# Error-preserving Automatic Speech Recognition of Young English Learners' Language

**Anonymous ACL submission**

## Abstract

One of the central skills that language learners need to practice is speaking the language. Currently, students in school do not get enough speaking opportunities and lack conversational practice. Recent advances in speech technology and natural language processing allow for the creation of novel tools to practice their speaking skills. In this work, we tackle the first component of such a pipeline, namely, the automated speech recognition module (ASR), which faces a number of challenges: first, state-of-the-art ASR models are often trained on adult read-aloud data by native speakers and do not transfer well to young language learners' speech. Second, most ASR systems contain a powerful language model, which smooths out mistakes made by the speakers. To give corrective feedback, which is a crucial part of language learning, the ASR systems in our setting need to preserve the mistakes made by the language learners. In this work, we build an ASR system that satisfies these requirements: it works on spontaneous speech by young language learners and preserves their mistakes. For this, we collected a corpus containing around 85 hours of English audio spoken by Swiss learners from grades 4 to 6 on different language learning tasks, which we used to train an ASR model. Our experiments show that our model benefits from direct fine-tuning on children's voices and has a much higher error preservation rate than other models.

## 1 Introduction

Speaking is one of the core competencies to be developed in foreign language classes and the second most widely used skill in everyday-life communication (Hedge, 2001). For students to successfully acquire speaking competencies, they must be trained from an early stage in the language learning process and in a systematic manner. However, speech production is a highly complex process that is often not addressed adequately in classrooms. The main issue is that students often do not get enough speaking opportunities (Kleinschroth and Oldham, 2014; Grimm et al., 2015), and lack extended conversational practice (Pfenninger and Lendl, 2017). The recent advancements in both speech processing (Malik et al., 2021), and conversational dialogue systems (Deriu et al., 2021; Ni et al., 2023) provide an opportunity to increase the speaking practice of language learners using automated tools.

The work presented in this paper is part of a larger effort to develop an interactive, voice-driven chatbot with which learners can practice their interactive speaking skills. The bot is designed as a conversation partner that adjusts to the skill level and interests of the students and provides corrective feedback to support their language development.

One key issue is the automated speech recognition (ASR) module, which transcribes the utterances of the language learners into text to be processed in downstream tasks (e.g., speaker-error analysis, dialogue systems, inter alia). The focus of this work is to adapt the ASR module to handle children's speech in a language learning environment. The core challenge for the ASR system in this setting is not only to transcribe the speech but to make sure that the mistakes made by the language learners are transcribed faithfully. This is needed to provide language learners with corrective feedback, which is a key component of second and foreign language development. It prompts learners to notice errors and is likely to lead to utterance repair, which, in turn, facilitates language development (Ellis, 2021). Our investigations showed that current state-of-the-art ASR models tend to correct the speakers' mistakes, which renders giving corrective feedback impossible.

The second challenge for the ASR system is handling spontaneous children's speech since most of these systems are trained on adult read-aloud error-free corpora recorded by native speakers (Panayotov et al., 2015; Ardila et al., 2020). Children's speech, especially spontaneous speech of language learners, differs significantly from read-aloud speech of native adult speakers (Shivakumar and Georgiou, 2020). Children's' speech has a different range of sound frequencies (Potamianos and Narayanan, 2003), a high within-subjects variability (Gerosa et al., 2006) and a high inter-speaker variability in different age groups (Lee et al., 1999).

These challenges yield three research questions, which we address in this work:

1. How can we measure error preservation, i.e. the "verbatimness" of an ASR transcript?
2. How well do current pre-trained ASR systems perform on learners' spontaneous English productions, with respect to error preservation and in general?
3. Does fine-tuning pre-trained systems with data from young learners lead to improved error preservation in the ASR transcripts?

**Contributions** In order to answer these questions, we first collected a dataset of young Swiss learners speaking English, consisting of 85 hours of recordings corresponding to 45'004 individual utterances by 327 distinct speakers. We subsequently created verbatim transcriptions of these recordings, in which learner errors are annotated using specific symbols. This dataset can be accessed as described in Section 3.4.2 below. We next developed a metric for error preservation, called *Word-Based Error Preservation Rate (WEPR)*, which takes into account only those reference words that contain an error annotation. Using WEPR and standard ASR metrics, we compared 7 pre-trained ASR systems with a custom fine-tuned model. Our results show that a) there are large differences between the pre-trained models both in terms of standard metrics and in terms of WEPR and b) fine-tuning significantly improves error preservation of learners' speech.

## 2 Related Work

**Children's Speech Corpora.** Corpora of children's speech can be divided into two types: i) corpora for native speaking children intended for building virtual tutors for non-language subjects, ii) corpora for young language learners that support building virtual tutors for language learning.

