Adaptation of automatic speech recognition models for child speech

Reeka Estacio and Allison Park
University of California, San Diego
Computational Social Science, 2025

Introduction

Background

Automatic Speech Recognition (ASR) is a powerful and versatile technology used across a wide range of applications—from everyday tasks like speech-to-text messaging and automated customer service systems to advanced research in linguistics and computer science. Despite its widespread utility, developing high-performing ASR systems that can successfully handle linguistic variability remains a complex challenge. These systems must not only achieve high accuracy, but also manage the significant demands of large-scale data and intensive computational resources.

Several methodologies have shaped the development of automatic speech recognition systems, drawing on interdisciplinary advances in machine learning, natural language processing (NLP), and linguistic theory to accurately transcribe and label spoken language. Modern, state-of-the-art ASR models are predominantly built on neural network architectures. Among the most influential is Facebook AI's wav2vec 2.0 (Baevski et al., 2020), a self-supervised model that learns speech representations directly from raw audio. Although it was not originally designed specifically for ASR, Facebook released pretrained versions—such as wav2vec2-base-960h—which are fine-tuned using supervised learning on labeled datasets like the 960-hour LibriSpeech corpus. This approach enables strong performance across diverse tasks and domains, including multilingual and accented speech, with relatively little task-specific adaptation. When evaluated on adult speech, Facebook reports that wav2vec2-base-960h achieves word error rates (WERs) of 1.8% on LibriSpeech test-clean (high-quality, clear speech) and 3.3% on LibriSpeech test-other (lower-quality, noisy speech). These WERs are highly competitive, indicating very strong performance on adult speech.

Despite their unprecedented accuracy, there are a number of studies that show that even the leading ASR models perform poorly on more variable types of speech, such as L2-accented speech and non-standard dialects. For the present study, we are particularly interested in the challenges posed by child speech. For instance, McGonigle et al. (2024) finds that OpenAI's

Whisper model accurately transcribes less than 50% of words produced by typically-developing children, attaining a WER between 54-74%. This is a steep decline from OpenAI's reported average WER of 12.8% across various adult speech datasets (Radford et al. 2022). One reason for this is that children's speech is acoustically more variable than adults—such as physiologically-driven differences in articulation and pitch—ultimately resulting in much lower reported WERs (Beckman et al. 2017). Another factor contributing to this disparity is the limited availability and sensitivity of child speech data. Unlike adult speech, which is widely collected and publicly available, child speech datasets are often smaller and subject to stricter ethical regulations. As a result, the amount of child speech data available for training is insufficient to meet the data-intensive demands of modern ASR models, making it difficult to achieve comparable performance.

While proven to be a technically challenging task, improving ASR performance for child speech is essential for creating more inclusive and accessible speech technology. Child speech recognition has important applications in the development of educational tools, language development research, and assistive technologies for children with speech or communication disorders. Learning apps, speech therapy tools, and interactive educational games increasingly feature ASR functionalities to support skill development. Without systems that can accurately interpret child speech, these technologies risk being ineffective—or entirely inaccessible—for younger users. This issue disproportionately affects children with language or learning delays, as inaccessibility to this technology widens the educational gap between them and their typically-developing peers.

These discrepancies are also evident in downstream tasks, such as forced alignment. Unlike the broader understanding of automatic speech recognition, which involves recognizing words audio without given labels, forced alignment is a task that finds the boundaries between words or sounds in the audio, given what those words or sounds are. The algorithm or model is required, or "forced", to use the given labels. Forced alignment requires accurate recognition of sounds in order to correctly identify the relevant boundaries.

Automation of transcription tasks like forced alignment are especially useful in research and development contexts that involve processing a high volume of data. Annotating transcriptions by hand takes three to eight times longer than the audio segment itself (Liu and Xiong, 2024), and requires human transcribers who have been trained for that specific task. Transcription task requirements themselves can vary, from phonetic-level tasks such as identification of vowels in speech segments to forced alignment of phones; to higher-level tasks such as identification of stutters or the forced alignment of words in sentences. The ability to automate at least some annotation and transcription tasks enables allocation of finite human resources to other aspects of research and development.

However, forced alignment of child speech suffers from similar problems to the broader automatic recognition of child speech. Forced alignment models that have been trained on adult productions of phones struggle with the differing acoustic properties of children's speech: higher pitch, differences in formant frequencies, and slower speed create problems, as does high variation between what phones (or mistakes) that different children might make when producing the same word (Knowles et al, 2015; Bharadwaj et al, 2022).

Efforts to increase ASR accuracy and usability on child speech include both adaptation of models to child speech and adaptation of child speech to the models. Techniques for transforming the input speech before processing it through ASR include vocal tract length normalization, maximum likelihood linear regression for adaptation to children's speech patterns, and speaker-adaptive training; however, these techniques are not usable for neural network architectures (Shivakumar and Narayanan, 2022). Large language models specifically for interaction with children are in development (Nayeem and Rafiei, 2024) but have not yet been used as the basis for ASR technology as with LLM-driven adult ASR.

Our goal

The current study aims to explore how ASR models perform on child speech after additional training on a child speech dataset provided by the Language Acquisition and Speech Recognition (LASR) Lab at UCSD. We evaluate the performance of two pretrained models and implement different adaptation techniques in effort to improve predictive performance on child speech. In particular, we investigate two questions:

1) Can we fine-tune neural ASR models for child speech, and what do systematic model errors reveal about the challenges of recognizing child speech?
2) What adjustments can be made to the non-neural components of hidden Markov models to improve forced alignment accuracy on child speech?

The initial goal of this project was to explore adaptation methods to improve forced alignment performance on child speech using two different models: Montreal Forced Aligner (MFA) and Charsiu. This proved feasible for MFA, due to its well-documented training pipeline and efficient HMM-based architecture. Adapting Charsiu, on the other hand, presented significant challenges. First, Charsiu's training pipeline is notably opaque. Beyond the technically-dense prose in Zhu et al. (2022), the Charsiu GitHub repository offers minimal guidance on training implementation. Even after reaching out to the authors for clarification, it became evident that further training of the model would require a far more complex and involved process than initially anticipated.

Furthermore, successful domain adaptation for neural ASR systems are exceptionally costly, typically requiring supervised training on hundreds of thousands of hours of labeled speech datasets. For instance, pretraining for Charsiu's Wav2Vec2 frame classification model relies on alignments generated by MFA from the 960h LibriSpeech dataset. In contrast, we had access to about 26 hours of child speech data, and only six of which were manually aligned. Given the poor performance of forced alignment tools on child speech, relying on MFA to generate automatic alignments was not a viable option. At the same time, producing manual alignments at scale would have been excessively time-consuming.

Training time and computational demands also posed significant challenges, since fine-tuning neural models typically requires several hours of processing. While GPU acceleration is typically the standard solution for this issue, GPUs were not readily available in the lab, and transferring data to external computers raised IRB concerns due to its sensitive nature. To address these constraints, we considered parameter-efficient fine-tuning (PEFT) methods, such as Low-Rank Adaptation (LoRA). Unlike a full fine-tuning procedure, which requires updating all model parameters (e.g., 94,691,232 parameters in the Wav2Vec2 base model), LoRA reduces computational load by freezing most parameters and updating only a small, trainable subset. However, implementing LoRA proved to be complex and finicky, particularly due to the high variability in audio sample lengths within our dataset.

With the consideration of these challenges, we ultimately chose to pivot away from examining neural forced alignment and broaden the scope of our analysis to child speech adaptation of ASR systems in a more general and exploratory sense. Since it is foundational to a variety of neural ASR tasks including forced alignment, we first examine Wav2Vec2 ASR performance and assess the effects of fine-tuning on our child speech dataset. We then shift focus to the downstream forced alignment task, evaluating the performance of various configurations of MFA acoustic models and dictionaries. Through this exploration, we aim to get first-hand experience in addressing challenges associated with child speech in ASR technology.

Data

For the examination ASR and forced alignment performance, we adapt pre-trained models using a 72-participant child speech dataset obtained from the LASR Lab at UCSD. This dataset consists of .wav speech recordings of 24 single-word utterances from a picture-naming task, produced by male and female child speakers ranging from ages 3 to 5 years old. In accordance with Institutional Review Board approval, this dataset is not available for public use. Of the 72 participants, 75% of participants were selected for training (*n*=54) and 25% (*n*=18) for evaluation. The training and evaluation sets are equally representative of each age-sex group.

1728 total speech files were available from the dataset. Each utterance was transcribed by two trained human coders and categorized for correctness or error type. For this project, 137 files were excluded due to participants producing the wrong word, or an incorrect word form (e.g. *"sock"* instead of the target word *"socks"*) that eliminated a phonemic contrast between that utterance and another target word from the dataset. We retain cases with errors categorized as misarticulated or mispronounced words, to capture the natural variability of child-produced speech. The final training set consists of 1192 audio files (~20 hours) and the final evaluation set consists of 399 audio files (~6 hours). For comparative analysis on the evaluation set, we also sampled 400 adult speech recordings. These recordings feature the same set of single-word utterances as the child speech dataset, but were instead produced by adult researchers.

Fine-tuning Wav2Vec 2.0 ASR for child speech

Overview

The ASR model of interest in this portion of the study is wav2vec2-base-960h, which is pretrained using the 960-hour LibriSpeech corpus and optimized using Connectionist Temporal Classification (CTC) loss. CTC is the standard technique used for training ASR models because it allows the model to receive sequential audio frames as input and output text transcriptions, bypassing the need for frame-level labels. As mentioned before, wav2vec2-base-960h achieves high accuracy on adult speech, but degrades drastically on child speech. On our evaluation set, wav2vec2-base-960h shows similar discrepancies between adult and child speech ASR predictions, as well as overall low word-level accuracy for single-word utterances (Figure 1). Note that accuracy, in this context, is at the word level—defined as the proportion of ASR outputs that exactly replicate the target transcription's grapheme sequence.

