IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 June 2023, accepted 14 July 2023, date of publication 10 August 2023, date of current version 17 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3304274

RESEARCH ARTICLE

Developing an AI-Assisted Low-Resource Spoken Language Learning App for Children

YAROSLAV GETMAN[1], NHAN PHAN[1],
RAGHEB AL-GHEZI[1], (Graduate Student Member, IEEE), EKATERINA VOSKOBOINIK[1],
MITTUL SINGH[1,2], TAMÁS GRÓSZ[1], MIKKO KURIMO[1],
GIAMPIERO SALVI[3,4], (Member, IEEE), TORBJØRN SVENDSEN[3], (Life Senior Member, IEEE),
SOFIA STRÖMBERGSSON[5], ANNA SMOLANDER[6], AND SARI YLINEN[6]

[1]Department of Information and Communications Engineering, Aalto University, 02150 Espoo, Finland
[2]Silo AI, 00180 Helsinki, Finland
[3]Department of Signal Processing, Norwegian University of Science and Technology, 7034 Trondheim, Norway
[4]EECS, KTH Royal Institute of Technology, 114 28 Stockholm, Sweden
[5]Department of Clinical Science, Intervention and Technology, Karolinska Institutet, 141 52 Huddinge, Sweden
[6]Logopedics, Welfare Sciences, Faculty of Social Sciences, Tampere University, 33100 Tampere, Finland

Corresponding author: Yaroslav Getman (yaroslav.getman@aalto.fi)

This work was supported by NordForsk through funding to technology-enhanced foreign and second-language learning of Nordic languages under Project 103893.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Regional Ethical Review Board in Stockholm.

ABSTRACT Computer-assisted Language Learning (CALL) is a rapidly developing area accelerated by advancements in the field of AI. A well-designed and reliable CALL system allows students to practice language skills, like pronunciation, any time outside of the classroom. Furthermore, gamification via mobile applications has shown encouraging results on learning outcomes and motivates young users to practice more and perceive language learning as a positive experience. In this work, we adapt the latest speech recognition technology to be a part of an online pronunciation training system for small children. As part of our gamified mobile application, our models will assess the pronunciation quality of young Swedish children diagnosed with Speech Sound Disorder, and participating in speech therapy. Additionally, the models provide feedback to young non-native children learning to pronounce Swedish and Finnish words. Our experiments revealed that these new models fit into an online game as they function as speech recognizers and pronunciation evaluators simultaneously. To make our systems more trustworthy and explainable, we investigated whether the combination of modern input attribution algorithms and time-aligned transcripts can explain the decisions made by the models, give us insights into how the models work and provide a tool to develop more reliable solutions.

INDEX TERMS ASR, children's speech, L2 speech, speech rating, SSD, wav2vec2.

I. INTRODUCTION

Learning foreign or second languages (L2) is a challenge for most adult learners, whereas children starting L2 learning early may eventually obtain better proficiency in different aspects of language [1], [2]. Therefore, it is often recommended to start language learning early on. For young illiterate children, L2 exposure and activities should be based on spoken language rather than text, yet also older children benefit from spoken language skills. Smart mobile systems coupled with automated game-based spoken language learning (GBLL) provide a unique opportunity to enhance spoken L2 acquisition for young learners. To be interactive, such systems are, however, dependent

The associate editor coordinating the review of this manuscript and approving it for publication was Ghulam Muhammad.

on speech technology. Therefore, there has been increasing interest in applying machine learning techniques to the field of spoken language learning for children. We aimed to develop a mobile application for children to practice language learning in an interactive and gamified fashion (see Section V for more details). It leverages recent advances in self-supervised machine learning to develop automatic speech recognition (ASR) for children's L2 speech in two low-resource languages such as L2 Finnish and Swedish spoken in Finland, which are scarcely studied in the field of computer-assisted language learning (CALL). In addition, this work has the potential to affect the language learning process in children with speech sound disorder (SSD). An engaging speech training app can help provide immediate feedback regarding their speech production. Usually, children with SSD constitute a large portion of speech-language pathologists' caseloads [3], and given the value of a high training dose frequency in speech intervention [4], an engaging speech training app could potentially relieve the burden of clinicians as well as children and their families alike.

Previous studies have shown that CALL and GBLL can be effective ways to learn L2 and particularly its vocabulary [5], [6], [7], [8], [9], [10]; for reviews, see [11], [12], [13], and [14]. Digital games have also been used to support speech therapy [15]. According to [11], the aspects of GBLL contributing to this include ease of use, challenge, reward-and-feedback, control/autonomy, goal-directedness, and interactivity. These factors may increase learners' motivation for rehearsal and activate the reward system of the brain [16]. The game features may also interact with players' individual characteristics and their learning styles [17], implying that the gaming effects vary across individuals. To train spoken language and pronunciation skills, DLL and GBLL have also been combined with speech technology [15], [18], [19]. ASR that provides feedback to learners has been shown to improve vocabulary and pronunciation [20], [21], [22]. In addition to learning effects observed in behavior, the effects of gaming with ASR-based feedback have been shown to improve children's neural representation of L2 speech sounds and words in the brain [23].

Based on these earlier findings, we expect that a system combining gaming and ASR, with its interactive and engaging nature, will prove to be a valuable teaching tool for developing the spoken language skills of L2 learners. Outside the educational setting, the system could also be beneficial for children with SSD, and encourage them to reach the high levels of practice and repetition that are often recommended in clinical intervention. This research will contribute to the body of knowledge on the use of technology in language teaching and clinical intervention, especially in low-resource languages, and provide valuable insights for educators, clinicians, researchers, and developers working in the field of spoken language learning. Furthermore, the described system will be a useful and practical resource for educators and parents, providing them with effective tools to assist children in second language acquisition.

A systematic literature review on automatic pronunciation assessment for L2 children and children with SSD has been conducted in [24]. However, the scope of this review is substantially different from our work, which makes it hard to compare the results and methods. To the best of our knowledge, in the context of Computer-Assisted Pronunciation Training (CAPT) for L2 Swedish and Finnish children, there are no previous work on automatic pronunciation assessment, not even for L2 Swedish and L2 Finnish adults. For other languages, pronunciation verification systems use a common method called Goodness of Pronunciation (GOP) [25], [26]. GOP is estimated as the probability that the expected phone is observed with respect to all the other observable phones. The GOP method has been successful in a wide range of pronunciation verification systems [27], [28]. However, it requires the development of a performant ASR system to function well. This is not always feasible in settings where labeled data are scarce, such as children's speech or low-resource languages. Children's speech differs from adults' speech in F0, speaking rate, and formant frequencies [29], making ASR systems for adults not applicable to the task. Nevertheless, recent developments in ASR, notably the availability of large mono- and multilingual pre-trained speech models, such as wav2vec2 [30], have demonstrated success in developing ASR systems using relatively small amounts of training data [31] on target tasks, making it possible to develop ASR systems for low-resource children's speech.

In this work, we make the following contributions. Firstly, we introduce a multitask wav2vec2 approach that performs ASR and pronunciation scoring simultaneously. Previously, these two steps were done separately, and this novel approach improves the computational efficiency and the speed of the system, making it fit for mobile-based gamified systems. Secondly, we investigate how well the proposed system can represent children's speech internally via an interpretability method called Integrated Gradients [32]. Finally, we evaluate the effectiveness of the proposed solutions on three different speech datasets: Swedish SSD and Swedish and Finnish L2 learning.

II. DATA

In this study, we employ three children's speech datasets for training the ASR and pronunciation rating systems. During data collection, the ethical boards provided approval and informed consent was obtained from the parents of all the children participants. After the data collection phase, the speech samples were manually rated by human annotators. Since our target users are very young (below the age of 10), we were quite limited in the options for feedback; in the end, we opted to use the well-known five-star rating system, which might be already familiar to the children. As a result, we unified the rating scales across the datasets and converted the scores accordingly, more details can be found in the

TABLE 1. Rating scale and descriptions.

