9th Workshop on Speech and Language Technology in Education (SLaTE)
18-20 August 2023, Dublin, Ireland

A Pronunciation Scoring System Embedded into Children's Foreign Language Learning Games with Experimental Verification of Learning Benefits

Reima Karhila[1,2]*, Sari Ylinen*[3]*, Anna-Riikka Smolander*[3]*, Aku Rouhe*[2]*, Ragheb Al-Ghezi*[2]*, Yaroslav Getman*[2]*, Tamas Grosz*[2]*, Maria Uther*[4]*, Mikko Kurimo*[2]

[1] Silo.AI, Finland, [2]Dpt. of Information and Communications Engineering, Aalto University, Finland
[3]Logopedics, Faculty of Social Sciences, Tampere University, Finland
[4] Birmingham City University, United Kingdom, maria.uther@bcu.ac.uk

reima.karhila@silo.ai, sari.p.ylinen@tuni.fi, mikko.kurimo@aalto.fi

Abstract

Over the years, language technology has become a valuable asset for foreign language learners. In this work, we introduce pronunciation feedback scoring systems for 6-12 year old children. The scoring systems were embedded in second-language (L2) English learning games that were designed to prompt children to repeat words. Speech and phone recognition models were used to validate utterances and extract phoneme-wise statistics, which were used to compute feedback scores of 0-5 stars. The scoring systems were trained to mimic the preferences of a single expert who evaluated all the training data. Our automatic scoring system reached a correlation of 0.59 to the human annotation. This system was also tested in a learning experiment, where EEG measurements indicated that children who played our learning game with our scoring engine for pronunciation feedback improved their perception of speech sounds. We release the game codes and the speech data used to train the scoring system.

Index Terms: CALL, CAPT Learning games, Pronunciation assessment; DNN; bilingual models.

1. Introduction

Carefully designed computer games have great potential to help children in foreign or second-language (L2) learning, including learning new words and practicing the pronunciation of new speech sounds with sufficient repetitions. The aim of the Say It Again, Kid! (SIAK) project was to investigate the effects of learning games and automatic pronunciation feedback on learning English as a foreign language [1, 2, 3]. We studied these effects by comparing learning in the game providing feedback and in a non-game condition without feedback but with the same amount of pronunciation training as in the game. The hypothesis was that using an engaging learning game with automatic feedback on pronunciation quality facilitates learning. As the study group consisted of children who were just beginning to learn English, and were only starting to grasp the pronunciation of new speech sounds, evaluating their pronunciation before and after the gaming period was not considered realistic. Instead, we measured the development of representations for new speech sounds in the brain.

Computer-assisted pronunciation training (CAPT) with language learning games requires fast and sufficiently accurate feedback with a system that does not require adaptation for a new player. The accuracy for each individual utterance is not as critical as in language skills assessment software. As long as there is enough consistency in the feedback, it is more important that the latency for feedback does not deteriorate the game experience. Our work concentrates on providing 1) an automatic scoring mechanism suitable for beginner and intermediate learners of a new language and 2) learning games suitable for regular sessions where the amount of gameplay and nature of tasks could be controlled for learning experiments.

This paper gives an overview of the data we collected, pronunciation scoring systems we built and describes experiments in validating their performance in real world learning experiments with children.

2. Related work

CAPT systems use computational measures to analyse speech segments in order to give a score or meaningful feedback about single mispronunciations or mispronunciation patterns. In a practical use case, the feedback is either immediate, or given as an analysis to the user or the user's teacher.

The majority of the early generation of CAPT systems were based on deducing some goodness score from likelihood scores derived from the ASR acoustic model set [4], but these reached their limit when the differences between correct and false pronunciations are subtler than the traditional ASR features could capture [5]. Several proposed systems first use ASR methods for creating a forced alignment of a canonical transcription of phonetic segments, and analyse these segments individually, allowing the use of different parameterisation of speech for each step [6, 7]. Often the alignment is done with extended recognition networks, that take into account typical pronunciation mistake patterns [8]. Several proposed systems use Recurrent Neural Networks (RNN) for predicting mispronunciations from features that represent articulatory or phonological properties of individual phonetic segments [9, 10, 11]. An upper limit for the performance of these *mispronunciation detection systems* is the level of disagreement between human evaluators. The disagreement has been reported to be around 20% of all the human labelled mispronunciations [12].