The MyST Children's Speech Corpus (Pradhan et al., 2016; Ward et al., 2019) contains 499 hours of conversational speech (out of which 233 hours are manually transcribed) for a virtual tutor for science topics targeted at young English native speakers. The OGI Kids' Speech Corpus (Shobaki et al., 2000) contains spontaneous speech from 1100 American children from kindergarten through grade 10, mainly consisting of scripted speech in the form of words and utterances, and a small sample of spontaneous speech. The AusKidTalk corpus (Ahmed et al., 2021) contains speech from Australian children ages 3 to 12 consisting of single words, utterances, and narrative speech. Other, smaller, datasets of native speaking children are available for different purposes such as read-aloud support (Eskenazi, 1996) or general analysis of English children's speech (Lee et al., 1999; Hagen et al., 2003). For German, the KidsTalk corpus (Rumberg et al., 2022) contains 25 hours of transcribed continuous speech from children aged 3 to 11. All these corpora are devised for settings with native speakers.

For language learners, there are far fewer datasets of children's speech. The TLT-school collection (Gretter et al., 2020) aims at assessing the proficiency of 9- to 16-year old Italian native speakers in English and German. TLT was recorded with a pool of 3000 students, resulting in approximately 275h of English and 265h of German data, out of which 16h for English and 8h for German have been transcribed. The corpus closest to our dataset is the CALL corpus (Baur et al., 2018), consisting of English utterances by Swiss German second and third year learners, where the task is to label the correctness of each utterance. In total, the corpus contains 38k utterances of students interacting with an online dialogue system, where they receive various prompts to produce speech. Across a series of shared tasks, subsets of around 6k annotated utterances have been released. The setting differs significantly from ours as we are interested in spontaneous speech with transcriptions to train an ASR system which can automatically transcribe learners' speech verbatim.

**ASR for children's speech and language learners.** The literature on ASR models for children's speech, especially for non-native language learners, is sparse. Most notably, Lu et al. (2022) investigated the performance of fine-tuning wav2vec 2.0 (Baevski et al., 2020) on children's speech (both native MyST and OGI), as well as non-native speech (TLT) compared to fine-tuning on adult-only data. The results show that ASR models trained on children's speech significantly outperform those models trained on adult-speech only, even in the case of non-native speakers. Similarly, Shivakumar and Narayanan (2022) investigated the impact of using children's data for fine-tuning ASR models. The conclusion is similar to Lu et al. (2022): adding children's data yields better performance; however, the performance of an adult ASR model on adult data is higher than the performance of an ASR model trained and applied on children's data. While both Lu et al. (2022) and Shivakumar and Narayanan (2022) are interested in the overall performance in terms of WER, our work focuses on the preservation of mistakes made by non-native children.

## 3 Dataset: Spontaneous Speech of Young Learners of English

We now describe the dataset that we collected for the purpose of this research. It contains 85 hours of audio recordings of spontaneous speech by young Swiss learners of English. Each recording is paired with a verbatim transcript that contains error annotations.

### 3.1 Audio Recording

The recording setup was designed such that the collected speech resembled the kind of conversations intended for the learners to hold with the chatbot. We used playful and engaging activities targeted to elicit extended authentic communication from young learners. Activities included role plays with problem-solving components (e.g. 'going shopping for a school trip'), guessing games (e.g. riddles), TV interviews with imaginary characters and asking/answering personal questions (e.g. 'if you could go into space, what would you take with you?'). All activities were piloted with a grade 4 class and maintained, adjusted (to yield more data) or rejected (e.g. because the task led to students communicating non-verbally and/or with much noise) for the main data collection period. To support learners, each activity further included visual and language support (e.g. cartoon characters they could choose from, sample dialogues, language chunks) as well as a preparation phase during which the students could familiarise themselves with the tasks by use of example sentences and model dialogues.[1]

**Speaker recruitment and consent** After receiving permission to collect audio data with minors from key government institutions that act as ethics review boards in Switzerland concerning research with schools and their learners, we recruited 20 primary school teachers interested in participating in our project with their classes (via personal and university networks, newsletters and direct contact with schools). Participation was entirely voluntary and could be withdrawn at any time. Participation further necessitated the approval of the school principal and the written consent of each student's legal caretaker.[2]

In the span of 9 months (March-November 2023), 337 primary school students aged 9 to 14 years (4th to 6th graders) enrolled in 8 different schools in German-speaking Switzerland performed our activities in pairs, trios or alone (if necessary) in three different settings: at school recorded by project members; on the university campus recorded by project members and student assistants; and at school recorded by teachers and sent to us via safe weblinks. School principals, teachers and students were not remunerated for participation but received small tokens of appreciation such as flowers and chocolates.

**Metadata** Each recording is associated with the following metadata:

- School area code: an integer between 1 and 8 (inclusive)
- School grade of the speakers: 4gr, 5gr, 6gr as well as combinations (4/5gr, 5/6gr, 4/6gr)
- Recording Device
- Recording Application
- Speaking activities
- Background Noise: a boolean indicating whether background noise is audible in the recording (set manually by project members).