Rating	Description
1	Not at all identifiable as the target word
2	Hard to identify as the target word
3	Slight phonemic error
4	Subphonemic error / "unexpected variant"
5	Prototypical / adult-like / correct

following sections. The final level distributions are shown in Figure 1, and the level descriptions are provided in Table 1. It should be noticed that separation between neighboring levels is expected to be a challenging task, both for human raters and automatic rating systems. In assessing speech accuracy in a clinical context, speech-language pathologists typically attend to specific features of speech, documented in phonetic transcription or as a percentage of consonants in a speech sample that are produced "correctly". Hence, a global rating of speech accuracy with regards to a 1-5 scale is an unfamiliar task even to expert listeners. Furthermore, the datasets are heavily imbalanced in terms of ratings towards the highest level: about half of the speech samples belong to level 5.

While the data have human ratings, it should be noted that the speech samples were not transcribed. Instead, only the target word was provided for each recording along with the rating. Therefore, we used all the data in the speech rating experiments, while samples only with ratings 4 and 5 were used for training and evaluating the ASR systems, expecting that the uttered word corresponds to the target word in these recordings.

A. SweSSD

The first dataset used in this study, named SweSSD, consists of 6027 isolated word recordings (2 hours) collected from 28 native Swedish children aged 4 to 10 years by the Functional consequences of misarticulation in children's connected speech project [33]. The vocabulary of 1109 words was compiled based on the articulation test LINUS [34] and the Swedish Test of Intelligibility for Children (STI-CH) [35]. Among the participants, 16 children had an SSD and the remaining 12 speakers had typical speech. After the data collection, the speech samples were pseudonymized to prevent the identification of speakers and rated on a 5-level scale by a native Swedish speech-language pathologist, for more details see Table 1. In addition, we asked the annotator to re-rate a randomly sampled 20% of the dataset half a year after the first data annotating trial. We compare the new ratings to the corresponding original ones and report the results in Section IV-C.

B. L2 SWEDISH AND FINNISH DATA

Our L2 data were collected from children of ages 7 to 11 who had not studied the target language at school yet. The Swedish samples were recorded from L1 Finnish speakers, while the Finnish recordings were collected from Ukrainian children whose mother tongue is Ukrainian or Russian.

For each language, we prepared a word list containing all the sounds that were considered important for the target language. These words either exist in the target (L2) language but not in the children's native language, or are expected to be difficult for L2 learners in general, such as Finnish words that contain front vowels ä, ö, and y (/æ/, /ø/, and /y/). The Swedish set was composed of 121 unique words, while the Finnish one had 90 words.

During the data collection, children were instructed to put on headsets and repeat the words that they heard. We used a toy animal as a proxy to which the child repeated the word. As the recording was relatively long from the kids' perspective, it was sometimes necessary to take small breaks every 3-4 minutes or change the toy animal, while some children were able to record all the words in one go. A quality check was implemented for each file by manually listening to them, modifying the labels to cover the child's utterances, and marking extra sounds and noises.

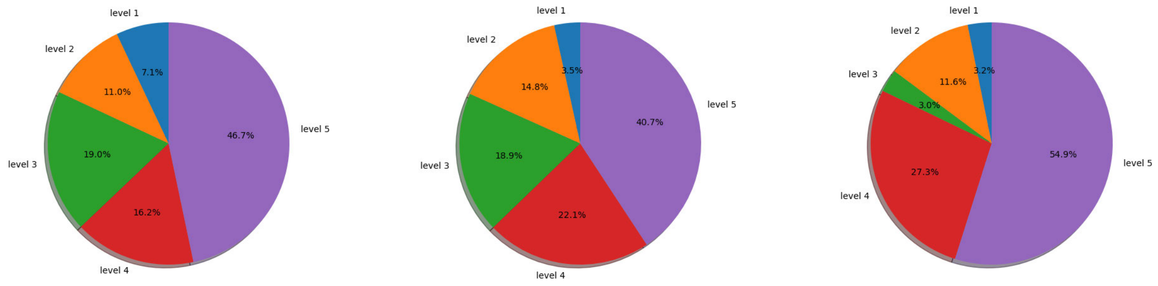
The collected data were rated by native Finnish university students in the last year of their master's studies, majoring in Swedish language and specializing in language teaching, with practical teacher experience. Furthermore, the annotators were trained by an experienced annotator with the same qualities and who was also one of the developers of the used speech technology rating platform [18], [23], [36], [37], [38]. At the beginning of the annotators' training period, they assessed a training set from the data together and discussed the rating procedure until a consensus was reached and they were feeling confident on their annotation.

The L2 Swedish data set consists of 2384 speech utterances collected from 20 children (90 minutes), while the Finnish one is composed of 2124 utterances from 24 speakers (83 minutes). The distribution of the data in the rating levels is shown in Figure 1.

III. METHODS

A. BASELINE

As a baseline for the end-to-end pronunciation assessment models, we applied a phoneme-level GOP score [25], [26] based on traditional Gaussian mixture models combined with Hidden Markov models (GMM-HMMs). This method compares the likelihood of the spoken utterance given the models and, assuming a canonical pronunciation, to the best possible likelihood when the model is free to choose any possible pronunciation. If the pronunciation is close to canonical, the two likelihoods are similar and the likelihood ratio is close to one (or, equivalently, the log-likelihood ratio is close to zero). If there is a significant deviation in pronunciation, the canonical likelihood is always lower than the best possible likelihood, giving negative log scores. The GMM-HMM models are phonetic, allowing for a detailed assessment of pronunciation deviations. However, because the task in this paper was to predict a human assessor score at the word level, only word-level likelihoods were considered.



(a) SweSSD (b) L2 Swedish (SweL2) (c) L2 Finnish (FinL2)

FIGURE 1. Data distribution in different rating levels.

The GMM-HMM models were trained on the NST Swedish database [39] and adapted on child speech from the PFstar dataset [40]. We used context-dependent HMMs (triphones) with 64 Gaussian components per state trained on MFCC features. The log-likelihood ratio scores obtained from the model had to be mapped to the 5-star rating scale. The decision tree was chosen as a simple tool to perform the transformation of GOP scores into rating categories, which made it directly comparable to other systems used in this study (explained later). In Section IV, we refer to this method as *MFCC+GMM-HMM+DT*.

B. Wav2vec2

Wav2vec2 [30] is a self-supervised framework consisting of a series of convolutional layers followed by a Transformer [41] network. The model learns general deep acoustic representations during pre-training from large amounts of unlabeled speech data. More precisely, it masks spans of consequent audio representations and is trained to distinguish between the true quantized latent representation and the negative examples randomly sampled from the same speech utterance. The second training phase involves fine-tuning the model with labeled data to a downstream task. For example, for ASR, a linear layer is added on top of the Transformer network, and the system is trained with a connectionist temporal classification (CTC) [42] loss. For utterance-level classification tasks, the representations of the last hidden layer are reduced in dimensionality by a projection layer, then combined using average pooling and fed into a classification layer. Cross-entropy (CE) loss is then used as a loss function. In addition, wav2vec2 can be fine-tuned for both ASR and speech pronunciation classification simultaneously by feeding the last hidden layer's outputs to multiple heads and jointly minimizing the corresponding loss functions. We refer to the single-task wav2vec2 models in Section IV as *W2V2 ASR* and *W2V2 rating* and the multi-task ones as *W2V2 multitask*.

C. MODEL INTERPRETATION

While AI tools are routinely developed for various tasks thanks to the numerous frameworks which streamline the training and deployment process, they are generally viewed as ''black boxes''. Unfortunately, not understanding how the system works could lead to catastrophic failures when the AI model learns to exploit some unintended artefact in the training data [43], thus resulting in a model that is not well generalized and performs under our expectations on unseen, real-life test data. Motivated by this, we perform model interpretation with the Integrated Gradients (IG) [32] tool to ensure our models do not rely on special properties of the training data.