For longer evaluation samples, from 3 minutes of speech onwards, computational systems to *evaluate spoken L2 language skills* have been found to correlate well with human scoring. Many systems use ASR for extracting contents of semi-spontaneous utterances and compute features like vocabulary size, number of pauses and phone posterior statistics and extract a score from these [13, 14]. Sometimes scoring based on pure phonological control may be desirable. In constrained tasks like reading prompts, phone posteriors alone can be used to extract a score for a speaker [15, 16]. The mapping from statistics to a score is done with neural networks, Support Vector Regressors (SVR), Gaussian Processes or even just linear regression. For shadowing tasks, where the user is repeating utterances after a model voice prompt, Dynamic Time Warping (DTW) path cost of DNN outputs has correlated well with human evaluators' scores [17].

10.21437/SLaTE.2023-5

Again the upper limit for the performance of a CAPT scoring system is the human annotators' disagreement. For shorter speech segments, the correlation between human evaluators - the inter-annotator agreement - starts to degrade, for example from a correlation of 0.9 for speaker evaluation based on the complete speech pool, to 0.6-0.7 for single items (single utterance or a collection of utterances consisting of a reply to a single prompt) [13, 18, 19]. For evaluation of samples that are shorter than a minute, let alone for items as short as a single word of one or two phonemes, the inter-annotator agreement is a limiting factor for the performance of the computational grading system that aims for widely accepted objective scoring. The scoring presented in this paper is computed from utterances as short as two phones, and we approach this more as a mispronunciation detection task, where different types of phonological alterations from canonical form affect the score of the utterance.

CAPT learning applications that give *immediate feedback* on single utterances have been built for L2 learners. [20] used ASR output as feedback on correct pronunciation for adults. [21] built a practise system for children with pronunciation difficulties in their native languages, where children get a correct/incorrect feedback from an ASR system. For children learning foreign languages, CAPT systems have been embedded into games. In [22] English pronunciation of short utterances is scored based on tone, speed, volume and timbre [23].

Speech technology is generally more difficult to build *for children*. Children's data are harder to acquire, and the variance of children's speaking styles is large. Several papers describe different methods and platforms to improve speech assessment in children. [24] proposes a child speech disorder detection system that uses a Siamese recurrent network trained with normal speech to detect speech sound disorder. The system measures the similarity and discrepancy of pronunciations and incorporates speech attribute features to provide diagnostic feedback. [25] presents a child speech verification platform designed to identify keywords and phrases in children's speech with high accuracy, even in noisy environments. [26] developed an interactive mobile application to develop primary students' initial skills in Sri Lanka. For child ASR, tuning models trained with adult data with a small amount of child audio has proven successful [27]. Models that are pretrained in a self-supervised manner have been finetuned with small amounts of child data for L2 learning and for children with speech sound disorder (SSD) [28, 29]. [30] found that automatic pronunciation feedback in a game improved pronounciation and engagement in speech therapy and outperformed offline mispronunciation detection tests. Finally, [31] measured effectiveness of automatic pronunciation feedback from children before and after training with a simple ASR-based CAPT system, and found the CAPT system was beneficial. The study was based on subjective evaluation of quality of pronunciation of complete words spoken before and after the training period. In our study, we are looking for more objective evidence of long-term representations of previously unfamiliar speech sounds forming in the brain.

