wav2vec2-based Speech Rating System for Children with Speech Sound Disorder

Yaroslav Getman1*, Ragheb Al-Ghezi1*, Ekaterina Voskoboinik1, Tamás Grósz1, Mikko Kurimo2, Giampiero Salvi2, Torbjörn Svendsen2, Sofia Strömbergsson3

1Department of Signal Processing and Acoustics, Aalto University, Finland
2Signal processing, Norwegian University of Science and Technology, Norway
3Department of Clinical Science, Intervention and Technology, Karolinska Institutet, Sweden

Abstract

Speaking is a fundamental way of communication, developed at a young age. Unfortunately, some children with speech sound disorder struggle to acquire this skill, hindering their ability to communicate efficiently. Speech therapies, which could aid these children in speech acquisition, greatly rely on speech practice trials and accurate feedback about their pronunciations. To enable home therapy and lessen the burden on speech-language pathologists, we need a highly accurate and automatic way of assessing the quality of speech uttered by young children. Our work focuses on exploring the applicability of state-of-the-art self-supervised, deep acoustic models, mainly wav2vec2, for this task. The empirical results highlight that these self-supervised models are superior to traditional approaches and close the gap between machine and human performance.

Index Terms: speech assessment, goodness of pronunciation, children speech, ASR, wav2vec2

1. Introduction

Speaking is a fundamental skill that enables humans to communicate efficiently. Typically, speech develops at a young age and is further tuned as the speaker gets older. However, for children with speech sound disorder (SSD), speech acquisition takes a different path. SSDs are characterized by developmentally inadequate speech production, often with negative impact on intelligibility [1]. It involves an increased risk of having associated difficulties with language and literacy [2], and for some, a risk of a negative social impact. SSD is a high-prevalence condition in preschool-aged children, and in clinical settings, children with SSD constitute a large portion of speech-language pathologists’ caseloads [3]. Clinical intervention serves to minimize the negative consequences of the SSD, most often focusing on increasing speech production accuracy [3]. Although multiple intervention approaches exist, many share as central ingredients a large number of speech practice trials for the child, and feedback regarding performance. Home-training is already an important ingredient in SSD intervention [4], and Covid-19 has further highlighted the need of intervention approaches that are not dependent on physical co-presence. An engaging speech game, with an embedded automatic rating system, may encourage the child to achieve the high-enough training dose and frequency needed to expedite change. Providing that such a speech game can give reliable feedback concerning the child’s speech production performance, it would constitute an attractive complement to traditional SSD intervention.

Computer-Assisted Pronunciation Training (CAPT) systems rely on a component that accurately predicts the Goodness of Pronunciation (GOP) [5], [6]. Traditional GOP solutions consist of a simple pipeline [7]. The first element is an ASR system that supplies the acoustic scores, which are then used by an aligner. The aligner decides which phones were uttered at a given time and also calculates the acoustic scores per phoneme. Lastly, the scoring module uses the phonetic scores to predict the GOP for a given speech sample. Recent works, aimed to develop GOP solutions for children, demonstrated that using the ASR log posterior probabilities to train classifiers is a superior solution compared to the traditional GOP pipeline method [8]. In line with those findings, we decided to test end-to-end solutions that rely on wav2vec2 models pre-trained on ASR tasks to estimate the level of the pronunciation. A huge advantage of our solution is that only a limited amount of annotated, in-domain data is needed to fine-tune the model.

Regardless of the approach chosen for building a GOP system for children’s speech, the main challenge at hand is the same: it is required to have good quality representation of children’s speech. It is not possible to use an adult model as is since children’s speech differs from adult speech both in its acoustic and linguistic content. First, children’s speech differs from adults’ in F0, speaking rate and formant frequencies [9], so adults’ ASR systems and training data are of limited use even after signal modifications [10]. Second, because all children grow and acquire speaking skills at an individual pace there is a large acoustic [11] and linguistic [12] variation within children’s speech itself. Lastly, there exists a shortage of publicly-available corpora [13] since the organization of large-scale recordings and collections of good quality children’s speech data is an exhaustive task due to many practical reasons.

The current work brings the following contributions. First, to the best of our knowledge, this is the first work that focuses on rating Swedish speech of children with SSDs. Second, this work demonstrates the applicability of using contextual features derived from self-supervised models for rating children’s speech.

2. Data

2.1. Swedish PF.STAR

The PF.STAR dataset [14] includes more than 60 hours of speech from British, German, Swedish and Italian children. In this study, we use the Swedish part of the data set. It contains recordings of 99 children aged 4-8 years uttering phrases of dif-

* The first two authors have equal contribution.
Table 1: Rating categories used in the SweSSD corpus and their description.