IG is a popular solution for visual interpretation as it is applicable to any differentiable model. It only requires a baseline input (in our case, it was complete silence) and calculates the attributions of each input feature toward the final output. In practice, IG uses the gradients of the output with respect to the input to estimate the attributions and to ensure that the attributions satisfy the Sensitivity criteria, it calculates the attributions of various inputs, which are an interpolation of the baseline and the actual input. Formally, IG attributions are defined as the path integral of the gradients along the straight line path between the baseline and input vectors [32]. These input attributions could offer insights into which parts of the audio are mostly considered when the model rates the pronunciation of the words. Furthermore, it could highlight potential weaknesses of the system, which should be addressed before we deploy the models in the mobile application.

We would like to note that while model interpretation techniques are commonly used to inspect models visually, employing them to discover systematic problems of the networks is still rare. Perhaps one rare but well-known example is the Husky or Wolf model, which learned to separate the two categories based on the presence of snow in the background [43]. Naturally, such a system is not trustworthy, as it makes predictions based on irrelevant parts of the input.

Inspired by this, we developed a novel solution to investigate our solutions and their potential weaknesses. As our system is essentially one single neural network, we could use IG to estimate the input attributions of each value in the raw audio input. To perform a systematic analysis, we also utilized the ASR component of our proposed solution to generate a so-called forced-aligned transcript of the expected word. This procedure provided us with timing information on when the expected word begins and when it ends, thus separating the relevant part of the audio from those that should not be considered when evaluating the pronunciation. Our experiments focused on the distribution of input attributions among these two recorded components to ensure that the most influential regions belong to the expected word and not to the environmental noises and other sounds recorded before and after the pronunciation.

IV. EXPERIMENTS

A. SPEECH RATING SYSTEM

Our primary goal was to design and train models that analyze recordings of single words uttered by children and estimate the goodness of their pronunciation. Due to the limited amount of training data, we selected the GOP baseline and our wav2vec2-based classifiers and compared them thoroughly using the largest dataset at our disposal (SweSSD). In Figure 2, we can see the overall workflow of all the systems employed in this study. First, we tested two relatively simple solutions that used the standard Mel-frequency cepstral coefficients (MFCC) extracted from the audio file. The traditional GOP approach used GMM-HMMs to estimate the ratings from the MFCC input, and the deep learning alternative replaced the GMM-HMM with a convolutional recurrent deep neural networks (CRDNN), which is a quite popular encoder model for audio processing [44]. The architecture of the CRDNN is a relatively simple one; the input is first processed by two convolutional layers along the time axis. Following the convolutional layers, a bidirectional LSTM layer summarizes the utterance-level information. Lastly, a feed-forward layer transforms the embeddings before the final softmax layer. We refer to this approach as *MFCC+CRDNN* in our experiments.

Next, we investigated modern, self-supervised wav2vec2 solutions. Although these models are very good at speech recognition [30], [31], spoken emotion recognition [45], disfluency detection [46], and many other related tasks, they have a considerable computational cost. Training such models require expensive infrastructure (GPUs and large memory) and a considerable amount of time, even if the training data is limited. Naturally, the most straightforward option is to fine-tune the models using all available data, and we demonstrate that it leads to the best systems, but it might not be a possibility for those with limited resources. To explore lightweight alternatives, we decided to use the wav2vec2 as a feature extractor only, thus avoiding the costly second fine-tuning procedure after the ASR training. Three different systems were compared; one used a simple Decision Tree (DT) to assign a rating based on the character error rate (CER) between the expected word and the transcript produced by wav2vec2, also referred to as *W2V2 CER+DT* in Section IV. Our GOP (MFCC+GMM-HMM+DT) is directly comparable and acts as a baseline for the W2V2 CER+DT system as both apply a speech model to produce a score, which is transformed into a rating by a DT.

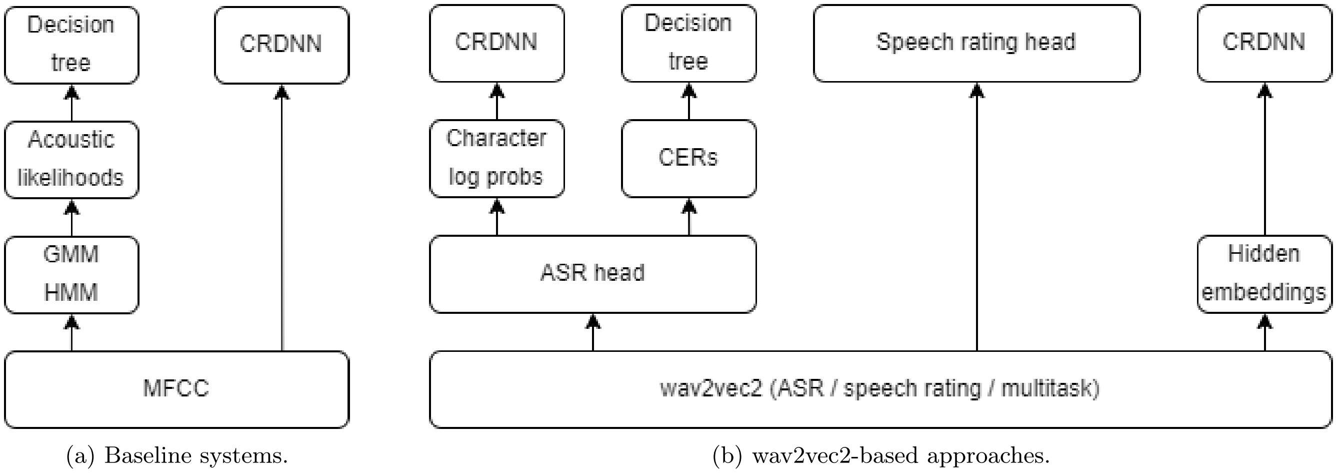
The second alternative employed a CRDNN that received the character-level log probabilities from the wav2vec2, while the third used a CRDNN and the so-called context embeddings to estimate the pronunciation quality. We refer to these methods in the results as *W2V2 logp+CRDNN* and *W2V2 emb+CRDNN*, respectively. We should note that although our DT and CRDNN models are orders of magnitude smaller than wav2vec2 and their training takes considerably less time and resources, they still required a computationally expensive fine-tuning of the wav2vec2 model for ASR. Moreover, the inference using all three approaches still requires a pass through the wav2vec2 network.

B. EVALUATION METRICS

In this study, we have access to only a limited amount of data, which we aim to utilize as much as possible for training and evaluation of the models. To achieve this, we opted for 6-fold cross-validation (CV) with no overlap between folds in the ASR and speech rating experiments. With CV, we were able to use the entire dataset to test our models, and the fact that we had to train 6 models for each task enabled us to assess the robustness of our proposed systems toward data selection.

The second major issue we had to address was the choice of evaluation metric. In our work, accuracy could be misleading because of the unbalanced nature of the corpora used in this study; for example, a classifier that assigns the most frequent class label (rating 5) to everything would achieve relatively good accuracy (approx. 50%) compared to random choice (approx. 20%), still, it would be useless in a real-world application. Unweighted average recall (UAR) [47] is designed to measure performance better in case of unbalanced data by calculating the recalls, or sensitivity, of each category and averaging them to get a final performance measure. Although UAR already addresses the unbalanced data problem, neither recall nor accuracy takes the distance between ratings into account. For instance, both recognizing the lowest level sample as the highest level and confusion between the neighboring classes are treated equally by these metrics. To solve this limitation, we also opted to use mean absolute error (MAE), which could be viewed as the expected difference between the model prediction and the human annotation.

Lastly, we employed the word and the character error rate (WER and CER) to evaluate our children ASR systems. One reason behind measuring the CER is that the WER could often mislead as it is sensitive to minor mistakes such as getting only one character wrong in the word would categorize it already as an error. Additionally, it should be noted that our children's speech corpora are single-word utterances, therefore, the CER would provide us with a more informative estimate of the ASR performance.