3. Data collection

Speech data were collected to train the pronunciation scoring system used in the learning experiments. Children's native-language (L1) and L2 English speech were collected with a simple recording program, a prototype of the SIAK game and a dedicated data collection game. L1 English speech samples were collected from 24 UK English native speakers between 6 and 12 years of age through the data collection game. The native English child participants were recruited from Surrey and Hampshire areas of England and had a Southern British accent. Language learners' utterances were collected from 148 children between 5 and 12 years of age living in Helsinki metropolitan area. Of these 130 had Finnish as native language, the rest were bilingual with Finnish and another language as their first language. 17969 utterances were collected. They were validated and scored on a subjective 0-100 point scale by a single annotator using a web interface. The annotator was a native Finnish speaker, who had experience in teaching English to Finnish primary school children. As the annotator was not a native speaker, the scoring is not intended to represent objective quality of English pronunciation. It reflects the opinion of a single individual for clarity of pronunciation that Finnish children are expected to develop after several months or years of English lessons in primary school. The downside of using a single annotator is that there is no simple way of finding annotation errors, and there are some inconsistencies in the annotations. Of the data, 1489 items were labelled as rejected (silence, interrupted, wrong word, spoken noise, lack of effort). Of the rest, 10 % were randomly selected as development set and 10 % as test set, the rest made up the scoring system training set. Each speaker only appeared in one of the sets. For training the scoring systems, the 0-100 discrete scale was mapped to a discrete 0-5 score. The anonymized annotated dataset can be downloaded for research purposes at `https://huggingface.co/datasets/rkarhila/SIAK`.

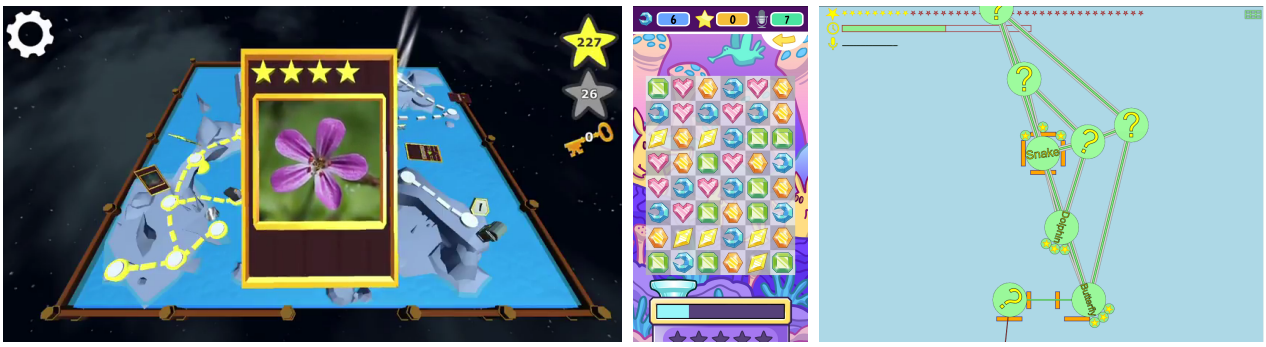
4. Game clients

We created three CAPT games. The first SIAK game, shown in Figure 1a, was an exploratory board game aimed at 9-12 year olds. The second game Pop2Talk shown in Figure 1b, had a simpler game mechanic aimed at younger children of 6-8, and had a better control over repetitions of stimuli and player tasks. Both research games were developed with Unity and compiled for Android and Windows platforms, consisting of levels with a number of challenges. In challenges, players hear an L1 translation and L2 example pronunciation of a word or phrase, and try to reproduce the L2 utterance as accurately as possible. The players then immediately receives a score in the form of 0-5 stars, reflecting the goodness of their pronunciation attempt. At the end of each level there is a special challenge, where the player will only hear the L1 translation, and needs to produce the L2 word without a model pronunciation.

The third game, a browser-based physics game Fysiak , shown in Figure 1c, was a small physics puzzle game that ran entirely on JavaScript in a modern desktop web browser, and was used to create a game-like athmosphere for data collection. All game clients are open source and available for research use at `https://github.com/rkarhila/siak-game-clients`.

5. Scoring models

We developed three scoring models used in different versions of the learning game. The accuracy of all scoring models as correlation to human reference scoring is summarised in Table 1. Training data for the systems consisted of read speech corpora with clear pronunciations. For English we used WSJCAM0 and PF-Starcorpora and for Finnish both adult and child parts of the Finnish SPEECON corpus. English word to phone mappings were from CMU dictionary and UK English CombiLex dictionary. Finnish mapping was rule-based. Each model uses a



(a) SIAK (b) Pop2talk (c) Fysiak

Figure 1: *Game clients*

Table 1: *Correlation betweeen prediction and human scoring for the collected test set.*

Model	Correlation
HMM-GMM + phone classifier + SVM	0.59
CTC + articulatory features + SVM	0.59
CTC + PWLD + SVM	0.61
CTC + PWLD Linear regression	0.54

simple energy-based voice activity detection based on Pocket Sphinx [32] to first validate presence of speech.