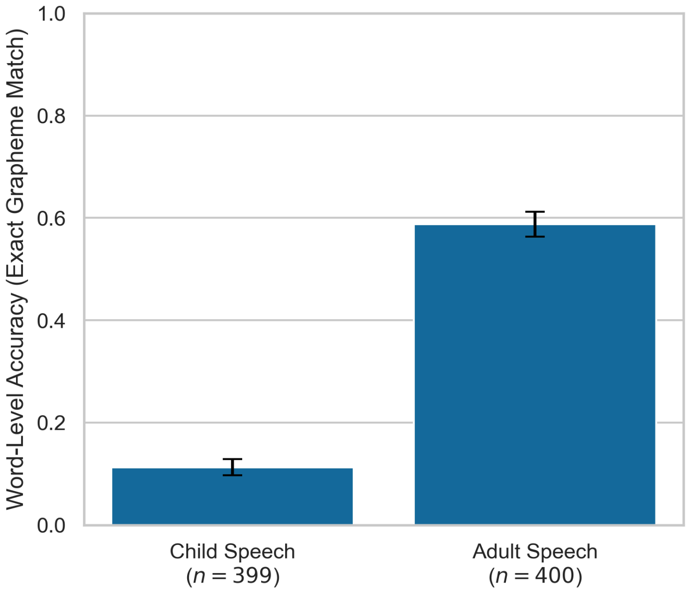


Figure 1: Baseline word-level accuracy of wav2vec2-base-960h ASR predictions for child speech evaluation set (n=399) and adult speech evaluation set (n=400).

These results highlight that even state-of-the-art ASR models struggle with robustness across variable speech inputs. In particular, they underscore the specific challenges posed by our dataset: wav2vec2-base-960h performs poorly not only on child speech, but also on single-word utterances more broadly, even at baseline.

In effort to bridge this gap, we fine-tune wav2vec2-base-960h on our child speech training set and assess performance across different learning rates. The learning rate is a hyperparameter in model training that controls how much the weights are updated in response to the loss (in this case, average CTC loss) calculated in training. The goal of the training process is to arrive at parameters that optimize the precision of model predictions and minimize loss. Larger learning rates run the risk of complete disruption of the model, potentially overfitting or forgetting the knowledge that it learned during its pretraining. Lower learning rates, on the other hand, may lead to very slow or negligible learning.

Our goal is to identify the optimal learning rate for our specific dataset to inform future fine-tuning efforts on similar child speech data. Beyond optimizing performance, we also aim to analyze the systematic errors made by the optimally fine-tuned model. This analysis will shed light on which linguistic features of child speech pose particular challenges for wav2vec2-base-960h, and potentially for ASR models more broadly. In doing so, we hope to highlight some of the fundamental acoustic and phonological differences between adult and child speech that impact model performance.

Methods

The fine-tuning procedure begins with preprocessing the training data. All audio files are resampled to a standardized sampling rate of 16kHz. The wav2vec2-base-960h pretrained processor converts the raw audio files into input features via the feature extractor. The corresponding text transcriptions are tokenized into label sequences for supervised training. To address that the audio inputs consist of varying lengths, we use the HuggingFace DataCollatorForCTC, which applies dynamic padding for each batch. This ensures that the audio features and labels are padded to the length of the longest sequence in each batch, allowing for efficient training across variable-length speech inputs.

We run the training loop across different learning rates, starting from a very slow learning rate of 10^{-10} and increasing logarithmically to 10^{-2}. A learning rate nearing 1 is very large for the purposes of model training, and will likely lead to model overfitting or forgetting. We strategically initialize training hyperparameters to accommodate for memory constraints and accelerate training. For instance, we use a small batch size ($n=5$) per device and enable mixed precision training (fp16=True). Mean CTC loss per batch is calculated and used to update model parameters. The model is trained for only a single epoch to quickly evaluate performance across the different learning rates. Ultimately, each training loop per learning rate ran for approximately

20 to 30 minutes. Finally, predictions are generated on the evaluation set, consisting of 399 child speech samples and 400 adult speech samples, for each fine-tuned model.

We acknowledge that the ASR model may produce homophones—predicted words that share the same phonemic sequence as the target but differ in their graphemic representation (e.g., predicting *"son"* instead of *"sun"*). In such cases, we consider the prediction successful. To account for this, we perform grapheme-to-phoneme (G2P) conversion using the `g2pE` package, which maps predicted text to ARPAbet phone sequences based on the Carnegie Mellon Pronouncing Dictionary (Park et al., 2019). To evaluate phone-level performance, we compute the normalized edit distance between the predicted and target phone sequences using the `editdistance` package in Python. This involves calculating the Levenshtein distance (i.e. the minimum number of insertions, deletions, and substitutions needed to transform the predicted sequence into the target sequence) and dividing by the number of phones in the target sequence. As a result, normalized edit distance ranges from 0 (perfect match) to values greater than 1 when the number of required edits exceeds the length of the target sequence.

Results

Identifying optimal learning rate for fine-tuning

The first (top) plot of Figure 2 illustrates the average normalized edit distance between the predicted and target phone sequences, grouped by learning rate and age group. A random sample of n=399 audio files were selected from the training set to assess in- and out-of-sample performance. The second (bottom) plot of Figure 2 shows the proportion of null responses to highlight instances where the model failed to generate any prediction, offering additional insight into overall model reliability across conditions. Note that this plot inverts the typical "higher is better" convention: higher edit distances reflect poorer model performance. A normalized edit distance of 1 or higher indicates that every predicted phone is incorrect, whereas a score of 0 signifies a perfect match between the predicted and target sequences.

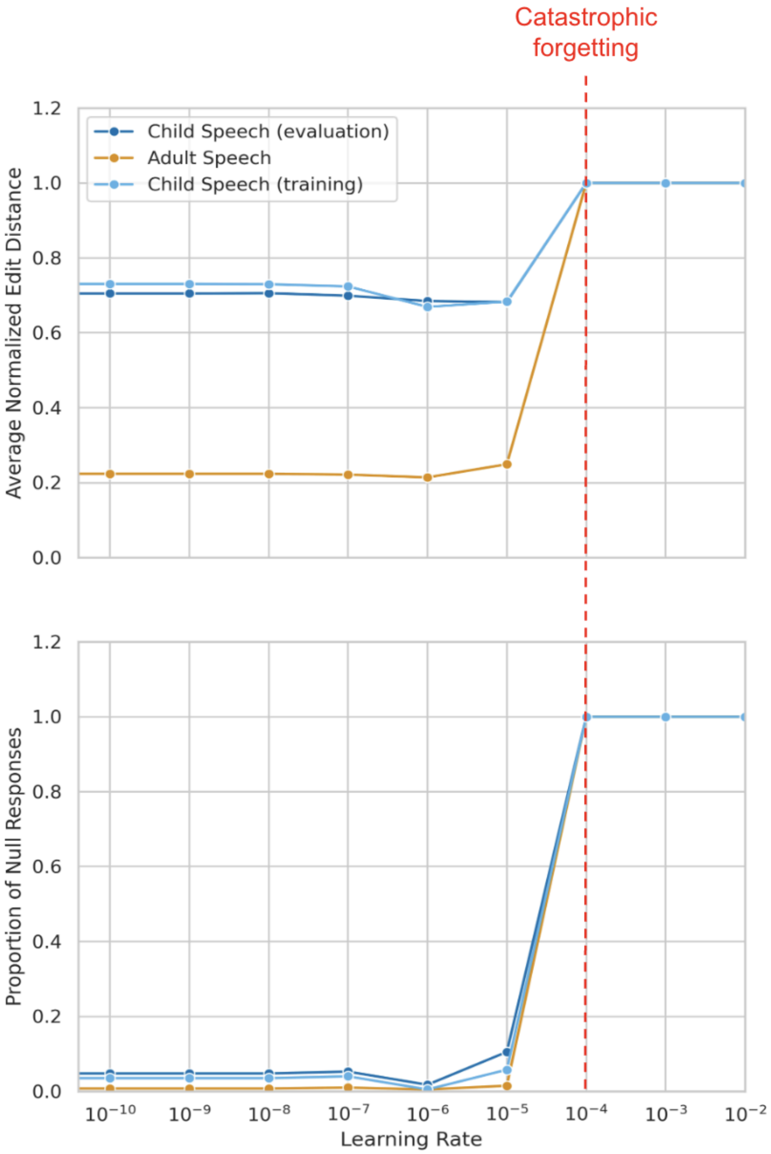


Figure 2: (Top) Average normalized edit distance between predicted phone sequence and target phone sequence, by learning rate and age group. Average normalized edit distance of 1 indicates that all predicted phones are inaccurate in reference to target phone sequence. Average normalized edit distance of 0 indicates that all predicted phones are accurate. (Bottom) Proportion of null responses generated by the models at each learning rate and age group.

Without additional fine-tuning—represented by a learning rate of zero on the far left of Figure 2—the wav2vec2-base-960h ASR model performs substantially better on adult speech than child speech. To further this point, the models produce less null responses to adult speech than child speech. Interestingly, the models consistently perform slightly worse on the child speech training set than the child speech evaluation set, but this is likely due to random variation in the training data sample.

The gap in performance between adult and child speech persists across learning rates from 0 to 10^{-5}, suggesting that the model retains a consistent advantage for adult speech despite exposure to child speech during fine-tuning. Both the average normalized edit distance and the proportion of null responses remain largely unchanged across these lower learning rates, indicating that the model is either not learning effectively or the learning rate is too small to drive meaningful parameter updates.

While average edit distance reaches a minimum at the learning rate of 10^{-5} for child speech and 10^{-6} for adult speech, the difference is minimal and may reflect random variation rather than learning. At a learning rate of 10^{-4}, model performance deteriorates catastrophically for both age groups—indicated by the dashed vertical red line going across both plots. At this learning rate, the average normalized edit distance is 1 and the model fails to generate any predictions. The proportion of null responses in the bottom figure demonstrate that this performance can be attributed to catastrophic forgetting of the model as opposed to overfitting to the training data. This is evidence that this learning rate is too high for effective fine-tuning of the model. Overall, these findings suggest that the optimal learning rate for fine-tuning for our dataset exists between 10^{-6} and 10^{-5}.

Inspecting systematic errors produced by the fine-tuned ASR model

Figure 3 depicts the average normalized edit distance for all target words across both age groups. ASR predictions are generated by the fine-tuned wav2vec2-base-960h model that resulted in the lowest average normalized edit distance for child speech (learning rate = 10^{-5}). The target words "*zoo"* and *"shoe"* demonstrate the highest average edit distance across all 24 words in the child dataset, at 1.088 and 1.029 respectively (Table 1). This is followed by *"leg"* (0.902), *"chip"* (0.833), and *"chick"* (0.833).