(a) Baseline systems.

(b) wav2vec2-based approaches.

FIGURE 2. The overall workflow of our experiments.

C. SweSSD

In the first phase of our experiments, we selected the SweSSD corpus and compared multiple solutions to determine the best. We used the entire dataset for the speech rating experiments without removing underrepresented classes or merging them into a single class.

In Table 2, we report the attempt to reproduce the experiments proposed in [48]. Some results deviate from the ones reported by [48] due to several reasons. First, we trained the wav2vec2-based ASR systems following the cross-validation setup instead of using all data with a high rating to train a single ASR model. Additionally, we selected a publicly available monolingual wav2vec2 model[1] already fine-tuned on adult Swedish speech as a base model in our experiments. According to [48], adapting a wav2vec2 model already fine-tuned for ASR on the same language but different domain leads to considerably lower error rates in comparison to fine-tuning a pre-trained model directly for the under-resourced target task. Second, we used different hyperparameters for the wav2vec2 models. The main observation is that wav2vec2-based solutions proved quite good, reducing the gap between human and AI performance considerably. Our main observation from these results is that the training of the wav2vec2 model is highly beneficial but also relatively slow. The baseline GOP system (MFCC+GMM-HMM+DT), in contrast, provided the lowest performance due to the small size of our training data compared to the wav2vec2-based methods that benefit from the large pre-trained models. As a result, it was not selected among the methods for further experiments.

We would like to note that the system with the second-best UAR score, W2V2 CER+DT, required only a few seconds to train, whereas training the other wav2vec2 sub-models took approximately 4.5 hours. Nevertheless, after the training procedure, the difference between their inference speed was negligible. One significant limitation of our best model (W2V2 rating) is that the wav2vec2 fine-tuned to provide ratings lost its ability to recognize the word spoken. In contrast, all solutions that are built upon the wav2vec2 ASR model (W2V2 logp+CRDNN, W2V2 emb+CRDNN, and W2V2 CER+DT) can still leverage the ASR capability. Maintaining the ASR capability is important because then the output of the same model can be used to verify that the child indeed tried to utter the expected word and not just said something completely different. In fact, by comparing the recognized text with the expected word, we can determine with about 90% accuracy whether the child uttered the expected word or not. We would like to note that most (approx 65%) of the false rejections are in the case of very low ratings (1 or 2) when even the human annotators had trouble determining whether the child said the target word or not.

TABLE 2. Previous results on the SweSSD of various techniques proposed in [48].

System	ACC (%)	UAR (%)	MAE
MFCC+GMM-HMM+DT	42.3	30.4	0.92
MFCC+CRDNN	47.4	35.1	0.91
W2V2 logp+CRDNN	49.2	33.1	0.81
W2V2 emb+CRDNN	52.4	38.1	0.72
W2V2 CER+DT	51.9	39.8	0.80
W2V2 rating	60.3	48.3	0.54
Human (20% data)	65.8	66.7	0.39

Unsurprisingly, the human expert (*Human (20% data)* in Table 2) considerably outperformed our automatic systems in this task. However, the obtained results clearly demonstrate the degree of complexity of the task: rating these samples on a 5-level scale is not trivial even for a human annotator.

Next, we focused on understanding how the best wav2vec2 model made its decisions. For this, we employed an input

[1] https://hf.co/KBLab/wav2vec2-large-voxrex-swedish

attribution method called Integrated Gradients (IG). This method aided us by revealing the most influential parts of the input by calculating their contributions towards the final decision of the network. IG is capable of generating so-called saliency maps, which enables us to gain insights into how large neural networks work [49]. Additionally, understanding how our methods function is a prerequisite of trustworthy systems [43], and the generated explanatory information could be helpful for experts to better understand the problem [50] and developers for building better solutions.

Naturally, input attribution methods, including IG, are known to be fragile [51], so we estimated the Infidelity [52] of the Integrated Gradients by perturbing the raw audio input with a small Gaussian noise ($\mu = 0$, $\sigma = 0.03$) to ensure that we can trust the observations that we made based on the outputs of IG. Our analysis revealed that the input attributions had an Infidelity of 2.4 (std 7.4), and only $\approx$ 14% of the samples showed a high response (infidelity>4) to the added noise. In a few extreme cases, the generated IG explanations became unreliable, which resulted in a high standard deviation compared to the mean. Based on this observation we concluded that the IG attribution maps can be trusted in general and proceeded with our further analysis of the models' behaviors.

In Figure 3, we show examples of how one can use input attribution to justify the predicted ratings and provide more detailed feedback to the users. Unfortunately, we also noticed that in some cases, the system learned to focus on the environmental noises (i.e. the non-speech parts at the beginning and the end of the recording), which is an undesirable behavior (see Figure 3d). After careful investigation, we determined that approx 35% of the samples were rated predominantly based on the non-speech parts before and after the word. To prevent this, we opted to add a Voice Activity Detector (VAD) [53] to our solution, which clipped away the silent parts of the recordings. This had two beneficial effects: it prevented our system from exploiting other information beyond the actual speech of the user and, at the same time, sped up the audio processing by reducing the duration of all recordings.

Lastly, to maintain the speech recognition capabilities of the W2V2, we employed multitask learning during the fine-tuning procedure. This ensured that the system could provide ratings and the recognized text simultaneously. Table 3 compares monotask models with the multitask one and demonstrates the effects of adding VAD to the system. Multitask training provides almost as good results as normal monotask training. The WER and the CER of the multitask system increase by 1.18% and 1.59% relative to the separate ASR model, and the UAR score increases by 1.47% relative to the individual classification model, while the accuracy and the mean absolute error stay unchanged. In contrast to the monotask models, using the multitask solution lets us get both outputs from the same model for the price of a small degradation in performance, which also brings us the advantages of halved memory and computational time requirements.