### 3.2 Transcription and Error Annotation

The transcription of our voice data was outsourced to a transcription agency. Services included both the transcription of the voice data and the annotation of lexical, grammatical and pronunciation

[1] Note: for the camera-ready version, we will share the descriptions of the speaking activities in the supplementary material. At this point, it is not possible to share them because the set of materials is very large and some questions, e.g. regarding copyright, have to be clarified first.

[2] Note: for the camera-ready version, we will share the consent forms in the supplementary material. At this point, it is not possible because they cannot be anonymised easily (they contain a lot of information about the participating institutions and people).

errors, as well as usage of German words. We developed a comprehensive data transcription guideline for the transcription agency which was first piloted on a small number of transcripts and then adjusted where necessary. Transcription guidelines included information about spelling conventions (British English), the frequency and nature of timestamps (start and end time of each word, in milliseconds), error codes (@! for errors of any kind and @g for German words) and disfluency markers (e.g. a hyphen "–" for verbatim repetitions, such as 'he's – he's really tall'). The complete transcription guidelines are provided in the supplementary material of this paper.

### 3.3 Data Aggregation and Filtering

The recording stage resulted in 1039 audio recordings. Of these, 23 were removed due to missing metadata or missing/retracted consent, so a total of 1016 recordings and their associated metadata and transcriptions were available for our experiments.

These recordings were split into individual utterances by a single speaker using the word-level timestamps provided in the transcripts, resulting in 49'608 utterances.We removed utterances shorter than 0.5 seconds and utterances attributed to adults (e.g. short interventions by teachers), creating a final dataset of 45'004 utterances corresponding to 85 hours of audio. Each utterance was paired with its reference transcription and metadata.

### 3.4 Final Dataset

The final dataset contains 45'004 utterances by 327 distinct speakers. Figure 1 shows the number of recordings and audio duration by school grades and school area codes. Almost half the data in terms of both utterances and hours comes from 6th graders, while the other half is split among the other grades. The dataset contains 485,770 tokens and 10,203 distinct types. There are 14,396 error-annotated tokens with 2,004 underlying types. Thus, our data contains a large amount of tokens and a relatively large amount of token diversity.

The length distribution is shown in Figure 2. It can be seen that most utterances are between 0.5 and 20 seconds long.

#### 3.4.1 Data Folds

For the experiments in this paper, we split the dataset into five distinct folds of similar duration (about 16h each), where each class (and therefore also each speaker) occurs in only one fold. To simulate the use case of the ASR system being confronted with a new class of learners, each fold contains data from a mix of grades. Figure 3 visualises the duration and grade distribution of each fold.

#### 3.4.2 Data Availability

The dataset that we collected contains sensitive data of minors and thus cannot be shared publicly. The data can, however, be accessed as part of a joint project with one or several of the original project partners, subject to a collaboration agreement.[3] Before sharing, all transcripts will undergo complete anonymisation such that any names and other personal information are removed.

## 4 Error-Preserving Automatic Speech Recognition

This section presents the metrics used for measuring error preservation and evaluating systems (Section 4.1), as well as the approaches to comparing pre-trained ASR systems (Section 4.2) and to fine-tuning existing systems using our learner dataset (Section 4.3). The qualitative results are presented and discussed in Section 4.4 and a qualitative evaluation is shared in 4.5.

### 4.1 Metrics

In order to measure error preservation, we use the error annotations that were manually added to each utterance (cp. Section 3.2) and a custom phonetic word-level alignment algorithm. This algorithm aligns two or more sequences (e.g., a reference and one or multiple hypotheses), identifying matches, substitutions (S), insertions (I), and deletions (D) at the word level. Our metric, WEPR (Word-Based Error Preservation Rate), considers only those word pairs where the reference word contains an error annotation. WEPR is calculated according to equation 1: $\mathcal{A}$ is the set of annotations that are considered (e.g. $\mathcal{A} = \{@!, @\}\}$), $\mathcal{S}$ and $\mathcal{D}$ are the number of substitutions and deletions, respectively, where the reference word contains an error annotation, and $\mathcal{N}$ is the total number of reference words that contain an error annotation.

$$WEPR(\mathcal{A}) = \frac{(\mathcal{S}+\mathcal{D})}{\mathcal{N}} \quad (1)$$

In addition to WEPR, we also compute the following general ASR metrics using all words in the

[3] Note: For anonymisation purposes, details regarding data access will only be shared upon acceptance, i.e., in the camera-ready version.