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>not at all identifiable as the target word</td>
</tr>
<tr>
<td>2</td>
<td>hard to identify as the target word</td>
</tr>
<tr>
<td>3</td>
<td>slight phonemic error</td>
</tr>
<tr>
<td>4</td>
<td>subphonemic error or &quot;unexpected variant&quot;</td>
</tr>
<tr>
<td>5</td>
<td>prototypical/adult-like/correct</td>
</tr>
</tbody>
</table>

These pre-trained acoustic representations can be fine-tuned for downstream tasks, such as speech recognition and general audio classification. For ASR purposes, they can be applied in an end-to-end ASR pipeline in which they are appended with a randomly-initialized classification head to predict transcripts directly using a connectionist temporal classification (CTC) loss. Standard wav2vec2 ASR fine-tuning pipeline [18], [20] incorporates freezing the feature extraction CNN layers because their weights are expected to be sufficiently trained during the pre-training step, eliminating the need for further adjustment.

3.2. Speech rating system
To automatically assess the speech samples, we built multiple solutions. The baseline GOP approach simply utilized the character error rate (CER) of the utterances instead of the phoneme posterior and employed decision trees to find the decision boundaries between classes (W2V2 CER+DT). As an alternative solution, we utilized the wav2vec2 model to produce the log probabilities for all possible characters and then used a convolutional recurrent deep neural network (CRDNN) to transform them into categorical labels (W2V2 logp). These models combine the advantages of convolutional and recurrent neurons and became a popular choice in paralinguistics and ASR [21]. To showcase the benefits of using the wav2vec2 model, we also trained CRDNNs that classify the examples based on their MFCC features (MFCC+CRDNN). Another baseline (MFCC+GMM-HMM+DT) is based on a GMM-HMM model originally trained on the NST Swedish database [22] and adapted to the target data. In this case, we use a confidence score given by the difference of log likelihoods when running word as opposed to phoneme recognition with the same acoustic models. Of the many models available from [22], triphone HMMs with 64 Gaussian components per state trained on MFCC features were used in this study. Decision trees were used to map scores to rating classes.

Beyond these baseline systems, we also wanted to fully utilize the wav2vec2 model. To achieve this, two separate approaches were tested. The first one is simply wav2vec2 CTC model further fine-tuned for the classification task (W2V2). As a lightweight alternative to avoid the costly fine-tuning procedure, we also used the wav2vec2 model as a feature extractor and replaced the MFCC inputs of the CRDNN with the embeddings produced by the last hidden layer (W2V2 emb). The advantage
of the latter approach is that the wav2vec2 is static, and we need to run the inference over the whole data only once. Furthermore, the training time of the CRDNN is negligible compared to wav2vec2 fine-tuning.

4. wav2vec2 children ASR

In this work, we used various publicly available wav2vec2-Large (317M parameters) models\(^1\). The ASR systems were evaluated in terms of word and character error rate (WER and CER). We decided to operate on the character level rather than the phoneme level, since the latter option would have required a hand-crafted lexicon, preferably with multiple pronunciations per word. Additionally, a recent study [23] showed that in low-resource scenarios, output units like characters lead to better performance than fine-grained phonemes. Table 2 summarizes their general properties, as well as the performance on Swedish children ASR after fine-tuning on the SweSSD data.

Table 2: ASR results for wav2vec2 models adapted to children speech. 3. column indicates whether the model was only pre-trained on unlabeled speech, or it was also fine-tuned on labeled adult Swedish data before adapting to Swedish children ASR. 4. column shows whether the CTC head weights of an adult ASR model were adjusted to children ASR along with the rest of the network, or they were initialized from scratch before fine-tuning on children speech.

<table>
<thead>
<tr>
<th>W2V2 Model</th>
<th>Pretrain size, h</th>
<th>Fine-tuned on adult speech</th>
<th>Adult LM head</th>
<th>WER, %</th>
<th>CER, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>KB_VoxPopuli</td>
<td>5.5K</td>
<td>✓</td>
<td>✓</td>
<td>42.33</td>
<td>13.89</td>
</tr>
<tr>
<td>KB_VoxPopuli</td>
<td>11.5K</td>
<td>✓</td>
<td></td>
<td>71.80</td>
<td>26.77</td>
</tr>
<tr>
<td>KB_VoxRex</td>
<td></td>
<td></td>
<td></td>
<td>37.91</td>
<td>14.75</td>
</tr>
<tr>
<td>FB_VoxPopuli</td>
<td>29.9K</td>
<td></td>
<td>✓</td>
<td>39.31</td>
<td>13.20</td>
</tr>
<tr>
<td>FB_VoxPopuli</td>
<td></td>
<td></td>
<td></td>
<td>76.30</td>
<td>29.20</td>
</tr>
</tbody>
</table>