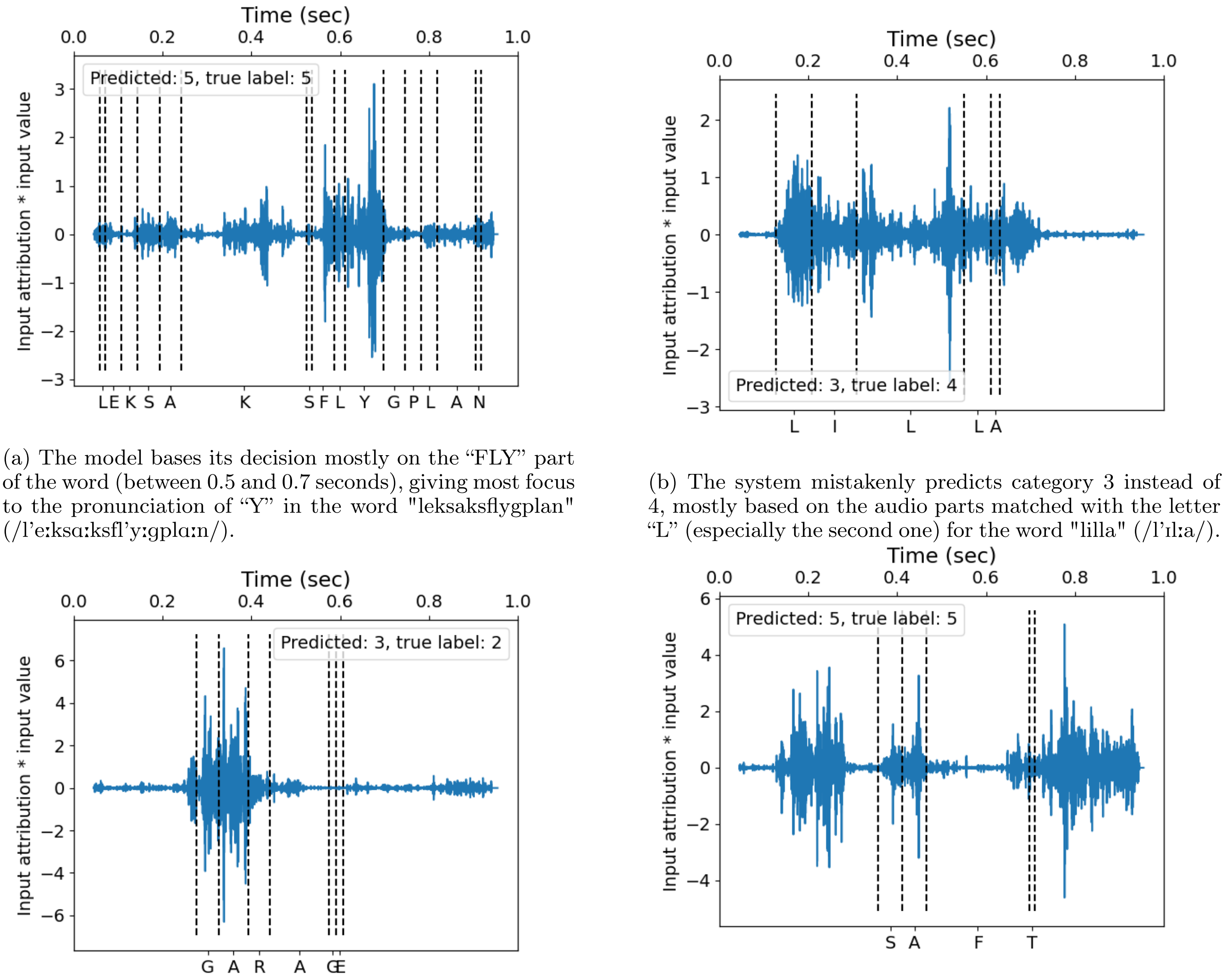
Training the models on the audio preprocessed by a VAD system did not improve these performance metrics, but actually degraded them a bit. The most evident reason for this slight degradation in the performance of all systems lies probably in the reduced overall amount of training data. These silent parts proved to be beneficial for our ASR systems and might have improved their ability to learn the blank label and to distinguish between the non-speech events ignored by the VAD algorithm and the actual speech. Additionally, some background information, such as acoustic conditions, might have been present in the parts cut by the VAD system, and, consequently, the wav2vec2 models were not able to exploit them. As a result, we decided to discard the VAD in further experiments, however, we still plan to use it in our mobile game application to avoid undesired biases.

D. SweL2

Our next experiments were performed on the children's Swedish L2 data. Similarly to the previous experiments, we followed the 6-fold cross-validation setup and excluded the samples with a rating lower than 4 when training and evaluating the ASR systems. Additionally, apart from SweSSD, we managed to split the folds here by speaker, since this dataset includes also child speaker IDs. We used the same adult Swedish wav2vec2 system as we did for SweSSD to serve as a base model for further training.

The monotask ASR model provides 8.95% WER and 3.70% CER, while the ASR component of the multitask system achieves slightly higher error rates of 9.95%/4.04% WER/CER. These results are very good for the difficult L2 children's speech, but this is probably because, as discussed in Section II-B, there were only 121 unique target words in the L2 Swedish corpus. Even though the ASR system does not use any lexicon or language model, the ASR models trained on these data are expected to know these words very well, and introducing new prompts would probably degrade the ASR performance.

Table 4 summarizes the results of the rating experiments performed on the L2 Swedish data. Similarly to the SweSSD experiments, the multitask system here slightly underperformed the monotask one for the price of preserved ASR capability. According to the recall results, our simple CER-based decision tree performs the best. However, it should be noted that the recall metric treats all misclassifications equally and does not take into account the distance between the true score and the predicted level. We analyzed the predictions made by the systems and discovered that our decision tree, despite producing the highest UAR score, predicts almost always either the lowest or the highest level, in other words, it basically learned to provide a binary decision. This can be seen also from a much higher MAE score of the decision tree compared to those of other systems. In contrast, the wav2vec2 models are



(a) The model bases its decision mostly on the "FLY" part of the word (between 0.5 and 0.7 seconds), giving most focus to the pronunciation of "Y" in the word "leksaksflygplan" (/l'eːksɑːksfl'yːgplɑːn/).

(b) The system mistakenly predicts category 3 instead of 4, mostly based on the audio parts matched with the letter "L" (especially the second one) for the word "lilla" (/l'ɪlːa/).

(c) The rating is decided based on the pronunciation of the first part of the word "garage" (/gar'ɑːɦ/).

(d) The model predicts the correct label but exploits some information present in the non-speech parts of the audio when rating the pronunciation of "saft" (/safːt/).

FIGURE 3. Input attributions (multiplied with the actual input values) of the best wav2vec2 model. The pictures demonstrate the influence of each individual input on the final model output, i.e. large values signify a considerable impact on the system's final decision.

TABLE 3. wav2vec2 results on SweSSD.

	ASR				Speech Rating					
System	WER (%)		CER (%)		ACC (%)		UAR (%)		MAE	
	orig.	VAD	orig.	VAD	orig.	VAD	orig.	VAD	orig.	VAD
W2V2 ASR	17.0	18.0	6.3	7.1	N/A	N/A	N/A	N/A	N/A	N/A
W2V2 Rating	N/A	N/A	N/A	N/A	60.3	59.9	48.3	47.8	0.54	0.55
W2V2 Multitask	17.2	18.1	6.4	7.2	60.3	59.8	47.6	46.7	0.54	0.5

mostly confused between the neighboring levels, moreover, they are prone to provide higher scores than humans. In other words, our wav2vec2 systems seem more lenient compared to human raters, which has been a beneficial feature for a CALL system to encourage language learners [54]. Lastly, we also had some data, which were rated by both human annotators, which allowed us to estimate how well human experts can perform this task. At first glance, based on accuracy and UAR, we could say that humans are far superior in this task, but once we consider the average

TABLE 4. Rating performances of automatic systems compared to those of the human raters on SweL2.

System	ACC (%)	UAR (%)	MAE
W2V2 CER + DT	44.5	37.7	1.0
W2V2 rating	48.9	36.4	0.7
W2V2 multitask	48.2	35.1	0.7
Human (12% data)	60.1	55.3	0.6

TABLE 5. Rating prediction results on FinL2.

System	ACC (%)	UAR (%)	MAE
W2V2 CER + DT	58.4	34.2	0.6
W2V2 rating	72.3	46.1	0.4
W2V2 multitask	72.1	43.1	0.4

errors (MAE) it demonstrates that although classifiers make more mistakes, those misrecognized samples, in general, receive a rating very close to the human annotation. Looking at the MAE results, we can see that the gap between W2V2 models and humans is much smaller on this challenging corpus compared to the one we observed in the case of SweSSD.

E. FinL2 (UKRAINIAN)

Next, we repeated the set of experiments done for L2 Swedish on the L2 Finnish data. The dataset was split into 6 folds with no speaker overlap, and the samples with a rating lower than 4 were excluded when training and evaluating the ASR systems. Because there is no monolingual Finnish wav2vec2 model available yet, we used as a base model the multi-lingual wav2vec2 model [55] pre-trained on the European parliamentary session recordings in 3 languages from the Uralic language family, including Finnish, Estonian, and Hungarian, and preliminarily fine-tuned on 100 hours of colloquial adult Finnish speech from the Lahjoita Puhetta (Donate Speech) corpus [56].

All our ASR models adapted to the FinL2 data have low error rates. Such good ASR performance is a consequence of having a very limited word vocabulary in the dataset. Since all children were asked to utter the same set of 90 words, our models memorized it during training and easily predicted these words in the evaluation step. Our model trained for ASR exclusively achieves the best results: it provides 4.47% WER and 1.36% CER. The multitask system still provides high ASR performance, although not as good as the monotask model: the WER and the CER are 6.30% and 2.13%, respectively.