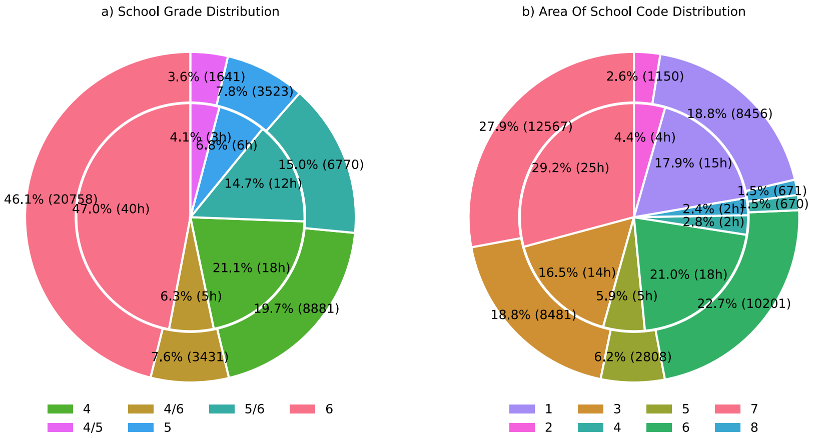

Figure 1: Number of utterances (outer ring) and audio hours (inner ring) by school grade (a) and school area code (b).

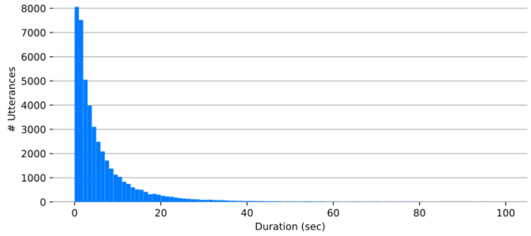

Figure 2: Distribution of utterance lengths.

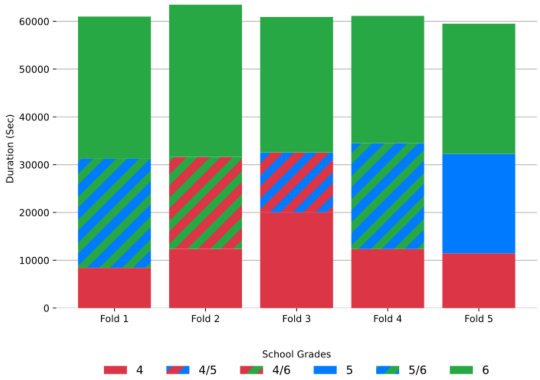

Figure 3: Duration and grade distribution of the data folds.

utterance: Word Error Rate (WER)[4], Character Error Rate (CER)[5], and character n-gram F-Score (chrF)[6] (Popović, 2015).

We evaluate all models on our dataset's five folds (cp. Section 3.4.1) and report for each model the mean and standard deviation across all folds.

For evaluation, all texts are normalised using the Whisper normalizer for English [7].

### 4.2 Pre-trained ASR Systems

We compare the performance of state-of-the-art ASR systems trained on datasets of adult English speakers. For this, we select seven different models, four based on a CTC decoding strategy, and three based on an encoder-decoder architecture. Our hypothesis is that CTC models are better at preserving speaker-errors as they do not rely on a language model, which potentially corrects such errors. Therefore, we do not use a n-gram language model during the CTC decoding phase, which is usually added for better WER performance. For the CTC-based models, we use the original Wav2VWec 2.0 large and base models (Baevski et al., 2020) fine-tuned on 960h of Librispeech (Panayotov et al., 2015) (English adult read-aloud data). We also use the fine-tuned Wav2Vec 2.0 models provided by Grosman (2021, 2022), which are based on the XLSR pretraining (Babu et al., 2021), and were fine-tuned on the CommonVoice 6.1 data (Ardila et al., 2020) consisting of approximately 2100 hours of English adult read-aloud data. For the encoder-decoder architecture, we used the Whisper medium, large, and large-v3 models provided by OpenAI (Radford et al., 2022).

[4] https://github.com/huggingface/evaluate/blob/main/metrics/wer/wer.py

[5] https://github.com/huggingface/evaluate/blob/main/metrics/cer/cer.py

[6] https://www.nltk.org/api/nltk.translate.chrf_score.html#nltk.translate.chrf_score.corpus_chrf

[7] https://github.com/openai/whisper/blob/main/whisper/normalizers/english.py

| System Name | #Param. | WER | CER | chrF | WEPR |
|---|---|---|---|---|---|
| Wav2Vec Base | 95M | 0.55 ± 0.02 | 0.34 ± 0.02 | 0.35 ± 0.02 | 0.57 ± 0.02 |
| Wav2Vec Large | 317M | 0.49 ± 0.02 | 0.29 ± 0.01 | 0.41 ± 0.02 | 0.50 ± 0.02 |
| XLSR-53 + CommonVoice 6.1 | 317M | 0.38 ± 0.01 | 0.26 ± 0.01 | 0.59 ± 0.01 | 0.50 ± 0.03 |
| XLSR-1B + CommonVoice 6.1 | 1B | 0.31 ± 0.01 | 0.21 ± 0.01 | 0.61 ± 0.01 | 0.44 ± 0.03 |
| Whisper Medium | 769M | 0.26 ± 0.02 | 0.20 ± 0.03 | **0.70** ± 0.02 | 0.46 ± 0.04 |
| Whisper Large | 1.5B | **0.25** ± 0.02 | 0.19 ± 0.01 | **0.70** ± 0.01 | 0.47 ± 0.03 |
| Whisper Large-v3 | 1.5B | 0.30 ± 0.04 | 0.23 ± 0.03 | **0.70** ± 0.02 | 0.45 ± 0.03 |
| ChaLL-300M (ours) | 300M | 0.30 ± 0.01 | **0.16** ± 0.01 | 0.68 ± 0.01 | **0.38** ± 0.03 |

