4%	77.4%	36.2%
XLS-R-10	33.5%	**78.5%**	36.8%
Finnish-NLP	**20.0%**	61.5%	**32.7%**
XLS-R-5	24.1%	**63.3%**	29.1%
XLS-R-10	24.7%	63.2%	28.9%
Finnish-NLP	**11.6%**	42.4%	**27.2%**
XLS-R-5	15.4%	**46.7%**	24.1%
XLS-R-10	17.6%	45.7%	22.2%
Finnish-NLP	**6.0%**	18.8%	**20.0%**
XLS-R-5	10.3%	**25.0%**	16.0%
XLS-R-10	13.6%	**25.0%**	10.0%

Table 2: CER, Recall and Precision for the pronunciation levels 1 to 4 (top to bottom: worst to best)

tion skill improves. Considering the practicality of smaller models, they would be suitable in MD for beginner L2 learners. While the Uralic model did, in our preliminary experiment on Common Voice 7.0 test set, produce lower CER on native Finnish speakers, it failed in both ASR and MD task on L2 speakers. One possible reason is that the Uralic models were not exposed to foreign language families, unlike the XLS-R models.

While it is possible to use the training part of the Digitala corpus for finetuning our wav2vec 2.0 models, we could not control the pronunciation quality, as the speakers are L2 learners and there is no phoneme annotation. In our preliminary experiments we found that finetuning with bad pronunciation data led to lower performance in MD.

5 Self-study assistant

APP_NAME (see figure 1) allows users to enter words into a text prompt to practise pronunciation. Their audio is sent to the server, and the device will display the obtained rating for each phoneme, with three possible ratings in colors (figure 2): flawed (phoneme is not recognizable), almost correct (improved, but not clear), and correct. The "almost correct" rating is given as positive feedback when user's phoneme score improves, but is still not considered correct. The users are also advised to refer to the app multimedia pronunciation instructions (figure 3).

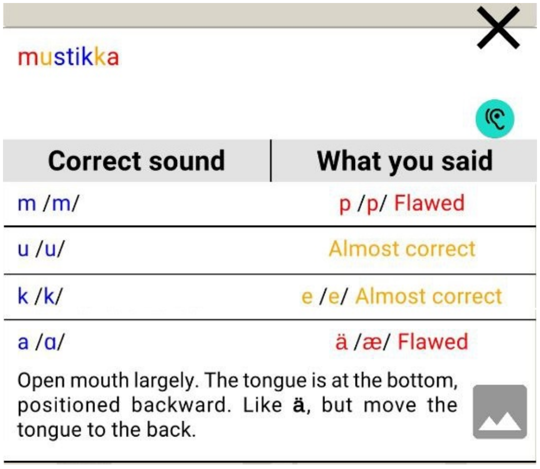

Figure 2: The result is coloured based on pronunciation score.

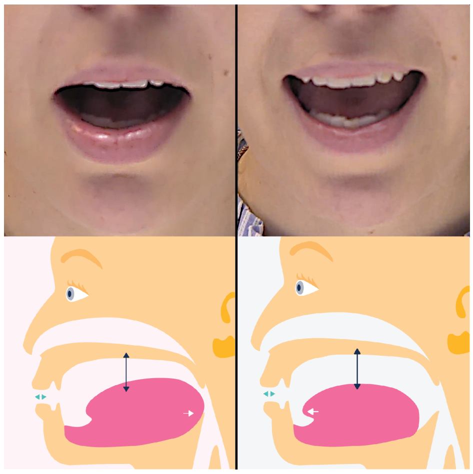

Figure 3: Visual pronunciation instructions for A [ɑ] (left) and Ä [æ] (right).

6 Conclusion

In this paper, we presented the prototype of APP_NAME, an app that helps language learners practise Finnish pronunciation. Because of the lack of data available for phoneme level pronunciation mistakes, our solution is based on multilingual wav2vec 2.0 models, which are finetuned for native Finnish ASR. By running the L2 learners' utterances through the ASR without a language model, we predict pronunciation errors and probability scores that indicate the success of pronunciation. The resulting models are validated by measuring CER, Recall and Precision for samples of different levels of pronunciation judged by human experts. In the future, we plan to collect user data (feedback and audio) with our app to update the models and improve the self-study application.

References

Ragheb Al-Ghezi, Yaroslav Getman, Ekaterina Voskoboinik, Mittul Singh, and Mikko Kurimo. 2023. Automatic rating of spontaneous speech for low-resource languages. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 339–345. IEEE.

Ioannis Arapakis, Souneil Park, and Martin Pielot. 2021. Impact of response latency on user behaviour in mobile web search. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 279–283.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. http://arxiv.org/abs/2006.11477 wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

Olov Engwall and Olle Bälter. 2007. Pronunciation feedback from real and virtual language teachers. *Computer Assisted Language Learning*, 20(3):235–262.

Jonas Fouz-González. 2020. Using apps for pronunciation training: An empirical evaluation of the english file pronunciation app. *Language, Learning and Technology*, 24.

Abbas Pourhosein Gilakjani. 2012. A study of factors affecting efl learners' english pronunciation learning and the strategies for instruction. *International journal of humanities and social science*, 2(3):119–128.

Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, et al. 2018. Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627*.

Adhan Kholis. 2021. Elsa speak app: automatic speech recognition (asr) for supplementing english pronunciation skills. *Pedagogy: Journal of English Language Teaching*, 9(1):01–14.

Kielipankki. 2022. http://urn.fi/urn:nbn:fi:lb-2022052002 Aalto Finnish Parliament ASR Corpus 2008-2020, version 2.

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. Ctc-segmentation of large corpora for german end-to-end speech recognition. In *International Conference on Speech and Computer*, pages 267–278. Springer.

Hu Liu, Sheng Jin, and Changshui Zhang. 2018. Connectionist temporal classification with maximum entropy regularization. *Advances in Neural Information Processing Systems*, 31.

Anja Virkkunen, Aku Rouhe, Nhan Phan, and Mikko Kurimo. 2022. Finnish parliament asr corpus-analysis, benchmarks and statistics. *arXiv preprint arXiv:2203.14876*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Steven Weinberger. 2015. http://accent.gmu.edu *Speech Accent Archive*. George Mason University.

Wellocution. 2023. https://www.boldvoice.com/ Boldvoice web page.

Silke M Witt and Steve J Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, 30(2-3):95–108.

Minglin Wu, Kun Li, Wai-Kim Leung, and Helen Meng. 2021. Transformer based end-to-end mispronunciation detection and diagnosis. In *Interspeech*, pages 3954–3958.

Xiaoshuo Xu, Yueteng Kang, Songjun Cao, Binghuai Lin, and Long Ma. 2021. Explore wav2vec 2.0 for mispronunciation detection. In *Interspeech*, pages 4428–4432.

Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2021. Why does ctc result in peaky behavior? *arXiv preprint arXiv:2105.14849*.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. https://doi.org/10.48550/ARXIV.2104.01378 speechocean762: An open-source non-native english speech corpus for pronunciation assessment.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-arctic: A non-native english speech corpus. In *Interspeech*, pages 2783–2787.