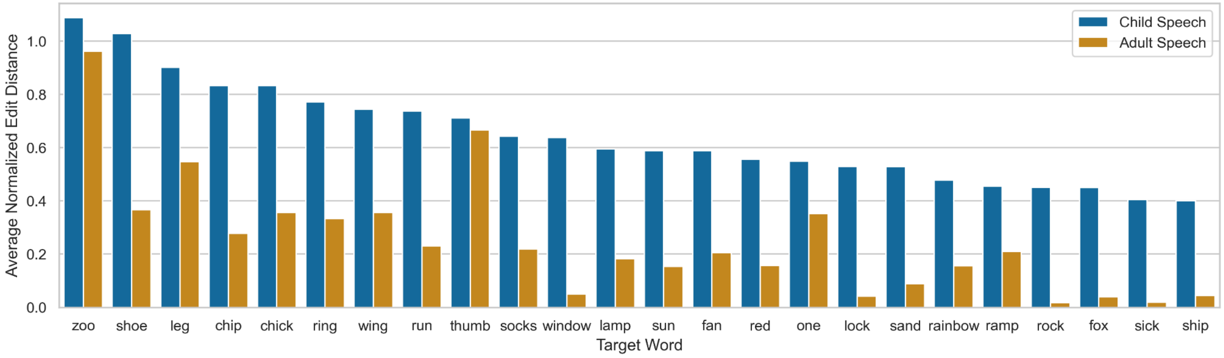
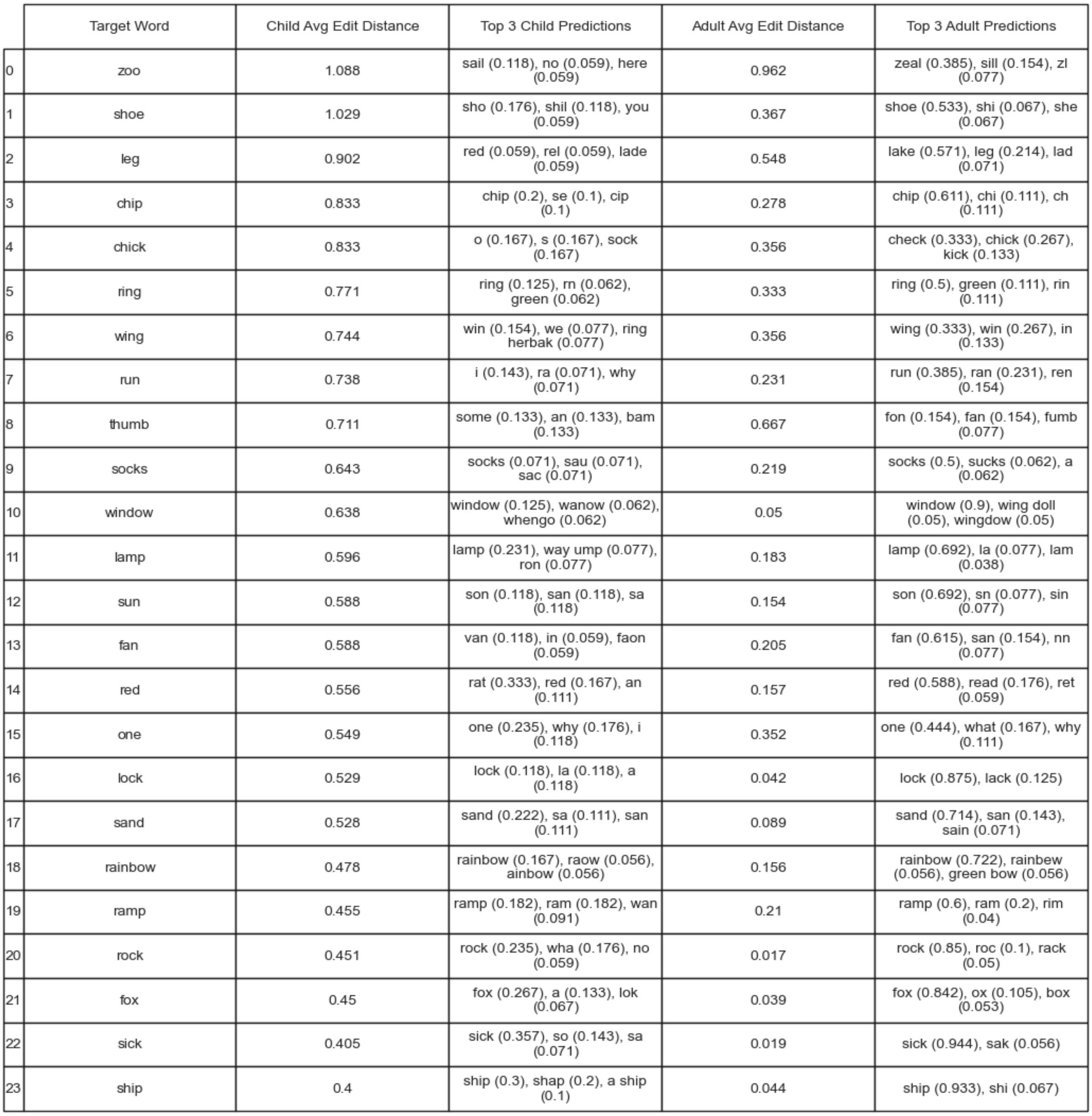


Figure 3: Average normalized edit distance for all target words. ASR Predictions on both child and adult speech are produced by the optimal fine-tuned model for child speech recognition (learning rate = 10^{-5}).

Although the fine-tuned model performs better overall on adult speech, its error distribution differs notably from that of child speech. For example, while the target word *"zoo"* yields the highest average edit distance in both groups—indicating consistently poor performance—the model also struggles disproportionately with *"thumb"* in the adult dataset, reaching an average edit distance of 0.667 (compared to 0.711 for the child dataset). In these cases, the model does not exhibit a clear adult speaker advantage, as performance is poor across both speaker groups. This pattern suggests that the model struggles with certain words regardless of speaker age. One possible explanation is that both *"zoo"* and *"thumb"* begin with relatively uncommon onset consonants: the voiced alveolar fricative /z/ and the voiceless interdental fricative /θ/. These phones may be underrepresented in the model's pretraining data, and are thus predicted less overall. In the case of adult-produced *"thumb"*, the model's top predictions frequently begin with the voiceless labiodental fricative /f/ (Table 1). This phone is a more common onset consonant and is acoustically similar to /θ/, suggesting that this is a systematic phone substitution for adult speech driven by frequency bias.

Table 1 shows the average normalized edit distance for all target words in both child and adult speech datasets. The top three most predicted phone sequences are also reported, as well as the proportion in which they occur relative to all other predictions made for that word. For child speech, these proportions are generally lower, indicating that the model tends to produce a wider variety of predictions, frequently generating unique or inconsistent outputs. In contrast, higher proportions are observed for adult speech, suggesting greater consistency in the model's predictions. This discrepancy highlights the model's continued difficulty in handling the increased variability characteristic of child speech, even after fine-tuning.



	Target Word	Child Avg Edit Distance	Top 3 Child Predictions	Adult Avg Edit Distance	Top 3 Adult Predictions
0	zoo	1.088	sail (0.118), no (0.059), here (0.059)	0.962	zeal (0.385), sill (0.154), zl (0.077)
1	shoe	1.029	sho (0.176), shil (0.118), you (0.059)	0.367	shoe (0.533), shi (0.067), she (0.067)
2	leg	0.902	red (0.059), rel (0.059), lade (0.059)	0.548	lake (0.571), leg (0.214), lad (0.071)
3	chip	0.833	chip (0.2), se (0.1), cip (0.1)	0.278	chip (0.611), chi (0.111), ch (0.111)
4	chick	0.833	o (0.167), s (0.167), sock (0.167)	0.356	check (0.333), chick (0.267), kick (0.133)
5	ring	0.771	ring (0.125), rn (0.062), green (0.062)	0.333	ring (0.5), green (0.111), rin (0.111)
6	wing	0.744	win (0.154), we (0.077), ring herbak (0.077)	0.356	wing (0.333), win (0.267), in (0.133)
7	run	0.738	i (0.143), ra (0.071), why (0.071)	0.231	run (0.385), ran (0.231), ren (0.154)
8	thumb	0.711	some (0.133), an (0.133), bam (0.133)	0.667	fon (0.154), fan (0.154), fumb (0.077)
9	socks	0.643	socks (0.071), sau (0.071), sac (0.071)	0.219	socks (0.5), sucks (0.062), a (0.062)
10	window	0.638	window (0.125), wanow (0.062), whengo (0.062)	0.05	window (0.9), wing doll (0.05), wingdow (0.05)
11	lamp	0.596	lamp (0.231), way ump (0.077), ron (0.077)	0.183	lamp (0.692), la (0.077), lam (0.038)
12	sun	0.588	son (0.118), san (0.118), sa (0.118)	0.154	son (0.692), sn (0.077), sin (0.077)
13	fan	0.588	van (0.118), in (0.059), faon (0.059)	0.205	fan (0.615), san (0.154), nn (0.077)
14	red	0.556	rat (0.333), red (0.167), an (0.111)	0.157	red (0.588), read (0.176), ret (0.059)
15	one	0.549	one (0.235), why (0.176), i (0.118)	0.352	one (0.444), what (0.167), why (0.111)
16	lock	0.529	lock (0.118), la (0.118), a (0.118)	0.042	lock (0.875), lack (0.125)
17	sand	0.528	sand (0.222), sa (0.111), san (0.111)	0.089	sand (0.714), san (0.143), sain (0.071)
18	rainbow	0.478	rainbow (0.167), raow (0.056), ainbow (0.056)	0.156	rainbow (0.722), rainbew (0.056), green bow (0.056)
19	ramp	0.455	ramp (0.182), ram (0.182), wan (0.091)	0.21	ramp (0.6), ram (0.2), rim (0.04)
20	rock	0.451	rock (0.235), wha (0.176), no (0.059)	0.017	rock (0.85), roc (0.1), rack (0.05)
21	fox	0.45	fox (0.267), a (0.133), lok (0.067)	0.039	fox (0.842), ox (0.105), box (0.053)
22	sick	0.405	sick (0.357), so (0.143), sa (0.071)	0.019	sick (0.944), sak (0.056)
23	ship	0.4	ship (0.3), shap (0.2), a ship (0.1)	0.044	ship (0.933), shi (0.067)

Table 1: Table showing the average normalized edit distance and the three most common predictions for each target word for both child and adult speech. The proportion in which the top predictions occur (over all other predictions for the target word) is calculated.