The results of speech rating experiments are summarized in Table 5. As in the Swedish experiments, the CER between the ASR output and the expected word does not solely comprise enough information needed to predict the pronunciation level, as a result, our decision tree underperformed in comparison to more complex wav2vec2-based solutions. The monotask model fine-tuned for classification provides the highest accuracy and recall, while the multitask one slightly outperforms it in terms of MAE.

V. MOBILE GAME APPLICATION

The goal of our project is to develop a mobile game application that facilitates pronunciation learning for young children. The idea is to observe a picture, listen to, and produce words and short phrases, initially with a model pronunciation and later without it. The speech task is combined with the visual game elements as follows. The players interact with multi-color shapes on the screen by simply pressing or dragging them. Each successful interaction triggers a voice sample replay to reinforce the word's auditory memorization and familiarity. After several interactions, with the actual number depending on the game's difficulty and the player's learning progress, a picture card representing the word will appear on the screen. The player can hear the word once again before a notification sound is played, indicating that the microphone is ready for recording. Players have a few seconds to say the word they heard and then receive immediate feedback on the success of their pronunciation. The feedback is simplified into a replay of their recorded attempt, followed by the correct pronunciation and 1 to 5 stars ratings. Collecting the stars is required for proceeding in the game, making the feedback an essential game element. The ratings could be personalized to give positive feedback if the player made several bad attempts.

The speaking task embedded in the game is aimed to tap speech, language, and reward processing in the brain. Speech production requires a neural representation for the to-be-produced speech sounds and words, as well as complex motor skills in speech production. Therefore, in the long term, repeated listening of speech sounds and practice of speech production with the game is expected to result in the establishment of the required representations, as well as improve and eventually automatize the motor skills needed in speech production. In addition, the perceive-and-produce design of the game is based on the fact that speech perception and production are closely connected in the brain [57], and the code of transformations between them is based on mappings between speech input and output that are learned in childhood during babbling [58]. In line with the models of the brain mechanisms of speech production [59], [60], the transformation between speech input and output through sensory-motor integration mechanisms is essential for long-term practice. Moreover, the feedback provided by the speech rating model of the game is expected to reinforce learning via the activation of the cortico-striatal reward system of the brain [61], [62].

The game is designed to help children learn by imitation; therefore, they are immersed in audio and visual illustrations without any textual explanation (see Figure 4 for some illustrations of the game user interface). It has playful characters that can appear on the screen to encourage players to practice. Guardians and teachers can control the use of the game with a username and password on the login screen

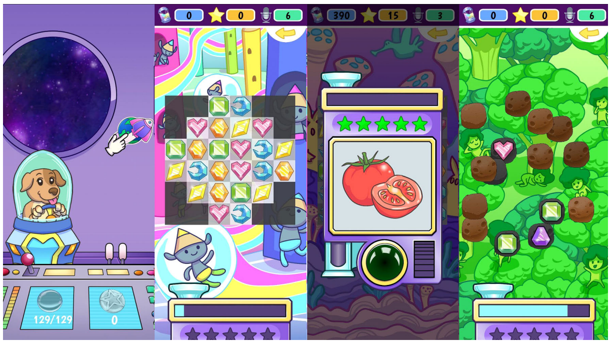


FIGURE 4. Screenshots of in-game experience.

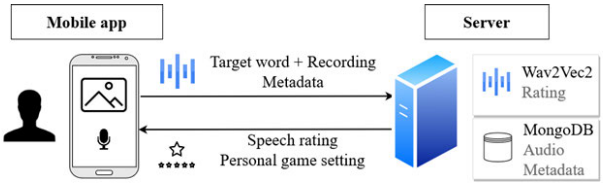


FIGURE 5. The game processing flowchart.

and can access details of their child's performance through a different platform.

There are several game modes available. The simplest one requires the player to press on random shapes appearing on the screen, and more challenging modes require a combination of hand and eye coordination skills, such as matching the color or catching the fast-moving shapes. There is also a classic memory-matching game, in which children need to match several face-down word cards from memory by flipping them one at a time. The child retention ability can also be tested with a silent mode, in which the child has to speak the word illustrated in the card without any audio cue. Similar to the rating system, the game modes and difficulties can also be personalized to motivate the child to engage with the game.

A. SPEECH SERVER

The speech server processes audio collected from the mobile app, provides speech rating, and collects other encrypted metadata (see Figure 5). As explained in Section III, the rating is mainly decided by wav2vec2 models, and the collected data are stored with the MongoDB program. Our server hardware and design allow us to handle a maximum of 20 children playing the game simultaneously without a significant latency in providing feedback.

Our server handles the feedback and game settings for all the user groups evaluated in this article. Different groups have different target languages, backgrounds, and, consequently, different game settings. Furthermore, the server contains our algorithms to analyze other data collected from the players. It can provide children with tailored lessons by instructing the game to adjust game mode, difficulty, and rating based on their performance.

VI. CONCLUSION

This work describes the steps taken to develop a CALL system for children practicing oral Swedish and Finnish with the use of the latest machine and deep learning methods. We demonstrated that a state-of-the-art deep learning model, wav2vec2, can be applied for speech assessment of specific target groups such as L2 speakers or children with SSD. Moreover, we implemented multitask learning of ASR and speech rating, which provided us with systems with considerably reduced overall latency and competitive performance compared to the constituent models. We also demonstrated the importance of understanding the automatic systems and used an input attribution algorithm to highlight the most influential parts of the recording, on which the model based its decision. This analysis revealed a weakness of the original system, namely that it learned to exploit some non-speech information present in the audio files. We addressed this problem by employing a VAD component to remove non-relevant information from the input. Finally, we integrated our best models into a speech-based mobile game application which we will next use for pedagogical studies.

Our future plans include the systematic pedagogical evaluation of the best models via the game app, and further data collection for other languages to broaden the capabilities of our game. Additionally, we wish to explore new ways of using the input attributions to provide detailed feedback to the experts and parents about the pronunciation problems of the children.

ACKNOWLEDGMENT

The computational resources were provided by Aalto ScienceIT.

REFERENCES

[1] J. K. Hartshorne, J. B. Tenenbaum, and S. Pinker, "A critical period for second language acquisition: Evidence from 2/3 million English speakers," *Cognition*, vol. 177, pp. 263–277, Aug. 2018.

[2] J. E. Flege, M. J. Munro, and I. R. A. MacKay, "Factors affecting strength of perceived foreign accent in a second language," *J. Acoust. Soc. Amer.*, vol. 97, no. 5, pp. 3125–3134, May 1995.

[3] Y. Wren, S. Harding, J. Goldbart, and S. Roulstone, "A systematic review and classification of interventions for speech-sound disorder in preschool children," *Int. J. Lang. Commun. Disorders*, vol. 53, no. 3, pp. 446–467, May 2018.

[4] A. Cummings, K. Giesbrecht, and J. Hallgrimson, "Intervention dose frequency: Phonological generalization is similar regardless of schedule," *Child Lang. Teaching Therapy*, vol. 37, no. 1, pp. 99–115, Feb. 2021.

[5] L. Aghlara and N. H. Tamjid, "The effect of digital games on Iranian children's vocabulary retention in foreign language acquisition," *Proc.-Social Behav. Sci.*, vol. 29, pp. 552–560, Jan. 2011.

[6] M.-H. Chen, W.-T. Tseng, and T.-Y. Hsiao, "The effectiveness of digital game-based vocabulary learning: A framework-based view of meta-analysis: The effectiveness of DGBL," *Brit. J. Educ. Technol.*, vol. 49, no. 1, pp. 69–77, Jan. 2018.

[7] S. J. Franciosi, "The effect of computer game-based learning on FL vocabulary transferability," *J. Educ. Technol. Soc.*, vol. 20, pp. 123–133, Jan. 2017.

[8] T.-Y. Liu and Y.-L. Chu, "Using ubiquitous games in an English listening and speaking course: Impact on learning outcomes and motivation," *Comput. Educ.*, vol. 55, no. 2, pp. 630–643, Sep. 2010.