The monolingual models, KB_VoxPopuli and KB_VoxRex, were pre-trained by KBLab\(^1\) (the data lab at KB, the National Library of Sweden) on Swedish local radio broadcasts. KB_VoxPopuli was also pre-trained on European Parliament plenary session recordings from the VoxPopuli corpus [24], while the KB_VoxRex pretraining data included audio books and other speech from KB collections. Transcribed speech data from Nordisk Språkteknologi (NST), Swedish Common Voice [25] and Swedish local radio were used in the fine-tuned version of KB_VoxRex. The multilingual model, FB_VoxPopuli, was pre-trained on VoxPopuli recordings in two North Germanic languages: Swedish (16.3K hours), and Danish (13.6K hours).

According to the results of the ASR experiments, the models benefit much from intermediate fine-tuning on adult speech. In addition, our experiments proved that preserving the language modeling head weights from the fine-tuned adult ASR model and adjusting them during fine-tuning for children ASR gives better performance than training them from scratch.

The most accurate model achieved 35.31%/13.20% WER/CER on the SweSSD test set and was applied in our speech rating experiments. In addition, we attempted to further improve the ASR results of this system by incorporating an additional intermediate step in our training pipeline: before fine-tuning the model on the target data (SweSSD), we tuned it on the PF_STAR corpus. The derived model yielded a WER of 34.36% and CER of 12.40% on the SweSSD test set.

5. Speech Rating Experiments

First, we investigated the wav2vec2 model used in this study. To train the wav2vec2-based speech rating systems, we utilized the ASR models described in Section 3.1 with a few minor modifications. The model CTC head was replaced with a projection layer, followed by an average pooling step before the classification layer. The resulted network was optimized by minimizing the cross-entropy loss, with the CNN part kept frozen.

As a CRDNN, we opted for a relatively simple architecture; the input is first processed by two convolutional layers, which are convolved along the time axis. After the CNN part, a bi-LSTM layer summarizes the utterance-level information. Before the final softmax layer, an additional feed-forward layer transforms the embeddings produced by the recurrent layer.

One critical part of all studies is the evaluation part, where appropriate steps must be used to showcase the real performance of the presented models. In our case, we need to consider several factors. First of all, we have a limited amount of data. Thus, we choose cross-validation (CV) to evaluate solutions. We opted for a 6-fold CV, this way, roughly 1000 randomly selected examples were part of each fold, with no overlap between folds. To train and evaluate the models 4 folds were used as training data, 1 fold as development data for hyperparameter tuning and 1 fold for testing. In this CV, each fold was used exactly once for testing and development (and not by the same model), and to get the final evaluation values we micro-aggregated the confusion matrices of the 6 models.

The next problem is the unbalanced nature of the data. As the best rating is severely over-represented (approx. 47% of the data belongs to this class, see Figure 1), the accuracy (ACC) metric could result in misleading results. Unweighted average recall (UAR) is a popular choice for scenarios like this [26], and it could give us a better understanding of the general performance. Additionally, we report two other metrics. The mean absolute error (MAE) showcases the expected difference between the predicted and annotated ratings, which could be considered as the expected error value between the predicted and the annotated level of speech. Lastly, the per-annotation accuracy (PACC) disregards mistakes between the top two levels (4 and 5), which, by definition are extremely hard to be separated.

5.1. Speech rating evaluations

Our first experiments were aimed to determine which wav2vec2 model is better suited for the rating task. Our initial experiments revealed that KB_VoxRex-based systems, when fine-tuned purely on in-domain data, consistently achieved better development results compared to the other models. One possible explanation might be that using the PF_STAR corpus unintentionally made the model robust against mispronunciation errors, which in turn degraded its ability to detect them and predict the pronunciation levels accurately. Another explanation is that PF_STAR has a different speech domain than SweSSD. The first consists of spontaneous and imitated speech while the latter is one-word and read-aloud speech.