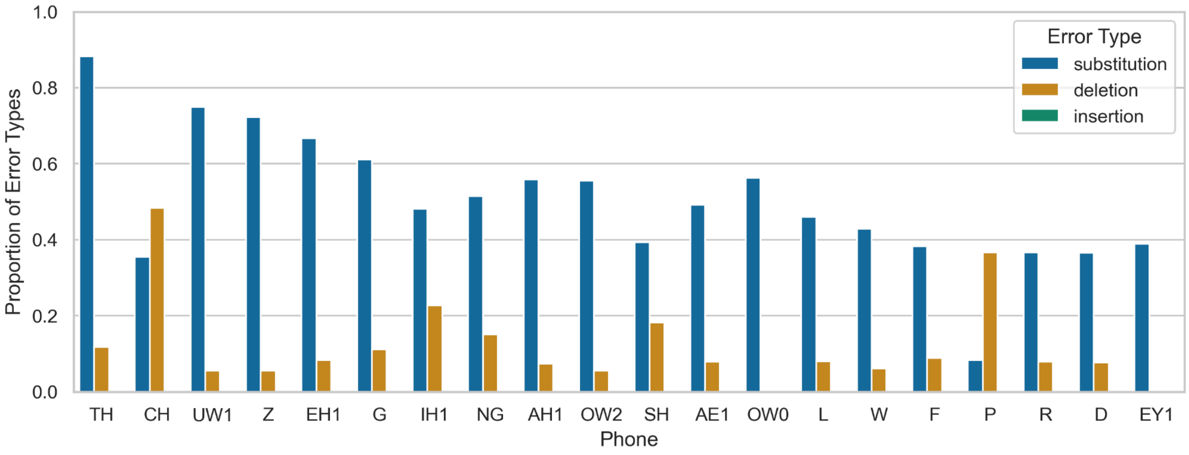


Figure 4: Proportion of error types (substitution, deletion, insertion) for the 15 most misidentified phones in the child speech dataset.

The distribution of the types of errors produced by the fine-tuned model for the most mispredicted phones in the child speech dataset is depicted in Figure 4. Substitutions comprise most of the errors, indicating the model frequently misidentifies phones rather than omitting or adding them. The "TH" or /θ/ phone shows the highest frequency of substitutions overall, contributing to the poor observed performance for the target word "thumb". The phones "CH" and "P" are more frequently omitted than substituted, suggesting that the model has difficulty recognizing these phones entirely instead of substituting them for another phone. Among the fifteen most misidentified phones depicted in Figure 4, insertions are not represented entirely, suggesting that these types of errors are generally rare across predictions compared to substitutions and deletions.

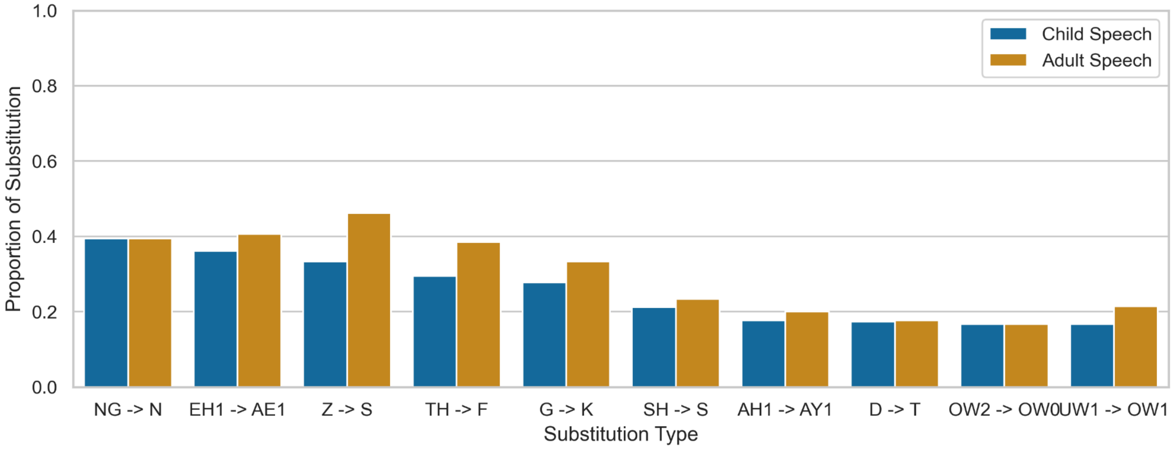


Figure 5. Top 10 substitution types for child speech recognition, with proportion relative to how often the target phone appears in the dataset. For comparison between speaker age groups, corresponding proportions of substitutions for adult speech are also plotted.

Taking a closer look at the most common error type, Figure 5 shows the ten most frequent phone substitutions made by the model, normalized by the number of occurrences of each corresponding target phone. At first glance, substitution rates appear slightly higher for adult speech. However, we interpret this not as evidence of poorer performance, but rather as an indication that the model makes more consistent and systematic errors when processing adult speech. In contrast, the lower substitution rates for child speech likely reflect the model's less stable and more variable performance. Due to the greater acoustic variability produced by younger speakers, the model generates a more scattered range of predictions, leading to fewer recurring, systematic substitutions.

Despite these differences, many of the most frequent substitutions occur at comparable rates across both age groups overall, suggesting that these errors are not primarily driven by speaker age. Instead, they reflect broader limitations in the model's ability to consistently recognize certain phones, even after fine-tuning. These errors often arise from confusion between acoustically similar sounds, including:

- Voicing errors, such as substituting the voiceless /s/ for the voiced /z/, or /k/ for /g/. These substitutions may occur because voicing contrasts can be difficult for the model to pick up on, especially in shorter or less clearly enunciated utterances.

- Vowel shifts, like substituting "EH1" for "AE1" (as in the word *"leg"*). These substitutions may reflect actual dialectal or idiosyncratic variation in vowel production, which the model does not generalize across during prediction.

- Place of articulation errors, such as substituting the alveolar nasal /n/ for the velar nasal /ŋ/. This is likely because the model overpredicts /n/ due to its higher frequency in the training data compared to /ŋ/, particularly in contexts where /ŋ/ is less clearly enunciated.

These common substitutions also help explain the especially poor performance observed for specific words such as *"thumb"* and *"zoo"*. The "TH" → "F" substitution is frequent across both groups and contributes to consistent errors in recognizing *"thumb"*. Interestingly, this error is more prevalent in adult speech, which may explain why the model often predicts words beginning with /f/ for adult productions of *"thumb"*. Similarly, for *"zoo"*, frequent substitutions in both the onset consonant ("Z" → "S") and the vowel ("UW1" → "OW1") are observed. These systematic errors likely account for the increased average edit distance for this word in both datasets.

Interim Discussion

In this section, we examine the performance of the wav2vec2-base-960h ASR model on child speech versus adult speech. As the model consistently underperforms on child speech, we explore fine-tuning it on child speech data to identify the optimal learning rate for improving recognition accuracy. Using the best-performing fine-tuned model, we then compare ASR output on both child and adult speech to investigate systematic error patterns. This analysis aims to determine whether these patterns reflect age-related variability in speech or reveal broader limitations of the model itself.

Fine-tuning efforts resulted in minimal improvement, even with examining different learning rates. The model that performs the best on child speech is trained at a learning rate of 10^{-5}, however fine-tuned model performance only does slightly better than baseline model performance. There are a number of reasons why this fine-tuning procedure did not produce particularly robust results. First, the child speech training dataset is very small and limited to only single-word utterances. Different results will likely be obtained by training on much more data, and examining effects in multiple-word utterances, like the model's pretraining data, will likely show better overall performance. Furthermore, the training loop is only executed for one epoch over the training dataset. While this was useful for quick, exploratory evaluation of different models, better learning will likely be observed if the model is trained over multiple epochs.

Error analysis of the fine-tuned wav2vec2-base-960h ASR model offers valuable insights into the extent to which prediction errors can be systematically attributed to speaker age, or alternatively reflects broader limitations of the model itself. Despite additional training on child speech, the model continues to exhibit a performance gap, with consistently higher accuracy (i.e. lower average edit distances) for adult speech. Moreover, recognition of adult speech exhibits more systematic and patterned errors while the model's performance on child speech remains highly variable, often producing a wider range of unique or inconsistent predictions. These findings suggest that, even after fine-tuning, the model continues to struggle with the inherent variability of child speech.

Notably, certain target words such as *"thumb"* and *"zoo"* are frequently mispredicted across both speaker groups. Closer examination reveals that the phones in these words are often substituted with more commonly-occurring phones, likely as a result of limited representation of certain phonological patterns—such as onset /θ/ and /z/—in the training data. Moreover, the most frequent substitutions appear at similar rates across adults and children. This indicates that these errors are not age-specific but instead reflect broader model limitations, highlighting its inability to recognize the subtle differences that distinguish acoustically-similar phones.

These findings demonstrate that the wav2vec2-base-960h ASR model continues to underperform on child speech compared to adult speech, even after fine-tuning. The relatively low average edit distances observed in the fine-tuned model highlight the need for more extensive training and a larger, more representative dataset to achieve robust performance. While the model produces more consistent outputs for adult speech, analysis of common phone-level errors indicates that many of its limitations are not age-related. Rather, the model seems to be susceptible to frequency-related biases arising from the uneven representation of certain phones and phonological patterns in the training data, leading to errors that persist across both child and adult speech.

Adapting Montreal Forced Aligner for child speech

Overview

The Montreal Forced Aligner (MFA) is a non-neural, hidden Markov model for forced align of words or phones, given an acoustic model for the language and a pronunciation dictionary (McAuliffe et al, 2017). The acoustic model component maps what acoustic features correspond to which phonetic features, or *phones*, given previous and following states; the dictionary component maps what words contain which phones (*MFA documentation: User guide: Glossary*). Given these two components and audio input labeled with the relevant words, MFA outputs TextGrid files that contain the model's timestamped estimation of the boundaries between the words or phones in the audio. MFA does not itself automatically recognize unlabeled audio.

The overarching goal for this portion of the project is to investigate how adjustments to the two MFA components can improve forced alignment for child speech. A sub-goal aims to create a relatively re-usable pipeline for future human alignment work. Accurately aligning boundaries by hand is intensely time-consuming work; adjusting boundaries that a model has already identified is less so. Therefore the investigated methods in this portion of the project especially avoid tasks that would not generalize well, such as the creation of an entirely new acoustic model over this dataset.

Methods

Both components of the essential MFA architecture can be adapted or modified to work with a given dataset. MFA includes utilities to adapt an existing acoustic model to a given corpus with a specific dictionary, adjusting the acoustic model's mappings of acoustic features to phones to better represent that corpus.

However, re-training MFA's acoustic model may not *fully* account for the variation in production of words in that corpus. While MFA performs speaker adaptation by default, this involves feature transforms that better reflect variation from individual speakers (*MFA documentation: Concepts in MFA: Speaker adaptation*) rather than adaptations that change what phones are expected in pronouncing words. In the case of child speech, which frequently involves relatively dramatic changes such as deletion, insertion, or transformation of phones, speaker adaptation will not cover the extent of expected changes in production. MFA is noted to struggle with cases that involve the outright deletion of phones, and improves performance when a dictionary is augmented with pronunciations that handle deletions present in specific datasets (*English MFA dictionary 3.1.0: Performance Factors*).