[9] J. Sandberg, M. Maris, and P. Hoogendoorn, "The added value of a gaming context and intelligent adaptation for a mobile learning application for vocabulary learning," *Comput. Educ.*, vol. 76, pp. 119–130, Jul. 2014.

[10] Y.-L. Tsai and C.-C. Tsai, "Digital game-based second-language vocabulary learning and conditions of research designs: A meta-analysis study," *Comput. Educ.*, vol. 125, pp. 345–357, Oct. 2018.

[11] E. O. Acquah and H. T. Katz, "Digital game-based L2 learning outcomes for primary through high-school students: A systematic literature review," *Comput. Educ.*, vol. 143, Jan. 2020, Art. no. 103667.

[12] P. Li and Y.-J. Lan, "Digital language learning (DLL): Insights from behavior, cognition, and the brain," *Bilingualism, Lang. Cognition*, vol. 25, no. 3, pp. 361–378, May 2022.

[13] P. Li and Y.-J. Lan, "Understanding the interaction between technology and the learner: The case of DLL," *Bilingualism, Lang. Cognition*, vol. 25, no. 3, pp. 402–405, May 2022.

[14] D. Zou, Y. Huang, and H. Xie, "Digital game-based vocabulary learning: Where are we and where are we going?" *Comput. Assist. Lang. Learn.*, vol. 34, nos. 5–6, pp. 751–777, Jul. 2021.

[15] 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 Trans. Accessible Comput.*, vol. 14, no. 1, pp. 1–26, Mar. 2021.

[16] M. Nahum and D. Bavelier, "Chapter 10—Video games as rich environments to foster brain plasticity," in *Handbook of Clinical Neurology*, vol. 168, F. N. Ramsey and J. D. R. Millán, Eds. Amsterdam, The Netherlands: Elsevier, 2020, pp. 117–136.

[17] T.-C. Hsu, "Learning English with augmented reality: Do learning styles matter?" *Comput. Educ.*, vol. 106, pp. 137–149, Mar. 2017.

[18] R. Karhila, S. P. Ylinen, S. Enarvi, K. Palomaki, A. Nikulin, O. Rantula, V. Viitanen, K. Dhinakaran, A. R. M. Smolander, H. H. Kallio, and M. Uther, "SIAK—A game for foreign language pronunciation learning," in *Proc. Interspeech*, 2017, pp. 3429–3430.

[19] C. Tejedor-García, D. Escudero-Mancebo, V. Cardeñoso-Payo, and C. González-Ferreras, "Using challenges to enhance a learning game for pronunciation training of English as a second language," *IEEE Access*, vol. 8, pp. 74250–74266, 2020.

[20] M. Bashori, R. van Hout, H. Strik, and C. Cucchiarini, "'Look, I can speak correctly': Learning vocabulary and pronunciation through websites equipped with automatic speech recognition technology," *Comput. Assist. Lang. Learn.*, pp. 1–29, May 2022.

[21] C. Cucchiarini, A. Neri, and H. Strik, "Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback," *Speech Commun.*, vol. 51, no. 10, pp. 853–863, Oct. 2009.

[22] E. M. Golonka, A. R. Bowles, V. M. Frank, D. L. Richardson, and S. Freynik, "Technologies for foreign language learning: A review of technology types and their effectiveness," *Comput. Assist. Lang. Learn.*, vol. 27, no. 1, pp. 70–105, Feb. 2014.

[23] 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 Lang.*, vol. 230, Jul. 2022, Art. no. 105124.

[24] J. McKechnie, B. Ahmed, R. Gutierrez-Osuna, P. Monroe, P. McCabe, and K. J. Ballard, "Automated speech analysis tools for children's speech production: A systematic literature review," *Int. J. Speech-Lang. Pathol.*, vol. 20, no. 6, pp. 583–598, Oct. 2018.

[25] S. M. Witt, "Use of speech recognition in computer-assisted language learning," Ph.D. dissertation, Dept. Eng., Univ. Cambridge, Cambridge, U.K., 2000.

[26] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.*, vol. 30, nos. 2–3, pp. 95–108, Feb. 2000.

[27] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Commun.*, vol. 67, pp. 154–166, Mar. 2015.

[28] J. Shi, N. Huo, and Q. Jin, "Context-aware goodness of pronunciation for computer-assisted pronunciation training," in *Proc. Interspeech*, Oct. 2020, pp. 3057–3061.

[29] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Amer.*, vol. 105, pp. 1455–1468, Mar. 1999.

[30] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12449–12460.

[31] R. Al-Ghezi, Y. Getman, A. Rouhe, R. Hildén, and M. Kurimo, "Self-supervised end-to-end ASR for low resource L2 Swedish," in *Proc. Interspeech*, Aug. 2021, pp. 1429–1433.

[32] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3319–3328.

[33] S. Strömbergsson, K. Holm, J. Edlund, T. Lagerberg, and A. McAllister, "Audience response system-based evaluation of intelligibility of children's connected speech—Validity, reliability and listener differences," *J. Commun. Disorders*, vol. 87, Sep. 2020, Art. no. 106037.

[34] C. Blumenthal and I. L. Hammarström, *LINUS. LINköpingsUnderSökningen: Ett Fonologiskt Testmaterial Från 3 År*. Linköping, Sweden: Linköping Univ. Electronic Press, 2014.

[35] T. B. Lagerberg, L. Hartelius, J. Å. Johnels, A.-K. Ahlman, A. Börjesson, and C. Persson, "Swedish test of intelligibility for children (STI-CH)—Validity and reliability of a computer-mediated single word intelligibility test for children," *Clin. Linguistics Phonetics*, vol. 29, no. 3, pp. 201–215, Mar. 2015.

[36] M. Uther, A. R. Smolander, K. Junttila, M. Kurimo, R. Karhila, S. Enarvi, and S. Ylinen, "User experiences from L2 children using a speech learning application," *Adv. Hum. Comput. Interact.*, vol. 6, Nov. 2018, Art. no. 7345397.

[37] R. Karhila, A.-R. Smolander, S. Ylinen, and M. Kurimo, "Transparent pronunciation scoring using articulatorily weighted phoneme edit distance," in *Proc. Interspeech*, Sep. 2019, pp. 1866–1870.

[38] S. Ylinen, A.-R. Smolander, R. Karhila, S. Kakouros, J. Lipsanen, M. Huotilainen, and M. Kurimo, "The effects of a digital articulatory game on the ability to perceive speech-sound contrasts in another language," *Frontiers Educ.*, vol. 6, May 2021, Art. no. 612457.

[39] N. Vanhainen and G. Salvi, "Free acoustic models for large vocabulary continuous speech recognition in Swedish," in *Proc. Lang. Resour. Eval. Conf. (LREC)*, 2014, pp. 388–392.

[40] A., Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF_STAR children's speech corpus," in *Proc. INTERSPEECH*, Lisbon, Portugal, Sep. 2005, pp. 2761–2764.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 6000–6010.

[42] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[43] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[44] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, and J. C. Chou, "SpeechBrain: A general-purpose speech toolkit," 2021, *arXiv:2106.04624*.

[45] L. W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, Jun. 2023, pp. 1–5.

[46] T. Grósz, D. Porjazovski, Y. Getman, S. Kadiri, and M. Kurimo, "wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering," in *Proc. 30th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2022, pp. 7026–7029.

[47] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. Interspeech*, Sep. 2009, pp. 312–315.

[48] Y. Getman, R. Al-Ghezi, K. Voskoboinik, T. Grósz, M. Kurimo, G. Salvi, T. Svendsen, and S. Strömbergsson, "wav2vec2-based speech rating system for children with speech sound disorder," in *Proc. Interspeech*, Sep. 2022, pp. 3618–3622.

[49] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze, "Evaluating saliency map explanations for convolutional neural networks: A user study," in *Proc. 25th Int. Conf. Intell. User Interfaces*, New York, NY, USA, Mar. 2020, pp. 275–285.