5.1. Baseline scoring system

The baseline in our experiments in validation and scoring utterances is the first generation of the SIAK game powered by an HMM-GMM forced alignment segmenter and a bilingual phonetic segment classifier recurrent neural network (RNN), used in demonstrations [1] and data collection. The hidden Markov model with Gaussian mixture model (HMM-GMM) segmenter was trained with native English data augmented with background noises. The segmenter produced alignments and likelihoods when it found a phone alignment of the target word via Viterbi search. Validating utterances by tuning the search beam or filtering by output likelihood did not work reliably, so all utterances were passed to the next component.

Each segment was separately fed to a phone classifier, a bidirectional long short term memory (LSTM) network four layers deep, with a width of 1000, 750, 500 and 250 units, trained on 36 dimensional Mel-spectral bins extracted from noise-augmented English and Finnish data. A support vector machine (SVM) regressor computed the pronunciation score based on the difference between reference and hypothesis phoneme sequence. The SVM training was augmented by adding artificial negative samples created by giving wrong prompts to training utterances. We reached a correlation of 0.59 between reference and model scoring on the scoring test set.

The pipeline consisted of Kaldi, Tensorflow and Scipy, and without proper integration, the delay between speaking and receiving score interrupted gameplay.

5.2. CTC + SVM system

In the second generation system, the segmenter and phone classifier are replaced by a connectionist temporal classification (CTC) phoneme/speech event recogniser multi-task trained to do framewise articulatory feature prediction. The network was small at 300 units wide and three layers deep to ensure light computational load. The incoming audio was decoded with two CTC outputs: A phone-based model and a speech event/landmark-based system. The speech event outputs were four broad phonetic classes adapted from [33], namely 1) vowel, 2) stop consonant, 3) fricative and 4) semi-vowels and nasals. The input was validated by passing under a preset error rate threshold with either the phone or the event output.

Scoring was based on outputs of the phonological feature predictor. The phonological features for UK English were adapted from [34], and appended with several features describing vowel and diphthong pronunciation more accurately. The phonological features were extracted for each phone based on boundaries found by Manhattan distance dynamic time warping (DTW) to the ideal phonological feature vectors of the target utterance. An SVM is used for mapping the feature difference to a score, and the same 0.59 correlation is reached but with a much faster, simpler pipeline that generally returned a score in 0.2 s to the player. This second generation system fast enough to use for the learning experiment.

5.3. CTC + PWLD system

In the third generation scoring model the articulatory feature DTW is replaced by phonetically weighted Levenshtein distance (PWLD) described in [35] . The more simplified pipeline allowed for a slightly larger, 3 layer 600 wide GRU for phonetic CTC, and the scoring mechanism attained a correlation of 0.61 with an SVM and 0.54 with a linear regression method. The linear regression method was developed to produce more stable and more interpretable outputs for the scoring. In internal playtests it was indeed found to perform more robustly than the SVM scorer and was chosen as the method to use for the next experiments.

6. Learning experiments and results

The scoring and the gaming approach to learning were validated in learning experiments with children [2, 36] . In addition, we studied children's user experience and affective ratings for the game app [3] . In [2] the participants were 37 native Finnish speaking children with no developmental, language, or learning disorders, whereas in [36] , the participants were 24 children with dyslexia and 24 control children with typical reading skills.

In both studies, the children were 7-11 years old, and they attended the Finnish comprehensive school. None of the children could understand or speak English fluently.

The participants practiced English words with the SIAK research game described in Section 2.2 with the second-generation scoring model. The game version had 24 levels with game features. In addition, three non-game like levels were embedded among the game levels, resulting in 27 levels in total. The three non-game levels included an identical speech listening and production task as the game levels, but they had no game elements, such as stars or other kind of feedback, or visual game features (these levels had just a black arrow on a white background to move forward). All children played all 24 game levels and three non-game levels. The game was played on either Windows laptops or Android tablets using a headset microphone. Gameplay was supervised by researchers and gaming equipment was provided by the project. The gaming period lasted on average 4.3 weeks. The children played the SIAK game on average 15.5 min a day 2.9 days a week.