The dictionaries available to download through MFA generally contain an extensive list of words, but only include one set of phones; this amounts to one expected and adult-like pronunciation per word. The dictionary component, being a simple text file of words and their phones, can be created from scratch given knowledge of what words are in the relevant corpus. MFA can itself handle more than one pronunciation per word, and also includes utilities to apply probabilities to different pronunciations of the same orthographic representation of a word. The latter function does not require knowledge of the specific probability beforehand, and is calculated by MFA automatically over a given corpus.

The adjustments to MFA selected for investigation are the adaptation of an acoustic model to a corpus, the creation of a dictionary specific to the corpus, and the application of probabilities to that customized dictionary.

This project works with the ARPABET set of phones. While the International Phonetic Alphabet, or IPA, representation of phones is usually preferred in linguistic work, MFA's use of IPA is self-described as "opinionated" in its documentation (*MFA documentation: Pronunciation dictionaries: MFA IPA phone set*). Selecting ARPABET instead allows the use of a typical English font in the output TextGrids and prevents an extra step in reconciling the model's opinionated use of IPA with human transcribers' use of IPA. ARPABET and IPA representations are not mutually exclusive; while ARPABET tends toward "broader" phonetic transcription and captures less variability in sound production, ARPABET can still be translated into IPA for future work.

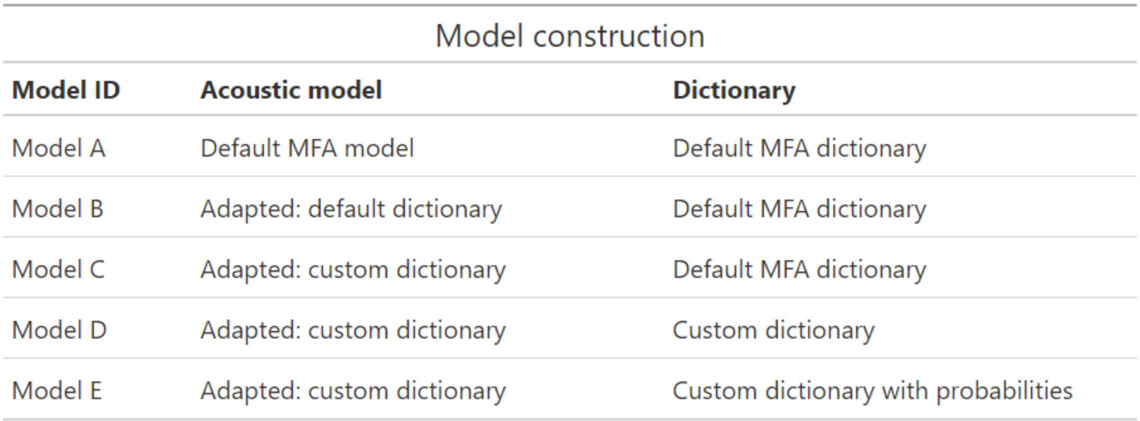
Dictionary creation began with transcription of the 24 corpus words into ARPABET. To select which production errors would be represented in the dictionary as alternate pronunciations, error types and rates were tabulated from the previously existing human transcriptions of the data. While this type of extensive human transcription would likely not exist for consultation during future work, knowledge of the error rates for alternate pronunciations used in this project may inform decisions about what to include in future dictionary creation.

Only the errors in training data were tabulated; the entire corpus was not considered when calculating error rates, in accordance with the spirit of the "train-test split". The customized dictionary therefore does not represent any data that should remain unknown to any trained or adapted models at alignment time.

Alternate pronunciations for a word were considered if pronunciations of that word erred at least 30% of the time in the training data (with an error defined as 'any error', rather than one specific misproduction): a relatively arbitrary threshold that separated approximately eleven words out of the twenty-four. A specific error was chosen for representation in the dictionary if that error occurred majority of times, out of all the errors made when pronouncing that word. Most words that involved at least a 30% error rate only involved one major error; other types of errors often only occurred in one or two utterances out of the fifty-four possible data points. Because most words with 30% error rates only involved one common error, only one alternate pronunciation per word was considered for this project. A word is therefore a candidate for being given an alternate pronunciation when the overall error rate is at least 30%; a specific error is chosen as that alternate pronunciation when it is the most common error made when producing that word.

The resulting dictionary includes one alternate pronunciation for eleven of the twenty-four words in the corpus. Nine of these alternate pronunciations are the liquid sounds “R” or “L” transformed into the glide “W”, covering the words “lamp”, “leg”, “lock”, “rainbow”, “ramp”, “red”, “ring”, “rock”, and “run”. The remaining two alternate pronunciations are initial alveolar fricative errors: “TH” in the word “thumb” is transformed into the labiodental fricative “F”, and “Z” in the word “zoo” is transformed into “S”.

This resulted in the creation of five overall models. Model A, the “baseline” model, is composed of the “default” English ARPABET acoustic model for MFA and the matching “default” English ARPABET dictionary. Model B uses the acoustic model adaptation feature, using the default English ARPABET dictionary to adapt the default English ARPABET acoustic model to produce an adapted model, with that same default dictionary used during alignment. Model C similarly adapts the acoustic model, but uses the customized dictionary for adaptation while using the default dictionary during alignment. Model D uses the customized dictionary both for adaptation and during alignment. Finally, Model E uses the customized dictionary for adaptation, and during alignment uses a version of the customized dictionary with probabilities applied.



Model construction

Model ID	Acoustic model	Dictionary
Model A	Default MFA model	Default MFA dictionary
Model B	Adapted: default dictionary	Default MFA dictionary
Model C	Adapted: custom dictionary	Default MFA dictionary
Model D	Adapted: custom dictionary	Custom dictionary
Model E	Adapted: custom dictionary	Custom dictionary with probabilities

Table 2. Montreal Forced Aligner model configurations. Note that while the adaptation of MFA acoustic models requires a provided dictionary, this dictionary need not be the same as the dictionary used at alignment.

Results

For evaluation of the forced alignment task, the evaluation set was manually aligned by two trained human transcribers. Each transcriber is randomly assigned to a set of 12 of the 18 participants in the evaluation set (approximately 280 individual audio files for each transcriber). Of the 12 participants assigned to each transcriber, each age-sex group is represented equally. Inter-rater reliability was then measured using the overlapping six participants (144 files) that both transcribers were tasked to adjust.

One participant aligned by coder 2 featured intensely unusual times, sometimes taking over ten seconds to speak a single word according to annotated boundaries. This participant was not one of the six randomly selected as a metric for inter-coder reliability, and so coder 1's alignments could not be consulted for reference. The sound files themselves do not reflect this extremely slow timing, and models aligned this participant without issue. Because the human alignments were so unusual, this participant was dropped as an outlier.

After processing silence intervals and dropping the outlier participant, the calculated inter-rater reliability is 0.719; that is, 71.9% of boundaries aligned by the two human coders were within 20ms of each other.

Models' performance is generally evaluated by calculation of the difference in milliseconds between the model-aligned and human-aligned boundaries. Differences are calculated between successive intervals regardless of label; i.e. a model's first interval is compared to the human's first interval, whether the labels assigned to that interval matched or not. Intervals representing leading (beginning of utterance) and trailing (end of utterance) silences were filtered out for these calculations to account for cases where human coders annotated silences and models did not. In cases of true interval addition or subtraction, where one aligner has a labeled interval where the other does not, the difference is calculated against the previous relevant interval. This results in an effective penalty of the entire interval's duration to that aligner.

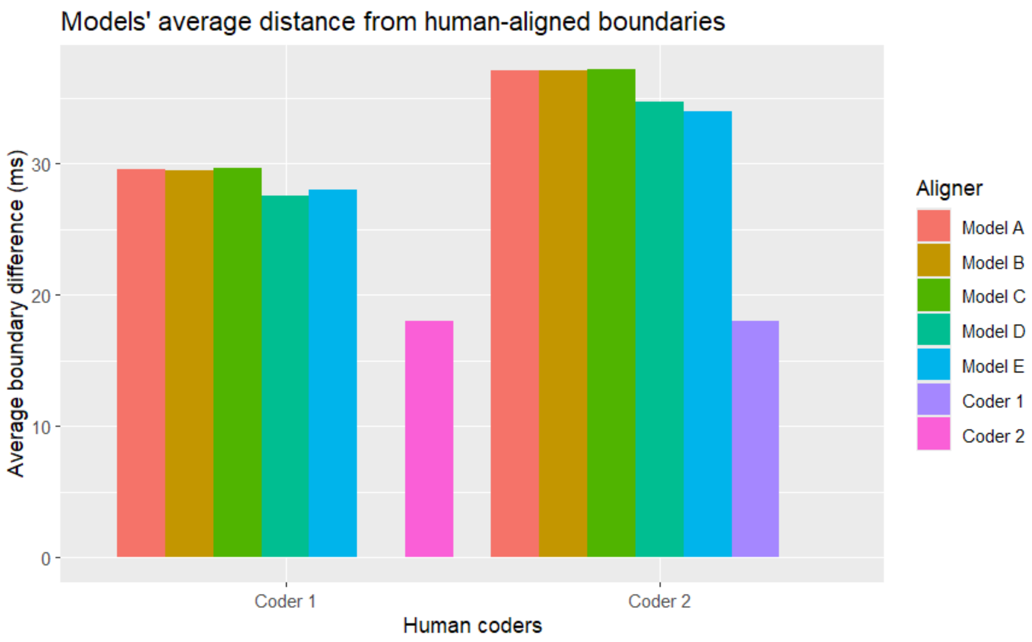
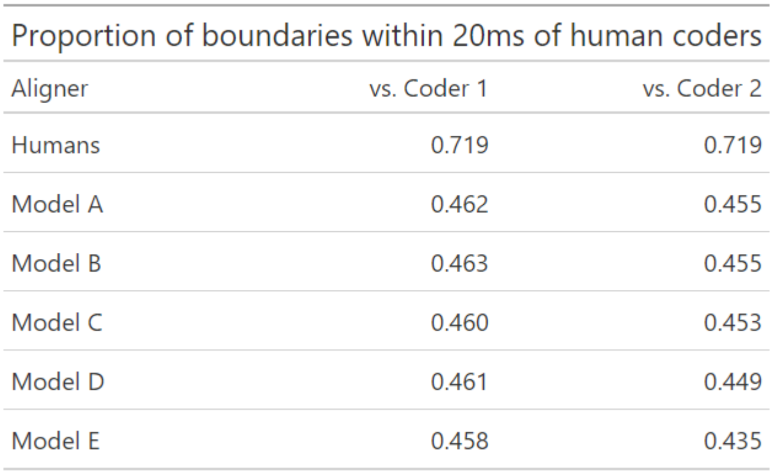


Figure 6. Average distance, in milliseconds, from human-aligned boundaries. Human coders averaged 18ms from each other; models averaged a distance of 29ms against coder 1 and 37ms against coder 2.