[50] G. Novakovsky, N. Dexter, M. W. Libbrecht, W. W. Wasserman, and S. Mostafavi, ''Obtaining genetics insights from deep learning via explainable artificial intelligence,'' *Nature Rev. Genet.*, vol. 24, no. 2, pp. 125–137, Feb. 2023.

[51] A. Ghorbani, A. Abid, and J. Zou, ''Interpretation of neural networks is fragile,'' in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 3681–3688.

[52] C. K. Yeh, C. Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar, ''On the (in)fidelity and sensitivity of explanations,'' in *Proc. Neural Inf. Process. Syst.*, 2019, pp. 1–12.

[53] H. Bredin and A. Laurent, ''End-to-end speaker segmentation for overlap-aware resegmentation,'' in *Proc. Interspeech*, Aug. 2021, pp. 3111–3115.

[54] M. Eskenazi, ''An overview of spoken language technology for education,'' *Speech Commun.*, vol. 51, no. 10, pp. 832–844, Oct. 2009.

[55] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, ''VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation,'' in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 993–1003.

[56] A. Moisio, D. Porjazovski, A. Rouhe, Y. Getman, A. Virkkunen, R. AlGhezi, M. Lennes, T. Grósz, K. Lindén, and M. Kurimo, ''*Lahjoita puhetta*: A large-scale corpus of spoken Finnish with some benchmarks,'' *Lang. Resour. Eval.*, pp. 1–33, Aug. 2022.

[57] G. Hickok, J. Houde, and F. Rong, ''Sensorimotor integration in speech processing: Computational basis and neural organization,'' *Neuron*, vol. 69, pp. 407–422, Feb. 2011.

[58] F. H. Guenther, ''Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production,'' *Psychol. Rev.*, vol. 102, no. 3, pp. 594–621, 1995.

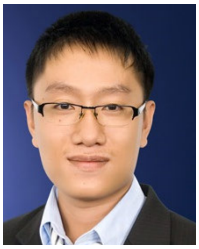
[59] X. Tian, ''Mental imagery of speech and movement implicates the dynamics of internal forward models,'' *Frontiers Psychol.*, vol. 1, p. 7029, Oct. 2010.

[60] G. Hickok, ''Computational neuroanatomy of speech production,'' *Nature Rev. Neurosci.*, vol. 13, no. 2, pp. 135–145, Feb. 2012.

[61] D. Shohamy, C. E. Myers, S. Grossman, J. Sage, M. A. Gluck, and R. A. Poldrack, ''Cortico-striatal contributions to feedback-based learning: Converging data from neuroimaging and neuropsychology,'' *Brain*, vol. 127, pp. 851–859, Apr. 2004.

[62] J. Mishra, J. A. Anguera, and A. Gazzaley, ''Video games for neurocognitive optimization,'' *Neuron*, vol. 90, no. 2, pp. 214–218, Apr. 2016.

YAROSLAV GETMAN received the M.Sc. degree in signal, speech and language processing from Aalto University, Finland, in 2021. From 2020 to 2021, he was a Research Assistant. Since 2022, he has been a Ph.D. Researcher with the Department of Signal Processing and Acoustics, Aalto University. His current research interests include self-supervised learning in automatic speech recognition and other speech-related tasks.



NHAN PHAN is currently pursuing the M.Sc. degree in computer, communication and information sciences with Aalto University, Finland. He joined the Department of Signal Processing and Acoustics, Aalto University, as a Research Assistant, in 2022. His current research interest includes developing practical CAPT applications.

RAGHEB AL-GHEZI (Graduate Student Member, IEEE) received the M.Sc. degree in human language technology from The University of Arizona, in 2019. Before that, he worked in educational technology and learning analytics for various international organizations. He is currently a Ph.D. Researcher with the Department of Information and Communications Engineering, Aalto University, where his research focuses on the application of natural language processing (NLP) and automatic speech recognition (ASR) in the field of automatic speaking assessment.

EKATERINA VOSKOBOINIK received the master's degree in signal, speech, and language processing from Aalto University, in 2019, where she is currently pursuing the Ph.D. degree with the Department of Information and Communications Engineering. Her research interest includes automatic speech assessment.

MITTUL SINGH received the Ph.D. degree in computer science from Saarland University, Germany. He is currently working as a Senior AI scientist at an AI startup SiloAI, developing speech and language solutions for their clients. He is simultaneously engages in collaborative endeavors with the Speech Recognition Group at Aalto University. He is passionate about developing deep learning solutions for restricted data scenarios to reach new markets and customers.

TAMÁS GRÓSZ received the Ph.D. degree in speech recognition from the University of Szeged, in 2018. From 2017 to 2018, he was an Assistant Research Fellow with the Research Group on Artificial Intelligence, Hungarian Academy of Sciences. From 2018 to 2019, he was a Senior Lecturer with the Department of Computer Algorithms and Artificial Intelligence, University of Szeged. He is currently a Research Fellow with the Department of Information and Communications Engineering, Aalto University. His current research interests include automatic speech recognition, deep learning, and computational paralinguistics.

MIKKO KURIMO received the D.Sc. (Ph.D.) degree in technology in computer science from the Helsinki University of Technology, Espoo, Finland, in 1997. He is currently a Professor of speech and language processing with the Department of Information and Communications Engineering, Aalto University. His research interests include speech recognition, machine learning, and language modeling.

GIAMPIERO SALVI (Member, IEEE) received the M.Sc. degree in electronic engineering from Universitã la Sapienza, Italy, and the Ph.D. degree in computer science from the KTH Royal Institute of Technology, Sweden. He was a Postdoctoral Fellow with the Institute of Systems and Robotics, IST, Portugal, and one of the co-founders of the company SynFace AB, Sweden. He is currently a Professor with the Department of Electronic Systems, Norwegian University of Science and Technology (NTNU), Norway, and an Associate Professor with the Department of Electrical Engineering and Computer Science, KTH Royal Institute of Technology. His main research interests include machine learning, speech technology, and cognitive systems.

TORBJØRN SVENDSEN (Life Senior Member, IEEE) received the Siv.Ing. (M.Sc.) and Dr.Ing. degrees from the Norwegian Institute of Technology (NTH), in 1980 and 1985, respectively. He was a Research Scientist with SINTEF before joining NTH, later the Norwegian University of Science and Technology (NTNU), in 1988. He is currently a Professor with the Department of Electronics and Telecommunications, NTNU. His research interests include automatic speech recognition, machine learning, and speech analysis and modeling.

SOFIA STRÖMBERGSSON received the degree from Lund University, Sweden, in 2007, and the Ph.D. degree in speech communication from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2014. She is a certified Speech-Language Pathologist. She is currently an Associate Professor with the Division of Speech-Language Pathology, CLINTEC, Karolinska Institutet. Her research interests include speech and language disorders in children and different ways of assessing speech and language during acquisition.

ANNA SMOLANDER received the B.Sc. degree in language technology, in 2019, and the M.Sc. degree in economy and business administration, in 2022. Since 2016, she has been involved in developing and assisting research projects in digital learning. She worked on speech data collection for speech recognition systems combined with EEG measurements and behavioral tests and gained experience in ethical protocols, GDPR policies, and research permission processes in Finland. Most of her studies concentrate on children's learning and voice. Still, there is also some research with vulnerable/minority target groups (e.g., Arabic-speaking illiterate adults) and dyslexic babies and children.

SARI YLINEN received the M.A. degree in linguistics and language pedagogy, including teacher's qualification, from the University of Jyväskylä, in 2000, and the Ph.D. degree in psychology from the University of Helsinki, in 2006. She specializes in cognitive neuroscience (Title of Docent corresponding to Adjunct Professorship with the University of Helsinki). She is currently an Associate Professor of speech-language pathology with Tampere University, Finland. Her research interests include language learning and acquisition, learning and technology, as well as neural aspects of language learning.