To quantify the effects of game-based language learning with pronunciation feedback in children as objectively as possible, we used electroencephalography (EEG) to measure children's brain responses (the mismatch negativity aka MMN; see [2] for details) that reflect the establishment of representations for new speech sounds in the brain. Such neural representations are prerequisites for both speech perception and production, yet perceptual learning is likely to precede progress in production. A benefit of our approach is that the brain responses we measured are elicited in the brain automatically regardless of children's effort or their attention allocation, which makes our method particularly suitable for children and more objective than the analysis of children's behavioral performance. We compared the activation of brain representations for L2 English speech sounds /θ/ and /ð/. To counterbalance the speech sounds, a half of the children learned words with /θ/ in the game and /ð/ in the non-game, and vice versa. The amount of exposure was kept equal for these sounds in each participant (for details in setup and equipment, see [2]). The statistical results of linear mixed model analysis showed that in typically developing children the brain responses had significantly increased after training the speech sounds with the game, but there was no statistically significant change after training with the non-game in the same children (see [2] for figures, amplitude values and statistics). An opposite pattern was found in children with dyslexia [36]. The statistical results of linear mixed model analysis showed that the brain responses had significantly increased after training the speech sounds with the game (from -0.32 to -0.99 μV), but there was no statistically significant change after training with the non-game in the same children (from -0.25 to -0.30 μV, see [35] for figures and statistics).

Regarding user experience, children expressed higher affective ratings for the game compared to non-game version of the application [3]. By combining the data in [2] and [3] for the current work and by analyzing their relationship with Pearson's r, we found that children's ratings for game engagement (liking the game and finding it easy) correlated positively with the degree of brain response change in [2] (r=0.38, p=0.021).

7. Discussion

To summarize the project outcome, we have shown the benefits of using speech technology in controlled, regular but short amounts of gamified learning for improving phonological representations in the brain.

The hypothesis of the SIAK project was that automatic pronunciation evaluation embedded in a game would improve learning results. To make sure that children did not just learn to use the scoring engine better, the learning effects were measured with EEG.The experimental setup and analysis were complicated, and the results for experiments that started in 2017 were finally published in 2022 and 2023, giving the evidence that the scoring system and the exploratory style game were both good enough to yield measureable learning benefits.

We did verify that children learn better with a game and automatic feedback than by using the time on control condition, namely, non-game with no feedback. What remains, however, is to factor this into the learning benefits brought by the game and benefits brought by automatic evaluation. Our results also suggest that user experience affects game-based learning: brain responses changed more as a result of gaming in those children with better engagement. This link to measurable changes in the brain highlights the importance of user experience in game-based learning. This is likely due to stronger activation of the reward system of the brain in those children who were more engaged in the game. We were expecting dyslexic children to benefit from text-free foreign language learning, but the results were opposite. Possible accounts for this are distraction by visual game features or atypical processing of rewards in the brain in dyslexia. For dyslexic children, possible follow up study would be to investigate the role of the feedback mechanism in a visually simpler game, or without a game at all.

We found that the SVM based scoring of the second generation system that was used in the learning experiment could not generalise at all. It worked reasonably well for the utterances in the game, but failed for most tests done with out-of-domain utterances. That is why the third generation system switched to a more constrained linear regressor scorer. This scoring did not reflect as well the annotator's preferences, but was more predictable and gave a direct breakdown of the score for any pronunciation errors. At this time we had developed a new game for conducting learning experiments. The second learning research game Pop2Talk game allowed better control of repeating stimuli and the game was easier to play, allowing us to target 6-8 year old children. However the planned experiments were cut short by COVID-19 lockdowns.

We are releasing the source code for our games and the training data used in our pronunciation evaluation systems, and are working on new scoring models to be released in the future.

8. Conclusions

We have described pronunciation feedback scoring methods to use in L2 pronunciation learning games for 6-12 year old children. We found the models to perform well enough to conduct a learning experiment with children learning to pronounce and distinguish speech sounds previously unfamiliar to them. The experiment showed that learning games with automatic scoring bring measurable benefits to learning.