Table 3 summarizes the results of different speech rating approaches using the best wav2vec2 ASR model from Table 2. The first thing that we can notice is that the simple MFCC-based solutions achieved the worst scores in all evaluation metrics. This result confirms the necessity of the ASR component for

\(\text{^1}https://huggingface.co/KBLab/wav2vec2-large-voxrex-swedish
\(\text{^2}https://github.com/facebookresearch/voxpopuli-pre-trained-models
\(\text{^3}https://www.kb.se/en-english/research-collaboration/kblab.html

3620
Table 3: Results of the speech rating solutions.

<table>
<thead>
<tr>
<th>System</th>
<th>ACC, %</th>
<th>PACC, %</th>
<th>UAR, %</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC+GMM-HMM+DT</td>
<td>42.3</td>
<td>59.5</td>
<td>30.4</td>
<td>0.92</td>
</tr>
<tr>
<td>MFCC+CRDNN</td>
<td>47.4</td>
<td>58.5</td>
<td>35.1</td>
<td>0.91</td>
</tr>
<tr>
<td>W2V2 logp+CRDNN</td>
<td>57.0</td>
<td>68.7</td>
<td>37.8</td>
<td>0.61</td>
</tr>
<tr>
<td>W2V2 emb+CRDNN</td>
<td>57.0</td>
<td>70.6</td>
<td>40.0</td>
<td>0.63</td>
</tr>
<tr>
<td>W2V2 CER+DT</td>
<td>60.4</td>
<td>71.9</td>
<td>43.6</td>
<td>0.51</td>
</tr>
<tr>
<td>Human (50% data)</td>
<td>65.8</td>
<td>72.3</td>
<td>66.7</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Figure 2: Recalls of each system per categories.

Figure 3: Confusion matrix of the best model (W2V2).

Building well-performing systems. Next, we can compare the different wav2vec2-based solutions. As expected, fine-tuning the whole model for the classification task worked the best, outperforming all other approaches. Interestingly, the second-best solution turned out to be the simplest one; using only the CER of each recording to estimate the pronunciation category. The remaining two rows showcase that the embeddings of wav2vec2 contain more information and thus are better suited for this task than the log probabilities. The best model (W2V2) seems to have low accuracy, and an even lower UAR reveals that it greatly overfitted to the overrepresented classes. To understand the difficulty of this task and provide an upper bound on the performance, we asked the original annotator to re-evaluate a randomly chosen subset (approx. 20%) of the data, six months after the original annotating process. This re-evaluation confirmed that the task is extremely difficult even for a human expert. We can see that the gap in accuracy is about 5%, which drops to only 1% when we merge the two hardest to separate classes (PACC). While these observations could suggest that we almost reached the human level performance with W2V2, the UAR and MAE metrics paint a different picture. According to the recall and average error, the human annotator is still considerably better at the rating task than the automatic systems.

The overall performance values might suggest that neither humans nor machines are good at this task. Contradicting this observation, the low MAE values indicate that a vast majority of the mistakes are just minor ones. Upon closer look, we observed that only 11.7% of the human errors are serious ones, where the difference between the original and re-annotated levels are more than one. Performing the same analysis on our best model (W2V2) revealed that it was also making predominantly minor errors; about 30% of its mistakes belong to the severe category.

Next, we took a closer look at the performances per category. Figure 2 displays a more detailed picture about the strengths and weaknesses of the different approaches. The first thing to notice is that all systems performed extremely well for level 5, outperforming even the human level. On the other hand, the annotator had a superior performance in recognizing level 4, which proved to be the most challenging class for all systems. Overall, the human expert proved to be the best for levels 1 and 4, while some models managed to reach the human performance in recognizing levels 2 and 3. W2V2 performs quite well compared to the other solutions but lags behind in recognizing level 1, while the CER-based decision trees proved to be most capable of recognizing the two extreme levels (1 and 5), but failed to properly separate the intermediate ones. Interestingly, using the log probabilities leads to better performance in case of the first two levels compared to the embeddings, which are more helpful in discriminating between the higher levels (3, 4 and 5).

Lastly, we investigated the performance of the best automatic rating solution. Figure 3 depicts its confusion matrix. As mentioned before, a majority (~70%) of the errors are made between consecutive levels. Confusing very different ratings is quite rare. Additionally, we can see that the model has difficulties separating level 4 from 5 and level 1 from 2, which is expected based on their definition in Table 1. Overall, we can say that despite its shortcomings, our solution manages to reach good performance, although not yet on par with a human expert.

6. Conclusions

This work demonstrated that wav2vec2-based approaches are viable for predicting the pronunciation level of children with speech sound disorders in comparison to other techniques. An additional advantage of our proposed solution is the fact that it requires only a limited amount of supervised data, which is quite hard to acquire in this field. Even though the best model fails to reach the performance level of a human expert, we are confident that it would be a valuable tool for non-expert parents during home treatment. To confirm this, we plan to investigate expert and non-expert human performances using this dataset. Furthermore, our approach can also be applied in other settings such as second language learning for children.

7. Acknowledgments

The computational resources were provided by Aalto ScienceIT. This work was supported by NordForsk through the funding to Technology-enhanced foreign and second-language learning of Nordic languages, project number 103893.
8. References