Models A through C performed extremely similarly in terms of average millisecond difference from the human-aligned boundaries, varying only tenths of a millisecond from each other.



Proportion of boundaries within 20ms of human coders

Aligner	vs. Coder 1	vs. Coder 2
Humans	0.719	0.719
Model A	0.462	0.455
Model B	0.463	0.455
Model C	0.460	0.453
Model D	0.461	0.449
Model E	0.458	0.435

Table 3. Proportion of boundaries within 20ms of the human coders.

While all models tended to have shorter average distances from coder 1's alignments compared to coder 2's alignments, models tended to add more intervals where coder 1 did not have them, compared to coder 2:

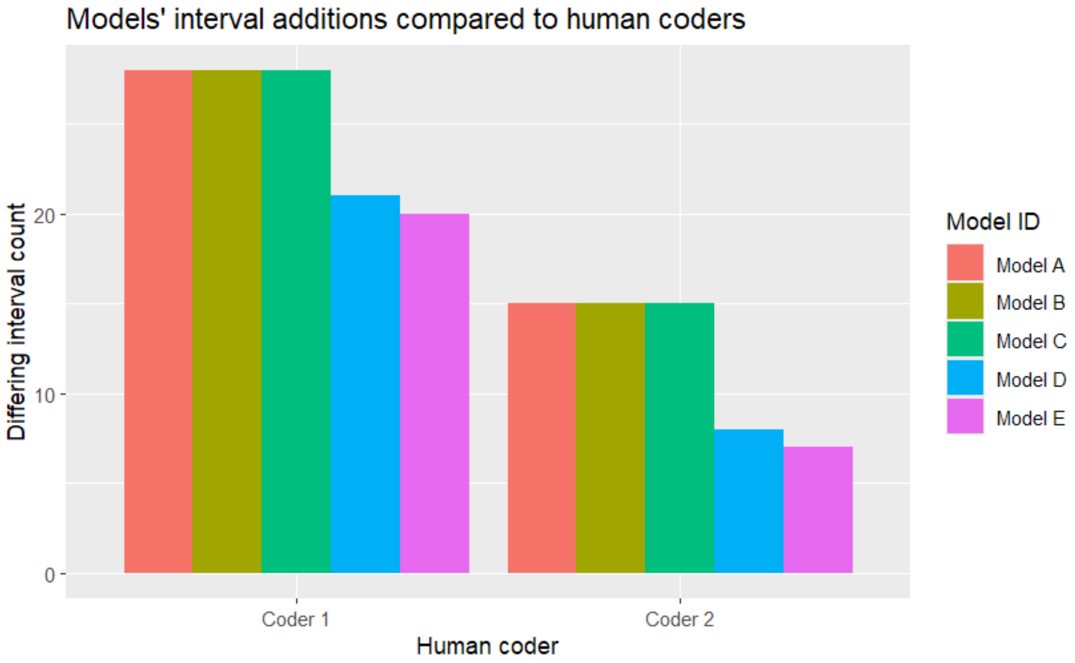


Figure 7. Counts of intervals where models had more intervals for a given utterance compared to human alignment. For comparison, coder 1 aligned 835 total intervals; coder 2 aligned 844 total intervals.

Models D and E added only 8 and 7 intervals, respectively, compared to coder 2. Out of the 844 intervals in common between the models and coder 2, the additional intervals represent less than one percent of the intervals.

All models tended to have fewer intervals compared to coder 2 than compared to coder 1:

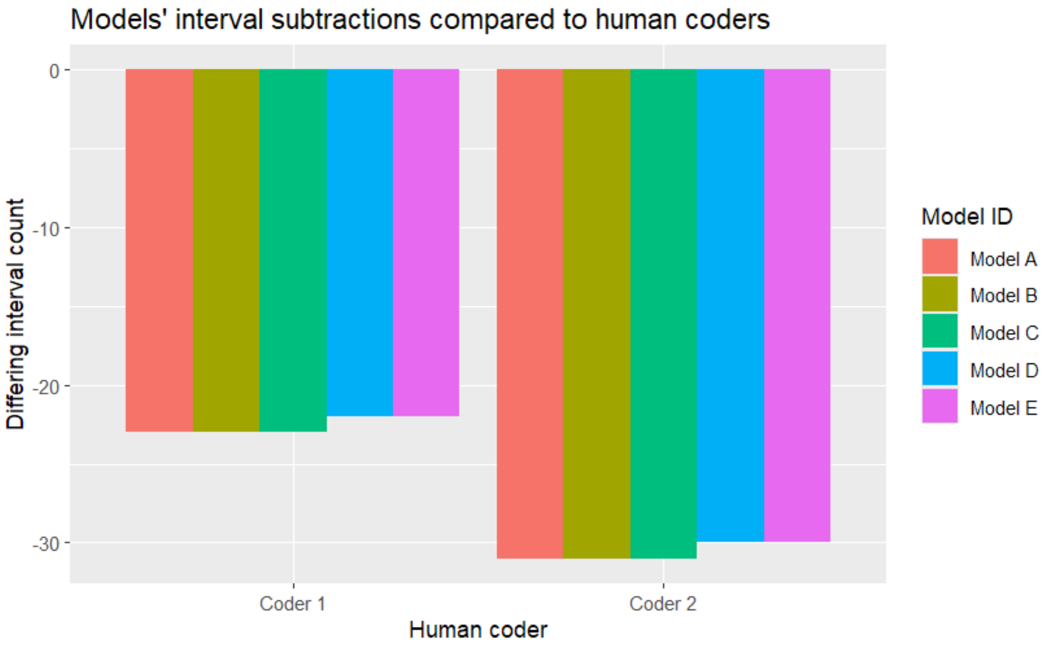


Figure 8 . Counts of intervals where models had fewer intervals for a given utterance compared to human alignment. For comparison, coder 1 aligned 835 total intervals; coder 2 aligned 844 total intervals.

A more thorough visualization of model performance relative to human alignment focuses again on the differences between model-aligned and human-aligned boundaries. The percentage of boundary differences that falls within the millisecond threshold on the x-axis is visualized by the following empirical cumulative distribution plots. A “perfect” agreement would be represented by a line drawing a steep 90° angle on the top left of the plot, indicating 100% agreement at a low millisecond threshold. Figure 9 visualizes the agreement between the two human coders; figures 10a and 10b visualize model performance vs. coder 1, and figures 11a and 11b visualize model performance vs. coder 2.

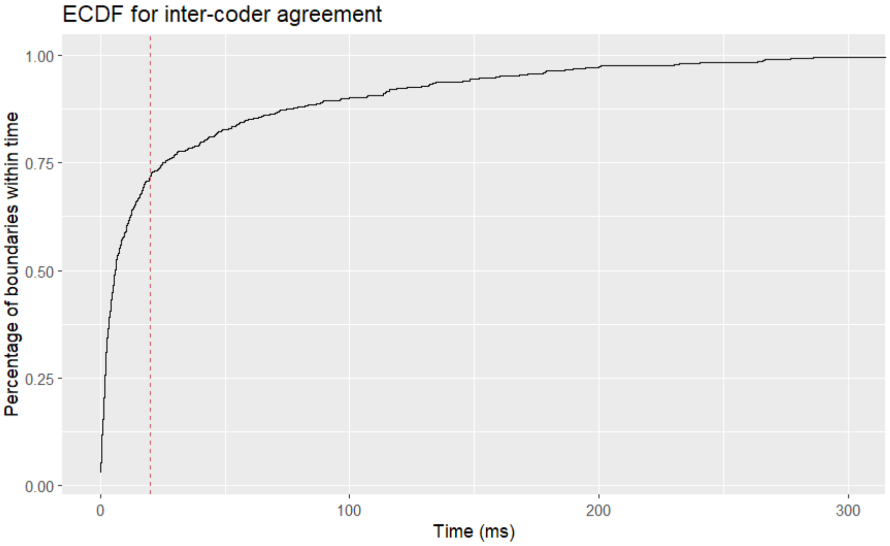
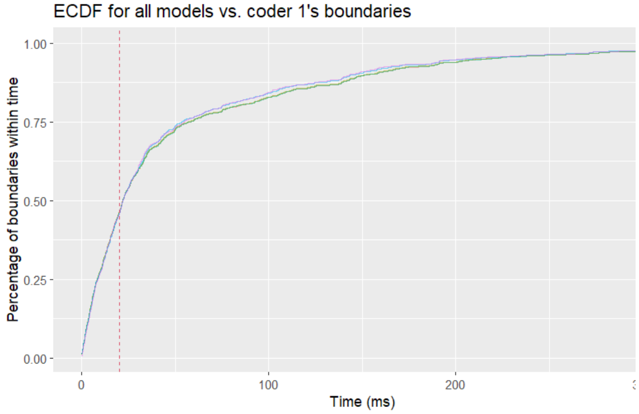
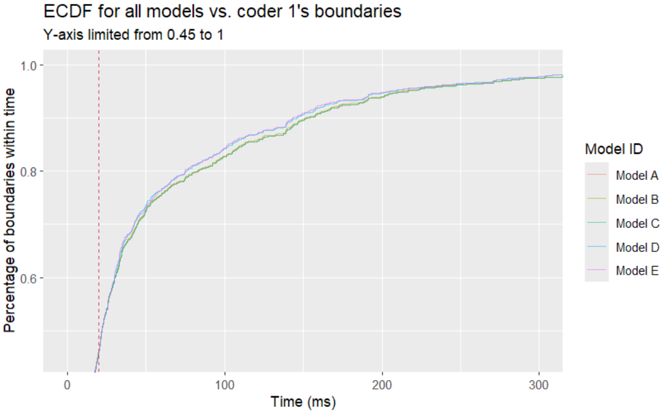
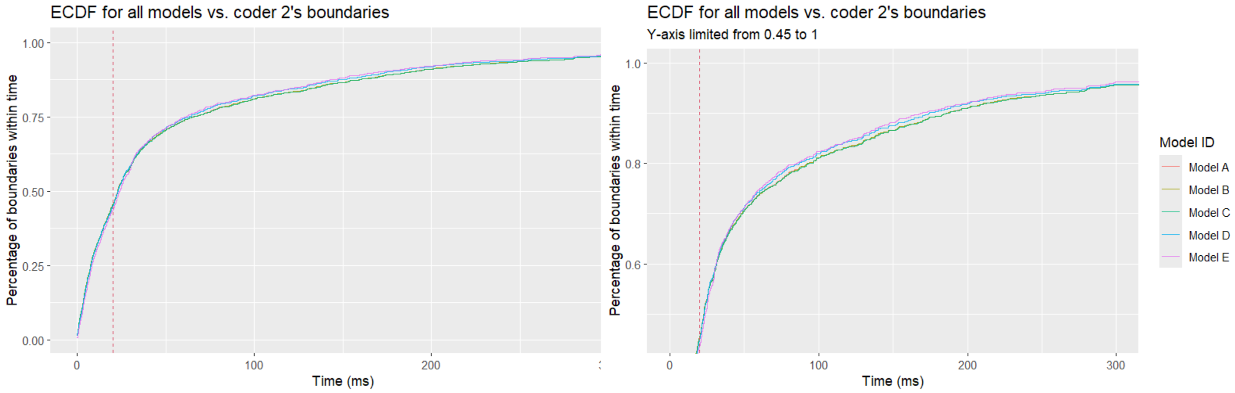


Figure 9. Empirical cumulative distribution plot for human inter-coder agreement. The dotted reference line at 20ms represents the threshold used above, at which coders achieve 71.9% agreement. Human coders approach 100% agreement by 300ms.