9. Acknowledgments

This work was financially supported by the Academy of Finland grant no. 274075 and 274058, NordForsk TEFLON-project grant no. 103893 and TELLme-project in Tekes Challenge Finland programme. Computational resources were provided by the Aalto Science-IT project. The authors would like to thank the other members of the SIAK-project and Rob Clark for help with data collection, experiments and building the game.

10. References

[1] R. Karhila, S. Ylinen, S. Enarvi, K. Palomäki, A. Nikulin, O. Rantula, V. Viitanen, K. Dhinakaran, A.-R. Smolander, H. Kallio, K. Junttila, M. Uther, P. Hämäläinen, and M. Kurimo, "SIAK - a game for foreign language pronunciation learning," in *Proc. Interspeech*, 2017.

[2] K. Junttila, A.-R. Smolander, R. Karhila, A. Giannakopoulou, M. Uther, M. Kurimo, and S. Ylinen, "Gaming enhances learning-induced plastic changes in the brain," *Brain and Language*, vol. 230, p. 105124, 2022.

[3] M. Uther, A.-R. Smolander, K. Junttila, M. Kurimo, R. Karhila, S. Enarvi, S. Ylinen *et al.*, "User experiences from l2 children using a speech learning application: implications for developing speech training applications for children," *Advances in Human-Computer Interaction*, vol. 2018, 2018.

[4] S. M. Witt *et al.*, *Use of speech recognition in computer-assisted language learning*. University of Cambridge Cambridge, 1999.

[5] H. Strik, K. Truong, F. De Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, no. 10, pp. 845–852, 2009.

[6] A. Lee and J. R. Glass, "Mispronunciation detection without non-native training data." in *Proc. Interspeech*, 2015, pp. 643–647.

[7] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in L2 English speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, Jan 2017.

[8] A. M. Harrison, W.-K. Lo, X.-j. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *Proc SLaTE*, 2009.

[9] R. Duan, T. Kawahara, M. Dantsuji, and J. Zhang, "Effective articulatory modeling for pronunciation error detection of L2 learner without non-native training data," in *IEEE ICASSP*. IEEE, 2017, pp. 5815–5819.

[10] V. Arora, A. Lahiri, and H. Reetz, "Phonological feature-based speech recognition system for pronunciation training in non-native language learning," *The Journal of the Acoustical Society of America*, vol. 143, no. 1, pp. 98–108, 2018.

[11] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Improving mispronunciation detection for non-native learners with multisource information and lstm-based deep models," *Proc. Interspeech*, pp. 2759–2763, 2017.

[12] H. Franco, H. Bratt, R. Rossier, V. Rao Gadde, E. Shriberg, V. Abrash, and K. Precoda, "Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Language Testing*, vol. 27, no. 3, pp. 401–418, 2010.

[13] J. Tao, S. Ghaffarzadegan, L. Chen, and K. Zechner, "Exploring deep learning architectures for automatically grading non-native spontaneous speech," in *IEEE ICASSP*, March 2016, pp. 6140–6144.

[14] Y. Qian, K. Evanini, X. Wang, C. M. Lee, and M. Mulholland, "Bidirectional LSTM-RNN for improving automated assessment of non-native children's speech," *Proc. Interspeech*, pp. 1417–1421, 2017.

[15] S.-K. Xu, S. Wei, Z.-H. Ling, Q.-Y. Gao, L.-R. Dai, and Q.-F. Liu, "A statistical modeling approach to automatic evaluation of mandarin pronunciation," *Journal of the Phonetic Society of Japan*, vol. 19, no. 1, pp. 44–52, 2015.

[16] R. Karhila, A. Rouhe, P. Smit, A. Mansikkaniemi, H. Kallio, E. Lindroos, R. Hildén, M. Vainio, M. Kurimo *et al.*, "Digitala: An augmented test and review process prototype for high-stakes spoken foreign language examination." in *Proc. Interspeech*, 2016, pp. 784–785.

[17] J. Yue, F. Shiozawa, S. Toyama, Y. Yamauchi, K. Ito, D. Saito, and N. Minematsu, "Automatic scoring of shadowing speech based on DNN posteriors and their DTW," *Proc. Interspeech*, pp. 1422–1426, 2017.