Table 1: Results of the 5-fold evaluation. We report for each model the mean and standard deviation (mean±std) of the scores on each of the 5 folds. The bottom row shows the scores of our fine-tuned model.

| Target | Prediction | ChaLL-300M | Whisper-Large | XLSR-1B |
|---|---|---|---|---|
| de@! | the | 0.946 | 0.869 | **0.805** |
| a@! | _ | **0.114** | 0.327 | 0.347 |
| a@! | an | **0.026** | 0.398 | 0.257 |
| have@! | has | **0.015** | 0.231 | 0.052 |
| have@! | _ | **0.034** | 0.128 | 0.129 |
| you@! | your | 0.244 | 0.306 | **0.099** |
| it's@! | it | **0.068** | 0.116 | 0.136 |
| it's@! | _ | **0.043** | 0.119 | 0.146 |
| is@! | _ | **0.05** | 0.125 | 0.136 |
| it's@! | is | **0.055** | 0.133 | 0.103 |
| are@! | _ | **0.072** | 0.162 | 0.144 |
| dis@! | this | 0.854 | 0.83 | **0.717** |
| it@! | _ | **0.094** | 0.311 | 0.193 |
| he@! | _ | **0.175** | 0.254 | 0.356 |
| de@! | _ | **0.029** | 0.123 | 0.143 |
| the@! | _ | **0.046** | 0.24 | 0.183 |
| in@! | _ | **0.027** | 0.127 | 0.107 |
| you@! | _ | **0.077** | 0.113 | 0.117 |
| i@! | _ | **0.133** | 0.248 | 0.294 |
| on@! | _ | **0.019** | 0.129 | 0.105 |
| Mean (n=20) | | **0.156** | 0.265 | 0.229 |

Table 2: System comparison on 20 most frequent incorrectly transcribed speaker-errors. For each system, the number indicates the fraction of cases in which the system incorrectly transcribes the error Target as Prediction (where "_" denotes deletion of Target). The lowest value of each row is set in boldface. The final row shows the mean across the 20 samples.

### 4.3 Fine-tuning Pre-trained ASR Systems Using Learner Data

To evaluate the impact of fine-tuning, we fine-tune the Wav2Vec-XLSR-300M model [8] (Babu et al., 2021) on our collected language learner data.

**Data Preprocessing.** For fine-tuning, we split longer utterances into chunks of a maximum of 12 seconds and removed trailing pauses. The transcripts were preprocessed as follows:

- Remove error annotations and other transcript conventions
- Convert to lowercase
- Standardise text (Remove text between brackets and parentheses. Standardise apostrophes by removing spaces before them. Remove commas between digits and periods not followed by numbers.)
- Clean and standardise whitespace
- Normalise/remove special characters.
- Transform numbers into words using *num2words*

**Approach.** We apply 5-fold cross-validation (cf. 3.4.1), that is, we train on four folds, and evaluate on the held-out fold. We trained each run on 6 nVidia Tesla V100 GPUs for 4000 steps using a learning rate of 3e-5, a per-device batch size of 14, and 15 gradient accumulation steps (for a total batch size of 1260, which corresponds to approx. 2 hours of audio per batch), and we used the 8-bit AdamW optimizer (Loshchilov and Hutter, 2017; Dettmers et al., 2021). Our fine-tuned model, called ChaLL-300M, is available on HuggingFace.[9]

[8] Due to the high computational cost, we decided to use the 300M model instead of the 1B model.

[9] Note: for anonymisation purposes, the link will only be shared upon acceptance, i.e. in the camera-ready version.