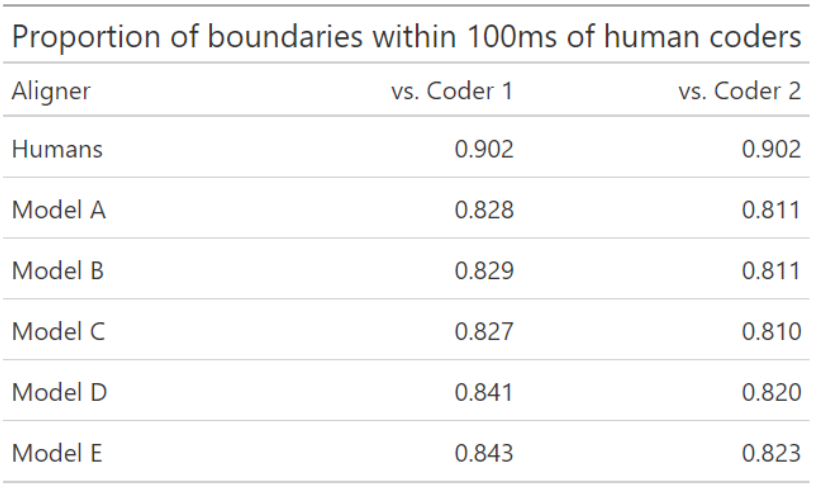


Figures 10*a* and 10*b*. Empirical cumulative distribution plot for all models versus coder 1's boundaries. Reference line at the 20ms threshold. Figure 10*a* shows the entire plot, with the y-axis from 0 to 1; figure 10*b* shows a zoomed-in view of the plot, including both the point where the reference line crosses all models' curves and a better view of where the models differ most, from 50ms to 200ms. Models A through C perform so similarly that they essentially draw a single visible line; models D and E group together in the same way.



Figures 11*a* and 11*b*. Empirical cumulative distribution plot for all models versus coder 2's boundaries. Reference line at the 20ms threshold. Figure 11*a* shows the entire plot, with the y-axis from 0 to 1; figure 11*b* shows a zoomed-in view of the plot, including both the point where the reference line crosses all models' curves and a better view of where the models differ most, from 50ms to 200ms. Models A through C again perform so similarly that they essentially draw a single visible line; models D and E group together in the same way.

From these figures, it is visually evident that model performance varies more after the 20ms threshold, beginning at approximately 100ms. Model performance at 100ms is shown in table 4:



Proportion of boundaries within 100ms of human coders

Aligner	vs. Coder 1	vs. Coder 2
Humans	0.902	0.902
Model A	0.828	0.811
Model B	0.829	0.811
Model C	0.827	0.810
Model D	0.841	0.820
Model E	0.843	0.823

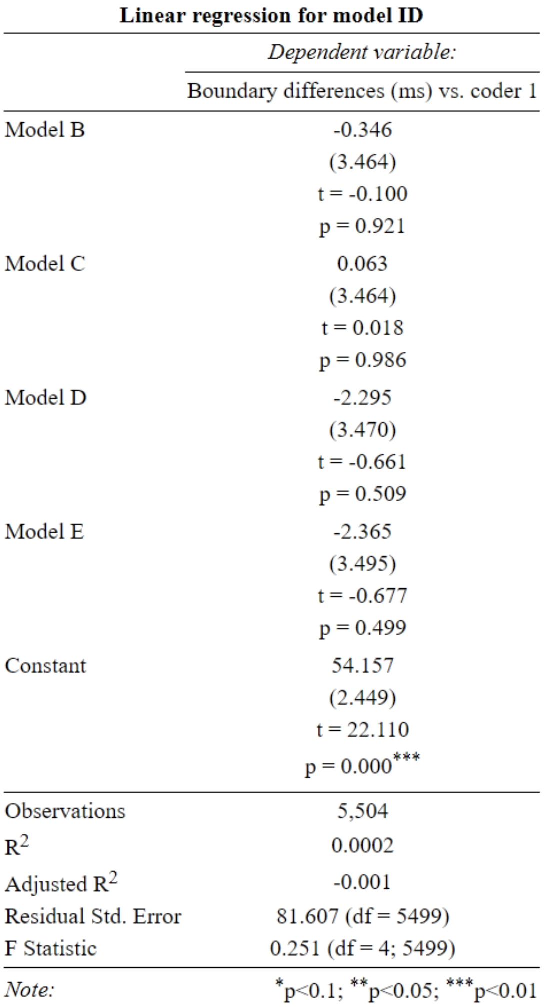
Table 4. Proportion of model boundaries within 100 milliseconds of human coders' boundaries.

We then fit linear regressions to analyze whether the differences in model performance are statistically significant.

Linear regression analyses

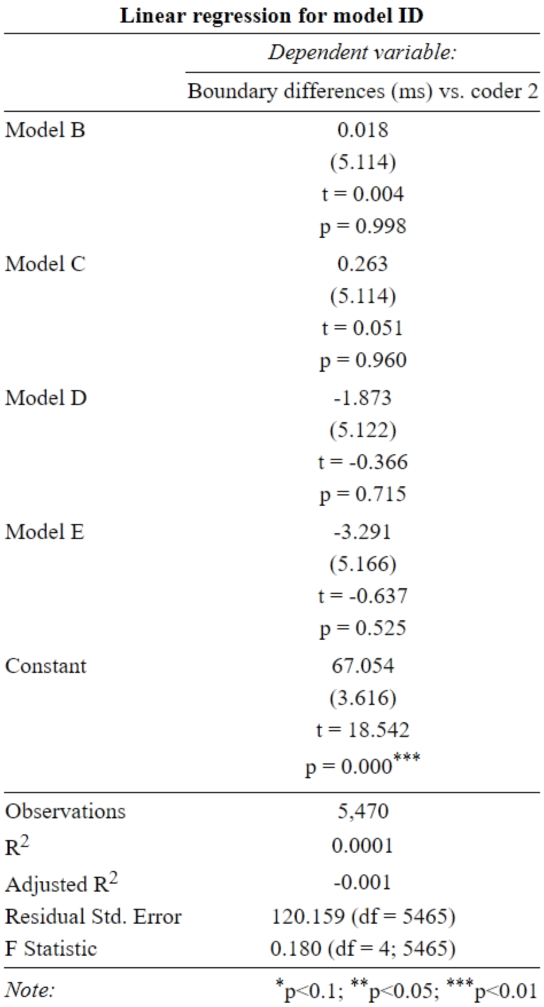
The performance of the different MFA models is measured above by the difference in milliseconds of the model-aligned boundaries and human-aligned boundaries. If models perform significantly differently from each other, there should be different patterns in the boundary differences, such that a linear regression with the boundary differences as a response variable can identify significance. Linear regression with acoustic model types or dictionary types as covariates may also identify significant differences in model performance fueled by the relevant adaptations, if such patterns exist.

Linear regressions were fit separately for boundary differences as compared to coder 1 and coder 2. First, the regressions were fit with only the model identity as covariates. Model A, the model utilizing both the default MFA acoustic model and default dictionary, was set as the reference level, represented by the constant.



Linear regression for model ID

	Dependent variable:
	Boundary differences (ms) vs. coder 1
Model B	-0.346
	(3.464)
	t = -0.100
	p = 0.921
Model C	0.063
	(3.464)
	t = 0.018
	p = 0.986
Model D	-2.295
	(3.470)
	t = -0.661
	p = 0.509
Model E	-2.365
	(3.495)
	t = -0.677
	p = 0.499
Constant	54.157
	(2.449)
	t = 22.110
	p = 0.000***
Observations	5,504
R^2	0.0002
Adjusted R^2	-0.001
Residual Std. Error	81.607 (df = 5499)
F Statistic	0.251 (df = 4; 5499)
Note:	*p<0.1; **p<0.05; ***p<0.01



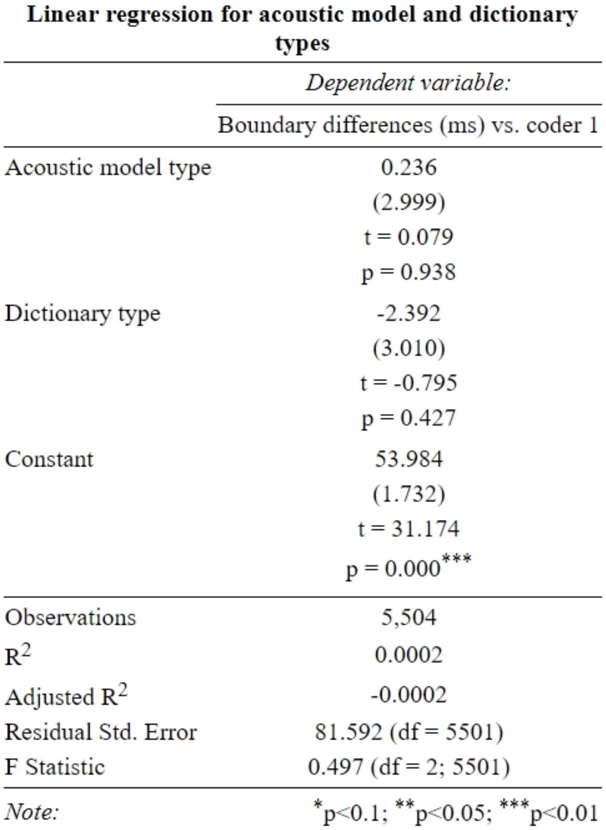
Linear regression for model ID

	Dependent variable:
	Boundary differences (ms) vs. coder 2
Model B	0.018
	(5.114)
	t = 0.004
	p = 0.998
Model C	0.263
	(5.114)
	t = 0.051
	p = 0.960
Model D	-1.873
	(5.122)
	t = -0.366
	p = 0.715
Model E	-3.291
	(5.166)
	t = -0.637
	p = 0.525
Constant	67.054
	(3.616)
	t = 18.542
	p = 0.000***
Observations	5,470
R^2	0.0001
Adjusted R^2	-0.001
Residual Std. Error	120.159 (df = 5465)
F Statistic	0.180 (df = 4; 5465)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5 & **table 6**. Linear regression results of boundary differences by model ID. Table 5 shows results of the regression for the boundaries aligned by coder 1; table 6 shows results for boundaries aligned by coder 2. Standard errors are given in parentheses below coefficients. No significant results were found by either regression.