[18] K. Evanini and X. Wang, "Automated speech scoring for non-native middle school students with multiple task types." in *Proc. Interspeech*, 2013, pp. 2435–2439.

[19] S. Papageorgiou, R. J. Tannenbaum, B. Bridgeman, and Y. Cho, "The association between TOEFL iBT® test scores and the common european framework of reference (cefr) levels," *Princeton, NJ: Educational Testing Service*, 2015.

[20] D. Liakin, W. Cardoso, and N. Liakina, "Learning L2 pronunciation with a mobile speech recognizer: French/y," *CALICO Journal*, vol. 32, no. 1, p. 1, 2015.

[21] I. Masuda-Katsuse, "Pronunciation practice support system for children who have difficulty correctly pronouncing words," in *Proc. Interspeech*, 2014, pp. 2144–2145.

[22] S. S.-C. Young and Y. H. Wang, "The game embedded CALL system to facilitate english vocabulary acquisition and pronunciation." *Educational Technology & Society*, vol. 17, no. 3, pp. 239–251, 2014.

[23] J.-C. Chen, J.-L. Lo, and J.-S. Jang, "Computer assisted spoken english learning for chinese in taiwan," in *Proc. ISCSLP*. IEEE, 2004, pp. 337–340.

[24] J. Wang, Y. Qin, Z. Peng, and T. Lee, "Child speech disorder detection with siamese recurrent network using speech attribute features." in *Proc. Interspeech*, 2019, pp. 3885–3889.

[25] A. C. Kelly, E. Karamichali, A. Saeb, K. Veselỳ, N. Parslow, A. Deng, A. Letondor, R. O'Regan, and Q. Zhou, "Soapbox labs verification platform for child speech." in *Proc. Interspeech*, 2020, pp. 486–487.

[26] C. Liyanage, U. Kavinda, D. Dasanayaka, P. Shehara, and D. De Silva, "Interactive mobile application for initial skills development of primary students in sri lanka," in *Proc. ICAC*. IEEE, 2022, pp. 358–362.

[27] E. Booth, J. Carns, C. Kennington, and N. Rafla, "Evaluating and improving child-directed automatic speech recognition," in *Proc. LREC*, 2020, pp. 6340–6345.

[28] Y. Getman, R. Al-Ghezi, E. Voskoboinik, T. Grósz, M. Kurimo, G. Salvi, T. Svendsen, S. Strömbergsson *et al.*, "wav2vec2-based speech rating system for children with speech sound disorder," in *Proc. Interspeech*, 2022.

[29] S. Bannò and M. Matassoni, "Proficiency assessment of l2 spoken english using wav2vec 2.0," *arXiv preprint arXiv:2210.13168*, 2022.

[30] A. Hair, K. J. Ballard, C. Markoulli, P. Monroe, J. Mckechnie, B. Ahmed, and R. Gutierrez-Osuna, "A longitudinal evaluation of tablet-based child speech therapy with apraxia world," *ACM TACCESS*, vol. 14, no. 1, pp. 1–26, 2021.

[31] A. Neri, O. Mich, M. Gerosa, and D. Giuliani, "The effectiveness of computer assisted pronunciation training for foreign language learning by children," *Computer Assisted Language Learning*, vol. 21, no. 5, pp. 393–408, 2008.

[32] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *IEEE ICASSP*, vol. 1. IEEE, 2006, pp. I–I.

[33] A. Juneja and C. Espy-Wilson, "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines," in *Proc. IJCNN*, vol. 1, July 2003, pp. 675–679 vol.1.

[34] V. Arora, A. Lahiri, and H. Reetz, "Phonological feature based mispronunciation detection and diagnosis using multi-task DNNs and active learning," *Proc. Interspeech*, pp. 1432–1436, 2017.

[35] R. Karhila, A.-R. Smolander, S. Ylinen, and M. Kurimo, "Transparent Pronunciation Scoring Using Articulatorily Weighted Phoneme Edit Distance," in *Proc. Interspeech*, 2019, pp. 1866–1870.

[36] K. Junttila, A.-R. Smolander, R. Karhila, M. Kurimo, and S. Ylinen, "Targeted training benefits spoken foreign-language processing in children with dyslexia," *Frontiers in Human Neuroscience*, 2023.