| TARGET | PREDICTION | CHALL-300M | WHISPER-LARGE | XLSR-1B |
|---|---|---|---|---|
| have@! | have | **0.875** | 0.587 | 0.699 |
| a@! | a | **0.804** | 0.215 | 0.325 |
| is@! | is | **0.79** | 0.667 | 0.769 |
| in@! | in | **0.897** | 0.773 | 0.807 |
| it's@! | it's | **0.703** | 0.568 | 0.482 |
| are@! | are | **0.739** | 0.688 | 0.699 |
| on@! | on | **0.917** | 0.749 | 0.79 |
| of@! | of | **0.922** | 0.705 | 0.848 |
| the@! | the | **0.815** | 0.632 | 0.678 |
| you@! | you | 0.606 | 0.55 | **0.735** |
| she@! | she | **0.867** | 0.713 | 0.774 |
| it@! | it | **0.772** | 0.579 | 0.659 |
| has@! | has | **0.825** | 0.775 | 0.774 |
| make@! | make | **0.95** | 0.746 | 0.808 |
| do@! | do | **0.82** | 0.744 | 0.748 |
| much@! | much | **0.98** | 0.96 | 0.96 |
| he@! | he | **0.679** | 0.627 | 0.561 |
| not@! | not | **0.89** | 0.75 | 0.777 |
| at@! | at | **0.811** | 0.612 | 0.759 |
| don't@! | don't | **0.885** | 0.826 | 0.811 |
| Mean (n=20) | | **0.827** | 0.673 | 0.723 |

Table 3: System comparison on 20 most frequent correctly preserved speaker-errors. For each system, the number indicates the fraction of cases in which the system correctly transcribes the error TARGET as PREDICTION. The highest value of each row is set in boldface. The final row shows the mean across the 20 samples.

### 4.4 Quantitative Results

**Performance Metrics.** The scores achieved by the different models are summarised in Table 1. Among the pre-trained models, *Whisper-Large* achieves the best overall WER and chrF scores. However, the best CERand WEPR scores were achieved by the *XLSR-1B* models fine-tuned on CommonVoice 6.1. This aligns with our expectations, as Whisper models are currently the most powerful ASR models, and we expected them to perform best in terms of WER. However, for our use-case, we are more interested in error preservation, thus, CTC-based models without language models are best for preserving the errors. The fine-tuning step on our dataset consisting of learner data yielded a significant boost in performance. It achieves the best WEPR score, which measures the error retention capability. The most comparable model in terms of number of parameters is the XLSR-53 model trained on adult read-aloud data. In comparison to this model, *Chall-300M* achieves an improvement of 8 points in WER and a 12-point improvement in WEPR. It is generally the case that larger models perform better. Thus, the interpretation of the results needs to factor this in. As most models are larger than ours, it becomes evident that fine-tuning on learner data increases the performance on this data in general, and the CTC architecture yields a better out-of-the-box preservation of speaker-errors .

**WEPR Analysis.** To show in more detail the reduction in WEPR, we compare the handling of specific speaker errors. Table 2 shows the confusion for the 20 most frequent examples, that is, the cases where the ASR system corrects a mistake it should have preserved. For each type of confusion, we report the rate at which it occurs. For instance, when the speaker mistakenly said "have" (denoted "have@!"), *Chall-300M* corrected it to "has" in 1.5% of cases, Whisper-Large corrected it in 23.1% of cases, and XLSR-1B in 12.9% of cases. Thus, *Chall-300M* preserved this particular kind of error the best. In total, it mistakenly corrected 15% of the 20 most frequent speaker-errors, while *Whisper-Large* corrected 26%, and *XLSR-1B* corrected 22.9%. It is interesting to note that two out of total three cases where *XLSR-1B* has the lowest rate of mis-correction is for pronunciation errors ("de@!" and "dis@!"). We also note that a majority of the most frequent unwanted error-corrections are deletions.

On the other hand, Table 3 shows the frequency at which the ASR systems correctly preserved the mistakes made by the speakers. For instance, when the speaker mistakenly says "have" (denoted as "have@!"), then *Chall-300M* preserves this mistake in 87.5% of cases, while *Whisper-Large* preserves it in only 58.7% and *XLSR-1B* in only 69.9% of cases. In total, *Chall-300M* is able to preserve 82.7% of the of the 20 most frequent mistakes made

| | Utterance | Err. Type. |
|---|---|---|
| Target | Yeah. Uhm it's – It have a Lampe. Uhm you can – | has/have, German |
| Chall300M | e uhm it's it's have a lampe you can | has/have, German |
| Whisper-Large | it has a lamp | - |
| Target | (...) What you're rather be a (...)- able for fly or be invisible- invisible? | for/to |
| Chall300M | wuld your reader be be aabble for fly or be invisible invisible | for/to |
| Whisper-Large | would your reader be able to fly or be invisible | - |
| Target | Do you have a enemy? | a/an |
| Chall300M | do you have a enemey | a/an |
| Whisper-Large | do you have an enemy | - |
| Target | What do you favourite food? | do/is |
| Chall300M | what do you favorite food | do/is |
| Whisper-Large | what's your favorite food | - |

Table 4: Manually selected examples.

by speakers, while *Whisper-Large* only preserved 67.3% of speaker mistakes and *XLSR-1B* preserved 72.3%.

Thus, *Chall-300M* displays a strong ability to preserve the mistakes made by speakers, which is crucial for the downstream task of providing automated corrective feedback.