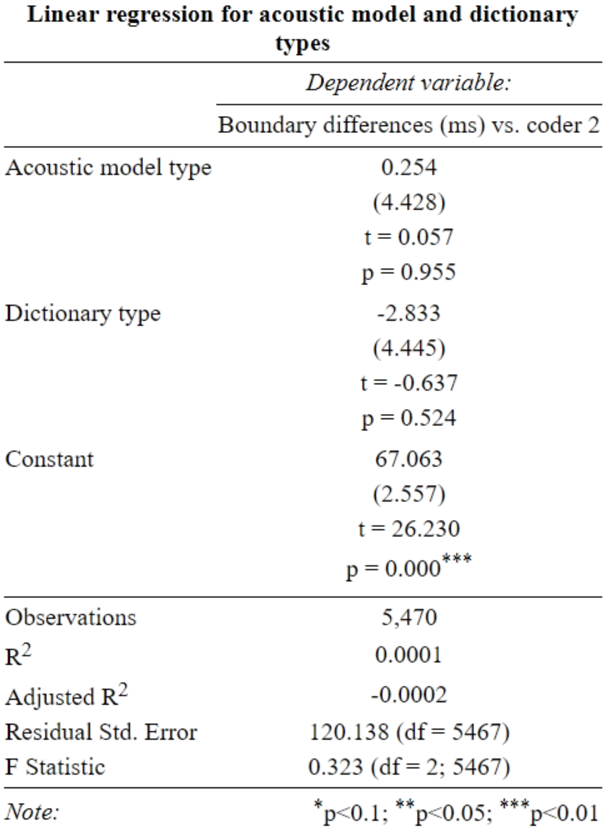
The regression identified no significant differences by model ID. (As an intercept, the constant is near-always identified as significant, and is not indicative of significant results for Model A.)

Next, linear regressions were fit with the acoustic model type and dictionary type as covariates, again fit separately for boundaries aligned by coder 1 and coder 2:



Linear regression for acoustic model and dictionary types

	Dependent variable:
	Boundary differences (ms) vs. coder 1
Acoustic model type	0.236
	(2.999)
	t = 0.079
	p = 0.938
Dictionary type	-2.392
	(3.010)
	t = -0.795
	p = 0.427
Constant	53.984
	(1.732)
	t = 31.174
	$p = 0.000^{***}$
Observations	5,504
R^2	0.0002
Adjusted R^2	-0.0002
Residual Std. Error	81.592 (df = 5501)
F Statistic	0.497 (df = 2; 5501)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$



Linear regression for acoustic model and dictionary types

	Dependent variable:
	Boundary differences (ms) vs. coder 2
Acoustic model type	0.254
	(4.428)
	t = 0.057
	p = 0.955
Dictionary type	-2.833
	(4.445)
	t = -0.637
	p = 0.524
Constant	67.063
	(2.557)
	t = 26.230
	$p = 0.000^{***}$
Observations	5,470
R^2	0.0001
Adjusted R^2	-0.0002
Residual Std. Error	120.138 (df = 5467)
F Statistic	0.323 (df = 2; 5467)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 7 & **table 8**. Linear regression results of boundary differences by acoustic model type and dictionary type. Table 7 shows results of the regression for the boundaries aligned by coder 1; table 8 shows results for boundaries aligned by coder 2. Standard errors are given in parentheses below coefficients. No significant results were found by either regression.

These regressions also did not identify any significant differences, indicating that despite the slight visual differences in the cumulative distribution plots above, these differences are not statistically relevant.

Interim Discussion

Adaptation of MFA's model components resulted in minimal differences. Despite some differences in performance when visualized with cumulative distribution plots, these differences were not statistically significant.

The slight difference in performance between adaptations of the acoustic model and customization of the dictionary indicates some direction for future work in this area. Models D and E visually distinguish themselves from models A through C on the cumulative distribution plots, and dictionary type as a predictor in linear regression finds somewhat more of a pattern than acoustic model type; editing dictionary entries to provide alternate pronunciations relevant to child speech shows slightly more promise than adapting MFA's acoustic models to specific corpora.

A significant drawback of customizing dictionaries with alternate pronunciations is the inability to scale to large datasets. This project utilized a corpus with twenty-four known words that had already been thoroughly transcribed; it was therefore feasible to tabulate error types and rates to create a dictionary specifically for this corpus. Larger datasets that include less strictly regulated utterances would require much more work to identify words with relevant errors and add them to a usable dictionary. However, as the errors included in our dictionary as alternate pronunciations represent error types that are known to be common in child speech, it may be possible to algorithmically identify words with phones that are frequently misarticulated, substituted, or deleted, and create relevant dictionary entries.

While non-neural architectures are being left behind in automatic speech recognition, models like MFA may still have some advantages over neural models in research settings. MFA has a much lower computational cost and a faster runtime, allowing its use on older laboratory machines that may not have the storage space or computational power for local neural network models. In scenarios where the words in a corpus are known, alignment may be made much faster by first running the corpus through MFA, then asking humans to adjust the boundaries, rather than aligning the corpus by hand entirely. Non-neural models, and adaptation of non-neural models to better recognize child speech, may still therefore have some place in academia.

General Discussion

While the current state-of-the-art ASR models perform exceptionally well on recognizing adult speech, they often fail to achieve the same robustness for child speech due to its inherent variability. Thus, the goal of this project is to examine methods to adapt ASR models to the domain of child speech. We first explore fine-tuning on the neural ASR model wav2vec2-base-960h and examine patterns of errors that that model makes, in a broader effort to reveal what phones or phonological patterns that are particularly challenging for the ASR model to capture. We then turn to the forced alignment—a downstream task of ASR—to examine how the acoustic model and dictionary components of non-neural models can be adapted to better align child speech.

Fine-tuning the wav2vec2-base-960h ASR model to improve performance on child speech proved challenging, largely due to the limited size of the training dataset and the need for more extensive training. Despite these efforts, a substantial performance gap between child and adult speech remained, indicating that significantly more data and training time are required to achieve more robust performance for child speech. While the phone-level error analysis did not fully capture the acoustic nuances that differentiate child speech from adult speech, the findings underscored a key limitation of the model itself. Susceptibility to frequency biases in the training data—driven by the overrepresentation of certain phones or phonological patterns—affects predictions across both age groups, pointing to a broader constraint of the model rather than an age-related shortcoming.

Forced alignment efforts resulted in similar shortcomings. While dictionary adaptation shows some promise for future work, neither acoustic model adaptation nor dictionary adaptation resulted in marked improvement over the default acoustic model or dictionary for forced alignment of child speech.

Overall, neither ASR or forced alignment model adaptation efforts demonstrated particularly significant improvement for child speech recognition, emphasizing the need for more extensive and sophisticated methods than investigated in the current exploration.

References

Bhardwaj, V., Ben Othman, M. T., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B. S., Rehman, A. U., Shafiq, M., & Hamam, H. (2022). Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review. *Applied Sciences*, *12*(9), 4419. https://doi.org/10.3390/app12094419

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

Beckman, M. E., Plummer, A. R., Munson, B., & Reidy, P. F. (2017). Methods for eliciting, annotating, and analyzing databases for child speech development. *Computer Speech and Language*, 45, 278–299. https://doi.org/10.1016/j.csl.2017.02.010

Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

Knowles, T., Clayards, M., Sonderegger, M., Wagner, M., Nadig, A., and Kristine H. Onishi; Automatic forced alignment on child speech: Directions for improvement. *Proc. Mtgs. Acoust.* 2 November 2015; 25 (1): 060001. https://doi.org/10.1121/2.0000125

Liu, D., & Xiong, J. (2024). FASA: a Flexible and Automatic Speech Aligner for Extracting High-quality Aligned Children Speech Data. *arXiv preprint arXiv:2406.17926*.

Mahr, T.J., Berisha V., Kawabata K., Liss J., Hustad K.C. Performance of Forced-Alignment Algorithms on Children's Speech. J Speech Lang Hear Res. 2021 Jun 18;64(6S):2213-2222. doi: 10.1044/2020_JSLHR-20-00268. Epub 2021 Mar 11. PMID: 33705675; PMCID: PMC8740721.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Conference of the International Speech Communication Association*.

McAuliffe, M. Release notes for the English MFA Dictionary 3.1.0. Accessed 12 June 2025. https://github.com/MontrealCorpusTools/mfa-models/releases/tag/dictionary-english_mfa-v3.1.0

McGonigle, E., Vandam, M., Wilkinson, C., & Johnson, K. T. (2024). Benchmarking Automatic Speech Recognition Technology for Natural Language Samples of Children With and Without Developmental Delays. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. https://doi.org/10.1109/EMBC53108.2024.10782773.

Montreal Forced Aligner documentation. Accessed 12 June 2025. https://montreal-forced-aligner.readthedocs.io/en/latest/index.html

Nayeem, M. T., & Rafiei, D. (2024). KidLM: Advancing Language Models for Children--Early Insights and Future Directions. *arXiv preprint arXiv:2410.03884*.

Park, K. & Kim, J. (2019). *g2pE* (Version 2.1.0). [GitHub repository]. GitHub. https://github.com/Kyubyong/g2p?tab=readme-ov-file.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. *International conference on machine learning* (pp. 28492-28518). PMLR. https://doi.org/10.48550/arXiv.2212.04356.

Shivakumar, P. G., & Narayanan, S. (2022). End-to-end neural systems for automatic children speech recognition: An empirical study. *Computer Speech & Language*, *72*, 101289. https://doi.org/10.1016/j.csl.2021.101289.

Zhu, J., Zhang, C., & Jurgens, D. (2022, May). Phone-to-audio alignment without text: A semi-supervised approach. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8167-8171). IEEE.