### 4.5 Qualitative Results

Table 4 shows four manually selected examples, highlighting some mistakes which the best-performing pre-trained model, *Whisper-Large*, corrects, and our model preserves. In the first example, it shows the mistake of using "have" instead of "has", as well as using the German pronunciation of the word "lamp" (i.e., "Lampe"). *Whisper-Large* corrects these mistakes, and creates a grammatically correct English utterance. The *ChaLL-300M* model preserves these errors as desired. The second error is a prepositional error, where the learner said "for fly" instead of "to fly". The *Chall-300M* model correctly preserved this error, while the language model used in *Whisper-Large* smoothed out the error. The third example is an error of the indefinite article: the learner used "a" instead of "an", which *ChaLL-300M* correctly preserved while *Whisper-Large* corrected the error. The final example contains the usage of the wrong verb "do" instead of "is", which again is correctly preserved by our model while Whisper corrects the mistake.

## 5 Conclusion and Outlook

Our work shows that state-of-the-art ASR systems have difficulties handling young learners' speech; furthermore, they tend to correct the mistakes made by the speakers, which makes the downstream identification of speaker mistakes and provision of corrective feedback impossible. Thus, we collected around 85 hours of children's language learner speech data, which we used to fine-tune a custom model. Our model outperforms all the others (including Whisper-Large) in terms of error preservation (Word-Based Error Preservation Rate, WEPR) and surpasses the English models of comparable size ($\approx 300M$ parameters) by a large margin in terms of Word Error Rate. Thus, our research shows the necessity of using targeted data (in this case, children who learn a foreign language) to fine-tune an ASR module, which is useful in downstream tasks. The focus of this work lies in a) investigating the utility of existing systems and b) creating an adequate ASR system that can be used as part of a language learning support tool to increase the students' speaking opportunities. As a next step, we will investigate how to enhance error preservation. For this, training larger models is the most straightforward approach. However, we also plan to train the ASR system jointly with error annotations. For this, we started the creation of more detailed error annotations. Initial results have shown that verbal errors are the largest error category for young Swiss learners of English (with about 22% of all errors) , and within these, wrong subject-verb agreement is most frequent. Similarly, investigating how to handle frequent code-switching to German words or sentence fragments is an unsolved issue that needs to be addressed to improve downstream tasks. Even *Whisper-Large*, which can handle multiple languages in principle, did not perform well in detecting code-switching.

Finally, we aim to evaluate ASR models in the context of integrating them with a conversational agent and corrective feedback.

## Limitations

While offering a unique tool for error-preserving ASR of young language learners, this work presents itself with a few limitations.

**Limited Demographic.** The dataset stems from a specific demographic of Swiss school children learning English in grades 4 to 6. An extension of the work would include language learners with different native languages or a larger range of ages. Thus, the transferability of our results must be confirmed with a different dataset.

**Outsourcing Error Annotation.** The outsourcing of transcription and error annotations always poses a risk of yielding erroneous data, since most transcribers are not trained in error annotation. We mitigated this risk by providing comprehensive guidelines and a steady exchange with the transcription agency. However, we plan to enhance the error annotations with a more detailed label set and annotators trained in this task.

**Small Model.** Due to the high computational cost of fine-tuning a 1B parameter model, we limited ourselves to fine-tuning the 300M parameter XLSR model. Most research indicates that the usage of larger models yields better results; thus, there is still potential in terms of increasing WER and WEPR. However, our results showed that even a small model can preserve errors better than state-of-the-art pre-trained models, which was the main scope of this work.

**No Performance Tuning.** Since the scope of this work is to understand if the usage of young learners' speech data is beneficial for our purposes, we did not tune the performance of our model. That is, we did not perform any hyper-parameter tuning or any other methods to increase performance (e.g., joint prediction of errors using a language model). Thus, there is still a large margin of improvement using our dataset.

**Data Availability.** Since our data consists of children's spontaneous speech, we must ensure its protection. Thus, we cannot make it freely available. While we publicly release the models trained on the data, access to the transcripts and recordings can only be granted in the scope of a joint project, subject to a collaboration agreement.

## Ethical Considerations

The main risks in this project have to do with data protection: all speakers are minors between 9 and 14 years of age, so their personal data must be very well safeguarded. Therefore, key government institutions approved the data collection before speakers were recruited, and informed consent was obtained from each speaker's legal caretaker (cp. details in Section 3.1). Consent forms entailed information about the nature of the project and data collection procedures, as well as a comprehensive description of the legal principles we followed to collect, use, and store voice data, transcripts, and annotations. The data protection measures we implemented for security and confidentiality were fully disclosed (e.g. password-protected documents, pseudonymisation, firewalls etc.) and risks to participants (e.g. potential voice recognition by project members) were outlined. Voice data and transcripts were pseudonymised by those project members who act as data owners before sharing them with other research partners and third parties. Third-party access to the collected data will be enabled in a closely controlled setting consisting of a joint project with a collaboration agreement.

## Use of AI Assistants

ChatGPT was used to support the creation of some figures. No AI assistants were used for writing the text of this paper.

## Acknowledgements

Acknowledgements will be added in the camera-ready version as they are not anonymous.

## References

Beena Ahmed, Kirrie J. Ballard, Denis Burnham, Tharmakulasingam Sirojan, Hadi Mehmood, Dominique Estival, Elise Baker, Felicity Cox, Joanne Arciuli, Titia Benders, Katherine Demuth, Barbara Kelly, Chloé Diskin-Holdaway, Mostafa Shahin, Vidhyasaharan Sethu, Julien Epps, Chwee Beng Lee, and Eliathamby Ambikairajah. 2021. AusKidTalk: An Auditory-Visual Corpus of 3- to 12-Year-Old Australian Children's Speech. In *Proc. Interspeech 2021*, pages 3680–3684.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

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

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Claudia Baur, Andrew Caines, Cathy Chua, Johanna Gerlach, Mengjie Qian, Manny Rayner, Martin Russell, Helmer Strik, and Xizi Wei. 2018. Overview of the 2018 spoken call shared task. In *Proceedings of Interspeech 2018*, Interspeech, pages 2354–2358. ISCA. Interspeech 2018 ; Conference date: 02-09-2018 Through 06-09-2018.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*.

Rod Ellis. 2021. Explicit and Implicit Oral Corrective Feedback. *The Cambridge Handbook of Corrective Feedback in Second Language Learning and Teaching*, pages 341–364.

Maxine S Eskenazi. 1996. Kids: a database of children's speech. *The Journal of the Acoustical Society of America*, 100(4_Supplement):2759–2759.

Matteo Gerosa, Diego Giuliani, and Shrikanth Narayanan. 2006. Acoustic analysis and automatic recognition of spontaneous children's speech. In *Ninth International Conference on Spoken Language Processing*.

Roberto Gretter, Marco Matassoni, Stefano Bannò, and Falavigna Daniele. 2020. TLT-school: a corpus of non native children speech. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 378–385, Marseille, France. European Language Resources Association.

Nancy Grimm, Michael Meyer, and Laurenz Volkmann. 2015. *Teaching English*. Narr Francke Attempto Verlag.

Jonatas Grosman. 2021. Fine-tuned XLSR-53 large model for speech recognition in English. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english.

Jonatas Grosman. 2022. Fine-tuned XLS-R 1B model for speech recognition in English. https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-english.

Andreas Hagen, Bryan Pellom, and Ronald Cole. 2003. Children's speech recognition with application to interactive books and tutors. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*, pages 186–191. IEEE.

Tricia Hedge. 2001. *Teaching and learning in the language classroom*, volume 106.

Robert Kleinschroth and Pete Oldham. 2014. *Sprechkompetenz-Training im Englischunterricht 7-8: Lebensnahe Sprechanlässe und vielfältige Aufgaben (7. und 8. Klasse)*. Auer Verlag.

Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Renee Lu, Mostafa Shahin, and Beena Ahmed. 2022. Improving children's speech recognition by fine-tuning self-supervised adult speech representations. *arXiv preprint arXiv:2211.07769*.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80:9411–9457.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56(4):3055–3155.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE.

Simone E. Pfenninger and Johanna Lendl. 2017. Transitional woes: On the impact of l2 input continuity from primary to secondary school. *Studies in Second Language Learning and Teaching*, 7(3):443–469.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

A. Potamianos and S. Narayanan. 2003. Robust recognition of children's speech. *IEEE Transactions on Speech and Audio Processing*, 11(6):603–616.

Sameer Pradhan, Ron Cole, and Wayne Ward. 2016. My science Tutor—Learning science with a conversational virtual tutor. In *Proceedings of ACL-2016 System Demonstrations*, pages 121–126, Berlin, Germany. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Lars Rumberg, Christopher Gebauer, Hanna Ehlert, Maren Wallbaum, Lena Bornholt, Jörn Ostermann, and Ulrike Lüdtke. 2022. kidsTALC: A Corpus of 3- to 11-year-old German Children's Connected Natural Speech. In *Proc. Interspeech 2022*, pages 5160–5164.

Prashanth Gurunath Shivakumar and Panayiotis Georgiou. 2020. Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. *Computer Speech & Language*, 63:101077.

Prashanth Gurunath Shivakumar and Shrikanth Narayanan. 2022. End-to-end neural systems for automatic children speech recognition: An empirical study. *Computer Speech & Language*, 72:101289.

Khaldoun Shobaki, John-Paul Hosom, and Ronald A. Cole. 2000. The OGI kids² speech corpus and recognizers. In *Proc. 6th International Conference on Spoken Language Processing (ICSLP 2000)*, pages vol. 4, 258–261.

Wayne Ward, Ron Cole, and Sameer Pradhan. 2019. My science tutor and the myst corpus. *Boulder Learn. Inc*.